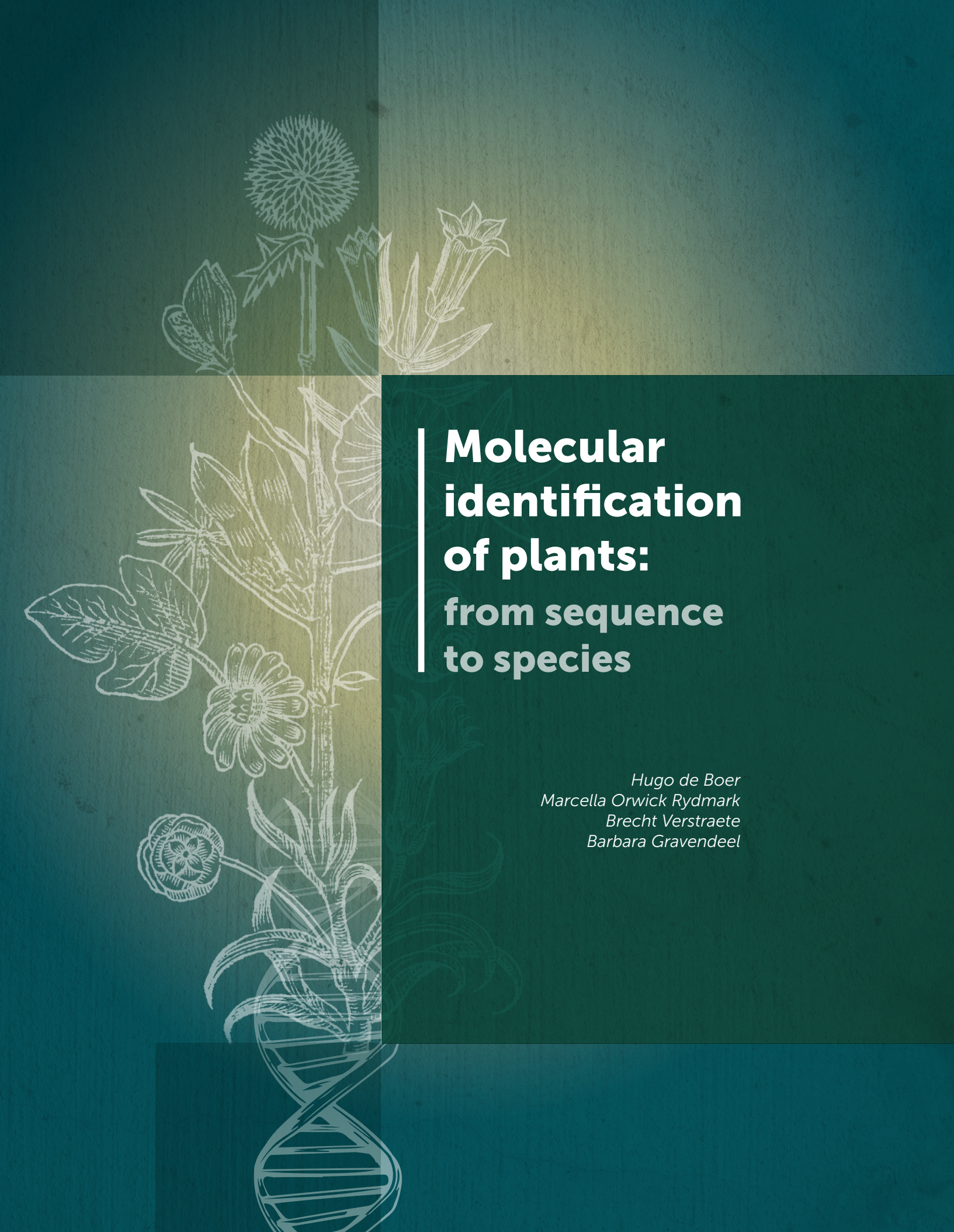

Molecular identification of plants:
from sequence to species

Hugo de Boer
Marcella Orwick Rydmark
Brecht Verstraete
Barbara Gravendeel

Molecular identification of plants: from sequence to species

Edited by
Hugo de Boer, Marcella Orwick Rydmark, Brecht Verstraete,
Barbara Gravendeel

Molecular identification of plants: from sequence to species
Edited by Hugo de Boer, Marcella Orwick Rydmark, Brecht Verstraete, Barbara Gravendeel

Cover design by Sven Bellanger

First published: 2022
eISBN: 978-619-248-091-2
DOI: 10.3897/ab.e98875

PENSOFT Publishers
12, Prof. Georgi Zlatarski Str. 12
1111 Sofia, Bulgaria
www.pensoft.net

Plant.ID

The publication of the book is a deliverable of the H2020 MSCA-ITN-ETN "Plant.ID", which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 765000.

Table of contents

Foreword 4
Introduction 5

Section 1: Design, sampling, and substrates 9

Chapter 1: DNA from plant tissue 10
Chapter 2: DNA from museum collections 29
Chapter 3: DNA from water 43
Chapter 4: DNA from soil 57
Chapter 5: DNA from pollen 68
Chapter 6: DNA from food and medicine 79
Chapter 7: DNA from faeces 89
Chapter 8: aDNA from sediments 103

Section 2: Methods 122

Chapter 9: Sequencing platforms 123
Chapter 10: DNA barcoding 137
Chapter 11: Amplicon metabarcoding 148
Chapter 12: Metagenomics 164
Chapter 13: DNA Barcoding - High Resolution Melting analysis (Bar-HRM) 178
Chapter 14: Target capture 190
Chapter 15: Transcriptomics 213
Chapter 16: Whole genome sequencing 233
Chapter 17: Species delimitation 245
Chapter 18: Sequence to species 270

Section 3: Applications 282

Chapter 19: Systematics and evolution 283
Chapter 20: Museomics 297
Chapter 21: Palaeobotany 308
Chapter 22: Healthcare 326
Chapter 23: Food safety 337
Chapter 24: Environmental and biodiversity assessments 354
Chapter 25: Wildlife trade 372
Chapter 26: Forensic genetics, botany, and palynology 387

Foreword

Alexandre Antonelli[1]

[1] Director of Science, Royal Botanical Gardens, Kew, United Kingdom

a.antonelli@kew.org

Names are the carriers of knowledge. Without names, much of science would be meaningless. Names give us insight into the diseases that affect our health; the objects that sustain our economies; the celestial bodies that travel in the Universe. Names solve ambiguity.

In botany, the name of a plant may provide the first clues as to its characteristics, also called traits. Is it edible, or poisonous? Beautiful, or ugly? While some traits are relative (edible by whom, ugly to whom?), others are absolute: thorny, succulent, epiphytic. Some are obvious, others elusive. From morphological descriptions and DNA sequences to historical accounts and traditional uses, they are all linked by the name.

Until recently, the reliable identification of plants was the task of a select few: the taxonomists. Today, this is less so. The molecular identification of plants through DNA barcodes has been shown to perform just as well, and in fact often better, than taxonomists for many taxa, particularly when specimens lack reproductive structures. Other techniques, such as image recognition through machine learning and the spectrophotometric signature of leaves, can yield similar results. Does this mean the demise of taxonomists is on the horizon?

Not at all. I believe it is very much the opposite: in the current environmental crisis, the need to document and protect the world's biodiversity has never been more acute. At the same time, some 20% of all plant species have not yet been scientifically described, and many of them may disappear even before we have identified and characterized them. The work of taxonomists remains therefore critical, but as molecular identification of species is underway and set to become routine across the private and public sectors, expert time can now be reallocated from bulk identifications to the training of students, build-up of physical and digital reference collections, and further development of identification methods. Technologies are here to help – not replace – taxonomy, by complementing the human strengths and compensating for some of our human weaknesses: an insufficient memory, a biased brain, and lack of time.

This book is for you who are curious about how plants can be identified using DNA: the most powerful source of information to link a plant to a name. This may sound trivial, but it is not. But don't despair in advance: it is doable, mostly fun, and always rewarding. You just need to learn how.

Here, you will not only learn how various types of materials containing plant fragments can be identified to species in the lab and how to execute sophisticated computer analyses, but also gain a deeper understanding of the complexities and challenges faced by taxonomy in general, and plant identification in particular, including the lack of comprehensive reference databases. Enforcing strict species concepts onto nature's inherent fluidity doesn't always work, and despite all recent advances in this field it still happens that some plant samples cannot be confidently named. Yet, if this ever happens to you, this initially frustrating insight can also be scientifically revealing, and help you design further experiments.

The applications of molecular identification are far more numerous and trans-disciplinary than most people would imagine. Several chapters take a deep dive at applications in fields as seemingly disparate as palaeobotany and healthcare, but as I argued at the start of this text, they are all unified by a common denominator: the name, the information-carrier.

I hope you will find this book as inspiring, informative, and revelatory as I have, and that you will choose to carry out your own projects using the molecular identification of plants. And if you do so, just don't forget to cite the chapters that inspired you!

Introduction

Hugo de Boer[1]
[1] Natural History Museum, University of Oslo, Norway
h.de.boer@nhm.uio.no

An estimated 340,000–390,000 vascular plant species are known to science (Lughadha et al. 2016; Govaerts et al. 2021), and on average an additional 2,000 species are described each year (IPNI, 2020). Many of these plant species are poorly known in terms of ecology, distribution, threats, and potential benefits. Less than 10% has been assessed for the IUCN Red List, with a strong bias towards trees and species that are considered to be threatened (Bachman et al. 2019). A study assessing a sample of a thousand species representing global plant diversity uncovered that more than one in five were threatened with extinction (Brummitt et al. 2015). Plant extinctions are shown to occur up to 500 times faster today than in pre-industrial times (Humphreys et al. 2019). We are currently in a situation where for a large number of plant species we are unaware that they are at risk of extinction because we know them so poorly. Although new species are continuously being described, at the same time, others are going extinct. Unfortunately, many more species are going extinct without us knowing about it or even having discovered them.

Organismal diversity is the foundation of all biological research, but species discovery and delimitation requires taxonomic skills. Even the most experienced taxonomists can rarely critically identify more than 0.01% of the estimated 10–15 million species (Hammond 1992; Hawksworth and Kalin-Arroyo 1995). The Convention on Biological Diversity (CBD) recognised this challenge at its 1992 Rio Earth Summit, and established the Global Taxonomy Initiative (GTI) a few years later at its 5th Conference of Parties (CBD COP5 1996). The GTI was created to reduce the taxonomic impediment and aims to advance taxonomy and address the lack of information and expertise. The taxonomic impediment consists of the knowledge gaps in our taxonomic system (including those associated with genetic systems), the shortage of trained taxonomists and curators, and the impact these deficiencies have on our ability to conserve, use, and share the benefits of our biological diversity. Achieving the Aichi Biodiversity Targets and the Sustainable Development Goals and contributing to the post-2020 Global Biodiversity Framework requires an acceleration of taxonomy beyond traditional morphology-based methods and further integration of DNA-based approaches.

The global scientific community lacks the expertise and continuity to identify all species diversity, and biodiversity is lost at a greater speed than we can discover and describe new taxa (Antonelli et al. 2020; Butchart et al. 2010; Dirzo and Raven 2003; Hooper et al. 2012). Species description is a rigorous and time consuming process that can be made more effective through open data sharing and integrative taxonomy (Riedel et al. 2013). Morphological species identification has four significant limitations as outlined by Hebert et al. (2003): (1) phenotypic plasticity and genetic variability in the characters employed for species recognition can lead to incorrect identifications; (2) morphologically cryptic taxa, which are common in many groups, can be overlooked (Burns et al. 2008; Jarman and Elliott 2000; Knowlton 1993; Ragupathy et al. 2009); (3) morphological keys are often effective only for a particular life stage or gender, and many individuals cannot be identified; (4) modern interactive keys represent a major advance, but the use of keys often demands such a high level of expertise that misdiagnoses are common.

DNA-based species identification, i.e., molecular identification, makes it possible to identify species precisely from trace fragments such as pollen (Bell et al. 2019; Hawkins et al. 2015), detecting substitution in herbal pharmaceuticals (Raclariu et al. 2018, 2017), authentication of sustainable tropical timber (Nithaniyal et al. 2014), monitoring invasive alien species (Armstrong and Ball 2005), uncovering illegal international trade in endangered species (de Boer et al. 2017; Ghorbani

et al. 2017), making rapid molecular biodiversity assessments (Bohmann et al. 2014; Thomsen and Willerslev 2015), and studying historical biodiversity through sedimentary DNA and ancient DNA (Anderson-Carpenter et al. 2011; Bálint et al. 2018).

These innovations in molecular identification enable us to detect and identify species in places and settings that were unimaginable only a few decades ago, or even in 2020 (Lynggaard et al. 2022). Molecular biodiversity assessments in fungi and insects especially have led to increasing numbers of "dark taxa", i.e., taxa detected from DNA sequences alone by lacking a physical reference and identity for morphological description (Chimeno et al. 2022; Hausmann et al. 2020; Ryberg and Nilsson 2018). Dark taxa pose a challenge for taxonomy (Page 2016), but also reveal how molecular biodiversity assessments can overtake and accelerate beyond traditional taxonomy. This acceleration of species detection and discovery is crucial to overcome our global taxonomic impediment and help systematics make a bigger contribution to the CBD post-2020 Global Biodiversity Framework. The actual description and giving of names to newly discovered taxa is dependent on traditional taxonomy, and should be done by combining morphology and DNA. However, when tropical rainforests – key ecosystems harboring mega diversity and unknown species – are lost at rates of millions of hectares annually (Brondizio et al. 2019), it is more important to rapidly assess the most crucial biodiversity to conserve than to put a name to each taxon. The current revolution in molecular identification will empower us to play a key role in identifying that biodiversity.

References

Anderson-Carpenter LL, McLachlan JS, Jackson ST, Kuch M, Lumibao CY, Poinar HN (2011) Ancient DNA from lake sediments: bridging the gap between paleoecology and genetics. BMC Evol. Biol. 11, 30. https://doi.org/10.1186/1471-2148-11-30

Antonelli A, Fry C, Smith RJ, Simmonds MSJ, Kersey PJ, Pritchard HW, Abbo MS, Acedo C, Adams J, Ainsworth AM, Allkin B, Annecke W, Bachman SP, Bacon K, Bárrios S, Barstow C, Battison A, Bell E, Bensusan K, Bidartondo MI, et al. (2020) State of the World's Plants and Fungi 2020. Royal Botanic Gardens, Kew. https://doi.org/10.34885/172

Armstrong KF, Ball SL (2005) DNA barcodes for biosecurity: invasive species identification. Philos. Trans. R. Soc. Lond. B Biol. Sci. 360, 1813–1823. https://doi.org/10.1098/rstb.2005.1713

Bachman SP, Field R, Reader T, Raimondo D, Donaldson J, Schatz GE, Lughadha EN (2019) Progress, challenges and opportunities for Red Listing. Biol. Conserv. 234, 45–55. https://doi.org/10.1016/j.biocon.2019.03.002

Bálint M, Pfenninger M, Grossart H-P, Taberlet P, Vellend M, Leibold MA, Englund G, Bowler D (2018) Environmental DNA time series in ecology. Trends Ecol. Evol. 33, 945–957. https://doi.org/10.1016/j.tree.2018.09.003

Bell KL, Burgess KS, Botsch JC, Dobbs EK, Read TD, Brosi BJ (2019) Quantitative and qualitative assessment of pollen DNA metabarcoding using constructed species mixtures. Mol. Ecol. 28, 431–455. https://doi.org/10.1111/mec.14840

Bohmann K, Evans A, Gilbert MTP, Carvalho GR, Creer S, Knapp M, Yu DW, de Bruyn M (2014) Environmental DNA for wildlife biology and biodiversity monitoring. Trends Ecol. Evol. 29, 358–367. https://doi.org/10.1016/j.tree.2014.04.003

Brummitt NA, Bachman SP, Griffiths-Lee J, Lutz M, Moat JF, Farjon A, Donaldson JS, Hilton-Taylor C, Meagher TR, Albuquerque S, Aletrari E, Andrews AK, Atchison G, Baloch E, Barlozzini B, Brunazzi A, Carretero J, Celesti M, Chadburn H, Cianfoni E, Nic Lughadha EM (2015) Green plants in the red: A baseline global assessment for the IUCN sampled red list index for plants. PLoS ONE 10, e0135152. https://doi.org/10.1371/journal.pone.0135152

Burns JM, Janzen DH, Hajibabaei M, Hallwachs W, Hebert PDN (2008) DNA barcodes and cryptic species of skipper butterflies in the genus Perichares in Area de Conservacion Guanacaste, Costa Rica. Proc. Natl. Acad. Sci. USA 105, 6350–6355. https://doi.org/10.1073/pnas.0712181105

Butchart SHM, Walpole M, Collen B, van Strien A, Scharlemann JPW, Almond REA, Baillie JEM, Bomhard B, Brown C, Bruno J, Carpenter KE, Carr GM, Chanson J, Chenery AM, Csirke J, Davidson NC, Dentener F, Foster M, Galli A, Galloway JN, Watson R (2010) Global biodiversity: indicators of recent declines. Science 328, 1164–1168. https://doi.org/10.1126/science.1187512

CBD COP5 (1996) Decision V/9. Global Taxonomy Initiative: implementation and further advance of the Suggestions for Action [WWW Document]. URL https://www.cbd.int/decision/cop/?id=7151 (accessed 12.21.20).

Chimeno C, Hausmann A, Schmidt S, Raupach MJ, Doczkal D, Baranov V, Hübner J, Höcherl A, Albrecht R, Jaschhof M, Haszprunar G, Hebert PDN (2022) Peering into the darkness: DNA barcoding reveals surprisingly high diversity of unknown species of Diptera (Insecta) in Germany. Insects 13, 82. https://doi.org/10.3390/insects13010082

de Boer HJ, Ghorbani A, Manzanilla V, Raclariu A-C, Kreziou A, Ounjai S, Osathanunkul M, Gravendeel B (2017) DNA metabarcoding of orchid-derived products reveals widespread illegal orchid trade. Proc. R. Soc. B 284, 20171182. https://doi.org/10.1098/rspb.2017.1182

Dirzo R, Raven PH (2003) Global state of biodiversity and loss. Annu. Rev. Environ. Resour. 28, 137–167. https://doi.org/10.1146/annurev.energy.28.050302.105532

Ghorbani A, Gravendeel B, Selliah S, Zarré S, de Boer HJ (2017) DNA barcoding of tuberous Orchidoideae: a resource for identification of orchids used in Salep. Mol. Ecol. Resour. 17, 342–352. https://doi.org/10.1111/1755-0998.12615

Govaerts R, Nic Lughadha E, Black N, Turner R, Paton A (2021) The World Checklist of Vascular Plants, a continuously updated resource for exploring global plant diversity. Sci. Data 8, 215. https://doi.org/10.1038/s41597-021-00997-6

Hammond PM (1992) Species inventory, in: Groombridge, B. (Ed.), Global Biodiversity: Status of the Earth's Living Resources. Springer, Houten, pp. 17–39.

Hausmann A, Krogmann L, Peters RS, Rduch V, Schmidt S (2020) GBOL III: DARK TAXA. barbull 10. https://doi.org/10.21083/ibol.v10i1.6242

Hawkins J, de Vere N, Griffith A, Ford CR, Allainguillaume J, Hegarty MJ, Baillie L, Adams-Groom B (2015) Using DNA metabarcoding to identify the floral composition of honey: A new tool for investigating honey bee foraging preferences. PLoS ONE 10, e0134735. https://doi.org/10.1371/journal.pone.0134735

Hawksworth DL, Kalin-Arroyo MT (1995) Magnitude and distribution of biodiversity, in: Heywood, V.H., Watson, R.T. (Eds.), Global Biodiversity Assessment. Cambridge University Press, Cambridge, pp. 107–191.

Hebert PDN, Cywinska A, Ball SL, de Waard JR (2003) Biological identifications through DNA barcodes. Proc. R. Soc. Lond. B 270, 313–322.

Hooper DU, Adair EC, Cardinale BJ, Byrnes JEK, Hungate BA, Matulich KL, Gonzalez A, Duffy JE, Gamfeldt L, O'Connor MI (2012) A global synthesis reveals biodiversity loss as a major driver of ecosystem change. Nature 486, 105–108. https://doi.org/10.1038/nature11118

Humphreys AM, Govaerts R, Ficinski SZ, Nic Lughadha E, Vorontsova MS (2019) Global dataset shows geography and life form predict modern plant extinction and rediscovery. Nat. Ecol. Evol. 3, 1043–1047. https://doi.org/10.1038/s41559-019-0906-2

IPBES (2019) Global assessment report of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. S. Díaz, J. Settele, E. Brondízio and H. T. Ngo. Bonn, Germany, IPBES Secretariat: 1753. https://doi.org/10.5281/zenodo.3831673 https://ipbes.net/global-assessment

IPNI (2020) International Plant Names Index. The Royal Botanic Gardens, Kew, Harvard University Herbaria & Libraries and Australian National Botanic Gardens. Published on the Internet; [WWW Document]. URL https://www.ipni.org/ (accessed 10.20.20).

Jarman SN, Elliott NG (2000) DNA evidence for morphological and cryptic Cenozoic speciations in the Anaspididae, 'living fossils' from the Triassic. J. Evol. Biol. 13, 624–633.

Knowlton N (1993) Sibling species in the sea. Annu. Rev. Ecol. Syst. 24, 189–216. https://doi.org/10.1146/annurev.es.24.110193.001201

Lughadha EN, Govaerts R, Belyaeva I, Black N, Lindon H, Allkin R, Magill RE, Nicolson N (2016) Counting counts: revised estimates of numbers of accepted species of flowering plants, seed plants, vascular plants and land plants with a review of other recent estimates. Phytotaxa 272, 82. https://doi.org/10.11646/phytotaxa.272.1.5

Lynggaard C, Bertelsen MF, Jensen CV, Johnson MS, Frøslev TG, Olsen MT, Bohmann K (2022) Airborne environmental DNA for terrestrial vertebrate community monitoring. Curr. Biol. 32, 701–707. https://doi.org/10.1016/j.cub.2021.12.014

Nithaniyal S, Newmaster SG, Ragupathy S, Krishnamoorthy D, Vassou SL, Parani M (2014) DNA barcode authentication of wood samples of threatened and commercial timber trees within the tropical dry evergreen forest of India. PLoS ONE 9, e107669. https://doi.org/10.1371/journal.pone.0107669

Page RDM (2016) DNA barcoding and taxonomy: dark taxa and dark texts. Philos. Trans. R. Soc. Lond. B Biol. Sci. 371. https://doi.org/10.1098/rstb.2015.0334

Raclariu AC, Heinrich M, Ichim MC, de Boer H (2018) Benefits and limitations of DNA barcoding and metabarcoding in herbal product authentication. Phytochem. Anal. 29, 123–128. https://doi.org/10.1002/pca.2732

Raclariu AC, Paltinean R, Vlase L, Labarre A, Manzanilla V, Ichim MC, Crisan G, Brysting AK, de Boer H (2017) Comparative authentication of Hypericum perforatum herbal products using DNA metabarcoding, TLC and HPLC-MS. Sci. Rep. 7, 1291. https://doi.org/10.1038/s41598-017-01389-w

Ragupathy S, Newmaster SG, Murugesan M, Balasubramaniam V (2009) DNA barcoding discriminates a new cryptic grass species revealed in an ethnobotany study by the hill tribes of the Western Ghats in southern India. Mol. Ecol. Resour. 9 Suppl s1, 164–171. https://doi.org/10.1111/j.1755-0998.2009.02641.x

Riedel A, Sagata K, Suhardjono YR, Tänzler R, Balke M (2013) Integrative taxonomy on the fast track - towards more sustainability in biodiversity research. Front. Zool. 10, 15. https://doi.org/10.1186/1742-9994-10-15

Ryberg, M, Nilsson, RH (2018) New light on names and naming of dark taxa. MycoKeys 30, 31–39. https://doi.org/10.3897/mycokeys.30.24376

Thomsen PF, Willerslev E (2015) Environmental DNA – An emerging tool in conservation for monitoring past and present biodiversity. Biol. Conserv. 183, 4–18. https://doi.org/10.1016/j.biocon.2014.11.019

SECTION 1

Design, sampling, and substrates

Chapter 1
DNA from plant tissue

Natalia Angela Saoirse Przelomska[1], Steven Dodsworth[1,2], Ana Rita Giraldes Simões[1], Panagiota Malakasi[1], Imalka Kahandawala[1], Roseina Woods[1], Timothy Fulcher[1], Yannick Woudstra[1,3,4,5], Olwen M. Grace[1]

1 Royal Botanic Gardens, Kew, Surrey, United Kingdom
2 University of Portsmouth, Portsmouth, United Kingdom
3 Natural History Museum Denmark, University of Copenhagen, Denmark
4 Gothenburg Global Biodiversity Center, Department of Biological and Environmental Sciences, University of Gothenburg, Sweden
5 Department of Plant Sciences, University of Oxford, United Kingdom

Natalia Angela Saoirse Przelomska **n.przelomska@kew.org**
Steven Dodsworth **steven.dodsworth@port.ac.uk**
Ana Rita Giraldes Simões **a.simoes@kew.org**
Panagiota Malakasi **p.malakasi@kew.org**
Imalka Kahandawala **i.kahandawala@kew.org**
Roseina Woods **r.woods@kew.org**
Timothy Fulcher **t.fulcher@kew.org**
Yannick Woudstra **yannickwoudstra@outlook.com**
Olwen M. Grace **o.grace@kew.org**

Citation: Przelomska NAS, Dodsworth S, Simões ARG, Malakasi P, Kahandawala I, Woods R, Fulcher T, Woudstra Y, Grace OM (2022) Chapter 1. DNA from plant tissue. In: de Boer H, Rydmark MO, Verstraete B, Gravendeel B (Eds) Molecular identification of plants: from sequence to species. Advanced Books. https://doi.org/10.3897/ab.e98875

Plant DNA

What is DNA?

Deoxyribonucleic acid (DNA) is the blueprint of life. DNA encodes genes which carry instructions for the production of proteins, the fundamental components of a cell's machinery. DNA was first isolated and confirmed as the genetic material in cells, and thereby the basis of heredity, in the 1940s (Avery et al. 1944). DNA is a polymer consisting of nucleotide monomers, each containing a phosphate group, a sugar group (ribose), and one of the four bases: adenine (A), thymine (T), cytosine (C), or guanine (G). The order of nucleotides determines the primary structure of DNA. Its secondary structure is dictated by hydrogen bonding between the purine-pyrimidine base pairs A-T (two bonds) and C-G (three bonds); these link the complementary antiparallel single DNA strands. The way in which nucleotide bases form pairs was discovered by Chargraff (1950). The two strands of chemically bonded nucleotides form the tertiary structure of DNA, which is a double helix in all known biological systems. The tertiary structure of DNA was first elucidated by Franklin, Watson, and Crick in the mid 20th century (Franklin and Gosling 1953; Watson and Crick 1953).

A fundamental tenet of molecular biology is that DNA is transcribed into ribonucleic acid (RNA), and subsequently translated into amino acids that form a protein sequence. We now have a much more detailed understanding of this framework, including the varied roles of RNA in gene expression and regulation, and the role of epigenetics—heritable changes in DNA that do not alter the base sequence (e.g., methylation). Since the discovery of DNA, there has been a steady increase in the use of DNA sequences as molecular markers in varied biological contexts, including medical and forensic applications, elucidation of genes encoding adaptive traits, understanding population genomic processes, as well as systematics of prokaryotic and eukaryotic organisms.

Distribution of plant DNA in the cell

Most DNA extraction protocols extract total cellular DNA. In certain experimental cases, it can also be preferable to target either DNA contained in the nucleus or DNA comprising organellar genomes (in plants: mitochondria and plastids). Organellar genomes are much smaller than any plant nuclear genome.

As with virtually all eukaryotes, plants have endosymbiotically derived mitochondria for cellular respiration and energy production. However, compared to other eukaryotic kingdoms (animals in particular), the mitochondrial genome of plants is quite large, ranging between 200 and 750 Kbp in size (Kubo and Newton 2008), and is characterised by a slow substitution rate and significant genome rearrangement events (Gualberto et al. 2014). Therefore, unlike in many other eukaryotes, the mitochondrial genome in plants is rarely used for molecular identification, including phylogenetics and systematics.

In contrast, plastid genomes (e.g.: found in chloroplasts of leaves or amyloplasts of cereal grains) have a very stable genomic structure and a size of around 150 Kbp in most cases (Twyford and Ness 2017). They have a high enough rate of substitution to serve as a useful molecular tool across different phylogenetic levels, including population-level and phylogeographical studies (Petit and Vendramin 2007), and harbour the plant DNA barcode genes *matK*, *rbcL*, and *trnH-psbA* (CBOL Plant Working Group 2009). Plant cells contain multiple plastids per cell, and each of these has several copies of their plastid genome, meaning that plastomes are present at a high copy number in DNA extracts. This makes them particularly useful for sequencing from total DNA extracts, using for instance a genome skimming approach (Dodsworth 2015; Twyford and Ness 2017).

Nuclear genomes, particularly in angiosperms, are highly variable in size, with the angiosperm mean and modal 1C (the amount of DNA in an unreplicated gametic nucleus) both at

around 5 pg/Gbp (Pellicer et al. 2018). The largest genomes are found mostly amongst monocots, particularly the Liliaceae and Melanthiaceae, including the record-holder *Paris japonica* with around 150 Gbp of DNA (Pellicer et al. 2010). The smallest genomes have been found in carnivorous members of the genus *Genlisea*, e.g.: 61 Mbp in *G. tuberosa* (Fleischmann et al. 2014). It has only recently become possible to perform genome-wide analyses on the largest plant genomes thanks to developments in molecular methods for high-throughput DNA sequencing, including those that reduce genomic complexity (Dodsworth et al. 2019). Methodologies for high quality (chromosome-level) assembly of large plant genomes have also advanced, one example being Hi-C technology (Putnam et al. 2016; Neale et al. 2022).

Experimental history and main principles of DNA extractions

The first isolation of DNA, by the Swiss physician Friedrich Miescher in 1869, happened accidentally while studying proteins from leukocyte nuclei (Dahm 2005). Miescher noted a substance that precipitated from solution when acid was added and that re-dissolved upon the addition of an alkaline solution. He called this precipitant "nuclein". While modern protocols for DNA isolation are considerably more refined than the very first trials, the general goal remains the same: to separate intact DNA from other plant cellular molecules, while minimising DNA degradation.

Plants possess a tough cell wall made up of cellulose and other compounds such as lignin, in addition to a cell membrane. This necessitates a robust first step for plant DNA extraction that disintegrates the structure of the plant tissue and breaks down cell walls. In a low-throughput scenario (or for samples that are tougher to disrupt), this could involve flash freezing the tissue with liquid nitrogen followed by grinding with a pestle and mortar. For higher throughput of samples, tissue-disrupting machinery can be applied. The ground material should then be taken forward immediately to the chemical steps of the process, which involve breakdown of the cellular membrane to release the lysate containing the soluble DNA. This is then separated from cell debris and other insoluble material. Various methods are subsequently used to separate DNA molecules from the remaining material, which can contain soluble proteins, nucleic acids, and small molecular metabolites (Doyle 1996). Cellulose and lignin derived from the cell wall, as well as polysaccharides, polyphenols, tannins, and other secondary metabolites (particularly prevalent in medicinal plants) are common endogenous impurities in DNA extractions. These compounds need to be separated and removed as much as possible; they may inhibit downstream laboratory steps and lead to poorer sequencing (Varma et al. 2007). However, extracting sufficient quantities of high-quality and high purity DNA from plants can often be challenging.

Numerous protocols and procedures have been developed to extract DNA from plant material of varying origins (Murray and Thompson 1980; Doyle and Doyle 1987; Rogers and Bendich 1989; Lodhi et al. 1994). Quite often, a protocol must be optimised or blended with others to obtain high-quality DNA from specific plant material. Refinement of the optimal isolation procedure will depend upon many factors, such as the source tissue, age of the material, and concentration of metabolites present in the plant.

A major innovation in DNA extraction protocols from plant material was developed by Doyle and Doyle (Doyle and Doyle 1987). This protocol uses the cationic detergent CTAB for extracting DNA from small amounts of plant tissues. This was a welcome alternative to the lengthy, expensive, and hazardous caesium chloride ethidium bromide density gradient centrifugation approach (Saghai-Maroof et al. 1984). This procedure quickly gained popularity due to its versatility and scalability, particularly in the volume of detergents, and the use of fresh instead of lyophilized tissue (Doyle and Doyle 1990; Doyle 1991). Today, there are numerous modifications of the original CTAB protocol for the isolation of pure, intact DNA from plants (Scott and

Playford 1996; Sharma et al. 2000; Pirttilä et al. 2001; Drábková et al. 2002; Shepherd et al. 2002; Mogg and Bond 2003; Agbagwa et al. 2012). In the era of next generation sequencing, current innovations in DNA extraction protocols tend to focus on the need for high-throughput DNA extractions from many different taxa simultaneously (Mavrodiev et al. 2021).

Chapter 1: Box 1. Important first steps in the collection of plant material

Plant material for any research project must be collected ethically and legally, and the preparation of DNA extracts is no exception. Permission, prior informed consent and mutually agreeable terms of use must be obtained before using plant tissue for DNA extraction according to the Convention on Biological Diversity. This includes the fair and equitable sharing of benefits arising from the utilisation of genetic resources (as outlined in the Nagoya Protocol). National and international law and conventions apply to derivatives of biological materials, including DNA extracts and their transportation. The same principles apply to botanical collections such as seeds, silica dried specimens stored in a tissue bank, herbarium specimens, or plants in living collections. The terms under which they are stored in a collection may restrict the use of specimens for research and require additional permissions (for instance, from the regulatory authority in the country of origin) before they can be used. The storage and future use of DNA extracts, likewise, must comply with the terms of the permissions granted, which could include being stored indefinitely for future research, returned to the country or institute of origin, or discarded. See Chapter 2 DNA from museum collections for guidance about your responsibilities as a researcher.

Storing and preparing plant material for DNA extraction

Plant material

DNA can be extracted from healthy plant tissues including leaves, flowers, buds, seeds, roots, bark, and even spines. Young leaf tissue is the preferred starting material (Gemeinholzer et al. 2010), particularly for herbaceous plants, and fresh leaf tissue usually yields high volumes of high-quality DNA (Guo et al. 2018). However, the type of material used for DNA extraction depends on availability. Access to plant material and availability due to plant life cycles and seasonal variation may require a pragmatic approach. Some plant tissues (e.g., roots, stems), clades (e.g., ferns; (Thomson 2002)), and morphological features (e.g., succulence (Neubig et al. 2014)) present specific challenges during sample collection and storage, requiring tailored processing approaches.

Successful extraction of high-quality DNA from any plant material depends on the material being prepared correctly, dried rapidly (without excessive heat treatment), and stored in a dark, dry place to minimise degradation of its DNA. DNA degradation prior to extraction is caused by the release of endogenous nucleases during cellular lysis, which may be accelerated by environmental factors such as heat and humidity (Savolainen et al. 1995).

The extraction method is determined by the plant material available. For most kit and CTAB based protocols, a 1 cm^2 section of herbaceous leaf tissue will suffice for a single extraction. Careful laboratory notes of the material used, including provenance data, sample weight, and extraction date, are vital for checking the quality of sequencing results against the specifics of the extraction

process in the lab and for pinpointing reasons for variation between samples. For some protocols, weighed tissue can be placed straight into a 1.5 ml tube labelled with a unique number or laboratory code and other information, ready for the DNA extraction process.

Silica drying

Plant material dried and stored in silica gel – including as specimens stored in tissue banks specifically for the purpose of DNA extraction – tends to be a good source of high-quality DNA. Silica gel (silicon dioxide xerogel) is a desiccant that removes moisture from the atmosphere, drying out the plant tissue. Indicator silica gel crystals change colour when the silica is saturated, signalling when the silica gel should be regenerated or replaced. These crystals can be used in a mixture with non-indicating silica gel.

The use of silica gel is a popular approach to dry fresh plant material for DNA extraction because it is low cost and convenient compared to liquid nitrogen or lyophilization, especially when preparing tissue in the field. To effectively preserve the DNA in plant tissue, the recommended minimum ratio between plant material and silica is 1:10 (Chase and Hills 1991). However, if the material collected is mucilaginous, thick, or hardy, the volume of tissue should be reduced and cut into pieces, bringing the desiccant into contact with the cut surface of the plant material to facilitate rapid desiccation. The environment in which plant material is collected also affects the amount of silica needed and the frequency at which it needs to be replaced; a humid environment will require frequent changes of the desiccant. Tissue samples can either be stored directly in individual, sealed plastic bags containing silica gel, or in a breathable material such as a folded tea bag or coffee filter in a sealed container containing silica gel. The latter method is recommended to prevent cross contamination between samples and avoid powdering of the sample due to friction with the silica gel beads, which makes it more difficult to extract the tissue from the container later. Each sample should be double labelled on the outside, with a second label placed within the sample bag.

Freezing

One approach is to freeze plant tissue until needed for DNA extraction, preferably at –80 °C, and otherwise in a standard laboratory freezer at –20 °C, if the sample is properly sealed. Alternatively, material can be flash frozen in liquid nitrogen. The resulting rapidly frozen material can yield high-quality DNA extractions, but liquid nitrogen is impractical for some settings due to handling considerations and cost (Till et al. 2015). Additionally, cycles of freezing and thawing of plant tissue should be avoided as this can damage plant cells, organelles, and DNA (Nagy 2010). It is therefore recommended that frozen plant material is only thawed once, right before the DNA is extracted.

Lyophilization

High-quality DNA can be extracted from lyophilized (or freeze-dried) tissue, such as leaves and roots (Guinn 1966). This method was developed in the 1960s and is still used when fresh material cannot be used immediately or is not available. When paired with the correct extraction technique, lyophilized plant material can yield DNA of high quality (Nunes et al. 2011). During lyophilization, plant tissue is maintained at low temperatures (< –50 °C) and pressures (< 0.1 mbar),

resulting in sublimation of the water in plant cells. A condenser is typically present that captures the vaporised water as ice. After removal of all water from the plant material (typically achieved within a few hours or overnight), the lyophilizer is brought to atmospheric conditions after which the dried plant tissues can be removed from the device. Proceeding with mechanical disruption of the tissue immediately after this is preferable, reabsorption of to avoid atmospheric moisture. However, the sample can alternatively be stored in silica gel before further use.

DNA extraction protocols

After the plant material has been prepared by drying and/or freezing using one of the above-mentioned techniques, a DNA extraction protocol can be implemented. Although there are a multitude of available protocols, the general methodology involves the following steps, discussed in more detail below:

- Weighing of plant tissue
- Mechanical disruption (grinding)
- (Optional) pre-treatment
- Extraction of nucleic acids from the cell
- DNA isolation and precipitation
- DNA purification

We place emphasis on the CTAB protocol due to its popularity, but also introduce other protocols that may be of interest to the reader.

General workflow for DNA extraction

Weighing plant tissue

The starting amount of plant tissue is important: too little will result in an unsatisfactory yield and too much may lead to poor grinding, saturation of the reaction and/or excessive debris which can also be detrimental to final yield. A useful starting ratio is a buffer quantity that is fivefold that of the weight of the leaf tissue (e.g., 0.2 g leaf tissue for 1 ml of buffer) (Kasajima 2018).

Mechanical disruption (grinding) of plant material

Plant tissue must be finely ground to a powder such that the cell walls are disrupted and the cell membranes are more accessible for the chemical reagents in subsequent steps to act successfully. It is advisable to scrape hairs or wax from the surface of the plant tissue before weighing and grinding. For herbarium specimens, special care should be taken that any glue that may be present is removed since this can interfere with the reagents used during the DNA extraction. Sterilised sand can also be used to increase the friction and enhance the disruption of the tissue; it will be separated later in the DNA extraction protocol. Fleshy tissue can be flash frozen in a mortar with a little liquid nitrogen before grinding. The dewar for transporting the liquid nitrogen should be clean and free of potential contaminants.

Manual grinding is inexpensive, yet time consuming and requires a sterilised mortar, pestle, and spatula for each sample. Use of a mechanical homogenizer, also called a tissue lyser, is more efficient. A steel ball bearing is added to each tube with a sample and shaken at high

frequency within the instrument. This allows multiple samples to be disrupted simultaneously with minimal degradation of the nucleic acids. It also minimises loss of material and the chances of contamination, as each sample is processed in the tube that it remains in for subsequent extraction steps. Metallic, ceramic, or silica beads of different sizes can be added to the sample tubes to increase the disruption of particularly tough or woody material. Metallic and ceramic beads must be removed before proceeding with the protocol, but silica beads can be separated later in the protocol.

Optional pre-treatment

This step can be included as an optimisation strategy for increased yield, quality, or purity of the extracted DNA. For example, when high amounts of polysaccharides and/or polyphenols in the plant material are a concern (as is the case for succulent plants and plants in high stress environments, respectively), the modified STE-CTAB protocol can be used (Shepherd and McLay 2011). The ground plant tissue is washed up to three times with a Sucrose-Tris-EDTA (STE) buffer that dissolves most of the polysaccharides and polyphenol, after which the standard CTAB protocol can be followed. An alternative sorbitol-based pre-wash can also be beneficial in polyphenol removal and hence obtaining DNA of higher purity (Inglis et al. 2018).

Extraction of nucleic acids from the cell

In this stage, the goal is to release nucleic acids from the cell, whilst also minimising risk of nucleic acid degradation and to commence the segregation of unwanted cellular compounds from the DNA molecules.

The hallmark of the most widely adopted method for DNA extraction from plants, originally developed by Doyle and Doyle (Doyle and Doyle 1987) and Doyle (Doyle 1991) is cetrimonium bromide (CTAB) extraction buffer, and this should contain:

- 2% w/v CTAB: a cationic detergent which, during DNA extraction, binds to the lipids in cell membranes, enhancing cell lysis, thus releasing intact nucleic acids from the nucleus and organelles
- 1.4 M NaCl: a salt which increases the ionic strength of the solution, which simultaneously induces plasmolysis, promotes separation of proteins from DNA, and aids in polysaccharide precipitation
- 100 mM Tris-HCl: a buffer (at pH ~8.0) which maintains the pH of the solution and stabilises the DNA by impeding degradation
- 20 mM EDTA (ethylenediaminetetraacetic acid): which protects the DNA by inhibiting the enzymatic activity of DNase and RNase (i.e., by chelating divalent cations, such as Mg^{2+} and Ca^{2+}, which are cofactors for these enzymes)
- 0.2% ß-mercaptoethanol: which denatures polyphenols and tannins (abundant in plants), rendering it possible to separate them from the DNA in subsequent steps

CTAB buffer is added to each sample tube containing ground plant tissue and the mixture is incubated at 60–65 °C for 15–60 minutes. This can be done in an automatic shaking incubator. Alternatively, the sample tubes can be periodically shaken manually.

Alternatively, methods involving an SDS buffer can be applied (Dellaporta et al. 1983). The buffer recipe also contains NaCl, Tris-HCl, EDTA, and ß-mercaptoethanol, but differs in the application of the anionic detergent sodium dodecyl sulphate (SDS) for the disruption of cellular membranes, as well as the addition of sodium acetate ($NaCH_3COO$).

DNA isolation and precipitation

The goal of this stage is the separation of DNA from other molecules in the lysate, by making use of the differing polarity of these molecules. This is followed by DNA precipitation from the solution.

In the CTAB protocol, the methodology is phase separation using organic solvent(s), where hydrophilic molecules, including DNA, can be isolated. A 24:1 solution of chloroform-isoamyl alcohol (SEVAG buffer) is added to the incubated CTAB/leaf tissue mixture. This solution is hazardous and must be prepared and added to the sample tubes in a fume hood to avoid inhalation. It is also highly volatile and evaporates very quickly, so it should be handled quickly to avoid evaporation during the work. The mixture is then centrifuged at room temperature, which results in the DNA becoming concentrated in the clear upper phase (i.e., the aqueous phase). The supernatant is very carefully drawn off with a pipette without disturbing or touching the organic phase (containing the chloroform with lipids, proteins, and other cellular debris) and transferred to a new tube. The supernatant is purified by adding RNase A and chilled isopropanol, where the latter induces precipitation of DNA. Samples are then transferred to a freezer at -20 °C, either overnight or for several days if sample input is low and maximum precipitation is desirable (at the cost of potential co-precipitation of salts).

In the SDS protocol, proteins and polysaccharides precipitate with the SDS itself. Sodium acetate in turn is used to precipitate the DNA; in solution this compound dissociates and the sodium ions (Na^+) neutralise the negative ions on the sugar phosphate backbone of DNA molecules, thus making it less hydrophilic and amenable to precipitation (Heikrujam et al. 2020).

As a final step to both methodologies, the samples are centrifuged to encourage the formation of a DNA pellet, optionally washed with 70% ethanol at least once and re-suspended, preferably in 10 mM Tris-EDTA buffer (which serves to protect the DNA from damage, as explained in the CTAB buffer recipe above).

DNA purification

The DNA isolation stage is not perfect. Since the extraction process involves steps that segregate compounds by binding properties and molecular weight, co-extraction of molecularly similar polysaccharides is common. Furthermore, the eluent can contain certain contaminants, including traces of chemicals added during the extraction process and precipitated salts, as well as endogenous proteins, tannins, polysaccharides, and other molecules. The presence of such compounds can negatively impact the downstream experimental use of the DNA (i.e., act as PCR inhibitors), and further purification of DNA using various clean-up steps may be necessary.

One strategy is using a silica column and centrifugation-based method, by adding a chaotropic agent (commonly guanidine hydrochloride), which disrupts the hydrogen bonds between water molecules, creating a more hydrophobic environment. This increases the solubility of non-polar compounds (often contaminants) and additionally breaks up the hydration shell that forms around the negatively charged DNA phosphate backbone and further promotes efficient adsorption to the column surface under high salt and moderately acidic conditions (Esser et al. 2006). This is followed by washing steps with alcohol-based solvents and centrifugation to remove unbound contaminants before final elution of the DNA in a suitable buffer, such as 10 mM Tris-EDTA (pH 8.0).

An alternative involves the use of Solid Phase Reverse Immobilisation (SPRI) beads (Hawkins et al. 1994). These beads are paramagnetic, meaning that they clump together when exposed to a magnetic field. Their magnetite surface is coated with carboxyl molecules that can reversibly bind to DNA under specific chemical conditions. Polyethylene glycol (PEG), in this context termed the 'crowding agent', promotes the binding of DNA to SPRI beads. The ratio of this crowding agent to the DNA eluent is key: the higher the concentration, the greater the attractive force of DNA molecules to the beads, meaning that progressively smaller fragments with molecules of lower charge can bind to the beads. Therefore, choosing a ratio of SPRI beads – which are in solution with the

crowding agent and salt (NaCl) – to DNA is the first step. A ratio of 1:1 is usually appropriate for DNA clean-up, though this ratio can be increased up to 2:1 for the retention of very short DNA fragments. Once the tube containing this mixture is placed into a paramagnetic plate, the DNA will remain immobilised to the SPRI beads, which are attracted to the sides of the tube, adjacent to the magnetic field. The supernatant containing any short nucleic acid remnants and contaminants can at this point be pipetted out from the tube. The beads are washed twice with an 80% ethanol solution before addition of an elution buffer (e.g., 10 mM Tris-HCl) to re-suspend the purified DNA.

Protocol optimization

When a DNA extraction protocol does not yield satisfactory results, in terms of quality or quantity of extracted DNA, modifications can be applied. A valuable strategy for this is conducting a search of the scientific literature for protocols that have been used for similar experimental purposes or have targeted the same taxonomic groups.

If using the CTAB protocol, understanding the biochemical actions and interactions of its components is a useful starting point to identifying what might need adjustment to help improve the outcome. CTAB acts according to the ionic strength of the solution; the concentration of NaCl must be at least 0.5 M so that it does not bind to nucleic acids, but does bind to proteins and neutrally charged polysaccharides as desired. NaCl is most commonly used at a concentration of 1.4 M. When working with a plant group that has a high content of polysaccharides, experimenting with higher concentrations of NaCl may improve the purity of the final DNA. Sometimes, other reagents such as N-Lauroylsarcosine (sarkosyl) buffer can be added, to enhance lysis (rupturing of the cell membrane) and to reduce the activity of DNase or RNase enzymes. Proteinase K can also be added to enhance the denaturation of proteins. The volume of 24:1 chloroform-isoamyl alcohol solution can also be adjusted. Phenol can be added as an additional non-polar, organic solvent that is highly effective in denaturing proteins and can aid in increasing the final DNA yield, as opposed to solely applying chloroform (Heikrujam et al. 2020), though it is very hazardous and requires careful handling.

Tris-HCl and EDTA are present in nearly all protocols. ß-mercaptoethanol is toxic and should thus be handled with care, and always in a fume hood with an extractor fan. One may consider simply not adding this reagent to the solution for plant tissues low in phenolic compounds. However, it is important to note that phenolic compounds co-precipitate with DNA and thus can be problematic in downstream steps of DNA laboratory work. ß-mercaptoethanol can be replaced with less toxic alternatives such as PVP (polyvinylpyrrolidone). PVP attaches to phenolic compounds via hydrogen bonding and can be removed together with them after centrifugation (Porebski et al. 1997; Varma et al. 2007). PVP has been found to improve DNA extraction from tissues such as wood (Rachmayanti et al. 2006). A similar compound – PVPP (polyvinylpolypyrrolidone), whose main characteristic compared to PVP is that it increases the pH of the extraction buffer – has also been found to increase the yield of DNA extracts (Kasajima et al. 2013). Finally, an optimization step for more recalcitrant plant tissues is the application of a 4–6 hour long or overnight incubation at 45–55 °C to increase the yield of the extracted DNA.

Commercial extraction kits

Most commercial kit-based protocols use a combination of buffers that perform similar functions to the components of the CTAB protocol, with a final step of elution through silica-columns, which tends to yield relatively clean DNA extracts. An added benefit of column-based kits is the use of filter columns at an earlier stage for the separation of crude plant material. Silica-based

columns bind DNA so that it can be washed multiple times with alcohol-containing solutions to wash away contaminants before DNA elution. This speeds up DNA extraction significantly, reducing the total time from multiple days – as is common in regular protocols – to 6 hours. Drawbacks of these approaches however include the reduced yields of purified DNA in comparison to CTAB + chloroform extractions, as well as the significantly higher (~3–4 fold greater) cost.

Commercial kits that use magnetic beads are also becoming increasingly popular. Magnetic bead extraction kits are highly versatile and provide high yields of DNA that are also highly pure, in the absence of the hazardous solvents chloroform and phenol. After plant tissue grinding and lysis with an appropriate buffer, DNA is bound to the surface of the magnetic particles. The magnetic particle-DNA system is then washed several times with alcohol-containing solutions before a final elution step with a low salt buffer or nuclease-free water. In contrast to the column-based extraction method, binding of DNA to the magnetic particles occurs in solution, thus enhancing the efficiency and kinetics of binding and simultaneously increasing the contact of the bead-DNA compounds with the wash buffer, which improves the purity of the DNA. Magnetic particle kits have also been applied in combination with steps from the CTAB extraction method to extract high quality DNA from sorghum leaves and seeds, cotton leaves and pine needles (Xin and Chen 2012).

Finally, a less common commercial method involves the use of Whatman FTA® PlantSaver cards and custom reagents. This method is very practical in terms of collection of samples in the field and their transportation. Furthermore, immediate mechanical disruption of the plant tissue can eliminate the need for obtaining permits. While this method has been predominantly applied to agricultural plant taxa, its performance in 15 phylogenetically diverse non-agricultural taxa has been demonstrated, where DNA from these samples was found to be less fragmented than that from replicate samples extracted alongside with the CTAB method (Siegel et al. 2017).

DNA quantification and quality assessment

Assessment of the properties of each genomic DNA (gDNA) sample post-extraction – its integrity, quantity, and purity – is imperative for making decisions regarding downstream molecular work. The methods described below have some overlapping uses in terms of assessing these different properties, but we highlight which is most appropriate for each DNA quality-related aspect.

DNA integrity - agarose gel electrophoresis

Agarose gel electrophoresis is an appropriate method for estimating DNA integrity, as well as for crudely estimating DNA concentration. This method requires a horizontal gel electrophoresis tank with an external power supply, agarose, a running buffer such as Tris-acetate-EDTA (TAE) or sodium borate (SB), a fluorescent intercalating DNA dye, a loading dye, and a DNA standard ('ladder'). The intercalating dye is added to the buffer (or sometimes to the loading dye) and serves to visualise the DNA in the agarose gel at the end point of electrophoresis. Historically, ethidium bromide was the standard intercalating agent, but it has now mostly been superseded by safer dyes that are less carcinogenic and do not require complex disposal procedures. Nonetheless, it is recommended that any compound that intercalates DNA be handled with care. The DNA standard is referred to as a ladder, since it is a complex of appropriately sized DNA standards of known concentrations which provide different benchmarks of size and concentration for comparison.

Each DNA sample and the DNA standard (ladder) are combined with loading dye and then pipetted into a well of the agarose gel, to then be subjected to an electric field. Due to the

negatively charged phosphate backbone, DNA molecules will migrate towards the positively charged anode. The DNA migration rate depends on the fragment size, where smaller DNA fragments migrate faster, leading to a size-associated separation of DNA molecules. Additionally, the percentage of agarose in the gel will determine the size range of DNA that will be resolved with the greatest clarity. A range of 0.5% to 3% encompasses most applications, where < 1% is best for examining the genomic DNA of plants and 3% would be suitable for examining fragments with small (e.g., ~20 bp) differences in length. Once the fragments have migrated sufficiently to ensure resolution of the DNA and ladder, the gel is transferred to a cabinet with a UV light and the DNA fragments are visualised due to the excitation of the intercalating dye when UV is applied. The approximate yield and concentration of genomic DNA in a gel are indicated by comparison of the sample's intensity of fluorescence to that of a standard.

Where a more precise estimation of the size of the DNA fragments is required, automated capillary electrophoresis can be used. Such systems (e.g., Agilent Bioanalyser, Agilent Tapestation) are more expensive to use, but – aside from precision – offer faster preparation and analysis time.

DNA quantity - fluorescence quantitation systems

Fluorescent measurements are considered the most accurate quantification method for measuring DNA concentration. These involve the addition of fluorescent dyes (in an accompanying buffer), which selectively intercalate into the DNA. Fluorescence measurements use excitation and emission values that vary depending on the dye used. The concentration of unknown samples is calculated by the fluorometer (e.g., Quantus™ or Qubit™) based on a comparison to a standard measurement from DNA of a known concentration (usually lambda bacteriophage DNA). Since the dyes are sensitive to light and degrade rapidly in its presence, sample tubes must be stored in the dark if readings are not taken imminently after their preparation in the buffer.

DNA purity - absorbance spectroscopy

A rough estimate of DNA yield and a more useful estimate of DNA purity can be measured via absorbance with a spectrophotometer that emits UV light through a UV-transparent cuvette containing the sample. Absorbance readings are conducted at 260 nm (A_{260}), the wavelength of maximum absorption for DNA. The A_{260} measurement is then adjusted for turbidity (measured by absorbance at 320 nm), multiplied by the dilution factor, and calibrated using the following conversion factor: A_{260} of 1.0 = 50 µg/ml pure dsDNA. This useful relationship between light absorption and DNA concentration can be defined according to the Beer-Lambert law. Total yield is obtained by multiplying the DNA concentration by the final total purified sample volume. However, it is key to note that RNA also has maximum absorbance at 260 nm and aromatic amino acids have a maximum absorbance at 280 nm. Both molecules can contribute to the total measured absorbance at 260 nm and thus provide a misleading overestimate of DNA yield.

DNA purity is evaluated by measuring absorbance in the 230–320 nm range. Since proteins are the contaminant of primary concern, absorbance at 260 nm divided by absorbance at 280 nm is the standard metric. DNA can be considered of high quality and suitable for most genomic applications, when it has an A_{260}/A_{280} ratio of 1.7–2.0. As a further step, the ratio of 260 nm to 230 nm can help evaluate the level of salt carryover in the purified DNA, where a A_{260}/A_{230} of > 1.5 is considered to be of good quality. Strong absorbance at around 230 nm, which would lower this ratio, suggests the presence of organic compounds or chaotropic salts.

Instruments such as the NanoDrop® 2000 spectrophotometer are highly accurate for evaluating the A_{260}/A_{280} and A_{260}/A_{230} ratios. This method is not as accurate as fluorescence quantitation, but is most suitable where information on DNA purity is sought and is also time efficient (the sample is loaded directly into the machine and requires no preparation of buffers).

Approaches to challenging DNA extractions

Particularly challenging types of plant tissue, as well as degraded plant material, can still yield high-quality DNA if suitably optimised protocols are followed.

For instance, seeds can be a good source of DNA if specialised protocols are used (Sudan et al. 2017). Similar to other plant tissues, seeds require different collection and storage techniques depending on their morphology. Dry seeds can usually be collected and stored for long periods without treatment before being ground and used in a DNA extraction protocol. Soft seeds in comparison may need to be flash frozen using liquid nitrogen and cryopreserved prior to DNA extraction. The watery components of fleshy or succulent plant tissues require modified approaches to speed up drying before extraction to remove polysaccharide contaminants from the DNA extract (Larridon et al. 2015; Malakasi et al. 2019).

Advances in the sensitivity of genomic sequencing and optimised DNA extraction methods make it possible to study herbarium and other dried botanical specimens (Bieker and Martin 2018; Brewer et al. 2019; Grace et al. 2021; Malakasi et al. 2019; Särkinen et al. 2012). However, using this material involves mining irreplaceable reservoirs of biological and cultural heritage (Austin et al. 2019; Freedman et al. 2018). Sampling should be restricted to the minimum size expected to yield sufficient DNA for the project and the decision on which part of the specimen to sample should be made in consultation with a collection manager or specialist

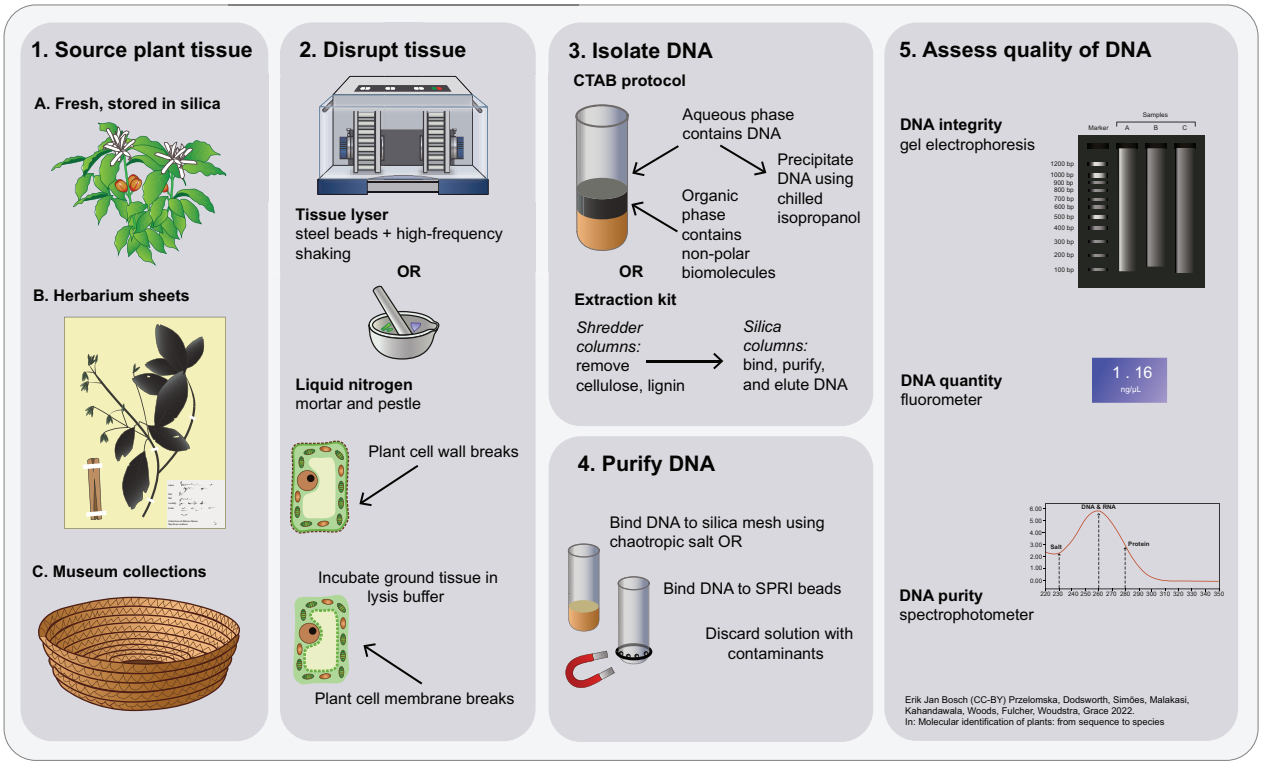

Figure 1. Chapter 1 Infographic: Visual representation of the content of this chapter.

(see Chapter 2 DNA from museum collections). Novel techniques have been developed for minimally destructive sampling of herbarium specimens (Shepherd 2017; Sugita et al. 2020), but these are not universally applicable. Archaeological and museum collections present similar challenges and sometimes even more sensitive decisions. Archaeological plant material can include plant micro- and macrofossils (reviewed in (Kistler et al. 2020)), such as for example grape pips (Wales et al. 2016), or even archaeological artefacts, e.g., a palm-leaf jar stopper (Pérez-Escobar et al. 2021) or paper-mulberry tapa (Peña-Ahumada et al. 2020). At the lab bench, two key obstacles should be considered: contamination and degradation of the DNA. Whereas contamination is a crucial consideration throughout the process of DNA extraction, the physical fragmentation of plant tissue requires the most consideration during experimental design and downstream HTS laboratory work since herbarium and museum samples are treated with an overall more sensitive set of protocols than standard plant tissue samples (see Chapter 2 DNA from museum collections). Any small amount of accidentally introduced nucleic acid contamination will hitchhike alongside the (most likely) degraded DNA present in the sample of interest throughout the purification procedures and has a high likelihood of being preferentially amplified. Crucially, a laboratory's DNA extraction area should not overlap with any area where PCR amplicons are generated. Finally, a simple way to test for persistent contamination is to include extraction blanks.

Physical and chemical degradation is to be expected in herbarium and museum specimens; DNA in deceased tissue breaks down over time. The rate of physical fragmentation is related to temperature and other environmental variables, as well as the composition of the plant tissue itself. In a study of herbarium specimens, it was shown that fragment length significantly regressed against sample age going back 300 years (Weiß et al. 2016), a proxy which can be exploited as a useful starting point for making DNA quality-based lab work decisions. This is more likely to hold true within a plant clade (e.g., plant family) and with a consistent method of sample preparation. However, the relationship of increasing fragmentation with sample age is not always linear. Fixation of the plant material for accessioning in the herbarium is often the single most damaging process (Staats et al. 2011).

The CTAB extraction protocol is generally preferable for extracting fragmented DNA, as it generally gives higher yields of DNA than kit-based methods. Where fragment size distribution is predicted to be very low, a high-volume chaotropic salt used as a binding buffer in the latter stage of extraction can improve the recovery of DNA molecules (Dabney et al. 2013). Alternatively, the ratio of SPRI beads to DNA during the clean-up step can be increased to retrieve more of the shorter DNA molecules. A hallmark of chemical DNA degradation, i.e. cytosine deamination, can be addressed in downstream steps by using repair enzymes in DNA library preparation and appropriate bioinformatic treatment (Kistler et al. 2020).

Concluding remarks

A wide variety of DNA extraction protocols are available in the literature. The structural, biochemical, and genomic characteristics of plants present a particular set of challenges; isolating high purity, undamaged DNA from plant tissue is non-trivial and requires a careful and patient approach in the laboratory. Therefore, researchers must often optimise a chosen protocol for their specific experiment. Success in the primary step of a molecular workflow is crucial, unlocking the downstream steps of plant molecular identification and characterisation, and hence possibilities for addressing many exciting questions in molecular and evolutionary biology.

Questions

1. For each of the DNA-containing compartments in a plant cell, which of its characteristics deserve most consideration during DNA extraction and analysis, and why?
2. Describe the main compound classes from plant extracts that need to be removed from DNA extracts for downstream analysis. How can they be removed?
3. Describe the main difference between DNA extraction using the CTAB protocol and using a column-based extraction kit. What are the advantages and disadvantages of both?

Glossary

Absorbance – A measure of the quantity of light absorbed by a sample, also referred to as optical density, measured using an absorbance spectrophotometer.

Beer-Lambert law – For a material through which light is travelling, the path length of light and concentration of the sample are both directly proportional to the absorbance of the light.

Chaotropic agent – A chemical substance which in an aqueous solution destroys the hydrogen bonds between water molecules (e.g., guanidine hydrochloride).

Cryopreservation – A preservation treatment for biological material, which involves cooling to very low temperatures (at least -80 °C, or -196 °C using e.g., liquid nitrogen).

Desiccant – A substance with a high affinity for water, such that it attracts moisture from surrounding materials, resulting in a state of dryness in its vicinity (e.g., silica gel).

DNA integrity – The level of fragmentation of extracted DNA, where minimal fragmentation of the original chromosomes equates to high DNA integrity.

Intercalating dye – A dye, whose molecular components stack between two bases of DNA, which is invaluable for DNA visualisation, yet at the same time implies a hazard for human health and demands laboratory safety considerations.

Lysate – A commonly fluid mixture of cellular contents that is the result of the disruption of cell walls and membranes via cell lysis.

Molecular marker (in a genetic context) – A sequence of DNA, which can be a single base pair, a gene, or repetitive sequence, with a known location in the genome, which tends to exhibit variation amongst individuals or taxa, such that it has useful research applications.

Organellar genome – The genetic material present in a plastid or mitochondrion, typically in the form of a small and circular genome and often in multiple copies within each organelle. These are thought to be present in eukaryotic cells as a result of endosymbiosis.

Plastome – The total genetic information contained by the plastid (e.g., chloroplast) of a plant cell.

References

Agbagwa IO, Datta S, Patil PG, Singh P, Nadarajan N (2012) A protocol for high-quality genomic DNA extraction from legumes. Genet. Mol. Res. 11, 4632–4639. https://doi.org/10.4238/2012.1

Austin RM, Sholts SB, Williams L, Kistler L, Hofman CA (2019) Opinion: To curate the molecular past, museums need a carefully considered set of best practices. Proc Natl Acad Sci USA 116, 1471–1474. https://doi.org/10.1073/pnas.1822038116

Avery OT, Macleod CM, McCarty M (1944) Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Inductions of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. J. Exp. Med. 79, 137–158. https://doi.org/10.1084/jem.79.2.137

Bieker VC, Martin MD (2018) Implications and future prospects for evolutionary analyses of DNA in historical herbarium collections. Botany Letters 165, 1–10. https://doi.org/10.1080/23818107.2018.1458651

Brewer GE, Clarkson JJ, Maurin O, Zuntini AR, Barber V, Bellot S, Biggs N, Cowan RS, Davies NMJ, Dodsworth S, Edwards SL, Eiserhardt WL, Epitawalage N, Frisby S, Grall A, Kersey PJ, Pokorny L, Leitch IJ, Forest F, Baker WJ (2019) Factors affecting targeted sequencing of 353 nuclear genes from herbarium specimens spanning the diversity of angiosperms. Front. Plant Sci. 10, 1102. https://doi.org/10.3389/fpls.2019.01102

CBOL Plant Working Group (2009) A DNA barcode for land plants. Proc Natl Acad Sci USA 106, 12794–12797. https://doi.org/10.1073/pnas.0905845106

Chargaff E (1950) Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. Experientia 6, 201–209. https://doi.org/10.1007/BF02173653

Chase MW, Hills HH (1991) Silica gel: An ideal material for field preservation of leaf samples for DNA studies. Taxon 40, 215. https://doi.org/10.2307/1222975

Dabney J, Meyer M, Pääbo S (2013) Ancient DNA damage. Cold Spring Harb. Perspect. Biol. 5. https://doi.org/10.1101/cshperspect.a012567

Dahm R (2005) Friedrich Miescher and the discovery of DNA. Dev. Biol. 278, 274–288. https://doi.org/10.1016/j.ydbio.2004.11.028

Dellaporta SL, Wood J, Hicks JB (1983) A plant DNA minipreparation: Version II. Plant Mol Biol Rep 1, 19–21.

Dodsworth S, Pokorny L, Johnson MG, Kim JT, Maurin O, Wickett NJ, Forest F, Baker WJ (2019) Hyb-Seq for flowering plant systematics. Trends Plant Sci. 24, 887–891. https://doi.org/10.1016/j.tplants.2019.07.011

Dodsworth S (2015) Genome skimming for next-generation biodiversity analysis. Trends Plant Sci. 20, 525–527. https://doi.org/10.1016/j.tplants.2015.06.012

Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemical Bulletin 19, 11–15.

Doyle JJ, Doyle JL (1990) A rapid total DNA preparation procedure for fresh plant tissue. Focus 12, 13–15.

Doyle J (1991) DNA protocols for plants, in: Hewitt, G.M., Johnston, A.W.B., Young, J.P.W. (Eds.), Molecular Techniques in Taxonomy. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 283–293. https://doi.org/10.1007/978-3-642-83962-7_18

Doyle K (1996) The source of discovery: protocols and applications guide.

Drábková L, Kirschner J, Vlĉek Ĉ (2002) Comparison of seven DNA extraction and amplification protocols in historical herbarium specimens of juncaceae. Plant Mol. Biol. Rep. 20, 161–175. https://doi.org/10.1007/BF02799431

Esser K-H, Marx WH, Lisowsky T (2006) maxXbond: first regeneration system for DNA binding silica matrices. Nat. Methods 3. https://doi.org/10.1038/nmeth845

Fleischmann A, Michael TP, Rivadavia F, Sousa A, Wang W, Temsch EM, Greilhuber J, Müller KF, Heubl G (2014) Evolution of genome size and chromosome number in the carnivorous plant genus Genlisea (Lentibulariaceae), with a new estimate of the minimum genome size in angiosperms. Ann. Bot. 114, 1651–1663. https://doi.org/10.1093/aob/mcu189

Franklin RE, Gosling RG (1953) Evidence for 2-chain helix in crystalline structure of sodium deoxyribonucleate. Nature 172, 156–157. https://doi.org/10.1038/172156a0

Freedman J, Dorp LB, Brace S (2018) Destructive sampling natural science collec-tions: an overview for museum professionals and researchers. Journal of Natural Science Collections 5, 21–34.

Gemeinholzer B, Rey I, Weising K, Grundman M, Muellner AN, Zetzsche H, Droege G, Seberg O, Petersen G, Rawson D, Weigt L (2010) Organizing specimen and tissue preservation in the field for subsequent molecular analyses, in: Eymann, J., Degreef, J., Hauser, C., Monje, J.C., Samyn, Y., VandenSpiegel, D. (Eds.), Manual on Field Recording Techniques and Protocols for All Taxa Biodiversity Inventories. Edgewater: ABCTaxa.

Grace OM, Pérez-Escobar OA, Lucas EJ, Vorontsova MS, Lewis GP, Walker BE, Lohmann LG, Knapp S, Wilkie P, Sarkinen T, Darbyshire I, Lughadha EN, Monro A, Woudstra Y, Demissew S, Muasya AM, Díaz S, Baker WJ,

Antonelli A (2021) Botanical monography in the anthropocene. Trends Plant Sci. 26, 433–441. https://doi.org/10.1016/j.tplants.2020.12.018

Gualberto JM, Mileshina D, Wallet C, Niazi AK, Weber-Lotfi F, Dietrich A (2014) The plant mitochondrial genome: dynamics and maintenance. Biochimie 100, 107–120. https://doi.org/10.1016/j.biochi.2013.09.016

Guinn G (1966) Extraction of nucleic acids from lyophilized plant material. Plant Physiol. 41, 689–695. https://doi.org/10.1104/pp.41.4.689

Guo Y, Yang G, Chen Y, Li D, Guo Z (2018) A comparison of different methods for preserving plant molecular materials and the effect of degraded DNA on ddRAD sequencing. Plant Diversity 40, 106–116. https://doi.org/10.1016/j.pld.2018.04.001

Hawkins TL, O'Connor-Morin T, Roy A, Santillan C (1994) DNA purification and isolation using a solid-phase. Nucleic Acids Res. 22, 4543–4544. https://doi.org/10.1093/nar/22.21.4543

Heikrujam J, Kishor R, Behari Mazumder P (2020) The chemistry behind plant DNA isolation protocols, in: Boldura, O.-M., Baltă, C., Sayed Awwad, N. (Eds.), Biochemical Analysis Tools - Methods for Bio-Molecules Studies. IntechOpen. https://doi.org/10.5772/intechopen.92206

Inglis PW, Pappas M de CR, Resende LV, Grattapaglia D (2018) Fast and inexpensive protocols for consistent extraction of high quality DNA and RNA from challenging plant and fungal samples for high-throughput SNP genotyping and sequencing applications. PLoS ONE 13, e0206085. https://doi.org/10.1371/journal.pone.0206085

Kasajima I, Sasaki K, Tanaka Y, Terakawa T, Ohtsubo N (2013) Large-scale extraction of pure DNA from mature leaves of Cyclamen persicum Mill. and other recalcitrant plants with alkaline polyvinylpolypyrrolidone (PVPP). Sci. Hortic. 164, 65–72. https://doi.org/10.1016/j.scienta.2013.09.011

Kasajima I (2018) Successful tips of DNA extraction and PCR of plants for beginners. Trends in Res 1. https://doi.org/10.15761/TR.1000115

Kistler L, Bieker VC, Martin MD, Pedersen MW, Ramos Madrigal J, Wales N (2020) Ancient plant genomics in archaeology, herbaria, and the environment. Annu. Rev. Plant Biol. 71, 605–629. https://doi.org/10.1146/annurev-arplant-081519-035837

Kubo T, Newton KJ (2008) Angiosperm mitochondrial genomes and mutations. Mitochondrion 8, 5–14. https://doi.org/10.1016/j.mito.2007.10.006

Larridon I, Walter HE, Guerrero PC, Duarte M, Cisternas MA, Hernández CP, Bauters K, Asselman P, Goetghebeur P, Samain M-S (2015) An integrative approach to understanding the evolution and diversity of Copiapoa (Cactaceae), a threatened endemic Chilean genus from the Atacama Desert. Am. J. Bot. 102, 1506–1520. https://doi.org/10.3732/ajb.1500168

Lodhi MA, Ye G-N, Weeden NF, Reisch BI (1994) A simple and efficient method for DNA extraction from grapevine cultivars andVitis species. Plant Mol. Biol. Rep. 12, 6–13. https://doi.org/10.1007/BF02668658

Malakasi P, Bellot S, Dee R, Grace OM (2019) Museomics clarifies the classification of aloidendron (asphodelaceae), the iconic African tree aloes. Front. Plant Sci. 10, 1227. https://doi.org/10.3389/fpls.2019.01227

Mavrodiev EV, Dervinis C, Whitten WM, Gitzendanner MA, Kirst M, Kim S, Kinser TJ, Soltis PS, Soltis DE (2021) A new, simple, highly scalable, and efficient protocol for genomic DNA extraction from diverse plant taxa. Appl. Plant Sci. 9, e11413. https://doi.org/10.1002/aps3.11413

Mogg RJ, Bond JM (2003) A cheap, reliable and rapid method of extracting high-quality DNA from plants. Mol. Ecol. Notes 3, 666–668. https://doi.org/10.1046/j.1471-8286.2003.00548.x

Murray MG, Thompson WF (1980) Rapid isolation of high molecular weight plant DNA. Nucleic Acids Res. 8, 4321–4325. https://doi.org/10.1093/nar/8.19.4321

Nagy ZT (2010) A hands-on overview of tissue preservation methods for molecular genetic analyses. Org. Divers. Evol. 10, 91–105. https://doi.org/10.1007/s13127-010-0012-4

Neale DB, Zimin AV, Zaman S, Scott AD, Shrestha B, Workman RE, Puiu D, Allen BJ, Moore ZJ, Sekhwal MK, De La Torre AR, McGuire PE, Burns E, Timp W, Wegrzyn JL, Salzberg SL (2022) Assembled and annotated 26.5 Gbp coast redwood genome: a resource for estimating evolutionary adaptive potential and investigating hexaploid origin. G3 (Bethesda) 12. https://doi.org/10.1093/g3journal/jkab380

Neubig KM, Whitten WM, Abbott JR, Elliott S, Soltis DE, Soltis PS (2014) Variables affecting DNA preservation in archival plant specimens, in: Applequist, W.L., Campbell, L.M. (Eds.), DNA Banking for the 21st Century: Proceedings of the US Workshop on DNA Banking. Presented at the Proceedings of the U.S. Workshop on DNA Banking, St. Louis, Missouri Botanical Garden, pp. 81–112.

Nunes CF, Ferreira JL, Fernandes MCN, Breves S de S, Generoso AL, Soares BDF, Dias MSC, Pasqual M, Borem A, Cançado GM de A (2011) An improved method for genomic DNA extraction from strawberry leaves. Cienc. Rural 41, 1383–1389. https://doi.org/10.1590/S0103-84782011000800014

Pellicer J, Fay M, Leitch I (2010) The largest eukaryotic genome of them all? Bot. J. Linn. Soc. 164, 10–15. https://doi.org/10.1111/j.1095-8339.2010.01072.x

Pellicer J, Hidalgo O, Dodsworth S, Leitch IJ (2018) Genome size diversity and its impact on the evolution of land plants. Genes (Basel) 9. https://doi.org/10.3390/genes9020088

Peña-Ahumada B, Saldarriaga-Córdoba M, Kardailsky O, Moncada X, Moraga M, Matisoo-Smith E, Seelenfreund D, Seelenfreund A (2020) A tale of textiles: Genetic characterization of historical paper mulberry barkcloth from Oceania. PLoS ONE 15, e0233113. https://doi.org/10.1371/journal.pone.0233113

Pérez-Escobar OA, Bellot S, Przelomska NAS, Flowers JM, Nesbitt M, Ryan P, Gutaker RM, Gros-Balthazard M, Wells T, Kuhnhäuser BG, Schley R, Bogarín D, Dodsworth S, Diaz R, Lehmann M, Petoe P, Eiserhardt WL, Preick M, Hofreiter M, Hajdas I, Baker WJ (2021) Molecular clocks and archeogenomics of a late period Egyptian date palm leaf reveal introgression from wild relatives and add timestamps on the domestication. Mol. Biol. Evol. 38, 4475–4492. https://doi.org/10.1093/molbev/msab188

Petit RJ, Vendramin GG (2007) Plant phylogeography based on organelle genes: an introduction, in: Weiss, S., Ferrand, N. (Eds.), Phylogeography of Southern European Refugia. Springer Netherlands, Dordrecht, pp. 23–97. https://doi.org/10.1007/1-4020-4904-8_2

Pirttilä AM, Hirsikorpi M, Kämäräinen T, Jaakola L, Hohtola A (2001) DNA isolation methods for medicinal and aromatic plants. Plant Mol. Biol. Rep. 19, 273–273. https://doi.org/10.1007/BF02772901

Porebski S, Bailey LG, Baum BR (1997) Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. Plant Mol. Biol. Rep. 15, 8–15. https://doi.org/10.1007/BF02772108

Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, Fields A, Hartley PD, Sugnet CW, Haussler D, Rokhsar DS, Green RE (2016) Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Res. 26, 342–350. https://doi.org/10.1101/gr.193474.115

Rachmayanti Y, Leinemann L, Gailing O, Finkeldey R (2006) Extraction, amplification and characterization of wood DNA from dipterocarpaceae. Plant Mol. Biol. Rep. 24, 45–55. https://doi.org/10.1007/BF02914045

Rogers SO, Bendich AJ (1989) Extraction of DNA from plant tissues, in: Gelvin, S.B., Schilperoort, R.A., Verma, D.P.S. (Eds.), Plant Molecular Biology Manual. Springer Netherlands, Dordrecht, pp. 73–83. https://doi.org/10.1007/978-94-009-0951-9_6

Saghai-Maroof MA, Soliman KM, Jorgensen RA, Allard RW (1984) Ribosomal DNA spacer-length polymorphisms in barley: mendelian inheritance, chromosomal location, and population dynamics. Proc Natl Acad Sci USA 81, 8014–8018. https://doi.org/10.1073/pnas.81.24.8014

Särkinen T, Staats M, Richardson JE, Cowan RS, Bakker FT (2012) How to open the treasure chest? Optimising DNA extraction from herbarium specimens. PLoS ONE 7, e43808. https://doi.org/10.1371/journal.pone.0043808

Savolainen V, Cuénoud P, Spichiger R, Martinez MDP, Crèvecoeur M, Manen J-F (1995) The use of herbarium specimens in DNA phylogenetics: Evaluation and improvement. Plant Syst. Evol. 197, 87–98. https://doi.org/10.1007/BF00984634

Scott KD, Playford J (1996) DNA extraction technique for PCR in rain forest plant species. BioTechniques 20, 974, 977, 979. https://doi.org/10.2144/96206bm07

Sharma KK, Lavanya M, Anjaiah V (2000) A method for isolation and purification of peanut genomic DNA suitable for analytical applications. Plant Mol. Biol. Rep. 18, 393–393. https://doi.org/10.1007/BF02825068

Shepherd LD, McLay TGB (2011) Two micro-scale protocols for the isolation of DNA from polysaccharide-rich plant tissue. J. Plant Res. 124, 311–314. https://doi.org/10.1007/s10265-010-0379-5

Shepherd LD (2017) A non-destructive DNA sampling technique for herbarium specimens. PLoS ONE 12, e0183555. https://doi.org/10.1371/journal.pone.0183555

Shepherd M, Cross M, Stokoe RL, Scott LJ, Jones ME (2002) High-throughput DNA extraction from forest trees. Plant Mol. Biol. Rep. 20, 425–425. https://doi.org/10.1007/BF02772134

Siegel CS, Stevenson FO, Zimmer EA (2017) Evaluation and comparison of FTA card and CTAB DNA extraction methods for non-agricultural taxa. Appl. Plant Sci. 5. https://doi.org/10.3732/apps.1600109

Staats M, Cuenca A, Richardson JE, Vrielink-van Ginkel R, Petersen G, Seberg O, Bakker FT (2011) DNA damage in plant herbarium tissue. PLoS ONE 6, e28448. https://doi.org/10.1371/journal.pone.0028448

Sudan J, Raina M, Singh R, Mustafiz A, Kumari S (2017) A modified protocol for high-quality DNA extraction from seeds rich in secondary compounds. Journal of Crop Improvement 31, 1–11. https://doi.org/10.1080/15427528.2017.1345028

Sugita N, Ebihara A, Hosoya T, Jinbo U, Kaneko S, Kurosawa T, Nakae M, Yukawa T (2020) Non-destructive DNA extraction from herbarium specimens: a method particularly suitable for plants with small and fragile leaves. J. Plant Res. 133, 133–141. https://doi.org/10.1007/s10265-019-01152-4

Thomson J (2002) An improved non-cryogenic transport and storage preservative facilitating DNA extraction from "difficult" plants collected at remote sites. Telopea 9, 755–760. https://doi.org/10.7751/telopea20024013

Till BJ, Jankowicz-Cieslak J, Huynh OA, Beshir MM, Laport RG, Hofinger BJ (2015) Sample collection and storage, in: Low-Cost Methods for Molecular Characterization of Mutant Plants. Springer International Publishing, Cham, pp. 9–11. https://doi.org/10.1007/978-3-319-16259-1_3

Twyford AD, Ness RW (2017) Strategies for complete plastid genome sequencing. Mol. Ecol. Resour. 17, 858–868. https://doi.org/10.1111/1755-0998.12626

Varma A, Padh H, Shrivastava N (2007) Plant genomic DNA isolation: an art or a science. Biotechnol. J. 2, 386–392. https://doi.org/10.1002/biot.200600195

Wales N, Ramos Madrigal J, Cappellini E, Carmona Baez A, Samaniego Castruita JA, Romero-Navarro JA, Carøe C, Ávila-Arcos MC, Peñaloza F, Moreno-Mayar JV, Gasparyan B, Zardaryan D, Bagoyan T, Smith A, Pinhasi R, Bosi G, Fiorentino G, Grasso AM, Celant A, Bar-Oz G, Gilbert MTP (2016) The limits and potential of paleogenomic techniques for reconstructing grapevine domestication. J. Archaeol. Sci. 72, 57–70. https://doi.org/10.1016/j.jas.2016.05.014

Watson JD, Crick FH (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171, 737–738. https://doi.org/10.1038/171737a0

Weiß CL, Schuenemann VJ, Devos J, Shirsekar G, Reiter E, Gould BA, Stinchcombe JR, Krause J, Burbano HA (2016) Temporal patterns of damage and decay kinetics of DNA retrieved from plant herbarium specimens. R. Soc. Open Sci. 3, 160239. https://doi.org/10.1098/rsos.160239

Xin Z, Chen J (2012) A high throughput DNA extraction method with high yield and quality. Plant Methods 8, 26. https://doi.org/10.1186/1746-4811-8-26

Answers

1. The nuclear genome of plants is hugely variable in size. To maximise retrieval of intact DNA for species with larger genomes, a higher DNA yield should be aimed for. This could affect decisions regarding input material and the number of total DNA extractions carried out per sample. The plastid genome is present in high copy numbers in plant cells, as well as being a useful unit for addressing a variety of biological questions. Therefore, it is ideal for genome skimming experiments and a valuable target in degraded material, where the (single copy) nuclear genome might be highly fragmented. The mitochondrial genome of plants is characterised by high plasticity in its genomic structure and therefore is not recommended for plant identification.

2. Problematic biomolecules in plant extracts include polyphenols, tannins, and polysaccharides. These interfere with DNA extraction buffers (such as CTAB) as well as with other buffers and enzymes used in downstream DNA analysis. They are removed from the solution by either SEVAG cleaning (in the CTAB protocol) or, basically, by column cleaning or magnetic particles (commercial kits). Polysaccharides can also be removed from the crude plant tissue prior to extraction using STE buffer. Phenolic compounds can often be removed using ß-mercaptoethanol and/or PVP. Further impurities such as secondary metabolic compounds that may interfere with enzymes in downstream protocols can often be removed using a SPRI bead clean-up protocol.
3. The CTAB protocol uses specific buffers (such as SEVAG) and DNA precipitation (involving isopropanol) to separate non-DNA and DNA biomolecules, whereas extraction kits rely on using DNA-binding columns. or magnetic particles Although the kits are much more expensive on a per-sample basis, they generally yield clean DNA with a short turnaround time (up to 6 hours). CTAB extractions are very cheap and highly scalable as they do not rely on the specifically manufactured columns or magnetic particles. However, the protocol takes at least two full days to progress from plant tissue to DNA extract. Co-precipitation of non-DNA biomolecules is often observed and therefore affects the purity of the final DNA extract. Sometimes, substantial yield losses are observed using extraction kits and this can be a key consideration when dealing with precious samples.

Chapter 2
DNA from museum collections

Nataly Allasi Canales[1,2], Andrew C. Clarke[3], Mark Nesbitt[2], Rafal Gutaker[2]

1 Natural History Museum of Denmark, University of Copenhagen, Denmark
2 Royal Botanic Gardens, Kew, United Kingdom
3 Future Food Beacon of Excellence & School of Biosciences, University of Nottingham, United Kingdom

Nataly O. Allasi Canales **allasicanales@gmail.com**
Andrew C. Clarke **andrew.clarke1@nottingham.ac.uk**
Mark Nesbitt **m.nesbitt@kew.org**
Rafal Gutaker **r.gutaker@kew.org**

Citation: Canales NA, Clarke AC, Nesbitt M, Gutaker R (2022) Chapter 2. DNA from museum collections. In: de Boer H, Rydmark MO, Verstraete B, Gravendeel B (Eds) Molecular identification of plants: from sequence to species. Advanced Books. https://doi.org/10.3897/ab.e98875

Introduction

Museum collections of plant origin include herbaria (pressed plants), xylaria (woods), and economic botany (useful plant) specimens. They are not only places of history and display, but also of research, and contain rich repositories of molecules, including DNA. Such DNA, retrieved from historical or ancient tissue, carries unique degradation characteristics and regardless of its age is known as ancient DNA (aDNA). Research into aDNA has developed rapidly in the last decade as a result of an improved understanding of its biochemical properties, the development of specific laboratory protocols for its isolation, and better bioinformatic tools. Why are museum collections useful sources of aDNA? We identify three main reasons: 1) specimens can play a key role in taxonomic and macroevolutionary inference when it is difficult to sample living material, for example, by giving us snapshots of extinct taxa (Van de Paer et al. 2016); 2) accurate identification of specimens that were objects of debate or scientific mystery, as exemplified by misidentified type specimens of the watermelon's progenitor (Chomicki and Renner 2015); 3) specimens can provide us with 'time machines' to study microevolutionary processes and diversity changes over decades- to millennia-long timeframes (Gutaker and Burbano 2017; Pont et al. 2019). In all three cases, specimens are often associated with evidence of their occurrence in space and time. For further examples see Chapter 20 Museomics, and the Glossary.

However, extracting DNA does mean the destruction of a part of the specimen. Museum curators therefore face challenges in balancing the conservation of specimens for future research with the rising demand for aDNA analysis. Increasingly, curators are also considering legal and ethical issues in sampling (Austin et al. 2019; Pálsdóttir et al. 2019). Close collaboration between the aDNA researcher and the curatorial staff of museums is therefore essential for appropriate management of these issues (Freedman et al. 2018).

Ethical and legal aspects

With few exceptions, plant material found in museums originally grew on lands tended or owned by people for many millennia (Ellis et al. 2021). Some specimens, such as artefacts or seeds of domesticated crops have an even more direct connection to human activities. Plant specimens, along with other living things, are therefore not simply assemblages of chemical compounds such as DNA, but also embody spiritual beliefs, diverse forms of ownership, traditional knowledge, and past histories of colonialism and other forms of harm (Anderson et al. 2011; Das and Lowe 2018; Pungetti et al. 2012). The implications of this are still being worked out in dialogues between museums and affected communities, often within a decolonising framework (McAlvay et al. 2021). There are, however, immediate steps that researchers and curators can take to ensure that the use of specimens is both legal and ethical.

A first consideration is whether the plant species or artefacts (such as baskets or wooden objects) are of special significance (e.g., sacred) to the source community. Examples of sacred material include *Banisteriopsis caapi*, used to make ayahuasca in South America (Rivier and Lindgren 1972), or *Duboisia hopwoodii* (pituri), used as tobacco in Australia (Ratsch et al. 2010). An online literature search or consultation with relevant experts will give a rapid pointer, which can be followed up with source communities in the study region. Collaboration with communities and scientists in source countries is essential for acknowledging the rights to plant material (even if not legally enshrined), and can be furthered by publication of results in local languages

and media. These communities also hold significant expertise on plants that will improve the quality and relevance of research (Gewin 2021).

There are international conventions that usually apply when accessing, researching, and moving plant material between institutions and countries. Researchers must also be aware of country-specific laws that may require further permits and inspections, e.g., for plants that produce controlled substances, require phytosanitary checks, or are considered invasive species. Legal elements of the Convention on Biological Diversity (CBD), Nagoya Protocol, and Convention on Trade in Endangered Species (CITES) are covered in Chapter 27 Legislation and policy as well as in other published works (e.g. McManis and Pelletier 2014, Iob and Botigué 2021). While the CBD applies to specimens received by museums from 1992, in ethical terms (and under some implementations of the Nagoya Protocol) its principles, such as benefit-sharing, also apply to pre-1992 specimens (cf. Sherman and Henry 2020).

Sampling museum collections

Locating collections and specimens

Botanical gardens hold living specimens and distribute seeds of these via seed lists (Index Seminum). Their global collections can be searched via PlantSearch, hosted by Botanic Gardens Conservation International. Gene banks hold seeds, and sometimes also tissue and living plants. While they originally focused on crop plants and their wild relatives, many have now broadened in scope to include wild plants, such as Royal Botanic Gardens Kew's Millennium Seed Bank. Many gene bank collections can be searched via Genesys. Herbaria hold dried plant specimens and can be located via Index Herbariorum. Although many herbaria are incompletely recorded in databases, substantial data can already be found in the Global Biodiversity Information Facility (GBIF) (Bieker and Martin 2018). Plants are present in abundance in almost all forms of human activity, and it is therefore not surprising that plant material can also be found outside the confines of herbaria, including in economic botany or ethnobotany collections (Salick et al. 2014), agricultural museums, and anthropology collections. Increasing awareness of the importance of biological collections, their uses, conservation efforts and crosslinks among them, is leading to important initiatives that integrate all digitised natural science collections from natural history museums, universities, and botanic gardens (Bakker et al. 2020).

There are a number of pitfalls when searching online catalogues. It may be necessary to search for accepted names and common synonyms: the same species may appear under different botanical names in a single collection, and accuracy of specimen identification varies. In general, herbarium specimens are the most reliable, as they bear diagnostic criteria such as flowers on which taxonomists rely. Garden material and seeds are often misidentified, or become confused in labelling, or are hybridised during repeated cultivations. Their identifications should be confirmed, for example growing on the seeds or by using morphological criteria (Nesbitt et al. 2003). Additionally, data may be missing, unspecific, or incorrectly transcribed or presented, in derived databases, for example in the case of georeferencing (Maldonado et al. 2015).

Researcher-curator collaboration

Research projects will benefit enormously from a close collaboration between researcher and curator. Museums should be approached early during a project, with the researcher providing

sufficient detail about its background, aims, methodology, and timetable. Museums are often under-staffed and persistence may be required in making contact. Curators' expertise will be crucial in identifying the most appropriate specimens for analysis, not only in their institutions, but in others with which they are familiar. The curator will also play a key role in assessing the provenance of specimens, using museum archives, and the implications for any of the ethical and legal issues addressed above. Curators often have good links to source communities and can advise on appropriate procedures.

After preliminary discussions, the researcher will usually need to fill in a 'destructive sampling' form. This acts as a permanent record of the justification for sampling, and allows the museum to make a detailed check on the aims and methodology of the project (see for example, British Museum form and policies). Requests that have unclear research aims or which employ inappropriate methodologies are unlikely to be approved. Researchers will likely need to sign a Material Transfer Agreement (MTA) or Material Supply Agreement (MSA) with the museum which sets out their legal responsibilities.

Sampling may be carried out by the researcher or the curator. If feasible, it is worthwhile for the researcher to carry out the sampling, as it allows for the investigation of the context of the specimen and for flexibility in choosing the samples. It may also speed up the process of obtaining samples, especially if a large number is required. It also allows samples to be safely hand-carried to the researcher's laboratory. Where materials must be sent, it is safest to use a courier service, with specimens marked "Scientific specimens of no commercial value".

It should be agreed with the museum whether, after sampling, surplus material should be returned or securely retained. Museums can require that they are informed about results and that they check manuscripts before publication. This is in any case good practice to ensure accurate reporting of sample details. Museum policies on co-authorship vary, and this topic should be discussed early. Significant contribution by the curator on the choice of appropriate samples, provenance research, or in technically complex sampling, merits co-authorship. Unless agreed otherwise, DNA sequencing data should be submitted to NCBI GenBank or other public repositories, taking care to give the correct specimen identifier. At a minimum, the museum's unique catalogue number (if one exists), and the name of the museum should be cited. This allows the DNA sequence data to be linked directly with the specimen or object. Other museum and laboratory information may be included with the DNA sequence data or in publications (e.g., the collector name, collection number, dates, locations, and laboratory extraction numbers). Additionally, most museum collections will require that vouchers are annotated in a way that links them to DNA sequencing data (see below). Some museums have also started to permanently store DNA isolates, and we encourage researchers to share their stocks on request. Integrated data management and accessibility of the raw data and results will ultimately bolster curatorial practices, develop a more ethical science, and safeguard collections for future generations (Schindel and Cook 2018). Useful guidance on documentation issues is available from the Global Genome Biodiversity Network (GGBN).

Choice of specimens and sampling

Sampling decisions will be determined both by the research design and the nature of the specimens, in addition to the legal and ethical factors mentioned above. Changes to agreed sampling lists are often necessary once specimens have been examined, for example when they are lost, in poor condition, inadequately annotated or georeferenced, present in small quantities, or of rare taxa. Bulk raw material is usually easy to sample, while objects are usually not subjected to destructive sampling unless the results will inform the history and significance of the

object. For herbarium specimens, preserving the morphological features, especially those that are diagnostic, for future research, is critical. Sampling should be targeted towards tissue types or organs at a given developmental state that are most numerous. For example, if there are many flowers and few leaves, it may be preferable to sample a petal. Or if there are few cauline and many rosette leaves, it may be preferable to sample a rosette leaf.

Different parts of a specimen may yield varying amounts, quality, and types of DNA. Wood, husks, and other tissues that were undergoing senescence at the time of preservation may yield less DNA. Young, immature leaves will have higher cell densities, and therefore are expected to yield more DNA. Seeds are often excellent sources of nuclear DNA, although the genotype of the seed will differ from the parent plant and might be of inconsistent ploidy. It may be necessary to extract DNA from individual seeds or to remove maternal tissue such as the testa. Some herbarium sheets will contain multiple individuals and, in most cases, it is better to sample individuals rather than mixed material. If individuals are pooled for DNA extraction, it may complicate downstream analyses that depend on individual genotypes.

The method of specimen preservation is another consideration for DNA isolation. Desiccation has been shown to preserve plant DNA remarkably well, while charring or ethanol preservation destroys plant DNA almost completely (Forrest et al. 2019; Nistelberger et al. 2016). Although not commonly used for aDNA analysis, ancient waterlogged (saturated with water) specimens have a potential for high endogenous contents as they are usually preserved in cold temperatures (Wagner et al. 2018; Wales et al. 2014).

Before sampling begins, the specimen's identifying data, such as its herbarium ID, should be recorded with great care, and double-checked on both the sample label and typed list of specimens. Additionally, the museum may require that vouchers are annotated with the sampling date, tissue type, sample identifier, and information about the researchers. The voucher, including any labels, should be photographed, ideally before and after sampling. Digital links between herbarium vouchers, imaging, and DNA sequences are very useful; they can be included in herbarium and nucleotide databases.

For desiccated leaves, the most commonly sampled tissue, the process is usually straightforward. Using forceps and a scalpel or scissors one can make a precise cut and remove 1 cm^2 or less of tissue. Generally, between 2 and 10 mg of dry leaf tissue is sufficient for the isolation of complex mixtures of genomic DNA fragments. It is preferable that leaves of lesser value are targeted, for example damaged, folded, or hidden, avoiding possible contamination by mould, lichen, or fungi. The sampling of detached "pocket" material should be conducted with caution, and only if the researcher and curator are confident that the detached material truly belongs to the voucher. For other tissue types, such as wood, researchers may need to develop tailored sampling methods on contemporary material first. After sampling, material should immediately be sealed in a labelled tube or envelope and packaged for transport.

Surface contamination

Potential contamination of the sample, specimen, or wider collection with exogenous DNA is an important consideration. For most museum collections, there will inevitably already be surface DNA contamination of specimens. Ask the curator about adhesives (e.g., wheat starch) and preservatives that were used with the specimen of interest. Curatorial staff and other users of the collections may not routinely wear gloves or, if they do, may not change them between specimens. In most cases, there is unlikely to be any benefit from the person undertaking sampling wearing protective equipment (e.g., face masks, hair nets) that is beyond that normally used by users of the collection. Contamination control is only as good as the weakest link.

Extra precautions may be taken for equipment that is used directly in the sampling process, for example, disposable scalpels that are changed between samples, or wiping of scalpel blades with bleach and ethanol. This will reduce the risk of cross-contamination between specimens. Further precautions may be beneficial if internal tissue is being sampled (e.g., inside a seed). In these cases, surface decontamination (see section below on pre-processing) followed by sampling with DNA-free equipment and while wearing personal protective equipment may be appropriate. In some cases where specialistic equipment such as microdrill is required, it may be beneficial for sampling to be undertaken within an ancient DNA laboratory, where contamination controls can be better implemented, however bringing large amounts of plant material into the laboratory should be limited as it is an additional contamination source.

Contamination of specimens and collections by 'modern' DNA and especially amplified DNA is perhaps the greatest risk, potentially compromising future research. Researchers are likely to have been using molecular laboratories, and steps should be taken to prevent the inadvertent transfer of modern DNA to museum collections. These precautions can include not visiting a collection directly from a modern laboratory, cleaning items that must move between modern laboratories and collections (e.g., clothes, phones, cameras), and using sampling equipment (scalpels, tubes, pens) that has not been taken from a modern laboratory.

Laboratory work with historical samples

Understanding aDNA traits

Before starting any experiments with historical and ancient plant samples, it is important to recognize challenges arising from the degraded nature of aDNA. Unlike DNA isolated from fresh samples, DNA from preserved specimens is fragmented, damaged, and contaminated post mortem (Gutaker and Burbano 2017), that includes even recently collected herbarium specimens (Weiß et al. 2016) and contamination with exogenous DNA (Bieker et al. 2020). Fragmentation describes the accumulation of breaks in the DNA backbone, leading to shorter DNA molecules. Breaks occur more often next to guanine or adenine bases, and this can be visualised in sequencing data with dedicated software (Jónsson et al. 2013). The median expected fragment length for aDNA from herbarium specimens is between 30-90 base pairs (bp) in unheated recent Arabidopsis extractions (Bakker 2019; Weiß et al. 2016). It is important to recognise that fragments shorter than 35 bp might generate spurious alignments due to microbial mismapping (Prüfer et al. 2010). The short length of aDNA fragments calls for special molecular methods that allow the retention of short molecules, as well as conservative bioinformatic settings during data processing.

aDNA is also affected by "damage", post mortem substitutions that convert cytosine to uracil residues through deamination (uracils are read by insensitive DNA polymerases as thymine, hence the commonly used term "C-to-T substitutions'') (Hofreiter et al. 2001). This process occurs preferentially at the ends of DNA molecules (Briggs et al. 2007), particularly with single-stranded DNA overhangs (Overballe-Petersen et al. 2012). Consequently, in the population of sequenced molecules, an elevated number of C-to-T substitutions are observed at the 5′ end, and complementary G-to-A substitutions at the 3′ end. Typically, herbarium-isolated DNA has around from 1 to 6% (in older samples) of cytosine residues converted to thymine (Durvasula et al. 2017; Gutaker et al. 2019; Weiß et al. 2016), while in archaeological material this number might be as high as 30%. These post mortem substitutions should be removed before downstream analyses.

Finally, it is important to recognize that aDNA from plants is in fact a mixture of bona fide endogenous DNA, exogenous DNA introduced pre mortem, (e.g., from endophytic microbes),

and exogenous DNA introduced post mortem (e.g., from microbes involved in decomposition, human-associated collection and museum practices; see above) (Pääbo et al. 2004). Quantification of contamination is commonly done by dividing the number of sequence reads that map to the target reference genome by the total number of sequenced reads from the museum sample. In fresh material, the ratio is often around 0.98; in degraded material it can vary from 0 to 0.95 (Gutaker et al. 2017). Several examples of aDNA successfully obtained from plants are illustrated in Table 1.

Table 1. Examples of selected successfully isolated and sequenced DNA from plant material. *BP: before present.

Species	Tissue	Age BP*	Endogenous DNA	Fragment length (bp)	Damage at 5′ end	Source
Thale cress (*Arabidopsis thaliana*)	Leaf	184	83%	~62	0.026	Durvasula et al. 2017
Potato (*Solanum tuberosum*)	Leaf	361	87%	~45	0.047	Gutaker et al. 2019
Maize (*Zea mays*)	Cobs	1863	80%	~52	0.052	Swarts et al. 2017
Wheat (*Triticum durum*)	Chaff	3150	40%	~53	0.095	Scott et al. 2019
Barley (*Hordeum vulgare*)	Seeds	4988	86%	~49	0.138	Mascher et al. 2016

Recommended working practices for aDNA

Given the characteristics of aDNA (Dabney et al. 2013) and the fact that it is very prone to contamination at any stage, guidelines have been proposed to facilitate the authentication process, and minimise potential contamination before, during, and after DNA extraction (Pääbo et al. 2004). We strongly recommend following gold-standard precautions when working with aDNA (Fulton and Shapiro 2019; Latorre et al. 2020).

The isolation and pre-amplification manipulation of aDNA should be carried out in a dedicated laboratory that is physically separated from labs where post-amplification steps are carried out. Ideally the aDNA laboratory should be supplied with HEPA-filtered air under positive pressure. Users should not move from a 'modern' laboratory (where amplified DNA is handled) to the aDNA laboratory on the same day. Reagents and materials in an aDNA lab should be DNA-free, disposable where possible, and never taken out of the clean lab. Surfaces should be cleaned before and after every experiment with 3–10% bleach, 70% ethanol, and overnight UV-C irradiation. To minimise contamination and ensure a DNA-free laboratory environment, users should wear full body suits, foot protectors, slippers, facemasks, sleeves, and double gloves (Fulton and Shapiro 2019). Together, these precautions limit cross-contamination from amplified and unamplified DNA.

Material preparation is an essential step before DNA can be isolated. Optional pre-processing of dirty samples can be done by gently cleaning the surface with a very low concentration (~3%) of bleach, and rinsing twice with ddH_2O (Cappellini et al. 2010). When handling waterlogged, fragile, or permeable material, avoid using bleach and carry out ddH_2O treatment only. To help identify contamination that might be introduced in the laboratory, samples should always be processed alongside negative controls, including for DNA isolation and library preparations. To reduce the likelihood of cross-contamination, small batches of up to 12 samples at a time are preferable (Latorre et al. 2020).

DNA extraction methods for different tissues should be considered. While plant materials tend to contain inhibitory substances like polyphenols, proteins, and polysaccharides, ancient

plant materials can additionally be rich in humic acids and salts. This set of macromolecules might prevent successful DNA amplification (Wales et al. 2014) by affecting polymerase activity (Schrader et al. 2012). To reduce this inhibitory effect, smaller amounts of sample can be extracted in parallel, and the resulting DNA pooled to achieve a sufficient yield (Wagner et al. 2018).

Here we will cover the basics of recovering the highest quality of DNA from ancient plant tissues. Using a two-day extraction protocol will greatly increase the recovery of endogenous DNA. The first day consists of grinding the plant material. Tissue can be disrupted by: grinding dry, grinding flash-frozen, or grinding material soaked in lysis buffer. In all cases, grinding to finer particles increases the recovery of aDNA. Ground tissue is incubated in a fresh lysis buffer. Three commonly used buffers include CTAB (Kistler 2012), DTT (Wales and Kistler 2019), or PTB mixtures (Latorre et al. 2020). The second day is dedicated to isolating DNA from the lysate. Initial removal of non-DNA particles can be achieved by centrifugation with a shredding column (Latorre et al. 2020) or phenol/chloroform mixture (Kistler 2012; Wales and Kistler 2019; Wagner et al. 2018). In all methods, DNA is then captured in various DNA-binding silica columns (for example QIAgen MinElute columns) and purified (Dabney et al. 2013). Elution from silica columns produces the final, isolated aDNA.

By contrast to primed amplification approaches, even low amounts of isolated DNA can be used for genomic library preparation (Staats et al. 2013) and hence we recommend that a genomic library is constructed using a well-established method (Carøe et al. 2017; Kircher et al. 2012; Meyer and Kircher 2010; Meyer et al. 2012). Quantification of genomic DNA before sequencing using RT-qPCR allows the number of amplification cycles for each sample to be adjusted, in turn allowing the complexity of sequenced DNA fragments to be maximised. Bioinformatic pre-processing is an essential part of aDNA analyses, and is summarised in three available pipelines (Latorre et al. 2020; Peltzer et al. 2016; Schubert et al. 2014). Authentication is another crucial step in bioinformatic analyses that can currently be best achieved with mapDamage software (Jónsson et al. 2013).

Choosing and authenticating aDNA samples

To help decide which sampled material is most promising for further DNA analyses it is necessary to obtain good estimates for fragmentation, damage, and contamination. This can be achieved through sequencing genomic libraries in low-throughput mode (about 10,000 DNA reads per sample), commonly referred to as "screening" and bioinformatic analyses that produce relevant summary statistics. Promising samples will contain aDNA with a median fragment length over 50 bp and endogenous content over 0.2. For samples of particular interest, mapping the accuracy for short aDNA reads can be improved with specialised procedures (de Filippo et al. 2018), and endogenous content can be increased by targeted enrichment on hybridization arrays (Hodges et al. 2009) or 'in solution' (Maricic et al. 2010). Finally, one should pay attention to the frequency of C-to-T substitutions at the ends of the sequenced reads. Samples with 2–6% C-to-Ts can be corrected bioinformatically (by trimming ends or filtering transitions), while a higher percentage of C-to-Ts can be remedied through more effective enzymatic removal of uracil (Briggs et al. 2010).

Characterising DNA fragmentation and damage is very useful for authentication and establishing historical provenance of degraded plant samples. DNA degradation advances with time (Weiß et al. 2016), although its rate is highly modulated by intrinsic and environmental factors. Old samples should be considered authentic only if they exhibit fragmentation and damage patterns congruent with their age, tissue type, and storage conditions. In contrast to library-based approaches, primer-based sequencing (such as Sanger sequencing) does not allow quantification of these characteristics and should not be used with aDNA (Gutaker and Burbano 2017).

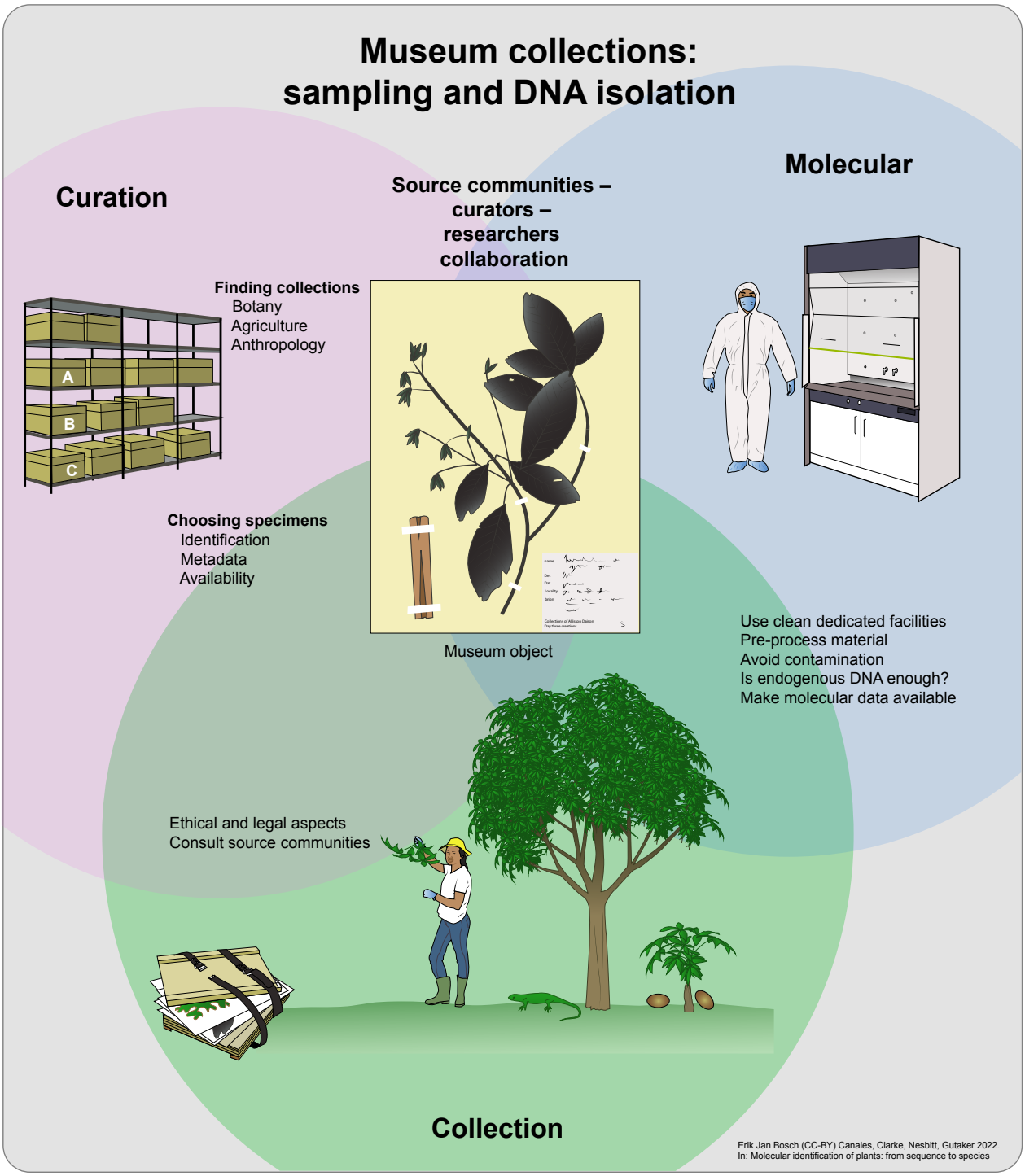

Figure 1. Chapter 2 Infographic: Overview of sampling and obtaining DNA from museum collections. An team effort of communities, curators and researchers (1) Collection of botanical material should have detailed consideration of its ethical and legal aspects and the consultation of source communities in advance, in accordance with CITES, CBD and Nagoya legal and ethical frameworks. (2) Curated botanical samples can be found in different types of museums that include botanic gardens, ethnobotany and anthropological collections. The next step is to find relevant specimens with preferably rich metadata, e.g. species identification, collection place and date. (3) Once the specimens have been identified, they should undergo molecular analyses in clean facilities. Where they will be pre-processed according to their traits, avoiding contamination with other samples, "modern" specimens, and amplicons. Then, it is crucial to identify samples that failed and passed quality controls for endogenous DNA. Finally, the data produced should be linked to their respective vouchers and made available in public repositories like NCBI and BOLD.

Responsible lab use for aDNA

Library-based methods assist with the responsible use of collections, as they preserve the total (non-selective) DNA and 'immortalise' it for future use. Immortalisation only has value if the DNA that has been amplified is truly historical/ancient and devoid of contemporary contamination and hence all the aforementioned precautions are necessary when working with aDNA. We recommend that extracts or library builds are precisely annotated with the methods used and are properly archived.

Questions

1. Name three legal considerations and their related ethical main issues that should be taken into account for aDNA research using museum material.
2. Why is it important to process herbarium samples in a dedicated clean lab?
3. Name three benefits of getting curators involved in the early stages of research using collections.

Glossary

aDNA – Ancient DNA, DNA that exhibits biochemical characteristics typical for DNA from old degraded material, i.e., damage and fragmentation, regardless of age.
Artefact – An object made by humans that is of historical or cultural importance, examples include: clothing, ornaments, utensils.
Authentication – Bioinformatic analyses that quantify damage and fragmentation of sequenced DNA to help rule out that DNA is derived from contemporary contamination.
Collection – Repository of curated biological material arranged in a systematic fashion.
Contamination – Introduction of alien tissue or DNA to a specimen or DNA isolate, examples include: microbial colonisation, human epithelium, plant-based foods, etc.
Curator – Custodian of a collection with expert knowledge about specimens, their organisation, and preservation.
Destructive sampling – Permanent removal of a fragment of a specimen of any size that will be irretrievable after biochemical characterization.
DNA damage – Typically conversion of cytosine to uracil in DNA through deamination, which accumulates with time. During sequencing, uracil is replaced with thymine, hence the common synonym, C-to-T substitutions.
Endogenous DNA – Authentic DNA from targeted individuals of a species, in contrast to exogenous DNA from associated microbes and contemporary plant and human DNA contamination.
Fragmentation – Breaks in the DNA backbone, most frequently caused by depurination, leading to shorter DNA fragments with time.
Immortalization – Molecular manipulation of DNA, for example the attachment of DNA adapters, that allows infinite re-amplification of the original DNA from a biological specimen.
Type specimen – Preserved individual plant that has defining features of that taxon that is used for the first taxonomic description of a species. This permanent feature-specimen link is recognized in a publication.

Voucher – Preserved botanical specimen kept in permanent collection and cited by research project. Vouchers will have been expertly identified and are usually annotated with collection time, place, and collector details.

References

Anderson EN, Pearsall D, Hunn E, Turner N (2011) Ethnobiology. John Wiley & Sons, Inc., Hoboken, NJ, USA. https://doi.org/10.1002/9781118015872

Austin RM, Sholts SB, Williams L, Kistler L, Hofman CA (2019) Opinion: To curate the molecular past, museums need a carefully considered set of best practices. Proc Natl Acad Sci USA 116, 1471–1474. https://doi.org/10.1073/pnas.1822038116

Bakker FT, Antonelli A, Clarke JA, Cook JA, Edwards SV, Ericson PGP, Faurby S, Ferrand N, Gelang M, Gillespie RG, Irestedt M, Lundin K, Larsson E, Matos-Maraví P, Müller J, von Proschwitz T, Roderick GK, Schliep A, Wahlberg N, Wiedenhoeft J, Källersjö M (2020) The Global Museum: natural history collections and the future of evolutionary science and public education. PeerJ 8, e8225. https://doi.org/10.7717/peerj.8225

Bakker FT (2019) Herbarium genomics: plant archival DNA explored, in: Lindqvist, C., Rajora, O.P. (Eds.) Paleogenomics: Genome-Scale Analysis of Ancient DNA, Population Genomics. Springer International Publishing, Cham, pp. 205–224. https://doi.org/10.1007/13836_2018_40

Bieker VC, Martin MD (2018) Implications and future prospects for evolutionary analyses of DNA in historical herbarium collections. Botany Letters 165, 1–10. https://doi.org/10.1080/23818107.2018.1458651

Bieker VC, Sánchez Barreiro F, Rasmussen JA, Brunier M, Wales N, Martin MD (2020) Metagenomic analysis of historical herbarium specimens reveals a postmortem microbial community. Mol. Ecol. Resour. 20, 1206–1219. https://doi.org/10.1111/1755-0998.13174

Briggs AW, Stenzel U, Johnson PLF, Green RE, Kelso J, Prüfer K, Meyer M, Krause J, Ronan MT, Lachmann M, Pääbo S (2007) Patterns of damage in genomic DNA sequences from a Neandertal. Proc Natl Acad Sci USA 104, 14616–14621. https://doi.org/10.1073/pnas.0704665104

Briggs AW, Stenzel U, Meyer M, Krause J, Kircher M, Pääbo S (2010) Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. Nucleic Acids Res. 38, e87. https://doi.org/10.1093/nar/gkp1163

Cappellini E, Gilbert MTP, Geuna F, Fiorentino G, Hall A, Thomas-Oates J, Ashton PD, Ashford DA, Arthur P, Campos PF, Kool J, Willerslev E, Collins MJ (2010) A multidisciplinary study of archaeological grape seeds. Naturwissenschaften 97, 205–217. https://doi.org/10.1007/s00114-009-0629-3

Carøe C, Gopalakrishnan S, Vinner L, Mak SST, Sinding MHS, Samaniego JA, Wales N, Sicheritz-Pontén T, Gilbert MTP (2017) Single-tube library preparation for degraded DNA. Methods Ecol. Evol. 9, 410–419. https://doi.org/10.1111/2041-210X.12871

Chomicki G, Renner SS (2015) Watermelon origin solved with molecular phylogenetics including Linnaean material: another example of museomics. New Phytol. 205, 526–532. https://doi.org/10.1111/nph.13163

Dabney J, Meyer M, Pääbo S (2013) Ancient DNA damage. Cold Spring Harb. Perspect. Biol. 5, a012567. https://doi.org/10.1101/cshperspect.a012567

Das S, Lowe M (2018) Nature read in black and white: decolonial approaches to interpreting natural history collections. Journal of Natural Science Collections 6, 1–14.

de Filippo C, Meyer M, Prüfer K (2018) Quantifying and reducing spurious alignments for the analysis of ultra-short ancient DNA sequences. BMC Biol. 16, 121. https://doi.org/10.1186/s12915-018-0581-9

Durvasula A, Fulgione A, Gutaker RM, Alacakaptan SI, Flood PJ, Neto C, Tsuchimatsu T, Burbano HA, Picó FX, Alonso-Blanco C, Hancock AM (2017) African genomes illuminate the early history and transition to selfing in *Arabidopsis thaliana*. Proc Natl Acad Sci USA 114, 5213–5218. https://doi.org/10.1073/pnas.1616736114

Ellis EC, Gauthier N, Klein Goldewijk K, Bliege Bird R, Boivin N, Díaz S, Fuller DQ, Gill JL, Kaplan JO, Kingston N, Locke H, McMichael CNH, Ranco D, Rick TC, Shaw MR, Stephens L, Svenning J-C, Watson JEM (2021) People have

shaped most of terrestrial nature for at least 12,000 years. Proc Natl Acad Sci USA 118, e2023483118. https://doi.org/10.1073/pnas.2023483118

Forrest LL, Hart ML, Hughes M, Wilson HP, Chung K-F, Tseng Y-H, Kidner CA (2019) The limits of Hyb-Seq for herbarium specimens: impact of preservation techniques. Front. Ecol. Evol. 7, 439. https://doi.org/10.3389/fevo.2019.00439

Freedman J, van Dorp L, Brace S (2018) Destructive sampling natural science collections: An overview for museum professionals and researchers. Journal of Natural Science Collections.

Fulton TL, Shapiro B (2019) Setting up an ancient DNA laboratory. Methods Mol. Biol. 1963, 1–13. https://doi.org/10.1007/978-1-4939-9176-1_1

Gewin V (2021) How to include Indigenous researchers and their knowledge. Nature 589, 315–317. https://doi.org/10.1038/d41586-021-00022-1

Gutaker RM, Burbano HA (2017) Reinforcing plant evolutionary genomics using ancient DNA. Curr. Opin. Plant Biol. 36, 38–45. https://doi.org/10.1016/j.pbi.2017.01.002

Gutaker RM, Reiter E, Furtwängler A, Schuenemann VJ, Burbano HA (2017) Extraction of ultrashort DNA molecules from herbarium specimens. BioTechniques 62, 76–79. https://doi.org/10.2144/000114517

Gutaker RM, Weiß CL, Ellis D, Anglin NL, Knapp S, Luis Fernández-Alonso J, Prat S, Burbano HA (2019) The origins and adaptation of European potatoes reconstructed from historical genomes. Nat. Ecol. Evol. 3, 1093–1101. https://doi.org/10.1038/s41559-019-0921-3

Hodges E, Rooks M, Xuan Z, Bhattacharjee A, Benjamin Gordon D, Brizuela L, Richard McCombie W, Hannon GJ (2009) Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. Nat. Protoc. 4, 960–974. https://doi.org/10.1038/nprot.2009.68

Hofreiter M, Serre D, Poinar HN, Kuch M, Pääbo S (2001) Ancient DNA. Nat. Rev. Genet. 2, 353–359. https://doi.org/10.1038/35072071

Iob A, Botigué L (2021) Crop archaeogenomics: a powerful resource in need of a well-defined regulation framework. Plants, People, Planet 4, 44–50. https://doi.org/10.1002/ppp3.10233

Jónsson H, Ginolhac A, Schubert M, Johnson PLF, Orlando L (2013) mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics 29, 1682–1684. https://doi.org/10.1093/bioinformatics/btt193

Kircher M, Sawyer S, Meyer M (2012) Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res. 40, e3. https://doi.org/10.1093/nar/gkr771

Kistler L (2012) Ancient DNA extraction from plants. Methods Mol. Biol. 840, 71–79. https://doi.org/10.1007/978-1-61779-516-9_10

Latorre SM, Lang PLM, Burbano HA, Gutaker RM (2020) Isolation, library preparation, and bioinformatic analysis of historical and ancient plant DNA. Curr. Protoc. Plant Biol. 5, e20121. https://doi.org/10.1002/cppb.20121

Maldonado C, Molina CI, Zizka A, Persson C, Taylor CM, Albán J, Chilquillo E, Rønsted N, Antonelli A (2015) Estimating species diversity and distribution in the era of Big Data: to what extent can we trust public databases? Glob. Ecol. Biogeogr. 24, 973–984. https://doi.org/10.1111/geb.12326

Maricic T, Whitten M, Pääbo S (2010) Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. PLoS ONE 5, e14004. https://doi.org/10.1371/journal.pone.0014004

Mascher M, Schuenemann VJ, Davidovich U, Marom N, Himmelbach A, Hübner S, Korol A, David M, Reiter E, Riehl S, Schreiber M, Vohr SH, Green RE, Dawson IK, Russell J, Kilian B, Muehlbauer GJ, Waugh R, Fahima T, Krause J, Stein N (2016) Genomic analysis of 6,000-year-old cultivated grain illuminates the domestication history of barley. Nat. Genet. 48, 1089–1093. https://doi.org/10.1038/ng.3611

McAlvay AC, Armstrong CG, Baker J, Elk LB, Bosco S, Hanazaki N, Joseph L, Martínez-Cruz TE, Nesbitt M, Palmer MA, Priprá de Almeida WC, Anderson J, Asfaw Z, Borokini IT, Cano-Contreras EJ, Hoyte S, Hudson M, Ladio AH, Odonne G, Peter S, Rashford J, Wall J, Wolverton S, Vandebroek I (2021) Ethnobiology phase VI: decolonizing institutions, projects, and scholarship. J. Ethnobiol. 41, 170–191. https://doi.org/10.2993/0278-0771-41.2.170

McManis CR, Pelletier JS (2014) Legal aspects of biocultural collections, in: Salick, J., Konchar, K., Nesbitt, M. (Eds.) Presented at the Curating biocultural collections: a handbook, Royal Botanic Gardens, Kew, Kew, pp. 229–243.

Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K, de Filippo C, Sudmant PH, Alkan C, Fu Q, Do R, Rohland N, Tandon A, Siebauer M, Green RE, Bryc K, Briggs AW, Pääbo S (2012) A high-coverage genome sequence from an archaic Denisovan individual. Science 338, 222–226. https://doi.org/10.1126/science.1224344

Meyer M, Kircher M (2010) Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb. Protoc. 2010, pdb.prot5448. https://doi.org/10.1101/pdb.prot5448

Nesbitt M, Colledge S, Murray MA (2003) Organisation and management of seed reference collections. Environmental Archaeology 8, 77–84. https://doi.org/10.1179/env.2003.8.1.77

Nistelberger HM, Smith O, Wales N, Star B, Boessenkool S (2016) The efficacy of high-throughput sequencing and target enrichment on charred archaeobotanical remains. Sci. Rep. 6, 37347. https://doi.org/10.1038/srep37347

Overballe-Petersen S, Orlando L, Willerslev E (2012) Next-generation sequencing offers new insights into DNA degradation. Trends Biotechnol. 30, 364–368. https://doi.org/10.1016/j.tibtech.2012.03.007

Pääbo S, Poinar H, Serre D, Jaenicke-Despres V, Hebler J, Rohland N, Kuch M, Krause J, Vigilant L, Hofreiter M (2004) Genetic analyses from ancient DNA. Annu. Rev. Genet. 38, 645–679. https://doi.org/10.1146/annurev.genet.37.110801.143214

Pálsdóttir AH, Bläuer A, Rannamäe E, Boessenkool S, Hallsson JH (2019) Not a limitless resource: ethics and guidelines for destructive sampling of archaeofaunal remains. R. Soc. Open Sci. 6, 191059. https://doi.org/10.1098/rsos.191059

Peltzer A, Jäger G, Herbig A, Seitz A, Kniep C, Krause J, Nieselt K (2016) EAGER: efficient ancient genome reconstruction. Genome Biol. 17, 60. https://doi.org/10.1186/s13059-016-0918-z

Pont C, Wagner S, Kremer A, Orlando L, Plomion C, Salse J (2019) Paleogenomics: reconstruction of plant evolutionary trajectories from modern and ancient DNA. Genome Biol. 20, 29. https://doi.org/10.1186/s13059-019-1627-1

Prüfer K, Stenzel U, Hofreiter M, Pääbo S, Kelso J, Green RE (2010) Computational challenges in the analysis of ancient DNA. Genome Biol. 11, R47. https://doi.org/10.1186/gb-2010-11-5-r47

Pungetti G, Oviedo G, Hooke D (2012) Sacred species and sites: advances in biocultural conservation. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781139030717

Ratsch A, Steadman KJ, Bogossian F (2010) The pituri story: a review of the historical literature surrounding traditional Australian Aboriginal use of nicotine in Central Australia. J. Ethnobiol. Ethnomed. 6, 26. https://doi.org/10.1186/1746-4269-6-26

Rivier L, Lindgren J-E (1972) "Ayahuasca," the South American hallucinogenic drink: an ethnobotanical and chemical investigation. Econ. Bot. 26, 101–129. https://doi.org/10.1007/BF02860772

Salick J, Konchar K, Nesbitt M (2014) Curating Biocultural Collections: A Handbook. Royal Botanic Gardens, Kew, Kew.

Schindel DE, Cook JA (2018) The next generation of natural history collections. PLoS Biol. 16, e2006125. https://doi.org/10.1371/journal.pbio.2006125

Schrader C, Schielke A, Ellerbroek L, Johne R (2012) PCR inhibitors - occurrence, properties and removal. J. Appl. Microbiol. 113, 1014–1026. https://doi.org/10.1111/j.1365-2672.2012.05384.x

Schubert M, Ermini L, Der Sarkissian C, Jónsson H, Ginolhac A, Schaefer R, Martin MD, Fernández R, Kircher M, McCue M, Willerslev E, Orlando L (2014) Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. Nat. Protoc. 9, 1056–1082. https://doi.org/10.1038/nprot.2014.063

Scott MF, Botigué LR, Brace S, Stevens CJ, Mullin VE, Stevenson A, Thomas MG, Fuller DQ, Mott R (2019) A 3,000-year-old Egyptian emmer wheat genome reveals dispersal and domestication history. Nat. Plants 5, 1120–1128. https://doi.org/10.1038/s41477-019-0534-5

Sherman B, Henry RJ (2020) The Nagoya Protocol and historical collections of plants. Nat. Plants 6, 430–432. https://doi.org/10.1038/s41477-020-0657-8

Staats M, Erkens RHJ, van de Vossenberg B, Wieringa JJ, Kraaijeveld K, Stielow B, Geml J, Richardson JE, Bakker FT (2013) Genomic treasure troves: complete genome sequencing of herbarium and insect museum specimens. PLoS ONE 8, e69189. https://doi.org/10.1371/journal.pone.0069189

Swarts K, Gutaker RM, Benz B, Blake M, Bukowski R, Holland J, Kruse-Peeples M, Lepak N, Prim L, Romay MC, Ross-Ibarra J, Sanchez-Gonzalez J de J, Schmidt C, Schuenemann VJ, Krause J, Matson RG, Weigel D, Buckler ES, Burbano HA (2017) Genomic estimation of complex traits reveals ancient maize adaptation to temperate North America. Science 357, 512–515. https://doi.org/10.1126/science.aam9425

Van de Paer C, Hong-Wa C, Jeziorski C, Besnard G (2016) Mitogenomics of Hesperelaea, an extinct genus of Oleaceae. Gene 594, 197–202. https://doi.org/10.1016/j.gene.2016.09.007

Wagner S, Lagane F, Seguin-Orlando A, Schubert M, Leroy T, Guichoux E, Chancerel E, Bech-Hebelstrup I, Bernard V, Billard C, Billaud Y, Bolliger M, Croutsch C, Čufar K, Eynaud F, Heussner KU, Köninger J, Langenegger F, Leroy F, Lima C, Orlando L (2018) High-Throughput DNA sequencing of ancient wood. Mol. Ecol. 27, 1138–1154. https://doi.org/10.1111/mec.14514

Wales N, Andersen K, Cappellini E, Avila-Arcos MC, Gilbert MTP (2014) Optimization of DNA recovery and amplification from non-carbonized archaeobotanical remains. PLoS ONE 9, e86827. https://doi.org/10.1371/journal.pone.0086827

Wales N, Kistler L (2019) Extraction of ancient DNA from plant remains. Methods Mol. Biol. 1963, 45–55. https://doi.org/10.1007/978-1-4939-9176-1_6

Weiß CL, Schuenemann VJ, Devos J, Shirsekar G, Reiter E, Gould BA, Stinchcombe JR, Krause J, Burbano HA (2016) Temporal patterns of damage and decay kinetics of DNA retrieved from plant herbarium specimens. R. Soc. Open Sci. 3, 160239. https://doi.org/10.1098/rsos.160239

Answers

1. Legal: CITES (restriction in international trade of endangered species), Nagoya Protocol (ownership and other significance to indigenous peoples), and Drug Act (controlled substances).
2. The decay of DNA from historical plant material makes it very susceptible to contamination with exogenous modern DNA.
3. Curators can contribute (1) high-quality metadata such as collection dates and provenance, (2) knowledge of collections in-house and elsewhere, (3) knowledge of source communities and ethical and legal issues, (4) advice on choice of specimens most suitable for sampling.

Chapter 3
DNA from water

Ozan Çiftçi[1], Sevgi Kaynar[2], Physilla Chua[3]

1 Institute of Environmental Sciences, Leiden University, The Netherlands
2 Institute for Biochemistry & Biology, University of Potsdam, Germany
3 Section for Evolutionary Genomics, Globe Institute, University of Copenhagen, Denmark

Ozan Çiftçi **ozancift@gmail.com**
Sevgi Kaynar **sevgi.kaynar@uni-potsdam.de**
Physilla Chua **physiliachua@gmail.com**

Citation: Çiftçi O, Kaynar S, Chua P (2022) Chapter 3. DNA from water. In: de Boer H, Rydmark MO, Verstraete B, Gravendeel B (Eds) Molecular identification of plants: from sequence to species. Advanced Books. https://doi.org/10.3897/ab.e98875

Introduction

The first studies conducted on DNA obtained from water samples were published in the 1990s. Cloning techniques were commonly used to investigate novel genes and functions of environmental communities at that time. Stein et al. (Stein et al. 1996) cloned DNA fragments obtained from water samples into *E. coli* vectors to investigate marine archaea metabolism. Relying on the development of high-throughput sequencing (HTS) technologies, it is now possible to capture and sequence almost all DNA fragments present in a water sample. A pioneering example represents the study, where Venter and colleagues sequenced sea water samples revealing diverse microbial compositions and functions (Venter et al. 2004). Further rapid development of sequencing technologies in the last decade, as well as the coinciding decrease in sequencing costs, has allowed for the incorporation of 'environmental DNA' (eDNA) methods (i.e., the analysis of DNA fragments isolated from environmental sample types such as water, air, and soil) into several applications, including in aquatic environmental surveys.

Conventionally, biomonitoring of freshwater and marine environments is based on direct observation of indicator taxa to compute biotic metrics/indices. This can be time and labour intensive (Pawlowski et al. 2018). Other methods such as depletion-based electrofishing, hydroacoustics, camera traps, and gillnets are also common (Deiner et al. 2017). In recent years, eDNA methods have been added to this toolbox of available methods for biomonitoring. Species-level information on key bioindicator species has for example been obtained by using the DNA obtained from water samples (Hajibabaei et al. 2011). Other applications include population quantification (Fukaya et al. 2020), invasive species detection (Anglès d'Auriac et al. 2019), water quality monitoring (Noyer et al. 2015), and revealing food web interactions (D'Alessandro and Mariani 2021).

The main advantage of water is the ease of sample collection compared to other aquatic sample types such as sediments or biofilms, as these substrates usually require more sophisticated tools and longer sampling times (Deiner et al. 2017). A potential disadvantage is that DNA in a water column decays into undetectable levels in two weeks at most (Dejean et al. 2011; Thomsen et al. 2012), whereas in sediments and ice cores it can persist much longer (Turner et al. 2015). Thus, DNA collected from water samples typically reflects contemporary communities, whereas those collected from sediments and ice cores reflect a longer temporal scale and can be used as a source for ancient DNA (Willerslev et al. 2007) (Chapter 8 DNA from ancient sediments).

Detecting DNA in water samples obtained from aquatic environments can be challenging because it is usually present at low concentrations with an uneven spatial distribution (Ficetola et al. 2008; Goldberg et al. 2016). In this chapter, we first explain the factors affecting the detection of DNA with a specific focus on plant species for environmental applications. The literature referenced here mostly focuses on vascular plants but the general approach might be suitable for a broader group of organisms as well (Alsos et al. 2018; Apothéloz-Perret-Gentil et al. 2021; Nowak et al. 2021). We then outline the general workflow and experimental setup for collecting DNA from water and strategies to optimise its detection.

Detection of DNA from aquatic environments

Natural processes influencing the composition and quantity of detectable DNA in a water sample can be categorised into 1) shedding of biological material from source organisms, 2) degradation, 3) transport across the water column, and 4) retention and resuspension (Harrison et al. 2019). Several biotic and abiotic environmental factors influence the rates of these processes.

This creates a complex and environment-specific relationship between DNA that is detected in the water and how well this can be related to the presence and relative abundance of an aquatic organism. As almost all DNA fragments in a water sample can be detected with current sequencing technologies, establishing an optimal sampling strategy is crucial for minimising the probability of contamination and obtaining an accurate representation of biodiversity.

Shedding

Senescence in aquatic plants releases free cells into the water column that will eventually break down into organic compounds, including DNA. However, degradation in many cells begins via apoptosis before shedding. Apoptosis involves the shrinkage of the cell and its nucleus in a programmed way, in contrast to necrosis, which is uncontrolled cell death due to loss of osmotic control typically by swelling and bursting (Hotchkiss et al. 2009; Toné et al. 2007). In general, plant and animal cells have similar mechanisms of apoptosis with tightly packed nuclear DNA in early stages, which is later hydrolyzed into smaller fragments of about 50 kb and multiples of approximately 180 bp (Reape et al. 2008; Vanyushin et al. 2004). Mitochondrial DNA, on the other hand, shows lower decay rates compared to nuclear DNA, which is attributed to the presence of the mitochondrial membrane or other localised factors (Foran 2006). Possibly owing to such similar mechanisms of cell death, Fujiwara et al. (Fujiwara et al. 2016) showed that temporal changes in the amount of DNA in water samples are similar for an aquatic plant, *Egeria densa,* and carp. However, this relationship is not always significant for plants, as opposed to fish, which is attributed to the differences in cell and tissue structures, cell functions, and metabolic systems (Matsuhashi et al. 2016).

DNA degradation

DNA is a highly stable molecule at neutral pH and moderate temperatures. However, there are several abiotic factors that directly and indirectly influence its stability in aquatic environments (Schroeder and Wolfenden 2007). High temperatures increase degradation rates either by denaturing DNA molecules directly or by increasing metabolic and enzymatic activities that lead to DNA degradation (Eichmiller et al. 2016; Okabe and Shimazu 2007). Ultraviolet light can either directly damage DNA or react with organic matter to form reactive molecules that indirectly damage DNA (Leech et al. 2009; Strickler et al. 2015). Pilliod et al. (2014) detected DNA in water samples after 18 days when kept in the dark after collection, but when exposed to light nothing was detected after eight days. Hypersaline and low oxygen environments can also affect the conformation and stability of DNA (Barnes et al. 2014; Hofreiter et al. 2001). Biotic factors depending on the source organism such as the type of shed tissue, age, size, or life history, or external biotic factors such as microbial activity, trophic state, or the concentration of extracellular nucleases might also influence DNA shedding and persistence in aquatic environments (Beng and Corlett 2020; de Souza et al. 2016; Eichmiller et al. 2016; Harrison et al. 2019). The effect of abiotic factors can be expected to be similar for all free extracellular DNA, so the detection probabilities of aquatic plants might be influenced more by shedding and transport rates prior to the release from the cell.

Transportation, retention, and resuspension

Hydrological characteristics of the water body are also critical to consider when inferring species presence and distribution. DNA can bind to particles of varying size in aquatic environments

(less than 0.2 µm to greater than 180 µm) and this particle association is one of many parameters that affect DNA transport and diffusion (Shogren et al. 2016). DNA is known to persist longer in sediment compared to water columns and this adsorbed portion can be resuspended into the water after aquatic DNA is degraded (Shogren et al. 2018). Microbial decomposition of plant material in freshwater sediments has also been shown to release extra plant DNA into the water column (Poté et al. 2009). The type of the sediment (e.g., clay vs. organic) and binding affinity of DNA are some of the factors that influence these processes (Beng and Corlett 2020; Harrison et al. 2019). In general, DNA transport in aquatic ecosystems follows similar dynamics with the particles categorised as fine particulate organic matter (i.e., between 0.5 µm to 1 mm) (Pont et al. 2018; Wilcox et al. 2016). Filtration methods are therefore usually designed to capture this size range (Harrison et al. 2019; Pont et al. 2018; Wilcox et al. 2016).

Considering the higher dilution and the effects of currents and waves in marine waters, DNA is generally less concentrated and more quickly dispersed compared to freshwater ecosystems (Foote et al. 2012; Thomsen et al. 2012). However, marine waters are also characterised by higher salinity and more stable temperatures which are known to have stabilising effects on DNA molecules (Okabe and Shimazu 2007; Tsuji et al. 2017). In addition to temperature, pH is also known to be more stable in seas and oceans compared to terrestrial aquatic ecosystems (Collins et al. 2018). Under favourable conditions, DNA obtained from marine water samples can distinguish communities less than 60 m apart for up until 6 hours and can persist above the detection limit for several days (Foote et al. 2012; Kelly et al. 2018; Thomsen et al. 2012).

In rivers and streaming waters, the probability of DNA detection is strongly correlated with downstream transportation rates. Retention, rather than degradation, appears to be a more important factor that limits the transport of DNA in streaming waters (Shogren et al. 2018; Wilcox et al. 2016). Considering the wide range of particle sizes that DNA has been found to be associated with, modelling its transport in flowing waters is not an easy task. The transport rate can be influenced by additional factors such as stream bed characteristics or the presence of biofilms (Shogren et al. 2018). More recently, hydrological models have been used to predict the transport and decay rates of DNA in aquatic ecosystems (Carraro et al. 2021, 2020; Mächler et al. 2021). In lakes and ponds, it can be distributed patchily and fall below the detection limit within just metres owing to the lack of horizontal mixing in the water column (Goldberg et al. 2016). This results in accumulation of DNA in comparatively small and stagnant waters (Harper et al. 2019). Vertical mixing, on the other hand, can be limited by thermal stratification in lakes. This results in each layer having different effects on DNA degradation. Collecting samples during periods when thermal stratification is released and mixing occurs (e.g., during spring and fall overturns in dimictic lakes) may lead to changes in biodiversity estimates, and this should be considered when designing sampling strategies (Bista et al. 2017; Harrison et al. 2019).

Targeted approaches and community analyses using plant DNA from water

Conventional sampling techniques often require a lot of time and effort for detecting indicator, rare, or invasive species. Keeping the target organism alive or intact might also be an important consideration in such cases. Detection of species via nucleic acids collected from environmental samples (eDNA/eRNA) is a relatively new approach that emerged in the last five years (Anglès d'Auriac et al. 2019). These methods offer a non-destructive and efficient complementary approach for the detection of aquatic organisms. They rely on reference sequences and the amount of available data varies among taxonomic groups and countries (Chapter 10 DNA barcoding and Chapter 11 Amplicon metabarcoding). For example, aquatic vascular plants used in biomonitoring are well represented in public databases (BOLD, GenBank), while this is hard to achieve for diatoms due to large propor-

tions of undescribed species and the problems with cultivation of monoclonal cultures (Weigand et al. 2019). Nevertheless, eDNA studies on plants are critical for our understanding of the dynamics of plant communities in aquatic environments. An important application of eDNA-based methods in recent years has been for the detection of invasive aquatic plants (Anglès d'Auriac et al. 2019; Fujiwara et al. 2016; Gantz et al. 2018; Kuehne et al. 2020; Miyazono et al. 2020; Muha et al. 2019; Scriver et al. 2015). In most of these studies, species-specific markers were used to obtain presence/absence data, and the downstream laboratory and data analysis steps are well known and efficient. These methods can also be useful to investigate seasonal or spatial distributions of species (Muha et al. 2019). As suggested for other types of environmental samples, increasing the number and spatial coverage of samples and the time of sampling may improve detection rates for species that are rare, have low biomass, or inhabit a relatively distant site (Alsos et al. 2018).

Although DNA from plant communities have been detected from environmental samples as parts of larger surveys (e.g., within coral reefs), biodiversity studies targeting a large number of plant species are still rare, possibly owing to issues with universal amplification and discriminatory power of single or multiple gene surveys in plants (DiBattista et al. 2019; Fraser et al. 2017). Yet, there are recent efforts to design primers and assays targeting larger groups of plants as well (Coghlan et al. 2020; Shackleton et al. 2019).

An important application for DNA-based methods is the quantification of species abundance and biomass since there are several environmental applications that rely on this information. Depending on the specific aim of the study, this information can be obtained at varying degrees of efficiency and reliability. Approaches employing species-specific methods are more suitable for abundance or biomass estimations (e.g., qPCR, ddPCR). However, they require a priori knowledge of the target group and are limited to already described species. On the other hand, high-throughput approaches can identify species that are rare or have low biomass (e.g., metabarcoding, metagenomics), but they suffer from biases introduced by downstream steps such as PCR amplification, sequencing (Chapter 9 Sequencing platforms and data types), availability of reference sequences, and even the bioinformatics analyses (Chapter 18 Sequence to species) (Alsos et al. 2018; Zhou et al. 2013)

Although molecular methods for species detection have been used as a tool for biodiversity management for more than a decade, only 2% of the available studies have focused on plants (Tsuji et al. 2019). One of the main reasons is the limited information on the dynamics of DNA released from plants to aquatic environments. However, several recent and exciting experimental studies have been published on the relationship between plant biomass and DNA concentrations (Fujiwara et al. 2016; Kuehne et al. 2020; Matsuhashi et al. 2016), temporal plant DNA degradation (Fujiwara et al. 2016; Gantz et al. 2018; Matsuhashi et al. 2016), and the seasonal variation of DNA concentrations (Anglès d'Auriac et al. 2019; Kuehne et al. 2020; Matsuhashi et al. 2019). Although these types of studies are relatively new and need further optimization, two important findings are already apparent: (i) there is so far no observed consistent positive relationship between biomass and DNA concentrations, and (ii) the detectability of DNA significantly increases in autumn in temperate regions when leaf senescence and degradation start. While there is no consensus yet that this is true for all plant species, this finding does imply that the optimal sampling season for a plant can vary depending on morphology, reproductive biology, and the life cycle of the target taxon.

Experimental design

Recent studies that detect plant species in aquatic ecosystems via eDNA are mainly about methodological adjustments (Fujiwara et al. 2016; Gantz et al. 2018; Kuehne et al. 2020; Matsuhashi

et al. 2019, 2016; Schabacker et al. 2020; Strickler et al. 2015). The technique involves three main steps: 1) collecting water samples, 2) DNA isolation and sequencing, and 3) taxonomic annotation of assembled sequences. Although substantial improvements are being made for the last step due to the developments in bioinformatics (Chapter 18 Sequence to species), sample collection, DNA isolation, and even the choice of sequencing platform (Chapter 9 Sequencing platforms and data types) can still introduce biases (Singer et al. 2019; Tsuji et al. 2019). Methodological research on these two steps is usually conducted using mesocosm and aquarium experiments focusing on spatial and temporal dynamics of DNA (Kuehne et al. 2020). In the next part of this chapter, we will describe the experimental workflow and discuss the issues related to the processes affecting DNA detection from water samples, as outlined in the previous section.

Sampling strategies: water collection, filtering, and transportation

There are three main steps in a field study for the collection of aqueous eDNA: water collection, transportation, and filtering. In designing sampling strategies for species identification from water samples, there are many factors to consider. These include, but are not limited to, the field conditions, the distance between sampling point and laboratory, the amount of water that is required, and the morphology and life cycle of the target organism (Tsuji et al. 2019). There are multiple methods that can be applied for each of these steps. In this section, we will discuss and compare these methods by focusing on their advantages and limitations.

After the selection of the sampling location, the next step is to decide on the transportation strategy. Water samples can either be directly transported to the laboratory or filtered in the field. If direct transportation is the chosen method, the samples are usually collected with sterilised glass or plastic bottles or disposable plastic tubes. After that, DNA in the water samples can be captured by filtration or ethanol precipitation in the laboratory. This method both reduces the effort and time spent in the field and researchers can perform additional analyses on water samples or store subsamples for further processing (Tsuji et al. 2019; Williams et al. 2016). Storage and preservation of these samples can be challenging, however, and the amount of obtained DNA can be lower compared to filtering in the field (Minamoto et al. 2016). Usually, 15 ml to 1 l of water samples are collected when the samples are transported, while thousands of litres can be processed via field filtration method when using filters with large pore sizes (Schabacker et al. 2020; Sepulveda et al. 2019). Additionally, another advantage of filtering is that a large number of filters can be easily transported in one go due to their small sizes. However, filtering can increase the required sampling time and effort in the field. For example, if muddy waters must be processed with small pore size filters, filtration of a single litre can take hours (Hunter et al. 2019). Another important consideration for field filtering is that lots of laboratory equipment should be carried to the site (Bruce et al. 2021). Keeping the equipment sterile and preventing contamination can be challenging in field conditions. The collection of filtered distilled water can be done to detect such issues, and is thus highly recommended. The choice of the transportation method, on the other hand, depends primarily on the distance between sampling locations and the laboratory (Thomas et al. 2019). For short distances, collecting water samples can be more practical, especially when resampling is possible, as the samples can be processed in sterile laboratory conditions. On the other hand, if the sampling site is far from the laboratory, field filtering can be more efficient due to the high volumes of water samples that can be processed in a single survey, and the protection of DNA in filters during the transportation (Harper et al. 2019; Hinlo et al. 2017; Minamoto et al. 2016).

Precipitation using ethanol or isopropanol can be used for capturing DNA after water collection, but filtration is the more widely used method (Tsuji et al. 2019). The aim of this tech-

nique, whether it is applied in the laboratory or in the field, is filtering the water through a relatively small pore size membrane to hold free extracellular or/and cellular DNA. There are different options for the filtering step based on the material, pore size, and filter type (Tsuji et al. 2019). Polyethersulfone (PES), cellulose nitrate, and glass fibre are the most commonly used types of filters in DNA research. Glass fibre filters are commonly suggested due to their higher capability to absorb DNA (Muha et al. 2019; Spens et al. 2017; Tsuji et al. 2019).

Pore sizes of filters used in eDNA studies range from 0.22 µm to 60 µm (Schabacker et al. 2020; Tsuji et al. 2019). Earlier studies were usually conducted using comparatively smaller pore size filters (0.22-0.7 µm). However, after it was shown that the actual particle sizes for eDNA collected from water samples vary between 0.2 µm and 180 µm (Minamoto et al. 2016; Schabacker et al. 2020; Turner et al. 2014), larger filter sizes were more commonly used. Although this makes it possible to process larger volumes of water, it also increases the probability of introducing PCR inhibitors. In this situation, extra steps for removing PCR inhibitors can be used during DNA isolation (Hunter et al. 2019).

The type of filter is one of the most important decisions to be made when designing the sampling strategy. Filters can be classified as open or encapsulated/cartridge filters (Spens et al. 2017; Tsuji et al. 2019). Open filters are membranes that are usually fixed on an immobilised manifold system that is connected to a vacuum pump for filtering the water. Open filters require more laboratory materials and are more easily contaminated, and they are therefore not practical for use in the field. Encapsulated filters can be used with vacuum pumps or simply by mechanical force (syringes). This reduces the effort and time required for field sampling. As contamination can also be prevented immediately after water filtering, encapsulated filters offer many advantages over filtering on-site (Spens et al. 2017; Thomas et al. 2019). A key drawback of encapsulated filters is cost: encapsulated filters generally cannot be used more than once, so the total cost of filters needs to be considered, in particular for large-scale projects.

Contamination of samples and the degradation of DNA are two critical processes that should be avoided as much as possible from water collection in the field to DNA isolation in the lab (Goldberg et al. 2016; Hinlo et al. 2017; Tsuji et al. 2019). Specific eDNA sampling guidelines have been published by environmental agencies with detailed instructions on how to avoid contamination and design the optimal collection protocol (Carim et al. 2016; Laramie et al. 2015).

DNA extraction

Choosing the correct DNA extraction protocol can be crucial in ensuring that the effect of PCR inhibitors in water samples will be minimised. The chemical and physical characteristics of samples can vary considerably, and therefore the quantity and purity of isolated DNA also vary (Goldberg et al. 2016). Plant DNA isolation protocols start after the capture of DNA from water samples via either precipitation or filtration. With precipitation, samples are usually mixed with ethanol and centrifuged to collect the precipitated DNA as pellets after removal of the supernatant. A critical point here is that ethanol should be totally removed from the samples as it can affect the efficiency of further downstream steps (Kuehne et al. 2020). When isolating DNA using filtration, the filters are usually incubated in a lysis solution to ensure the DNA is free, centrifuged to remove other molecules, and isolated from the supernatant. Commercial DNA isolation kits designed for environmental tissues and plant samples (e.g., DNeasy PowerWater, PowerPlant or Blood & Tissue, Qiagen) are usually preferred by researchers with some small modifications on the recommended protocols based on the sample types (Coghlan et al. 2020; Kuehne et al. 2020; Matsuhashi et al. 2016; Miyazono et al. 2020; Schabacker et al. 2020).

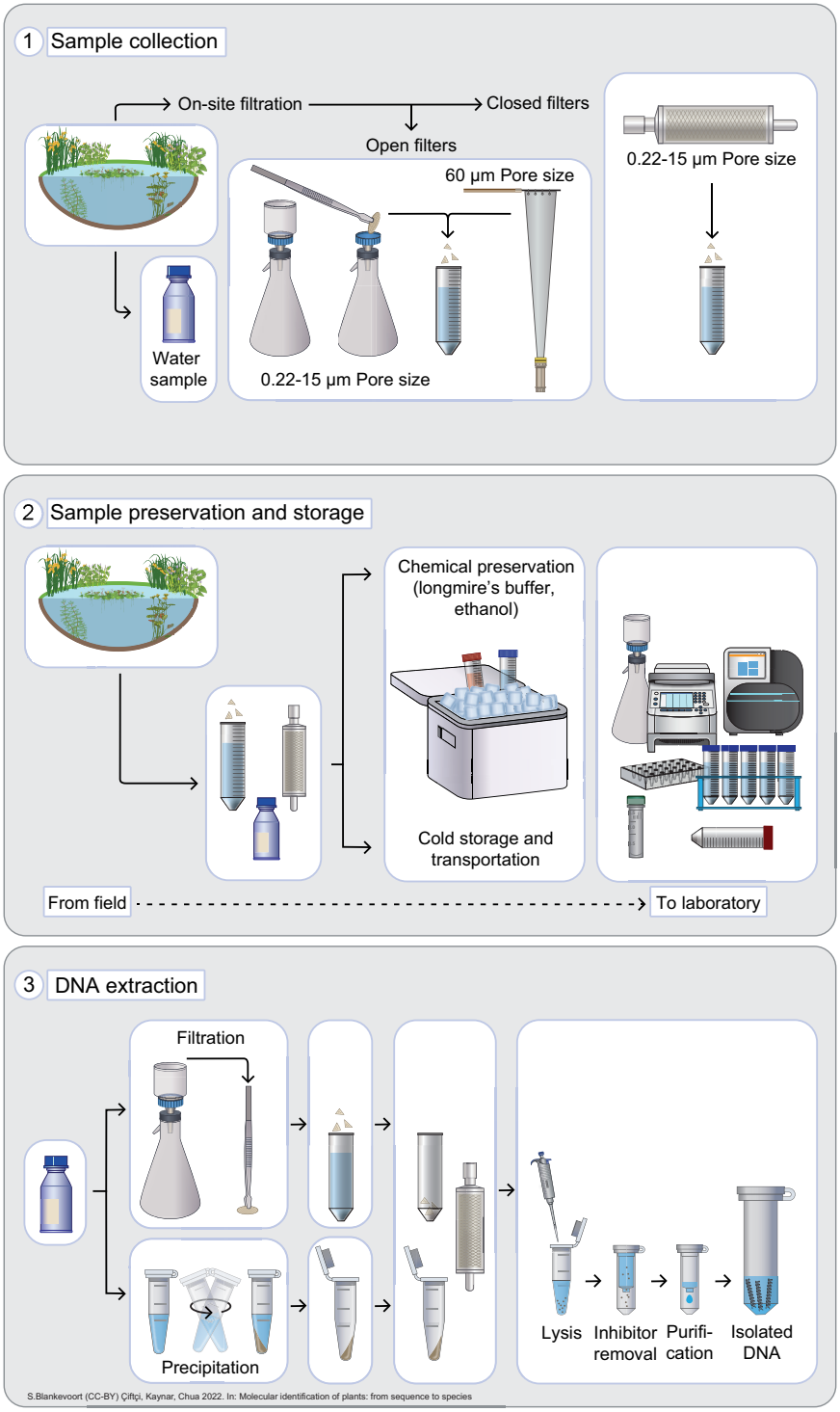

Figure 1. Chapter 3 Infographic: Summary of steps from field collection of water samples to DNA extraction in the laboratory. (1) Open or closed (encapsulated/cartridge) filters can be used for filtering water samples on-site. Large filters (e.g., plankton net with 60 µm pore size) are preferred for filtering larger volumes of water, while small pore size filters can usually process a few litres. Closed filters offer the advantage of preventing contamination, therefore they are more commonly used for on-site filtration. (2) Degradation is another important issue that should be prevented until DNA extraction. Water or filter samples can either be preserved in a chemical buffer or transported in cold and dark conditions to the laboratory for further processing. (3) Plant DNA in water samples can be captured by filtration or precipitation. When using filtration, samples are usually incubated in a lysis solution to extract DNA, while in precipitation samples are mixed with ethanol and DNA is collected in the pellet. Commercial DNA isolation kits specifically designed for environmental sample types are commonly used with some small modifications.

Conclusion and prospects

DNA isolated from water samples can be used for several downstream applications based on the specific aim of the study or survey. Currently, qPCR methods are the most commonly used method for detecting specific target taxa in water samples, while metabarcoding is used for community analyses (Chapter 11 Metabarcoding). The studies comparing the efficiency of these DNA methods with more conventional methods show varying results. For some species or taxa, DNA-based detection methods appear to outperform more conventional methods (Deiner et al. 2016; Tingley et al. 2018), though in other cases there is not necessarily an improvement in detection compared to more conventional surveys (Rose et al. 2019; Wood et al. 2019). Although DNA-based methods are constantly being improved, there are still challenges related to both false positives (when DNA is detected for a species or taxa that is known to be absent) and false negatives (when DNA is not detected for a species or taxa that is known to be present) (Beng and Corlett 2020). Therefore, at least for now, DNA-based methods for aquatic studies focusing on plants are still best coupled with conventional surveys. Based on the further development of sequencing methods and the increasing availability of reference sequence data in public databases, however, there are additional opportunities such as application of metagenomics (Chapter 12 Metagenomics), target capture (Chapter 14 Target capture), or analysis of whole plastomes (Chapter 16 Whole genome sequencing). These methods might provide the additional benefit of integrating functional information besides species detection.

Questions

1. Are water samples best collected at a single point representative of the habitat diversity as a whole? Motivate your answer.
2. Describe three biotic and three abiotic factors that can affect DNA detection rates in aquatic environments. Explain in a few sentences how these factors can result in the detection of false positives and false negatives in streaming waters.
3. List five factors that should be taken into account while designing a sampling strategy for detection of DNA from water samples.

Glossary

Apoptosis – Controlled cell death which involves cell shrinkage, nuclear fragmentation, chromatin condensation, and chromosomal DNA fragmentation.

Biofilm – A consortium of microorganisms where cells stick to each other and often also to a surface.

Dimictic lake – A body of freshwater whose difference in temperature between surface and bottom layers becomes negligible twice per year.

Extracellular nucleases – Enzymes that can work outside of the cell and are capable of cleaving the phosphodiester bonds between nucleotides of nucleic acids.

Mesocosm – Any outdoor experimental system that simulates the natural environment under controlled conditions.

Necrosis – Uncontrolled cell death due to the loss of osmotic control typically by swelling and bursting.

PCR inhibitors – Any factor which prevents the amplification of nucleic acids through the polymerase chain reaction.

Primer – A short single stranded nucleic acid sequence used by all living organisms in the initiation of DNA synthesis.

qPCR (Quantitative PCR) – An extension of the PCR technique which allows estimation of the initial quantity of nucleic acids in a biological sample.

Senescence – The gradual deterioration of functional characteristics with ageing (can be used both for organismal or cellular ageing).

Thermal stratification – The phenomenon in which lakes develop two discrete layers of water of different temperatures; warm on top (epilimnion) and cold below (hypolimnion).

Vector (i.e., cloning vectors) – A small piece of DNA that can be stably maintained in an organism that a foreign DNA fragment can be inserted into for cloning purposes.

References

Alsos IG, Lammers Y, Yoccoz NG, Jørgensen T, Sjögren P, Gielly L, Edwards ME (2018) Plant DNA metabarcoding of lake sediments: How does it represent the contemporary vegetation. PLoS ONE 13, e0195403. https://doi.org/10.1371/journal.pone.0195403

Anglès d'Auriac MB, Strand DA, Mjelde M, Demars BOL, Thaulow J (2019) Detection of an invasive aquatic plant in natural water bodies using environmental DNA. PLoS ONE 14, e0219700. https://doi.org/10.1371/journal.pone.0219700

Apothéloz-Perret-Gentil L, Bouchez A, Cordier T, Cordonier A, Guéguen J, Rimet F, Vasselon V, Pawlowski J (2021) Monitoring the ecological status of rivers with diatom eDNA metabarcoding: A comparison of taxonomic markers and analytical approaches for the inference of a molecular diatom index. Mol. Ecol. 30, 2959–2968. https://doi.org/10.1111/mec.15646

Barnes MA, Turner CR, Jerde CL, Renshaw MA, Chadderton WL, Lodge DM (2014) Environmental conditions influence eDNA persistence in aquatic systems. Environ. Sci. Technol. 48, 1819–1827. https://doi.org/10.1021/es404734p

Beng KC, Corlett RT (2020) Applications of environmental DNA (eDNA) in ecology and conservation: opportunities, challenges and prospects. Biodivers. Conserv. 29, 2089–2121. https://doi.org/10.1007/s10531-020-01980-0

Bista I, Carvalho GR, Walsh K, Seymour M, Hajibabaei M, Lallias D, Christmas M, Creer S (2017) Annual time-series analysis of aqueous eDNA reveals ecologically relevant dynamics of lake ecosystem biodiversity. Nat. Commun. 8, 14087. https://doi.org/10.1038/ncomms14087

Bruce K, Blackman R, Bourlat SJ, Hellstrom AM, Bakker J, Bista I, Bohmann K, Bouchez A, Brys R, Clark K, Elbrecht V, Fazi S, Fonseca V, Hänfling B, Leese F, Mächler E, Mahon AR, Meissner K, Panksep K, Pawlowski J, Deiner K (2021) A practical guide to DNA-based methods for biodiversity assessment. Pensoft Publishers. https://doi.org/10.3897/ab.e98875

Carim KJ, McKelvey KS, Young MK, Wilcox TM, Schwartz MK (2016) A protocol for collecting environmental DNA samples from streams. U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station, Ft. Collins, CO. https://doi.org/10.2737/RMRS-GTR-355

Carraro L, Mächler E, Wüthrich R, Altermatt F (2020) Environmental DNA allows upscaling spatial patterns of biodiversity in freshwater ecosystems. Nat. Commun. 11, 3585. https://doi.org/10.1038/s41467-020-17337-8

Carraro L, Stauffer JB, Altermatt F (2021) How to design optimal eDNA sampling strategies for biomonitoring in river networks. Environmental DNA 3, 157–172. https://doi.org/10.1002/edn3.137

Coghlan SA, Shafer ABA, Freeland JR (2020) Development of an environmental DNA metabarcoding assay for aquatic vascular plant communities. Environmental DNA. https://doi.org/10.1002/edn3.120

Collins RA, Wangensteen OS, O'Gorman EJ, Mariani S, Sims DW, Genner MJ (2018) Persistence of environmental DNA in marine systems. Commun. Biol. 1, 185. https://doi.org/10.1038/s42003-018-0192-6

D'Alessandro S, Mariani S (2021) Sifting environmental DNA metabarcoding data sets for rapid reconstruction of marine food webs. Fish Fish. https://doi.org/10.1111/faf.12553

Deiner K, Bik HM, Mächler E, Seymour M, Lacoursière-Roussel A, Altermatt F, Creer S, Bista I, Lodge DM, de Vere N, Pfrender ME, Bernatchez L (2017) Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. Mol. Ecol. 26, 5872–5895. https://doi.org/10.1111/mec.14350

Deiner K, Fronhofer EA, Mächler E, Walser J-C, Altermatt F (2016) Environmental DNA reveals that rivers are conveyer belts of biodiversity information. Nat. Commun. 7, 12544. https://doi.org/10.1038/ncomms12544

Dejean T, Valentini A, Duparc A, Pellier-Cuit S, Pompanon F, Taberlet P, Miaud C (2011) Persistence of environmental DNA in freshwater ecosystems. PLoS ONE 6, e23398. https://doi.org/10.1371/journal.pone.0023398

de Souza LS, Godwin JC, Renshaw MA, Larson E (2016) Environmental DNA (edna) detection probability is influenced by seasonal activity of organisms. PLoS ONE 11, e0165273. https://doi.org/10.1371/journal.pone.0165273

DiBattista JD, Reimer JD, Stat M, Masucci GD, Biondi P, De Brauwer M, Bunce M (2019) Digging for DNA at depth: rapid universal metabarcoding surveys (RUMS) as a tool to detect coral reef biodiversity across a depth gradient. PeerJ 7, e6379. https://doi.org/10.7717/peerj.6379

Eichmiller JJ, Best SE, Sorensen PW (2016) Effects of temperature and trophic state on degradation of environmental DNA in lake water. Environ. Sci. Technol. 50, 1859–1867. https://doi.org/10.1021/acs.est.5b05672

Ficetola GF, Miaud C, Pompanon F, Taberlet P (2008) Species detection using environmental DNA from water samples. Biol. Lett. 4, 423–425. https://doi.org/10.1098/rsbl.2008.0118

Foote AD, Thomsen PF, Sveegaard S, Wahlberg M, Kielgast J, Kyhn LA, Salling AB, Galatius A, Orlando L, Gilbert MTP (2012) Investigating the potential use of environmental DNA (eDNA) for genetic monitoring of marine mammals. PLoS ONE 7, e41781. https://doi.org/10.1371/journal.pone.0041781

Foran DR (2006) Relative degradation of nuclear and mitochondrial DNA: an experimental approach. J. Forensic Sci. 51, 766–770. https://doi.org/10.1111/j.1556-4029.2006.00176.x

Fraser CI, Connell L, Lee CK, Cary SC (2017) Evidence of plant and animal communities at exposed and subglacial (cave) geothermal sites in Antarctica. Polar Biol. 41, 1–5. https://doi.org/10.1007/s00300-017-2198-9

Fujiwara A, Matsuhashi S, Doi H, Yamamoto S, Minamoto T (2016) Use of environmental DNA to survey the distribution of an invasive submerged plant in ponds. Freshwater Science 35, 748–754. https://doi.org/10.1086/685882

Fukaya K, Murakami H, Yoon S, Minami K, Osada Y, Yamamoto S, Masuda R, Kasai A, Miyashita K, Minamoto T, Kondoh M (2020) Estimating fish population abundance by integrating quantitative data on environmental DNA and hydrodynamic modelling. Mol. Ecol. https://doi.org/10.1111/mec.15530

Gantz CA, Renshaw MA, Erickson D, Lodge DM, Egan SP (2018) Environmental DNA detection of aquatic invasive plants in lab mesocosm and natural field conditions. Biol. Invasions 20, 1–18. https://doi.org/10.1007/s10530-018-1718-z

Goldberg CS, Turner CR, Deiner K, Klymus KE, Thomsen PF, Murphy MA, Spear SF, McKee A, Oyler-McCance SJ, Cornman RS, Laramie MB, Mahon AR, Lance RF, Pilliod DS, Strickler KM, Waits LP, Fremier AK, Takahara T, Herder JE, Taberlet P (2016) Critical considerations for the application of environmental DNA methods to detect aquatic species. Methods Ecol. Evol. https://doi.org/10.1111/2041-210X.12595

Hajibabaei M, Shokralla S, Zhou X, Singer GAC, Baird DJ (2011) Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. PLoS ONE 6, e17497. https://doi.org/10.1371/journal.pone.0017497

Harper LR, Buxton AS, Rees HC, Bruce K, Brys R, Halfmaerten D, Read DS, Watson HV, Sayer CD, Jones EP, Priestley V, Mächler E, Múrria C, Garcés-Pastor S, Medupin C, Burgess K, Benson G, Boonham N, Griffiths RA, Lawson Handley L, Hänfling B (2019) Prospects and challenges of environmental DNA (eDNA) monitoring in freshwater ponds. Hydrobiologia 826, 25–41. https://doi.org/10.1007/s10750-018-3750-5

Harrison JB, Sunday JM, Rogers SM (2019) Predicting the fate of eDNA in the environment and implications for studying biodiversity. Proc. Biol. Sci. 286, 20191409. https://doi.org/10.1098/rspb.2019.1409

Hinlo R, Gleeson D, Lintermans M, Furlan E (2017) Methods to maximise recovery of environmental DNA from water samples. PLoS ONE 12, e0179251. https://doi.org/10.1371/journal.pone.0179251

Hofreiter M, Serre D, Poinar HN, Kuch M, Pääbo S (2001) Ancient DNA. Nat. Rev. Genet. 2, 353–359. https://doi.org/10.1038/35072071

Hotchkiss RS, Strasser A, McDunn JE, Swanson PE (2009) Cell death. N. Engl. J. Med. 361, 1570–1583. https://doi.org/10.1056/NEJMra0901217

Hunter ME, Ferrante JA, Meigs-Friend G, Ulmer A (2019) Improving eDNA yield and inhibitor reduction through increased water volumes and multi-filter isolation techniques. Sci. Rep. 9, 5259. https://doi.org/10.1038/s41598-019-40977-w

Kelly RP, Gallego R, Jacobs-Palmer E (2018) The effect of tides on nearshore environmental DNA. PeerJ 6, e4521. https://doi.org/10.7717/peerj.4521

Kuehne LM, Ostberg CO, Chase DM, Duda JJ, Olden JD (2020) Use of environmental DNA to detect the invasive aquatic plants *Myriophyllum spicatum* and *Egeria densa* in lakes. Freshwater Science 39, 521–533. https://doi.org/10.1086/710106

Laramie MB, Pilliod DS, Goldberg CS, Strickler KM (2015) Environmental DNA sampling protocol - filtering water to capture DNA from aquatic organisms. Techniques and Methods.

Leech DM, Snyder MT, Wetzel RG (2009) Natural organic matter and sunlight accelerate the degradation of 17ss-estradiol in water. Sci. Total Environ. 407, 2087–2092. https://doi.org/10.1016/j.scitotenv.2008.11.018

Mächler E, Salyani A, Walser J-C, Larsen A, Schaefli B, Altermatt F, Ceperley N (2021) Environmental DNA simultaneously informs hydrological and biodiversity characterization of an Alpine catchment. Hydrol. Earth Syst. Sci. 25, 735–753. https://doi.org/10.5194/hess-25-735-2021

Matsuhashi S, Doi H, Fujiwara A, Watanabe S, Minamoto T (2016) Evaluation of the environmental DNA method for estimating distribution and biomass of submerged aquatic plants. PLoS ONE 11, e0156217. https://doi.org/10.1371/journal.pone.0156217

Matsuhashi S, Minamoto T, Doi H (2019) Seasonal change in environmental DNA concentration of a submerged aquatic plant species. Freshwater Science 38, 654–660. https://doi.org/10.1086/704996

Minamoto T, Naka T, Moji K, Maruyama A (2016) Techniques for the practical collection of environmental DNA: filter selection, preservation, and extraction. Limnology 17, 23–32. https://doi.org/10.1007/s10201-015-0457-4

Miyazono S, Kodama T, Akamatsu Y, Nakao R, Saito M (2020) Application of environmental DNA methods for the detection and abundance estimation of invasive aquatic plant Egeria densa in lotic habitats. Limnology. https://doi.org/10.1007/s10201-020-00636-w

Muha TP, Skukan R, Borrell YJ, Rico JM, Garcia de Leaniz C, Garcia-Vazquez E, Consuegra S (2019) Contrasting seasonal and spatial distribution of native and invasive Codium seaweed revealed by targeting species-specific eDNA. Ecol. Evol. 9, 8567–8579. https://doi.org/10.1002/ece3.5379

Nowak P, Wiebe C, Karez R, Schubert H (2021) Applications of environmental DNA methods for charophyte biodiversity. ACA 4. https://doi.org/10.3897/aca.4.e64944

Noyer C, Abot A, Trouilh L, Leberre VA, Dreanno C (2015) Phytochip: development of a DNA-microarray for rapid and accurate identification of Pseudo-nitzschia spp and other harmful algal species. J. Microbiol. Methods 112, 55–66. https://doi.org/10.1016/j.mimet.2015.03.002

Okabe S, Shimazu Y (2007) Persistence of host-specific Bacteroides-Prevotella 16S rRNA genetic markers in environmental waters: effects of temperature and salinity. Appl. Microbiol. Biotechnol. 76, 935–944. https://doi.org/10.1007/s00253-007-1048-z

Pawlowski J, Kelly-Quinn M, Altermatt F, Apothéloz-Perret-Gentil L, Beja P, Boggero A, Borja A, Bouchez A, Cordier T, Domaizon I, Feio MJ, Filipe AF, Fornaroli R, Graf W, Herder J, van der Hoorn B, Iwan Jones J, Sagova-Mareckova M, Moritz C, Barquín J, Kahlert M (2018) The future of biotic indices in the ecogenomic era: Integrating (e)DNA metabarcoding in biological assessment of aquatic ecosystems. Sci. Total Environ. 637–638, 1295–1310. https://doi.org/10.1016/j.scitotenv.2018.05.002

Pilliod DS, Goldberg CS, Arkle RS, Waits LP (2014) Factors influencing detection of eDNA from a stream-dwelling amphibian. Mol. Ecol. Resour. 14, 109–116. https://doi.org/10.1111/1755-0998.12159

Pont D, Rocle M, Valentini A, Civade R, Jean P, Maire A, Roset N, Schabuss M, Zornig H, Dejean T (2018) Environmental DNA reveals quantitative patterns of fish biodiversity in large rivers despite its downstream transportation. Sci. Rep. 8, 10361. https://doi.org/10.1038/s41598-018-28424-8

Poté J, Ackermann R, Wildi W (2009) Plant leaf mass loss and DNA release in freshwater sediments. Ecotoxicol. Environ. Saf. 72, 1378–1383. https://doi.org/10.1016/j.ecoenv.2009.04.010

Reape TJ, Molony EM, McCabe PF (2008) Programmed cell death in plants: distinguishing between different modes. J. Exp. Bot. 59, 435–444. https://doi.org/10.1093/jxb/erm258

Rose JP, Wademan C, Weir S, Wood JS, Todd BD (2019) Traditional trapping methods outperform eDNA sampling for introduced semi-aquatic snakes. PLoS ONE 14, e0219244. https://doi.org/10.1371/journal.pone.0219244

Schabacker JC, Amish SJ, Ellis BK, Gardner B, Miller DL, Rutledge EA, Sepulveda AJ, Luikart G (2020) Increased eDNA detection sensitivity using a novel high-volume water sampling method. Environmental DNA 2, 244–251. https://doi.org/10.1002/edn3.63

Schroeder GK, Wolfenden R (2007) Rates of spontaneous disintegration of DNA and the rate enhancements produced by DNA glycosylases and deaminases. Biochemistry 46, 13638–13647. https://doi.org/10.1021/bi701480f

Scriver M, Marinich A, Wilson C, Freeland J (2015) Development of species-specific environmental DNA (eDNA) markers for invasive aquatic plants. Aquatic Botany 122, 27–31. https://doi.org/10.1016/j.aquabot.2015.01.003

Sepulveda AJ, Schabacker J, Smith S, Al-Chokhachy R, Luikart G, Amish SJ (2019) Improved detection of rare, endangered and invasive trout in using a new large-volume sampling method for eDNA capture. Environmental DNA 1, 227–237. https://doi.org/10.1002/edn3.23

Shackleton ME, Rees GN, Watson G, Campbell C, Nielsen D (2019) Environmental DNA reveals landscape mosaic of wetland plant communities. Glob. Ecol. Conserv. 19, e00689. https://doi.org/10.1016/j.gecco.2019.e00689

Shogren AJ, Tank JL, Andruszkiewicz EA, Olds B, Jerde C, Bolster D (2016) Modelling the transport of environmental DNA through a porous substrate using continuous flow-through column experiments. J. R. Soc. Interface 13. https://doi.org/10.1098/rsif.2016.0290

Shogren AJ, Tank JL, Egan SP, August O, Rosi EJ, Hanrahan BR, Renshaw MA, Gantz CA, Bolster D (2018) Water flow and biofilm cover influence environmental DNA detection in recirculating streams. Environ. Sci. Technol. 52, 8530–8537. https://doi.org/10.1021/acs.est.8b01822

Singer GAC, Fahner NA, Barnes JG, McCarthy A, Hajibabaei M (2019) Comprehensive biodiversity analysis via ultra-deep patterned flow cell technology: a case study of eDNA metabarcoding seawater. Sci. Rep. 9, 5991. https://doi.org/10.1038/s41598-019-42455-9

Spens J, Evans AR, Halfmaerten D, Knudsen SW, Sengupta ME, Mak SST, Sigsgaard EE, Hellström M (2017) Comparison of capture and storage methods for aqueous macrobial eDNA using an optimized extraction protocol: advantage of enclosed filter. Methods Ecol. Evol. 8, 635–645. https://doi.org/10.1111/2041-210X.12683

Stein JL, Marsh TL, Wu KY, Shizuya H, DeLong EF (1996) Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. J. Bacteriol. 178, 591–599. https://doi.org/10.1128/jb.178.3.591-599.1996

Strickler KM, Fremier AK, Goldberg CS (2015) Quantifying effects of UV-B, temperature, and pH on eDNA degradation in aquatic microcosms. Biological Conservation 183, 85–92. https://doi.org/10.1016/j.biocon.2014.11.038

Thomas AC, Nguyen PL, Howard J, Goldberg CS (2019) A self-preserving, partially biodegradable eDNA filter. Methods Ecol. Evol. 10, 1136–1141. https://doi.org/10.1111/2041-210X.13212

Thomsen PF, Kielgast J, Iversen LL, Wiuf C, Rasmussen M, Gilbert MTP, Orlando L, Willerslev E (2012) Monitoring endangered freshwater biodiversity using environmental DNA. Mol. Ecol. 21, 2565–2573. https://doi.org/10.1111/j.1365-294X.2011.05418.x

Tingley R, Greenlees M, Oertel S, van Rooyen AR, Weeks AR (2018) Environmental DNA sampling as a surveillance tool for cane toad Rhinella marina introductions on offshore islands. Biol. Invasions 1–6. https://doi.org/10.1007/s10530-018-1810-4

Toné S, Sugimoto K, Tanda K, Suda T, Uehira K, Kanouchi H, Samejima K, Minatogawa Y, Earnshaw WC (2007) Three distinct stages of apoptotic nuclear condensation revealed by time-lapse imaging, biochemical and electron microscopy analysis of cell-free apoptosis. Exp. Cell Res. 313, 3635–3644. https://doi.org/10.1016/j.yexcr.2007.06.018

Tsuji S, Takahara T, Doi H, Shibata N, Yamanaka H (2019) The detection of aquatic macroorganisms using environmental DNA analysis—A review of methods for collection, extraction, and detection. Environmental DNA 1, 99–108. https://doi.org/10.1002/edn3.21

Tsuji S, Ushio M, Sakurai S, Minamoto T, Yamanaka H (2017) Water temperature-dependent degradation of environmental DNA and its relation to bacterial abundance. PLoS ONE 12, e0176608. https://doi.org/10.1371/journal.pone.0176608

Turner CR, Barnes MA, Xu CCY, Jones SE, Jerde CL, Lodge DM (2014) Particle size distribution and optimal capture of aqueous macrobial eDNA. Methods Ecol. Evol. 5, 676–684. https://doi.org/10.1111/2041-210X.12206

Turner CR, Uy KL, Everhart RC (2015) Fish environmental DNA is more concentrated in aquatic sediments than surface water. Biological Conservation 183, 93–102. https://doi.org/10.1016/j.biocon.2014.11.017

Vanyushin BF, Bakeeva LE, Zamyatnina VA, Aleksandrushkina NI (2004) Apoptosis in plants: specific features of plant apoptotic cells and effect of various factors and agents. Int. Rev. Cytol. 233, 135–179. https://doi.org/10.1016/S0074-7696(04)33004-4

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Smith HO (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304, 66–74. https://doi.org/10.1126/science.1093857

Weigand H, Beermann AJ, Čiampor F, Costa FO, Csabai Z, Duarte S, Geiger MF, Grabowski M, Rimet F, Rulik B, Strand M, Szucsich N, Weigand AM, Willassen E, Wyler SA, Bouchez A, Borja A, Čiamporová-Zaťovičová Z, Ferreira S, Dijkstra K-DB, Ekrem T (2019) DNA barcode reference libraries for the monitoring of aquatic biota in Europe: Gap-analysis and recommendations for future work. Sci. Total Environ. 678, 499–524. https://doi.org/10.1016/j.scitotenv.2019.04.247

Wilcox TM, McKelvey KS, Young MK, Sepulveda AJ, Shepard BB, Jane SF, Whiteley AR, Lowe WH, Schwartz MK (2016) Understanding environmental DNA detection probabilities: A case study using a stream-dwelling char Salvelinus fontinalis. Biological Conservation 194, 209–216. https://doi.org/10.1016/j.biocon.2015.12.023

Willerslev E, Cappellini E, Boomsma W, Nielsen R, Hebsgaard MB, Brand TB, Hofreiter M, Bunce M, Poinar HN, Dahl-Jensen D, Johnsen S, Steffensen JP, Bennike O, Schwenninger J-L, Nathan R, Armitage S, de Hoog C-J, Alfimov V, Christl M, Beer J, Collins MJ (2007) Ancient biomolecules from deep ice cores reveal a forested southern Greenland. Science 317, 111–114. https://doi.org/10.1126/science.1141758

Williams KE, Huyvaert KP, Piaggio AJ (2016) No filters, no fridges: a method for preservation of water samples for eDNA analysis. BMC Res. Notes 9, 298. https://doi.org/10.1186/s13104-016-2104-5

Wood SA, Pochon X, Ming W, von Ammon U, Woods C, Carter M, Smith M, Inglis G, Zaiko A (2019) Considerations for incorporating real-time PCR assays into routine marine biosecurity surveillance programmes: a case study targeting the Mediterranean fanworm (Sabella spallanzanii) and club tunicate (Styela clava) 1. Genome 62, 137–146. https://doi.org/10.1139/gen-2018-0021

Zhou X, Li Y, Liu S, Yang Q, Su X, Zhou L, Tang M, Fu R, Li J, Huang Q (2013) Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. Gigascience 2, 4. https://doi.org/10.1186/2047-217X-2-4

ANSWERS

1. No. The probability of species detection depends on the presence and concentration of DNA collected in a water sample. Therefore, multiple sampling sites with replicates is highly encouraged to obtain a broader overview of the local species diversity.
2. The life history, age, and size of an organism are some of the biotic factors which can affect DNA detection rates. After release from the cell, abiotic factors such as UV, temperature, and pH can influence these rates. In streaming waters, including currents, false positives might be detected in downstream regions due to the transport of DNA. Similarly, if the DNA of the target organism degrades too quickly, it cannot be detected resulting in false negatives.
3. The scientific question, environmental conditions (physical and chemical), distance between sampling point and laboratory, and morphology and life cycle of the target organism should be considered when designing sampling strategies.

Chapter 4
DNA from soil

Maria Ariza Salazar[1], Sandra Garcés-Pastor[2], Hugo de Boer[1]

1 Natural History Museum, University of Oslo, Norway
2 The Arctic University Museum of Norway, UiT - The Arctic University of Norway, Tromsø, Norway

Maria Ariza Salazar **mariadelosangelesariza@gmail.com**
Sandra Garcés-Pastor **sandra.garces-pastor@uit.no**
Hugo de Boer **h.de.boer@nhm.uio.no**

Citation: Salazar MA, Garcés-Pastor S, de Boer H (2022) Chapter 4. DNA from soil. In: de Boer H, Rydmark MO, Verstraete B, Gravendeel B (Eds) Molecular identification of plants: from sequence to species. Advanced Books. https://doi.org/10.3897/ab.e98875

Introduction

The natural presence of any plant entails the existence of a substrate where it can anchor itself and absorb nutrients for its development and survival (Wardle et al. 2004). This is most commonly the ground and specifically, soil. Nevertheless, the link between soil and plants goes beyond soil supporting plants as plants are one of the main soil-forming forces of pedogenesis through the accumulation of organic matter as well as modification of the soil biochemistry surrounding the roots (Corti et al. 2005). This process over time leads to the formation of soil layers, termed horizons, that can commonly be visibly identified (Schulz et al. 2013; Shlemon 1985; Vogt et al. 1995). Near the ground surface, the first soil horizon is an organic layer composed of growing roots and decomposing vegetative and reproductive plant material from local or regional origins, i.e., fallen debris, pollen particles, seeds (Vogt et al. 1995). Hence, this soil horizon is particularly rich in plant DNA from the environment (soil eDNA in short; Taberlet et al. 2018) and can be used as a proxy for plant identification and other biodiversity assessments (Fahner et al. 2016; Taberlet et al. 2018; Yoccoz et al. 2012).

Since the first isolation of DNA from soil bacteria, soil eDNA has gained attention for the assessment of terrestrial environments for several reasons: soil is virtually everywhere, it is easy to collect and transport, harbors signals from above and below biota including both active and dormant cells, and is a non-invasive sample collection technique (Torsvik et al. 1990; Yoccoz 2012); for more on soil eDNA applications see Chapter 24 Environment and biodiversity assessments). Soil eDNA assessments targeting modern plant diversity commonly employ samples that are collected near the surface (organic horizont). However, some studies may refer to sediments which can lead to confusing eDNA samples coexisting in underground environments (Kristensen and Rabenhorst 2015). Although both soil and sediments are products of mineral weathering (Wood 1987), in soils the deposition of these products happens in situ and remains on the surface, while in sediments these products are transported and redeposited elsewhere in layers over time, e.g., the ground or the bottom of a lake or stream (Burdige 2020). Moreover, sediments in general have very different organic content, particle size and mineralogy, and lesser organismal activity than soil, although the transition from soil to sediment can be gradual and depends on the eco-physiological characteristics of the regional environment (e.g., tropical vs. boreal forest; Shackley 1975; Smol et al. 2001). Yet, during flooding events sediments can be transported very rapidly from one place to another while sedimenting in new layers mixed with soil (Baldwin and Mitchell 2000). In these contexts, soil and sedimentary eDNA samples may have a mix of different spatio-temporal signals when it comes to the reconstruction of terrestrial or aquatic environments (Deiner et al. 2017; Thomsen and Willerslev 2015). Ancient sedimentary DNA (sedaDNA) is commonly sampled from bottom sediment layers in either aquatic or terrestrial environments (Parducci et al. 2018), and its temporal signal is usually correlated with sampling depth (Willerslev and Cooper 2005). For more on sedaDNA and its applications see Chapter 8 aDNA from sediments. Sedimentary DNA (sedDNA) usually refers to modern sediments that were either recently deposited or signal contemporary environments. Plant biodiversity assessments of modern environments often employ surface lake sediments (Andersen et al. 2012; Pedersen et al. 2015; Willerslev et al. 2014) as it captures current biodiversity from the entire watershed catchment area (Alsos et al. 2018). This chapter focuses on modern DNA isolated from soil eDNA.

Further, studies may also refer to bulk soil DNA when using soil samples to identify unknown communities, especially in forensic contexts (Boggs et al. 2019; Gothwal et al. 2007; Meiklejohn et al. 2018). Bulk DNA is commonly used in contexts where known taxa are mixed, molecularly identified (usually by metabarcoding), and then studied. There is no consensus

on the precise use of these different terms, and the terminology often reflects disciplinary backgrounds and study approaches (Kristensen and Rabenhorst 2015). Yet, it is worth noting that all terms mentioned so far are not mutually exclusive nor encapsulate a particular environment. For example, soil may also be used in aquatic contexts when pedogenic processes lead to horizon differentiation, e.g., estuarine substrata (Wardle et al. 2004). Thus, careful interpretation of the context in which the term is employed is recommended to ensure correct interpretation of data and studies.

Soil DNA: degradation, persistence, and decay

Molecular (plant) identification using soil or sediment eDNA relies on isolating DNA traces from roots, debris, seeds, and pollen (Levy-Booth et al. 2007), which signal diverse spatial and temporal origins, i.e., local or regional, ancient or contemporary. When these plant parts settle into the ground, DNA can be present either in intact cells (intracellular DNA or iDNA) or free in the environment following cell lysis or rupture (extracellular DNA or exDNA; Nagler et al. 2018). The largest fraction of eDNA in underground environments is exDNA that originates from bacteria and fungal soil communities (Levy-Booth et al. 2007; Nagler et al. 2018; Pietramellara et al. 2009; Poté et al. 2009).

The state of DNA in the soil is subject to intrinsic and extrinsic DNA properties related to the origins of the DNA as well as factors influencing its decay (Barnes et al. 2014; Lacoursière-Roussel and Deiner 2021; Sirois and Buckley 2019). For more on leaf DNA decay together with organic horizon formation, see the infographic. Soil eDNA is therefore a combination of iDNA and exDNA, that can degrade rapidly or persist over time. Intrinsic DNA properties that can affect its persistence in the ground include characteristics such as DNA GC content, purity, and weight (Nielsen et al. 2000; Pietramellara et al. 2009; Sirois and Buckley 2019; Taberlet et al. 2018; Vuillemin et al. 2017). Intrinsic DNA properties are those of the organism that affect the magnitude of DNA deposition such as life history traits like biomass, feeding, social, nesting, burrowing, hibernation, etc. Extrinsic DNA properties are more related to abiotic and biotic processes operating in the ground, e.g., soil mineralogy, organic components, pH, electrostatic properties, moisture, the presence/absence of UV radiation, bioturbation**,** enzymatic activity by microbial communities, and decomposition (Cozzolino et al. 2007; Gardner and Gunsch 2017; Gulden et al. 2005; Levy-Booth et al. 2007; Prosser and Hedgpeth 2018; Saeki et al. 2011). Examples of biotic processes operating in natural environments can be found in the infographic.

iDNA persists due to protection from the cell wall and membranes against abiotic processes. Cells are more likely to remain intact in the ground if there is decreased enzymatic activity as a result of rapid soil desiccation, low temperatures, or extreme pH values (Pietramellara et al. 2009; Taberlet et al. 2018). exDNA is more likely to persist when it binds to surface-reactive particles and hydrophobic soil components such as clay, sand, silt, and humic acids (Levy-Booth et al. 2007; Pietramellara et al. 2009). DNA may also indirectly persist via bacterial integration of DNA fragments (Levy-Booth et al. 2007). Bacterial enzymatic activity plays a central role in DNA degradation in soil (Blum et al. 1997). DNase is secreted copiously to access the phosphorus and nitrogen from the DNA and acts more rapidly on DNA at higher temperatures (Levy-Booth et al. 2007). Since both the temperature and underground biota activity levels are higher in tropical climates, there are generally increased degradation rates in tropical vs. boreal soils. Soil types may also affect degradation rates (Sirois and Buckley 2019) using a controlled microcosm reported that synthetic DNA degraded slower in forest than in agricultural soils where tillage and other disruptive processes can affect persistence. Predicting the origins and persistence of

eDNA remains a thorny issue, mainly because of the complex nature of the properties involved (Barnes and Turner 2015; Deiner et al. 2017).

Soil memory

Plant eDNA bound to soil particles can originate from multiple taxa and multiple vegetative parts, each one with particular mechanisms to bind, persist and degrade in soil substrates. Plant DNA persistence within soil allows us to harvest its botanical memory for identifying vegetation through time. Indeed, comparisons of plant identifications through both visual vegetation surveys and soil eDNA assessments have shed light on the temporal signals stored in top soils. In boreal areas, plant identification through soil eDNA signal mostly registered contemporary vegetation (Ariza et al. 2022; Edwards et al. 2018; Yoccoz et al. 2012), however, taxa surveyed up to 30 years ago was also reported, suggesting that soil eDNA harbors more of a contemporary memory (Ariza et al. 2022). The extent of this memory effect across soil types and environments is poorly understood while its implications are relevant for society (e.g., biodiversity assessments and monitoring, forensics, biosafety). For more on applications of soil eDNA see Chapter 24 Environment and biodiversity assessments.

Designing a soil eDNA study

The flora and study area are key in any study to ensure sound conclusions. Below you will find considerations that can help you to answer common questions when designing field and wet lab experiments.

How to sample and how much?

Soil sampling can be done either by scooping out the soil, drilling down a tube, i.e., a 50 ml falcon tube, or with a soil core sampler. We recommend to use sampling protocols specifically validated in an environment similar to your study site, e.g., woodlands, grasslands, meadows, boreal temperate, and tropical forest (Bienert et al. 2012; Dopheide et al. 2019; Fahner et al. 2016; Taberlet et al. 2012; Yoccoz et al. 2012). It is also recommended to sample in flat areas as slopes can cause erosion and colluvium that can interfere with soil stratification. Soil and sedimentary particles are deposited in sequence, thus we can expect the bottom soil horizons to harbor older eDNA signals than those at the top. However, mixing across vertical layers can be expected as a result of bioturbation, and it is thus very important to assess the stratigraphy of the soil/sediment that is being investigated. If bioturbation is absent, sampling specific soil horizons can thus be used to capture vegetation with particular time signals (Dickie et al. 2018). Similarly, the amount of soil collected, as well as the number of samples and replicates, can affect the spatial and time signal captured (Calderón-Sanou et al. 2020; Dopheide et al. 2019; Taberlet et al. 2012; Zinger et al. 2019a). We recommend sampling at least 10 g of soil, but power analysis and rarefaction curves can aid to determine and optimize this parameter (Dickie et al. 2018; Dopheide et al. 2019). If one prefers to reduce the effect of local heterogeneity in the sampling strategy, several dozens of subsamples (between 20 and 50 g) can be mixed (Dickie et al. 2018; Taberlet et al. 2012). This strategy is however not suitable for studies dealing with patterns at small spatial scales (< 1 m^2; Edwards et al. 2018).

How to process the soil samples?

Obtaining clean DNA samples as well as avoiding cross contamination is challenging when sampling soil eDNA. Collection instruments should therefore be decontaminated between each sample (e.g., flaming, chlorine cleaning), gloves and masks should be worn and changed regularly to avoid introduction of DNA, and samples should be stored in separate plastic bags. In order to stop (or greatly reduce) enzymatic activity, samples should be stored cold or frozen, preferably at -20 °C, if immediate sample processing is not possible (Taberlet et al. 2012). Post-collection treatment of soil samples can also include air drying or freeze-drying to stop enzymatic activity and preserve DNA integrity in the sample (Nocker et al. 2012; Ritter et al. 2018). Soil samples are usually a mix of both above and below ground fragments of fauna and flora, i.e., debris, manure, roots, seeds, pollen, insects, and worms. DNA from organisms that are present in large total biomass may complicate detection of DNA signals from rare organisms. Thus, particularly for plant identification studies, it is worth considering whether root and leaf fragments should be sieved out from the soil samples. This will also contribute towards amplifying the signal from those low abundant taxa and normalize amplifications for all organisms present in a sample.

Extraction of iDNA or exDNA?

DNA extraction is a key bottleneck when capturing molecular data, and protocols need to be tailored to both the study area and the question(s). At a minimum, you need to decide which fraction of the total soil eDNA (iDNA or exDNA) you want to isolate to answer your research question. In general, isolating exDNA is preferred when targeting non-microorganisms and avoiding diversity patterns across short temporal scales (Taberlet et al. 2012; Zinger et al. 2009). While both extraction protocols are generally similar, iDNA extraction requires a cell lysis step. Breaking the cell wall or pollen exine can be achieved with soil grinding, sonication, thermal shocks, or chemical treatments (Frostegård et al. 1999; Zhou et al. 2007). For DNA extraction protocols specifically for pollen DNA, see Chapter 5 DNA from pollen. Commercial kits for DNA extraction are readily available for joint or separate extraction of iDNA and exDNA from soil, and these are commonly used in soil eDNA studies (Alsos et al. 2018; Edwards et al. 2018; Fahner et al. 2016; Foucher et al. 2020; Yoccoz et al. 2012; Zinger et al. 2019b). Taberlet et al. (2012) proposed an extraction protocol targeting exDNA that is suitable for tropical and nontropical areas, and can be performed with material that is commonly found in molecular laboratories. Depending on the soil properties in your study area, you can adapt commercial kits to increase the quality and quantity of DNA. For example, adding chloroform can increase the separation of the organic phase and aqueous phase, which in turn optimizes DNA quality (Fatima et al. 2014). However, chloroform is highly abrasive and can induce cell lysis. Alternatively, slightly alkaline solutions of phosphate buffers can remove soil particles to which exDNA might be bound while simultaneously preventing lysis of the cells (Nagler et al. 2018).

Which DNA marker(s) to use?

If (meta)barcoding is used for identification, there are three desired features for a barcode in any study: sufficient polymorphism for identification at the desired taxonomic resolution, conserved primer binding sites for universal amplification, and available reference sequences for

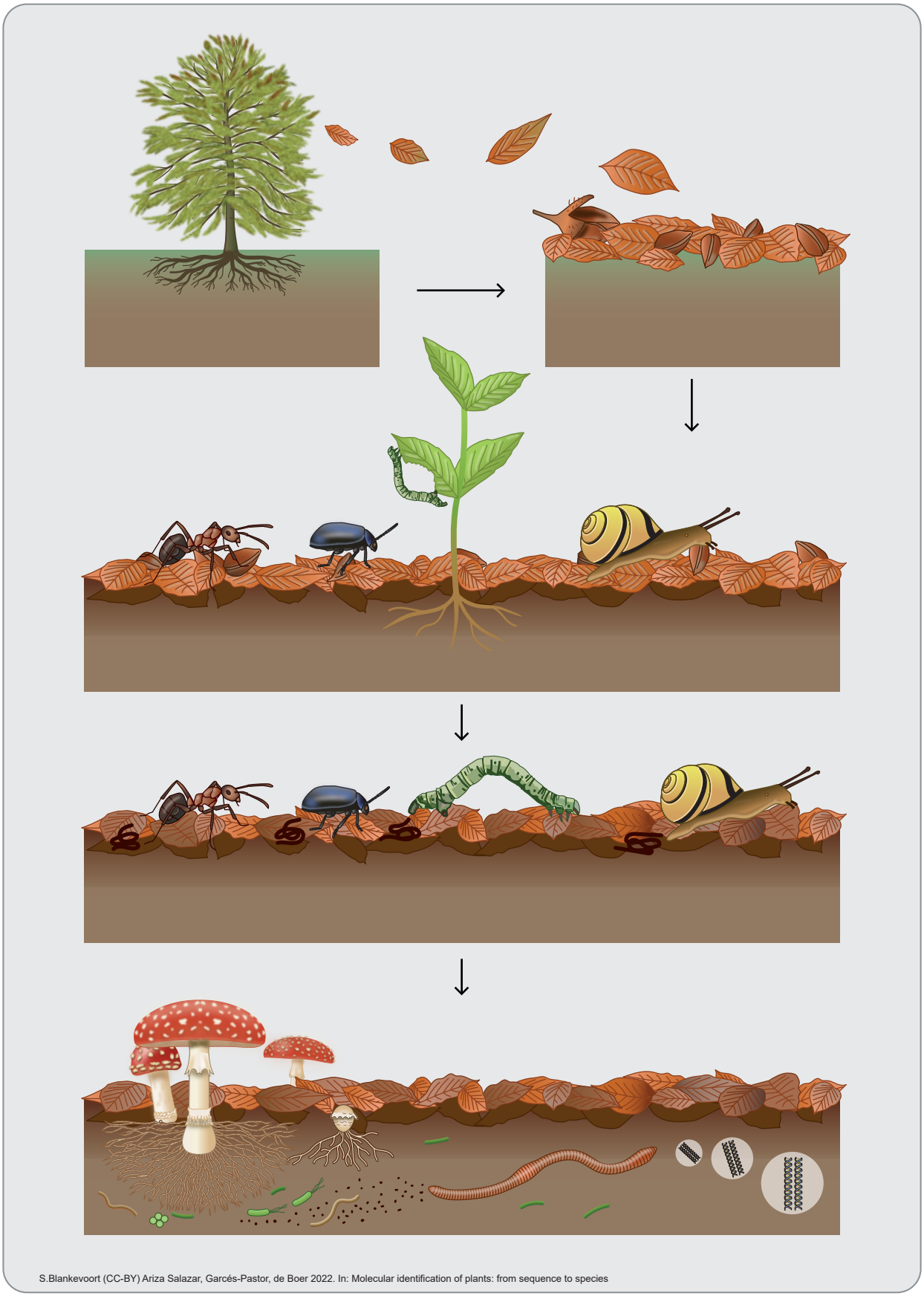

Figure 1. Chapter 4 Infographic: From leaf DNA to soil environmental DNA. One of the ways in which plant DNA is deposited in soil surfaces is through the accumulation of fallen leaves from trees.

the target organism. In many cases, not all features can be met. You may therefore need to decide on which features are most important for your research question. For more general information about choosing suitable markers and available reference databases, see Chapter 10 DNA barcoding and Chapter 11 Amplicon metabarcoding. Soil eDNA studies targeting plants have used markers found in chloroplast DNA (*trnL* P6 loop, *matK*, *rbcL*) and in ribosomal DNA (ITS2; Epp et al. 2018; Fahner et al. 2016; Yoccoz et al. 2012). However, metagenomic and target enrichment approaches are also starting to gain popularity as these avoid bias by PCR amplification and reduce the noise from non-target organisms (Johnson et al. 2019; Murchie et al. 2021). Fahner et al. (2016) compared the performance of plant barcodes (long vs. short barcodes) and recommended ITS2 and *rbcL* when identifying plants through soil eDNA metabarcoding, because these outperformed other markers in terms of recovery, reference completeness and identification resolution. Since the nuclear region, ITS2, is shared across plants and fungi, and the latter are abundantly present in soil, increased amplification of fungi can be expected. To avoid this, plant-specific primers targeting these regions can be used (Cheng et al. 2016). Furthermore, to avoid biased assessments towards particular plant groups when using ITS2, i.e., flowering plants or mosses, a combination of both TS2F/ITSp4 and ITSp3/ITSu4 primers pairs, is recommended to yield most of the land plant communities (Cheng et al. 2016; Timpano et al. 2020). In addition, the *trnL* P6 loop is the most commonly used marker in plant eDNA studies for a number of reasons: it has sufficient variability across both angiosperms and gymnosperms, there are a number of available reference databases as well as taxa-specific primers, and its small size works well for degraded eDNA (Alsos et al. 2020; Epp et al. 2018; Foucher et al. 2020).

Questions

1. The laboratory technician hands you an extraction protocol that has been used previously to extract DNA from soil and sediments. How do you know if this protocol will extract both iDNA and exDNA? Motivate your answer.
2. You are designing your soil eDNA study for a plant taxon that is distributed heterogeneously across plots. Describe the soil sampling strategy that will take into account the target taxon distribution.
3. You want to reconstruct vegetation types based on soil eDNA targeting the *trnL* P6 loop. This marker will not allow you to identify all taxa to species level. Will this affect your ability to determine the vegetation types? Motivate why or why not?

Glossary

Bioturbation – Biological processes involved in the dissemination of genetic media through terrestrial media.
DNA degradation – Refers to the physical changes of the DNA molecule.
DNA decay – Refers to the reduction in detectable quantity of eDNA.
DNA persistence – Refers to the amount of DNA that remains detectable across time.
DNA polymorphism – Presence of two or more variants of a particular DNA sequence.
Horizon – A layer parallel to the soil surface whose physical, chemical and biological characteristics differ from the layers above and beneath.

Power analysis – Probability of detecting an effect, given that the effect is really there. Can also be seen as rejecting the null hypothesis when it is in fact false.

Pedogenesis – The process of soil formation as regulated by the effects of place, environment, and history.

Rarefaction curves (in ecology) – A technique to assess species richness given the number of samples collected.

References

Alsos IG, Lammers Y, Yoccoz NG, Jørgensen T, Sjögren P, Gielly L, Edwards ME (2018) Plant DNA metabarcoding of lake sediments: how does it represent the contemporary vegetation. PLoS ONE 13, e0195403. https://doi.org/10.1371/journal.pone.0195403

Alsos IG, Lavergne S, Merkel MKF, Boleda M, Lammers Y, Alberti A, Pouchon C, Denoeud F, Pitelkova I, Pușcaș M, Roquet C, Hurdu B-I, Thuiller W, Zimmermann NE, Hollingsworth PM, Coissac E (2020) The treasure vault can be opened: large-scale genome skimming works well using herbarium and silica gel dried material. Plants 9, 432. https://doi.org/10.3390/plants9040432

Andersen K, Bird KL, Rasmussen M, Haile J, Breuning-Madsen H, Kjaer KH, Orlando L, Gilbert MTP, Willerslev E (2012) Meta-barcoding of "dirt" DNA from soil reflects vertebrate biodiversity. Mol. Ecol. 21, 1966–1979. https://doi.org/10.1111/j.1365-294X.2011.05261.x

Ariza M, Fouks B, Mauvisseau Q, Halvorsen R, Alsos IG, de Boer H (2022) Plant biodiversity assessment through soil eDNA reflects temporal and local diversity. Methods Ecol. Evol. https://doi.org/10.1111/2041-210X.13865

Baldwin DS, Mitchell AM (2000) The effects of drying and re-flooding on the sediment and soil nutrient dynamics of lowland river-floodplain systems: a synthesis. Regul. Rivers: Res. Mgmt. 16, 457–467. https://doi.org/10.1002/1099-1646(200009/10)16:5<457::AID-RRR597>3.0.CO;2-B

Barnes MA, Turner CR, Jerde CL, Renshaw MA, Chadderton WL, Lodge DM (2014) Environmental conditions influence eDNA persistence in aquatic systems. Environ. Sci. Technol. 48, 1819–1827. https://doi.org/10.1021/es404734p

Barnes MA, Turner CR (2015) The ecology of environmental DNA and implications for conservation genetics. Conserv. Genet. 17, 1–17. https://doi.org/10.1007/s10592-015-0775-4

Bienert F, De Danieli S, Miquel C, Coissac E, Poillot C, Brun J-J, Taberlet P (2012) Tracking earthworm communities from soil DNA. Mol. Ecol. 21, 2017–2030. https://doi.org/10.1111/j.1365-294X.2011.05407.x

Blum SAE, Lorenz MG, Wackernagel W (1997) Mechanism of retarded DNA degradation and prokaryotic origin of dnases in nonsterile soils. Syst. Appl. Microbiol. 20, 513–521. https://doi.org/10.1016/S0723-2020(97)80021-5

Boggs LM, Scheible MKR, Machado G, Meiklejohn KA (2019) Single fragment or bulk soil DNA metabarcoding: which is better for characterizing biological taxa found in surface soils for sample separation? Genes (Basel) 10, 431. https://doi.org/10.3390/genes10060431

Burdige DJ (2007) Geochemistry of marine sediments. Princeton University Press, Princeton. 624 pp.

Calderón-Sanou I, Münkemüller T, Boyer F, Zinger L, Thuiller W (2020) From environmental DNA sequences to ecological conclusions: how strong is the influence of methodological choices? J. Biogeogr. 47, 193–206. https://doi.org/10.1111/jbi.13681

Cheng T, Xu C, Lei L, Li C, Zhang Y, Zhou S (2016) Barcoding the kingdom Plantae: new PCR primers for ITS regions of plants with improved universality and specificity. Mol. Ecol. Resour. 16, 138–149. https://doi.org/10.1111/1755-0998.12438

Corti G, Agnelli A, Cuniglio R, Sanjurjo MF, Cocco S (2005) Characteristics of rhizosphere soil from natural and agricultural environments, in: Huang, P.M., Gobran, G.R. (Eds.), Biogeochemistry of Trace Elements in the Rhizosphere. Elsevier, pp. 57–128. https://doi.org/10.1016/B978-044451997-9/50005-2

Cozzolino S, Cafasso D, Pellegrino G, Musacchio A, Widmer A (2007) Genetic variation in time and space: the use of herbarium specimens to reconstruct patterns of genetic variation in the endangered orchid *Anacamptis palustris*. Conserv. Genet. 8, 629–639. https://doi.org/10.1007/s10592-006-9209-7

Deiner K, Bik HM, Mächler E, Seymour M, Lacoursière-Roussel A, Altermatt F, Creer S, Bista I, Lodge DM, de Vere N, Pfrender ME, Bernatchez L (2017) Environmental DNA metabarcoding: transforming how we survey animal and plant communities. Mol. Ecol. 26, 5872–5895. https://doi.org/10.1111/mec.14350

Dickie IA, Boyer S, Buckley HL, Duncan RP, Gardner PP, Hogg ID, Holdaway RJ, Lear G, Makiola A, Morales SE, Powell JR, Weaver L (2018) Towards robust and repeatable sampling methods in eDNA-based studies. Mol. Ecol. Resour. 18, 940–952. https://doi.org/10.1111/1755-0998.12907

Dopheide A, Xie D, Buckley TR, Drummond AJ, Newcomb RD (2019) Impacts of DNA extraction and PCR on DNA metabarcoding estimates of soil biodiversity. Methods Ecol. Evol. 10, 120–133. https://doi.org/10.1111/2041-210X.13086

Edwards ME, Alsos IG, Yoccoz N, Coissac E, Goslar T, Gielly L, Haile J, Langdon CT, Tribsch A, Binney HA, von Stedingk H, Taberlet P (2018) Metabarcoding of modern soil DNA gives a highly local vegetation signal in Svalbard tundra. The Holocene 28, 2006–2016. https://doi.org/10.1177/0959683618798095

Epp LS, Kruse S, Kath NJ, Stoof-Leichsenring KR, Tiedemann R, Pestryakova LA, Herzschuh U (2018) Temporal and spatial patterns of mitochondrial haplotype and species distributions in Siberian larches inferred from ancient environmental DNA and modeling. Sci. Rep. 8, 17436. https://doi.org/10.1038/s41598-018-35550-w

Fahner NA, Shokralla S, Baird DJ, Hajibabaei M (2016) Large-scale monitoring of plants through environmental DNA metabarcoding of soil: recovery, resolution, and annotation of four DNA markers. PLoS ONE 11, e0157505. https://doi.org/10.1371/journal.pone.0157505

Fatima F, Pathak N, Rastogi Verma S (2014) An improved method for soil DNA extraction to study the microbial assortment within rhizospheric region. Mol. Biol. Int. 2014, 518960. https://doi.org/10.1155/2014/518960

Foucher A, Evrard O, Ficetola GF, Gielly L, Poulain J, Giguet-Covex C, Laceby JP, Salvador-Blanes S, Cerdan O, Poulenard J (2020) Persistence of environmental DNA in cultivated soils: implication of this memory effect for reconstructing the dynamics of land use and cover changes. Sci. Rep. 10, 10502. https://doi.org/10.1038/s41598-020-67452-1

Frostegård A, Courtois S, Ramisse V, Clerc S, Bernillon D, Le Gall F, Jeannin P, Nesme X, Simonet P (1999) Quantification of bias related to the extraction of DNA directly from soils. Appl. Environ. Microbiol. 65, 5409–5420. https://doi.org/10.1128/AEM.65.12.5409-5420.1999

Gardner CM, Gunsch CK (2017) Adsorption capacity of multiple DNA sources to clay minerals and environmental soil matrices less than previously estimated. Chemosphere 175, 45–51. https://doi.org/10.1016/j.chemosphere.2017.02.030

Gothwal RK, Nigam VK, Mohan MK, Sasmal D, Ghosh P (2007) Extraction of bulk DNA from Thar Desert soils for optimization of PCR-DGGE based microbial community analysis. Electron. J. Biotechnol. 10, 400–408. https://doi.org/10.2225/vol10-issue3-fulltext-6

Gulden RH, Lerat S, Hart MM, Powell JR, Trevors JT, Pauls KP, Klironomos JN, Swanton CJ (2005) Quantitation of transgenic plant DNA in leachate water: real-time polymerase chain reaction analysis. J. Agric. Food Chem. 53, 5858–5865. https://doi.org/10.1021/jf0504667

Johnson MG, Pokorny L, Dodsworth S, Botigué LR, Cowan RS, Devault A, Eiserhardt WL, Epitawalage N, Forest F, Kim JT, Leebens-Mack JH, Leitch IJ, Maurin O, Soltis DE, Soltis PS, Wong GK-S, Baker WJ, Wickett NJ (2019) A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. Syst. Biol. 68, 594–606. https://doi.org/10.1093/sysbio/syy086

Kristensen E, Rabenhorst MC (2015) Do marine rooted plants grow in sediment or soil? A critical appraisal on definitions, methodology and communication. Earth-Science Reviews 145, 1–8. https://doi.org/10.1016/j.earscirev.2015.02.005

Lacoursière-Roussel A, Deiner K (2021) Environmental DNA is not the tool by itself. J. Fish Biol. 98, 383–386. https://doi.org/10.1111/jfb.14177

Levy-Booth DJ, Campbell RG, Gulden RH, Hart MM, Powell JR, Klironomos JN, Pauls KP, Swanton CJ, Trevors JT, Dunfield KE (2007) Cycling of extracellular DNA in the soil environment. Soil Biology and Biochemistry 39, 2977–2991. https://doi.org/10.1016/j.soilbio.2007.06.020

Meiklejohn KA, Jackson ML, Stern LA, Robertson JM (2018) A protocol for obtaining DNA barcodes from plant and insect fragments isolated from forensic-type soils. Int. J. Legal Med. 132, 1515–1526. https://doi.org/10.1007/s00414-018-1772-1

Murchie TJ, Kuch M, Duggan AT, Ledger ML, Roche K, Klunk J, Karpinski E, Hackenberger D, Sadoway T, MacPhee R, Froese D, Poinar H (2021) Optimizing extraction and targeted capture of ancient environmental DNA for reconstructing past environments using the PalaeoChip Arctic-1.0 bait-set. Quaternary Research 99, 305–328. https://doi.org/10.1017/qua.2020.59

Nagler M, Insam H, Pietramellara G, Ascher-Jenull J (2018) Extracellular DNA in natural environments: features, relevance and applications. Appl. Microbiol. Biotechnol. 102, 6343–6356. https://doi.org/10.1007/s00253-018-9120-4

Nielsen KM, Smalla K, van Elsas JD (2000) Natural transformation of Acinetobacter sp. strain BD413 with cell lysates of Acinetobacter sp., Pseudomonas fluorescens, and Burkholderia cepacia in soil microcosms. Appl. Environ. Microbiol. 66, 206–212. https://doi.org/10.1128/aem.66.1.206-212.2000

Nocker A, Fernández PS, Montijn R, Schuren F (2012) Effect of air drying on bacterial viability: A multiparameter viability assessment. J. Microbiol. Methods 90, 86–95. https://doi.org/10.1016/j.mimet.2012.04.015

Parducci L, Nota K, Wood J (2018) Reconstructing past vegetation communities using ancient DNA from lake sediments, in: Lindqvist, C., Rajora, O.P. (Eds.), Paleogenomics: Genome-Scale Analysis of Ancient DNA. Springer International Publishing, Cham, pp. 163–187. https://doi.org/10.1007/13836_2018_38

Pedersen MW, Overballe-Petersen S, Ermini L, Sarkissian CD, Haile J, Hellstrom M, Spens J, Thomsen PF, Bohmann K, Cappellini E, Schnell IB, Wales NA, Carøe C, Campos PF, Schmidt AMZ, Gilbert MTP, Hansen AJ, Orlando L, Willerslev E (2015) Ancient and modern environmental DNA. Philos. Trans. R. Soc. Lond. B Biol. Sci. 370, 20130383. https://doi.org/10.1098/rstb.2013.0383

Pietramellara G, Ascher J, Borgogni F, Ceccherini MT, Guerri G, Nannipieri P (2009) Extracellular DNA in soil and sediment: fate and ecological relevance. Biol. Fertil. Soils 45, 219–235. https://doi.org/10.1007/s00374-008-0345-8

Poté J, Ackermann R, Wildi W (2009) Plant leaf mass loss and DNA release in freshwater sediments. Ecotoxicol. Environ. Saf. 72, 1378–1383. https://doi.org/10.1016/j.ecoenv.2009.04.010

Prosser CM, Hedgpeth BM (2018) Effects of bioturbation on environmental DNA migration through soil media. PLoS ONE 13, e0196430. https://doi.org/10.1371/journal.pone.0196430

Ritter CD, Zizka A, Roger F, Tuomisto H, Barnes C, Nilsson RH, Antonelli A (2018) High-throughput metabarcoding reveals the effect of physicochemical soil properties on soil and litter biodiversity and community turnover across Amazonia. PeerJ 6, e5661. https://doi.org/10.7717/peerj.5661

Saeki K, Ihyo Y, Sakai M, Kunito T (2011) Strong adsorption of DNA molecules on humic acids. Environ. Chem. Lett. 9, 505–509. https://doi.org/10.1007/s10311-011-0310-x

Schulz S, Brankatschk R, Dümig A, Kögel-Knabner I, Schloter M, Zeyer J (2013) The role of microorganisms at different stages of ecosystem development for soil formation. Biogeosciences 10, 3983–3996. https://doi.org/10.5194/bg-10-3983-2013

Shackley M (1975) Archaeological sediments: a survey of analytical methods. Butterworth, London and Boston.

Shlemon RJ (1985) Application of soil-stratigraphic techniques to engineering geology. Environmental & Engineering Geoscience xxii, 129–142. https://doi.org/10.2113/gseegeosci.xxii.2.129

Sirois SH, Buckley DH (2019) Factors governing extracellular DNA degradation dynamics in soil. Environ. Microbiol. Rep. 11, 173–184. https://doi.org/10.1111/1758-2229.12725

Smol JP, Birks HJB, Last WM, Bradley RS, Alverson K (Eds) (2001) Tracking environmental change using lake sediments: terrestrial, algal, and siliceous indicators, Developments in paleoenvironmental research. Springer Netherlands, Dordrecht. https://doi.org/10.1007/0-306-47668-1

Taberlet P, Bonin A, Zinger L, Coissac E (Eds) (2018) Environmental DNA: for biodiversity research and monitoring. Oxford University Press. https://doi.org/10.1093/oso/9780198767220.001.0001

Taberlet P, Prud'Homme SM, Campione E, Roy J, Miquel C, Shehzad W, Gielly L, Rioux D, Choler P, Clément J-C, Melodelima C, Pompanon F, Coissac E (2012) Soil sampling and isolation of extracellular DNA from large amount of starting material suitable for metabarcoding studies. Mol. Ecol. 21, 1816–1820. https://doi.org/10.1111/j.1365-294X.2011.05317.x

Thomsen PF, Willerslev E (2015) Environmental DNA – An emerging tool in conservation for monitoring past and present biodiversity. Biological Conservation 183, 4–18. https://doi.org/10.1016/j.biocon.2014.11.019

Timpano EK, Scheible MKR, Meiklejohn KA (2020) Optimization of the second internal transcribed spacer (ITS2) for characterizing land plants from soil. PLoS ONE 15, e0231436. https://doi.org/10.1371/journal.pone.0231436

Torsvik V, Goksøyr J, Daae FL (1990) High diversity in DNA of soil bacteria. Appl. Environ. Microbiol. 56, 782–787. https://doi.org/10.1128/aem.56.3.782-787.1990

Vogt KA, Vogt DJ, Palmiotto PA, Boon P, O'Hara J, Asbjornsen H (1995) Review of root dynamics in forest ecosystems grouped by climate, climatic forest type and species. Plant Soil 187, 159–219. https://doi.org/10.1007/BF00017088

Vuillemin A, Horn F, Alawi M, Henny C, Wagner D, Crowe SA, Kallmeyer J (2017) Preservation and significance of extracellular DNA in ferruginous sediments from Lake Towuti, Indonesia. Front. Microbiol. 8, 1440. https://doi.org/10.3389/fmicb.2017.01440

Wardle DA, Bardgett RD, Klironomos JN, Setälä H, van der Putten WH, Wall DH (2004) Ecological linkages between aboveground and belowground biota. Science 304, 1629–1633. https://doi.org/10.1126/science.1094875

Willerslev E, Cooper A (2005) Ancient DNA. Proc. Biol. Sci. 272, 3–16. https://doi.org/10.1098/rspb.2004.2813

Willerslev E, Davison J, Moora M, Zobel M, Coissac E, Edwards ME, Lorenzen ED, Vestergård M, Gussarova G, Haile J, Craine J, Gielly L, Boessenkool S, Epp LS, Pearman PB, Cheddadi R, Murray D, Bråthen KA, Yoccoz N, Binney H, Taberlet P (2014) Fifty thousand years of Arctic vegetation and megafaunal diet. Nature 506, 47–51. https://doi.org/10.1038/nature12921

Wood JM (1987) Biological processes involved in the cycling of elements between soil or sediments and the aqueous environment. Hydrobiologia 149, 31–42. https://doi.org/10.1007/BF00048644

Yoccoz NG, Bråthen KA, Gielly L, Haile J, Edwards ME, Goslar T, Von Stedingk H, Brysting AK, Coissac E, Pompanon F, Sønstebø JH, Miquel C, Valentini A, De Bello F, Chave J, Thuiller W, Wincker P, Cruaud C, Gavory F, Rasmussen M, Taberlet P (2012) DNA from soil mirrors plant taxonomic and growth form diversity. Mol. Ecol. 21, 3647–3655. https://doi.org/10.1111/j.1365-294X.2012.05545.x

Yoccoz NG (2012) The future of environmental DNA in ecology. Mol. Ecol. 21, 2031–2038. https://doi.org/10.1111/j.1365-294X.2012.05505.x

Zhou L-J, Pei K-Q, Zhou B, Ma K-P (2007) A molecular approach to species identification of Chenopodiaceae pollen grains in surface soil. Am. J. Bot. 94, 477–481. https://doi.org/10.3732/ajb.94.3.477

Zinger L, Bonin A, Alsos IG, Bálint M, Bik H, Boyer F, Chariton AA, Creer S, Coissac E, Deagle BE, De Barba M, Dickie IA, Dumbrell AJ, Ficetola GF, Fierer N, Fumagalli L, Gilbert MTP, Jarman S, Jumpponen A, Kauserud H, Taberlet P (2019a) DNA metabarcoding-Need for robust experimental designs to draw sound ecological conclusions. Mol. Ecol. 28, 1857–1862. https://doi.org/10.1111/mec.15060

Zinger L, Shahnavaz B, Baptist F, Geremia RA, Choler P (2009) Microbial diversity in alpine tundra soils correlates with snow cover dynamics. ISME J. 3, 850–859. https://doi.org/10.1038/ismej.2009.20

Zinger L, Taberlet P, Schimann H, Bonin A, Boyer F, De Barba M, Gaucher P, Gielly L, Giguet-Covex C, Iribar A, Réjou-Méchain M, Rayé G, Rioux D, Schilling V, Tymen B, Viers J, Zouiten C, Thuiller W, Coissac E, Chave J (2019b) Body size determines soil community assembly in a tropical forest. Mol. Ecol. 28, 528–543. https://doi.org/10.1111/mec.14919

Answers

1. By checking if there is a step that can lyse the cells to extract iDNA. This step can be grinding, sonication, thermal shocks, or chemical treatments such as with chloroform.
2. To take into account heterogeneity the strategy is to take many subsamples and mix them.
3. Soil eDNA using *trnL* P6 loop will not give you accurate species lists in most floras, but rather lists of genera with occasional low-level or higher-level identifications. Most vegetation types are characterized by a few key species only, so having limited taxonomic resolution of your identifications is unlikely to affect the overall vegetation type calling. However in some floras or vegetation types this approach will be insufficient, e.g., for those characterized by specific taxa in locally speciose genera.

Chapter 5
DNA from pollen

Marcel Polling[1,2]

1 Naturalis Biodiversity Center, Leiden, The Netherlands
2 Natural History Museum, University of Oslo, Oslo, Norway

Marcel Polling **marcel.polling@wur.nl**

Citation: Polling M (2022) Chapter 5. DNA from pollen. In: de Boer H, Rydmark MO, Verstraete B, Gravendeel B (Eds) Molecular identification of plants: from sequence to species. Advanced Books. https://doi.org/10.3897/ab.e98875

Background

Why use DNA from pollen instead of morphology?

To identify pollen, spores, and other plant-related microremains, the field of palynology has traditionally relied on microscope-based analyses. This is a time-consuming process that requires highly trained specialists. Additionally, pollen grains from many plant families are morphologically indistinguishable using light microscopy (Beug 2004). Therefore, pollen can often not be distinguished beyond the genus- or family-level. Using more advanced microscopy techniques, the finer and potentially species-specific details on the pollen surface (i.e., exine) can be visualised (e.g., scanning electron microscope (SEM) and super-resolution microscopy (see e.g. Sivaguru et al. 2018)). However, these techniques often require extensive sample preparation, highly trained palynologists, and require costly microscopes. Moreover, some pollen grain features are so fine (less than 500 nm) that not even these sophisticated imaging techniques can visualise them. A combination of high-resolution imaging and automatic image detection using sufficiently trained neural networks is another emerging method to increase taxonomic resolution with pollen morphology (Polling et al. 2021; Romero et al. 2020). This technique, however, requires an extensively trained network with a large and varied pollen image reference database.

These challenges highlight the necessity for innovative methods within the field of palynology, to increase both the speed and accuracy of pollen identifications. DNA-based methods for the molecular identification of pollen grains have the potential to be of complementary value. However, the extraction of DNA from pollen is non-trivial. This chapter therefore focuses on how DNA can be extracted from pollen, the common problems encountered, and the qualitative and quantitative molecular possibilities for analyses.

Applications of DNA-based methods for pollen identification

Using pollen grain DNA for identification has shown promising results in a number of applications, including the study of provenance and authentication of honey (Hawkins et al. 2015; Prosser and Hebert 2017; Utzeri et al. 2018), plant-pollinator networks (Pornon et al. 2017; Richardson et al. 2019), hay fever predictions (Campbell et al. 2020; Kraaijeveld et al. 2015; Leontidou et al. 2018), forensic science (Bell et al. 2016a, and references therein), and environmental reconstructions from pollen in soil (Parducci et al. 2017) (see Section 3 for full information on applications). Ancient DNA can be extracted from pollen grains as old as 150 kyr (Suyama et al. 1996), and has also been used for reconstructing ancient plant-pollinator networks (Gous et al. 2019) (see Chapter 21 Palaeobotany).

Collecting pollen for DNA analysis

Collecting pollen for DNA analysis is mostly similar to collecting pollen for microscopic analysis, though more care should be taken to avoid contamination from other potential sources of DNA. This is because pollen generally contains low quantities of DNA and is therefore prone to contamination. Pollen grains can either be collected directly from the environment (air, water, soil, etc.) or from pollinators (pollen baskets, honey). Pollen collected

from the environment will most often (though not always) be derived from anemophilous (wind pollinated) plants, while pollinators collect the majority of pollen from so-called entomophilous (insect pollinated) plants. Pollinators may, however, also have anemophilous pollen accidentally sticking to their bodies. For studies looking at pollen from pollinators, either all pollen grains on the animal's body are collected by washing off the pollen or, when present, only the corbicular pollen baskets are collected (Bell et al. 2017; Richardson et al. 2015). Pollinators can either be collected in the field using aerial netting or collected from natural history collections (Gous et al. 2019). Insect-collected pollen baskets contain many hundreds of thousands of pollen grains, and collecting even a small subset of this basket is sufficient for molecular analysis. Honey also contains huge numbers of pollen grains, but it can be more challenging to work with for DNA analyses. This is because there are many compounds in honey such as polyphenols and flavonoids that can chemically inhibit methods used for DNA sequencing (Prosser and Hebert 2017). In contrast, while airborne pollen grains lack these inhibitors, it is present in only relatively low concentrations in the ambient air. Therefore, to collect sufficient amounts of pollen for molecular analyses, most of the sampling methods focus on air filtration methods. These include both volumetric (e.g., Hirst type; Hirst 1952) and gravimetric methods (for an overview please see Banchi et al. 2020; Levetin 2004).

Pollen DNA extraction

Pollen lysis

Pollen grains can be referred to as "natural plastic": they have a very hard outer cell wall called an exine, which is made of sporopollenin (Brooks and Shaw 1968). Pollen exine is very resistant to non-oxidative physical, biological, and chemical degradation. This is evidenced by their ubiquitous presence in the fossil record and some fossil pollen exines have been found preserved for over 243 million years (Hochuli and Feist-Burkhardt 2013). Extracting DNA from pollen grains is thus not trivial, since the exine must be broken to release the inner DNA. Entomophilous pollen grains also contain DNA-rich pollenkitt outside the exine, but this DNA is usually heavily degraded, and it is the DNA inside the pollen grains that remains intact (Pornon et al. 2017; Pacini and Hesse 2005). A lysis step using mechanical bead-beating and a lysis buffer is often used before DNA extraction of pollen grains, and has been shown to improve DNA quantity (Swenson and Gemeinholzer 2021). However, if the lysis time is too long, or the bead-beating too vigorous, DNA yield may actually decrease. (Swenson and Gemeinholzer 2021) found that best results can be obtained at 33 to 67% exine rupture, instead of 100% exine rupture and using 2 hours of lysis incubation instead of 24 hours. Various different bead-beating strategies have been adopted (Table 1), including using a single relatively large bead (5 mm) or different mixtures of large and small beads. Many different types of material have also been used, including stainless steel, tungsten carbide, glass, and zirconium beads, but the choice of material does not seem to influence the extraction. It is always recommended to test the lysis efficiency, which can be done by checking the fraction of broken (i.e., lysed) pollen grains under the microscope after the bead beating process (e.g. Kraaijeveld et al. 2015).

It should be noted that other methods for DNA extraction from pollen exist in which the pollen grains are not destroyed, and in some specific cases, excluding the bead-beating step has even given better results (Ghitarrini et al. 2018; Gous et al. 2019).

Table 1. Overview of selected studies since 2017 that have used molecular techniques to identify pollen, including the aim, strategy for pollen lysis, extraction method, amount of PCR cycles, sequencing method, and marker choice.

Study	Aim	Pollen lysis step	Extraction method	PCR cycles	Sequencing method	Markers
Leontidou et al. 2018	Airborne pollen identification	Bead beating (one 5 mm stainless steel bead), two 1-min cycles at 30 Hz	DNeasy Plant Mini Kit (Qiagen) and Nucleomag kit (Macherey-Nagel)	30	Sanger sequencing	*trnL*
Lang et al. 2019	Pollen quantification	Bead beating (mix of 0.5 and 1 mm silica beads), 2 min	Wizard (Promega)	N/A	Genome skimming	N/A
Bell et al. 2019	Pollen quantification	Bead beating (mini-bead beater), 3 min	FastDNA SPIN Kit for Soil (MP Biomedicals)	30	Metabarcoding	nrITS2, *rbcL*
Peel et al. 2019	Pollen quantification	Bead beating (five 1 mm stainless steel beads), 2 min at 22.5 Hz	Adapted CTAB	N/A	Genome skimming	N/A
Gous et al. 2019	Plant pollinator interactions over time	Bead beating (one 3 mm stainless steel bead + lysis buffer), 2 min at 25 Hz	QIAamp DNA Micro Kit and DNeasy Plant Mini Kit (Qiagen), Nucleospin DNA Trace Kit (Macherey-Nagel)	30	Metabarcoding	nrITS1, nrITS2, *rbcL*
Brennan et al. 2019	Airborne pollen identification	Bead beating (3 mm tungsten beads), 4 min at 30 Hz	DNeasy Plant Mini Kit (Qiagen)	35	Metabarcoding	nrITS2, *rbcL*
Richardson et al. 2019	Bee pollen diet	Bead beating (3.355 mg 0.7 mm zirconia beads), 5 min	DNeasy Plant Mini kit (Qiagen)	Three steps (55 cycles in total)	Metabarcoding	nrITS2, *rbcL*, *trnL*, *trnH*
Suchan et al. 2019	Insect migration analysis	Bead beating (five zirconium beads), 1 min at 30 Hz	No extraction, using Phire Plant Direct Polymerase	Two steps (32 cycles in total)	Metabarcoding	nrITS2
Baksay et al. 2020	Pollen quantification	CF lysis buffer (Nucleospin Food Kit)	DNeasy Plant Mini Kit (Qiagen)	25, 30, 35	Metabarcoding	nrITS1, *trnL*
Campbell et al. 2020	Airborne pollen identification	Bead beating (0.2 g 425–600 µm glass beads + lysis buffer), two 1-min cycles (3450 oscillations/min)	Adapted CTAB	40	Metabarcoding	*rbcL*
Bänsch et al. 2020; Leidenfrost et al. 2020	Bee pollen diet	Bead beating (150 g mix of 1.4 mm ceramic and 3 mm tungsten beads + lysis buffer), two 45 second cycles at 6.5 m/s	DNeasy Plant Mini Kit (Qiagen)	37	Metabarcoding	nrITS2

DNA extraction

Several commercially available DNA extraction protocols have been used for DNA extraction from pollen grains after the lysis step. Table 1 gives an overview of protocols used in recent literature (for a full overview see Bell et al. 2016b). DNA is most commonly extracted from pollen using the DNeasy Plant Mini Kit (Qiagen) due to its ease of use and high success rate. However, while this is the most commonly used method, recent papers comparing different methods suggest that the best DNA extraction protocol should be empirically found. In one recent paper, several extraction protocols were compared for airborne pollen collected using air samplers (Leontidou et al. 2018). The highest DNA yield was obtained by using a DNA lysis step with steel beads and the Nucleomag Kit. For bee-collected pollen grains, however, the DNeasy Mini Kit gave the best results amongst several different protocols (Gous et al. 2019). Thus, it is always recommended to test several different DNA extraction methods for optimal DNA yield within the chosen study system.

The quality of DNA that can be extracted from pollen samples is critical for any molecularly-based identification method, and particularly when working with very small amounts of DNA. Therefore, avoiding contamination is critical and it is essential to work in a clean lab, to keep windows closed, use sterilised tools in a laminar flow cabinet, and to keep the DNA extraction lab separate from the post-PCR environment.

Molecular methods for pollen identification

Molecular methods can contribute to the analysis of pollen both by identifying which species are present (qualitative) as well as by giving a measure of the abundance of different pollen species (quantification). While DNA metabarcoding methods are currently most often used (Table 1), DNA barcoding techniques have also been applied to target specific species from a mixture, while metagenomics now allows for pollen quantification. For a review of these different sequencing methods, see Chapter 10 DNA barcoding, Chapter 11 Amplicon metabarcoding, and Chapter 12 Metagenomics.

Qualitative pollen analysis

DNA barcoding

Species-resolution in pollen grain identifications is critical for studies that try to answer specific research questions including: what particular species of flower does a common carder bee prefer? What grass species is responsible for most of the pollen in the ambient air in early May? Species-specific markers and qPCR techniques can be used for the identification of specific species within a mixture of different pollen types (see Chapter 10 DNA barcoding). One study used custom-made primers for the nuclear Internal Transcribed Spacer (nrITS) to differentiate between mugwort (*Artemisia vulgaris*) and ragweed (*Ambrosia artemisiifolia*), two notoriously allergenic species from the Asteraceae family (Müller-Germann et al. 2017). These newly constructed primers were then applied on aerobiological samples to show that ragweed pollen can travel long distances, since it was detected outside of the local pollination period. Barcoding was also used to show that allergenic *Juniperus ashei* pollen grains could be found in Canada, even if the closest plants that they could have originated from were located in Texas and Oklahoma, USA (Mohanty et al. 2017). These are two studies that illustrate the potential to identify pollen grains at the species level using DNA-based methods, though this level of resolution is not always necessary. In the

grass family (Poaceae) for example, all species from certain subfamilies are known to have much higher allergenic prevalence than other subfamilies, and therefore subfamily resolution is sufficient for hay fever predictions (Frenguelli et al. 2010). Ghitarrini et al. (2018), for example, used species- but also subfamily-specific primers with real-time PCR to target the most allergenic types of grasses. Pooideae (a subfamily of grasses with many allergenic species) and individual species within this subfamily were detected in aerobiological samples on a presence/absence basis.

DNA metabarcoding

DNA barcoding can be used to target specific species, yet it is rare that a pollen sample contains only a single pollen species. DNA metabarcoding is therefore the most-often used method for the molecular identification of the different species of pollen grains from mixed samples (see Chapter 11 Amplicon metabarcoding). Both nuclear and chloroplast DNA can be amplified in pollen DNA (Bell et al. 2016b), and amongst the many different markers that have been tested, *rbcL*, *trnL*, *matK*, and *trnH-psbA* from the chloroplast, as well as nuclear ribosomal ITS2 (nrITS2), have so far shown the most promise for the molecular identification of pollen grains. Since no universal barcode exists that would allow detection of all plant lineages, a combination of a nuclear and chloroplast marker has been advised (Hollingsworth 2011). nrITS2 (~450 bp) is particularly relevant for the identification of pollen grains when relatively fresh (and non-degraded) DNA is available. In one example, pollen was collected from the bodies of the migratory butterfly species *Vanessa cardui* and identified based on nrITS2, providing geographical information on where the butterflies were migrating from (Suchan et al. 2019). Because several Saharan endemic plants were identified to the species level, this provided excellent evidence for the butterflies originating from the Sahara region.

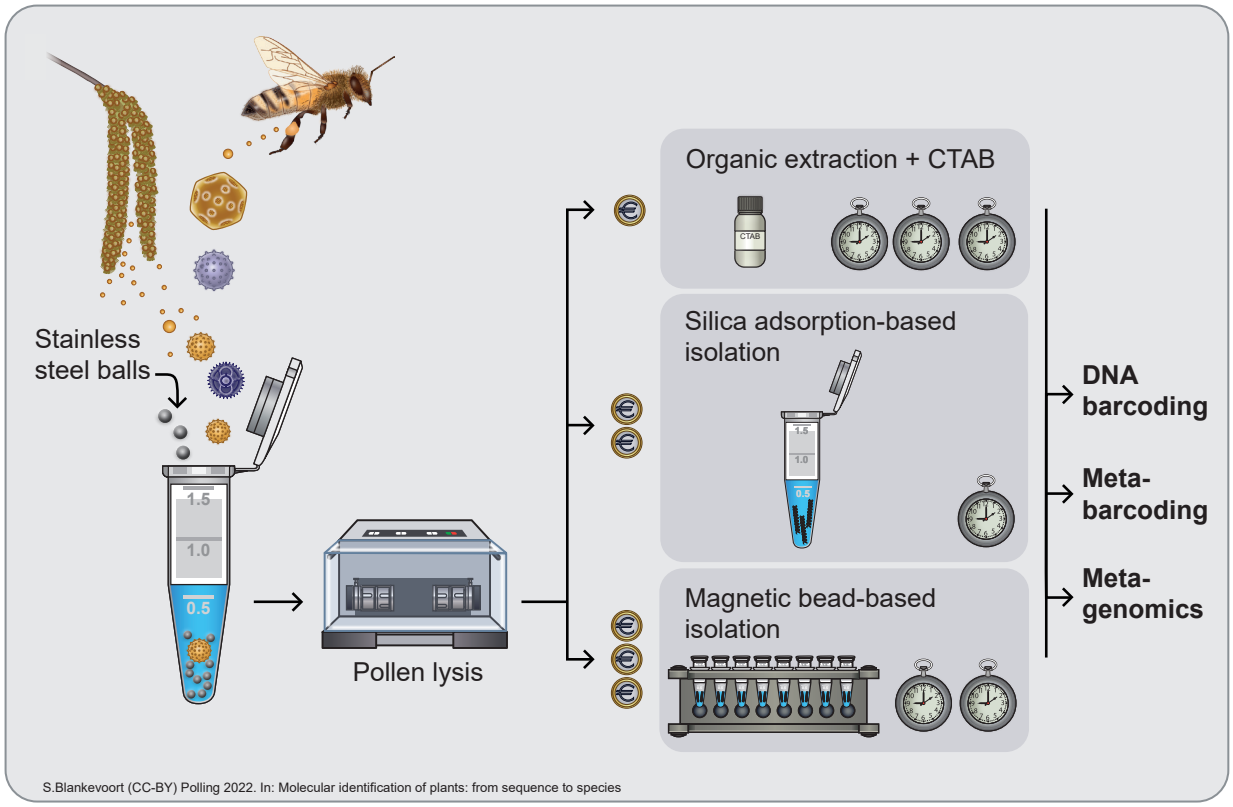

Figure 1. Chapter 5 Infographic: Overview of pollen sources, DNA extraction, and downstream analytical methods for the molecular identification of plants from pollen DNA.

While research into targeting different barcoding regions and primers is ongoing (*trnT-F*; Alan et al. 2019; and ITS1; Baksay et al. 2020), another development is the use of more specific reference databases. The commonly used NCBI GenBank returns many untrustworthy hits since it is not curated (see e.g. Meiklejohn et al. 2019). Brennan et al. (2019) designed a metabarcoding study with two common markers (*rbcL* and nrITS2), but using a strictly curated reference library containing sequences only from those grass species that occurred locally. They further customised this database to include all other invasive as well as cultivated species in the UK. Using their customised database, the authors showed signals in temporally restricted grass genera throughout the grass pollen season, with minimal background from unexpected species that often results from mismatches when using a more generic reference database. Furthermore, they identified that while some genera of grass may flower early in summer in one location, it could be months later for flowering to occur in other locations. This information can be used by hay fever patients to figure out what specific grass genus they are allergic to, and additionally illustrates the relationship between flowering phenology and airborne pollen incidence.

It is important to use positive controls with known concentrations of different pollen species in any DNA metabarcoding study. This is because the amount of DNA that can be extracted from different pollen types has been shown to vary. For example, it can be easier to extract DNA from pollen with a thinner exine and from plant species that are richer in chloroplast DNA than from those having a more 'sturdy' exine (Leontidou et al. 2018). Furthermore, in-silico testing of the chosen primers on target plant species, and making sure reference sequences are available can help to improve the efficiency of the study.

Quantitative pollen analysis

Beyond identifying which pollen species are present in a particular sample, pollen grain quantification is equally important. For example, for hay fever forecasts, it is not just important to know *if* there are certain allergenic pollen in the air, but also how many pollen grains there are at a given point in time. The golden standard for palynology has been to count a certain number of pollen grains under the microscope (e.g., 200 to 500) to obtain a semi-quantitative measure of the pollen types in a sample. While DNA-based methods for pollen quantification are less developed than DNA-based methods for identification, DNA-based pollen quantification using metagenomics (reviewed in Chapter 12 Metagenomics) seems feasible, while there is still strong debate about using DNA metabarcoding reads for this purpose.

DNA metabarcoding reads

In a recent study on the use of DNA to quantify pollen grains, Bell and colleagues found a very weak correlation between pollen counts recorded by palynologists and the proportion of metabarcoding reads (Bell et al. 2019). They constructed different mixtures of known pollen species, and then amplified the marker regions *rbcL* and nrITS2. The authors showed that it depends not only on the species studied, but also on the presence of other species in the mock mixture whether or not this correlation was higher or lower. They identified four metabarcoding related factors that influenced this quantitative bias: copy number, preservation, DNA isolation technique, and amplification bias. Indeed, in many other studies that explore quantification using metabarcoding reads, these factors are often identified as major problems, and DNA metabarcoding reads are therefore mostly used only for relative read abundances in other fields of science (Deagle et al. 2019; Lamb et al. 2019; Pawluczyk et al. 2015).

Another group of scholars, however, are finding more promising results in using DNA metabarcoding to quantify pollen grains. Baksay et al. (2020) for example studied the influence of several factors on quantifying species abundance using mock pollen mixtures, with two commonly found bee-collected pollen species. First, the marker regions nrITS1 and *trnL* were chosen and the amplification results were compared to the number of pollen grains counted using flow cytometry. They found the best results using *trnL* and 30 PCR cycles, or with a high-fidelity PCR polymerase and nrITS1 to circumvent the high GC content in the nuclear ribosomal nrITS region. It is important to note that while *trnL* overall gave the best results for quantification, species-level resolution was only possible with the nrITS1 marker region. Similarly promising results were obtained by Richardson et al. (2019) where a multi-locus approach was used to quantify bee-collected pollen. The amplification results for *trnL* and *rbcL* matched well with the microscopy results, while nrITS2 showed a weak correlation. The authors therefore recommended using the median or mean abundance from several loci to improve the quantification accuracy. Bänsch et al. (2020) in contrast found a high correlation between read count and microscopy count using the nrITS2 region on pollen collected by honey bees and bumblebees. The authors suggested that the correlation depends on the specific type of pollen species studied.

Metagenomic approaches

Since using DNA metabarcoding approaches for pollen abundance may not give quantitative results with complex, multi-species samples, other molecular methods such as genome skimming and shotgun sequencing are being used to circumvent some of the drawbacks. The major advantage of these two methods is that they do not include a PCR-step and therefore do not introduce amplification bias (see Chapter 12 Metagenomics). Genome skimming has already been used to show that quantification is feasible, even for pollen from species that are very rare in mock mixtures (Lang et al. 2019). Because full genomes are only available for less than 1% of all plant species, Peel et al. (2019) developed a method where only partial genome skims are used (0.5X coverage). They found a high correlation between their partial genome skimming results and the expected relative abundance for each pollen type in the mixture. Moreover, the authors indicate that while genome skimming a single pollen sample is still relatively expensive (€70), the advancements made in sequencer technology will help to reduce this price significantly in the near future.

Questions

1. What are the main advantages of molecular pollen identification over traditional (microscopic) methods? Justify your answer.
2. Pollen is dispersed by various vectors. There are two main types of pollination strategies in land plants, please name them and also explain the importance of the difference between the two in terms of DNA yield.
3. Which four factors make the quantification of pollen grains using metabarcoding problematic?

Glossary

Anemophilous – Wind-pollinated.
Bead beating – The application of beads to break open the outer cell wall of pollen grains.

Hirst-type pollen trap – Volumetric air sampler that is one of the standard devices for monitoring airborne pollen and spores.
cpDNA – Chloroplast DNA.
Entomophilous – Insect-pollinated.
Exine – Outer wall of pollen grains. Composed mainly of sporopollenin that is extremely resistant to degradation. The exine of pollen grains has to be broken to release the DNA from the organic material within the grains.
Palynology – The science that studies both living and fossil spores, pollen grains, and other microscopic structures (e.g., chironomids, dinocysts, acritarchs, chitinozoans, scolecodonts).
Pollen grains – The male gametophyte of seed plants; source and carrier for the male gametes (spermatozoids or sperm cells).
Pollenkitt – The outermost hydrophobic lipid layer mostly present on entomophilous pollen grains.
Sporopollenin – A chemically inert biological polymer that is a component of the outer wall (see Exine) of a pollen grain.
Super-resolution microscopy – Technique in optical microscopy that allows visualisation of images with resolutions up to 140 nm, much higher than those imposed by the diffraction limit. This technique also allows visualisation of internal structures.

References

Alan Ş, Sarışahin T, Şahin AA, Kaplan A, Erdoğan İ, Pınar NM (2019) A new method to quantify atmospheric Poaceae pollen DNA based on the trnT-F cpDNA region. Turk. J. Bioch. 44, 248–253. https://doi.org/10.1515/tjb-2018-0020

Baksay S, Pornon A, Burrus M, Mariette J, Andalo C, Escaravage N (2020) Experimental quantification of pollen with DNA metabarcoding using ITS1 and trnL. Sci. Rep. 10, 4202. https://doi.org/10.1038/s41598-020-61198-6

Banchi E, Pallavicini A, Muggia L (2020) Relevance of plant and fungal DNA metabarcoding in aerobiology. Aerobiologia (Bologna) 36, 9–23. https://doi.org/10.1007/s10453-019-09574-2

Bänsch S, Tscharntke T, Wünschiers R, Netter L, Brenig B, Gabriel D, Westphal C (2020) Using ITS2 metabarcoding and microscopy to analyse shifts in pollen diets of honey bees and bumble bees along a mass-flowering crop gradient. Mol. Ecol. 29, 5003–5018. https://doi.org/10.1111/mec.15675

Bell KL, Burgess KS, Botsch JC, Dobbs EK, Read TD, Brosi BJ (2019) Quantitative and qualitative assessment of pollen DNA metabarcoding using constructed species mixtures. Mol. Ecol. 28, 431–455. https://doi.org/10.1111/mec.14840

Bell KL, Burgess KS, Okamoto KC, Aranda R, Brosi BJ (2016a) Review and future prospects for DNA barcoding methods in forensic palynology. Forensic Sci. Int. Genet. 21, 110–116. https://doi.org/10.1016/j.fsigen.2015.12.010

Bell KL, de Vere N, Keller A, Richardson RT, Gous A, Burgess KS, Brosi BJ (2016b) Pollen DNA barcoding: current applications and future prospects. Genome 59, 629–640. https://doi.org/10.1139/gen-2015-0200

Bell KL, Fowler J, Burgess KS, Dobbs EK, Gruenewald D, Lawley B, Morozumi C, Brosi BJ (2017) Applying pollen DNA metabarcoding to the study of plant-pollinator interactions. Appl. Plant Sci. 5, 1600124. https://doi.org/10.3732/apps.1600124

Beug H-J (2004) Leitfaden der Pollenbestimmung für Mitteleuropa und angrenzende Gebiete. Friedrich Pfeil, München.

Brennan GL, Potter C, de Vere N, Griffith GW, Skjøth CA, Osborne NJ, Wheeler BW, McInnes RN, Clewlow Y, Barber A, Hanlon HM, Hegarty M, Jones L, Kurganskiy A, Rowney FM, Armitage C, Adams-Groom B, Ford CR, Petch GM, PollerGEN Consortium Creer S (2019) Temperate airborne grass pollen defined by spatio-temporal shifts in community composition. Nat. Ecol. Evol. 3, 750–754. https://doi.org/10.1038/s41559-019-0849-7

Brooks J, Shaw G (1968) Chemical structure of the exine of pollen walls and a new function for carotenoids in nature. Nature 219, 532–533. https://doi.org/10.1038/219532a0

Campbell BC, Al Kouba J, Timbrell V, Noor MJ, Massel K, Gilding EK, Angel N, Kemish B, Hugenholtz P, Godwin ID, Davies JM (2020) Tracking seasonal changes in diversity of pollen allergen exposure: Targeted metabarcoding of a subtropical aerobiome. Sci. Total Environ. 747, 141189. https://doi.org/10.1016/j.scitotenv.2020.141189

Deagle BE, Thomas AC, McInnes JC, Clarke LJ, Vesterinen EJ, Clare EL, Kartzinel TR, Eveson JP (2019) Counting with DNA in metabarcoding studies: how should we convert sequence reads to dietary data? Mol. Ecol. 28, 391–406. https://doi.org/10.1111/mec.14734

Frenguelli G, Passalacqua G, Bonini S, Fiocchi A, Incorvaia C, Marcucci F, Tedeschini E, Canonica GW, Frati F (2010) Bridging allergologic and botanical knowledge in seasonal allergy: a role for phenology. Ann. Allergy Asthma Immunol. 105, 223–227. https://doi.org/10.1016/j.anai.2010.06.016

Ghitarrini S, Pierboni E, Rondini C, Tedeschini E, Tovo GR, Frenguelli G, Albertini E (2018) New biomolecular tools for aerobiological monitoring: Identification of major allergenic Poaceae species through fast real-time PCR. Ecol. Evol. 8, 3996–4010. https://doi.org/10.1002/ece3.3891

Gous A, Swanevelder DZH, Eardley CD, Willows-Munro S (2019) Plant-pollinator interactions over time: Pollen metabarcoding from bees in a historic collection. Evol. Appl. 12, 187–197. https://doi.org/10.1111/eva.12707

Hawkins J, de Vere N, Griffith A, Ford CR, Allainguillaume J, Hegarty MJ, Baillie L, Adams-Groom B (2015) Using DNA metabarcoding to identify the floral composition of honey: A new tool for investigating honey bee foraging preferences. PLoS ONE 10, e0134735. https://doi.org/10.1371/journal.pone.0134735

Hirst JM (1952) AN AUTOMATIC VOLUMETRIC SPORE TRAP. Ann. Applied Biology 39, 257–265. https://doi.org/10.1111/j.1744-7348.1952.tb00904.x

Hochuli PA, Feist-Burkhardt S (2013) Angiosperm-like pollen and Afropollis from the Middle Triassic (Anisian) of the Germanic Basin (Northern Switzerland). Front. Plant Sci. 4, 344. https://doi.org/10.3389/fpls.2013.00344

Hollingsworth PM (2011) Refining the DNA barcode for land plants. Proc Natl Acad Sci USA 108, 19451–19452. https://doi.org/10.1073/pnas.1116812108

Kraaijeveld K, de Weger LA, Ventayol García M, Buermans H, Frank J, Hiemstra PS, den Dunnen JT (2015) Efficient and sensitive identification and quantification of airborne pollen using next-generation DNA sequencing. Mol. Ecol. Resour. 15, 8–16. https://doi.org/10.1111/1755-0998.12288

Lamb PD, Hunter E, Pinnegar JK, Creer S, Davies RG, Taylor MI (2019) How quantitative is metabarcoding: A meta-analytical approach. Mol. Ecol. 28, 420–430. https://doi.org/10.1111/mec.14920

Lang D, Tang M, Hu J, Zhou X (2019) Genome-skimming provides accurate quantification for pollen mixtures. Mol. Ecol. Resour. 19, 1433–1446. https://doi.org/10.1111/1755-0998.13061

Leidenfrost RM, Bänsch S, Prudnikow L, Brenig B, Westphal C, Wünschiers R (2020) Analyzing the dietary diary of bumble bee. Front. Plant Sci. 11, 287. https://doi.org/10.3389/fpls.2020.00287

Leontidou K, Vernesi C, De Groeve J, Cristofolini F, Vokou D, Cristofori A (2018) DNA metabarcoding of airborne pollen: new protocols for improved taxonomic identification of environmental samples. Aerobiologia (Bologna) 34, 63–76. https://doi.org/10.1007/s10453-017-9497-z

Levetin E (2004) Methods for aeroallergen sampling. Curr. Allergy Asthma Rep. 4, 376–383. https://doi.org/10.1007/s11882-004-0088-z

Meiklejohn KA, Damaso N, Robertson JM (2019) Assessment of BOLD and GenBank - Their accuracy and reliability for the identification of biological materials. PLoS ONE 14, e0217084. https://doi.org/10.1371/journal.pone.0217084

Mohanty RP, Buchheim MA, Levetin E (2017) Molecular approaches for the analysis of airborne pollen: A case study of Juniperus pollen. Ann. Allergy Asthma Immunol. 118, 204-211.e2. https://doi.org/10.1016/j.anai.2016.11.015

Müller-Germann I, Pickersgill DA, Paulsen H, Alberternst B, Pöschl U, Fröhlich-Nowoisky J, Després VR (2017) Allergenic Asteraceae in air particulate matter: quantitative DNA analysis of mugwort and ragweed. Aerobiologia (Bologna) 33, 493–506. https://doi.org/10.1007/s10453-017-9485-3

Pacini E, Hesse M (2005) Pollenkitt – its composition, forms and functions. Flora - Morphology, Distribution, Functional Ecology of Plants 200, 399–415. https://doi.org/10.1016/j.flora.2005.02.006

Parducci L, Bennett KD, Ficetola GF, Alsos IG, Suyama Y, Wood JR, Pedersen MW (2017) Ancient plant DNA in lake sediments. New Phytol. 214, 924–942. https://doi.org/10.1111/nph.14470

Pawluczyk M, Weiss J, Links MG, Egaña Aranguren M, Wilkinson MD, Egea-Cortines M (2015) Quantitative evaluation of bias in PCR amplification and next-generation sequencing derived from metabarcoding samples. Anal. Bioanal. Chem. 407, 1841–1848. https://doi.org/10.1007/s00216-014-8435-y

Peel N, Dicks LV, Clark MD, Heavens D, Percival-Alwyn L, Cooper C, Davies RG, Leggett RM, Yu DW (2019) Semi-quantitative characterisation of mixed pollen samples using MinION sequencing and Reverse Metagenomics (RevMet). Methods Ecol. Evol. 10, 1690–1701. https://doi.org/10.1111/2041-210X.13265

Polling M, Li C, Cao L, Verbeek F, de Weger LA, Belmonte J, De Linares C, Willemse J, de Boer H, Gravendeel B (2021) Neural networks for increased accuracy of allergenic pollen monitoring. Sci. Rep. 11, 11357. https://doi.org/10.1038/s41598-021-90433-x

Pornon A, Andalo C, Burrus M, Escaravage N (2017) DNA metabarcoding data unveils invisible pollination networks. Sci. Rep. 7, 16828. https://doi.org/10.1038/s41598-017-16785-5

Prosser SWJ, Hebert PDN (2017) Rapid identification of the botanical and entomological sources of honey using DNA metabarcoding. Food Chem. 214, 183–191. https://doi.org/10.1016/j.foodchem.2016.07.077

Richardson RT, Curtis HR, Matcham EG, Lin C-H, Suresh S, Sponsler DB, Hearon LE, Johnson RM (2019) Quantitative multi-locus metabarcoding and waggle dance interpretation reveal honey bee spring foraging patterns in Midwest agroecosystems. Mol. Ecol. 28, 686–697. https://doi.org/10.1111/mec.14975

Richardson RT, Lin C-H, Sponsler DB, Quijia JO, Goodell K, Johnson RM (2015) Application of ITS2 metabarcoding to determine the provenance of pollen collected by honey bees in an agroecosystem. Appl. Plant Sci. 3, 1400066. https://doi.org/10.3732/apps.1400066

Romero IC, Kong S, Fowlkes CC, Jaramillo C, Urban MA, Oboh-Ikuenobe F, D'Apolito C, Punyasena SW (2020) Improving the taxonomy of fossil pollen using convolutional neural networks and superresolution microscopy. Proc. Natl. Acad. Sci. USA 17, 28496–28505. https://doi.org/10.1073/pnas.2007324117

Sivaguru M, Urban MA, Fried G, Wesseln CJ, Mander L, Punyasena SW (2018) Comparative performance of airyscan and structured illumination superresolution microscopy in the study of the surface texture and 3D shape of pollen. Microsc. Res. Tech. 81, 101–114. https://doi.org/10.1002/jemt.22732

Suchan T, Talavera G, Sáez L, Ronikier M, Vila R (2019) Pollen metabarcoding as a tool for tracking long-distance insect migrations. Mol. Ecol. Resour. 19, 149–162. https://doi.org/10.1111/1755-0998.12948

Suyama Y, Kawamuro K, Kinoshita I, Yoshimura K, Tsumura Y, Takahara H (1996) DNA sequence from a fossil pollen of Abies spp. from Pleistocene peat. Genes Genet. Syst. 71, 145–149. https://doi.org/10.1266/ggs.71.145

Swenson SJ, Gemeinholzer B (2021) Testing the effect of pollen exine rupture on metabarcoding with Illumina sequencing. PLoS ONE 16, e0245611. https://doi.org/10.1371/journal.pone.0245611

Utzeri VJ, Schiavo G, Ribani A, Tinarelli S, Bertolini F, Bovo S, Fontanesi L (2018) Entomological signatures in honey: an environmental DNA metabarcoding approach can disclose information on plant-sucking insects in agricultural and forest landscapes. Sci. Rep. 8, 9996. https://doi.org/10.1038/s41598-018-27933-w

Answers

1. A higher taxonomic resolution can be achieved using molecular methods such as metabarcoding. Furthermore, pollen analysis requires highly trained experts that have to spend considerable time to analyse a single sample and therefore molecular techniques are faster, especially with a large number of samples.
2. Entomophilous (insect collected) and anemophilous (wind dispersed) pollen. The presence of pollenkitt on entomophilous pollen grains influences the amount of DNA that can be obtained per pollen grain.
3. Copy number, DNA preservation, DNA isolation technique, and amplification bias.

Chapter 6
DNA from food and medicine

Felicitas Mück[1], Carlos A. Vásquez-Londoño[2]

1 Department of Pharmacy, Faculty of Mathematics and Natural Sciences, University of Oslo, Norway
2 Department of Pharmacy, Faculty of Sciences, National University of Colombia, Colombia

Felicitas Mück **felicitas.mueck@farmasi.uio.no**
Carlos A. Vásquez-Londoño **cavasquezl@unal.edu.co**

Citation: Mück F, Vásquez-Londoño CA (2022) Chapter 6. DNA from food and medicine. In: de Boer H, Rydmark MO, Verstraete B, Gravendeel B (Eds) Molecular identification of plants: from sequence to species. Advanced Books. https://doi.org/10.3897/ab.e98875

Why use DNA for the identification of food and medicine?

DNA-based methods for the molecular identification of plant products can help us to address food and medicine authenticity issues at each stage in the supply chain (Di Bernardo et al. 2007). Documentation and requirements for DNA-based detection methods for food authentication are defined in collaborative activities by the European Committee for Standardization (CEN) and the International Organization for Standardization (ISO). Both rapid and accurate identification of plant products are crucial for the the herbal drug industry (Mishra et al. 2016), where DNA-based authentication is recognised as a sensitive approach to identify edible and medicinal plant species, cultivars, and to detect their substitutes and adulterants in crude or processed products independent from life stage, tissue type, and physiological conditions of their constituents (Howard et al. 2020; Lo and Shaw 2018a, 2019; Mishra et al. 2016; Pawar et al. 2017; Raclariu et al. 2018; Techen et al. 2014). Molecular methods were integrated in the Pharmacopoeia of China in 2020 (Pharmacopoeia Committee of P. R. China 2020) and validated by the British Pharmacopoeia Commission in 2018 (British Pharmacopoeia Commission 2018) . In addition to species authentication, genetic identification of medicinal plants can assist in the field of pharmacophylogenomics and bioprospecting to discover new plant pharmaceutical resources (Hao et al. 2015) (Chapter 22 Healthcare).

DNA-based methods to identify plants in food and medicine

The majority of standardised DNA-based authentication methods for the inspection and regulation of food and plant-medicines use well-established PCR-based techniques for DNA amplification as these are sensitive, specific, and simple (Hirst et al. 2019). PCR-based authentication methods are standardised in the CEN international legislation mainly for the detection of genetically modified foods (GMOs) including soybean, hazelnut, almond, rapeseed, etc. (Grohmann and Seiler 2019).

DNA barcoding methods are also established for the identification of unique medicinal and edible plant species (Ichim 2019). High resolution melting (HRM) in combination with DNA barcoding (Bar-HRM) can also be used to identify barcode differences in complex botanical matrices and assess the quality of crude materials in the herbal supply chain (Mezzasalma et al. 2017) (See Chapter 13 Barcoding - High resolution melting).

High-throughput sequencing (HTS) methods such as amplicon metabarcoding are also powerful tools for the authentication of herbal end products, post-marketing control, pharmacovigilance, and the assessment of species composition in botanical medicines, such as in traditional Chinese medicines (TCMs) (Arulandhu et al. 2017; De Boer et al. 2015; Juul et al. 2015; Lo and Shaw 2018b; Omelchenko et al. 2019; Raclariu et al. 2018; Seethapathy et al. 2019). An example of the discriminatory power of DNA metabarcoding is revealed in a study in which 15 highly processed TCM ingredients could be identified as species and genera listed on CITES appendices I and II (Coghlan et al. 2012).

In addition to PCR-based techniques, the detection of single nucleotide polymorphisms (SNPs) is frequently used for the molecular identification and authentication of various food commodities using small DNA fragments (Di Bernardo et al. 2007; Lo and Shaw 2019). Several

assay types are commercially available for SNP chips and similar technologies (Hirst et al. 2019). Metagenomics is also promising for the qualitative and quantitative analysis of processed food and medicine matrices (Raime et al. 2020).

DNA-based methods for molecular plant identification depend on well-curated nucleotide sequence repositories. In addition to GenBank (Benson et al. 2018) and the Barcode of Life database (BOLD) (Ratnasingham and Hebert 2007), the Medicinal Materials DNA Barcode Database (MMDBD) has been proposed as a sequence reference platform to identify medicinal plant, animal, and fungi species (Wong et al. 2018).

DNA isolation from food and medicines

Successful DNA extraction is the foundation for any further downstream analysis (Corrado 2016; Elsanhoty et al. 2011; Pinto et al. 2007; Turkec et al. 2015). Since food and medicine products can differ in molecular characteristics and structural form, the choice for a DNA isolation strategy must be sample specific. DNA extraction from most food products are based on DNA isolation techniques originally designed in the 1980s (Dellaporta et al. 1983; Lockley and Bardsley 2000), though these protocols are now typically adapted to include polymerases resistant to the inhibitors commonly found in a wide range of food and medicinal products (Omelchenko et al. 2019). Frequently used DNA extraction procedures are phenol-chloroform, detergent, and protease-based extraction methods and solid-phase extraction methods (see Chapter 1 DNA from plant tissue).

Factors affecting the efficacy of DNA extraction

Four main factors that affect the efficacy of DNA isolation from food and medicine samples are the sample source and processing, collection and storage, homogenisation, and the presence of contaminants. Generally it is easier to extract high-quality DNA from fresh samples (Peterson et al. 1997) since processing techniques often involve factors (e.g., high temperatures and changes in pH) that reduce the quality of DNA (Gryson 2010; Gryson et al. 2004). Secondly, samples need to be stored in low temperature conditions to reduce nuclease activity. Chemical inhibitors can also be used during collection and storage to lower the risk of hydrolysis and block nuclease activity. Sample homogenization is necessary to ensure that the purified DNA samples are representative of the complete original sample as well as to reduce DNA interactions with high molecular weight compounds such as polysaccharides (Wood 2002). Mechanical grinding with a mortar and pestle, disruption via agitation in the presence of ceramic and metal beads, and mechanical shearing with the help of grinding mills can be used. Alternatively, hydrolysing enzymes, or grinding in presence of liquid nitrogen can disrupt problematic plant material, such as samples with a high content of hardened cell walls. Fourthly, optimization of DNA extraction protocols is often necessary to reduce contaminating constituents, like plant secondary metabolites, proteins, etc. (see Table 1) (Wilkes 2019). In particular, spices and teas are rich in secondary metabolites, bark, roots, hard seeds, etc. (Omelchenko et al. 2019). The cetyltrimethylammonium bromide (CTAB) isolation method (Murray and Thompson 1980) is mostly used for unknown multi-herbal samples or samples with high quantities of polysaccharides (Arulandhu et al. 2017). This usually includes serine protease within the extraction buffer to remove proteins. The enzymatic activity of proteinase K is accelerated by sample incubation with the extraction buffer at 56 °C. Additionally, the initial lysis can be prolonged for optimal results.

Table 1. Removal of frequent contaminants that can reduce the yield of extracted DNA from edible and medicinal plants.

Proteins and RNA			
What compounds define the chemical composition of your samples?	**Polysaccharides (starch, sugars)**		**Polyphenolics**
	RNA		**(plant secondary metabolites like: tannins, flavonoids, terpenoids, etc.)**
Understand the specific properties of your samples for DNA extraction	Can co-purify with DNA	Can co-precipitate with DNA	When bound to DNA very hard to remove in extraction
	dependending on the age of the samples and how they were conserved	Results in a sticky viscous consistency to DNA pellet after centrifugation	
		Inhibition of enzymes used for molecular techniques (restriction endonucleases, polymerases, and ligases (Pandey et al. 1996))	Results in contaminated pellets not usable for many downstream analyses (John 1992; Peterson et al. 1997)
		Adherence to wells in agarose gel residing in long smears of bands detected in gel (Sharma et al. 2002)	
Consider applying mitigation strategies to overcome difficulties in extracting DNA from your samples	RNA removable with DNase-free RNase A or ethanol precipitation using lithium chloride	Removal via highly concentrated sodium chloride (NaCl) in extraction buffers leading to increased solubility in ethanol	Binder compounds polyvinyl pyrrolidone (PVP) or polypyrrolidone (PVPP) can be used in extraction buffers to absorb polyphenols before polymerization with DNA
	Proteins can be removed by i) inclusion of detergents (cetyltrimethylammonium bromide (CTAB), SDS) in extraction buffer	Combination of NaCl and cationic detergent CTAB	
		CTAB with differential precipitation (Murray and Thompson 1980)	Use of antioxidant compounds (BME, DDT, ascorbic acid, iso-ascorbate) in buffer to prevent polymerization (Pich and Schubert 1993; Puchooa 2004)
	ii) protein denaturants e.g., β-mercaptoethanol (BME), dithiothreitol (DTT)		
	iii) enzymatic proteases e.g., proteinase K		

Although CTAB-based methods usually result in DNA extraction from plants and processed food and medicine products, the quantity is often quite low and the protocols are time consuming (Costa et al. 2015; Grazina et al. 2020). Many commercial DNA extraction kits are based on solid phase DNA purification (Boom et al. 1990) and have been well adopted for DNA isolation from specific matrices from various organic materials. Optimization and specification of such extraction protocols can be achieved by modifying wash buffer composition, extraction reagents, etc. in-house.

Commercial vs. in-house DNA isolation techniques

Several studies exist that compare commercial and in-house DNA isolation techniques for food and medicine (Costa et al. 2015; Di Bernardo et al. 2007; Omelchenko et al. 2019; Pafundo et al. 2011; Pinto et al. 2007; Smith et al. 2005). These studies indicate that the best method for DNA isolation is highly sample dependent. For example, silica membrane spin column based

kits and sorbent-based kits produce higher DNA yields for teas and spices, whereas the CTAB method based on liquid-phase segregation was superior for more processed herbal remedies (Omelchenko et al. 2019), while extremely processed foods or medicinal extracts could hardly be analysed as a result of total DNA degradation (Grazina et al. 2020; Llongueras et al. 2013; Parveen et al. 2016). In one comparative study, eight different DNA extraction kits were tested for 13 medicinal plant products. Nucleospin plant methods overall yielded the best purity and amplification results for DNA extraction from degraded samples, while DNeasy kits resulted in the highest yields of extracted DNA from botanicals (Llongueras et al. 2013). This suggests that DNA extraction using commercial kits is highly sample dependent, and that there is no universal protocol to extract DNA from herbal products (Grazina et al. 2020; Llongueras et al. 2013). Nevertheless, the European Union Reference Laboratory for GM Food and Feed (EU-RL GMFF) recommends the use of certain extraction methods (Table 2), which can be a helpful guideline for choosing the right extraction protocol and adapting it to a specific sample source.

Table 2. Overview of different DNA extraction methods recommended for use with food by the European Union Reference Laboratory for GM Food and Feed (EU-RL GMFF).

Plant source	Method of choice	Reference
Maize Maize seeds and grains	**CTAB precipitate (in-house)** (Rogers and Bendich 1985) For isolation of genomic DNA from a wide variety of maize tissues and derived matrices for high-quality genomic DNA from processed plant tissue (e.g., leaf, grain, or seed). Lysis step (thermal lysis in the presence of Tris HCl, EDTA, CTAB, and β-mercaptoethanol). Tissues processed prior to extraction procedure. Possible methods of processing include a mortar and pestle with liquid nitrogen (leaf) or commercial blender (grain or seed).	CRLVL16/05XP corrected version 2 01/03/2018
Soybean Soybean seeds	**CTAB precipitate (in-house)** (Dellaporta et al. 1983) "Dellaporta-derived" method starts with a lysis step (thermal lysis in the presence of Tris HCl, EDTA, NaCl, and β-mercaptoethanol). Isopropanol precipitation and removal of contaminants such as lipophilic molecules and proteins by extraction with phenol:chloroform:isoamyl alcohol.	CRLVL13/05XP 14/05/2007
Potato Freeze-dried potato tubers	**"CTAB/Microspin" method** Lysis step (thermal lysis in the presence of CTAB, EDTA, and proteinase K). Removal of RNA by digestion with RNase A and removal of contaminants such as lipophilic molecules and proteins by extraction with chloroform. Remaining inhibitors are removed by a gel filtration step using the commercially available product S-300 HR Microspin Columns (Amersham Pharmacia).	CRLVL09/05XP Corrected Version 1 20/01/2009
Rapeseed	**CTAB precipitate (in-house) (Dellaporta et al. 1983)** Lysis step (thermal lysis in the presence of Tris HCl, EDTA, SDS, and β-mercaptoethanol). Removal of contaminants such as lipophilic molecules and proteins by extraction with phenol and chloroform. DNA precipitate is generated by using isopropanol. The pellet is dissolved in TE buffer.	CRLVL14/04XP Corrected Version 1 15/01/2007

Plant source	Method of choice	Reference
Rapeseed	Inhibitors are removed by an anion exchange chromatography step using the DNA Clean & Concentrator 25 kit (Zymo Research).	CRLVL14/04XP Corrected Version 1 15/01/2007
Multi-herbal products	**CTAB precipitate (in-house)** (Murray and Thompson 1980) Technique is ideal for the rapid isolation of small amounts of DNA from many different species and is also useful for large scale isolations. Lysis step (thermal lysis in the presence of Tris HCl, EDTA, CTAB, and β-mercaptoethanol). Removal of contaminants such as lipophilic molecules and proteins by extraction with phenol and chloroform. Samples processed prior to extraction procedure (mortar and pestle, liquid nitrogen, or commercial blender).	Arulandhu et al. 2017

Analysing the quantity and purity of extracted DNA

After DNA extraction, measuring both the DNA concentration and purity is important before continuing with further downstream analysis. Isolated DNA can be tested for quality using absorbance methods, agarose gel electrophoresis, and fluorescent DNA-intercalating dyes (Wilkes 2019). DNA concentration can be determined with the help of optical density (when the ratios of 260/280 nm and 260/230 nm are between 1.5 and 2.0, the isolated DNA can be used for amplification) (Lo and Shaw 2018a; Matsuoka et al. 2001; Wilkes 2019). Additionally, for sequencing, it is recommended to include a positive control to avoid false negatives that could be due to the presence of PCR inhibitors (Hoorfar et al. 2004; Lo and Shaw 2018a).

The reality of DNA-based identification

It is in the interest of both biodiversity conservation and public safety that DNA-based techniques are further developed to screen food and medicine sourced from the global market (Han et al. 2016; Ichim and de Boer 2020; Seethapathy et al. 2019). Standards for taxon-specific PCR techniques are described for some food plants, like soybean, hazelnut, almond, and rapeseed, but these are established mostly for allergen and GMP testing (Grohmann and Seiler 2019). HTS techniques are the only control tests that ensure the potential identification of all species in complex medicine and food products without prior knowledge on expected adulterants. These have become increasingly popular and more affordable methods for commercial laboratories. Today, laboratories offer analysis for meat, plants (including spices and herbs), fish, and crustaceans (Hirst et al. 2019). However, market tests have come with a higher detection limit and cannot guarantee a 100% confirmation of the presence or absence of an adulterant species (Hirst et al. 2019). Thus, the gap between the expectations and reality for DNA testing of food and medicine is the availability of commercial tests, the limit for detection and quantification, the specificity and the comparison with accurate reference materials and databases. In all these applications, however, the extraction of sufficient quantities of (reasonably) high-quality DNA are necessary. We can expect however that as the interest in using DNA-based methods

for the authentication of food and medicine grows, that further development in methods related to extracting DNA from these often-challenging sources will also progress.

Questions

1. The quality of DNA from food and medicinal sources is a critical factor for DNA-based analyses. Which factors can influence the quality of nucleic acids extracted from foods and plant-based medicines?
2. What is the first step when choosing a DNA isolation technique for your samples?
3. What methods can be used for measuring DNA quality after isolation?

Glossary

Bioprospecting - The exploration of biodiversity for new resources of social and commercial value.

Pharmacophylogenomics - Plant pharmacophylogenomics is a field established by combining the fields of ethnopharmacology, plant systematics, phytochemistry, pharmacology, and bioinformatics. It is the application of phylogenomics to the study of pharmaceuticals.

Pharmacopoeia - From the obsolete typography *pharmacopœia*, literally, "drug-making". In its modern technical sense, it is a book containing directions for the identification of compound medicines, and is published by the authority of a government or a medical or pharmaceutical society.

Pharmaphylogenetics - Field of research focusing on the phylogenetic correlation between phylogeny, chemical constituents, and pharmaceutical effects of medicinal plants.

References

Arulandhu AJ, Staats M, Hagelaar R, Voorhuijzen MM, Prins TW, Scholtens I, Costessi A, Duijsings D, Rechenmann F, Gaspar FB, Barreto Crespo MT, Holst-Jensen A, Birck M, Burns M, Haynes E, Hochegger R, Klingl A, Lundberg L, Natale C, Niekamp H, Kok E (2017) Development and validation of a multi-locus DNA metabarcoding method to identify endangered species in complex samples. Gigascience 6, 1–18. https://doi.org/10.1093/gigascience/gix080

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Ostell J, Pruitt KD, Sayers EW (2018) GenBank. Nucleic Acids Res. 46, D41–D47. https://doi.org/10.1093/nar/gkx1094

Boom R, Sol CJ, Salimans MM, Jansen CL, Wertheim-van Dillen PM, van der Noordaa J (1990) Rapid and simple method for purification of nucleic acids. J. Clin. Microbiol. 28, 495–503. https://doi.org/10.1128/jcm.28.3.495-503.1990

British Pharmacopoeia Commission (2018) Supplementary chapter SC VII D, in: DNA Barcoding as a Tool for Botanical Identification of Herbal Drugs. The Stationary Office, London, United Kingdom.

Coghlan ML, Haile J, Houston J, Murray DC, White NE, Moolhuijzen P, Bellgard MI, Bunce M (2012) Deep sequencing of plant and animal DNA contained within traditional Chinese medicines reveals legality issues and health safety concerns. PLoS Genet. 8, e1002657. https://doi.org/10.1371/journal.pgen.1002657

Corrado G (2016) Advances in DNA typing in the agro-food supply chain. Trends Food Sci. Technol. 52, 80–89. https://doi.org/10.1016/j.tifs.2016.04.003

Costa J, Amaral JS, Fernandes TJR, Batista A, Oliveira MBPP, Mafra I (2015) DNA extraction from plant food supplements: influence of different pharmaceutical excipients. Mol. Cell. Probes 29, 473–478. https://doi.org/10.1016/j.mcp.2015.06.002

Dellaporta SL, Wood J, Hicks JB (1983) A plant DNA minipreparation: Version II. Plant Mol. Biol. Rep. 1, 19–21. https://doi.org/10.1007/BF02712670

De Boer HJ, Cross HB, De Wilde WJJO, Duyfjes-de Wilde BEE, Gravendeel B (2015) Molecular phylogenetic analyses of Cucurbitaceae tribe Benincaseae urge for merging of Pilogyne with Zehneria. Phytotaxa 236, 173. https://doi.org/10.11646/phytotaxa.236.2.6

Di Bernardo G, Del Gaudio S, Galderisi U, Cascino A, Cipollaro M (2007) Comparative evaluation of different DNA extraction procedures from food samples. Biotechnol. Prog. 23, 297–301. https://doi.org/10.1021/bp060182m

Elsanhoty RM, Ramadan MF, Jany KD (2011) DNA extraction methods for detecting genetically modified foods: a comparative study. Food Chem. 126, 1883–1889. https://doi.org/10.1016/j.foodchem.2010.12.013

Grazina L, Amaral JS, Mafra I (2020) Botanical origin authentication of dietary supplements by DNA-based approaches. Comp. Rev. Food Sci. Food Safety 19, 1080–1109. https://doi.org/10.1111/1541-4337.12551

Grohmann L, Seiler C (2019) CHAPTER 21. Standardization of DNA-based Methods for Food Authenticity Testing, in: Burns, M., Foster, L., Walker, M. (Eds.), DNA Techniques to Verify Food Authenticity: Applications in Food Fraud, Food Chemistry, Function and Analysis. Royal Society of Chemistry, Cambridge, pp. 227–234. https://doi.org/10.1039/9781788016025-00227

Gryson N, Messens K, Dewettinck K (2004) Evaluation and optimisation of five different extraction methods for soy DNA in chocolate and biscuits. Extraction of DNA as a first step in GMO analysis. J. Sci. Food Agric. 84, 1357–1363. https://doi.org/10.1002/jsfa.1767

Gryson N (2010) Effect of food processing on plant DNA degradation and PCR-based GMO analysis: a review. Anal. Bioanal. Chem. 396, 2003–2022. https://doi.org/10.1007/s00216-009-3343-2

Han J, Pang X, Liao B, Yao H, Song J, Chen S (2016) An authenticity survey of herbal medicines from markets in China using DNA barcoding. Sci. Rep. 6, 18723. https://doi.org/10.1038/srep18723

Hao D, Xiao P, Liu L, Peng Y, He C (2015) [Essentials of pharmacophylogeny: knowledge pedigree, epistemology and paradigm shift]. Zhongguo Zhong Yao Za Zhi 40, 3335–3342.

Hirst B, Fernandez-Calvino L, Weiss T (2019) Chapter 24. Commercial DNA testing, in: Burns, M., Foster, L., Walker, M. (Eds.), DNA Techniques to Verify Food Authenticity: Applications in Food Fraud, Food Chemistry, Function and Analysis. Royal Society of Chemistry, Cambridge, pp. 264–282. https://doi.org/10.1039/9781788016025-00264

Hoorfar J, Cook N, Malorny B, Wagner M, De Medici D, Abdulmawjood A, Fach P (2004) Letter to the editor. Lett. Appl. Microbiol. 38, 79–80. https://doi.org/10.1046/j.1472-765X.2003.01456.x

Howard C, Lockie-Williams C, Slater A (2020) Applied barcoding: the practicalities of DNA testing for herbals. Plants 9. https://doi.org/10.3390/plants9091150

Ichim MC, de Boer HJ (2020) A review of authenticity and authentication of commercial ginseng herbal medicines and food supplements. Front. Pharmacol. 11, 612071. https://doi.org/10.3389/fphar.2020.612071

Ichim MC (2019) The DNA-based authentication of commercial herbal products reveals their globally widespread adulteration. Front. Pharmacol. 10, 1227. https://doi.org/10.3389/fphar.2019.01227

John ME (1992) An efficient method for isolation of RNA and DNA from plants containing polyphenolics. Nucleic Acids Res. 20, 2381. https://doi.org/10.1093/nar/20.9.2381

Juul S, Izquierdo F, Hurst A, Dai X, Wright A, Kulesha E, Pettett R, Turner DJ (2015) What's in my pot? Real-time species identification on the MinION. BioRxiv. https://doi.org/10.1101/030742

Llongueras JP, Nair S, Salas-Leiva D, Schwarzbach AE (2013) Comparing DNA extraction methods for analysis of botanical materials found in anti-diabetic supplements. Mol. Biotechnol. 53, 249–256. https://doi.org/10.1007/s12033-012-9520-0

Lockley AK, Bardsley RG (2000) DNA-based methods for food authentication. Trends Food Sci. Technol. 11, 67–77. https://doi.org/10.1016/S0924-2244(00)00049-2

Lo Y-T, Shaw P-C (2018a) DNA-based techniques for authentication of processed food and food supplements. Food Chem. 240, 767–774. https://doi.org/10.1016/j.foodchem.2017.08.022

Lo Y-T, Shaw P-C (2018b) DNA barcoding in concentrated Chinese medicine granules using adaptor ligation-mediated polymerase chain reaction. J. Pharm. Biomed. Anal. 149, 512–516. https://doi.org/10.1016/j.jpba.2017.11.048

Lo YT, Shaw PC (2019) Application of next-generation sequencing for the identification of herbal products. Biotechnol. Adv. 37, 107450. https://doi.org/10.1016/j.biotechadv.2019.107450

Matsuoka T, Kuribara H, Akiyama H, Miura H, Goda Y, Kusakabe Y, Isshiki K, Toyoda M, Hino A (2001) A multiplex PCR method of detecting recombinant DNAs from five lines of genetically modified maize. Shokuhin Eiseigaku Zasshi 42, 24–32. https://doi.org/10.3358/shokueishi.42.24

Mezzasalma V, Ganopoulos I, Galimberti A, Cornara L, Ferri E, Labra M (2017) Poisonous or non-poisonous plants? DNA-based tools and applications for accurate identification. Int. J. Legal Med. 131, 1–19. https://doi.org/10.1007/s00414-016-1460-y

Mishra P, Kumar A, Nagireddy A, Mani DN, Shukla AK, Tiwari R, Sundaresan V (2016) DNA barcoding: an efficient tool to overcome authentication challenges in the herbal market. Plant Biotechnol. J. 14, 8–21. https://doi.org/10.1111/pbi.12419

Murray MG, Thompson WF (1980) Rapid isolation of high molecular weight plant DNA. Nucleic Acids Res. 8, 4321–4325. https://doi.org/10.1093/nar/8.19.4321

Omelchenko DO, Speranskaya AS, Ayginin AA, Khafizov K, Krinitsina AA, Fedotova AV, Pozdyshev DV, Shtratnikova VY, Kupriyanova EV, Shipulin GA, Logacheva MD (2019) Improved Protocols of ITS1-Based Metabarcoding and Their Application in the Analysis of Plant-Containing Products. Genes (Basel) 10. https://doi.org/10.3390/genes10020122

Pafundo S, Gullì M, Marmiroli N (2011) Comparison of DNA extraction methods and development of duplex PCR and real-time PCR to detect tomato, carrot, and celery in food. J. Agric. Food Chem. 59, 10414–10424. https://doi.org/10.1021/jf202382s

Pandey RN, Adams RP, Flournoy LE (1996) Inhibition of random amplified polymorphic DNAs (RAPDs) by plant polysaccharides. Plant Mol. Biol. Rep. 14, 17–22. https://doi.org/10.1007/BF02671898

Parveen I, Gafner S, Techen N, Murch SJ, Khan IA (2016) DNA barcoding for the identification of botanicals in herbal medicine and dietary supplements: strengths and limitations. Planta Med. 82, 1225–1235. https://doi.org/10.1055/s-0042-111208

Pawar RS, Handy SM, Cheng R, Shyong N, Grundel E (2017) Assessment of the authenticity of herbal dietary supplements: comparison of chemical and DNA barcoding methods. Planta Med. 83, 921–936. https://doi.org/10.1055/s-0043-107881

Peterson DG, Boehm KS, Stack SM (1997) Isolation of milligram quantities of nuclear DNA from tomato (Lycopersicon esculentum), A plant containing high levels of polyphenolic compounds. Plant Mol. Biol. Rep. 15, 148–153. https://doi.org/10.1007/BF02812265

Pharmacopoeia Committee of P. R. China (2020) Pharmacopoeia of People's Republic of China. China Medical Science and Technology Press, Beijing.

Pich U, Schubert I (1993) Midiprep method for isolation of DNA from plants with a high content of polyphenolics. Nucleic Acids Res. 21, 3328. https://doi.org/10.1093/nar/21.14.3328

Pinto AD, Forte V, Guastadisegni MC, Martino C, Schena FP, Tantillo G (2007) A comparison of DNA extraction methods for food analysis. Food Control 18, 76–80. https://doi.org/10.1016/j.foodcont.2005.08.011

Puchooa D (2004) A simple, rapid and efficient method for the extraction of genomic DNA from lychee (Litchi chinensis Sonn.). Afr. J. Biotechnol. 3, 253–255. https://doi.org/10.5897/AJB2004.000-2046

Raclariu AC, Heinrich M, Ichim MC, de Boer H (2018) Benefits and limitations of DNA barcoding and metabarcoding in herbal product authentication. Phytochem. Anal. 29, 123–128. https://doi.org/10.1002/pca.2732

Raime K, Krjutškov K, Remm M (2020) Method for the Identification of Plant DNA in Food Using Alignment-Free Analysis of Sequencing Reads: A Case Study on Lupin. Front. Plant Sci. 11, 646. https://doi.org/10.3389/fpls.2020.00646

Ratnasingham S, Hebert PDN (2007) BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). Mol. Ecol. Notes 7, 355–364. https://doi.org/10.1111/j.1471-8286.2007.01678.x

Rogers SO, Bendich AJ (1985) Extraction of DNA from milligram amounts of fresh, herbarium and mummified plant tissues. Plant Mol. Biol. 5, 69–76. https://doi.org/10.1007/BF00020088

Seethapathy GS, Raclariu-Manolica A-C, Anmarkrud JA, Wangensteen H, de Boer HJ (2019) DNA metabarcoding authentication of ayurvedic herbal products on the European market raises concerns of quality and fidelity. Front. Plant Sci. 10, 68. https://doi.org/10.3389/fpls.2019.00068

Sharma AD, Gill PK, Singh P (2002) DNA isolation from dry and fresh samples of polysaccharide-rich plants. Plant Mol. Biol. Rep. 20, 415–415. https://doi.org/10.1007/BF02772129

Smith DS, Maxwell PW, De Boer SH (2005) Comparison of several methods for the extraction of DNA from potatoes and potato-derived products. J. Agric. Food Chem. 53, 9848–9859. https://doi.org/10.1021/jf051201v

Techen N, Parveen I, Pan Z, Khan IA (2014) DNA barcoding of medicinal plant material for identification. Curr. Opin. Biotechnol. 25, 103–110. https://doi.org/10.1016/j.copbio.2013.09.010

Turkec A, Kazan H, Karacanli B, Lucas SJ (2015) DNA extraction techniques compared for accurate detection of genetically modified organisms (GMOs) in maize food and feed products. J. Food Sci. Technol. 52, 5164–5171. https://doi.org/10.1007/s13197-014-1547-8

Unable to find information for 9645779, n.d.

Wilkes T (2019) CHAPTER 3. DNA Extraction from Food Matrices, in: Burns, M., Foster, L., Walker, M. (Eds.), DNA Techniques to Verify Food Authenticity: Applications in Food Fraud, Food Chemistry, Function and Analysis. Royal Society of Chemistry, Cambridge, pp. 29–49. https://doi.org/10.1039/9781788016025-00029

Wong T-H, But GW-C, Wu H-Y, Tsang SS-K, Lau DT-W, Shaw P-C (2018) Medicinal Materials DNA Barcode Database (MMDBD) version 1.5-one-stop solution for storage, BLAST, alignment and primer design. Database (Oxford) 2018. https://doi.org/10.1093/database/bay112

Wood EJ (2002) Principles and techniques of practical biochemistry (5th Ed.): Wilson, K., Walker, J. (eds.). Biochem. Mol. Biol. Educ. 30, 214–215. https://doi.org/10.1002/bmb.2002.494030030062

Answers

1. Sample source and processing, collection and storage, homogenisation, and the presence of contaminants.
2. One should firstly consider whether it is a complex mixture or a pure product, and the degree of processing (form and degree of homogeneity).
3. Absorbance methods, agarose gel electrophoresis, and fluorescent DNA-intercalating dyes.

Chapter 7
DNA from faeces

Physilia Chua[1,2], Christina Lynggaard[1], Kristine Bohmann[1]

1 Section for Evolutionary Genomics, Globe Institute, University of Copenhagen, Denmark
2 Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK

Physilia Chua **physiliachua@gmail.com**
Christina Lynggaard **christina.lynggaard@sund.ku.dk**
Kristine Bohmann **kbohmann@sund.ku.dk**

Citation: Chua P, Lynggaard C, Bohmann K (2022) Chapter 7. DNA from faeces. In: de Boer H, Rydmark MO, Verstraete B, Gravendeel B (Eds) Molecular identification of plants: from sequence to species. Advanced Books. https://doi.org/10.3897/ab.e98875

Background

What are faecal samples?

Do you know that faeces are windows to the natural world? Faeces, although not the most glamorous thing in the world, are worth their weight in gold when it comes to providing information about the host(s) they are derived from. Faeces, also commonly known as scat, poop, droppings, excreta, or stools are solid remains of the ingested food that were not digested in the intestine. They are composed of water, protein, polysaccharides, fats, solids (e.g., fibres from plants), and bacteria (Rose et al. 2015). From mites to elephants, faeces provide researchers with useful information about the animal and its environment (Tovey et al. 1981; Webber et al. 2018). Although fresh samples are usually used, information can also be retrieved from coprolites (fossilised faecal remains) even when they are 237 million years old (Qvarnström et al. 2019; van Geel et al. 2011; Welker et al. 2014).

Types of information that can be retrieved from faeces

Different types of information can be obtained from faeces. Chemical analyses provide information on hormonal changes that can occur from stress (Barja et al. 2008; Turner and Mathews 2010). Home-range and behaviour can be studied using the location of the faeces (Penteriani and Delgado 2008; Stewart et al. 2001). The appearance and size of faeces can even provide sex and species identification. For example, male capercaillies have larger dropping diameters than females (Thiel et al. 2007), and wombats are the only mammal with cube-shaped droppings (Yang et al. 2019). Molecular methods can be used for sex and species identification, identifying intestinal parasites, microbiome studies, and host genetics (A'Hara et al. 2009; Medeiros et al. 2012; Oliveira et al. 2020; Palomares et al. 2012; Soares et al. 2020). Additionally, studying faeces can give insights into animals' diet, providing information on the composition of plants ingested by herbivores and omnivores (Robeson et al. 2018; Valentini et al. 2009).

Non-molecular methods for analysing diet in faecal samples

Non-molecular methods have traditionally been used for the analysis of contents from faecal samples. An example is microhistology, where small amounts of faecal samples are mounted on a microscope slide, and digested remains of plant cuticle fragments are identified based on morphology (Baumgartner and Martin 1939). However, this method is extremely time-consuming and requires trained experts to be able to identify partial fragments of plants. Another disadvantage is that the abundance of easily digested plants are often underestimated using this technique (Shrestha and Wegge 2006). Near-infrared reflectance spectroscopy (NIRS) is another method used to determine the composition of plants in faecal samples (Norris et al. 1976), but this technique requires validation with reference samples of the diet and constant monitoring of equipment calibration (Dixon and Coates 2009). Stable isotope analysis and plant cuticular wax alkane measurements of faecal samples have also been carried out (Carnahan 2011; Mayes and Dove 2000). However, species-level resolution is not possible in stable isotope analysis (Mayes and Dove 2000), while alkane measurements require specialised equipment for extraction and detection that is often not available as standard laboratory services (Garnick et al. 2018). Additionally, cuticular wax alkane measurements are not suitable for assessing complex compositions (Garnick et al. 2018). These challenges coupled with ad-

vancements in high-throughput sequencing (HTS) techniques have resulted in a shift towards molecular methods for analysing faecal samples.

Applications of extracted DNA from faeces

In plant molecular applications, a common use of faecal samples is in herbivore/omnivore diet studies. The goal of most plant-focused diet studies is to characterise the diet profile of the host, which can be used to answer research questions concerning for example, resource competition and partitioning (Kartzinel et al. 2015; Lopes et al. 2015; Soininen et al. 2015), herbivore impact on local vegetation (Hibert et al. 2011), how livestock diets can be monitored (Lee et al. 2018; Pegard et al. 2009), temporal variability in diet compositions (Aziz et al. 2017), and dietary foraging plasticity (Kowalczyk et al. 2019; Quéméré et al. 2013). DNA extracted from faecal samples can also be used for other types of plant identification applications including palaeobotany (Chame 2003; Poinar et al. 1998) (see Chapter 21 Palaeobotany), faecal contamination of food in food safety (Jay-Russell 2013) (see Chapter 24 Food safety), environment and biodiversity assessments (Best 2008; Eycott et al. 2007; Green et al. 2018; Kartzinel et al. 2015) (see Chapter 25 Environment and biodiversity assessments), and forensics genetics (Norris and Bock 2000) (see Chapter 26 Forensics genetics, botany and palynology). Another potential application is the study of plant diseases such as parasitic fungi to assess the health of a particular ecosystem. Parasitic fungi that are found in plants can be ingested by herbivores/omnivores when the plants are eaten and derived fungal DNA is subsequently found in their faeces.

Advantages and limitations

The main advantage of using faecal samples for molecular plant identification as compared to other types of samples such as whole animals/insects (Staudacher et al. 2011) or gut contents (Junnila et al. 2010) is that it is non-invasive, and removes the need to capture or locate the animals for obtaining samples (Taberlet et al. 1999). In addition to being easily collected, faeces are constantly produced and therefore not considered rare (except for coprolites). Moreover, they are relatively easy to detect as they are normally the most persistent remnant from scarce or elusive animals (Hibert et al. 2013; Iwanowicz et al. 2016). Trained dogs can be used for the detection of faeces if required (Arandjelovic et al. 2015).

One limitation when using faecal samples for molecular plant identification is that it can be difficult to obtain fresh faecal samples collected immediately after defecation, especially when working with wild animals. Age of samples can have an impact on the amount and quality of DNA that can be extracted due to DNA degradation caused by exposure to environmental conditions (Taberlet et al. 1999). DNA degradation is particularly problematic when working with large DNA markers (>300 bp) as degradation results in short DNA fragments, which will not be amplified using large DNA markers (Frantzen et al. 1998; Taberlet et al. 1999). The availability of fresh faecal samples can also have an impact on the choice of downstream molecular techniques used for analysis (Chua et al. 2021). Another consideration is that if closely-related species have overlapping habitats, additional molecular work is needed to distinguish and identify the host of the droppings (A'Hara et al. 2009), which increases the budget and time required to process the samples. Finally, information obtained from faecal samples provides only a snapshot of the diet and can be influenced by individual preferences (Lopes et al. 2015), sex (Mata et al. 2016), or seasonal differences (Clare et al. 2014), therefore, more samples per individual or species are needed to obtain a full overview of the diet (Trites and Joy 2005) (Table 1).

Table 1. Advantages and limitations of using DNA from faeces to reconstruct plant communities.

Advantages	Limitations
Non-invasive	Fresh samples may be challenging to obtain from wild animals
Easy to detect and collect	Presence of PCR inhibitors
Not considered rare	DNA degradation
Does not require capturing or locating animal of interest	Hard to distinguish morphologically with closely related species • Additional molecular work needed • Increased cost and time

Experimental design

Sampling strategies

Before designing any sampling strategies for the collection of faecal samples, there are at least six factors that researchers must take into consideration:

1. The research question(s) and the required data to achieve the research objectives
2. The ecology of the species to be studied
3. The feasibility of sampling in the study area (is accessing the terrain a safety risk?)
4. The duration and spatial extent of the project (long term or short term? Does it span across different seasons?)
5. Budget constraints
6. Ethical considerations

Based on the research question(s) and objectives (i.e., quantitative, presence/absence, composition), researchers must decide how many samples and replicates are needed from each individual and/or population to sufficiently meet their research objectives. The choice of downstream molecular methods used for reconstructing herbivore/omnivore diet will also have an impact on how many samples are required. In quantitative studies where the objective is to quantify the ingested biomass, the number of different individuals sampled is not as important as in composition studies, where more individuals are required to obtain a better overview of the dietary range of the studied species. This is due to the effect of individual food preference, which can lead to biases in retrieving the whole range of a dietary profile for a given species if only a few individuals are studied (Watanabe 1984). In studies determining the presence/absence of a specific dietary component, it may be prudent to sample in larger numbers from both different individuals and populations to prevent false negatives caused by small sample size. While not always possible due to the ecology of the studied species, collecting replicates is also a good practice to evaluate any variation in the study and this aspect should be incorporated into sampling strategies (Mata et al. 2019). Other sampling variables such as seasonal effects (Ait Baamrane et al. 2012), age of faecal samples (McInnes et al. 2017), and differences in diet between sexes (Du Toit 2006), should also be taken into consideration, as these factors can affect the reconstructed diet (Chua et al. 2021).

Generally, the more ecological information gathered and incorporated into sampling strategies, the higher the chance of successful faecal collection. For wild species, prior ecological information regarding the species of interest is essential for designing sound sampling strategies, to optimise and streamline sample collection. Researchers can use the following questions as a guide in planning their sample collection strategy:

- Is the target species localised to a certain area?
- What is the extent of its daily range (does it differ between seasons)?
- Is it a generalist or a specialist?
- What is its foraging behaviour (does it differ between seasons)?
- Is the habitat easily accessible for sample collection?
- What is the density of the population in the study sites?
- Does its habitat overlap with closely-related species and will this lead to possible collection of faeces from non-target species?

Without this information, it is challenging to narrow down specific study sites for field collection. Additionally, such information can reduce the necessary man-power, resources, and time spent in the field while increasing the probability of finding sufficient numbers of faecal samples. Knowledge of habitat range and population density can prevent excessive amounts of samples collected from a single individual when the research question requires samples from multiple individuals. Differences in home-range and diet between seasons can also impact sample collection strategy (Rodrìguez and Obeso 2000). For example, a higher population density may be present during breeding seasons as compared to non-breeding seasons, which can impact the sampling strategies. For wide species whose faecal samples may be hard to find, sampling can be aided by the use of pointing dogs (Arandjelovic et al. 2015). For captive species, this information is not as important, as the study area is significantly reduced and faecal samples can be easily collected. Other considerations for both wild and captive animal faecal collection strategies include issues such as disturbances to the studied animals and risk to personal safety from aggressive animals.

Sampling strategies are also heavily dependent on budget constraints, which may reduce the time spent on sample collection, the number of samples processed, and also the molecular techniques used in analysing the faecal samples. Therefore, it is prudent to ensure that the budget fits the research objectives or that research objectives should be tailored to fit the research budget. While there are many different approaches to sampling, two commonly used approaches are systematic sampling and opportunistic sampling. In systematic sampling, the study area is divided into grids or transects, and samples are taken at each grid point or fixed intervals (Osborne 1942). While simple to carry out, it is not always feasible for faecal collection as animals do not defecate along a grid point/transect line. In contrast, in opportunistic sampling, researchers simply collect faecal samples in a study area when they come across it without being confined to grids or transects. While time-consuming, this method can result in the collection of more samples (De Barba et al. 2010). Depending on the ecology of the animal, sometimes more than one dropping is deposited at a single location. Researchers can choose to collect all droppings or only a few, depending on the research question.

Finally, ethical consideration of minimising distress to studied animals is one of the main concerns in animal studies and there are legal restrictions as implemented in the EU Directive 2010/63/EU on the protection of animals used for scientific purposes (Zemanova 2019). As faecal samples are collected non-invasively and often without the presence of the studied animal, researchers are less bound by this restriction as it does not pose any welfare harm to the animal. However, permits may be required for the collection of faecal samples from protected species, for entering protected habitats, and/or for transportation across borders. The possibility of receiving these permits should be checked at the beginning of any project, and be organised well in advance of the planned collection period. When all these factors have all been considered, sampling strategies can then be developed to cover the essential questions such as where, when, and how many samples to collect.

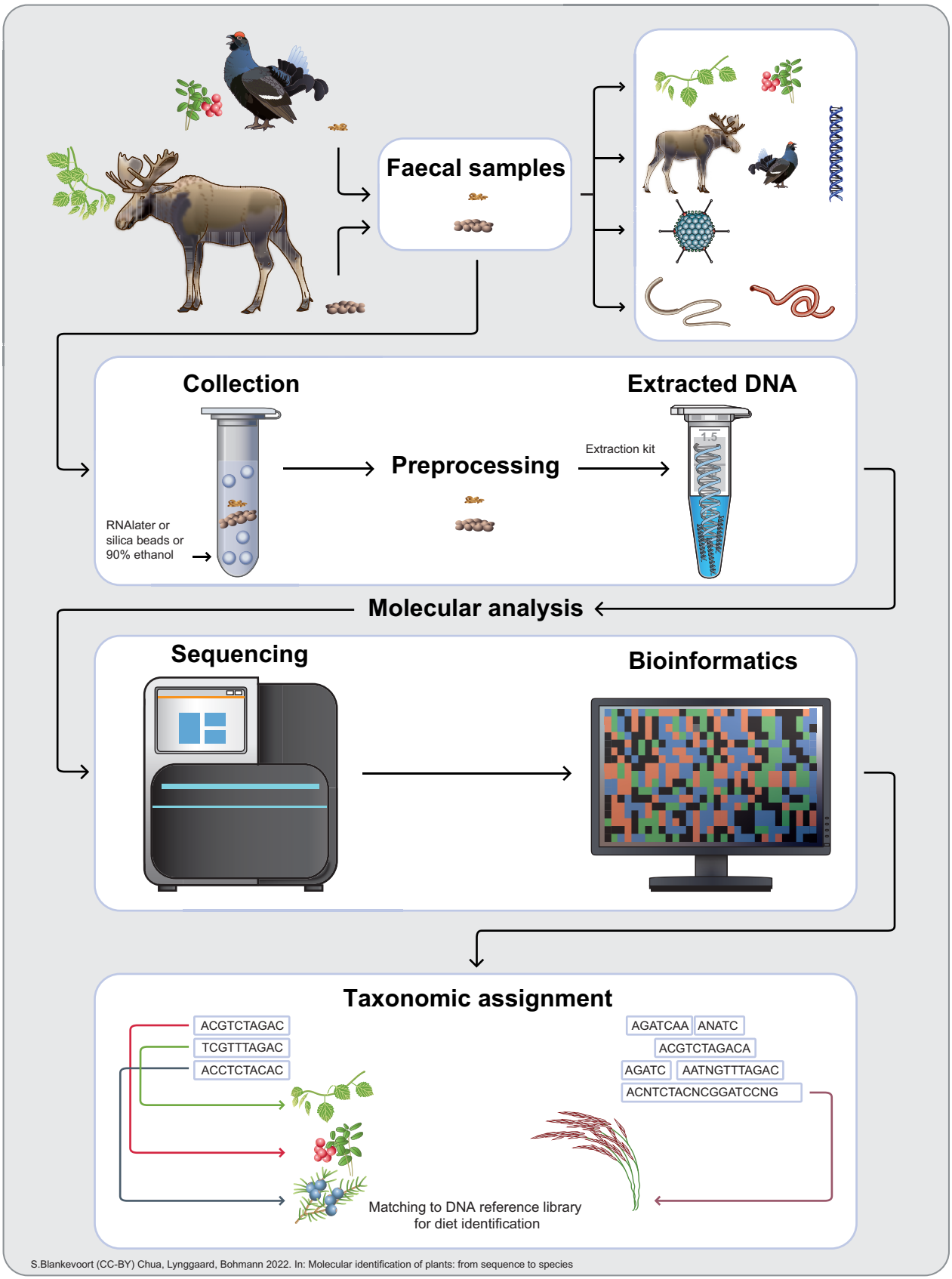

Figure 1. Chapter 7 Infographic: Visual representation of the content of this chapter.

Collection, transportation, and storage

Once the sampling strategy has been determined, the sampling in the field can start. The first step is to locate the faecal samples in the field. Once faecal samples have been located, collection can begin. When collecting faecal samples, there are a few materials that will be needed no matter what animal and habitat the faecal samples are derived from; sterile tubes filled with e.g. RNAlater™, silica beads or 90% ethanol, gloves, and a device to collect the samples. Sterile tubes will be necessary for sample storage. Tubes can have either removable screw-lids or hinged lids. Removable screw-lids have the advantage that the lids will not come off during transport. However, there is an increased risk of environmental contamination with these lids since they are separate from the tube and must be placed somewhere before collection. Tubes with hinged lids are easier to work with in that sense, though they can open during transport if not sealed (e.g., with parafilm™). Proper use of gloves and a collection device are also important to limit the risk of a collector becoming sick from directly handling faeces, as well as reducing the risk of sample contamination. The size and type of the sampling device can differ depending on the size of the faecal dropping and can range from a toothpick to a large spoon.

DNA-based diet analyses are very sensitive to contamination, and the trace amounts of digested plant material that can be extracted from faecal samples is easily contaminated. Contamination can occur between samples, by plant DNA from the surrounding environment, or even from the collector's (plant-based) lunch (Lusk 2014). Therefore, it is important to practice good sample collection hygiene. Although it is important to wear gloves while collecting samples, the gloves themselves can be a source of contamination. Care must be taken to not contaminate the gloves through touching other plants or plant-based items in the environment, or from handling another faecal sample. If they are contaminated, they must be changed or sterilised with bleach-solution (at least 5%). If the latter is performed, it is important to carry the bleach solution back to the lab site for proper disposal. Similar to the gloves, the devices for sample collection can also be a source of sample contamination. Thus, one-use disposable devices or those that are sterilised through the use of flame or a bleach solution should be used and properly disposed of (Champlot et al. 2010; Kemp and Smith 2005). Finally, collection needs to be done on surfaces with minimal contamination such as rocks or ice (McInnes et al. 2017), and avoid collecting samples on wet soil (Ando et al. 2018). To identify any potential sample contamination, it is necessary to include negative collection controls, which does not include any faecal sample. For these negative controls to be useful in downstream analyses, it is important to treat them identically to the 'real' samples. Thus, the negative controls should include the same storage buffers used, and be collected under the same conditions with the same collections devices and storage tubes that were used for the 'real' samples (Deiner et al. 2017; Zinger et al. 2019).

To avoid DNA degradation, faecal samples should be preserved as soon as possible upon collection and stored under the same conditions (Nsubuga et al. 2004). This can be achieved in a variety of ways, including freezing the samples with or without storage in e.g. RNAlater, 90% ethanol, or silica with or without prior ethanol addition (Alberdi et al. 2019; Nsubuga et al. 2004; Roeder et al. 2004).

DNA extraction

To avoid contamination, extractions should be carried out in a room free of PCR amplified DNA. Due to the risk of zoonotic disease transmission, extraction should ideally be carried out in a flow-hood to avoid inhaling dust from dry faeces (Lear et al. 2018). Before extracting DNA from faecal samples,

pre-processing steps are required. This entails removal of the outer faecal layers which have been in contact with the environment and thus exposed to environmental contaminants (Van Geel et al. 2014). Outer layers are enriched for host epithelial DNA (Creamer et al. 1961), which reduces the proportion of starting plant DNA material, so it would be prudent to remove it. Depending on the research question, pooling of faecal samples collected from the same or different individuals may be necessary. If this is required, samples should be well-mixed. Reduction of faecal sample volume through sub-sampling of faecal dropping may also be necessary for DNA extraction.

Faecal samples from plant-eating animals usually contain high levels of PCR inhibitors such as humic acid, which can lead to amplification failure during downstream analysis (Ramón-Laca et al. 2015). Minimising the carryover of PCR inhibitors is thus one of the key considerations in the extraction process, particularly when using metabarcoding (see Chapter 11 Amplicon metabarcoding). Several commercial kits have been developed to deal with the removal of inhibitors (Johnson et al. 2005), and some commonly used kits for extracting plant DNA from faecal samples are i.e. QIAGEN DNeasy Blood and Tissue Kit, and QIAGEN DNeasy PowerFecal kit. However, some kits such as the QIAGEN stool kit can contain plant contaminants such as potato, so it is recommended to avoid using such kits when identifying plant DNA extracted from faecal samples (Valentini et al. 2009). Similar to the sample collection controls, it is also important to include extraction controls for each extraction day, so that any possible contaminants can be identified (Zinger et al. 2019). Additionally, the use of extraction replicates (two or more DNA extraction from the same sample) allows for a better overview of the plant communities present within one sample as compared to not having any replicates (Hernandez-Rodriguez et al. 2018).

Molecular methods for faecal analysis

Depending on the research question(s), several different HTS methods can be used for analysing DNA extracted from faecal samples including metabarcoding (Valentini et al. 2009) (see Chapter 11 Amplicon metabarcoding), metagenomics (Srivathsan et al. 2016, 2015) (see Chapter 12 Metagenomics), and target capture (Perry 2014) (see Chapter 14 Target capture). The advantages and limitations of each method can be found in the aforementioned chapters. Another less commonly used non-HTS molecular approach is based on PCR amplification with selected primers coupled to electrophoresis, called PCR capillary electrophoresis (PCR-CE) (Czernik et al. 2013; Pegard et al. 2009). One advantage of this approach compared to HTS methods is that it is faster and cheaper and when complementary genes are targeted, high species resolution can be achieved. However, this approach is sensitive to contamination from extraction kits (Valentini et al. 2009), such as potato DNA which has similar peak sizes to some plant species, making it challenging for accurate species identification. It is also only useful when fresh faecal samples are available (Czernik et al. 2013), which is not always possible with fieldwork.

Questions

1. Name one sampling limitation of working with faecal samples as compared to other types of samples (e.g., gut contents) and give suggestions on how to overcome this limitation.
2. How does prior information of studied species ecology aid in sampling design?
3. Contamination can occur during sample collection, sample preprocessing, and DNA extraction. Describe the main type of contamination during each phase and how it can be prevented.

Glossary

Coprolites – Fossilised faeces.
Near-infrared spectroscopy (NIRS) – A non-destructive and fast technique utilising the near-infrared region of the electromagnetic spectrum.
RNAlater – Non-toxic aqueous reagent for storage purposes, preserving RNA and DNA.
Stable isotopes – Non-radioactive elements.
Zoonotic disease – Infectious disease caused by pathogens jumping from non-human hosts to humans.

References

A'Hara SW, Hancock M, Piertney SB, Cottrell JE (2009) The development of a molecular assay to distinguish droppings of black grouse *Tetrao tetrix* from those of capercaillie *Tetrao urogallus* and red grouse *Lagopus Lagopus Scoticus*. Wildlife Biol. 15, 328–337. https://doi.org/10.2981/08-046

Ait Baamrane MA, Shehzad W, Ouhammou A, Abbad A, Naimi M, Coissac E, Taberlet P, Znari M (2012) Assessment of the food habits of the Moroccan dorcas gazelle in M'Sabih Talaa, west central Morocco, using the trnL approach. PLoS ONE 7, e35643. https://doi.org/10.1371/journal.pone.0035643

Alberdi A, Aizpurua O, Bohmann K, Gopalakrishnan S, Lynggaard C, Nielsen M, Gilbert MTP (2019) Promises and pitfalls of using high-throughput sequencing for diet analysis. Mol. Ecol. Resour. 19, 327–348. https://doi.org/10.1111/1755-0998.12960

Ando H, Fujii C, Kawanabe M, Ao Y, Inoue T, Takenaka A (2018) Evaluation of plant contamination in metabarcoding diet analysis of a herbivore. Sci. Rep. 8, 15563. https://doi.org/10.1038/s41598-018-32845-w

Arandjelovic M, Bergl RA, Ikfuingei R, Jameson C, Parker M, Vigilant L (2015) Detection dog efficacy for collecting faecal samples from the critically endangered Cross River gorilla (Gorilla gorilla diehli) for genetic censusing. R. Soc. Open Sci. 2, 140423. https://doi.org/10.1098/rsos.140423

Aziz SA, Clements GR, Peng LY, Campos-Arceiz A, McConkey KR, Forget P-M, Gan HM (2017) Elucidating the diet of the island flying fox (Pteropus hypomelanus) in Peninsular Malaysia through Illumina Next-Generation Sequencing. PeerJ 5, e3176. https://doi.org/10.7717/peerj.3176

Barja I, Silván G, Illera JC (2008) Relationships between sex and stress hormone levels in feces and marking behavior in a wild population of Iberian wolves (Canis lupus signatus). J. Chem. Ecol. 34, 697–701. https://doi.org/10.1007/s10886-008-9460-0

Baumgartner LL, Martin AC (1939) Plant Histology as an Aid in Squirrel Food-Habit Studies. The Journal of Wildlife Management 3, 266. https://doi.org/10.2307/3796113

Best RJ (2008) Exotic grasses and feces deposition by an exotic herbivore combine to reduce the relative abundance of native forbs. Oecologia 158, 319–327. https://doi.org/10.1007/s00442-008-1137-4

Carnahan AM (2011) Determining the diets of moose (Alces alces) in Alaska using plant wax components (Undergraduate thesis). University of Alaska Anchorage.

Chame M (2003) Terrestrial mammal feces: a morphometric summary and description. Mem. Inst. Oswaldo Cruz 98 Suppl 1, 71–94. https://doi.org/10.1590/s0074-02762003000900014

Champlot S, Berthelot C, Pruvost M, Bennett EA, Grange T, Geigl E-M (2010) An efficient multistrategy DNA decontamination procedure of PCR reagents for hypersensitive PCR applications. PLoS ONE 5. https://doi.org/10.1371/journal.pone.0013042

Chua PYS, Crampton-Platt A, Lammers Y, Alsos IG, Boessenkool S, Bohmann K (2021) Metagenomics: a viable tool for reconstructing herbivore diet. Mol. Ecol. Resour. 21, 2249–2263. https://doi.org/10.1111/1755-0998.13425

Chua PYS, Lammers YYS, Menoni E, Ekrem T, Bohmann K, Boessenkool S, Alsos IG (2021) Molecular dietary analyses of western capercaillies (*Tetrao urogallus*) reveal a diverse diet. Environ. DNA 3, 1156–1171. https://doi.org/10.1002/edn3.237

Clare EL, Symondson WOC, Broders H, Fabianek F, Fraser EE, MacKenzie A, Boughen A, Hamilton R, Willis CKR, Martinez-Nuñez F, Menzies AK, Norquay KJO, Brigham M, Poissant J, Rintoul J, Barclay RMR, Reimer JP (2014) The diet of Myotis lucifugus across Canada: assessing foraging quality and diet variability. Mol. Ecol. 23, 3618–3632. https://doi.org/10.1111/mec.12542

Creamer B, Shorter RG, Bamforth J (1961) The turnover and shedding of epithelial cells: Part I The turnover in the gastro-intestinal tract. Gut 2, 110–116. https://doi.org/10.1136/gut.2.2.110

Czernik M, Taberlet P, Swisłocka M, Czajkowska M, Duda N, Ratkiewicz M (2013) Fast and efficient DNA-based method for winter diet analysis from stools of three cervids: moose, red deer, and roe deer. Acta Theriol. 58, 379–386. https://doi.org/10.1007/s13364-013-0146-9

De Barba M, Waits LP, Genovesi P, Randi E, Chirichella R, Cetto E (2010) Comparing opportunistic and systematic sampling methods for non-invasive genetic monitoring of a small translocated brown bear population. Journal of Applied Ecology 47, 172–181. https://doi.org/10.1111/j.1365-2664.2009.01752.x

Deiner K, Bik HM, Mächler E, Seymour M, Lacoursière-Roussel A, Altermatt F, Creer S, Bista I, Lodge DM, de Vere N, Pfrender ME, Bernatchez L (2017) Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. Mol. Ecol. 26, 5872–5895. https://doi.org/10.1111/mec.14350

Dixon R, Coates D (2009) Review: near infrared spectroscopy of faeces to evaluate the nutrition and physiology of herbivores. J. Near Infrared Spectrosc. 17, 1–31. https://doi.org/10.1255/jnirs.822

Du Toit JT (2006) Sex differences in the foraging ecology of large mammalian herbivores, in: Ruckstuhl, K., Neuhaus, P. (Eds.), Sexual Segregation in Vertebrates: Ecology of the Two Sexes. Cambridge University Press, Cambridge, pp. 35–52. https://doi.org/10.1017/CBO9780511525629.004

Eycott AE, Watkinson AR, Hemami MR, Dolman PM (2007) The dispersal of vascular plants in a forest mosaic by a guild of mammalian herbivores. Oecologia 154, 107–118. https://doi.org/10.1007/s00442-007-0812-1

Frantzen MA, Silk JB, Ferguson JW, Wayne RK, Kohn MH (1998) Empirical evaluation of preservation methods for faecal DNA. Mol. Ecol. 7, 1423–1428. https://doi.org/10.1046/j.1365-294x.1998.00449.x

Garnick S, Barboza PS, Walker JW (2018) Assessment of animal-based methods used for estimating and monitoring rangeland herbivore diet composition. Rangeland Ecology & Management 71, 449–457. https://doi.org/10.1016/j.rama.2018.03.003

Green AJ, Lovas-Kiss Á, Stroud RA, Tierney N, Fox AD (2018) Plant dispersal by Canada geese in Arctic Greenland. Polar Res. 37, 1508268. https://doi.org/10.1080/17518369.2018.1508268

Hernandez-Rodriguez J, Arandjelovic M, Lester J, de Filippo C, Weihmann A, Meyer M, Angedakin S, Casals F, Navarro A, Vigilant L, Kühl HS, Langergraber K, Boesch C, Hughes D, Marques-Bonet T (2018) The impact of endogenous content, replicates and pooling on genome capture from faecal samples. Mol. Ecol. Resour. 18, 319–333. https://doi.org/10.1111/1755-0998.12728

Hibert F, Sabatier D, Andrivot J, Scotti-Saintagne C, Gonzalez S, Prévost M-F, Grenand P, Chave J, Caron H, Richard-Hansen C (2011) Botany, genetics and ethnobotany: a crossed investigation on the elusive tapir's diet in French Guiana. PLoS ONE 6, e25850. https://doi.org/10.1371/journal.pone.0025850

Hibert F, Taberlet P, Chave J, Scotti-Saintagne C, Sabatier D, Richard-Hansen C (2013) Unveiling the diet of elusive rainforest herbivores in next generation sequencing era? The tapir as a case study. PLoS ONE 8, e60799. https://doi.org/10.1371/journal.pone.0060799

Iwanowicz DD, Vandergast AG, Cornman RS, Adams CR, Kohn JR, Fisher RN, Brehme CS (2016) Metabarcoding of fecal samples to determine herbivore diets: a case study of the endangered pacific pocket mouse. PLoS ONE 11, e0165366. https://doi.org/10.1371/journal.pone.0165366

Jay-Russell M (2013) What is the risk from wild animals in food-borne pathogen contamination of plants? CAB Reviews 8. https://doi.org/10.1079/PAVSNNR20138040

Johnson DJ, Martin LR, Roberts KA (2005) STR-typing of human DNA from human fecal matter using the QIAGEN QIAamp stool mini kit. J. Forensic Sci. 50, 802–808.

Junnila A, Müller GC, Schlein Y (2010) Species identification of plant tissues from the gut of An. sergentii by DNA analysis. Acta Trop. 115, 227–233. https://doi.org/10.1016/j.actatropica.2010.04.002

Kartzinel TR, Chen PA, Coverdale TC, Erickson DL, Kress WJ, Kuzmina ML, Rubenstein DI, Wang W, Pringle RM (2015) DNA metabarcoding illuminates dietary niche partitioning by African large herbivores. Proc Natl Acad Sci USA 112, 8019–8024. https://doi.org/10.1073/pnas.1503283112

Kemp BM, Smith DG (2005) Use of bleach to eliminate contaminating DNA from the surface of bones and teeth. Forensic Sci. Int. 154, 53–61. https://doi.org/10.1016/j.forsciint.2004.11.017

Kowalczyk R, Wójcik JM, Taberlet P, Kamiński T, Miquel C, Valentini A, Craine JM, Coissac E (2019) Foraging plasticity allows a large herbivore to persist in a sheltering forest habitat: DNA metabarcoding diet analysis of the European bison. Forest Ecology and Management 449, 117474. https://doi.org/10.1016/j.foreco.2019.117474

Lear G, Dickie I, Banks J, Boyer S, Buckley H, Buckley T, Cruickshank R, Dopheide A, Handley K, Hermans S, Kamke J, Lee C, MacDiarmid R, Morales S, Orlovich D, Smissen R, Wood J, Holdaway R (2018) Methods for the extraction, storage, amplification and sequencing of DNA from environmental samples. N. Z. J. Ecol. https://doi.org/10.20417/nzjecol.42.9

Lee T, Alemseged Y, Mitchell A (2018) Dropping Hints: estimating the diets of livestock in rangelands using DNA metabarcoding of faeces. MBMG 2, e22467. https://doi.org/10.3897/mbmg.2.22467

Lopes CM, De Barba M, Boyer F, Mercier C, da Silva Filho PJS, Heidtmann LM, Galiano D, Kubiak BB, Langone P, Garcias FM, Gielly L, Coissac E, de Freitas TRO, Taberlet P (2015) DNA metabarcoding diet analysis for species with parapatric vs sympatric distribution: a case study on subterranean rodents. Heredity 114, 525–536. https://doi.org/10.1038/hdy.2014.109

Lusk RW (2014) Diverse and widespread contamination evident in the unmapped depths of high throughput sequencing data. PLoS ONE 9, e110808. https://doi.org/10.1371/journal.pone.0110808

Mata VA, Amorim F, Corley MFV, McCracken GF, Rebelo H, Beja P (2016) Female dietary bias towards large migratory moths in the European free-tailed bat (Tadarida teniotis). Biol. Lett. 12, 20150988. https://doi.org/10.1098/rsbl.2015.0988

Mata VA, Rebelo H, Amorim F, McCracken GF, Jarman S, Beja P (2019) How much is enough? Effects of technical and biological replication on metabarcoding dietary analysis. Mol. Ecol. 28, 165–175. https://doi.org/10.1111/mec.14779

Mayes RW, Dove H (2000) Measurement of dietary nutrient intake in free-ranging mammalian herbivores. Nutr. Res. Rev. 13, 107–138. https://doi.org/10.1079/095442200108729025

McInnes JC, Alderman R, Deagle BE, Lea M-A, Raymond B, Jarman SN (2017) Optimised scat collection protocols for dietary DNA metabarcoding in vertebrates. Methods Ecol. Evol. 8, 192–202. https://doi.org/10.1111/2041-210X.12677

Medeiros RJ, King RA, Symondson WOC, Cadiou B, Zonfrillo B, Bolton M, Morton R, Howell S, Clinton A, Felgueiras M, Thomas RJ (2012) Molecular evidence for gender differences in the migratory behaviour of a small seabird. PLoS ONE 7, e46330. https://doi.org/10.1371/journal.pone.0046330

Norris DO, Bock JH (2000) Use of Fecal Material to Associate a Suspect with a Crime Scene: Report of Two Cases. J. Forensic Sci. 45, 14657J. https://doi.org/10.1520/JFS14657J

Norris KH, Barnes RF, Moore JE, Shenk JS (1976) Predicting forage quality by infrared replectance spectroscopy. J. Anim. Sci. 43, 889–897. https://doi.org/10.2527/jas1976.434889x

Nsubuga AM, Robbins MM, Roeder AD, Morin PA, Boesch C, Vigilant L (2004) Factors affecting the amount of genomic DNA extracted from ape faeces and the identification of an improved sample storage method. Mol. Ecol. 13, 2089–2094. https://doi.org/10.1111/j.1365-294X.2004.02207.x

Oliveira BCM, Murray M, Tseng F, Widmer G (2020) The fecal microbiota of wild and captive raptors. Anim Microbiome 2, 15. https://doi.org/10.1186/s42523-020-00035-7

Osborne JG (1942) Sampling Errors of Systematic and Random Surveys of Cover-Type Areas. J. Am. Stat. Assoc. 37, 256–264. https://doi.org/10.1080/01621459.1942.10500634

Palomares F, Roques S, Chávez C, Silveira L, Keller C, Sollmann R, do Prado DM, Torres PC, Adrados B, Godoy JA, de Almeida Jácomo AT, Tôrres NM, Furtado MM, López-Bao JV (2012) High proportion of male faeces in jaguar populations. PLoS ONE 7, e52923. https://doi.org/10.1371/journal.pone.0052923

Pegard A, Miquel C, Valentini A, Coissac E, Bouvier F, François D, Taberlet P, Engel E, Pompanon F (2009) Universal DNA-based methods for assessing the diet of grazing livestock and wildlife from feces. J. Agric. Food Chem. 57, 5700–5706. https://doi.org/10.1021/jf803680c

Penteriani V, Delgado M del M (2008) Owls may use faeces and prey feathers to signal current reproduction. PLoS ONE 3, e3014. https://doi.org/10.1371/journal.pone.0003014

Perry GH (2014) The Promise and Practicality of Population Genomics Research with Endangered Species. Int. J. Primatol. 35, 55–70. https://doi.org/10.1007/s10764-013-9702-z

Poinar HN, Hofreiter M, Spaulding WG, Martin PS, Stankiewicz BA, Bland H, Evershed RP, Possnert G, Pääbo S (1998) Molecular coproscopy: dung and diet of the extinct ground sloth Nothrotheriops shastensis. Science 281, 402–406. https://doi.org/10.1126/science.281.5375.402

Quéméré E, Hibert F, Miquel C, Lhuillier E, Rasolondraibe E, Champeau J, Rabarivola C, Nusbaumer L, Chatelain C, Gautier L, Ranirison P, Crouau-Roy B, Taberlet P, Chikhi L (2013) A DNA metabarcoding study of a primate dietary diversity and plasticity across its entire fragmented range. PLoS ONE 8, e58971. https://doi.org/10.1371/journal.pone.0058971

Qvarnström M, Wernström JV, Piechowski R, Tałanda M, Ahlberg PE, Niedźwiedzki G (2019) Beetle-bearing coprolites possibly reveal the diet of a Late Triassic dinosauriform. R. Soc. Open Sci. 6, 181042. https://doi.org/10.1098/rsos.181042

Ramón-Laca A, Soriano L, Gleeson D, Godoy JA (2015) A simple and effective method for obtaining mammal DNA from faeces. Wildlife Biol. 21, 195–203. https://doi.org/10.2981/wlb.00096

Robeson MS, Khanipov K, Golovko G, Wisely SM, White MD, Bodenchuck M, Smyser TJ, Fofanov Y, Fierer N, Piaggio AJ (2018) Assessing the utility of metabarcoding for diet analyses of the omnivorous wild pig (*Sus scrofa*). Ecol. Evol. 8, 185–196. https://doi.org/10.1002/ece3.3638

Rodrìguez AE, Obeso JR (2000) Diet of the Cantabrian Capercaillie: geographic variation and energetic content. Ardeola 47, 77–83.

Roeder AD, Archer FI, Poinar HN, Morin PA (2004) A novel method for collection and preservation of faeces for genetic studies. Mol. Ecol. Notes 4, 761–764. https://doi.org/10.1111/j.1471-8286.2004.00737.x

Rose C, Parker A, Jefferson B, Cartmell E (2015) The characterization of feces and urine: A review of the literature to inform advanced treatment technology. Crit. Rev. Environ. Sci. Technol. 45, 1827–1879. https://doi.org/10.1080/10643389.2014.1000761

Shrestha R, Wegge P (2006) Determining the composition of herbivore diets in the trans-Himalayan rangelands: a comparison of field methods. Rangeland Ecology & Management 59, 512–518. https://doi.org/10.2111/06-022R2.1

Soares FA, Benitez A do N, Santos BMD, Loiola SHN, Rosa SL, Nagata WB, Inácio SV, Suzuki CTN, Bresciani KDS, Falcão AX, Gomes JF (2020) A historical review of the techniques of recovery of parasites for their detection in human stools. Rev. Soc. Bras. Med. Trop. 53, e20190535. https://doi.org/10.1590/0037-8682-0535-2019

Soininen EM, Gauthier G, Bilodeau F, Berteaux D, Gielly L, Taberlet P, Gussarova G, Bellemain E, Hassel K, Stenøien HK, Epp L, Schrøder-Nielsen A, Brochmann C, Yoccoz NG (2015) Highly overlapping winter diet in two sympatric lemming species revealed by DNA metabarcoding. PLoS ONE 10, e0115335. https://doi.org/10.1371/journal.pone.0115335

Srivathsan A, Ang A, Vogler AP, Meier R (2016) Fecal metagenomics for the simultaneous assessment of diet, parasites, and population genetics of an understudied primate. Front. Zool. 13, 17. https://doi.org/10.1186/s12983-016-0150-4

Srivathsan A, Sha JCM, Vogler AP, Meier R (2015) Comparing the effectiveness of metagenomics and metabarcoding for diet analysis of a leaf-feeding monkey (*Pygathrix nemaeus*). Mol. Ecol. Resour. 15, 250–261. https://doi.org/10.1111/1755-0998.12302

Staudacher K, Wallinger C, Schallhart N, Traugott M (2011) Detecting ingested plant DNA in soil-living insect larvae. Soil Biol. Biochem. 43, 346–350. https://doi.org/10.1016/j.soilbio.2010.10.022

Stewart PD, Macdonald DW, Newman C, Cheeseman CL (2001) Boundary faeces and matched advertisement in the European badger (*Meles meles*): a potential role in range exclusion. J. Zool. 255, 191–198. https://doi.org/10.1017/S0952836901001261

Taberlet P, Waits LP, Luikart G (1999) Noninvasive genetic sampling: look before you leap. Trends Ecol. Evol. 14, 323–327. https://doi.org/10.1016/s0169-5347(99)01637-7

Thiel D, Jenni-Eiermann S, Braunisch V, Palme R, Jenni L (2007) Ski tourism affects habitat use and evokes a physiological stress response in capercaillie *Tetrao urogallus*: a new methodological approach. Journal of Applied Ecology 45, 845–853. https://doi.org/10.1111/j.1365-2664.2008.01465.x

Tovey ER, Chapman MD, Platts-Mills TA (1981) Mite faeces are a major source of house dust allergens. Nature 289, 592–593. https://doi.org/10.1038/289592a0

Trites AW, Joy R (2005) Dietary analysis from fecal samples: how many scats are enough? J. Mammal. 86, 704–712. https://doi.org/10.1644/1545-1542(2005)086[0704:DAFFSH]2.0.CO;2

Turner DH, Mathews DH (2010) NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. Nucleic Acids Res. 38, D280-2. https://doi.org/10.1093/nar/gkp892

Valentini A, Miquel C, Nawaz MA, Bellemain E, Coissac E, Pompanon F, Gielly L, Cruaud C, Nascetti G, Wincker P, Swenson JE, Taberlet P (2009) New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the *trnL* approach. Mol. Ecol. Resour. 9, 51–60. https://doi.org/10.1111/j.1755-0998.2008.02352.x

van Geel B, Guthrie RD, Altmann JG, Broekens P, Bull ID, Gill FL, Jansen B, Nieman AM, Gravendeel B (2011) Mycological evidence of coprophagy from the feces of an Alaskan Late Glacial mammoth. Quat. Sci. Rev. 30, 2289–2303. https://doi.org/10.1016/j.quascirev.2010.03.008

Van Geel BAS, Protopopov A, Bull IAN, Duijm E, Gill F, Lammers Y, Nieman A, Rudaya N, Trofimova S, Tikhonov AN, Vos R, Zhilich S, Gravendeel B (2014) Multiproxy diet analysis of the last meal of an early Holocene Yakutian bison. J. Quaternary Sci. 29, 261–268. https://doi.org/10.1002/jqs.2698

Watanabe JM (1984) Food preference, food quality and diets of three herbivorous gastropods (Trochidae: *Tegula*) in a temperate kelp forest habitat. Oecologia 62, 47–52. https://doi.org/10.1007/BF00377371

Webber, JT, Henley, MD, Pretorius, Y, Somers, MJ, Ganswindt, A (2018) Changes in African elephant (*Loxodonta africana*) faecal steroid concentrations post-defaecation. Bothalia 48. https://doi.org/10.4102/abc.v48i2.2312

Welker, F, Duijm, E, van der Gaag, KJ, van Geel, B, de Knijff, P, van Leeuwen, J, Mol, D, van der Plicht, J, Raes, N, Reumer, J, Gravendeel, B (2014) Analysis of coprolites from the extinct mountain goat *Myotragus balearicus*. Quaternary Research 81, 106–116. https://doi.org/10.1016/j.yqres.2013.10.006

Yang, P, Lee, A, Chan, M, Martin, A, Edwards, A, Carver, S, Hu, D (2019) How, and why, do wombats make cube-shaped poo?

Zemanova, MA (2019) Poor implementation of non-invasive sampling in wildlife genetics studies. ReEco 4, 119–132. https://doi.org/10.3897/rethinkingecology.4.32751

Zinger, L, Bonin, A, Alsos, IG, Bálint, M, Bik, H, Boyer, F, Chariton, AA, Creer, S, Coissac, E, Deagle, BE, De Barba, M, Dickie, IA, Dumbrell, AJ, Ficetola, GF, Fierer, N, Fumagalli, L, Gilbert, MTP, Jarman, S, Jumpponen, A, Kauserud, H, Taberlet, P (2019) DNA metabarcoding-Need for robust experimental designs to draw sound ecological conclusions. Mol. Ecol. 28, 1857–1862. https://doi.org/10.1111/mec.15060

Answers

1. Possible challenges and solutions: It is difficult to obtain fresh faecal samples → one can use pointing dogs; Problem of using relatively long DNA barcoding fragments → use primers that can amplify shorter regions; Overlapping habitats of closely related species → use additional molecular markers to identify species (though this increases the cost and time necessary); Faecal samples only provide a snapshot of the entire diet → take multiple samples from the same individual and/or sample a larger number of individuals over a longer period and larger geographical area.

2. It helps to narrow down the study areas for field collection. This reduces the manpower, resources, and time needed, increasing the chance of finding samples. When applying for permits, you can point out how to keep the disturbance of animals in the field to a minimum with this knowledge, which will increase the chances of obtaining permission.
3. During sample collection → wear gloves, ensure that samples are not collected from wet soil, practice good collection hygiene; During sample preprocessing → remove outer layers that were in close contact with the environment, work in flow-hood and a PCR-free lab; During DNA extraction → include extraction controls, avoid using extraction kits with plant-based or other types of contaminants.

Chapter 8
aDNA from sediments

Anneke T.M. ter Schure[1], Yi Wang[2], Anna L.J. Chagas[2], Laura S. Epp[2]

1 Centre for Ecological and Evolutionary Synthesis, University of Oslo, Oslo, Norway
2 University of Konstanz, Konstanz, Germany

Anneke T.M. ter Schure **anneke.terschure@gmail.com**
Yi Wang **yiwang@uni-konstanz.de**
Anna Luiza Jaime Chagas **anna.jaime-chagas@uni-konstanz.de**
Laura S. Epp **laura.epp@uni-konstanz.de**

Citation: Ter Schure ATM, Wang Y, Chagas ALJ, Epp LS (2022) Chapter 8. aDNA from sediments. In: de Boer H, Rydmark MO, Verstraete B, Gravendeel B (Eds) Molecular identification of plants: from sequence to species. Advanced Books. https://doi.org/10.3897/ab.e98875

Background

Sedimentary ancient DNA studies aim to reconstruct the biology and ecology of past environments using the DNA present in the sediment record. Compared to modern soil and sedimentary DNA (see Chapter 4 DNA from soil), these analyses can be more challenging due to the prolonged exposure of the DNA to degradation processes. This has major implications for the scope of the study and the appropriate study design, which will be discussed in this chapter.

What is sedimentary ancient DNA?

In order to use sedimentary ancient DNA for paleoecological studies (*sed*aDNA; Haile et al. 2009) it is important to understand some aspects of its physical nature and the local environment's role in transforming modern DNA into *sed*aDNA. We will start by breaking down the term into its components.

Ancient DNA is the hereditary genetic content of cells from organisms that died a long time ago. There is no consensus on how old DNA should be in order to be called ancient, as the age is generally less important than the exposure to degradation processes that make it more degraded than modern DNA. *Sed*aDNA degradation processes are primarily related to environmental and sedimentary properties, such as temperature, pH, water content, oxygen levels, and minerals present in the sediment (Giguet-Covex et al. 2019; Torti et al. 2015), whereas time plays a secondary role: providing opportunity for these processes to take place. Permafrost in general provides excellent conditions for preserving DNA, due to its neutral pH, anaerobic conditions, and near-constant subzero temperatures that ensure it remains constantly frozen for 2 years or longer. Optimal conditions in ice cores from Greenland have allowed the detection of plant DNA as old as 450 to 800 thousand years (Willerslev et al. 2007). To date, the oldest amplifiable DNA from sediments is from ca. 400 thousand years old permafrost (Willerslev et al. 2004, 2003).

How does DNA end up in the sediment? Sediment is a result of erosion, weathering and biological processes and consists of organic and inorganic particles (e.g., sand and silt) that are transported by wind, water, or people (Masselink et al. 2014). These transportation processes also explain the main distinctive quality between sediments and soils: soils develop precisely because of the absence of horizontal transport, allowing biological, physical, and chemical weathering of the local substrate, thereby forming soil horizons rich in organic matter (see Chapter 4 DNA from soil). Deposition of sediment happens when the sediments stop being transported and stay in place. The incorporation of organismal remains into the sediment are similarly a result of transportation by wind, water, or people, or a result of organisms living at that location (Alsos et al. 2018; Parducci et al. 2018). The processes involved in the transfer, deposition, and preservation of organismic remains are called taphonomic processes. Bacterial and fungal DNA make up a very large part of sedimentary DNA, since they are natural inhabitants of sediments and outrate macroorganisms in terms of total biomass. Animal DNA that is found in sediment typically comes from skin flakes, faeces, urine, saliva, hair, feathers, and other animal tissues, while plant DNA typically originates from plant debris, leaves, seeds, fruits, and other plant tissues. Living cells can actively secrete DNA into sediment (e.g., plant root tips; Wen et al. 2017), while dead tissues can degrade, releasing the intracellular DNA (iDNA), along with the rest of the cell contents, when cell lysis occurs. Both active secretion of DNA as well as cell lysis result in iDNA becoming extracellular DNA (exDNA).

Once exposed to the sedimentary environment, exDNA can undergo different post-depositional taphonomic processes that determine the quality of the DNA on longer timescales. ExDNA can be internalised by microbial cells (Overballe-Petersen and Willerslev 2014), degraded by

extracellular microbial nucleases that break it up into smaller fragments, damaged by abiotic processes such as hydrolysis and oxidation, or preserved by adsorption onto particles such as humic acids, sand and clay minerals (Torti et al. 2015; Willerslev and Cooper 2005). An overview of DNA degradation processes is provided in Figure 1. Chemical alkylation can lead to cross-links within (intra) and between (inter) DNA molecules making it impossible to PCR amplify the DNA (Fulton and Shapiro 2019). Low pH, high temperatures, high oxygen and water content can also lead to strand breaks, deamination of nitrogen bases, and base modifications (Dabney et al. 2013; Willerslev and Cooper 2005). These processes can result in a decrease in the amount of detectable DNA, shorter DNA fragments, and changes in chemical properties as damage accumulates over time. DNA is better preserved in sediments with a high mineral content and at low temperatures. Minerals can inactivate nucleases as well as bind to and protect DNA, while low temperatures thermally stabilise DNA against chemical degradation (Torti et al. 2015). Desiccated dry and anoxic sediments will putatively also strongly decrease the effects of hydrolysis and oxidation, respectively. The preserved exDNA together with the iDNA preserved in dead cells make up the total DNA that can be recovered using *sed*aDNA methods.

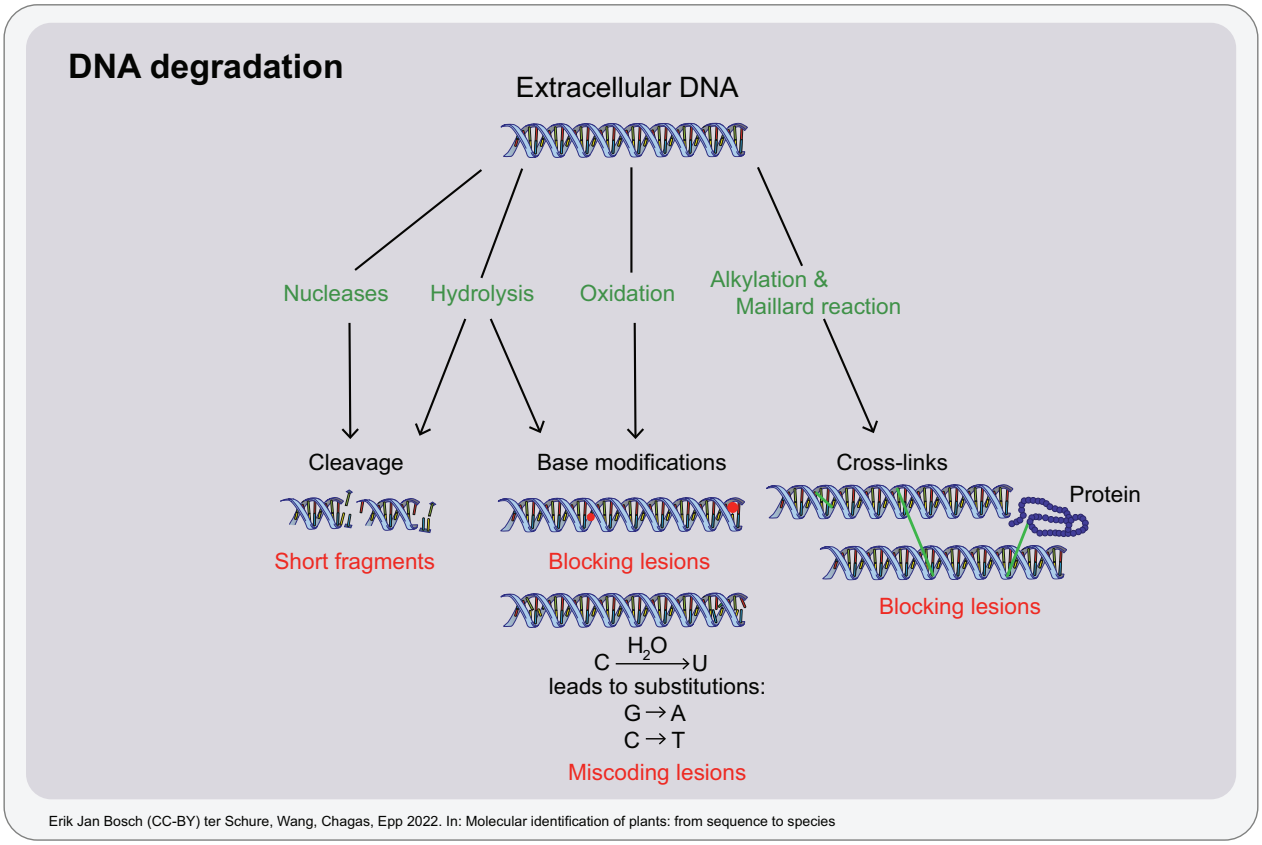

Figure 1. Schematic overview of DNA degradation processes (hydrolysis, oxidation, alkylation and Maillard reaction) that can cause DNA damage in the form of cleavage, base modifications or cross-links. The major mechanism leading to miscoding lesions in aDNA is the hydrolysis of cytosine to uracil, which leads to G to A and C to T substitutions by DNA polymerases, whereas blocking lesions can obstruct the movement of DNA polymerases during PCR (Dabney et al. 2013).

Advantages and limitations of *sed*aDNA as palaeoecological proxy

By analysing the ancient DNA present in the sediment (Haile et al. 2009; Slon et al. 2017) it is possible to identify the source species of archaeological artefacts and deposits, and even detect organisms in the absence of any visible remains. For plants, the detection of taxa that do

not leave traces in the fossil record (e.g., Alsos et al. 2016; Bremond et al. 2017; Brown et al. 2021; Pedersen et al. 2013) opens up new ways of studying past vegetation complementary to more traditional palaeoecological proxies such as pollen and macrofossils.

Macrofossils and plant *sed*aDNA originate close to the sample location and give a similar local signal (Alsos et al. 2018; Jørgensen et al. 2012; Niemeyer et al. 2017), while the pollen record generally includes taxa that originated from further away from the sample location (Parducci et al. 2018) as pollen, especially of wind-pollinated species, may originate from a wide area as they are distributed regionally through the air (Birks and Bjune 2010). Pollen does not contribute much to the total pool of *sed*aDNA (Clarke et al. 2020; Sjögren et al. 2017). This can be partially explained by the low DNA content of pollen grains and the robustness of the pollen grain wall, hindering the retrieval of the DNA. At the source, DNA can be considered more consistent than pollen, as all plant tissues contain DNA, but not all plants produce pollen, and insect-pollinated plants produce fewer pollen than wind-pollinated plants.

In general, palaeovegetation data are the result of the attributes of the original vegetation, combined with depositional factors and preservation, as well as the experimental procedures to produce the data. For *sed*aDNA analyses, this includes every step of the data generation itself: sampling, transport, storage, processing of the DNA in the laboratory, and finally, the bioinformatic pipelines used. In terms of the data generation, pollen analyses and macrofossil analyses rely on taxonomic identification by microscopy, which is labour-intensive and requires a high level of taxonomic knowledge. Although some training is needed to work in an ancient DNA laboratory, in principle, taxonomic identification by DNA can be carried out without prior taxonomic knowledge. However, familiarity with plant taxonomy, phylogenetic placement, and biology of different groups is invaluable in the interpretation of the automated identifications. For example, it is important to check if the automated DNA identifications make sense for the sample location, because contamination, DNA degradation, and the quality of the reference library can cause false DNA identifications (see Chapter 18 Sequence to species for details).

A combination of *sed*aDNA, macrofossils, and pollen proxies gives the most complete overview of plant diversity and community composition through time. The choice for these proxies is dependent on the aims of the study. Table 1 summarises the main differences.

*Sed*aDNA research applications

The first study using *sed*aDNA of macroorganisms was published in 2003, demonstrating the possibility to detect plant and animal DNA in both permafrost sediments and temperate cave sediments (Willerslev et al. 2003). Since then, the number of *sed*aDNA studies and applications has increased as enhanced understanding of ancient DNA and methodological developments allowed better reconstructions, as also illustrated by a recent comprehensive synthesis of current analytical procedures (Capo et al. 2021). *Sed*aDNA methods are relevant for a range of research fields across biology, conservation, and archaeology and have been applied for roughly two main purposes: understanding natural environmental processes and reconstructing past human-environmental interactions.

Environmental reconstructions can range from polar, to temperate and tropical regions, although they are limited to sampling sites that allow preservation of *sed*aDNA, such as permafrost, lake sediments, and dry cave sediments. Permafrost sediment can be used to assess vegetational development in polar regions under climate change (e.g., Willerslev et al. 2014; Zimmermann et al. 2017). *Sed*aDNA from archaeological sites can reveal human past activities such as plant and animal cultivation, migration and settlement history (e.g., Hebsgaard et al. 2009; Smith et al. 2015), and Neanderthal and Denisovan DNA have been recovered from

cave sediments (Slon et al. 2017; Vernot et al. 2021). Lake sediments can be reliable archives of the palaeoenvironment, integrating environmental information across the lake catchment area and displaying a very clear temporal stratification. Many *sed*aDNA studies use lake sediments to focus on past vegetation dynamics, which can be used to establish natural baselines for conservation (e.g., Boessenkool et al. 2014; Wilmshurst et al. 2014), reconstruct the effects of past climate change on the environment (e.g., Alsos et al. 2016, 2020; Clarke et al. 2020; Jørgensen et al. 2012), show long-lasting effects of biological invasions (e.g., Ficetola et al. 2018), or track past human impacts (e.g., Giguet-Covex et al. 2014; Pansu et al. 2015). This list illustrates the wide range of potential applications, but for further discussion, please see Section 3 of this book, especially Chapter 21 Palaeobotany and Chapter 24 Environment and biodiversity assessments can be relevant for *sed*aDNA.

Experimental design

*Sed*aDNA research strategy

Due to its low concentration, retrieving ancient DNA from sediment samples requires strict protocols to avoid contamination by modern DNA or further degradation (Cooper and Poinar 2000; Capo et al. 2021). However, once these protocols are followed *sed*aDNA can be a powerful tool providing novel insights to palaeoecology reconstructions that are not possible through traditional methods.

The previous section described some *sed*aDNA studies focusing on palaeoecological and archaeological questions. In both cases, choices of location and methods are very much steered by the research focus and what is already known about the area, such as past changes in climate, geology, ecology, or human impacts. Although details in the study design can differ, all *sed*aDNA studies follow the same steps: site selection, collection of samples and metadata,

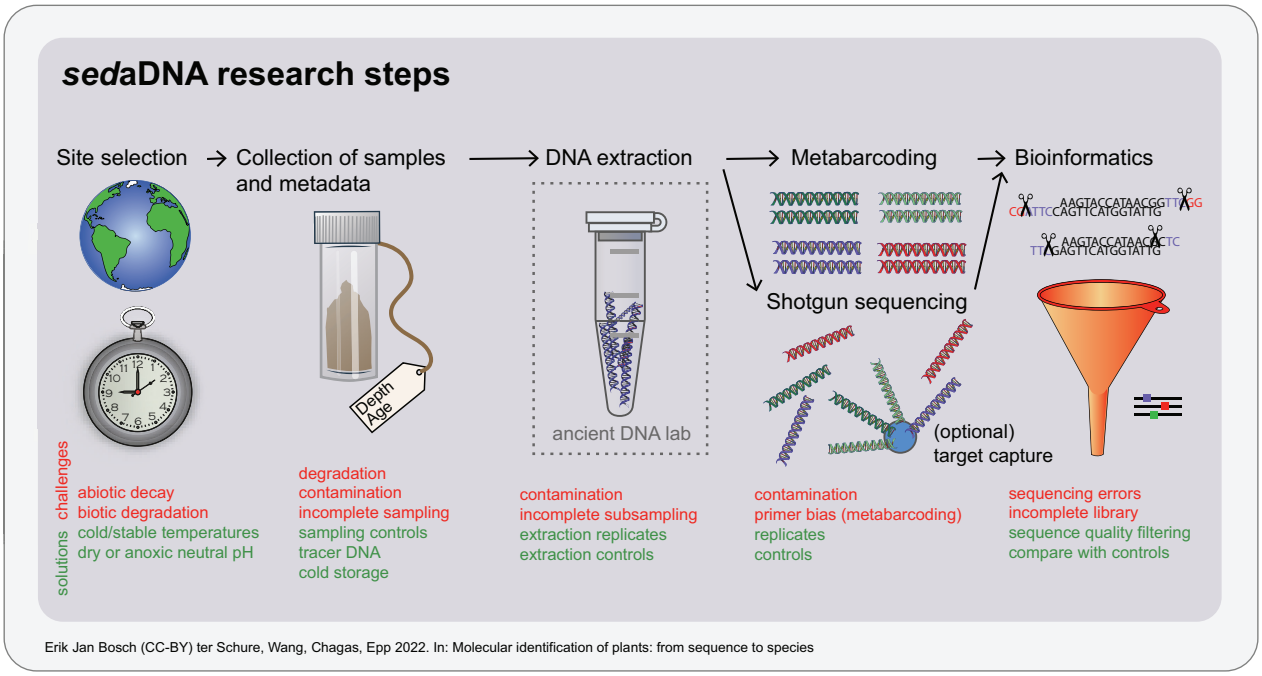

Figure 2. Simplified overview of the *sed*aDNA research process, including some of the major challenges and potential solutions indicated at each step.

DNA extraction, further processing of the DNA in the lab, sequencing, and finally, bioinformatic sequence quality filtering and data analyses (Figure 2).

Choices for the different options at each step depend on the aims of the study. For example, when performing a reconstruction of overall plant community dynamics with universal plant metabarcoding primers, the most common taxa and major trends in community change will be reliably retrieved in the first PCR performed (Alsos et al. 2016), with no specific sampling strategy. However, the detection of rare plant species will require a number of repeats (Alsos et al. 2016), and possibly sampling at several locations (Capo et al. 2021). The following questions can help to develop a *sed*aDNA research strategy and these topics will be discussed throughout this chapter:

1. What is my study aim?
2. What spatial and temporal scale do I need to cover?
3. What contextual information and metadata do I need?
4. What taxa should I target and at what taxonomic resolution?
5. What laboratory and analytical methods should I use?
6. How will I minimise / control for contamination, biases, and false positives?

Table 1. Comparison of pollen, plant macrofossils, and *sed*aDNA as proxies for palaeoecological reconstructions on the levels of: source and sediment, data generation, and data interpretation. Sources: Ahmed et al. 2018; Birks and Bjune 2010; Parducci et al. 2018, 2017.

Category	Pollen	Plant macrofossils	*Sed*aDNA
Source and sediment			
- Scale	Regional	Local	Local
- Taxonomic groups	Pollen-producers	All plants	All organisms
- *Potential sources of bias*	*High pollen-producing plants; vegetation cover close to sampling area; differential preservation*	*Differential preservation of tissue-types and species*	*Differential DNA degradation and decay*
Data generation			
- Labour-intensive	Yes	Yes	No
- Need for taxonomic knowledge	Yes	Yes	No
- Taxonomic resolution	Limited to identifiable pollen types, generally to genus level	Generally to species-level	Depends on the marker, possible to species-level *DNA contamination;*
- *Potential sources of bias*	*Identifiability of the remains*	*Identifiability of the remains; random occurrence*	*choice of lab techniques; completeness of reference library*
Data interpretation			
- Qualitative	Yes	Yes	Yes
- Quantitative	Partial	Limited	Debated

Site selection

The aims of the study define the temporal and spatial scale needed to achieve them, thereby steering the selection of relevant sampling sites. Lake sediments provide a record of the plants

that occurred in the lake catchment, being the area of land from which water and surface runoff drains into the lake (Giguet-Covex et al. 2019). A lake sediment record can only go as far back as the formation of the lake itself. Other terrestrial sediments may primarily contain the DNA that is deposited by plants growing at that particular location, or by humans, animals, or abiotic factors such as wind and water. For example, DNA in cave sediments will come primarily from organisms that have lived or died in the cave, or from remains that are transported into the cave (Hofreiter et al. 2003). The likelihood of finding *sed*aDNA should also be considered. However, more often than not the sampling location is opportunity driven, especially when it comes to archaeological sites, and *sed*aDNA retrieval can prove difficult.

General conditions under which *sed*aDNA preserves well are: cold and stable temperatures, neutral pH, dry or anoxic sediments with a high mineral content. Sediments from rockshelters, dry caves, and lake sediments are generally preferred as they are protected and provide stable conditions: rockshelter and dry cave sediments are sheltered from rain and have stable temperatures and there is some evidence that calcite has a high adsorption capacity for DNA (Capo et al. 2021; Freeman et al. 2020). Lake sediments on the other hand are often anoxic and generally undisturbed, especially when they are below the wave disturbance depth and subsurface slopes are gentle.

Dating of sediments

Dating is important in any study that involves ancient samples. Only with accurate dating can the timing of events be compared and their rates of change estimated. Commonly applied sediment dating methods are radioisotopic dating (in particular ^{210}Pb, ^{14}C, and luminescence dating) and dating based on chemostratigraphy or marker minerals (in particular tephrochronology), and the choice for a method depends on the type and age of the sediments (see Table 2 for an overview). Many sources describe these methods in detail (e.g., Bradley 1999) and we provide a brief introduction here.

Radioisotopic dating is based on the principle of radioactive decay. When a nucleus breaks down, it emits energy and forms a daughter product. The time this takes is expressed as the half-life, i.e., the time that it takes for 50% of a parent element to transmute into the daughter product. The relative quantity of a radioactive parent element in a sample can be used to infer its age. Relatively young aquatic sediments, with ages up to 150 years are commonly dated with ^{210}Pb (half-life: 22.27 years; Barsanti et al. 2020). ^{210}Pb occurs naturally in the atmosphere and settles in sediments through dry fallout or precipitation. The supply of this ^{210}Pb is not constant but the decline of this excess ^{210}Pb along a sediment sequence is a proxy for the sedimentation rate. Additionally, if the age at a point of the sequence is known, a chronology can be determined. Radiocarbon (^{14}C, half-life: 5730 years) is a radioactive isotope of carbon that naturally occurs in the atmosphere. Plants fix atmospheric carbon during photosynthesis, so the level of ^{14}C in plants and animals upon death approximately equals the level of ^{14}C in the atmosphere at that time. After death, it decreases as ^{14}C decays to ^{14}N at a rate of 50% per 5730 years, allowing the date of death to be estimated. Limited by its half-life, radiocarbon dating is only possible for samples younger than 50,000 years. As the concentration of atmospheric ^{14}C is not constant over time, radiocarbon dates are calibrated against a global calibration curve obtained from tree rings and varved lake sediments (Reimer et al. 2020). This produces calendrical dates, which are expressed as calibrated years before present (cal years BP) with present being 1950 (before large-scale testing of nuclear weapons). The most reliable age-depth models for both marine and lake sediments use accelerator mass-spectrometry (AMS) dating of macroscopic plant or animal fragments (as little as 0.1 mg) as this can avoid the problems of both mixed material and also the so-called hard-water error associated with carbonate waters.

Luminescence dating is based on the phenomenon that mineral crystals absorb electrons from the ionising radiation of surrounding sediments over time, and when stimulated in a laboratory by heat or light, they release the accumulated radiation as luminescence. The intensity of measured luminescence indicates the length of time between this in-lab stimulation and the last natural event of similar stimulation. Heat stimulated or thermoluminescence (TL) dating is used to date baked pottery from archeological sites or sediments once in contact with molten lava; optically stimulated luminescence (OSL) dating is used to date sediments once exposed to sunlight. The time range for luminescence dating can be from a few decades to over 1 Ma, depending on the ability of a mineral to absorb radiation over time. For studies concerning relatively young samples, OSL dating of quartz grains are generally used, covering from a few decades to ~150 ka.

Tephrochronology uses the chemical signature of tephra (volcanic ash) to pinpoint the age of that specific layer in a sediment sequence by reference to known or unknown dated volcanic eruptions. Terrestrial sediments (Froese et al. 2006), marine deposits (Larsen et al. 2002), and ice cores (Davies et al. 2008) from areas once under the influence of dated volcanic eruption events can be dated with this method. With accurate geochemical fingerprinting, tephrochronology can be used to corroborate or even extend the dating limits of other techniques.

Table 2. Summary of sediment dating methods, their applicability and limitations. Sources: Barsanti et al. 2020; Bradley 1999; Fattahi and Stokes 2003.

Dating method	Suitable sample types	Age limit	Sources of error and uncertainty
^{210}Pb dating	Materials from aquatic environments such as lacustrine and marine deposits	~100 to 150 years	Complex sedimentation processes that break the dating model assumptions, such as compaction, local mixing, erosion etc.
^{14}C (radiocarbon) dating	Organic remains (charcoal, wood, animal tissue), carbonates (corals, sediments, stalagmites and stalactites), water, air and organic matter from various sediments, soil, paleosol and peat deposits	Up to 50,000 years	Atmospheric ^{14}C content fluctuation due to changes in cosmogenic production rate and exchange between the atmosphere and ocean
Luminescence dating: - Thermoluminescence (TL) - Optical stimulated luminescence (OSL)	TL: materials containing crystalline minerals, such as sediments, lava, clay, and ceramics OSL: materials containing quartz or potassium feldspar sand-sized grains, or fine-grained mineral deposits	TL: A few years to over 1,000,000 years OSL: A few decades to ~150,000 years for quartz.	Variations in environmental radiation dose; saturation of electron traps in sample minerals
Tephrochronology	Terrestrial and lake sediments, marine deposits and ice cores that contain tephra	Up to 35,000 years, extendable under good conditions	Can only obtain indirect dates within the ^{14}C age range

Prepare to work cleanly

DNA is everywhere - including in the air - and contamination can come from many different sources. When collecting and working with *sed*aDNA samples, it is important to keep in mind

that the DNA you are interested in will probably be present in very low concentrations. Contamination with modern DNA can easily overpower the *sed*aDNA signal in which you are interested. Therefore it is important to absolutely minimise the amount of modern DNA coming into your samples and limit further degradation of the sedaDNA.

The precautions you can take include: work cleanly, use equipment that is free of DNA and nucleases, and try to keep the samples in a stable and cold environment. In practice this is not so easy, which is why dedicated ancient DNA facilities are set up to avoid any form of contamination. These facilities should be physically isolated - ideally in a separate building - from any location where PCRs are performed (Fulton and Shapiro 2019) and strict cleaning regimes and clean lab practices should be upheld. How to set up and work in an ancient DNA lab is described in detail by e.g., Cooper and Poinar (2000) and Fulton and Shapiro (2019). Here we summarise general clean lab practices. We note that working cleanly and consistently will require practice and adequate training.

You should assume that everything that you bring into the lab is contaminated with DNA. Therefore, before entering the lab, you should have showered and changed into clean clothes and everything you bring into the lab should be decontaminated. Inside the lab, you should wear a hairnet, face mask, full body suit with hood, shoe covers, and gloves at all times. Wearing two layers of gloves will allow you to change the outer gloves while still covering your hands, and you should change your outer gloves regularly while working. All tools and equipment should be decontaminated before use, and regular cleaning of the aDNA workspace is needed. Decontamination can be achieved by using a DNA decontamination product (e.g., 3-10% bleach or DNA-ExitusPlus™) for surfaces, ideally supplemented with UV irradiation of the workspace. To prevent cross-contamination, tools should be cleaned between working with each sample or sample-extract. Tools should be left in a DNA decontamination product for at least 10 minutes, rinsed with UV irradiated milliQ water, and ideally also UV irradiated using a UV cross-linker with irradiation at the shortest distance possible to the UV source (Champlot et al. 2010).

Collection, transport, and storage of ancient sediment samples

Choices for sampling and personal protective equipment will depend on the setting, as the sampling of sediments at an archaeological site can be very different from the sub-sampling of a lake sediment core in a lab facility. It is important to try to limit the amount of potential contamination, but practical considerations and the target DNA can also be leading. For example, a study aiming to recover human aDNA will require stricter use of personal protective equipment than a study focussing on plant aDNA. Sampling of sediments can be done directly in the field or by subsampling of sediment cores in a clean, sheltered environment. When collecting sediment cores for *sed*aDNA, closed-chamber piston-type corers are preferred (Parducci et al. 2017) as they enclose the sediment in a plastic tube that can be opened in the laboratory. As frozen sediments should be kept at freezing temperatures, subsampling of these types of cores requires a climate chamber (Epp et al. 2019).

A general *sed*aDNA sampling kit contains personal protective equipment, sampling equipment, and cleaning products, including: full bodysuits, face masks, hairnets, nitrile gloves, sterile scalpels, sample tubes, clean ziplock bags, DNA decontamination products, distilled water, 70% ethanol, trays or beakers for cleaning the tools, paper towels, trash bags and pens for labelling. To limit potential contamination, much of the preparation for the sampling kit takes place in the ancient DNA lab facility: making sure the sampling tools and collection tubes are prepared and DNA-free. Aluminium foil can be helpful for covering your workspace and provides a clean surface for all of the sampling materials at a sampling site. Sterile syringes with the tip cut off can be useful mini-corers, speeding up the sample-taking (Epp et al. 2019). If you are taking sub-samples in a lab facility, make sure it is isolated from any PCR machine as the high

number of DNA copies produced with PCR can become airborne and may enter your samples through the building air supply (Fulton and Shapiro 2019; Willerslev and Cooper 2005). Tracing of contamination during sampling can be done by placing several open sample tubes with DNA-free water in your work area (Parducci et al. 2017), or using tracer DNA during coring or on the outside of the sediment core (Epp et al. 2019; Pedersen et al. 2016).

The sampling itself follows aDNA lab procedures where possible, even if it takes place elsewhere: clean the workspace, use personal protective equipment, do not hover over the sediment you are sampling and change outer gloves and tools between each individual sample. In order to avoid contamination, sampling should start at the oldest part of the sediment, working your way up to the youngest parts and subsamples from sediment cores should be taken from inside the undisturbed centre (Parducci et al. 2017). Sampling procedures for both non-frozen and frozen sediment cores are described in detail by Epp et al. (2019). Collected samples should be kept in a stable and low-temperature environment (i.e. freeze at -20 for longer term storage), as degradation slows down with lower temperatures and temperature fluctuations can be additionally damaging to the DNA. An ice-box with ice packs can be used for temporary storage and transport of the taken samples. Further processing of the *sed*aDNA samples should be done in a laboratory dedicated to working with ancient DNA.

Sedimentary ancient DNA extraction

The choice for a specific DNA extraction protocol depends on a range of factors, including the aim of your study, sample characteristics, available laboratory facilities and equipment, and costs of the reagents or extraction kits. The latter can be a consideration of investing either time or finances as it can be cheaper to make the buffers needed for extraction yourself, but this also increases the preparation time and could introduce additional contamination to your samples. There are several protocols that can be used for *sed*aDNA extraction (see Capo et al. 2021; for

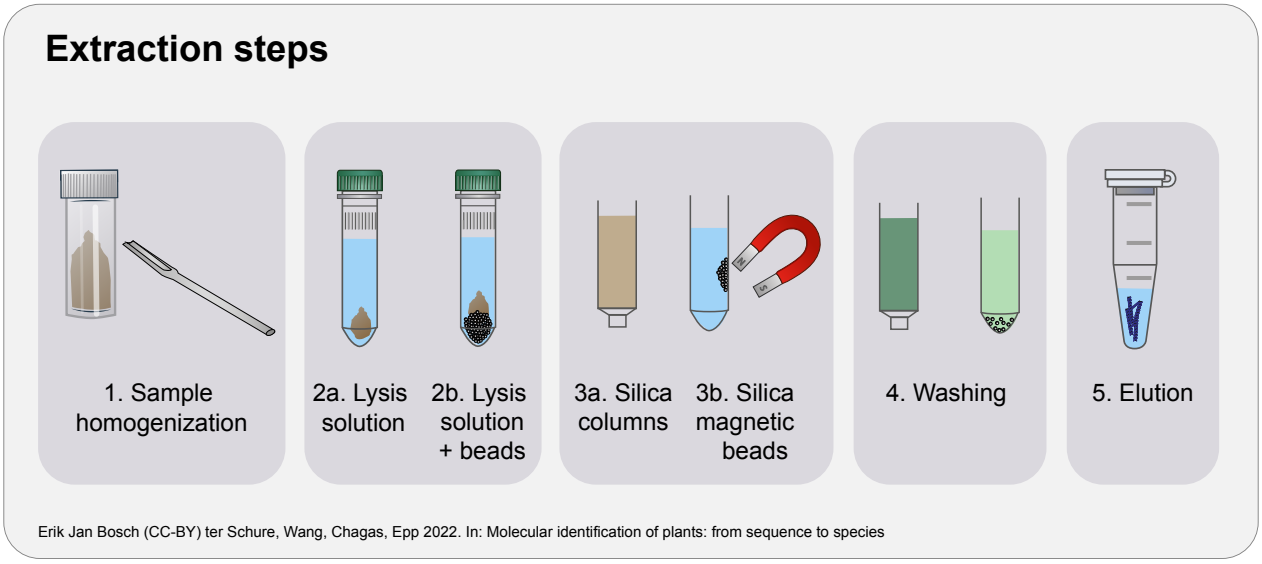

Figure 3. Common DNA extraction steps: (1) samples are first homogenized using a sterile scalpel and later on go through a step, in which either (2a) extracellular DNA is washed off the sedimentary matrix (Taberlet et al. 2012) and/or (2b) intracellular DNA is freed through lysis, which can include beating with garnet beads. The free DNA suspended in a high salt buffer can now bind to either (3a) a silica column or (3b) silica magnetic beads, (4) samples are washed with an ethanol based buffer to remove impurities, and finally (5) DNA is eluted in an elution buffer. Figure based on Rohland et al. (2018).

a detailed review) and general steps are: sample homogenization, lysis, binding, washing, and elution of the DNA. Here we discuss some of the most commonly used extraction protocols and we summarise their main advantages and limitations in Table 3.

All extraction protocols include similar steps for the isolation of sedimentary DNA (Figure 3), but due to the differences in chemical composition of the buffers, input volume, use of equipment, and targeted DNA (total DNA, iDNA, or exDNA), results of these protocols can vary. You can decide to extract only exDNA using the "Taberlet protocol", where samples are first incubated in a saturated phosphate buffer and later on purified with an extraction kit, skipping the lysis step (Taberlet et al. 2012). An advantage is that a large sample volume can be processed, minimising the possible effects of heterogeneous distribution of DNA in the sediment. However, DNA yield and purity can be lower in comparison to the DNeasy PowerMax Kit (Qiagen), formerly known as the PowerMax Soil DNA Isolation Kit (MO BIO Laboratories, Inc.; Zinger et al. 2016) and probably also to other protocols targeting total DNA (e.g., the Rohland protocol; Rohland et al. 2018).

*Sed*aDNA studies employing protocols developed for the extraction of modern environmental DNA from soils and sediments generally add additional steps to increase the yield of DNA from low concentration ancient sediment samples. A lysis step can be added to extract iDNA from intact cells present in the samples through chemical lysis, and/or mechanical shearing of cell membranes using beads. Adding certain chemicals to the lysis buffer can also

Table 3. Overview of the advantages and limitations of several commonly used extraction protocols and some example publications using these protocols.

Extraction protocol	Sample size	Advantages	Limitations	Used by
DNeasy PowerMax kit (Qiagen)	≤ 10 g	- Large initial sample volume - Few inhibitors in the resulting extract	- Expensive - DNA can be lost with inhibitor removal solution	Epp et al. 2018; Zimmermann et al. 2017
DNeasy PowerSoil kit (Qiagen)	≤ 250 mg	- Few amplification and sequencing inhibitors in the resulting extract - Easy processing of large sets of samples	- DNA can be lost with inhibitor removal solution - Smaller initial sample volume compared to the PowerMax kit	Lejzerowicz et al. 2013; Monchamp et al. 2016; Dommain et al. 2020
Rohland protocol (Rohland et al. 2018)	≤ 50 mg	- Developed to recover small DNA fragments - Easy processing of large sets of samples	- Small starting amount of sediment - Potential coextraction of inhibitors - Homemade buffers can increase contamination risk	Zavala et al. 2021; Vernot et al. 2021
Phosphate buffer + NucleoSpin® Soil kit (Taberlet et al. 2012)	≤ 15 g	- Large initial sample volume	- Extracts only extracellular DNA - Processes a 2 ml subsample of the phosphate buffer and sample mixture	Giguet-Covex et al. 2014; Pansu et al. 2015
Murchie protocol (Murchie et al. 2020)	≤ 250 mg	- High DNA yields - Uses a high volume binding buffer to improve the recovery of small DNA fragments	- Optimised for permafrost samples and may not perform as well in lake sediment	Murchie et al. 2020

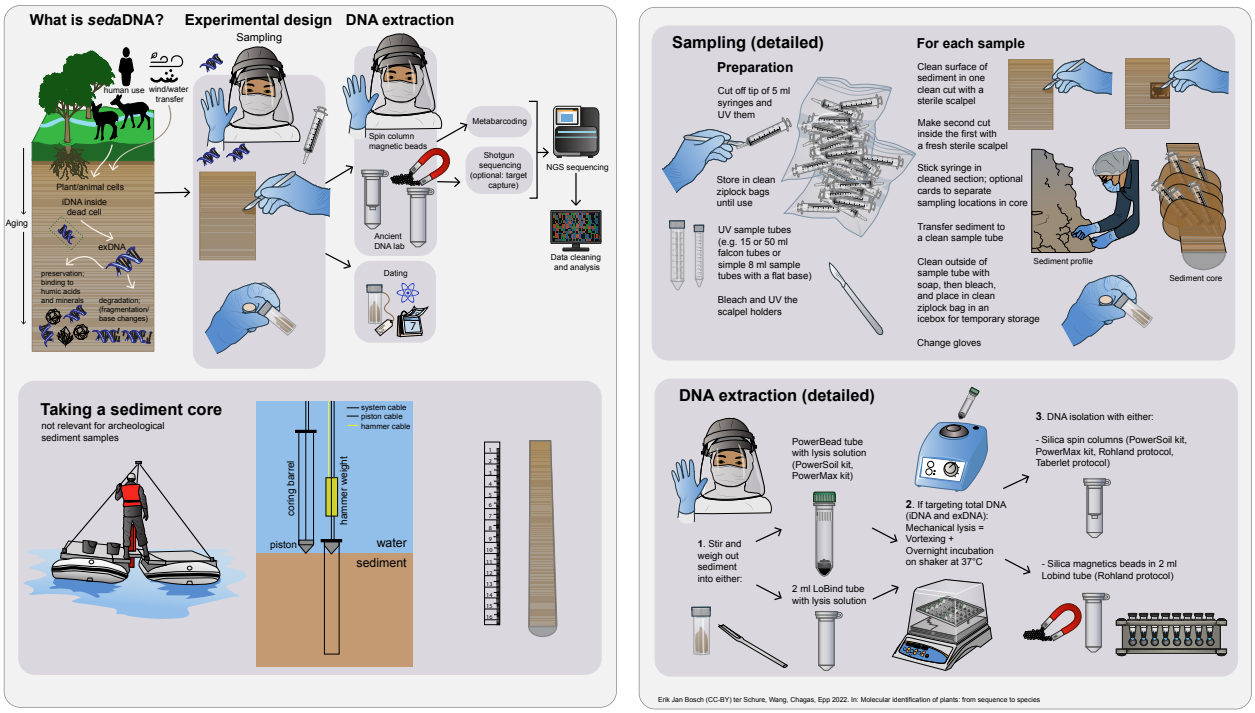

Figure 4. Chapter 8 Infographic: Visual representation of the content of this chapter. Top left image based on Pederson et al. (2015).

increase yield: *N*-phenacylthiazolium bromide (PTB) breaks down cross-links between DNA and proteins (Vasan et al. 1996; Poinar et al. 1998), and adding proteinase K and dithiothreitol (DTT) during the lysis step of the PowerMax and PowerSoil kits allows better recovery of DNA (Epp et al. 2019). It has also been suggested to concentrate the DNA before further processing (Taberlet et al. 2018), as *sed*aDNA concentrations are likely to be low (Zimmermann et al. 2020). The Rohland protocol is specifically designed to target degraded DNA from ancient samples (Rohland et al. 2018) and should yield a higher concentration of short fragments compared to the other extraction protocols, especially when silica magnetic beads are used for DNA binding.

Be aware that the presence of certain substances may inhibit further amplification or sequencing steps. These can be derived from humic substances (important components of humus), which are commonly present in sediments and might inhibit downstream analysis. Moreover, the amount of humic substances is site-specific, and it might be necessary to repurify the samples or use inhibitor removal columns. During DNA extraction, contamination may be introduced from the laboratory facilities, tools, reagents and other consumables. It is essential to track this contamination by including a negative control. It is suggested to add one such extraction control for each batch of 11 samples, and include it in all subsequent steps (e.g., metabarcoding, library preparation, sequencing; Rohland et al. 2018). It is common for the extraction of modern DNA to add a positive control with a known DNA content, but due to the contamination risk this is not recommended for *sed*aDNA (Willerslev and Cooper 2005).

Molecular methods for *sed*aDNA

After extracting the DNA, the *sed*aDNA needs to be further processed before sequencing and several approaches are continuously being improved and new ones developed.

Most *sed*aDNA studies apply a DNA metabarcoding approach, using PCR amplification primers to target short DNA sequences (< 300 bp, preferentially around or below 100 bp) from

taxonomic marker genes to identify specific taxonomic groups (see Chapter 11 Amplicon metabarcoding). It is relatively low cost and some of the metabarcoding primers give high taxonomic resolution. However, this method can introduce amplification bias (Bellemain et al. 2010) and is susceptible to errors introduced in the PCR. More recently, shotgun sequencing became another option for these types of samples (Pedersen et al. 2016). This approach converts the DNA extracts directly to a library for sequencing, allowing the analyses of the entire diversity of taxonomic groups in the samples including microorganisms (Ahmed et al. 2018), plants (Parducci et al. 2019; Pedersen et al. 2016), animals (Graham et al. 2016; Pedersen et al. 2016), and humans (Slon et al. 2017; Vernot et al. 2021). Shotgun sequencing requires a high sequencing depth and can be costly as most sequences will be from non-target organisms. Target capture has recently been applied to *sed*aDNA samples to enrich the concentration of taxa of interest in a shotgun approach by using DNA (Schulte et al. 2020) or RNA (Murchie et al. 2020; Seeber et al. 2019) baits. These methods are described in detail in Chapter 11 Amplicon metabarcoding, Chapter 12 Metagenomics, and Chapter 14 Target capture, and are followed by library preparation and sequencing (see Chapter 9 Sequencing platforms and data types).

Sequencing data can be processed using bioinformatic tools, where strict quality filtering of the sequence data is followed by taxonomic assignment. Further filtering allows removal of sequences with low identity scores, contaminants (i.e., sequences present in the controls), and false-positives (see Chapter 18 Sequence to species for details). False identifications can be caused by the quality of the reference library, but also by technical errors, contamination, or errors in the DNA sequences, especially as *sed*aDNA is generally highly degraded and of low concentration. It is therefore important to check if the identifications make sense for the sampling location and age before further analyses of the *sed*aDNA data.

Questions

1. Name and explain two main advantages of using *sed*aDNA as a proxy for past plant presence compared to pollen. Motivate your answer.
2. Imagine you have a long lake sediment core that is thought to be between 50 000 and 10 000 years old. What dating methods could be used to date this core and why?
3. What are the main sources of bias when working with *sed*aDNA (name at least 3) and how can you limit the resulting false positives?

Glossary

Alkylation – Addition or substitution of an alkyl group (C_nH_{2n+1}) to an organic molecule.
Accelerator Mass-Spectrometry (AMS) dating – A dating method that determines the age of an organic material (i.e., macroscopic remains of plants or animals) by measuring their radiocarbon concentration.
Cell lysis – The process whereby the membrane(s) of a cell breaks down, thereby releasing the cell contents.
exDNA – Extracellular DNA; all DNA located outside cell membranes.
Geochemical fingerprinting – A method using chemical signals to infer the origin, the formation and/or the environment of a geological sample.

Half-life – The time necessary for half of a radioactive atom's nucleus to decay by emission of matter and energy to form a new daughter product. The half-life is specific to a radioactive element, and can be used for dating purposes.
iDNA – Intracellular DNA; all DNA present within cell membranes.
Lake catchment – Area of land from which water and surface runoff drains into a lake.
Luminescence dating – A group of methods to determine how long ago mineral grains were last exposed to sunlight or sufficient heating by measuring the luminescence emitted by the mineral grain upon stimulation.
Metabarcoding – Method for the simultaneous identification of many taxa within the same complex DNA extract. This is achieved by high throughput sequencing (HTS) of amplicons from taxonomic marker genes (barcodes).
Next Generation Sequencing (NGS) – Massively parallel sequencing technology allowing high throughput of DNA.
Nucleases – Diverse group of enzymes able to hydrolyze the phosphodiester bonds of DNA and RNA thereby cleaving them into smaller fragments.
Optically stimulated luminescence (OSL) dating – Dating method that determines the age of a sample by measuring the luminescence it emits in response to visible or infrared light.
Palaeoecology – The study of the relationship between past organisms and their ancient environments.
Permafrost – Soil, sediment, or rock that is continuously exposed to temperatures of < 0 °C for at least two consecutive years.
Radioactive isotope – An atom with excess nuclear energy and prone to undergo radioactive decay.
Reference library – A database of known DNA sequences with their taxonomic identifications, used in bioinformatics as a reference to identify the DNA sequences obtained in a *sed*aDNA study.
***sed*aDNA** – Sedimentary ancient DNA; this is the aged and degraded DNA from dead organisms now incorporated in the sediment record, either as iDNA in dead tissues, or as exDNA free in the sediment matrix or adsorbed to sediment particles.
Shotgun sequencing – A method for the random sequencing of all of the DNA within a DNA extract.
Taphonomic processes – The processes involved in the transfer, deposition and preservation or organismal remains, including DNA.
Target capture – A technique that allows the capture of the DNA of interest by hybridization to target-specific probes (baits).
Tephrochronology – A geochronological technique that uses layers of tephra (volcanic ash from a single volcanic eruption) to create a chronological framework for the sedimentary record.
Thermoluminescence (TL) dating – Dating method that determines the age of a sample by measuring the luminescence it emits in response to heat.
Total DNA – The intracellular and extracellular DNA combined.
Tree-ring dating – Also called dendrochronology; a method of dating tree rings to the exact year they were formed.

References

Ahmed E, Parducci L, Unneberg P, Ågren R, Schenk F, Rattray JE, Han L, Muschitiello F, Pedersen MW, Smittenberg RH, Yamoah KA, Slotte T, Wohlfarth B (2018) Archaeal community changes in Lateglacial lake sediments: Evidence from ancient DNA. Quat. Sci. Rev. 181, 19–29. https://doi.org/10.1016/j.quascirev.2017.11.037

Alsos IG, Lammers Y, Yoccoz NG, Jørgensen T, Sjögren P, Gielly L, Edwards ME (2018) Plant DNA metabarcoding of lake sediments: How does it represent the contemporary vegetation. PLoS ONE 13, e0195403. https://doi.org/10.1371/journal.pone.0195403

Alsos IG, Sjögren P, Brown AG, Gielly L, Merkel MKF, Paus A, Lammers Y, Edwards ME, Alm T, Leng M, Goslar T, Langdon CT, Bakke J, van der Bilt WGM (2020) Last Glacial Maximum environmental conditions at Andøya, northern Norway; evidence for a northern ice-edge ecological "hotspot." Quat. Sci. Rev. 239, 106364. https://doi.org/10.1016/j.quascirev.2020.106364

Alsos IG, Sjögren P, Edwards ME, Landvik JY, Gielly L, Forwick M, Coissac E, Brown AG, Jakobsen LV, Foreid MK, Pedersen MW (2016) Sedimentary ancient DNA from Lake Skartjorna, Svalbard: Assessing the resilience of arctic flora to Holocene climate change. The Holocene 26, 627–642. https://doi.org/10.1177/0959683615612563

Barsanti M, Garcia-Tenorio R, Schirone A, Rozmaric M, Ruiz-Fernández AC, Sanchez-Cabeza JA, Delbono I, Conte F, De Oliveira Godoy JM, Heijnis H, Eriksson M, Hatje V, Laissaoui A, Nguyen HQ, Okuku E, Al-Rousan SA, Uddin S, Yii MW, Osvath I (2020) Challenges and limitations of the 210Pb sediment dating method: Results from an IAEA modelling interlaboratory comparison exercise. Quat. Geochronol. 59, 101093. https://doi.org/10.1016/j.quageo.2020.101093

Bellemain E, Carlsen T, Brochmann C, Coissac E, Taberlet P, Kauserud H (2010) ITS as an environmental DNA barcode for fungi: an in silico approach reveals potential PCR biases. BMC Microbiol. 10, 189. https://doi.org/10.1186/1471-2180-10-189

Birks HH, Bjune AE (2010) Can we detect a west Norwegian tree line from modern samples of plant remains and pollen? Results from the DOORMAT project. Veg. Hist. Archaeobot. 19, 325–340. https://doi.org/10.1007/s00334-010-0256-0

Boessenkool S, McGlynn G, Epp LS, Taylor D, Pimentel M, Gizaw A, Nemomissa S, Brochmann C, Popp M (2014) Use of ancient sedimentary DNA as a novel conservation tool for high-altitude tropical biodiversity. Conserv. Biol. 28, 446–455. https://doi.org/10.1111/cobi.12195

Bradley RS (1999) Paleoclimatology: reconstructing climates of the Quaternary. Elsevier.

Bremond L, Favier C, Ficetola GF, Tossou MG, Akouégninou A, Gielly L, Giguet-Covex C, Oslisly R, Salzmann U (2017) Five thousand years of tropical lake sediment DNA records from Benin. Quat. Sci. Rev. 170, 203–211. https://doi.org/10.1016/j.quascirev.2017.06.025

Brown AG, Van Hardenbroek M, Fonville T, Davies K, Mackay H, Murray E, Head K, Barratt P, McCormick F, Ficetola GF, Gielly L, Henderson ACG, Crone A, Cavers G, Langdon PG, Whitehouse NJ, Pirrie D, Alsos IG (2021) Ancient DNA, lipid biomarkers and palaeoecological evidence reveals construction and life on early medieval lake settlements. Sci. Rep. 11, 11807. https://doi.org/10.1038/s41598-021-91057-x

Capo E, Giguet-Covex C, Rouillard A, Nota K, Heintzman PD, Vuillemin A, Ariztegui D, Arnaud F, Belle S, Bertilsson S, Bigler C, Bindler R, Brown AG, Clarke CL, Crump SE, Debroas D, Englund G, Ficetola GF, Garner RE, Gauthier J, Parducci L (2021) Lake sedimentary DNA research on past terrestrial and aquatic biodiversity: overview and recommendations. Quaternary 4, 6. https://doi.org/10.3390/quat4010006

Champlot S, Berthelot C, Pruvost M, Bennett EA, Grange T, Geigl E-M (2010) An efficient multistrategy DNA decontamination procedure of PCR reagents for hypersensitive PCR applications. PLoS ONE 5. https://doi.org/10.1371/journal.pone.0013042

Clarke CL, Alsos IG, Edwards ME, Paus A, Gielly L, Haflidason H, Mangerud J, Regnéll C, Hughes PDM, Svendsen JI, Bjune AE (2020) A 24,000-year ancient DNA and pollen record from the Polar Urals reveals temporal dynamics of arctic and boreal plant communities. Quat. Sci. Rev. 247, 106564. https://doi.org/10.1016/j.quascirev.2020.106564

Cooper A, Poinar HN (2000) Ancient DNA: do it right or not at all. Science 289, 1139. https://doi.org/10.1126/science.289.5482.1139b

Dabney J, Meyer M, Pääbo S (2013) Ancient DNA damage. Cold Spring Harb. Perspect. Biol. 5. https://doi.org/10.1101/cshperspect.a012567

Davies SM, Wastegård S, Rasmussen TL, Svensson A, Johnsen SJ, Steffensen JP, Andersen KK (2008) Identification of the Fugloyarbanki tephra in the NGRIP ice core: a key tie-point for marine and ice-core sequences during the last glacial period. J. Quaternary Sci. 23, 409–414. https://doi.org/10.1002/jqs.1182

Dommain R, Andama M, McDonough MM, Prado NA, Goldhammer T, Potts R, Maldonado JE, Nkurunungi JB, Campana MG (2020) The Challenges of Reconstructing Tropical Biodiversity With Sedimentary Ancient DNA: A 2200-Year-Long Metagenomic Record From Bwindi Impenetrable Forest, Uganda. Front. Ecol. Evol. 8. https://doi.org/10.3389/fevo.2020.00218

Epp LS, Kruse S, Kath NJ, Stoof-Leichsenring KR, Tiedemann R, Pestryakova LA, Herzschuh U (2018) Temporal and spatial patterns of mitochondrial haplotype and species distributions in Siberian larches inferred from ancient environmental DNA and modeling. Sci. Rep. 8, 17436. https://doi.org/10.1038/s41598-018-35550-w

Epp LS, Zimmermann HH, Stoof-Leichsenring KR (2019) Sampling and Extraction of Ancient DNA from Sediments. Methods Mol. Biol. 1963, 31–44. https://doi.org/10.1007/978-1-4939-9176-1_5

Fattahi M, Stokes S (2003) Dating volcanic and related sediments by luminescence methods: a review. Earth-Science Reviews 62, 229–264. https://doi.org/10.1016/S0012-8252(02)00159-9

Ficetola GF, Poulenard J, Sabatier P, Messager E, Gielly L, Leloup A, Etienne D, Bakke J, Malet E, Fanget B, Støren E, Reyss J-L, Taberlet P, Arnaud F (2018) DNA from lake sediments reveals long-term ecosystem changes after a biological invasion. Sci. Adv. 4, eaar4292. https://doi.org/10.1126/sciadv.aar4292

Freeman, Dieudonné, Collins, Sand (2020) Survival of environmental DNA in natural environments: Surface charge and topography of minerals as driver for DNA storage. BioRxiv. https://doi.org/10.1101/2020.01.28.922997

Froese DG, Zazula GD, Reyes AV (2006) Seasonality of the late Pleistocene Dawson tephra and exceptional preservation of a buried riparian surface in central Yukon Territory, Canada. Quat. Sci. Rev. 25, 1542–1551. https://doi.org/10.1016/j.quascirev.2006.01.028

Fulton TL, Shapiro B (2019) Setting up an ancient DNA laboratory. Methods Mol. Biol. 1963, 1–13. https://doi.org/10.1007/978-1-4939-9176-1_1

Giguet-Covex C, Ficetola GF, Walsh K, Poulenard J, Bajard M, Fouinat L, Sabatier P, Gielly L, Messager E, Develle AL, David F, Taberlet P, Brisset E, Guiter F, Sinet R, Arnaud F (2019) New insights on lake sediment DNA from the catchment: importance of taphonomic and analytical issues on the record quality. Sci. Rep. 9, 14676. https://doi.org/10.1038/s41598-019-50339-1

Giguet-Covex C, Pansu J, Arnaud F, Rey P-J, Griggo C, Gielly L, Domaizon I, Coissac E, David F, Choler P, Poulenard J, Taberlet P (2014) Long livestock farming history and human landscape shaping revealed by lake sediment DNA. Nat. Commun. 5, 3211. https://doi.org/10.1038/ncomms4211

Graham RW, Belmecheri S, Choy K, Culleton BJ, Davies LJ, Froese D, Heintzman PD, Hritz C, Kapp JD, Newsom LA, Rawcliffe R, Saulnier-Talbot É, Shapiro B, Wang Y, Williams JW, Wooller MJ (2016) Timing and causes of mid-Holocene mammoth extinction on St. Paul Island, Alaska. Proc Natl Acad Sci USA 113, 9310–9314. https://doi.org/10.1073/pnas.1604903113

Haile J, Froese DG, Macphee RDE, Roberts RG, Arnold LJ, Reyes AV, Rasmussen M, Nielsen R, Brook BW, Robinson S, Demuro M, Gilbert MTP, Munch K, Austin JJ, Cooper A, Barnes I, Möller P, Willerslev E (2009) Ancient DNA reveals late survival of mammoth and horse in interior Alaska. Proc Natl Acad Sci USA 106, 22352–22357. https://doi.org/10.1073/pnas.0912510106

Hebsgaard MB, Gilbert MTP, Arneborg J, Heyn P, Allentoft ME, Bunce M, Munch K, Schweger C, Willerslev E (2009) 'The Farm Beneath the Sand' – an archaeological case study on ancient 'dirt' DNA. Antiquity 83, 430–444. https://doi.org/10.1017/S0003598X00098537

Hofreiter M, Mead JI, Martin P, Poinar HN (2003) Molecular caving. Curr. Biol. 13, R693-5.

Jørgensen T, Haile J, Möller P, Andreev A, Boessenkool S, Rasmussen M, Kienast F, Coissac E, Taberlet P, Brochmann C, Bigelow NH, Andersen K, Orlando L, Gilbert MTP, Willerslev E (2012) A comparative study of ancient sedimentary DNA, pollen and macrofossils from permafrost sediments of northern Siberia reveals long-term vegetational stability. Mol. Ecol. 21, 1989–2003. https://doi.org/10.1111/j.1365-294x.2011.05287.x

Kromer B (2009) Radiocarbon and dendrochronology. Dendrochronologia 27, 15–19. https://doi.org/10.1016/j.dendro.2009.03.001

Larsen G, Eiríksson J, Knudsen KL, Heinemeier J (2002) Correlation of late Holocene terrestrial and marine tephra markers, north Iceland: implications for reservoir age changes. Polar Res. 21, 283–290. https://doi.org/10.1111/j.1751-8369.2002.tb00082.x

Lejzerowicz F, Esling P, Majewski W, Szczuciński W, Decelle J, Obadia C, Arbizu PM, Pawlowski J (2013) Ancient DNA complements microfossil record in deep-sea subsurface sediments. Biol. Lett. 9, 20130283. https://doi.org/10.1098/rsbl.2013.0283

Masselink G, Hughes M, Knight J (2014) Introduction to Coastal Processes and Geomorphology.

Monchamp M-E, Walser J-C, Pomati F, Spaak P (2016) Sedimentary DNA Reveals Cyanobacterial Community Diversity over 200 Years in Two Perialpine Lakes. Appl. Environ. Microbiol. 82, 6472–6482. https://doi.org/10.1128/AEM.02174-16

Murchie TJ, Kuch M, Duggan AT, Ledger ML, Roche K, Klunk J, Karpinski E, Hackenberger D, Sadoway T, MacPhee R, Froese D, Poinar H (2020) Optimizing extraction and targeted capture of ancient environmental DNA for reconstructing past environments using the PalaeoChip Arctic-1.0 bait-set. Quaternary Research 1–24. https://doi.org/10.1017/qua.2020.59

Niemeyer B, Epp LS, Stoof-Leichsenring KR, Pestryakova LA, Herzschuh U (2017) A comparison of sedimentary DNA and pollen from lake sediments in recording vegetation composition at the Siberian treeline. Mol. Ecol. Resour. 17, e46–e62. https://doi.org/10.1111/1755-0998.12689

Overballe-Petersen S, Willerslev E (2014) Horizontal transfer of short and degraded DNA has evolutionary implications for microbes and eukaryotic sexual reproduction. Bioessays 36, 1005–1010. https://doi.org/10.1002/bies.201400035

Pansu J, Giguet-Covex C, Ficetola GF, Gielly L, Boyer F, Zinger L, Arnaud F, Poulenard J, Taberlet P, Choler P (2015) Reconstructing long-term human impacts on plant communities: an ecological approach based on lake sediment DNA. Mol. Ecol. 24, 1485–1498. https://doi.org/10.1111/mec.13136

Parducci L, Alsos IG, Unneberg P, Pedersen MW, Han L, Lammers Y, Salonen JS, Väliranta MM, Slotte T, Wohlfarth B (2019) Shotgun environmental DNA, pollen, and macrofossil analysis of lateglacial lake sediments from southern sweden. Front. Ecol. Evol. 7. https://doi.org/10.3389/fevo.2019.00189

Parducci L, Bennett KD, Ficetola GF, Alsos IG, Suyama Y, Wood JR, Pedersen MW (2017) Ancient plant DNA in lake sediments. New Phytol. 214, 924–942. https://doi.org/10.1111/nph.14470

Parducci L, Nota K, Wood J (2018) Reconstructing Past Vegetation Communities Using Ancient DNA from Lake Sediments, in: Lindqvist, C., Rajora, O.P. (Eds.), Paleogenomics: Genome-Scale Analysis of Ancient DNA, Population Genomics. Springer International Publishing, Cham, pp. 163–187. https://doi.org/10.1007/13836_2018_38

Pedersen MW, Ginolhac A, Orlando L, Olsen J, Andersen K, Holm J, Funder S, Willerslev E, Kjær KH (2013) A comparative study of ancient environmental DNA to pollen and macrofossils from lake sediments reveals taxonomic overlap and additional plant taxa. Quat. Sci. Rev. 75, 161–168. https://doi.org/10.1016/j.quascirev.2013.06.006

Pedersen MW, Overballe-Petersen S, Ermini L, Sarkissian CD, Haile J, Hellstrom M, Spens J, Thomsen PF, Bohmann K, Cappellini E, Schnell IB, Wales NA, Carøe C, Campos PF, Schmidt AM, Gilbert MT, Hansen AJ, Orlando L, Willerslev E (2015) Ancient and modern environmental DNA. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 370, 20130383. Doi: 10.1098/rstb.2013.0383

Pedersen MW, Ruter A, Schweger C, Friebe H, Staff RA, Kjeldsen KK, Mendoza MLZ, Beaudoin AB, Zutter C, Larsen NK, Potter BA, Nielsen R, Rainville RA, Orlando L, Meltzer DJ, Kjær KH, Willerslev E (2016) Postglacial viability and colonization in North America's ice-free corridor. Nature 537, 45–49. https://doi.org/10.1038/nature19085

Poinar HN, Hofreiter M, Spaulding WG, Martin PS, Stankiewicz BA, Bland H, Evershed RP, Possnert G, Pääbo S (1998) Molecular coproscopy: dung and diet of the extinct ground sloth Nothrotheriops shastensis. Science 281, 402–406. https://doi.org/10.1126/science.281.5375.402

Reimer PJ, Austin WEN, Bard E, Bayliss A, Blackwell PG, Bronk Ramsey C, Butzin M, Cheng H, Edwards RL, Friedrich M, Grootes PM, Guilderson TP, Hajdas I, Heaton TJ, Hogg AG, Hughen KA, Kromer B, Manning SW, Muscheler R, Palmer JG, Talamo S (2020) The IntCal20 Northern Hemisphere radiocarbon age calibration curve (0–55 cal kBP). Radiocarbon 1–33. https://doi.org/10.1017/RDC.2020.41

Rohland N, Glocke I, Aximu-Petri A, Meyer M (2018) Extraction of highly degraded DNA from ancient bones, teeth and sediments for high-throughput sequencing. Nat. Protoc. 13, 2447–2461. https://doi.org/10.1038/s41596-018-0050-5

Schulte L, Bernhardt N, Stoof-Leichsenring K, Zimmermann HH, Pestryakova LA, Epp LS, Herzschuh U (2020) Hybridization capture of larch (Larix Mill) chloroplast genomes from sedimentary ancient DNA reveals past changes of Siberian forest. Mol. Ecol. Resour. https://doi.org/10.1111/1755-0998.13311

Seeber PA, McEwen GK, Löber U, Förster DW, East ML, Melzheimer J, Greenwood AD (2019) Terrestrial mammal surveillance using hybridization capture of environmental DNA from African waterholes. Mol. Ecol. Resour. 19, 1486–1496. https://doi.org/10.1111/1755-0998.13069

Sjögren P, Edwards ME, Gielly L, Langdon CT, Croudace IW, Merkel MKF, Fonville T, Alsos IG (2017) Lake sedimentary DNA accurately records 20th Century introductions of exotic conifers in Scotland. New Phytol. 213, 929–941. https://doi.org/10.1111/nph.14199

Slon V, Hopfe C, Weiß CL, Mafessoni F, de la Rasilla M, Lalueza-Fox C, Rosas A, Soressi M, Knul MV, Miller R, Stewart JR, Derevianko AP, Jacobs Z, Li B, Roberts RG, Shunkov MV, de Lumley H, Perrenoud C, Gušić I, Kućan Ž, Meyer M (2017) Neandertal and Denisovan DNA from Pleistocene sediments. Science 356, 605–608. https://doi.org/10.1126/science.aam9695

Smith O, Momber G, Bates R, Garwood P, Fitch S, Pallen M, Gaffney V, Allaby RG (2015) Archaeology. Sedimentary DNA from a submerged site reveals wheat in the British Isles 8000 years ago. Science 347, 998–1001. https://doi.org/10.1126/science.1261278

Taberlet P, Bonin A, Zinger L, Coissac E (Eds) (2018) Environmental DNA: For Biodiversity Research and Monitoring.

Taberlet P, Prud'Homme SM, Campione E, Roy J, Miquel C, Shehzad W, Gielly L, Rioux D, Choler P, Clément J-C, Melodelima C, Pompanon F, Coissac E (2012) Soil sampling and isolation of extracellular DNA from large amount of starting material suitable for metabarcoding studies. Mol. Ecol. 21, 1816–1820. https://doi.org/10.1111/j.1365-294X.2011.05317.x

Torti A, Lever MA, Jørgensen BB (2015) Origin, dynamics, and implications of extracellular DNA pools in marine sediments. Mar. Genomics 24 Pt 3, 185–196. https://doi.org/10.1016/j.margen.2015.08.007

Vasan S, Zhang X, Zhang X, Kapurniotu A, Bernhagen J, Teichberg S, Basgen J, Wagle D, Shih D, Terlecky I, Bucala R, Cerami A, Egan J, Ulrich P (1996) An agent cleaving glucose-derived protein crosslinks in vitro and in vivo. Nature 382, 275–278. https://doi.org/10.1038/382275a0

Vernot B, Zavala EI, Gómez-Olivencia A, Jacobs Z, Slon V, Mafessoni F, Romagné F, Pearson A, Petr M, Sala N, Pablos A, Aranburu A, de Castro JMB, Carbonell E, Li B, Krajcarz MT, Krivoshapkin AI, Kolobova KA, Kozlikin MB, Shunkov MV, Meyer M (2021) Unearthing Neanderthal population history using nuclear and mitochondrial DNA from cave sediments. Science 372. https://doi.org/10.1126/science.abf1667

Wen F, Curlango-Rivera G, Huskey DA, Xiong Z, Hawes MC (2017) Visualization of extracellular DNA released during border cell separation from the root cap. Am. J. Bot. 104, 970–978. https://doi.org/10.3732/ajb.1700142

Willerslev E, Cappellini E, Boomsma W, Nielsen R, Hebsgaard MB, Brand TB, Hofreiter M, Bunce M, Poinar HN, Dahl-Jensen D, Johnsen S, Steffensen JP, Bennike O, Schwenninger J-L, Nathan R, Armitage S, de Hoog C-J, Alfimov V, Christl M, Beer J, Collins MJ (2007) Ancient biomolecules from deep ice cores reveal a forested southern Greenland. Science 317, 111–114. https://doi.org/10.1126/science.1141758

Willerslev E, Cooper A (2005) Ancient DNA. Proc. Biol. Sci. 272, 3–16. https://doi.org/10.1098/rspb.2004.2813

Willerslev E, Davison J, Moora M, Zobel M, Coissac E, Edwards ME, Lorenzen ED, Vestergård M, Gussarova G, Haile J, Craine J, Gielly L, Boessenkool S, Epp LS, Pearman PB, Cheddadi R, Murray D, Bråthen KA, Yoccoz N, Binney H, Taberlet P (2014) Fifty thousand years of Arctic vegetation and megafaunal diet. Nature 506, 47–51. https://doi.org/10.1038/nature12921

Willerslev E, Hansen AJ, Binladen J, Brand TB, Gilbert MTP, Shapiro B, Bunce M, Wiuf C, Gilichinsky DA, Cooper A (2003) Diverse plant and animal genetic records from Holocene and Pleistocene sediments. Science 300, 791–795. https://doi.org/10.1126/science.1084114

Willerslev E, Hansen AJ, Rønn R, Brand TB, Barnes I, Wiuf C, Gilichinsky D, Mitchell D, Cooper A (2004) Long-term persistence of bacterial DNA. Curr. Biol. 14, R9–R10. https://doi.org/10.1016/j.cub.2003.12.012

Wilmshurst JM, Moar NT, Wood JR, Bellingham PJ, Findlater AM, Robinson JJ, Stone C (2014) Use of pollen and ancient DNA as conservation baselines for offshore islands in New Zealand. Conserv. Biol. 28, 202–212. https://doi.org/10.1111/cobi.12150

Zavala EI, Jacobs Z, Vernot B, Shunkov MV, Kozlikin MB, Derevianko AP, Essel E, de Fillipo C, Nagel S, Richter J, Romagné F, Schmidt A, Li B, O'Gorman K, Slon V, Kelso J, Pääbo S, Roberts RG, Meyer M (2021) Pleistocene sediment DNA reveals hominin and faunal turnovers at Denisova Cave. Nature. https://doi.org/10.1038/s41586-021-03675-0

Zimmermann HH, Raschke E, Epp LS, Stoof-Leichsenring KR, Schirrmeister L, Schwamborn G, Herzschuh U (2017) The History of Tree and Shrub Taxa on Bol'shoy Lyakhovsky Island (New Siberian Archipelago) since the Last Interglacial Uncovered by Sedimentary Ancient DNA and Pollen Data. Genes (Basel) 8. https://doi.org/10.3390/genes8100273

Zimmermann HH, Stoof-Leichsenring KR, Kruse S, Müller J, Stein R, Tiedemann R, Herzschuh U (2020) Changes in the composition of marine and sea-ice diatoms derived from sedimentary ancient DNA of the eastern Fram Strait over the past 30 000 years. Ocean Sci. 16, 1017–1032. https://doi.org/10.5194/os-16-1017-2020

Zinger L, Chave J, Coissac E, Iribar A, Louisanna E, Manzi S, Schilling V, Schimann H, Sommeria-Klein G, Taberlet P (2016) Extracellular DNA extraction is a fast, cheap and reliable alternative for multi-taxa surveys based on soil DNA. Soil Biology and Biochemistry 96, 16–19. https://doi.org/10.1016/j.soilbio.2016.01.008

Answers

1. Possible advantages of *sed*aDNA compared to pollen as a proxy for past plant presence are: the possibility of detecting past plant presence even in the absence of visible remains; less labour-intensive as taxonomic identification is automated; in principle, no prior taxonomic knowledge is needed for the data generation with *sed*aDNA (although it is highly called for in the interpretation of the data); and it is possible to obtain a higher taxonomic resolution depending on the choice of marker.
2. For mineral-rich sediments, luminescence dating can be used as this method can be applied to sediments from a few decades old to over a million years old, and is based on the phenomenon that mineral crystals absorb electrons from ionising radiation of surrounding sediments over time. For sediment rich in organic materials, AMS radiocarbon dating of identified macroscopic remains (with calibration) is a good option. Radiocarbon dating is based on the concentration of C^{14} in organismic remains. The half-life of C^{14} (5730 years) makes it an appropriate method for samples under 50,000 years old. To increase confidence in the dating results, multiple dating techniques could be used for creating an age model for the core.
3. Biases when working with *sed*aDNA can come from: taphonomic processes including differential DNA degradation and preservation, choice of metabarcoding primers, completeness of reference library, and contamination during sampling, DNA extraction and other lab processes. False positives can be limited by inclusion of multiple replicates and controls and prevention of contamination at every step of the experimental design, preparation of an appropriate reference database, and checking if the identifications fit with what is known for the age and location of the sample by a taxonomic expert.

SECTION 2

Methods

Chapter 9
Sequencing platforms

Marcella Orwick Rydmark[1], Yannick Woudstra[2,3,4,5], Hugo de Boer[1]

1 Natural History Museum, University of Oslo, Norway
2 Royal Botanic Gardens, Kew, United Kingdom
3 Natural History Museum Denmark, University of Copenhagen, Copenhagen, Denmark
4 Gothenburg Global Biodiversity Center, Department of Biological and Environmental Sciences, University of Gothenburg, Gothenburg, Sweden
5 Department of Plant Sciences, University of Oxford, Oxford, United Kingdom

Marcella Orwick Rydmark **m.c.o.rydmark@nhm.uio.no**
Yannick Woudstra **yannickwoudstra@outlook.com**
Hugo de Boer **h.de.boer@nhm.uio.no**

Citation: Rydmark MO, Woudstra Y, de Boer H (2022) Chapter 9. Sequencing platforms. In: de Boer H, Rydmark MO, Verstraete B, Gravendeel B (Eds) Molecular identification of plants: from sequence to species. Advanced Books. https://doi.org/10.3897/ab.e98875

Introduction

The revolution in genome-wide screening has vastly reduced the price for sequencing, with enormous implications in the biomedical field, industry, biodiversity monitoring, as well as in plant identification. The first plant genome (*Arabidopsis thaliana* L.) was sequenced using Sanger sequencing. This took 10 years to complete with an associated cost of approximately $100,000,000 (Arabidopsis Genome Initiative 2000). With current high-throughput sequencing (HTS) methods, this same genome now takes 1 week to sequence and assemble, and costs $1000 (Michael et al. 2018). Plant genomes and DNA sequences are however under-represented in the literature in comparison to other organisms such as microorganisms and animals, and most reported plant genomes belong to angiosperms with relatively small genomes. This is due to a number of confounding factors that make the sequencing of plant genomes particularly difficult including the extraction of sufficient quantities of high-quality DNA that is not irreparably damaged (Inglis et al. 2018) (see Chapter 1 DNA from plant tissue), the size and complexity of the genome (i.e., gene islands, high GC content, transposable elements), heterozygosity, and polyploidy (Chen et al. 2018) (see Chapter 16 Whole genome sequencing). Advances in high molecular weight DNA extraction, high throughput sequencing technologies, and bioinformatics approaches to deal with heterozygosity and assembly in recent years have alleviated these challenges tremendously, with enormous strides in the generation of high quality genomic datasets to test research hypotheses. In this chapter, we review the different sequencing technologies most commonly utilised today, beginning with Sanger sequencing, the current method of choice for small-scale projects. We then discuss HTS approaches, including next-generation sequencing and third-generation PCR-free sequencing methods (van Dijk et al. 2018). The underlying principles of these technologies are discussed and linked to their advantages and disadvantages in considering which method is best suited for different sequencing projects.

Sequencing platforms

Sanger sequencing

Sanger sequencing was introduced in 1977 by Sanger and colleagues, and for over 40 years, it was the most commonly-used form of sequencing (Heather and Chain 2016). Sanger sequencing is a PCR-based technique that uses chain-terminating fluorescently-labelled dideoxyribonucleotides (ddNTPs) to determine the sequence of Polymerase Chain Reactionamplified target DNA. The target DNA is mixed with both standard deoxyribonucleotide triphosphates (dNTPs) and a much lower concentration (around 1%) of four differently labelled fluorescent ddNTPs (ddATP, ddTTP, ddCTP, and ddGTP), that correspond to the 4 different dNTPs. The ddNTPs lack the chemical 3′-OH group that is required for phosphodiester bond formation. Thus, in the PCR reaction, when a fluorescently labelled ddNTP is added, the polymerase can no longer add another dNTP, and extension ceases. This results in chain termination with oligonucleotide copies of the target DNA terminated at random lengths (up to 1000 bp) by the fluorescently-labelled ddNTPs (Hagemann 2015).

In the second step of Sanger sequencing, the oligonucleotides are separated by size using capillary gel electrophoresis. A laser excites the terminal fluorescent nucleotide in each oligonucleotide, resulting in fluorescence emission that is detected and read by a computer. By reading the gel bands from smallest to largest, the 5′ to 3′ sequence of the target DNA can be determined at single base pair resolution. The data output for Sanger sequencing is a chromatogram

which is automatically read by a computer to generate the DNA sequence. Primer sequences should be trimmed off the reads as these are not part of the target DNA, and the quality of the chromatogram should be assessed to determine the reliability of the generated DNA sequence. There are a number of online tutorials from both industrial and academic sources that we refer the reader to for assessing a chromatogram quality (University of Michigan, Biomedical Research Core Facilities, n.d.). Base calling accuracy can also be measured using Phred quality scores (Ewing and Green 1998; Ewing et al. 1998). A Phred quality score indicates the quality of an oligonucleotide assignment that is generated during DNA sequencing. A Phred score of 20 indicates 99% accuracy in the assignment, which is generally considered acceptable.

Sanger sequencing is not used today for large-scale genomic projects due its low throughput. The requirement of needing specific primers for a region of interest limits its easy use and application across divergent plant taxa. Additionally, the amplification of multicopy genes, such as the commonly used DNA barcode ITS (see Chapter 10 DNA barcoding), as well as markers in taxa of allopolyploid hybrid origin, result in difficult-to-interpret chromatograms. This is because nucleotide polymorphisms between different copies result in double peaks in the resulting chromatogram (Hughes et al. 2013).Nevertheless, it is still a widely-used technique for smaller scale projects on DNA barcoding and phylogenetics, especially incremental studies where new sequences are added to existing phylogenetic frameworks. Additionally, due to its high accuracy and relatively long reads, Sanger sequencing is also sometimes used in conjunction with HTS techniques that may have shorter reads and/or higher error rates to aid in the proper assembly of contigs and check the accuracy of the final sequence (Slatko et al. 2018).

Illumina sequencing

Illumina was the second HTS technique that became commercially available in the early 2000s (Heather and Chain 2016; McGinn and Gut 2013). Illumina was preceded by the Roche 454 pyrosequencing by synthesis sequencing, though this system has since been discontinued (Edwards et al. 2006; Thomas et al. 2012). As in Sanger sequencing, Illumina also uses fluorescently-labelled dNTPs though in Illumina sequencing they do not permanently block further synthesis of a growing nucleotide strand. Additionally, Illumina sequencing is done in an enormously parallel fashion. This results in dramatic time and cost reductions compared to Sanger sequencing for large-scale genomic projects. Today, Illumina, along with PacBio and Nanopore technologies, is one of the most widely used technologies for large-scale genomic projects (van Dijk et al. 2018).

In Illumina sequencing, like in other high throughput sequencing approaches, the target DNA is initially broken into shorter fragments that match the optimal fragment sequencing length of the platform, if not already present as shorter segments. These fragments are then PCR-amplified with adaptors that can be individually chemically tethered to the flow cell surface. Using bridge amplification (Clark et al. 2018), these fragments are amplified to form millions of dense clusters of DNA strands in the flow cell, as the platform cannot read single DNA strands but needs thousands of identical strands for accurate base calling. After this initial amplification step, fluorescently-labelled dNTPs are added to the flow cells. Depending on the synthesis by sequencing technology four, two or one dyes are added respectively. The newer NovaSeq, NextSeq 550, and MiniSeq platforms use the faster two dye two-channel technology.

Dyed dNTPs are added in a controlled fashion through the use of reversible blocking group chemistry, so that the emission of each added fluorescent dNTP is read before the addition of the next fluorescently-labelled dNTP. This process is done on millions of fragments simultaneously, making it a far more efficient method than Sanger sequencing for large-scale genomic projects (Slatko et al. 2018). In addition to large-scale genomic projects, Illumina is also

an important technology for gene-targeted applications, including barcoding (see Chapter 10 DNA barcoding), metabarcoding (see Chapter 11 Amplicon metabarcoding), and target capture (see Chapter 14 Target capture). This is because these methods include library preparations where using HTS methods such as Illumina sequencing offers major time advantages in comparison to Sanger sequencing (Head et al. 2014).

Two limitations to consider with Illumina sequencing however are that the produced reads are relatively short (50 to 300 bp), and similarly to Sanger sequencing, most applications require a PCR amplification step. However, PCR free library kits and protocols provide increasingly good results, and have the important advantage of reducing typical PCR-induced biases. Assembling whole genomes using short read Illumina methods, especially if they are highly repetitive, can be challenging (Kyriakidou et al. 2018). As well, the requirement for a PCR amplification step introduces the possibility for bias in mixed samples (i.e., DNA from different sources may be amplified to different degrees) (Aird et al. 2011). Nevertheless, its high throughput and accurate reads make Illumina the standard choice for sequencing amplicon libraries and genome resequencing. It is also the approach of choice for degraded herbarium material where long-read platforms would bring no benefits. In addition, Illumina's MiSeq and Miniseq instruments offer desktop solutions that produce long reads in comparison to other illumina platforms (up to 300 bp) with integrated software for data analysis (Twyford 2016). Thus, for projects more focused on targeted gene sequencing (for example amplicon barcoding, metabarcoding, and target enrichment) this platform offers an in-house integrated solution with rapid turnaround times (Ravi et al. 2018). Scaling the sequencing needs of your project to the sequencing platform is essential to avoid obtaining either too little or too much data. Many metabarcoding projects require relatively little reads to obtain all available OTUs with high numbers of reads per cluster, so costs could be optimised by combining libraries with projects that have higher output demands. Multiplexing different samples can be an efficient – and in most cases essential – way to optimise the output from a sequencing run. Samples for metabarcoding require relatively few little reads, whereas target-capture, genome skimming or deep sequencing require more sequencing depth for their applications. Care needs to be taken when multiplexing samples of different fragments lengths, such as for example metabarcoding amplicons of 150 and 300 bp, as the shorter fragments will be preferentially sequenced and thus dominate results. The same applies to normalisation of individual samples in a mixed library, where combining samples with expected low concentrations or highly degraded DNA templates will yield fewer reads when mixed with high concentration DNA. As a rule combining old and new DNA in the same library is avoided, as well as long and short amplicons. Separating these samples will necessitate extra sequencing runs, but yield the best results. Finally, it is worth mentioning that some of the super high-throughput platforms like the Novaseq S4 yield more data than necessary for many applications, or at least are hard to fill with sufficient multiplexed samples to make it worth the compounded risk of combining a large number of samples.

Table 1. Current examples of Illumina sequencing platforms, specifications, and suitability for different applications in plant identification.

Illumina sequencing platform	MiSeq	HiSeq 2500*	HiSeq 3000*	HiSeq 4000*	NextSeq 1000 and 2000	NovaSeq 6000
Specifications						
Maximum read length (pair ended)	2 x 300	2 x 250	2 x 150	2 x 150	2 x 150	2 x 250
Maximum reads per run (single reads)	25 million	600 million	2.5 billion	5 billion	1.1 billion	20 billion
Flow Cell output	15 Gb	300 Gb	750 Gb	1.5 Tb	330 Gb	6 Tb

Illumina sequencing platform	MiSeq	HiSeq 2500*	HiSeq 3000*	HiSeq 4000*	NextSeq 1000 and 2000	NovaSeq 6000
Method suitability						
Metabarcoding	+++	+++	+	+	+	++
Target Capture	+	+	+	+++	+	+++
Shotgun sequencing	+	++	+++	+++	++	+++
Genome skimming	+	++	+++	+++	++	+++
Organellar sequencing (plastids)	+	++	+++	+++	++	++
Transcriptomics:gene targeted	+++	+++	+	+	+	++
Transcriptomics: total RNA/mRNA seq	+	+	++	++	++	+++

*The HiSeq series has been discontinued but is still widely available. We thus include the general specifications and which applications they are best suited for.

Pacific Biosciences

Pacific Biosciences (PacBio) sequencing is based on single molecule real time (SMRT) technologies for reading DNA and RNA sequences. No PCR amplification is required, which for certain applications can be advantageous. This includes if PCR inhibitors are/may be present, the sequence is GC rich, or if PCR bias should be avoided. Additionally, PacBio reads are considerably longer than in either Sanger or Illumina sequencing (up to 25 kb) ((Pacific Biosciences, n.d.). This reduces computational challenges related to assembling contigs into full sequences. PacBio is considered a third generation sequencing technology, as it reads the nucleotide sequence both in real-time and at the single molecule level (Amarasinghe et al. 2020).

Similarly to Illumina and Sanger sequencing, PacBio also uses fluorescently-labelled dNTPs for determining a target DNA sequence. PacBio however employs a technology called zero mode waveguides (ZMW) to read nucleotide sequences at the single molecule level. ZMWs are nanosized wells that can be etched into different materials, with attoliter (10^{-21} L) volumes. ZMW technology differentiates a fluorescent molecule that is floating in solution from a fluorescently-labelled nucleotide that is located at the bottom of the well. A single DNA polymerase is tethered to the bottom of each well, and when a fluorescently-labelled dNTP is incorporated into the growing DNA strand, the fluorescent label is cleaved off. There is a unique fluorescent marker for each of the 4 nucleotides, and each cleavage event is read and directly linked to a specific nucleotide (van Dijk et al. 2018). Additionally, the rate of addition can be used to infer whether the target DNA is modified (i.e., post-translationally phosphorylated or methylated), since a modified DNA strand moves more slowly through the DNA polymerase, resulting in a reduced incorporation rate for a fluorescent nucleotide. This information is extremely powerful for predicting epigenetic modifications that are critical for a variety of biological functions. In addition, chemical modifications that are often present in aDNA can also be detected, making PacBio a particularly useful technique for assessing aDNA damage (Flusberg et al. 2010) (See Chapter 8 aDNA from sediments).

While previously PacBio suffered from a high error rate in comparison to Illumina sequencing, this has been dramatically reduced by the introduction of circular consensus sequencing (CCS), also known as long high-fidelity (HiFi) reads (Eid et al. 2009). In circular consensus sequencing, the ends of a DNA strand are ligated together to circularise it. This DNA template strand is called a SMRTbell. This circularization allows for multiple reads of the same

sequence (so long as the strand is not too long) through a DNA polymerase, dramatically reducing the error rate, and can provide read lengths up to 25 kB. In one recent study, large gene fragments (circa 40 kB) were read with up to 99.91% accuracy when CCS was combined with a carefully optimised protocol for the handling of DNA to reduce any fragmentation/nicks. (Wenger et al. 2019). In addition to CCS, continuous long read (CLR) techniques are especially useful for gene assemblies (Vollger et al. 2020). CLR lengths are approximately equivalent to the polymerase read length. The sequence is generated from a single continuous template from start to finish, thus emphasising the longest read possible (up to 175 kB for CLR vs. 25 kB for CCS), though the overall CLR accuracy is lower than with CLS reads (90% vs. 99% read accuracy).

Oxford Nanopore

Oxford Nanopore (or simply Nanopore) sequencing is also a third generation SMRT technology that is single-molecule based and measured in real time. Nanopore is unique from the other sequencing technologies discussed here in that no DNA polymerase is required, and no expensive chemically modified dNTPs are necessary for reading the target sequence. The system consists of an electrolytic solution and a nanosized, biologically-derived pore in an insulating solid (a material that does not conduct electricity). The biological nanopores used in this technology are derived from proteins that form pores in biological membranes that naturally function to allow for the passage of ions and biomolecules across the membrane. When an electric field is applied, ions in the electrolytic solution pass through the pore, resulting in a stable current that can be detected. When larger molecules pass through the pore, such as DNA strand, detectable disruptions in the current occur. With a DNA strand, sequences of 6–7 nucleotides move through the pore and the movement of these bases yield a changing detectable disruption. This disruption has a unique signature with a specific current change for a specific length of time that can be linked to each of the four individual nucleotides. From the current disruption pattern it is possible to deduce the sequence. As well, since it is the change in current through the pore that is detected, no other chemical markers are necessary (Jain et al. 2016; Kono and Arakawa 2019). This is an important advantage over the other technologies discussed, where the fluorescently-labelled nucleotides are expensive.

Nanopore technologies, with a read length up to 4 Mb, are rapidly becoming important due to their scalability and portability. The MinION sequencing platform (theoretical output up to 50 Gb/flow cell) is a portable and cost-effective option (87 g, available from $1000) that can be used in the field. Already, a number of excellent examples of biodiversity studies (and plant-based studies in particular) are available in the literature (Bethune et al. 2019; Maestri et al. 2019; Srivathsan et al. 2021). As well, the GridION (theoretical output up to 50 Gb/flow cell) and PromethION (theoretical output up to 290 Gb/flow cell) are both desktop-sized sequencing platforms for mid and high-throughput data generation and analysis for in-house sequencing projects. A historical drawback with Nanopore technologies is the high error rate, with a reported raw read error rate between 10 and 22% (Kono and Arakawa 2019; Krehenwinkel et al. 2019). One method to overcome this is rolling circle amplification (RCA). For sequencing experiments, a linear single-stranded DNA molecule is firstly circularised and then copied multiple times as a single sequence (Johne et al. 2009). Thus, the same DNA sequence may be read multiple times using Nanopore technologies, with resulting read accuracies as high as 99.3% (Baloğlu et al. 2020). RCA combined with neural network and machine learning approaches such as Guppy, Bonito, Sacall, SquiggleNet, DeepNano-blitz, can raise base calling accuracies even further (Wick et al. 2019; Bao et al. 2021; Boža et al. 2020; Huang et al. 2022; (Vereecke et al. 2020).

Chapter 9: Box 1. Library preparation - tips and considerations

Library preparations are essential for all experiments involving HTS. General points to consider are discussed here and we also refer to Chapter 12 Metagenomics and Chapter 15 Transcriptomics for more details.

i) DNA fragmentation. Short-read Illumina sequencing requires target DNA in the correct size range (50–600 bp, depending on the specific platform). High molecular weight DNA can be sheared either with ultrasonication (e.g., Covaris platforms) or (more economically) with library preparation kits that incorporate a fragmentase enzyme. However, fragmentase activity is highly dependent on genome organisation, and may require optimisation for each analysed species. Different fragment lengths may be considered if the desired study requires long-read sequencing.

ii) Input DNA quality assessment. After DNA extraction using a tissue-specific protocol (see section 1 of this book), the quality of the target DNA needs to be assessed. Fragment length distribution is an important consideration to produce libraries with even DNA fragment size distributions. Fresh or silica-dried plant material may yield high molecular weight DNA which can be checked visually using agarose gel electrophoresis, but DNA isolated from herbarium material is often highly fragmented and this requires careful inspection to decide on the optimal fragmentation protocol (see below) (Chapter 1 DNA from plant tissue). Problematic samples can be assessed with a high-precision automated electrophoresis tool (e.g., Agilent TapeStation, Bioanalyzer or Fragment Analyzer).

iii) Library preparation. The library preparation protocols depend on the sequencing platform being used as well as the sort of experiment being performed (e.g., metabarcoding, target capture, or metagenomics) For Illumina platforms, dual-indexed libraries can be generated with Illumina TruSeq (Illumina), third-party kits such as NEBNext Ultra II, or non-kit based protocols (Meyer and Kircher 2010; Troll et al. 2019), including protocols for degraded DNA (Troll et al. 2019). It is often possible to use half-volume reactions with these kits to reduce the per-sample cost without significant yield loss. Steps involved in library preparation often involve the trimming of fragment termini (necessary for target capture), ligation of adaptor sequences, optimization of the library fragment size, and the addition of unique index sequences through PCR amplification using multiplex primers. This can be done using single index sequences on one side of the fragment (up to 12 samples) or using dual index sequences where different index sequences are added to each side of the fragment. This last step is also important for bringing the library concentration up to an acceptable level again, as the number of DNA fragments in the library is diminished significantly during size selection.

Ion Torrent

Unlike in other forms of sequencing, Ion Torrent technologies are not based upon optical outputs, but rather on changes in pH. When a DNA polymerase adds a nucleotide to a growing DNA strand, a proton is released upon each addition. It is this release of protons into solution, and the resulting change in the pH of the solution, that is detected in Ion Torrent technologies (Rothberg et al. 2011; Slatko et al. 2018).

Similarly to Illumina sequencing, the target DNA is initially fragmented (200–600 bps) and PCR-amplified with adaptors that can be tethered to micro-machined wells on a semiconductor chip. The plates are then flooded with one of the 4 nucleotides. If a nucleotide is added across

from the complementary base in the single-stranded DNA by the DNA polymerase, it results in the release of a proton and a subsequent change in solution pH. This shift in solution pH is detected by an ion-sensitive field-effect transistor (ISEFT), which can detect changes in proton concentration. This is done in a massively parallel fashion, with 1000s of microwell plates being used simultaneously. The pH change that results from the addition of multiple nucleotides in a repetitive sequence is also detectable using this technology, as the addition of two nucleotides will result in double the voltage change as the addition of a single nucleotide. The data output with Ion Torrent technologies can provide an approximate readout of 10 MBb in a single run with conventional machines, and up to 10 GBb with the newest models. The platform however struggles with base calling of homopolymers, and for these sequences it can be a challenge to obtain accurate reads.

The Ion Torrent machine and sequencing chips are relatively inexpensive compared to Illumina and PacBio, and this made it popular in smaller labs without access to high throughput sequencing core facility sequencing, though its use is no longer as common.

Which sequencing platform?

The sequencing platform that is ultimately chosen by a scientist depends on a number of factors. This can include (but is not limited to) the scientific question being considered, the quality of target DNA (see Chapter 1 DNA from plant tissue), costs, as well as in-house expertise and/or availability of existing platforms. In all cases, however, the quality and sequencing depth of target DNA should be considered. For DNA that is primarily expected to exist in shorter sequences (i.e., samples that are expected to be degraded from herbarium or ancient sources), then technologies requiring long reads are often not necessary, and Illumina sequencing or Ion Torrent technologies may be sufficient. If however one wishes to avoid any PCR bias or acquire long reads, then using PacBio or Nanopore is advisable. Finally, it may even be useful to use

Table 2. Sequencing platform choices for different experimental questions and sample types.

Experiment or sample considerations	Recommended method(s)	Comments
Whole genome or organellar sequencing project (genome skimming, genome resequencing, de novo genome assembly)	Illumina, PacBio, or a combination of both	Illumina is the method of choice for resequencing for high throughput short read projects due to its high read accuracy
Barcoding	Sanger sequencing or PacBio CCS	Larger projects are moving to PacBio CCS to reduce costs. Multiplexing very large numbers of samples is necessary to optimise costs
Metabarcoding/Target capture	Illumina, MGI, DNBSEQ, or Ion Torrent	PacBio and/or Nanopore may also be considered if the sequence is expected to be highly repetitive
Heavily degraded samples (i.e., herbarium or ancient DNA samples)	Illumina (or Ion Torrent)	PacBio may also be relevant for the study of post-genetic modifications often found in ancient DNA samples, or if dealing with hard-to-phase sequences
On-site sequencing	Nanopore (MinION) Hi-C/3C-Seq/ Capture-C (Illumina)	

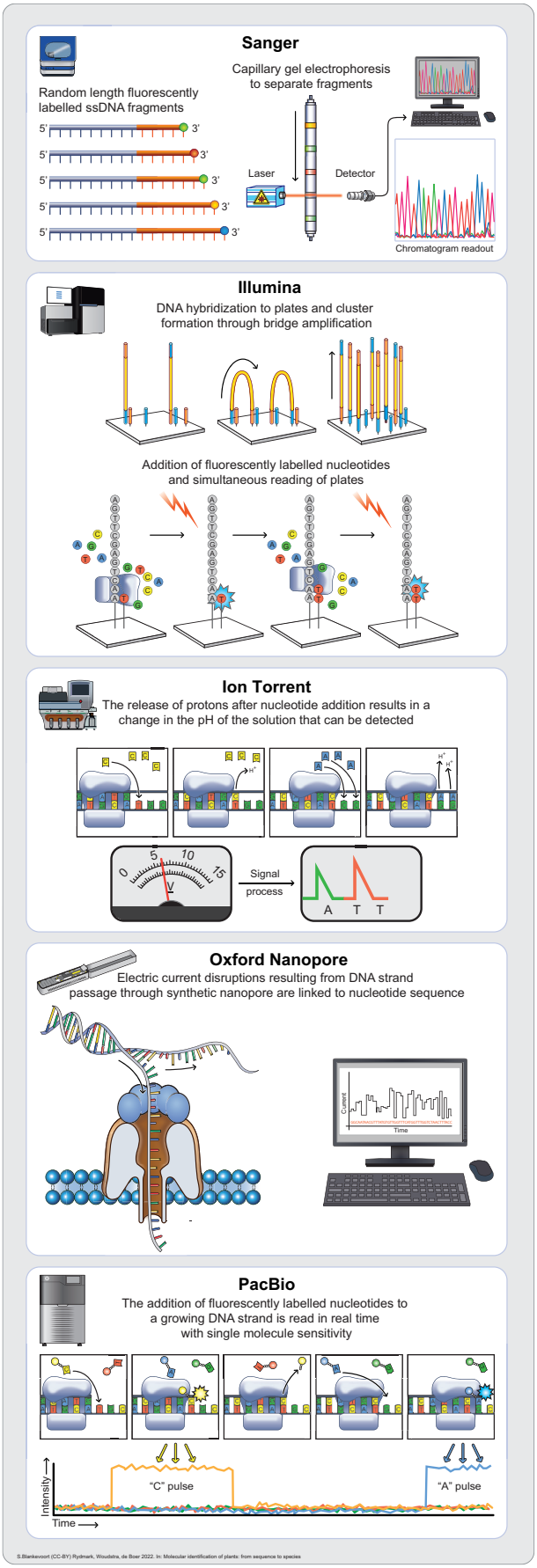

Figure 1. Chapter 9 Infographic: Visual representation of the content of this chapter.

two different types of sequencing to overcome each technology's respective limitations. For example, in whole genome sequencing, hybrid methods combining Illumina with PacBio are commonly used to ensure long reads and high accuracy.

Perspectives

In the last decades, developments in sequencing platforms have primarily focused on increasing the throughput and accuracy of sequencing output, increasing the length of reads, and reducing costs. We can expect the field to continue developing further in this direction, with a focus in particular on the miniaturisation of these platforms for more on-site work, as well as better automation and integration of analytical software and data analysis pipelines. In particular, miniaturisation and automatization of data analysis can be expected to have major impacts in regulatory fields related to both food safety and trade, where the ability for non-specialists to rapidly test on-site for the presence/absence of species will be extremely useful (see Chapter 22 Healthcare and Chapter 23 Food safety). Further development of HTS technologies to be used at the single-cell level and in functional studies can be also expected.

Questions

1. What method(s) are most commonly used for whole genome projects of plants and why? Which sequencing method(s) are most currently most commonly used for library preparations and why?
2. In a scenario where you may want to include amplicons and primers of different length when creating a library for sequencing, how would you adapt your library setup? How would it affect your sequencing costs?
3. Why do Nanopore and PacBio-based technologies provide longer reads than Illumina or Ion torrent-based technologies? Why are these longer reads especially useful for projects in plant identification?

Glossary

Allopolyploid hybrids – A polyploid species with multiple sets of chromosomes that originate from different species. If the hybrid is derived from two diploid species, the resulting tetraploid is fertile. These allopolyploid hybrids may be at least partially reproductively isolated from the parent species from which they are derived, and allopolyploid speciation is the best known route to hybrid speciation in plants.

Bridge amplification – A method used in Illumina sequencing to create DNA clusters with 1000s of double-stranded copies of the target DNA in flow cells. After amplification and generation of these clusters is complete, the reverse strand is washed away and sequencing by synthesis takes place.

Capillary gel electrophoresis (CGE) – An analytical method for the separation of charged molecules. DNA is separated according to size with this technique, with only nanogram quantities

necessary for the input. Single-base pair resolution can be achieved on fragments up to several hundred base pairs in length.

Circular consensus sequencing (CCS) – Developed by PacBio and also known as HiFi reads, involves the circulation of a target DNA strand by ligating the ends of the strand (called a SMRTbell). This SMRTbell can be read multiple times by a DNA polymerase, dramatically reducing the error rate in the generated sequence.

Electrolytic solution – An electrically conductive solution. This conductivity is often due to the presence of ions in solution (for example dissociated Na^+ and Cl^- ions), though non-ionic solutions can also be conductive.

Epigenetic modifications – Alterations in gene expression and cellular function without changes to the original DNA sequence. Three mechanisms for epigenetic modifications so far identified include DNA methylation, histone modification, and non-coding RNA (ncRNA)-associated gene silencing.

Insulating solid – A solid material that an electric current cannot pass through.

Ion-sensitive field-effect transistor (ISEFT) – A field effect transistor that can measure ion concentrations in solution. Changes in the H^+ concentration result in a pH change in solution that results in changes in the current that is detected. This technology is used in Ion Torrent sequencing platforms to identify when a base pair is added to a growing DNA double strand and is the basis for identifying the target DNA sequence.

Phred quality scores – Scores to measure the confidence of the nucleobase identifications generated from DNA sequencing methods. They are widely accepted for assessing the quality of reads.

Rolling circular amplification (RCA) – Where a linear single-stranded DNA molecule is firstly circularised and then copied multiple times as a single sequence (Johne et al. 2009). With the nanopore platform, RCA allows the same DNA sequence to be read multiple times to give a higher read accuracy.

Single molecular real time sequencing (SMRT) – A term coined by PacBio to describe their sequencing technologies. In contrast to second generation sequencing methods, SMRT technologies possess single-molecule sensitivity and provide the sequence readout in real time, dramatically increasing the sensitivity and turnaround times for DNA sequencing.

Zero mode waveguide (ZMW) – Nanosized wells that can be etched into different materials, with attoliter (10^{-21} L) volumes. ZMW technology differentiates a fluorescent molecule that is floating in solution from a fluorescently-labelled nucleotide that is located at the bottom of the well. This technology is used by PacBio for the single-molecule detection of fluorescently-labelled nucleotides that are added to immobilised DNA at the bottom of these wells so that nucleotide incorporation can be detected in real time.

References

Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. 12, R18. https://doi.org/10.1186/gb-2011-12-2-r18

Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q (2020) Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 21, 30. https://doi.org/10.1186/s13059-020-1935-5

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408, 796–815. https://doi.org/10.1038/35048692

Baloğlu B, Chen Z, Elbrecht V, Braukmann T, MacDonald S, Steinke D (2020) A workflow for accurate metabarcoding using nanopore MinION sequencing. BioRxiv. https://doi.org/10.1101/2020.05.21.108852

Bethune K, Mariac C, Couderc M, Scarcelli N, Santoni S, Ardisson M, Martin J-F, Montúfar R, Klein V, Sabot F, Vigouroux Y, Couvreur TLP (2019) Long-fragment targeted capture for long-read sequencing of plastomes. Appl. Plant Sci. 7, e1243. https://doi.org/10.1002/aps3.1243

Chen F, Dong W, Zhang J, Guo X, Chen J, Wang Z, Lin Z, Tang H, Zhang L (2018) The sequenced angiosperm genomes and genome databases. Front. Plant Sci. 9, 418. https://doi.org/10.3389/fpls.2018.00418

Clark DP, Pazdernik NJ, McGehee MR (2018) Molecular Biology, 3rd ed. Academic Cell.

Edwards RA, Rodriguez-Brito B, Wegley L, Haynes M, Breitbart M, Peterson DM, Saar MO, Alexander S, Alexander EC, Rohwer F (2006) Using pyrosequencing to shed light on deep mine microbial ecology. BMC Genomics 7, 57. https://doi.org/10.1186/1471-2164-7-57

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Turner S (2009) Real-time DNA sequencing from single polymerase molecules. Science 323, 133–138. https://doi.org/10.1126/science.1162986

Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8, 186–194.

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 8, 175–185. https://doi.org/10.1101/gr.8.3.175

Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat. Methods 7, 461–465. https://doi.org/10.1038/nmeth.1459

Hagemann IS (2015) Overview of technical aspects and chemistries of next-generation sequencing, in: Clinical Genomics. Elsevier, pp. 3–19. https://doi.org/10.1016/B978-0-12-404748-8.00001-0

Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, Ordoukhanian P (2014) Library construction for next-generation sequencing: overviews and challenges. BioTechniques 56, 61–4, 66, 68, passim. https://doi.org/10.2144/000114133

Heather JM, Chain B (2016) The sequence of sequencers: The history of sequencing DNA. Genomics 107, 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003

Hughes KW, Petersen RH, Lodge DJ, Bergemann SE, Baumgartner K, Tulloss RE, Lickey E, Cifuentes J (2013) Evolutionary consequences of putative intra-and interspecific hybridization in agaric fungi. Mycologia 105, 1577–1594. https://doi.org/10.3852/13-041

Inglis PW, Pappas M de CR, Resende LV, Grattapaglia D (2018) Fast and inexpensive protocols for consistent extraction of high quality DNA and RNA from challenging plant and fungal samples for high-throughput SNP genotyping and sequencing applications. PLoS ONE 13, e0206085. https://doi.org/10.1371/journal.pone.0206085

Jain M, Olsen HE, Paten B, Akeson M (2016) The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol. 17, 239. https://doi.org/10.1186/s13059-016-1103-0

Johne R, Müller H, Rector A, van Ranst M, Stevens H (2009) Rolling-circle amplification of viral DNA genomes using phi29 polymerase. Trends Microbiol. 17, 205–211. https://doi.org/10.1016/j.tim.2009.02.004

Kono N, Arakawa K (2019) Nanopore sequencing: Review of potential applications in functional genomics. Dev. Growth Differ. 61, 316–326. https://doi.org/10.1111/dgd.12608

Krehenwinkel H, Pomerantz A, Henderson JB, Kennedy SR, Lim JY, Swamy V, Shoobridge JD, Graham N, Patel NH, Gillespie RG, Prost S (2019) Nanopore sequencing of long ribosomal DNA amplicons enables portable and simple biodiversity assessments with high phylogenetic resolution across broad taxonomic scale. Gigascience 8. https://doi.org/10.1093/gigascience/giz006

Kyriakidou M, Tai HH, Anglin NL, Ellis D, Strömvik MV (2018) Current strategies of polyploid plant genome sequence assembly. Front. Plant Sci. 9, 1660. https://doi.org/10.3389/fpls.2018.01660

Maestri S, Cosentino E, Paterno M, Freitag H, Garces JM, Marcolungo L, Alfano M, Njunjić I, Schilthuizen M, Slik F, Menegon M, Rossato M, Delledonne M (2019) A Rrapid and accurate MinION-based workflow for tracking species biodiversity in the field. Genes (Basel) 10. https://doi.org/10.3390/genes10060468

McGinn S, Gut IG (2013) DNA sequencing - spanning the generations. N. Biotechnol. 30, 366–372. https://doi.org/10.1016/j.nbt.2012.11.012

Meyer M, Kircher M (2010) Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb. Protoc. 2010, pdb.prot5448. https://doi.org/10.1101/pdb.prot5448

Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, Loudet O, Weigel D, Ecker JR (2018) High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. Nat. Commun. 9, 541. https://doi.org/10.1038/s41467-018-03016-2

Pacific Biosciences (n.d.) SMRT SCIENCE SMRT SEQUENCING [WWW Document]. URL https://www.pacb.com/smrt-science/smrt-sequencing/ (accessed 3.22.21).

Ravi RK, Walton K, Khosroheidari M (2018) Miseq: A next generation sequencing platform for genomic analysis. Methods Mol. Biol. 1706, 223–232. https://doi.org/10.1007/978-1-4939-7471-9_12

Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M, Hoon J, Simons JF, Marran D, Myers JW, Davidson JF, Branting A, Nobile JR, Puc BP, Light D, Clark TA, Bustillo J (2011) An integrated semiconductor device enabling non-optical genome sequencing. Nature 475, 348–352. https://doi.org/10.1038/nature10242

Slatko BE, Gardner AF, Ausubel FM (2018) Overview of Next-Generation Sequencing Technologies. Curr. Protoc. Mol. Biol. 122, e59. https://doi.org/10.1002/cpmb.59

Srivathsan A, Lee L, Katoh K, Hartop E, Kutty SN, Wong J, Yeo D, Meier R (2021) MinION barcodes: biodiversity discovery and identification by everyone, for everyone. BioRxiv. https://doi.org/10.1101/2021.03.09.434692

Thomas T, Gilbert J, Meyer F (2012) Metagenomics - a guide from sampling to data analysis. Microb. Inform. Exp. 2, 3. https://doi.org/10.1186/2042-5783-2-3

Troll CJ, Kapp J, Rao V, Harkins KM, Cole C, Naughton C, Morgan JM, Shapiro B, Green RE (2019) A ligation-based single-stranded library preparation method to analyze cell-free DNA and synthetic oligos. BMC Genomics 20, 1023. https://doi.org/10.1186/s12864-019-6355-0

Twyford AD (2016) Will Benchtop Sequencers Resolve the Sequencing Trade-off in Plant Genetics? Front. Plant Sci. 7, 433. https://doi.org/10.3389/fpls.2016.00433

University of Michigan, Biomedical Research Core Facilities (n.d.) Interpretation of Sequencing Chromatograms [WWW Document]. Interpretation of Sequencing Chromatograms. URL https://brcf.medicine.umich.edu/cores/advanced-genomics/faqs/sanger-sequencing-faqs/interpretation-of-sequencing-chromatograms/ (accessed 3.15.21).

van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C (2018) The third revolution in sequencing technology. Trends Genet. 34, 666–681. https://doi.org/10.1016/j.tig.2018.05.008

Vereecke N, Bokma J, Haesebrouck F, Nauwynck H, Boyen F, Pardon B, Theuns S (2020) High quality genome assemblies of Mycoplasma bovis using a taxon-specific Bonito basecaller for MinION and Flongle long-read nanopore sequencing. BMC Bioinformatics 21, 517. https://doi.org/10.1186/s12859-020-03856-0

Vollger MR, Logsdon GA, Audano PA, Sulovari A, Porubsky D, Peluso P, Wenger AM, Concepcion GT, Kronenberg ZN, Munson KM, Baker C, Sanders AD, Spierings DCJ, Lansdorp PM, Surti U, Hunkapiller MW, Eichler EE (2020) Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads. Ann. Hum. Genet. 84, 125–140. https://doi.org/10.1111/ahg.12364

Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, Töpfer A, Alonge M, Mahmoud M, Qian Y, Chin C-S, Phillippy AM, Schatz MC, Myers G, DePristo MA, Ruan J, Hunkapiller MW (2019) Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat. Biotechnol. 37, 1155–1162. https://doi.org/10.1038/s41587-019-0217-9

Answers

1. For whole genome projects, PacBio and Nanopore technologies are the most commonly used technologies. This is because both are long read technologies, which reduces the bioinformatic challenges related to assembling 1000s of short contigs together for assembling a whole genome. For library preparations, Illumina platforms are still the most commonly

used due to their relatively competitive costs, high accuracy, and available support from a range of analysis tools and pipelines.

2. It is important to create equimolar pools so that the total number of DNA molecules is normalised across a library, so that one result does not dominate the others. However, even after normalisation of concentrations, it may still be the case that amplicons of very different length will not be amplified with the same efficiency. Additionally, using primers of roughly the same length so that their annealing temperatures are approximately the same is also important in order to avoid PCR bias within the same library. Thus, in a scenario with amplicons and/or primers of very different length, it is often best to put those amplicons in separate libraries. However, when amplicons and primers are of reasonably similar size, pooling the library samples can be an effective method to reduce sequencing costs.
3. Nanopore and PacBio based technologies provide longer reads than Illumina or Ion torrent based technologies since both have platforms available that do not require for a sample to be fragmented. These long reads can be especially useful for projects in plant identification when working with sequences that are particularly repetitive or have very large genomes. Additionally, PacBio technologies can also be used when longer amplicons are required for the phasing of haplotypes for instance or when tracing polyploid ancestry.

Chapter 10
DNA barcoding

Hugo de Boer[1]

1 Natural History Museum, University of Oslo, Oslo, Norway

Hugo de Boer **h.de.boer@nhm.uio.no**

Citation: de Boer H (2022) Chapter 10. DNA barcoding. In: de Boer H, Rydmark MO, Verstraete B, Gravendeel B (Eds) Molecular identification of plants: from sequence to species. Advanced Books. https://doi.org/10.3897/ab.e98875

DNA barcoding

The method of identifying living organisms to species level using DNA sequences has been coined DNA barcoding (Hebert et al. 2003). It makes use of short (< 1000 bp), agreed-upon regions of the genome (a 'barcode') that evolve quickly enough to differ among closely related species (Kress et al. 2005). A generated barcode sequence from a sample allows for identification by matching the sequence against a reference library of sequences. Reference sequence libraries comprise sequences generated from vouchered and expert-identified materials in natural history collections, which are available through public sequence repositories or tailored databases. In other words, DNA barcodes function as molecular identifiers for individual species, in the same way as machine-readable black-and-white barcodes are used in the retail industry to identify products (Veldman et al. 2014).

DNA-based typing for species identification focused first on microbial organisms (Olive and Bean 1999). DNA barcoding as a concept distinct from DNA-based typing or phylogenetic analysis of taxon accessions was popularised by Hebert et al. (2003), who proposed to use the mitochondrial gene CO1 as the standard barcode for all animals. Despite initial scepticism (Rubinoff et al. 2006; Will and Rubinoff 2004), DNA barcoding was readily embraced by the scientific community. Assessments have since shown that CO1 can be used to distinguish over 90% of species in many animal groups: among these spiders (Barrett and Hebert 2005), birds (Hebert et al. 2004b), amphibians (Smith et al. 2008), and butterflies (Burns et al. 2008).

In recent years, the barcoding movement has grown substantially, and worldwide efforts coordinated by the Consortium for the Barcode of Life (CBOL) are now being focused on barcoding all organisms (Hobern and Hebert 2019; Hobern 2020). The amount of sequencing data derived from DNA barcoding is exponentially increasing, and it is now considered a mainstream taxonomic tool. Although DNA barcoding does not replace the need for traditional taxonomy, it does highlight the need for robust species descriptions to enable accurate identification of species from "orphan" barcodes (sequences from unnamed species). Integrative taxonomy, which is achieved by combining evidence from morphology, ecology, phylogenetics and DNA barcoding, is critically important to speed up species discovery in the light of biodiversity loss (Padial et al. 2010; Schlick-Steiner et al. 2010).

DNA barcoding and species delimitation

Species delimitation is a central tenet of taxonomy (see Chapter 17 Species delimitation). Traditionally, species were identified, described and classified based mainly on their morphological characters. This is more difficult when it comes to cryptic, hybridising or highly convergent species (Struck et al. 2018). Combining characters, such as molecular data and behaviour, can provide further confidence when attempting to distinguish between species (Schlick-Steiner et al. 2010). However, species delimitation remains fundamentally difficult due to the fact that it is unclear how a species should be defined (de Queiroz 2007). The assumption that species are fixed entities underpins every international agreement on biodiversity conservation, all national environmental legislation and the efforts of many individuals and organisations to safeguard plants and animals (Garnett and Christidis 2017). However, one of the major unresolved questions in science is 'What is a species?', even though this is one of the most important concepts in biology (Kennedy and Norman 2005). Species concepts differ, and a single definition that fits all organisms has not been found (de Queiroz 2007).

Most species concepts agree on species being evolving metapopulations (de Queiroz 2007), and this implies that genetic variation exists both within and between species. Advanced approaches using many accessions as well as many loci, such as species delimitation based on multispecies coalescent theory, can enhance species identification resolution. However, more data also adds new challenges, and inferred structure due to population-level processes and that due to species boundaries are hard to distinguish (Sukumaran and Knowles 2017). Initial studies on DNA barcoding suggested a significant barcoding 'gap' between intra- and interspecific variation (Barrett and Hebert 2005; Hebert et al. 2004a, 2003), but these studies have been criticised for undersampling both intraspecific and interspecific divergence (Meyer and Paulay 2005). A DNA barcoding reference database for identification that would include all species references should also contain multiple accessions of populations to ensure that intraspecific and interspecific variation can be distinguished In absence of this ideal situation, many studies use more or less arbitrary cut-off percentages for sequence divergence (Blaxter et al. 2005; Ghorbani et al. 2017a; Veldman et al. 2017). Species assignments in DNA barcoding are hypotheses similar to species assignments based on morphology.

To identify an unknown DNA barcode using a reference library, one can use several approaches to look at the interrelatedness of the samples (see Chapter 18 Sequence to species). Many databases including GenBank and BOLD (Ratnasingham and Hebert 2007) make use of the similarity based BLAST (Altschul et al. 1990). Similarity based on genetic distance can be used as above and for phylogenetic tree reconstruction (Hebert et al. 2003). Disadvantages of using distance based information are 1) these do not yield diagnostic characters for species distinction (DeSalle et al. 2005); 2) similarity scores do not always give the nearest neighbour as the closest relative (Koski and Golding 2001); and 3) a lack of an objective set of criteria to delineate taxa when using distances (Goldstein et al. 2000). Other common approaches are based on characters instead of distances, and these rely on phylogenetic methods using maximum likelihood (Felsenstein 1973), parsimony (Nixon 1999), Bayesian statistics (Huelsenbeck and Ronquist 2001), or multispecies coalescent methods (Yang and Rannala 2017). These phylogenetic methods are implemented in RAxML (Stamatakis 2006; Swofford 2002), PAUP* (Swofford 2002), MrBayes (Huelsenbeck and Ronquist 2001) and BPP (Yang 2015), respectively. Character-based tree building overcomes many of the shortcomings of distance-based results, but tree-based methods have limitations if single gene trees are used to infer phylogenetic relationships. Another limitation of tree building for species identification is that evolution at the species level is not hierarchical. Applying hierarchical methods and terms, such as trees, classification and monophyly for delimitation of species, is not reflective of the evolutionary history of individuals and populations within a species (DeSalle et al. 2005). Veldman et al. (2014) summarised several suggestions to minimise the effect of these drawbacks, such as a diagnostic system including other lines of evidence (DeSalle et al. 2005), a probabilistic modelling approach (Knowles and Carstens 2007), the use of dominant and codominant multi-locus markers (Hausdorf and Hennig 2010) and new heuristic methods without fixed species assignments (O'Meara 2010).

DNA barcoding for plants

The mitochondrial genome in plants evolves far too slowly to allow it to distinguish between species (Cho et al. 2004). Phylogenetic studies of plants focused early on plastid markers as well as the ribosomal DNA (Palmer et al. 1988; Ritland and Clegg 1987). In the search for alternatives to the popular animal marker COI, various genes and non-coding regions in the plastid genome were proposed (CBOL Plant Working Group 2009; Fazekas et al. 2008, 2009; Hollingsworth

2011; Kress and Erickson 2007; Kress et al. 2005). In its most basic definition, a barcode must differ between species so that species can be identified. However, a barcode should not differ much within species, and not be too different between species within the same genus or family because this would make it more difficult to assign a unknown to a group with confidence.

The plastid marker *rbcL* was for example good to infer relationships between angiosperm families (Soltis et al. 1999) but varies too little for species discrimination in many plant genera (China Plant BOL Group et al. 2011). In addition to being sufficiently rapidly evolving, a barcode must also be flanked by conserved regions of the genome that can function as universal amplification primer binding sites. A single primer pair that would amplify any of over 350,000 species of plants would be ideal (Kress et al. 2005). The plastid coding region *matK* for example has variation between species, but it can be difficult to amplify universally (de Boer et al. 2014; Kool et al. 2012) (de Boer et al. 2014; Piredda et al. 2011; Sass et al. 2007). Insufficient primer universality makes this marker difficult to use in large-scale studies across families, although using target-group specific-primers for amplification is often successful (Mahadani and Ghosh 2013; Palhares et al. 2015; Purushothaman et al. 2014; Wallace et al. 2012). Other considerations can also affect barcode marker choice.

The nuclear ribosomal marker ITS, and specifically nrITS2, is used commonly in barcoding and metabarcoding studies (China Plant BOL Group et al. 2011; Ivanova et al. 2016; Raclariu et al. 2017, 2018). nrITS2 has been long advocated as a secondary marker to plastid barcodes (China Plant BOL Group et al. 2011). However, nrDNA has limitations for phylogenetic inference that also apply to barcoding (Álvarez and Wendel 2003), including alignment difficulties and limited utility in phylogenetic inference between closely related and/or recently diverged taxa (Manzanilla et al. 2018). It is also a challenge to determine the orthology and the paralogy of nrDNA sequences in the case of hybridization events or incomplete lineage sorting (Bailey 2003; Fehrer et al. 2007; Soltis and Kuzoff 1995). nrITS is also present in multiple copies, and these copies can belong to different parental lineages in hybrids, and PCR amplification success of these copies is unrelated to whether these copies are functionally transcribed or not, which in turn has an influence on the substitution rate of these sequences (Kool et al. 2012). Bailey et al. (2003) emphasised that especially for allopolyploids nrDNA might not be the optimal choice to assess species trees, which applies equally well to species assignment in DNA barcoding studies.

The strict requirements for both universality and high variability for potential universal barcodes has led some to label DNA barcoding a "search for the Holy Grail" (Rubinoff et al. 2006). Since there is still no single plant barcoding locus combining variability and universality, the current consensus is that a combination of two or more markers should be used for standard barcoding applications (CBOL Plant Working Group 2009; China Plant BOL Group et al. 2011; Hollingsworth 2011). Thus, where the animal community is entirely focused on using standardized and defined markers for species discrimination, the plant community has a looser vision for DNA-based identification of plants with tailored solutions based on their study objective. Plant DNA-based identification incorporates plastid, nuclear ribosomal and nuclear sequence data ranging from barcodes through plastomes, genome skimming and target capture to whole genomes (Bohmann et al. 2020; Coissac et al. 2016; Hollingsworth et al. 2016; Manzanilla et al. 2018).

Hands-on plant DNA barcoding

The core plant DNA barcoding markers are *rbcL* and *matK* (CBOL Plant Working Group 2009). nrITS (or nrITS2 only) is the third most commonly used barcode (China Plant BOL Group et al. 2011; Hollingsworth 2011; Kress et al. 2005). The *trnL-F* spacer, *psbA-trnH*, and *rpoC1*

Table 1. The most commonly used primers for plant DNA barcoding.*

Barcode	Primer	Sequence (5′-3′)	Dir.	Reference
rbcLa	rbcLa_f	ATGTCACCACAAACAGAGACTAAAGC	F	Levin et al. (2003)
	rbcLa_rev	GTAAAATCAAGTCCACCRCG	R	Kress et al. (2009)
matK	matk-3F	CGTACAGTACTTTTGTGTTTACGAG	F	CBOL Plant Working Group (2009)
	matk-1R	ACCCAGTCCATCTGGAAATCTTGGTTC	R	CBOL Plant Working Group (2009)
nrITS	ITS5a	CCTTATCATTTAGAGGAAGGAG	F	Wurdack in Stanford et al. (2000)
	ITS4	TCCTCCGCTTATTGATATGC	R	White et al. (1990)
nrITS2	S2F	ATGCGATACTTGGTGTGAAT	F	Chen et al. (2010)
	S3R	GACGCTTCTCCAGACTACAAT	R	Chen et al. (2010)
trnL P6	trnL-g	GGGCAATCCTGAGCCAA	F	Taberlet et al. (2007)
	trnL-h	CCATTGAGTCTCTGCACCTATC	R	Taberlet et al. (2007)
psbA-trnH	psbA	GTTATGCATGAACGTAATGCTC	F	Sang et al. (1995)
	trnH	CGCGCATGGTGGATTCACAATCC	R	Tate et al. (2005)

*These are some of the most commonly used primers, but there are many more primers and markers that have been used for specific applications. Never use these primers blindly, but always check for appropriate markers and primers for your target group.

(Ghorbani et al. 2017b; Kool et al. 2012; Kress et al. 2005) are also reported in literature, and the *trnL* P6 loop is the standard barcode for plant metabarcoding studies (Taberlet et al. 2012, 2007).

When choosing appropriate markers for a plant DNA barcoding study it is important to consider the following questions:

What is the necessary taxonomic level of identification? For composition studies of a flora or vegetation, genus-level identifications are often sufficient. Species-level identification can however be important for other questions. Identifying all angiosperms in Greenland is more straightforward than in a Neotropical rainforest. Also, although family-level identifications in Greenland provide useful insights into the local flora, this information most often does not have meaningful applications in rainforests. After deciding on the appropriate level of identification, the researcher then needs to determine whether multiple markers are necessary to ensure that all species can be distinguished.

What kind of a reference library will you use to identify the target barcodes? Query identification in a database that contains all plants is more challenging than with a tailored reference library. For example, identifying a sequence of *Oxalis* (Oxalidaceae) is easy in a database of Scandinavian plant sequences because there is only a single native *Oxalis* species. Any queried *Oxalis* sequence would match the Scandinavian *Oxalis acetosella* because it would be the only reference *Oxalis* sequence in a local database. In contrast, a database with South American *Oxalis* species has hundreds of taxa, and identification requires a marker with sufficient variation to discriminate between these species. Thus, for Scandinavia, one could use a marker with limited variation but universal primers, whereas for South America a specific marker or markers should be sought that can distinguish all *Oxalis* species present in a global database. It is therefore critical to pick your marker(s) based on the expected diversity in your reference library.

What is your source of reference sequences? If you want to identify species, which is common in studies aiming to authenticate herbal drugs and supplements, you need to include all putative species in your reference library. For example, if your goal is to identify a European wild collected *Hypericum*, your reference library should ideally include all European *Hypericum* species that could be confused or substituted for *Hypericum perforatum*. A reference library can

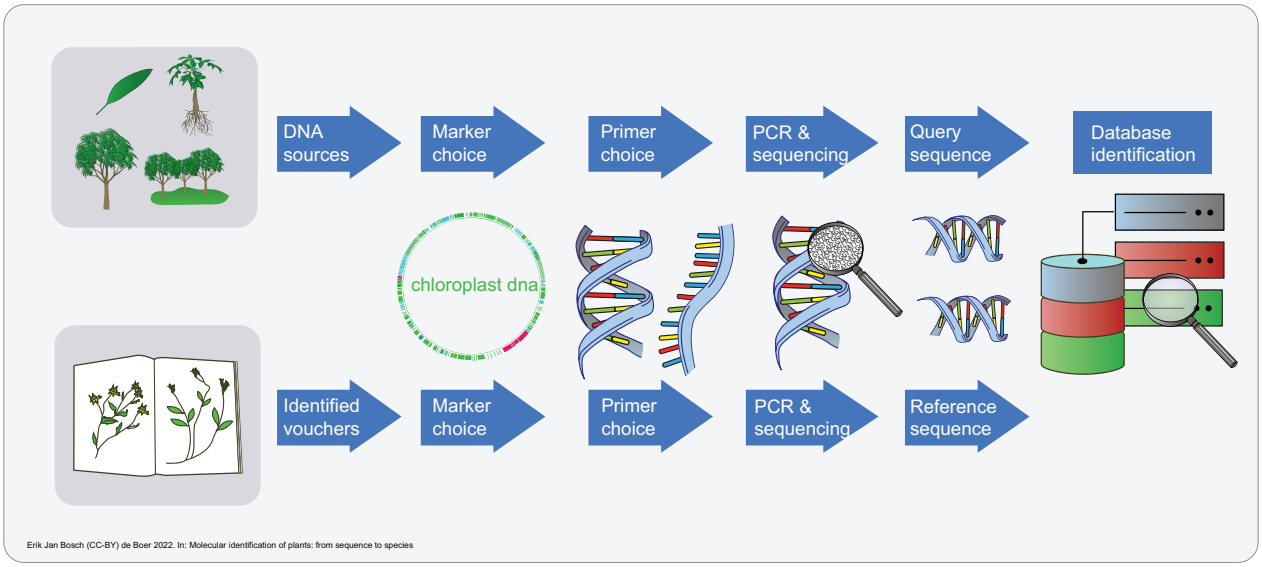

Figure 1. Chapter 10 Infographic: DNA barcoding of plants encompasses two streams of data from organism to DNA, one for the query sequence that should be identified and one for the reference sequence that is part of the reference library for identification. DNA source, marker choice, primer choice, sequencing approach and identification strategy all influence the ability and resolution of identification.

be compiled from *de novo* sequenced amplicons from voucher accessions or from reference sequences mined from public repositories.

After choosing one or several markers, it is important to consider the following:

Are universal primers available? If yes, this facilitates your project. However, are these primers really universal? Check this by seeing whether the study publishing the primers gets cited by relevant studies and look for larger studies and reviews that might provide more information about (1) amplification success with these primers; (2) ability to amplify from degraded or poor DNA extracts, a common challenge when working with older herbarium vouchers or processed herbal products; and (3) the need to tweak amplification protocols to make these primers work. If no universal primers are available, try to find studies using this marker and see which primers were used, to find suitable primers that you can then test. If possible, use studies targeting the same target order, family or genus. If there are no previously published primers for your marker, then it is necessary to design your own. If your primers target a widely used marker, then the primer performance that is assessed based on matching these novel primers to multiple sequence alignments of published data (*in-silico* testing) is generally reliable. If only genomic data is available, however, the accuracy of in-silico testing will be highly dependent on the relatedness of the reference genomes.

Do the primers amplify the right part of the marker? Primers can target fragments of longer loci, i.e., parts of *rbcL*, *matK*, nrITS. It is thus important that the segment the primers amplify is useful for your study. It should generate sequences that are identifiable in your reference library and variable enough for your intended level of identification. For example, targeting *trnL* intron with the universal g-h primers will yield short amplicons, and these have less variation than the entire *trnL-F* region. Make sure you reassess your marker choice after selecting suitable primers.

How many primers per marker will you use? Long markers can be hard to amplify from degraded templates and can be split up into multiple primer pairs. Degraded DNA is a common challenge when working with a common challenge when working with older herbarium vouchers or processed herbal products. Different combinations of forward and reverse primers can also increase the chance of successful amplification as having multiple different primers can increase the chance that one of these has a good fit to the organism being tested. However, the

primer pair with the best fit and targeting the shorter marker will amplify more effectively than other pairs or longer fragments, and can lead to amplification bias.

Once a suitable combination of markers has been found and suitable primers or primer panels have been selected, it is important to test the primers on a sufficient number of your samples. Template DNA quality, DNA concentration, and the effects of inhibiting secondary metabolites can all influence the efficacy of the PCR and might require optimization to obtain the best possible results for the largest number of samples. This is beyond the scope of this book, but sufficient online resources are available to help you with optimization. In addition, there are many online discussion forums to troubleshoot PCR optimization.

The subsequent chapters in section 2 describe different sequencing platforms and approaches to obtain DNA sequences for downstream analysis, and section 3 provides an overview of applications of molecular identification of plants. Depending on whether one chooses standard DNA barcoding using Sanger sequencing, DNA metabarcoding using Ion Torrent, Illumina, or other platforms, or a variety of whole or reduced library representation genome sequencing approaches, one will need to choose different wet lab steps to create the relevant sequencing libraries. Check out the relevant chapter for your application to find out more.

Questions

4. An author writes that she used DNA barcoding to identify *Bellis perennis* using *rbcL*. The generated query sequence matched 100% with the reference of *Bellis perennis* in GenBank. Can you think of two situations that would falsify this finding?
5. You are planning to use DNA barcoding to distinguish herbal medicines based on *Paeonia*. In the literature you find that five *Paeonia* species are commonly used in herbal medicines and that these can be distinguished using nrITS2 sequences. In your study of 37 herbals, you find that 15 contain species A, 7 B, 5 C, 3 D, and 2 E, whereas five samples fail to amplify nrITS2. 2A) How can you be sure that these 37 products contain only these five species? 2B) What does your experiment tell you about the five samples that failed to amplify?
6. You want to investigate if DNA barcoding can outperform morphology-based biodiversity assessments in terms of species identification. For what material do you expect DNA barcoding to be more useful than morphology-based identification?

Glossary

matK – Plastid gene coding for maturase K. *matK* is one of the core plant DNA barcodes.

nrITS – Internal transcribed spacer (ITS) is a spacer situated between the small-subunit rDNA and large-subunit rDNA genes. In plants, it flanks the 18S and 26S rDNA genes. nrITS is split into two spacers, nrITS1 and nrITS2 with the 5.8S rDNA gene in between. nrITS is highly variable, and primers are designed in the conservated 18S, 5.8S, and 26S rDNA genes.

psbA-trnH – Plastid intergenic spacer region between the coding genes *psbA* and *trnH*. *psbA-trnH* has been advocated as a plant DNA barcoding marker.

Primer – Short DNA sequence used to amplify a marker.

rbcL – Plastid gene coding for ribulose-1,5-bisphosphate carboxylase-oxygenase. Most barcoding studies target the rbcLa region, but will refer to *rbcL*. *rbcL* is one of the core plant DNA barcodes. Plastids in plants are often incorrectly referred to as chloroplasts.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J. Mol. Biol. 215, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Álvarez I, Wendel JF (2003) Ribosomal ITS sequences and plant phylogenetic inference. Mol. Phylogenet. Evol. 29, 417–434. https://doi.org/10.1016/S1055-7903(03)00208-2

Bailey C (2003) Characterization of angiosperm nrDNA polymorphism, paralogy, and pseudogenes. Mol. Phylogenet. Evol. 29, 435–455. https://doi.org/10.1016/j.ympev.2003.08.021

Barrett RDH, Hebert PDN (2005) Identifying spiders through DNA barcodes. Can. J. Zool. 83, 481–491. https://doi.org/10.1139/z05-024

Blaxter M, Mann J, Chapman T, Thomas F, Whitton C, Floyd R, Abebe E (2005) Defining operational taxonomic units using DNA barcode data. Philos. Trans. R. Soc. Lond. B Biol. Sci. 360, 1935–1943. https://doi.org/10.1098/rstb.2005.1725

Bohmann K, Mirarab S, Bafna V, Gilbert MTP (2020) Beyond DNA barcoding: The unrealized potential of genome skim data in sample identification. Mol. Ecol. 29, 2521–2534. https://doi.org/10.1111/mec.15507

Burns JM, Janzen DH, Hajibabaei M, Hallwachs W, Hebert PDN (2008) DNA barcodes and cryptic species of skipper butterflies in the genus Perichares in Area de Conservacion Guanacaste, Costa Rica. Proc Natl Acad Sci USA 105, 6350–6355. https://doi.org/10.1073/pnas.0712181105

CBOL Plant Working Group (2009) A DNA barcode for land plants. Proc Natl Acad Sci USA 106, 12794–12797. https://doi.org/10.1073/pnas.0905845106

Chen S, Yao H, Han J, Liu C, Song J, Shi L, Zhu Y, Ma X, Gao T, Pang X, Luo K, Li Y, Li X, Jia X, Lin Y, Leon C (2010) Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. PLoS ONE 5, e8613. https://doi.org/10.1371/journal.pone.0008613

China Plant BOL Group, Li D-Z, Gao L-M, Li H-T, Wang H, Ge X-J, Liu J-Q, Chen Z-D, Zhou S-L, Chen S-L, Yang J-B, Fu C-X, Zeng C-X, Yan H-F, Zhu Y-J, Sun Y-S, Chen S-Y, Zhao L, Wang K, Yang T, Duan G-W (2011) Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. Proc Natl Acad Sci USA 108, 19641–19646. https://doi.org/10.1073/pnas.1104551108

Cho Y, Mower JP, Qiu Y, Palmer JD (2004) Mitochondrial substitution rates are extraordinarily elevated and variable in a genus of flowering plants. PNAS 101, 17741–17746.

Coissac E, Hollingsworth PM, Lavergne S, Taberlet P (2016) From barcodes to genomes: extending the concept of DNA barcoding. Mol. Ecol. 25, 1423–1428. https://doi.org/10.1111/mec.13549

DeSalle R, Egan MG, Siddall M (2005) The unholy trinity: taxonomy, species delimitation and DNA barcoding. Philos. Trans. R. Soc. Lond. B Biol. Sci. 360, 1905–1916. https://doi.org/10.1098/rstb.2005.1722

de Boer HJ, Ouarghidi A, Martin G, Abbad A, Kool A (2014) DNA barcoding reveals limited accuracy of identifications based on folk taxonomy. PLoS ONE 9, e84291. https://doi.org/10.1371/journal.pone.0084291

de Queiroz K (2007) Species concepts and species delimitation. Syst. Biol. 56, 879–886. https://doi.org/10.1080/10635150701701083

Fazekas AJ, Burgess KS, Kesanakurti PR, Graham SW, Newmaster SG, Husband BC, Percy DM, Hajibabaei M, Barrett SCH (2008) Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. PLoS ONE 3, e2802. https://doi.org/10.1371/journal.pone.0002802

Fazekas AJ, Kesanakurti PR, Burgess KS, Percy DM, Graham SW, Barrett SCH, Newmaster SG, Hajibabaei M, Husband BC (2009) Are plant species inherently harder to discriminate than animal species using DNA barcoding markers? Mol. Ecol. Resour. 9 Suppl s1, 130–139. https://doi.org/10.1111/j.1755-0998.2009.02652.x

Fehrer J, Gemeinholzer B, Chrtek J, Bräutigam S (2007) Incongruent plastid and nuclear DNA phylogenies reveal ancient intergeneric hybridization in Pilosella hawkweeds (Hieracium, Cichorieae, Asteraceae). Mol. Phylogenet. Evol. 42, 347–361. https://doi.org/10.1016/j.ympev.2006.07.004

Felsenstein J (1973) Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst. Zool. 22, 240. https://doi.org/10.2307/2412304

Garnett ST, Christidis L (2017) Taxonomy anarchy hampers conservation. Nature 546, 25–27. https://doi.org/10.1038/546025a

Ghorbani A, Gravendeel B, Selliah S, Zarré S, de Boer HJ (2017a) DNA barcoding of tuberous Orchidoideae: a resource for identification of orchids used in Salep. Molecular ecology resources 17, 342–352.

Ghorbani A, Saeedi Y, de Boer HJ (2017b) Unidentifiable by morphology: DNA barcoding of plant material in local markets in Iran. PLoS ONE 12, e0175722. https://doi.org/10.1371/journal.pone.0175722

Goldstein PZ, Desalle R, Amato G, Vogler AP (2000) Conservation genetics at the species boundary. Conserv. Biol. 14, 120–131. https://doi.org/10.1046/j.1523-1739.2000.98122.x

Hausdorf B, Hennig C (2010) Species delimitation using dominant and codominant multilocus markers. Syst. Biol. 59, 491–503. https://doi.org/10.1093/sysbio/syq039

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. Proc. Biol. Sci. 270, 313–321. https://doi.org/10.1098/rspb.2002.2218

Hebert PDN, Penton EH, Burns JM, Janzen DH, Hallwachs W (2004a) Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly Astraptes fulgerator. Proc Natl Acad Sci USA 101, 14812–14817. https://doi.org/10.1073/pnas.0406166101

Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM (2004b) Identification of birds through DNA Barcodes. PLoS Biol. 2, e312. https://doi.org/10.1371/journal.pbio.0020312

Hobern D, Hebert P (2019) BIOSCAN - revealing eukaryote diversity, dynamics, and interactions. BISS 3. https://doi.org/10.3897/biss.3.37333

Hobern D (2020) BIOSCAN: DNA barcoding to accelerate taxonomy and biogeography for conservation and sustainability. Genome 1–4. https://doi.org/10.1139/gen-2020-0009

Hollingsworth PM, Li D-Z, van der Bank M, Twyford AD (2016) Telling plant species apart with DNA: from barcodes to genomes. Philos. Trans. R. Soc. Lond. B Biol. Sci. 371. https://doi.org/10.1098/rstb.2015.0338

Hollingsworth PM (2011) Refining the DNA barcode for land plants. Proc Natl Acad Sci USA 108, 19451–19452. https://doi.org/10.1073/pnas.1116812108

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17, 754–755. https://doi.org/10.1093/bioinformatics/17.8.754

Ivanova NV, Kuzmina ML, Braukmann TWA, Borisenko AV, Zakharov EV (2016) Authentication of herbal supplements using next-generation sequencing. PLoS ONE 11, e0156426. https://doi.org/10.1371/journal.pone.0156426

Kennedy D, Norman C (2005) What don't we know? Science 309, 75. https://doi.org/10.1126/science.309.5731.75

Knowles LL, Carstens BC (2007) Delimiting species without monophyletic gene trees. Syst. Biol. 56, 887–895. https://doi.org/10.1080/10635150701701091

Kool A, de Boer HJ, Krüger A, Rydberg A, Abbad A, Björk L, Martin G (2012) Molecular identification of commercialized medicinal plants in southern Morocco. PLoS ONE 7, e39459. https://doi.org/10.1371/journal.pone.0039459

Koski LB, Golding GB (2001) The closest BLAST hit is often not the nearest neighbor. J. Mol. Evol. 52, 540–542. https://doi.org/10.1007/s002390010184

Kress WJ, Erickson DL, Jones FA, Swenson NG, Perez R, Sanjur O, Bermingham E (2009) Plant DNA barcodes and a community phylogeny of a tropical forest dynamics plot in Panama. Proc Natl Acad Sci USA 106, 18621–18626. https://doi.org/10.1073/pnas.0909820106

Kress WJ, Erickson DL (2007) A two-locus global DNA barcode for land plants: the coding rbcL gene complements the non-coding trnH-psbA spacer region. PLoS ONE 2, e508. https://doi.org/10.1371/journal.pone.0000508

Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH (2005) Use of DNA barcodes to identify flowering plants. Proc Natl Acad Sci USA 102, 8369–8374. https://doi.org/10.1073/pnas.0503123102

Levin RA, Wagner WL, Hoch PC, Nepokroeff M, Pires JC, Zimmer EA, Sytsma KJ (2003) Family-level relationships of Onagraceae based on chloroplast rbcL and ndhF data. Am. J. Bot. 90, 107–115. https://doi.org/10.3732/ajb.90.1.107

Mahadani P, Ghosh SK (2013) DNA Barcoding: A tool for species identification from herbal juices. DNA Barcodes 1, 35–38. https://doi.org/10.2478/dna-2013-0002

Manzanilla V, Kool A, Nguyen Nhat L, Nong Van H, Le Thi Thu H, de Boer HJ (2018) Phylogenomics and barcoding of *Panax*: toward the identification of ginseng species. BMC Evol. Biol. 18, 44. https://doi.org/10.1186/s12862-018-1160-y

Meyer CP, Paulay G (2005) DNA barcoding: error rates based on comprehensive sampling. PLoS Biol. 3, e422. https://doi.org/10.1371/journal.pbio.0030422

Nixon KC (1999) The parsimony ratchet, a new method for rapid parsimony analysis. Cladistics 15, 407–414. https://doi.org/10.1111/j.1096-0031.1999.tb00277.x

O'Meara BC (2010) New heuristic methods for joint species delimitation and species tree inference. Syst. Biol. 59, 59–73. https://doi.org/10.1093/sysbio/syp077

Olive DM, Bean P (1999) Principles and applications of methods for DNA-based typing of microbial organisms. J. Clin. Microbiol. 37, 1661–1669. https://doi.org/10.1128/JCM.37.6.1661-1669.1999

Padial JM, Miralles A, De la Riva I, Vences M (2010) The integrative future of taxonomy. Front. Zool. 7, 16. https://doi.org/10.1186/1742-9994-7-16

Palhares RM, Gonçalves Drummond M, Dos Santos Alves Figueiredo Brasil B, Pereira Cosenza G, das Graças Lins Brandão M, Oliveira G (2015) Medicinal plants recommended by the world health organization: DNA barcode identification associated with chemical analyses guarantees their quality. PLoS ONE 10, e0127866. https://doi.org/10.1371/journal.pone.0127866

Palmer JD, Jansen RK, Michaels HJ, Chase MW, Manhart JR (1988) Chloroplast DNA variation and plant phylogeny. Ann Mo Bot Gard 75, 1180. https://doi.org/10.2307/2399279

Piredda R, Simeone MC, Attimonelli M, Bellarosa R, Schirone B (2011) Prospects of barcoding the Italian wild dendroflora: oaks reveal severe limitations to tracking species identity. Mol. Ecol. Resour. 11, 72–83. https://doi.org/10.1111/j.1755-0998.2010.02900.x

Purushothaman N, Newmaster SG, Ragupathy S, Stalin N, Suresh D, Arunraj DR, Gnanasekaran G, Vassou SL, Narasimhan D, Parani M (2014) A tiered barcode authentication tool to differentiate medicinal Cassia species in India. Genet Mol Res 13, 2959–2968.

Raclariu AC, Heinrich M, Ichim MC, de Boer H (2018) Benefits and limitations of DNA barcoding and metabarcoding in herbal product authentication. Phytochem. Anal. 29, 123–128. https://doi.org/10.1002/pca.2732

Raclariu AC, Paltinean R, Vlase L, Labarre A, Manzanilla V, Ichim MC, Crisan G, Brysting AK, de Boer H (2017) Comparative authentication of Hypericum perforatum herbal products using DNA metabarcoding, TLC and HPLC-MS. Sci. Rep. 7, 1291. https://doi.org/10.1038/s41598-017-01389-w

Ratnasingham S, Hebert PDN (2007) BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). Mol. Ecol. Notes 7, 355–364. https://doi.org/10.1111/j.1471-8286.2007.01678.x

Ritland K, Clegg MT (1987) Evolutionary analysis of plant DNA sequences. Am. Nat. 130, S74–S100. https://doi.org/10.1086/284693

Rubinoff D, Cameron S, Will K (2006) Are plant DNA barcodes a search for the Holy Grail? Trends Ecol. Evol. 21, 1–2. https://doi.org/10.1016/j.tree.2005.10.019

Sang T, Crawford DJ, Stuessy TF (1995) Documentation of reticulate evolution in peonies (Paeonia) using internal transcribed spacer sequences of nuclear ribosomal DNA: implications for biogeography and concerted evolution. Proc Natl Acad Sci USA 92, 6813–6817. https://doi.org/10.1073/pnas.92.15.6813

Sass C, Little DP, Stevenson DW, Specht CD (2007) DNA barcoding in the cycadales: testing the potential of proposed barcoding markers for species identification of cycads. PLoS ONE 2, e1154. https://doi.org/10.1371/journal.pone.0001154

Schlick-Steiner BC, Steiner FM, Seifert B, Stauffer C, Christian E, Crozier RH (2010) Integrative taxonomy: a multisource approach to exploring biodiversity. Annu. Rev. Entomol. 55, 421–438. https://doi.org/10.1146/annurev-ento-112408-085432

Smith MA, Poyarkov NA, Hebert PDN (2008) DNA BARCODING: CO1 DNA barcoding amphibians: take the chance, meet the challenge. Mol. Ecol. Resour. 8, 235–246. https://doi.org/10.1111/j.1471-8286.2007.01964.x

Soltis DE, Kuzoff RK (1995) Discordance between nuclear and chloroplast phylogenies in the heuchera group (saxifragaceae). Evolution 49, 727–742. https://doi.org/10.1111/j.1558-5646.1995.tb02309.x

Soltis PS, Soltis DE, Chase MW (1999) Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. Nature 402, 402–404. https://doi.org/10.1038/46528

Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688–2690. https://doi.org/10.1093/bioinformatics/btl446

Stanford AM, Harden R, Parks CR (2000) Phylogeny and biogeography of Juglans (Juglandaceae) based on matK and ITS sequence data. Am. J. Bot. 87, 872–882.

Struck TH, Feder JL, Bendiksby M, Birkeland S, Cerca J, Gusarov VI, Kistenich S, Larsson K-H, Liow LH, Nowak MD, Stedje B, Bachmann L, Dimitrov D (2018) Finding evolutionary processes hidden in cryptic species. Trends Ecol. Evol. 33, 153–163. https://doi.org/10.1016/j.tree.2017.11.007

Sukumaran J, Knowles LL (2017) Multispecies coalescent delimits structure, not species. Proc Natl Acad Sci USA 114, 1607–1612. https://doi.org/10.1073/pnas.1607921114

Swofford DL (2002) PAUP*: phylogenetic analysis using parsimony (* and other methods). Version. 4. Sinauer Associates, Sunderland, Massachusetts.

Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. Mol. Ecol. 21, 2045–2050. https://doi.org/10.1111/j.1365-294X.2012.05470.x

Taberlet P, Coissac E, Pompanon F, Gielly L, Miquel C, Valentini A, Vermat T, Corthier G, Brochmann C, Willerslev E (2007) Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding. Nucleic Acids Res. 35, e14. https://doi.org/10.1093/nar/gkl938

Tate JA, Fuertes Aguilar J, Wagstaff SJ, La Duke JC, Bodo Slotta TA, Simpson BB (2005) Phylogenetic relationships within the tribe Malveae (Malvaceae, subfamily Malvoideae) as inferred from ITS sequence data. Am. J. Bot. 92, 584–602. https://doi.org/10.3732/ajb.92.4.584

Veldman S, Gravendeel B, Otieno JN, Lammers Y, Duijm E, Nieman A, Bytebier B, Ngugi G, Martos F, van Andel TR, de Boer HJ (2017) High-throughput sequencing of African chikanda cake highlights conservation challenges in orchids. Biodivers. Conserv. 26, 2029–2046. https://doi.org/10.1007/s10531-017-1343-7

Veldman S, Otieno J, Gravendeel B, Andel T van Boer H de (2014) Conservation of Endangered Wild Harvested Medicinal Plants: Use of DNA Barcoding. Novel Plant Bioresources: Applications in Food, Medicine and Cosmetics 81–88.

Wallace LJ, Boilard SMAL, Eagle SHC, Spall JL, Shokralla S, Hajibabaei M (2012) DNA barcodes for everyday life: Routine authentication of Natural Health Products. Food Res. Int 49, 446–452. https://doi.org/10.1016/j.foodres.2012.07.048

White TJ, Bruns T, Lee S, Taylor J (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics, in: Innis, M.A., Gelfand, J.J., Sninsky, D.H., White, T.J. (Eds.) PCR Protocols: A Guide to Methods and Applications. Academic Press San Diego, CA, pp. 315–322.

Will KW, Rubinoff D (2004) Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. Cladistics 20, 47–55. https://doi.org/10.1111/j.1096-0031.2003.00008.x

Yang Z, Rannala B (2017) Bayesian species identification under the multispecies coalescent provides significant improvements to DNA barcoding analyses. Mol. Ecol. 26, 3028–3036. https://doi.org/10.1111/mec.14093

Yang Z (2015) The BPP program for species tree estimation and species delimitation. Curr. Zool. 61, 854–865. https://doi.org/10.1093/czoolo/61.5.854

Answers

1. Some things that might have been overlooked: (1) Does NCBI GenBank list more than one species of *Bellis*? If not, then it might be any other *Bellis* species not present in this database; (2) How much variation does rbcL have in *Bellis*? Does the query match 100% with more than one species of *Bellis*? If yes, then a more variable marker should be used.
2. Answer for 2A) The study can ascertain that only these five species are present if it includes a sequence reference database of all other *Paeonia* species (or those possibly present). If the sequence reference database contains only the five common species, then no such conclusion can be made. 2B) The failed samples could: not include *Paeonia*; contain degraded *Paeonia* DNA that is not amplifiable; or contain inhibitors that make the DNA nonamplifiable.
3. Answers could include: vegetative material such as roots, leaves, and seedlings, DNA extracts from bulk samples, soil DNA, faecal DNA, pollen DNA, or air-captured eDNA.

Chapter 11
Amplicon metabarcoding

Physilia Chua[1,2*], Marcel Polling[3,4*], Christina Lynggaard[1], Maria Ariza Salazar[4], Kristine Bohmann[1]

1 Section for Evolutionary Genomics, Globe Institute, University of Copenhagen, Copenhagen, Denmark
2 Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK
3 Naturalis Biodiversity Center, Leiden, The Netherlands
4 Natural History Museum, University of Oslo, Oslo, Norway

Physilia Chua **physiliachua@gmail.com**
Marcel Polling **marcel.polling@wur.nl**
Christina Lynggaard **christina.l@sund.ku.dk**
Maria Ariza Salazar **m.a.salazar@nhm.uio.no**
Kristine Bohmann **kbohmann@sund.ku.dk**

* These authors contributed equally.

Citation: Chua P, Polling M, Lynggaard C, Salazar MA, Bohmann K (2022) Chapter 11. Amplicon metabarcoding. In: de Boer H, Rydmark MO, Verstraete B, Gravendeel B (Eds) Molecular identification of plants: from sequence to species. Advanced Books. https://doi.org/10.3897/ab.e98875

Background

What is metabarcoding?

DNA metabarcoding is an approach where taxonomically informative regions in the DNA are amplified from mixed-template samples containing DNA from different taxa for identification (Pompanon et al. 2012; Riaz et al. 2011). These taxonomically informative regions, also referred to as DNA barcodes or markers, ideally have low intraspecific variability and high interspecific variability to be able to discriminate between species, and conservative regions for universal amplification of the targeted community (Coissac et al. 2016). To target these DNA barcode regions, some prior knowledge is required for the design of primers that are complementary to flanking conservative regions of barcodes. Additionally, dependent on the metabarcoding approach used, primers can contain unique nucleotide tags to discern between samples during downstream bioinformatics processes (Binladen et al. 2007; Valentini et al. 2009b). After PCR amplification, amplicons are built into libraries where library indexes are added to allow for multiple amplicon libraries to be sequenced in one flow cell (Elbrecht and Leese 2015; Elbrecht et al. 2017). Adapters specific to the sequencing platforms are added to the PCR products (amplicons) and sequenced on a high-throughput sequencing (HTS) platform. The resulting sequences can be taxonomically identified by matching them to a reference database (De Barba et al. 2014; Kress and Erickson 2008; Taberlet et al. 2018, 2012). This method is useful for identifying different taxa from bulk samples of organismal DNA (Yu et al. 2012), and specifically to detect plants from environmental DNA (eDNA) samples including water, soil, sediment, air, and organic remains such as faeces (Deiner et al. 2017; Taberlet et al. 2012).

Plant metabarcoding

Metabarcoding is based on the DNA barcoding concept (see Chapter 10 DNA barcoding). However, for metabarcoding, samples containing DNA from a mix of different taxa are typically used. One of the first studies that used metabarcoding on a parallel sequencing system (herein referred to as DNA barcoding) to identify plants was by Valentini and colleagues (Valentini et al. 2009a) who analysed the diet of a variety of animals using their faeces. Earlier attempts at diet analyses were also made using chloroplastic (Poinar et al. 2001) and nuclear regions (Bradley et al. 2007), though these are not strictly speaking metabarcoding studies since they did not use high-throughput sequencing. Identification of plants through barcoding has had a turbulent history due to the lack of consensus on which plant barcodes should be used as standards (Pennisi 2007). In the landmark paper by Hebert and colleagues (Hebert et al. 2003), it was shown that animal species can be confidently identified through a short and highly variable piece of mitochondrial DNA called cytochrome oxidase subunit 1 (*CO1*). This has led many research groups to search for a similar barcode for the identification of plants (Chase et al. 2007; Kress et al. 2005). For plant species identification, the metabarcoding community has heavily relied on short fragments of plastid barcodes *rbcL*, *trnH-psbA*, *matK*, the P6 loop of the *trnL* intron and the nuclear ribosomal internal transcribed spacers nrITS1 and nrITS2 (China Plant BOL Group et al. 2011; Hollingsworth et al. 2016). There is, however, still no consensus on which plant DNA barcode(s) perform best. Studies that test various DNA barcodes for specific groups of plants find big differences between them (e.g., Braukmann et al. 2017), while others find that none of the available DNA barcodes provides species discrimination in certain plant groups (Zarrei et al. 2015). The search for the universal plant barcode is thus still ongoing.

Sample types and application

Plant metabarcoding is widely used to study the taxonomic composition of mixed template samples such as water (Zimmermann et al. 2015) (see Chapter 3 DNA from water), soil and sediments (Yoccoz et al. 2012; Ariza et al. 2022) (see Chapter 4 DNA from soil), bryophyte spores (Stech et al. 2011) airborne pollen from ambient air (Sickel et al. 2015; Kraaijeveld et al. 2015; Polling et al. 2022) (see Chapter 5 DNA from pollen), honey, food and medicine (Hawkins et al. 2015; Raclariu et al. 2018) (see Chapter 6 DNA from food and medicine), faeces and coprolites (Valentini et al. 2009a + Polling et al. 2021) (see Chapter 7 DNA from faeces), ancient sediments (Alsos et al. 2016) (see Chapter 8 DNA from ancient sediments), ice and snow (Thomsen and Willerslev 2015; Varotto et al. 2021) plant macrofossils (Murray et al. 2012), whole insects (Kajtoch 2014), gut contents (McClenaghan et al. 2015), and epilithic samples (Apothéloz-Perret-Gentil et al. 2017). DNA extraction methods are highly dependent on the type of material used and this is covered separately in Section 1 of this book.

Plant metabarcoding has been used in various types of applications including species delimitation (see Chapter 17 Species delimitation), archaeo- and palaeo-botany (Parducci et al. 2017) (see Chapter 21 Palaeobotany), healthcare (Reese et al. 2019) (see Chapter 23 Healthcare), food safety (Raclariu et al. 2017) (see Chapter 24 Food safety), environmental and biodiversity assessments (Fahner et al. 2016) (see Chapter 24 Environment and biodiversity assessments), wildlife trade (de Boer et al. 2017) (see Chapter 25 Wildlife trade), hay fever forecasts (Kraaijeveld et al. 2015) (see Chapter 5 DNA from pollen), water quality assessments (Smucker et al. 2020; Zimmermann et al. 2015) (see Chapter 3 DNA from water), and documenting environmental change (Jørgensen et al. 2012). These are some examples of plant-specific applications where metabarcoding has proven its value, though further detailed information can be found in the chapters referred to here.

Advantages and limitations of metabarcoding

DNA metabarcoding is a cost-effective method as compared to metagenomics (Chua et al. 2021a) (see Chapter 12 Metagenomics) or target capture (see Chapter 14 Target capture) as only DNA from targeted taxa is amplified and sequenced (Taberlet et al. 2012). The tagging system makes it possible to process large numbers of samples simultaneously, further decreasing the sequencing costs and increasing the total sample throughput. DNA present in low quantities (e.g., from rare species) can be targeted and amplified using specific primers and PCR-amplified. It is also a useful method for samples with low-quality DNA (i.e., degraded DNA) since it targets small barcodes that are relatively stable through time (Goldberg et al. 2016; Deiner et al. 2017). For example, plant DNA can be sequenced from ice core samples as old as 500 000 years old (Willerslev et al. 2007).

However, DNA metabarcoding also has its limitations, and the PCR amplification step has previously proven to be particularly problematic (Taberlet et al. 2012). This step can cause stochasticity (Murray et al. 2015) and create false positives (Ficetola et al. 2015), which stresses the need for both PCR and extraction replicates. However, depending on the specific research question, it may also be advisable to limit the number of PCR replicates and instead focus on sequencing depth (Smith and Peay 2014), although this would decrease species richness estimates (Dopheide et al. 2018).

Another drawback of DNA metabarcoding is primer binding bias due to mismatches between the primer and the template DNA. This can result in discrepancies between the proportion of the original taxa in the DNA extract and the amplified DNA sequences (Bista et al. 2018; Elbrecht and Leese 2015). Although quantitative results can be obtained from some primers using certain laboratory and bioinformatic controls (Ji et al. 2020; Piñol et al. 2019), this is still taxa-dependent and therefore not commonly used. Depending on the metabarcoding strate-

gy, tag jumps during library building should also be taken into consideration as they can cause false sequence-to-sample assignments (Carøe and Bohmann 2020; Schnell et al. 2015).

Finally, the taxonomic assignment of sequences to species is heavily dependent on the DNA reference database used for sequence matching. When the reference database to which the resulting sequences are compared to is incomplete and/or consists of inaccurately identified species, this results in erroneously identified species and/or false negatives (Banchi et al. 2020; Meiklejohn et al. 2019). This also affects the species resolution of the results. For example, a reference database based on the *trnL* barcode region may give a resolution of 33% species identification on a large circum-arctic scale, but within a localised area, this resolution may increase to 77–93% (Sønstebø et al. 2010; Alsos et al. 2018; Chua et al. 2021b). Thus, both the plant marker of choice as well as the reference database used are important and often limiting factors in metabarcoding studies for species identification. Lastly, taxonomic assignments between different species can have the same highest identity scores, but this can be handled by using a Last Common Ancestor approach (e.g., using MEGAN Huson et al. 2006 or OBITools Boyer et al. 2016).

Setting up a metabarcoding study

At the start of any (plant) metabarcoding study lies a clearly defined research question. A study design should furthermore encompass a clear sampling strategy, and identification of suitable DNA extraction techniques for the sample type used before carrying out downstream analysis (Zinger et al. 2019). As the chapters in Section 1 already details DNA extraction methods based on specific starting materials, this section will cover the subsequent steps, starting with selecting the plant barcodes to best answer the research question, choosing a nucleotide tagging strategy, sequencing and finally analysing the sequence output using bioinformatics pipelines.

Barcode choice

Barcode choice is one of the most important aspects of metabarcoding studies as it will determine which taxa are identified and to what resolution. Considerable efforts have gone into constructing libraries for these plant barcodes and in assessing their limitations (CBOL Plant Working Group 2009; Cowan et al. 2006; Fazekas et al. 2012; Hollingsworth et al. 2011; Kress 2017). Metabarcoding studies are often heavily dependent on reducing the potentially identifiable species, e.g., using *trnL* P6 loop one can make species-specific identifications of the Greenland flora, but family level identification in a tropical rainforest. The objective of the study determines the level of taxonomic resolution needed, and thus the approach (marker, replicates, etc), e.g., if only relative abundances at the family level are desired or if specific species in a vegetation plot need to be identified from soil. Different research groups use different 'preferred' barcodes that they consider best suited for their specific target plants. Despite this lack of consensus, the efficacy of metabarcoding for identifying the majority of plant species from plant mixtures still makes this a very useful tool. When choosing barcodes for metabarcoding studies, three factors must be considered: 1) sequence availability and presence in a reference library, 2) discriminatory power / taxonomic resolution, and 3) degree of DNA degradation in the sample (Hollingsworth et al. 2011). These three steps will be briefly explained below.

1. The first step is to check whether or not reference libraries exist for the sequences of the targeted organism(s). This is because barcodes are only useful if the sequences for the targeted

organism(s) are available in sequence repositories or reference libraries (Weigand et al. 2019). For some barcodes and specific geographic regions, optimised plant reference libraries exist that minimise inaccurate identification of sequences. One such example is the arctic boreal vascular plant and bryophyte database that is based on the P6 loop of *trnL* (Sønstebø et al. 2010). A curated global plant database is also available for nrITS2 (Banchi et al. 2020). Pre-made reference databases are not complete and it is therefore recommended to compare several databases to obtain the best resolution. Another option is to construct a tailored reference database, for example using the BOLD data portal or in GenBank using the e-utilities tool kit. The use of the publicly available GenBank database is generally discouraged as it contains many erroneous sequences (e.g., Steinegger and Salzberg 2020). If the target organisms are not present in any public sources, then one would opt for constructing de novo reference libraries. The idea behind it is to sequence barcodes from specimens collected in the study site, which are then assigned taxonomical annotations/identification (see Chapter 10 DNA barcoding). The construction of regional reference libraries usually employs a combination of both strategies described above. Last, one would opt for blasting the obtained sequences to a public source. This strategy would incur multiple taxonomic assignments to one single sequence and thus a threshold of blasting similarity would have to be arbitrarily designed.

2. Discriminatory power refers to how effectively the barcodes can discriminate between closely related species and is linked to the variability of the locus. Typically, barcodes can only identify plants up to a certain taxonomic level (resolution) depending on the barcode used and the group of plants targeted. Moreover, because reference libraries are incomplete for all DNA barcodes, some species may only be detected using one DNA barcode while others may only be detected by another. Therefore, using a single primer set will most often not result in the recovery of all species present in a sample. We recommend adopting a multilocus approach to gain highly resolved taxonomic coverage for complex samples (see e.g. Arulandhu et al. 2017).
3. DNA is relatively unstable in the environment and can degrade quickly depending on certain factors such as age, transport, and abiotic factors (Deiner et al. 2017). In highly degraded and/or old materials, the use of very short, highly distinctive barcodes is recommended (e.g., P6 loop of *trnL* intron). Although this can provide a good indication of the plant community from mixed samples, some taxa cannot be identified beyond the family level (e.g., Asteraceae and Poaceae). Therefore, when possible, it is recommended to use the longer and in some cases more distinctive nuclear ribosomal barcodes ITS1 (De Barba et al. 2014; Omelchenko et al. 2019) and/or ITS2 (Yao et al. 2010). However, the nuclear ITS region is also present in fungi and in order to avoid amplification of fungal DNA, plant-specific primers should be used (Cheng et al. 2016; Chen et al. 2010; Moorhouse-Gann et al. 2018; Omelchenko et al. 2019; Timpano et al. 2020).

Metabarcoding nucleotide tagging strategies

In the metabarcoding laboratory workflow, unique nucleotide tags are added to amplicons, and these tags are used to assign sequences to the sample they originate from (Binladen et al. 2007). This allows for the pooling of many labelled PCR replicates for sequencing, and dramatically increases the throughput. Labelling amplicons with unique nucleotide tags can be done at two stages during a metabarcoding workflow: prior to library building as 5′ nucleotide tags added to the amplicons, and/or after library completion as library indexes. The strategies to achieve this labelling can be condensed into three main approaches: the 'one-step PCR' approach, the 'two-step PCR' approach, and the 'tagged PCR approach'.

In the 'one-step PCR' approach, the metabarcoding barcode is amplified and built into libraries during one PCR. This is achieved through the use of metabarcoding primers that carry both adapters and library indexes (Elbrecht and Leese 2015; Elbrecht et al. 2017), though unique nucleotide tags instead of library indexes can also be added in the one-step PCR approach (Elbrecht and Steinke 2018). In this approach, each PCR replicate is a library.

In the 'two-step PCR' approach, sample extracts are PCR-amplified with metabarcoding primers that only carry 5' tails. These are added to act as templates for the following second PCR and do not include any labelling. The second PCR is carried out on each PCR product with primers that carry adapters and indexes (Galan et al. 2018; Miya et al. 2015; Swift et al. 2018), although unique nucleotide tags can also be added in the first PCR (Kitson et al. 2019). In the two-step PCR approach, each PCR replicate is also a library.

In the 'tagged PCR' approach, DNA extracts are PCR amplified with metabarcoding primers that carry 5' unique nucleotide tags. Next, the individually 5' tagged PCR products are pooled and library preparation is carried out on the pools (first demonstrated by (Binladen et al. 2007) on the 454 FLX platform). Library preparation can be with (Drinkwater et al. 2019; Hibert et al. 2013) or without (Carøe and Bohmann 2020; Sigsgaard et al. 2017) an indexing PCR step. Care should be taken with using this approach, as several studies have shown it to be prone to so-called tag-jumping where amplicon sequences carry false combinations of nucleotide tags after amplification (Schnell et al. 2015). This can be avoided using specific library preparation protocols (Carøe and Bohmann 2020; Sigsgaard et al. 2017)). Finally, indexes can also be ligated to the amplicons with the primers, a technique used for example in Nanopore sequencing.

With the cost of sequencing decreasing exponentially, more effort can be put into applying technical PCR replicates to circumvent sequencing errors and other PCR related issues. When using PCR replicates they should be sequenced in separate locations on the same 96-well plate or, ideally, with replicates in separate plates.Taxa identification lies at the core of any ecological research question. Thus, it is crucial to perform a reliable and reproducible identification workflow to ensure correct identification. In general, care should be taken to avoid cross-contamination between samples by working in clean laboratories with filter-tipped pipettes and separate pre- and post-PCR labs. Normalisation of the amplicons prior to library construction is crucial to avoid overamplification of the most represented taxa in the sample. Since some often-used plant-specific marker regions are very short (e.g., *trnL* P6 loop, 8 to152 bp), they are prone to picking up the slightest contaminants from the environment. It is therefore recommended to work in a clean environment, e.g. an ancient DNA laboratory with protective clothing.

Sequencing platforms

The preferred platforms for sequencing are currently IonTorrent and Illumina. Both platforms require an additional post-ligation PCR-step or PCR-free ligation of platform-specific adapters to the amplicons before sequencing. However, due to the different technologies behind both platforms, both the error rates and error types can differ. For Illumina (optical sequencing), a substitution error rate of 0.1% has been identified, while IonTorrent (based on detection of hydrogen ions) can show up to 1% indel errors (Quail et al. 2012; Shin et al. 2017). The IonTorrent platform has a slightly higher error rate when the material contains high amounts of homopolymers because no good correlation exists between the number of identical bases incorporated and the observed voltage change (Bragg et al. 2013). Illumina is the most often used platform in metabarcoding studies due to its lower error rates, and the generation of relatively long reads by paired-ending (Forin-Wiart et al. 2018). Since IonTorrent and Illumina are limited in the maximum length of amplicons that can be generated (up to 600 bp), more recent sequencing

platforms like Nanopore and PacBio are increasingly being used. These long read technologies have the advantage of being able to retrieve for example the whole nuclear ITS or plastid *matK* regions. For more information on sequencing platforms, please refer to Chapter 9 Sequencing platforms and data types.

Bioinformatics tools

Several different bioinformatic tools can be used to analyse the sequence output. Some commonly used packages are OBITools (Boyer et al. 2016), BEGUM (Yang et al. 2020), MOTHUR (Schloss et al. 2009), QIIME (Caporaso et al. 2010), and DADA2 (Callahan et al. 2016). The bioinformatics workflow includes these common steps: quality check of raw reads, removal of adapter sequences, demultiplexing, filtering of erroneous sequences, sequence dereplication, removal of singletons and PCR/sequencing errors, clustering/denoising, and taxonomic annotations using reference databases (most commonly using BLASTn). Depending on the pipelines used, sequences are either clustered into OTUs based on sequence similarity level (often 97%) such as in QIIME, MOTHUR, VSEARCH, or denoised into strictly unique sequences called ASVs such as in DADA2. The choice to cluster sequences into OTUs or denoise into ASVs is dependent on the research question. Clustering sequences into OTUs reduces sequencing errors, but increases false negatives as multiple similar species are clustered into a single OTU. In datasets where it is expected that closely related species are present, such as species with homopolymers (e.g., *Vaccinium spp*), denoising sequences into ASVs would be preferred since these homopolymers can be sorted out into separate sequence variants. However, using this technique may also result in artificially inflating diversity as species may have more than one sequence variant, especially if the reference database used is incomplete. Alternatively, sequences can also be assigned directly to taxons such as in OBITools, one of the most frequently used open-source programs for plant metabarcoding studies. OBITools was specifically designed for the analysis of metabarcoding data generated from HTS. It relies on filtering and sorting algorithms, which allows users to customise their pipelines tailored to their needs. A distinct feature of OBITools is its ability to account for taxonomic annotations, which allows the sorting of sequences based on taxonomy instead of OTUs/ASVs.

Future of metabarcoding

Currently, metabarcoding is the dominant technique used in the identification of plants from mixed samples. Developments and improvements in addressing methodological challenges such as PCR bias may one day allow for unbiased quantitative inferences from metabarcoding datasets. This would be a huge step forward for the metabarcoding community since it is still controversial to use read counts as an indication for biomass (Deagle et al. 2019). With the continued advances in HTS technologies coupled with the inherent limitations of metabarcoding, there is also a possibility that alternative HTS techniques can be used in the future. For example, the development of more regional DNA reference databases based on whole organelle genomes instead of single barcode regions (Coissac et al. 2016) (see Chapter 10 DNA barcoding) would encourage the use of HTS techniques that rely on whole genomes or multiple non-standard barcode regions for taxonomic identification. Particularly, if sequencing becomes cheaper and if the limitations of metagenomics (see Chapter 12 Metagenomics) or target capture (see Chapter 14 Target capture) are addressed, we may see an increase in other types of methods used to identify plants in mixed

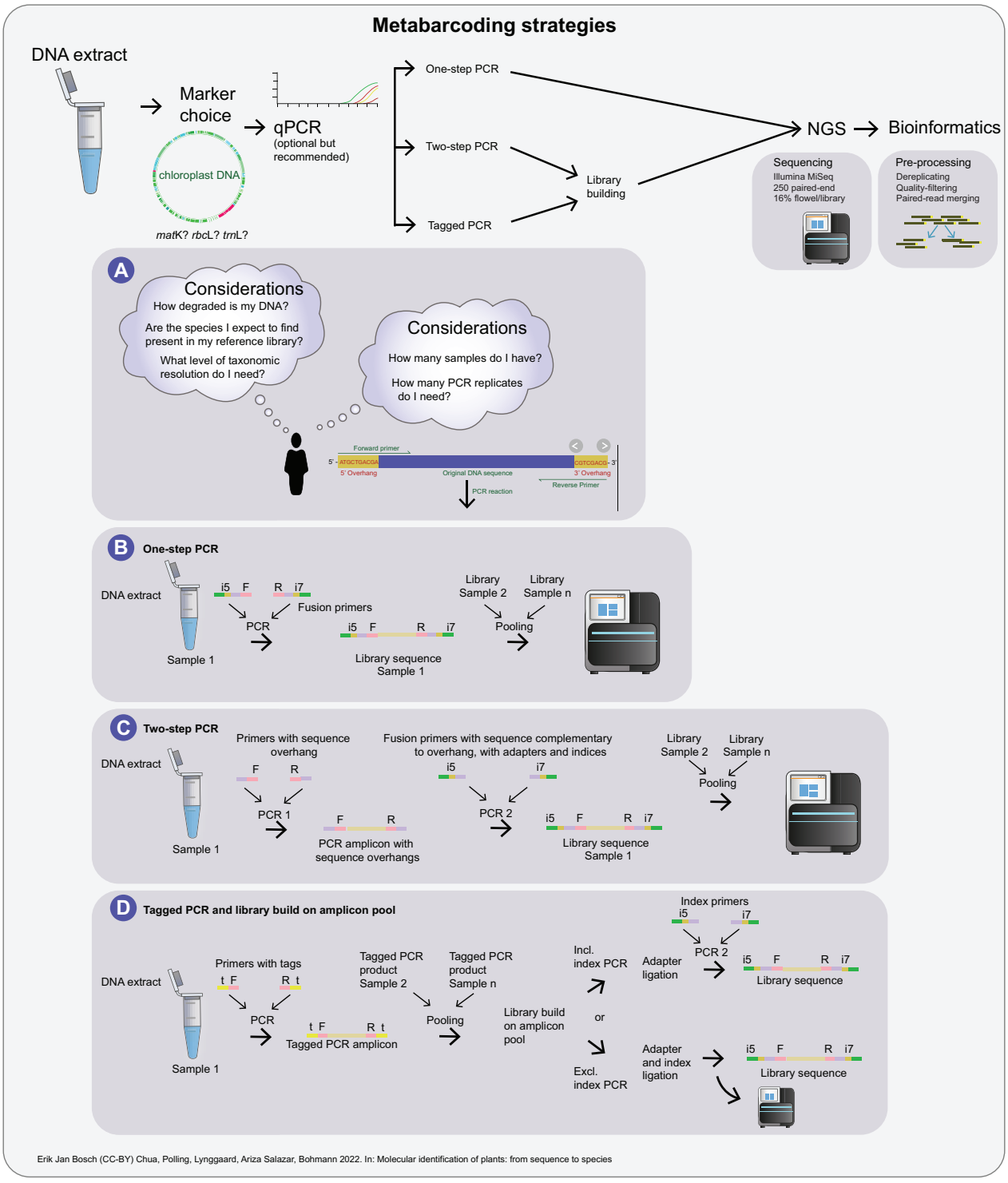

Figure 1. Chapter 11 Infographic: Visual representation of the content of this chapter.

templates. However, metabarcoding has the advantage of being a cheaper option, where large numbers of samples can be processed for meaningful statistical analysis. Bioinformatics pipelines are also well-established and better reference databases are available for mini barcodes as compared to whole organelles. This makes metabarcoding the preferred technique for many applications. In addition, ongoing efforts to build curated reference databases, design better primers, and detect potential plant-specific barcode regions might increase species resolution and circumvent many of the drawbacks associated with metabarcoding (Chua et al. 2021c).

Metabarcoding could potentially be used to determine plant composition in a landscape from bulk arthropod samples. Bulk arthropod samples have been used for biodiversity monitoring of vertebrates (Lynggaard et al. 2019), but it has not been used for any plant-related studies. Another potential application of metabarcoding is in forensic genetics (see Chapter 26 Forensic genetics, botany and palynology), where plants are used as evidence in criminal investigations (Bryant 2013). For example, morphological identification of pollen grains has been used to solve murders and determine marijuana distribution locations (Alotaibi et al. 2020; Bryant 2013). However, metabarcoding is underutilised in these applications where morphological identification is still the main technique. One possible limiting factor for this lack of utilisation could be that pollen DNA extraction destroys the samples and therefore cannot be stored as evidence (Bell et al. 2016). Metabarcoding could also potentially be used in meta-phylogeographic studies to simultaneously study the phylogeographic features and intraspecies patterns of many species (Turon et al. 2019).

Questions

1. How can overamplification of the most represented taxa in a single sequencing run of multiple complex mixtures be avoided?
2. Which DNA barcode region is most suitable for dealing with plant DNA from samples where DNA is expected to be degraded?
3. The nuclear ribosomal ITS region is shared between plants and fungi. How can undesirable fungal DNA amplification be avoided?

Glossary

Adapters – Specific nucleotide sequences unique to different types of sequencing platforms that are added to amplicon libraries to allow for the attachment of library fragments to the flow cell for sequencing.
Amplicons – Products of PCR amplification.
ASVs – Amplicon sequence variants are also known as exact sequence variants or zero-radius OTUs. Although sometimes considered synonymous to OTUs, they correspond to all the unique reads in a dataset and do not require clustering used in creating OTUs.
Barcode – Targeted gene region, see Locus.
Demultiplexing – Bioinformatics step of assigning sequences to samples based on assigned nucleotide tags and/or library indexes.
Epilithic – Plant growing on surfaces of rocks, e.g., seaweeds.
Homopolymers – Nucleotide repetition, usually in tandem of more than 7 nucleotides.
Indel errors – Insertions or deletions in sequences resulting from mutations.
ITS – The internal transcribed spacer is a nuclear ribosomal region found between the small subunit ribosomal RNA (rRNA) and large-subunit rRNA genes.
Library indexes – Nucleotide index added to amplicon libraries to allow for the parallel sequencing of multiple libraries, which can be used bioinformatically to assign reads to the correct amplicon libraries.
Locus – Section and position in a chromosome where a particular DNA sequence is located. It can also be referred to as a barcode.

Macrofossils – Preserved plant remains large enough to be seen without a microscope.
matK – Maturase K is a gene found in the chloroplast genome.
Meta-phylogeography – Study of phylogeographic features and intraspecies variation.
Multiplexing – Parallel amplification of barcodes in one PCR reaction.
OTU – Operational taxonomic unit. The term is used to categorise clusters of similar sequences.
Overhangs – Stretch of unpaired nucleotides at the end of DNA fragments.
PCR – Polymerase chain reaction.
PCR stochasticity – Uneven amplification of molecules during PCR that can be a result of some sequences being present in lower copy numbers than others.
Phylogeography – Investigate the origin of genetic variation within closely related species across a landscape.
Primers – A short single-stranded nucleic acid sequence that serves as a starting point for the DNA replication in the PCR.
Primer set – Nucleic acid sequences explained above complementary to the 5′ end and 3′ end of the flanking regions of a loci.
Primer bias – Differences in DNA amplification due to a primer inefficiently binding to the target template. This can result from sequence divergence in the primer binding sites.
qPCR – Polymerase chain reaction used for quantifying DNA.
rbcL – The ribulose-1,5-bisphosphate carboxylase large subunit gene is found in the chloroplast genome.
Singletons – A sequence only present in one copy.
Nucleotide tags – Short nucleotide sequences added at the 5′ end of the primer in metabarcoding studies.
Tag jumps – Generation of amplicons with different tags than originally used, resulting in false positives in the data. For more detail see (Schnell et al. 2015).
Taxa – Plural of taxon. A taxon is a group of organisms that form a taxonomic group.
Taxonomic assignment – Matching the obtained sequences to taxa names.
trnH-psbA – An intergenic spacer region found in the chloroplast genome.
trnL – The *trnL* gene is part of the *trnL-F* region of the chloroplast genome.

References

Alotaibi SS, Sayed SM, Alosaimi M, Alharthi R, Banjar A, Abdulqader N, Alhamed R (2020) Pollen molecular biology: applications in the forensic palynology and future prospects: A review. Saudi J. Biol. Sci. 27, 1185–1190. https://doi.org/10.1016/j.sjbs.2020.02.019

Alsos IG, Ehrich D, Seidenkrantz M-S, Bennike O, Kirchhefer AJ, Geirsdottir A (2016) The role of sea ice for vascular plant dispersal in the Arctic. Biol. Lett. 12. https://doi.org/10.1098/rsbl.2016.0264

Alsos IG, Lammers Y, Yoccoz NG, Jørgensen T, Sjögren P, Gielly L, Edwards ME (2018) Plant DNA metabarcoding of lake sediments: How does it represent the contemporary vegetation. PLoS ONE 13, e0195403. https://doi.org/10.1371/journal.pone.0195403

Apothéloz-Perret-Gentil L, Cordonier A, Straub F, Iseli J, Esling P, Pawlowski J (2017) Taxonomy-free molecular diatom index for high-throughput eDNA biomonitoring. Mol. Ecol. Resour. 17, 1231–1242. https://doi.org/10.1111/1755-0998.12668

Ariza M, Fouks B, Mauvisseau Q, Halvorsen R, Alsos IG, de Boer HJ (2022) Plant biodiversity assessment through soil eDNA reflects temporal and local diversity. Methods Ecol. Evol., 00, 1–16. https://doi.org/10.1111/2041-210X.13865

Arulandhu AJ, Staats M, Hagelaar R, Voorhuijzen MM, Prins TW, Scholtens I, Costessi A, Duijsings D, Rechenmann F, Gaspar FB, Barreto Crespo MT, Holst-Jensen A, Birck M, Burns M, Haynes E, Hochegger R, Klingl A, Lundberg L, Natale

C, Niekamp H, Kok E (2017) Development and validation of a multi-locus DNA metabarcoding method to identify endangered species in complex samples. Gigascience 6, 1–18. https://doi.org/10.1093/gigascience/gix080

Banchi E, Ametrano CG, Greco S, Stanković D, Muggia L, Pallavicini A (2020) PLANiTS: a curated sequence reference dataset for plant ITS DNA metabarcoding. Database (Oxford) 2020. https://doi.org/10.1093/database/baz155

Bell KL, Burgess KS, Okamoto KC, Aranda R, Brosi BJ (2016) Review and future prospects for DNA barcoding methods in forensic palynology. Forensic Sci. Int. Genet. 21, 110–116. https://doi.org/10.1016/j.fsigen.2015.12.010

Binladen J, Gilbert MTP, Bollback JP, Panitz F, Bendixen C, Nielsen R, Willerslev E (2007) The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. PLoS ONE 2, e197. https://doi.org/10.1371/journal.pone.0000197

Bista I, Carvalho GR, Tang M, Walsh K, Zhou X, Hajibabaei M, Shokralla S, Seymour M, Bradley D, Liu S, Christmas M, Creer S (2018) Performance of amplicon and shotgun sequencing for accurate biomass estimation in invertebrate community samples. Mol. Ecol. Resour. https://doi.org/10.1111/1755-0998.12888

Boyer F, Mercier C, Bonin A, Le Bras Y, Taberlet P, Coissac E (2016) obitools: a unix-inspired software package for DNA metabarcoding. Mol. Ecol. Resour. 16, 176–182. https://doi.org/10.1111/1755-0998.12428

Bradley BJ, Stiller M, Doran-Sheehy DM, Harris T, Chapman CA, Vigilant L, Poinar H (2007) Plant DNA sequences from feces: potential means for assessing diets of wild primates. Am. J. Primatol. 69, 699–705. https://doi.org/10.1002/ajp.20384

Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW (2013) Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. PLoS Comput. Biol. 9, e1003031. https://doi.org/10.1371/journal.pcbi.1003031

Braukmann TWA, Kuzmina ML, Sills J, Zakharov EV, Hebert PDN (2017) Testing the efficacy of DNA barcodes for identifying the vascular plants of canada. PLoS ONE 12, e0169515. https://doi.org/10.1371/journal.pone.0169515

Bryant VM (2013) Use of quaternary proxies in forensic science | analytical techniques in forensic palynology, in: Encyclopedia of Quaternary Science. Elsevier, pp. 556–566. https://doi.org/10.1016/B978-0-444-53643-3.00363-0

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP (2016) DADA2: High-resolution sample inference from Illumina amplicon data. Nat. Methods 13, 581–583. https://doi.org/10.1038/nmeth.3869

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Knight R (2010) QIIME allows analysis of high-throughput community sequencing data. Nat. Methods 7, 335–336. https://doi.org/10.1038/nmeth.f.303

Carøe C, Bohmann K (2020) Tagsteady: a metabarcoding library preparation protocol to avoid false assignment of sequences to samples. BioRxiv. https://doi.org/10.1101/2020.01.22.915009

CBOL Plant Working Group (2009) A DNA barcode for land plants. Proc Natl Acad Sci USA 106, 12794–12797. https://doi.org/10.1073/pnas.0905845106

Chase MW, Cowan RS, Hollingsworth PM, van den Berg C, Madriñán S, Petersen G, Seberg O, Jørgsensen T, Cameron KM, Carine M, Pedersen N, Hedderson TAJ, Conrad F, Salazar GA, Richardson JE, Hollingsworth ML, Barraclough TG, Kelly L, Wilkinson M (2007) A proposal for a standardised protocol to barcode all land plants. Taxon 56, 295–299. https://doi.org/10.1002/tax.562004

Cheng T, Xu C, Lei L, Li C, Zhang Y, Zhou S (2016) Barcoding the kingdom Plantae: new PCR primers for ITS regions of plants with improved universality and specificity. Mol. Ecol. Resour. 16, 138–149. https://doi.org/10.1111/1755-0998.12438

Chen S, Yao H, Han J, Liu C, Song J, Shi L, Zhu Y, Ma X, Gao T, Pang X, Luo K, Li Y, Li X, Jia X, Lin Y, Leon C (2010) Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. PLoS ONE 5, e8613. https://doi.org/10.1371/journal.pone.0008613

China Plant BOL Group, Li D-Z, Gao L-M, Li H-T, Wang H, Ge X-J, Liu J-Q, Chen Z-D, Zhou S-L, Chen S-L, Yang J-B, Fu C-X, Zeng C-X, Yan H-F, Zhu Y-J, Sun Y-S, Chen S-Y, Zhao L, Wang K, Yang T, Duan G-W (2011) Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. Proc Natl Acad Sci USA 108, 19641–19646. https://doi.org/10.1073/pnas.1104551108

Chua PYS, Crampton-Platt A, Lammers Y, Alsos IG, Boessenkool S, Bohmann K (2021a) Metagenomics: a viable tool for reconstructing herbivore diet. Mol. Ecol. Resour. 21, 2249–2263. https://doi.org/10.1111/1755-0998.13425

Chua PYS, Lammers Y, Menoni E, Ekrem T, Bohmann K, Boessenkool S, Alsos IG (2021b) Molecular dietary analyses of western capercaillies (*Tetrao urogallus*) reveal a diverse diet. Environmental DNA 3, 1156–1171. https://doi.org/10.1002/edn3.237

Chua PYS, Leerhøi F, Langkjær EMR, Noer CL, Richter SR, Marlene E, Margaryan A, Gilbert MTP, Coissac E, Alsos IG, Boessenkool S, Bohmann K (2021c) Towards the extended barcode concept: Generating DNA reference data through genome skimming of Danish plants. BioRxiv. https://doi.org/10.1101/2021.08.11.456029

Coissac E, Hollingsworth PM, Lavergne S, Taberlet P (2016) From barcodes to genomes: extending the concept of DNA barcoding. Mol. Ecol. 25, 1423–1428. https://doi.org/10.1111/mec.13549

Cowan RS, Chase MW, Kress WJ, Savolainen V (2006) 300,000 species to identify: problems, progress, and prospects in DNA barcoding of land plants. Taxon 55, 611–616. https://doi.org/10.2307/25065638

Deagle BE, Thomas AC, McInnes JC, Clarke LJ, Vesterinen EJ, Clare EL, Kartzinel TR, Eveson JP (2019) Counting with DNA in metabarcoding studies: how should we convert sequence reads to dietary data? Mol. Ecol. 28, 391–406. https://doi.org/10.1111/mec.14734

De Barba M, Miquel C, Boyer F, Mercier C, Rioux D, Coissac E, Taberlet P (2014) DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: application to omnivorous diet. Mol. Ecol. Resour. 14, 306–323. https://doi.org/10.1111/1755-0998.12188

de Boer HJ, Ghorbani A, Manzanilla V, Raclariu A-C, Kreziou A, Ounjai S, Osathanunkul M, Gravendeel B (2017) DNA metabarcoding of orchid-derived products reveals widespread illegal orchid trade. Proc. Biol. Sci. 284. https://doi.org/10.1098/rspb.2017.1182

Deiner K, Bik HM, Mächler E, Seymour M, Lacoursière-Roussel A, Altermatt F, Creer S, Bista I, Lodge DM, de Vere N, Pfrender ME, Bernatchez L (2017) Environmental DNA metabarcoding: transforming how we survey animal and plant communities. Mol. Ecol. 26, 5872–5895. https://doi.org/10.1111/mec.14350

Drinkwater R, Schnell IB, Bohmann K, Bernard H, Veron G, Clare E, Gilbert MTP, Rossiter SJ (2019) Using metabarcoding to compare the suitability of two blood-feeding leech species for sampling mammalian diversity in North Borneo. Mol. Ecol. Resour. 19, 105–117. https://doi.org/10.1111/1755-0998.12943

Elbrecht V, Leese F (2015) Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass-sequence relationships with an innovative metabarcoding protocol. PLoS ONE 10, e0130324. https://doi.org/10.1371/journal.pone.0130324

Elbrecht V, Steinke D (2018) Scaling up DNA metabarcoding for freshwater macrozoobenthos monitoring. Freshw. Biol. https://doi.org/10.1111/fwb.13220

Elbrecht V, Vamos EE, Meissner K, Aroviita J, Leese F (2017) Assessing strengths and weaknesses of DNA metabarcoding-based macroinvertebrate identification for routine stream monitoring. Methods Ecol. Evol. 8, 1265–1275. https://doi.org/10.1111/2041-210X.12789

Fahner NA, Shokralla S, Baird DJ, Hajibabaei M (2016) Large-scale monitoring of plants through environmental DNA metabarcoding of soil: recovery, resolution, and annotation of four DNA markers. PLoS ONE 11, e0157505. https://doi.org/10.1371/journal.pone.0157505

Fazekas AJ, Kuzmina ML, Newmaster SG, Hollingsworth PM (2012) DNA barcoding methods for land plants. Methods Mol. Biol. 858, 223–252. https://doi.org/10.1007/978-1-61779-591-6_11

Ficetola GF, Pansu J, Bonin A, Coissac E, Giguet-Covex C, De Barba M, Gielly L, Lopes CM, Boyer F, Pompanon F, Rayé G, Taberlet P (2015) Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. Mol. Ecol. Resour. 15, 543–556. https://doi.org/10.1111/1755-0998.12338

Forin-Wiart M-A, Poulle M-L, Piry S, Cosson J-F, Larose C, Galan M (2018) Evaluating metabarcoding to analyse diet composition of species foraging in anthropogenic landscapes using Ion Torrent and Illumina sequencing. Sci. Rep. 8, 17091. https://doi.org/10.1038/s41598-018-34430-7

Galan M, Pons J-B, Tournayre O, Pierre É, Leuchtmann M, Pontier D, Charbonnel N (2018) Metabarcoding for the parallel identification of several hundred predators and their prey: Application to bat species diet analysis. Mol. Ecol. Resour. 18, 474–489. https://doi.org/10.1111/1755-0998.12749

Goldberg CS, Turner CR, Deiner K, Klymus KE, Thomsen PF, Murphy MA, Spear SF, McKee A, Oyler-McCance SJ, Cornman RS, Laramie MB, Mahon AR, Lance RF, Pilliod DS, Strickler KM, Waits LP, Fremier AK, Takahara T, Herder

JE, Taberlet P (2016) Critical considerations for the application of environmental DNA methods to detect aquatic species. Methods Ecol. Evol. https://doi.org/10.1111/2041-210X.12595

Hawkins J, de Vere N, Griffith A, Ford CR, Allainguillaume J, Hegarty MJ, Baillie L, Adams-Groom B (2015) Using DNA metabarcoding to identify the floral composition of honey: A new tool for investigating honey bee foraging preferences. PLoS ONE 10, e0134735. https://doi.org/10.1371/journal.pone.0134735

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. Proc. Biol. Sci. 270, 313–321. https://doi.org/10.1098/rspb.2002.2218

Hibert F, Taberlet P, Chave J, Scotti-Saintagne C, Sabatier D, Richard-Hansen C (2013) Unveiling the diet of elusive rainforest herbivores in next generation sequencing era? The tapir as a case study. PLoS ONE 8, e60799. https://doi.org/10.1371/journal.pone.0060799

Hollingsworth PM, Graham SW, Little DP (2011) Choosing and using a plant DNA barcode. PLoS ONE 6, e19254. https://doi.org/10.1371/journal.pone.0019254

Hollingsworth PM, Li D-Z, van der Bank M, Twyford AD (2016) Telling plant species apart with DNA: from barcodes to genomes. Philos. Trans. R. Soc. Lond. B Biol. Sci. 371. https://doi.org/10.1098/rstb.2015.0338

Ji Y, Huotari T, Roslin T, Schmidt NM, Wang J, Yu DW, Ovaskainen O (2020) SPIKEPIPE: a metagenomic pipeline for the accurate quantification of eukaryotic species occurrences and intraspecific abundance change using DNA barcodes or mitogenomes. Mol. Ecol. Resour. 20, 256–267. https://doi.org/10.1111/1755-0998.13057

Jørgensen T, Kjaer KH, Haile J, Rasmussen M, Boessenkool S, Andersen K, Coissac E, Taberlet P, Brochmann C, Orlando L, Gilbert MTP, Willerslev E (2012) Islands in the ice: detecting past vegetation on Greenlandic nunataks using historical records and sedimentary ancient DNA meta-barcoding. Mol. Ecol. 21, 1980–1988. https://doi.org/10.1111/j.1365-294X.2011.05278.x

Kajtoch Ł (2014) A DNA metabarcoding study of a polyphagous beetle dietary diversity: the utility of barcodes and sequencing techniques. Folia Biol (Krakow) 62, 223–234. https://doi.org/10.3409/fb62_3.223

Kitson JJN, Hahn C, Sands RJ, Straw NA, Evans DM, Lunt DH (2019) Detecting host-parasitoid interactions in an invasive Lepidopteran using nested tagging DNA metabarcoding. Mol. Ecol. 28, 471–483. https://doi.org/10.1111/mec.14518

Kraaijeveld K, de Weger LA, Ventayol García M, Buermans H, Frank J, Hiemstra PS, den Dunnen JT (2015) Efficient and sensitive identification and quantification of airborne pollen using next-generation DNA sequencing. Mol. Ecol. Resour. 15, 8–16. https://doi.org/10.1111/1755-0998.12288

Kress WJ, Erickson DL (2008) DNA barcodes: genes, genomics, and bioinformatics. Proc Natl Acad Sci USA 105, 2761–2762. https://doi.org/10.1073/pnas.0800476105

Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH (2005) Use of DNA barcodes to identify flowering plants. Proc Natl Acad Sci USA 102, 8369–8374. https://doi.org/10.1073/pnas.0503123102

Kress WJ (2017) Plant DNA barcodes: Applications today and in the future. J. Syst. Evol. 55, 291–307. https://doi.org/10.1111/jse.12254

Lynggaard C, Nielsen M, Santos-Bay L, Gastauer M, Oliveira G, Bohmann K (2019) Vertebrate diversity revealed by metabarcoding of bulk arthropod samples from tropical forests. Environmental DNA 1, 329–341. https://doi.org/10.1002/edn3.34

McClenaghan B, Gibson JF, Shokralla S, Hajibabaei M (2015) Discrimination of grasshopper (Orthoptera: Acrididae) diet and niche overlap using next-generation sequencing of gut contents. Ecol. Evol. 5, 3046–3055. https://doi.org/10.1002/ece3.1585

Meiklejohn KA, Damaso N, Robertson JM (2019) Assessment of BOLD and GenBank - Their accuracy and reliability for the identification of biological materials. PLoS ONE 14, e0217084. https://doi.org/10.1371/journal.pone.0217084

Miya M, Sato Y, Fukunaga T, Sado T, Poulsen JY, Sato K, Minamoto T, Yamamoto S, Yamanaka H, Araki H, Kondoh M, Iwasaki W (2015) MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: detection of more than 230 subtropical marine species. R. Soc. Open Sci. 2, 150088. https://doi.org/10.1098/rsos.150088

Moorhouse-Gann RJ, Dunn JC, de Vere N, Goder M, Cole N, Hipperson H, Symondson WOC (2018) New universal ITS2 primers for high-resolution herbivory analyses using DNA metabarcoding in both tropical and temperate zones. Sci. Rep. 8, 8542. https://doi.org/10.1038/s41598-018-26648-2

Murray DC, Coghlan ML, Bunce M (2015) From benchtop to desktop: important considerations when designing amplicon sequencing workflows. PLoS ONE 10, e0124671. https://doi.org/10.1371/journal.pone.0124671

Murray DC, Pearson SG, Fullagar R, Chase BM, Houston J, Atchison J, White NE, Bellgard MI, Clarke E, Macphail M, Gilbert MTP, Haile J, Bunce M (2012) High-throughput sequencing of ancient plant and mammal DNA preserved in herbivore middens. Quat. Sci. Rev. 58, 135–145. https://doi.org/10.1016/j.quascirev.2012.10.021

Omelchenko DO, Speranskaya AS, Ayginin AA, Khafizov K, Krinitsina AA, Fedotova AV, Pozdyshev DV, Shtratnikova VY, Kupriyanova EV, Shipulin GA, Logacheva MD (2019) Improved protocols of ITS1-based metabarcoding and their application in the analysis of plant-containing products. Genes (Basel) 10. https://doi.org/10.3390/genes10020122

Parducci L, Bennett KD, Ficetola GF, Alsos IG, Suyama Y, Wood JR, Pedersen MW (2017) Ancient plant DNA in lake sediments. New Phytol. 214, 924–942. https://doi.org/10.1111/nph.14470

Pennisi E (2007) Taxonomy. Wanted: a barcode for plants. Science 318, 190–191. https://doi.org/10.1126/science.318.5848.190

Piñol J, Senar MA, Symondson WOC (2019) The choice of universal primers and the characteristics of the species mixture determine when DNA metabarcoding can be quantitative. Mol. Ecol. 28, 407–419. https://doi.org/10.1111/mec.14776

Poinar HN, Kuch M, Sobolik KD, Barnes I, Stankiewicz AB, Kuder T, Spaulding WG, Bryant VM, Cooper A, Pääbo S (2001) A molecular analysis of dietary diversity for three archaic Native Americans. Proc Natl Acad Sci USA 98, 4317–4322. https://doi.org/10.1073/pnas.061014798

Polling M, ter Schure ATM, van Geel B, van Bokhoven T, Boessenkool S, MacKay G, Langeveld BW, Ariza M, van der Plicht H, Protopopov AV, Tikhonov A, de Boer H, Gravendeel B (2021) Multiproxy analysis of permafrost preserved faeces provides an unprecedented insight into the diets and habitats of extinct and extant megafauna. Quat. Sci. Rev. 267, 107084. https://doi.org/10.1016/j.quascirev.2021.107084

Polling M, Sin M, de Weger LA, Speksnijder AGCL, Koenders MJF, de Boer H, Gravendeel B (2022) DNA metabarcoding using nrITS2 provides highly qualitative and quantitative results for airborne pollen monitoring.Sci. Total Environ. 806(Part 1), 150468. https://doi.org/10.1016/j.scitotenv.2021.150468 Pompanon, F., Deagle, B.E., Symondson, W.O.C., Brown, D.S., Jarman, S.N., Taberlet, P., 2012. Who is eating what: diet assessment using next generation sequencing. Mol. Ecol. 21, 1931–1950. https://doi.org/10.1111/j.1365-294X.2011.05403.x

Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13, 341. https://doi.org/10.1186/1471-2164-13-341

Raclariu AC, Heinrich M, Ichim MC, de Boer H (2018) Benefits and limitations of DNA barcoding and metabarcoding in herbal product authentication. Phytochem. Anal. 29, 123–128. https://doi.org/10.1002/pca.2732

Raclariu AC, Paltinean R, Vlase L, Labarre A, Manzanilla V, Ichim MC, Crisan G, Brysting AK, de Boer H (2017) Comparative authentication of Hypericum perforatum herbal products using DNA metabarcoding, TLC and HPLC-MS. Sci. Rep. 7, 1291. https://doi.org/10.1038/s41598-017-01389-w

Reese AT, Kartzinel TR, Petrone BL, Turnbaugh PJ, Pringle RM, David LA (2019) Using DNA metabarcoding to evaluate the plant component of human diets: a proof of concept. mSystems 4. https://doi.org/10.1128/mSystems.00458-19

Riaz T, Shehzad W, Viari A, Pompanon F, Taberlet P, Coissac E (2011) ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis. Nucleic Acids Res. 39, e145. https://doi.org/10.1093/nar/gkr732

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. 75, 7537–7541. https://doi.org/10.1128/AEM.01541-09

Schnell IB, Bohmann K, Gilbert MTP (2015) Tag jumps illuminated--reducing sequence-to-sample misidentifications in metabarcoding studies. Mol. Ecol. Resour. 15, 1289–1303. https://doi.org/10.1111/1755-0998.12402

Shin S, Kim Y, Chul Oh S, Yu N, Lee S-T, Rak Choi J, Lee K-A (2017) Validation and optimization of the Ion Torrent S5 XL sequencer and Oncomine workflow for BRCA1 and BRCA2 genetic testing. Oncotarget 8, 34858–34866. https://doi.org/10.18632/oncotarget.16799

Sigsgaard EE, Nielsen IB, Carl H, Krag MA, Knudsen SW, Xing Y, Holm-Hansen TH, Møller PR, Thomsen PF (2017) Seawater environmental DNA reflects seasonality of a coastal fish community. Mar. Biol. 164, 128. https://doi.org/10.1007/s00227-017-3147-4

Smith DP, Peay KG (2014) Sequence depth, not PCR replication, improves ecological inference from next generation DNA sequencing. PLoS ONE 9, e90234. https://doi.org/10.1371/journal.pone.0090234

Smucker NJ, Pilgrim EM, Nietch CT, Darling JA, Johnson BR (2020) DNA metabarcoding effectively quantifies diatom responses to nutrients in streams. Ecol. Appl. 30, e02205. https://doi.org/10.1002/eap.2205

Sønstebø JH, Gielly L, Brysting AK, Elven R, Edwards M, Haile J, Willerslev E, Coissac E, Rioux D, Sannier J, Taberlet P, Brochmann C (2010) Using next-generation sequencing for molecular reconstruction of past Arctic vegetation and climate. Mol. Ecol. Resour. 10, 1009–1018. https://doi.org/10.1111/j.1755-0998.2010.02855.x

Steinegger M, Salzberg SL (2020) Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. Genome Biol. 21, 115. https://doi.org/10.1186/s13059-020-02023-1

Swift JF, Lance RF, Guan X, Britzke ER, Lindsay DL, Edwards CE (2018) Multifaceted DNA metabarcoding: Validation of a noninvasive, next-generation approach to studying bat populations. Evol. Appl. 11, 1120–1138. https://doi.org/10.1111/eva.12644

Taberlet P, Bonin A, Zinger L, Coissac E (2018) Environmental DNA, Oxford Scholarship Online. Oxford University Press. https://doi.org/10.1093/oso/9780198767220.001.0001

Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. Mol. Ecol. 21, 2045–2050. https://doi.org/10.1111/j.1365-294X.2012.05470.x

Timpano EK, Scheible MKR, Meiklejohn KA (2020) Optimization of the second internal transcribed spacer (ITS2) for characterizing land plants from soil. PLoS ONE 15, e0231436. https://doi.org/10.1371/journal.pone.0231436

Turon X, Antich A, Palacín C, Præbel K, Wangensteen OS (2019) From metabarcoding to metaphylogeography: separating the wheat from the chaff. BioRxiv. https://doi.org/10.1101/629535

Valentini A, Miquel C, Nawaz MA, Bellemain E, Coissac E, Pompanon F, Gielly L, Cruaud C, Nascetti G, Wincker P, Swenson JE, Taberlet P (2009a) New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the trnL approach. Mol. Ecol. Resour. 9, 51–60. https://doi.org/10.1111/j.1755-0998.2008.02352.x

Valentini A, Pompanon F, Taberlet P (2009b) DNA barcoding for ecologists. Trends Ecol. Evol. 24, 110–117. https://doi.org/10.1016/j.tree.2008.09.011

Weigand H, Beermann AJ, Čiampor F, Costa FO, Csabai Z, Duarte S, Geiger MF, Grabowski M, Rimet F, Rulik B, Strand M, Szucsich N, Weigand AM, Willassen E, Wyler SA, Bouchez A, Borja A, Čiamporová-Zaťovičová Z, Ferreira S, Dijkstra K-DB, Ekrem T (2019) DNA barcode reference libraries for the monitoring of aquatic biota in Europe: Gap-analysis and recommendations for future work. Sci. Total Environ. 678, 499–524. https://doi.org/10.1016/j.scitotenv.2019.04.247

Willerslev E, Cappellini E, Boomsma W, Nielsen R, Hebsgaard MB, Brand TB, Hofreiter M, Bunce M, Poinar HN, Dahl-Jensen D, Johnsen S, Steffensen JP, Bennike O, Schwenninger J-L, Nathan R, Armitage S, de Hoog C-J, Alfimov V, Christl M, Beer J, Collins MJ (2007) Ancient biomolecules from deep ice cores reveal a forested southern Greenland. Science 317, 111–114. https://doi.org/10.1126/science.1141758

Yang C, Bohmann K, Wang X, Cai W, Wales N, Ding Z, Gopalakrishnan S, Yu DW (2020) Biodiversity Soup II: A bulk-sample metabarcoding pipeline emphasizing error reduction. BioRxiv. https://doi.org/10.1101/2020.07.07.187666

Yao H, Song J, Liu C, Luo K, Han J, Li Y, Pang X, Xu H, Zhu Y, Xiao P, Chen S (2010) Use of ITS2 region as the universal DNA barcode for plants and animals. PLoS ONE 5. https://doi.org/10.1371/journal.pone.0013102

Yoccoz NG, Bråthen KA, Gielly L, Haile J, Edwards ME, Goslar T, Von Stedingk H, Brysting AK, Coissac E, Pompanon F, Sønstebø JH, Miquel C, Valentini A, De Bello F, Chave J, Thuiller W, Wincker P, Cruaud C, Gavory F, Rasmussen M, Taberlet P (2012) DNA from soil mirrors plant taxonomic and growth form diversity. Mol. Ecol. 21, 3647–3655. https://doi.org/10.1111/j.1365-294X.2012.05545.x

Yu DW, Ji Y, Emerson BC, Wang X, Ye C, Yang C, Ding Z (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. Methods in Ecology and Evolution 3, 613–623. https://doi.org/10.1111/j.2041-210X.2012.00198.x

Zarrei M, Talent N, Kuzmina M, Lee J, Lund J, Shipley PR, Stefanović S, Dickinson TA (2015) DNA barcodes from four loci provide poor resolution of taxonomic groups in the genus Crataegus. AoB Plants 7. https://doi.org/10.1093/aobpla/plv045

Zimmermann J, Glöckner G, Jahn R, Enke N, Gemeinholzer B (2015) Metabarcoding vs. morphological identification to assess diatom diversity in environmental studies. Mol. Ecol. Resour. 15, 526–542. https://doi.org/10.1111/1755-0998.12336

Zinger L, Bonin A, Alsos IG, Bálint M, Bik H, Boyer F, Chariton AA, Creer S, Coissac E, Deagle BE, De Barba M, Dickie IA, Dumbrell AJ, Ficetola GF, Fierer N, Fumagalli L, Gilbert MTP, Jarman S, Jumpponen A, Kauserud H, Taberlet P (2019) DNA metabarcoding-Need for robust experimental designs to draw sound ecological conclusions. Mol. Ecol. 28, 1857–1862. https://doi.org/10.1111/mec.15060

Answers

1. By using equimolar pooling of individual samples.
2. The highly stable P6 loop can best be targeted in this case, using *trnL* primers.
3. By using plant-specific ITS primers that minimise the amplification of fungal DNA.

Chapter 12
Metagenomics

Physilia Chua[1], Youri Lammers[3], Ozan Çiftçi[4]

1 Section for Evolutionary Genomics, Globe Institute, University of Copenhagen, Copenhagen, Denmark
2 Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK
3 The Arctic University Museum of Norway, UiT - The Arctic University of Norway, Tromsø, Norway
4 Institute of Environmental Sciences, Leiden University, Leiden, The Netherlands

Physilia Chua **physiliachua@gmail.com**
Youri Lammers **youri.lammers@uit.no**
Ozan Çiftçi **ozancift@gmail.com**

Citation: Chua P, Lammers Y, Çiftçi O (2022) Chapter 12. Metagenomics. In: de Boer H, Rydmark MO, Verstraete B, Gravendeel B (Eds) Molecular identification of plants: from sequence to species. Advanced Books. https://doi.org/10.3897/ab.e98875

Background

Metagenomics is the study of genetic material recovered directly from environmental samples such as air, water, soil, or sediments (Bashir et al. 2014). It is also referred to as environmental genomics, ecogenomics, or community genomics (Guazzaroni et al. 2009). The DNA found in environmental samples are usually a mixture of genetic materials from multiple organisms. Typically, genomic DNA extracted from environmental samples is shotgun sequenced to identify organisms and/or make metabolic and other protein predictions (Porter and Hajibabaei 2018). Prior to sequencing, DNA molecules are first fragmented into smaller pieces. DNA molecules from samples are first randomly fragmented into size-controlled fragments. These fragments are then subsequently converted into libraries which consist of the DNA fragments attached to adapters specific to the sequencing platform used. Each library is then sequenced using the shotgun sequencing approach, often without targeted PCR amplification (Noonan et al. 2005). This provides a distinct advantage over PCR-based methods by enabling a less biassed investigation of a community, and the detection of all genes in a sample.

History of metagenomics

The term 'metagenome' was first coined in 1998 by Handelsman et al. (Handelsman et al. 1998). Their approach involved cloning environmental DNA extracted from soil into *E. coli* vectors, and screening the phenotypes for functional analysis of the soil microbiome. Earlier studies employed cloning techniques from environmental samples, although the term 'metagenome' had not been in use yet. For example, Stein et al. used the DNA extracted directly from seawater to investigate novel metabolisms in the marine Archaea clade Crenarchaeota (Stein et al. 1996). Similarly, Healy et al. cloned gene libraries obtained from thermophilic anaerobic microbes to discover new enzymes for biotechnological applications (Healy et al. 1995). These earlier metagenomic studies employed a functional approach by cloning genes into vectors and screening for biochemical functions, which is now more commonly referred to as functional metagenomics.

With the development of high-throughput sequencing (HTS) technologies, the need for cloning to increase the amount of starting material was eliminated. An early study recovered the first near-complete genomes of five dominant members of a natural acidophilic biofilm using an insert plasmid library and shotgun sequencing (Tyson et al. 2004). The first application of HTS to capture all representative sequences from an environmental sample was led by Venter et al. (Venter et al. 2004) in the same year. They applied shotgun sequencing to water samples collected from the Sargasso Sea, demonstrating the potential of the method to reveal the composition and function of a diverse group of microbial organisms.

The immense amount of data collected by these methods introduced challenges in data analysis, resulting in several innovations in comparative metagenomics such as clustering orthologs (Tyson et al. 2004; Yooseph et al. 2007), use of GC content to distinguish genomes (Tyson et al. 2004), and single-copy genes to check for completeness and genome size predictions (Ciccarelli et al. 2006; Raes et al. 2007; Turnbaugh et al. 2007). With these innovations and an increasing number of available reference genomes, the application of metagenomics has expanded outside its traditional use in microbial research.

Suitable samples types

Similar to metabarcoding, substrates that can be used for metagenomics in plant identification include environmental samples, fragmented template materials (i.e., dental calculus and faeces) (Weyrich et al. 2017), mixed food templates (i.e., herbal medicines, protein powder), and complex samples like honey (Bovo et al. 2018). Soil samples in particular are promising for metagenomics, as the method can also provide insight into the root microbiomes of plants (Molina-Montenegro et al. 2019; Simões et al. 2015). However, metagenomic sequencing of environmental samples can be challenging as the starting material is a mixture of DNA from viral, bacterial, archeal, and eukaryotic species. Typically, the abundance of those species is also different within a sample, complicating downstream data analysis. Although it depends on the study aim, a sample with unequal abundances can be more problematic as the reads cannot be easily assembled into longer contigs (reads coming from different organisms that do not overlap). This results in a lower probability that the correct taxonomic or functional annotations are assigned (Ayling et al. 2020). To reduce the complexity of the sample and ensure that enough target DNA is obtained, fractionation, size selection, selective lysis, or enrichment can be performed (Teeling and Glöckner 2012; Thomas et al. 2012). The amount of DNA obtained from certain types of samples can be very small depending on the degradation or the amount of starting material. As library preparation protocols require a certain amount of DNA, a prior whole-genome amplification or concentration step might be desirable. Amplification can however introduce biases for metagenomic community analysis, so one should consider whether it is necessary. Another problem associated with environmental samples is the presence of inhibitors such as humic acids present in soil, but this issue has been addressed extensively where protocols have been developed to remove such inhibitors (Delmont et al. 2011).

Uses of metagenomics

Several promising applications exist for plant-related metagenomics as compared to conventional targeted genomic approaches. Dietary studies are one such application. While dietary studies have been revolutionised by conventional metabarcoding (see Chapter 11 Amplicon metabarcoding; Kartzinel et al. 2015; Moorhouse-Gann et al. 2018; Soininen et al. 2009), they can benefit from the additional sequences and genes that metagenomics provides. In addition to plant identification in diet studies, the sequenced data can also be used to simultaneously identify and genotype the host, categorise the gut microbiome, and detect parasites (Srivathsan et al. 2016). Besides dietary studies, metagenomics has also been applied for plant authentication in herbal medicines (see Chapter 22 Healthcare; Xin et al. 2018), for detection of contaminants in the food supply chain (see Chapter 23 Food safety; Haiminen et al. 2019), in palaeobotany to characterise historical environments (see Chapter 21 Palaeobotany; Pedersen et al. 2016; Stahlschmidt et al. 2019), and to describe the shifts in ecosystems with environmental change (Parducci et al. 2019). Furthermore, the ecological information collected from various samples can be applied in conservation management (see Chapter 24 Environment and biodiversity assessments).

Similar to metabarcoding (see Chapter 11 Amplicon metabarcoding), metagenomics can potentially be used to reconstruct plant compositions from bulk arthropods samples, and to solve crimes in forensic genetics (see Chapter 26 Forensic genetics, botany, and palynology), especially by uncovering taxa that are not normally amplified in metabarcoding studies. It can also potentially be applied to plant resources for the retrieval of plant population genetic information from mixed templates (which has already been shown in mammals; Srivathsan et al.

2016, 2015). Additionally, metagenomics can also be applied in water quality studies, through both quantitative and qualitative assessment of diatoms present in water bodies (Chessman et al. 2007).

Advantages and limitations

Metagenomics is an untargeted method that captures all genetic material in a sample, which is advantageous over targeted methods as no prior knowledge of the taxa and their genes is required (Pedersen et al. 2015; Quince et al. 2017). These data can be used to identify a wide range of different taxonomic groups, including bacteria, archaea, and eukaryotes (Bovo et al. 2018; Stat et al. 2017). Furthermore, metagenomics avoids biases that can be introduced during marker amplification and thus can provide a more reliable abundance estimate compared to metabarcoding (Ziesemer et al. 2015). Metagenomics can also be used to extract information from degraded material since long templates are not necessary (Parducci et al. 2019; Pedersen et al. 2016). Finally, there is the opportunity to use the metagenomic data for alternative types of analyses, such as genomic reconstruction or gene discovery (Molina-Montenegro et al. 2019; Quince et al. 2017). A highly significant advantage of metagenomics is that it can also be used in functional ecology, where gene expression can be studied (Mackelprang et al. 2011).

Metagenomics does, however, come with some disadvantages that need to be considered. The main downside is the taxonomic inefficiency of the method. Sequenced material can originate from any part of the genome, but full nuclear genome references for most species are currently lacking. Thus, only a small proportion of species can currently be identified (Chua et al. 2021; Parducci et al. 2019; Srivathsan et al. 2016; Stat et al. 2017). This problem is exacerbated for multicellular organisms, which have a lower abundance compared to microbes in an environmental sample (Azam and Malfatti 2007) and can therefore have a smaller proportion of reads (Stat et al. 2017). However, with the sequencing of whole genomes currently underway for more plant species, issues related to unavailable reference data are becoming less problematic (Alsos et al. 2020; Chua et al. 2021; Li et al. 2019; Nevill et al. 2020). Furthermore, the low number of assigned metagenomic reads can be addressed by increasing the sequencing depth, though at an increased cost.

The process of metagenomics

Step by step laboratory workflow

DNA fragmentation

DNA fragmentation is an essential step in the metagenomic workflow, and the size of the DNA fragments required depends on the sequencing platform used. Broadly speaking, there are two methods for DNA fragmentation to obtain size-controlled DNA fragments: enzyme-based and mechanical. Each method has its associated advantages and disadvantages (Li et al. 2017). Enzyme-based methods generally use transposons, restriction enzymes, or nicking enzymes to fragment the DNA (Anderson 1981; Hoheisel et al. 1989; Seed et al. 1982; Wong et al. 1997). Although these enzyme-based methods are precise and efficient for fragmenting DNA, the fragments are not randomly fragmented, and enzymatic digestion is less efficient for DNA with high GC content (Kasoji et al. 2015; Thorstenson et al. 1998). Mechanical methods typically use sonication (Deininger 1983; Kasoji et al. 2015; Tseng et al. 2012), nebulisation (Lentz et al. 2005; Sambrook and Russell 2006), or hydrodynamic shearing (Joneja and Huang 2009; Shui et al.

2011). These methods provide better random fragmentation with increased size and distribution control as compared to enzyme-based fragmentation methods (Hengen 1997). However, while sonication is efficient and easy to use, it can cause breaks within AT-rich regions, resulting in damaged DNA fragments that cannot be sequenced (Hengen 1997). Nebulisation is a fast method for DNA fragmentation, but the DNA fragments have a wider size range and require expensive equipment. Hydrodynamic shearing produces short DNA fragments with less damage, and with a narrow size distribution. However, this method requires complex machinery and trained users (Hengen 1997; Shui et al. 2011). The choice of method for DNA fragmentation thus depends on the final fragment size required, the choice of sequencing platform, the amount of input DNA, funding, and scalability. The most important consideration is that the method must sufficiently randomly fragment the DNA so that the sequencing libraries will fully represent the starting DNA template.

Library preparation

Library preparation is another important step in the metagenomics workflow as it can affect the results of the sequencing output. The addition of adapters to the ends of DNA fragments lets it bind to the sequencing flow cell, which allows for the identification of the reads (DeWitt 2019). There are two types of library preparation: ligation-based and tagmentation. In ligation-based library preparation, DNA fragmentation and adapter ligation occur in two separate steps. Library preparation usually entails the use of double-stranded fragmented DNA as input, end-repair, 5′ end phosphorylation and A-tailing of 3′ end, adapters ligation, and PCR enrichment of adapters-ligated DNA fragments (optional) (Carøe et al. 2017; Head et al. 2014). For this method, an optimal adapter to fragment ratio (~10:1) has to be calculated as an excess of adapters may lead to the formation of adapter-dimers that can be over-amplified in the PCR step. Depending on the amount of starting DNA, amplification-free library building can be carried out if enough DNA material can be extracted (~250 ng) (Genohub 2018). The more starting material there is, the less amplification is required. In tagmentation library preparation, such as the Nextera DNA Sample Prep Kit (Illumina), DNA fragmentation and adapter ligation occur together in one reaction (Hennig et al. 2018). Libraries are prepared using a transposase enzyme which simultaneously fragments and adds adapters to the DNA (Adey et al. 2010). The first step is the tagmentation reaction, where the transposase enzyme cleaves and tags the input double-stranded unfragmented DNA with a universal overhang. This tagmentation step determines the success of the library preparation, and successful tagmentation is highly sensitive towards and dependent on the amount of input DNA (< 1 ng overtagmentation, > 1 ng undertagmentation) (Illumina 2015). This method is also sensitive to temperature and reaction time.

Sequencing approaches and platforms

DNA sequencing has gradually shifted from Sanger to HTS technologies in the last decades. These new sequencing technologies can provide much higher yields of reads at a much lower cost (see Chapter 9 Sequencing platforms and data types). Initially, 454/Roche pyrosequencing (discontinued) was the most widely used platform (Edwards et al. 2006; Thomas et al. 2012). However, the generation of artificial replicate reads and systematic homopolymer errors limits its use for metagenomic applications. Illumina sequencing offers short read lengths up to 300 bp (paired-end), generates high output (up to 1.5 billion bp per run) and high accuracy (error rates < 1%). Given the platform's wide availability, it became the dominant choice for shotgun metagenomics. Out of the available Illumina platforms, only the MiSeq provides 300 bp read

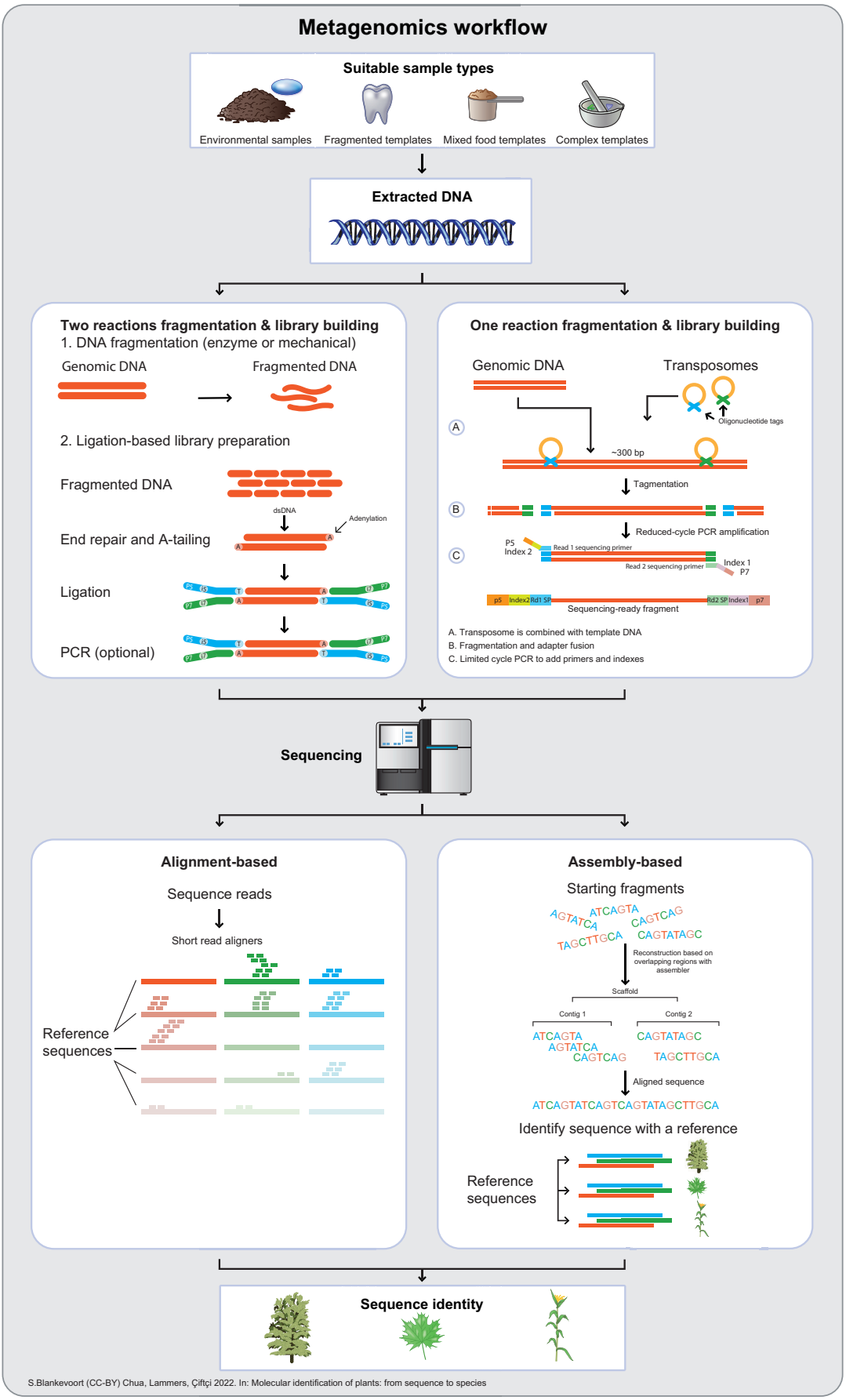

Figure 1. Chapter 12 Infographic: Visual representation of the content of this chapter.

lengths, though the total output is somewhat lower, making it more suitable for single marker surveys. Illumina HiSeq 2500/4000, NextSeq, and NovaSeq all produce higher outputs and are well suited for metagenomic applications (Quince et al. 2017).

Short reads are bioinformatically challenging for metagenomic assembly because genes and chromosomal regions can be difficult to span, especially if they are long or composed of repetitive elements. Certain protocols have been developed to overcome such challenges (e.g., assembly after binning and taxonomic assignment), but long-read sequencing technologies offer excellent alternatives for metagenomics. PacBio and Oxford Nanopore technologies offer longer read lengths but can be accompanied by higher error rates and higher costs. In contrast to the other platforms which introduce inherent systematic errors (e.g., homopolymer regions, index hopping), errors in these platforms are mostly random, which might be overcome with technological improvements (Teeling and Glöckner 2012). Additionally, they provide read lengths long enough to span multiple genes making them a promising alternative for metagenomics.

The exact number of reads required to effectively characterise a sample using metagenomics will be highly variable, and as such, no one number for the total number of reads required can be given universally. In principle, the total number of species in the sample, the genome sizes, and the relative abundance of each species should be known to make such an estimation. As a rule of thumb, it is suggested to maximise the output to capture as many reads as possible from the rare members of the community (Quince et al. 2017).

Bioinformatics strategies

There are currently two main strategies to identify the contents of a metagenomic sample: identification of individual reads by alignment to a reference, or by assembling the reads into longer contigs prior to identification.

Alignment

The most straightforward method for identification is by aligning the reads to a known reference dataset. BLAST and related tools such as MegaBLAST (Zhang et al. 2000) are commonly used for this alignment and identification of reads. Though accurate, these tools are computationally inefficient and do not scale well with increasing sizes of current metagenomic datasets and reference databases (Ye et al. 2019).

Two alternative approaches aim to speed up the identification of metagenomic datasets. These either use more compressed reference databases in combination with more efficient aligners or rely on exact alignments of k-mers between the reads and the reference (Ye et al. 2019). The Burrows-Wheeler transform (BWT) in combination with FM indexes is a good way to compress references and speed up alignments. These techniques are common for genomic mapping programs such as BWA (Li and Durbin 2009) and bowtie (Langmead and Salzberg 2012), but have also been applied in metagenomic tools such as Centrifuge (Kim et al. 2016) and MetaPhlAn (Truong et al. 2015). The alternative k-mer method uses smaller subsequences of k-length that are extracted from the metagenomic reads. The read k-mers can be directly compared to a set of k-mers from the reference database, which simplifies and speeds up the identifications. Metagenomic tools such as Kraken (Wood and Salzberg 2014) or CLARK (Ounit et al. 2015) use k-mer matching for their identifications. Both strategies are substantially faster than their traditional alignment counterparts (Ye et al. 2019), but the various tools differ from each other in terms of memory requirements, speed, and additional features. BWT-based align-

ers generally require less memory but are marginally slower than the more memory demanding k-mer aligners. The results, regardless of the method, are a set of identifications to the Last Common Ancestor (LCA), to account for conserved or homologous sequences, after which the results can be explored in tools such as MEGAN (Huson et al. 2007).

Assembly

Assembly methods attempt to generate longer contigs before downstream analysis. These longer contigs can be used for gene identifications (Quince et al. 2017) or can result in better identifications compared to shorter reads (Vestergaard et al. 2017). De Bruijn graphs are commonly used to generate de novo contigs from genomic data (Zerbino and Birney 2008). First, a de novo assembler constructs a graph of all overlapping k-mers, which are obtained from the read data. The assembler then attempts to find a path through the graph that corresponds to contigs. Metagenomic datasets can be problematic for de Bruijn graph-based assemblers, where a large pool of (possibly) closely related species with uneven coverage between taxa can result in fragmented or incorrect contigs (Quince et al. 2017; Chua et al. 2022). Dedicated metagenomic assemblers such as MetaSpades (Nurk et al. 2017), IDBA-UD (Peng et al. 2012), and MEGAHIT (Li et al. 2015) attempt to overcome these problems. These tools construct multiple graphs at different k-mer lengths to resolve the aforementioned issues. Graphs based on smaller k-mer sizes can be beneficial for the assembly of low-abundance taxa (Quince et al. 2017), while those constructed out of larger k-mers can bridge gaps and yield longer contigs for more abundant species (Peng et al. 2012). These methods have proven to be useful for microbial datasets (Pasolli et al. 2019; Qin et al. 2010), but their application in the assembly of eukaryotic genomes can be problematic given their low abundance in environmental samples and because they have more complex genomes (Azam and Malfatti 2007).

Bioinformatic summary

Each bioinformatic strategy has its pros and cons, and the decision about which strategy to use depends on the starting material available as well as the research questions to aim to be answered. The alignment method works well when there is ample reference material available for the taxa of interest, when working with older and more fragmented material, or when the target taxa are sparse in a sample. The assembly method on the other hand performs best when there is abundant material available, which is often not the case for environmental datasets.

Future of metagenomics

As sequencing costs continue to significantly decrease, bioinformatics pipelines are optimised, and more comprehensive DNA reference libraries are available (Alsos et al. 2020; Chua et al. 2021; Li et al. 2019; Nevill et al. 2020), we can expect that the number of metagenomics studies will increase. This is because metagenomics can provide better taxonomic resolution over PCR-based methods as longer and larger reference data can be utilised. The ability to retrieve almost all the DNA content present in samples without targeted enrichment or any prior knowledge of the dataset can potentially make metagenomics a powerful tool for biomonitoring, where large amounts of ecological data are often required from minute samples. Metagenomics can also simultaneously characterise the entire microbiome and infer functional information (Chua and Rasmussen 2022), taking it beyond metabarcoding by allowing for more biological questions to be explored in more detail.

Questions

1. What are the two main steps in the metagenomics laboratory workflow and why are they necessary?
2. What are the challenges of using short-read sequencing for metagenomics applications and how do you overcome these challenges?
3. What are problems caused by using environmental samples with unequal abundances in metagenomics applications?

Glossary

Basic Local Alignment Search Tool (BLAST) – An alignment tool commonly used in conjunction with the NCBI nucleotide reference database for sequence identifications. Different BLAST versions exist for nucleotide or protein alignments.
Binning – Clustering sequences based on their nucleotide composition or similarity to a reference database.
Burrows-Wheeler transform – Data transformation algorithm to make transformed data more compressible.
Community genetics – Study of genetic interactions between species and their environment in complex communities.
Contigs – A longer assembled DNA sequence.
Coverage – The mean number of times a nucleotide is sequenced in a genome.
De Bruijn graphs – A popular method for the de novo assembly of contigs. The graph is built up out of k-mers that overlap, which can be solved to construct contigs.
De novo assembly – The assembly of contigs or genomes from sequenced data without the aid of a reference.
DNA fragmentation – Separating or breaking DNA molecules into smaller fragments.
DNA libraries – DNA libraries are a collection of DNA fragments with specific sequencing-platform adapters ligated to both ends.
Ecogenomics – Study of the influence of environmental factors on the genome.
Environmental genomics – Prediction of organism responses at the genetic level.
FM-index – A compressed data structure for full-text pattern searching based on the Burrows-Wheeler transform.
Functional metagenomics – Study of gene functions from DNA extracted from mixed communities.
Hydrodynamic shearing – Fragmentation of DNA molecules by forcing them through a small tube or small gauge needle at high velocity.
K-mer – A short subsequence of length *k* that is generated from longer sequencing reads. The shorter k-mers allow for faster alignments and assemblies.
Last Common Ancestor (LCA) – A point on the tree of life from which a set of taxa are descended.
MegaBLAST – A faster, though less accurate, version of the BLAST tool.
Metagenome – All genetic material found in an environmental sample. It contains the genomes of many different organisms.
Nebulisation – Process of breaking DNA molecules into small fragments by passing DNA solution into a nebuliser unit, resulting in a fine mist that is collected.
Orthologs – Genes in different species that evolved from a common ancestral gene.

Paired-end sequencing – Sequencing of a DNA fragment from both ends. Both sequences can either be merged into a single larger fragment, if overlap is present, or kept separate.
Read – A DNA sequence generated by a sequencer.
Shotgun sequencing – A technique that randomly fragments DNA and then reassembles the fragments by searching for overlapping regions.
Sonication – Application of sound energy to break up DNA strands into smaller fragments.

References

Adey A, Morrison HG, Asan Xun X, Kitzman JO, Turner EH, Stackhouse B, MacKenzie AP, Caruccio NC, Zhang X, Shendure J (2010) Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biol. 11, R119. https://doi.org/10.1186/gb-2010-11-12-r119

Alsos IG, Lavergne S, Merkel MKF, Boleda M, Lammers Y, Alberti A, Pouchon C, Denoeud F, Pitelkova I, Pușcaș M, Roquet C, Hurdu B-I, Thuiller W, Zimmermann NE, Hollingsworth PM, Coissac E (2020) The treasure vault can be opened: large-scale genome skimming works well using herbarium and silica gel dried material. Plants 9, 432. https://doi.org/10.3390/plants9040432

Anderson S (1981) Shotgun DNA sequencing using cloned DNase I-generated fragments. Nucleic Acids Res. 9, 3015–3027. https://doi.org/10.1093/nar/9.13.3015

Ayling M, Clark MD, Leggett RM (2020) New approaches for metagenome assembly with short reads. Brief. Bioinformatics 21, 584–594. https://doi.org/10.1093/bib/bbz020

Azam F, Malfatti F (2007) Microbial structuring of marine ecosystems. Nat. Rev. Microbiol. 5, 782–791. https://doi.org/10.1038/nrmicro1747

Bashir Y, Pradeep Singh S, Kumar Konwar B (2014) Metagenomics: an application based perspective. Chinese Journal of Biology 2014, 1–7. https://doi.org/10.1155/2014/146030

Bovo S, Ribani A, Utzeri VJ, Schiavo G, Bertolini F, Fontanesi L (2018) Shotgun metagenomics of honey DNA: Evaluation of a methodological approach to describe a multi-kingdom honey bee derived environmental DNA signature. PLoS ONE 13, e0205575. https://doi.org/10.1371/journal.pone.0205575

Carøe C, Gopalakrishnan S, Vinner L, Mak SST, Sinding MHS, Samaniego JA, Wales N, Sicheritz-Pontén T, Gilbert MTP (2017) Single-tube library preparation for degraded DNA. Methods Ecol. Evol. https://doi.org/10.1111/2041-210X.12871

Chessman BC, Bate N, Gell PA, Newall P (2007) A diatom species index for bioassessment of Australian rivers. Mar. Freshwater Res. 58, 542. https://doi.org/10.1071/MF06220

Chua PYS, Leerhøi F, Langkjær EMR, Noer CL, Richter SR, Marlene E, Margaryan A, Gilbert MTP, Coissac E, Alsos IG, Boessenkool S, Bohmann K (2021) Towards the extended barcode concept: Generating DNA reference data through genome skimming of Danish plants. BioRxiv. https://doi.org/10.1101/2021.08.11.456029

Chua PYS, Rasmussen JA (2022) Taking metagenomics under the wings. Nat. Rev. Microbiol. https://doi.org/10.1038/s41579-022-00746-5

Chua PYS, Carøe C, Crampton-Platt A, Reyes-Avila C S, Jones G, Streicker D, Bohmann K (2022) A two-step metagenomics approach for the identification and mitochondrial DNA contig assembly of vertebrate prey from the blood meals of common vampire bats (*Desmodus rotundus*). Metabarcoding Metagenom. 6:e78756. https://doi.org/10.3897/mbmg.6.78756

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P (2006) Toward automatic reconstruction of a highly resolved tree of life. Science 311, 1283–1287. https://doi.org/10.1126/science.1123061

Deininger PL (1983) Random subcloning of sonicated DNA: application to shotgun DNA sequence analysis. Anal. Biochem. 129, 216–223. https://doi.org/10.1016/0003-2697(83)90072-6

Delmont TO, Robe P, Clark I, Simonet P, Vogel TM (2011) Metagenomic comparison of direct and indirect soil DNA extraction approaches. J. Microbiol. Methods 86, 397–400. https://doi.org/10.1016/j.mimet.2011.06.013

DeWitt J (2019) A simple library prep workflow for many sequencing applications. IDT.

Edwards RA, Rodriguez-Brito B, Wegley L, Haynes M, Breitbart M, Peterson DM, Saar MO, Alexander S, Alexander EC, Rohwer F (2006) Using pyrosequencing to shed light on deep mine microbial ecology. BMC Genomics 7, 57. https://doi.org/10.1186/1471-2164-7-57

Genohub (2018) Metagenomics sequencing guide. Genohub.

Guazzaroni M-E, Beloqui A, Golyshin PN, Ferrer M (2009) Metagenomics as a new technological tool to gain scientific knowledge. World J. Microbiol. Biotechnol. 25, 945–954. https://doi.org/10.1007/s11274-009-9971-z

Haiminen N, Edlund S, Chambliss D, Kunitomi M, Weimer BC, Ganesan B, Baker R, Markwell P, Davis M, Huang BC, Kong N, Prill RJ, Marlowe CH, Quintanar A, Pierre S, Dubois G, Kaufman JH, Parida L, Beck KL (2019) Food authentication from shotgun sequencing reads with an application on high protein powders. npj Sci. Food 3, 24. https://doi.org/10.1038/s41538-019-0056-6

Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem. Biol. 5, R245-9. https://doi.org/10.1016/s1074-5521(98)90108-9

Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, Ordoukhanian P (2014) Library construction for next-generation sequencing: overviews and challenges. BioTechniques 56, 61–4, 66, 68, passim. https://doi.org/10.2144/000114133

Healy FG, Ray RM, Aldrich HC, Wilkie AC, Ingram LO, Shanmugam KT (1995) Direct isolation of functional genes encoding cellulases from the microbial consortia in a thermophilic, anaerobic digester maintained on lignocellulose. Appl. Microbiol. Biotechnol. 43, 667–674. https://doi.org/10.1007/BF00164771

Hengen PN (1997) Shearing DNA for genomic library construction. Trends Biochem. Sci. 22, 273–274. https://doi.org/10.1016/s0968-0004(97)01080-3

Hennig BP, Velten L, Racke I, Tu CS, Thoms M, Rybin V, Besir H, Remans K, Steinmetz LM (2018) Large-scale low-cost NGS library preparation using a robust Tn5 purification and tagmentation protocol. G3 (Bethesda) 8, 79–89. https://doi.org/10.1534/g3.117.300257

Hoheisel JD, Nizetic D, Lehrach H (1989) Control of partial digestion combining the enzymes dam methylase and MboI. Nucleic Acids Res. 17, 9571–9582. https://doi.org/10.1093/nar/17.23.9571

Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. Genome Res. 17, 377–386. https://doi.org/10.1101/gr.5969107

Illumina (2015) Nextera XT library prep: tips and troubleshooting. Illumina.

Joneja A, Huang X (2009) A device for automated hydrodynamic shearing of genomic DNA. BioTechniques 46, 553–556. https://doi.org/10.2144/000113123

Kartzinel TR, Chen PA, Coverdale TC, Erickson DL, Kress WJ, Kuzmina ML, Rubenstein DI, Wang W, Pringle RM (2015) DNA metabarcoding illuminates dietary niche partitioning by African large herbivores. Proc Natl Acad Sci USA 112, 8019–8024. https://doi.org/10.1073/pnas.1503283112

Kasoji SK, Pattenden SG, Malc EP, Jayakody CN, Tsuruta JK, Mieczkowski PA, Janzen WP, Dayton PA (2015) Cavitation enhancing nanodroplets mediate efficient DNA fragmentation in a bench top ultrasonic water bath. PLoS ONE 10, e0133014. https://doi.org/10.1371/journal.pone.0133014

Kim D, Song L, Breitwieser FP, Salzberg SL (2016) Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. 26, 1721–1729. https://doi.org/10.1101/gr.210641.116

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359. https://doi.org/10.1038/nmeth.1923

Lentz YK, Worden LR, Anchordoquy TJ, Lengsfeld CS (2005) Effect of jet nebulization on DNA: identifying the dominant degradation mechanism and mitigation methods. J. Aerosol Sci. 36, 973–990. https://doi.org/10.1016/j.jaerosci.2004.11.017

Li D, Liu C-M, Luo R, Sadakane K, Lam T-W (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31, 1674–1676. https://doi.org/10.1093/bioinformatics/btv033

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li H-T, Yi T-S, Gao L-M, Ma P-F, Zhang T, Yang J-B, Gitzendanner MA, Fritsch PW, Cai J, Luo Y, Wang H, van der Bank M, Zhang S-D, Wang Q-F, Wang J, Zhang Z-R, Fu C-N, Yang J, Hollingsworth PM, Chase MW, Li D-Z (2019) Origin of angiosperms and the puzzle of the Jurassic gap. Nat. Plants 5, 461–470. https://doi.org/10.1038/s41477-019-0421-0

Li L, Jin M, Sun C, Wang X, Xie S, Zhou G, van den Berg A, Eijkel JCT, Shui L (2017) High efficiency hydrodynamic DNA fragmentation in a bubbling system. Sci. Rep. 7, 40745. https://doi.org/10.1038/srep40745

Mackelprang R, Waldrop MP, DeAngelis KM, David MM, Chavarria KL, Blazewicz SJ, Rubin EM, Jansson JK (2011) Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. Nature 480, 368–371. https://doi.org/10.1038/nature10576

Molina-Montenegro MA, Ballesteros GI, Castro-Nallar E, Meneses C, Gallardo-Cerda J, Torres-Díaz C (2019) A first insight into the structure and function of rhizosphere microbiota in Antarctic plants using shotgun metagenomic. Polar Biol. 42, 1825–1835. https://doi.org/10.1007/s00300-019-02556-7

Moorhouse-Gann RJ, Dunn JC, de Vere N, Goder M, Cole N, Hipperson H, Symondson WOC (2018) New universal ITS2 primers for high-resolution herbivory analyses using DNA metabarcoding in both tropical and temperate zones. Sci. Rep. 8, 8542. https://doi.org/10.1038/s41598-018-26648-2

Nevill PG, Zhong X, Tonti-Filippini J, Byrne M, Hislop M, Thiele K, van Leeuwen S, Boykin LM, Small I (2020) Large scale genome skimming from herbarium material for accurate plant identification and phylogenomics. Plant Methods 16, 1. https://doi.org/10.1186/s13007-019-0534-5

Noonan JP, Hofreiter M, Smith D, Priest JR, Rohland N, Rabeder G, Krause J, Detter JC, Pääbo S, Rubin EM (2005) Genomic sequencing of Pleistocene cave bears. Science 309, 597–599. https://doi.org/10.1126/science.1113485

Nurk S, Meleshko D, Korobeynikov A, Pevzner PA (2017) metaSPAdes: a new versatile metagenomic assembler. Genome Res. 27, 824–834. https://doi.org/10.1101/gr.213959.116

Ounit R, Wanamaker S, Close TJ, Lonardi S (2015) CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics 16, 236. https://doi.org/10.1186/s12864-015-1419-2

Parducci L, Alsos IG, Unneberg P, Pedersen MW, Han L, Lammers Y, Salonen JS, Väliranta MM, Slotte T, Wohlfarth B (2019) Shotgun environmental DNA, pollen, and macrofossil analysis of lateglacial lake sediments from southern Sweden. Front. Ecol. Evol. 7. https://doi.org/10.3389/fevo.2019.00189

Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, Beghini F, Manghi P, Tett A, Ghensi P, Collado MC, Rice BL, DuLong C, Morgan XC, Golden CD, Quince C, Huttenhower C, Segata N (2019) Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. Cell 176, 649-662.e20. https://doi.org/10.1016/j.cell.2019.01.001

Pedersen MW, Overballe-Petersen S, Ermini L, Sarkissian CD, Haile J, Hellstrom M, Spens J, Thomsen PF, Bohmann K, Cappellini E, Schnell IB, Wales NA, Carøe C, Campos PF, Schmidt AMZ, Gilbert MTP, Hansen AJ, Orlando L, Willerslev E (2015) Ancient and modern environmental DNA. Philos. Trans. R. Soc. Lond. B Biol. Sci. 370, 20130383. https://doi.org/10.1098/rstb.2013.0383

Pedersen MW, Ruter A, Schweger C, Friebe H, Staff RA, Kjeldsen KK, Mendoza MLZ, Beaudoin AB, Zutter C, Larsen NK, Potter BA, Nielsen R, Rainville RA, Orlando L, Meltzer DJ, Kjær KH, Willerslev E (2016) Postglacial viability and colonization in North America's ice-free corridor. Nature 537, 45–49. https://doi.org/10.1038/nature19085

Peng Y, Leung HCM, Yiu SM, Chin FYL (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28, 1420–1428. https://doi.org/10.1093/bioinformatics/bts174

Porter TM, Hajibabaei M (2018) Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. Mol. Ecol. 27, 313–338. https://doi.org/10.1111/mec.14478

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Wang J (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464, 59–65. https://doi.org/10.1038/nature08821

Quince C, Walker AW, Simpson JT, Loman NJ, Segata N (2017) Shotgun metagenomics, from sampling to analysis. Nat. Biotechnol. 35, 833–844. https://doi.org/10.1038/nbt.3935

Raes J, Korbel JO, Lercher MJ, von Mering C, Bork P (2007) Prediction of effective genome size in metagenomic samples. Genome Biol. 8, R10. https://doi.org/10.1186/gb-2007-8-1-r10

Sambrook J, Russell DW (2006) Fragmentation of DNA by nebulization. CSH Protoc. 2006. https://doi.org/10.1101/pdb.prot4539

Seed B, Parker RC, Davidson N (1982) Representation of DNA sequences in recombinant DNA libraries prepared by restriction enzyme partial digestion. Gene 19, 201–209.

Shui L, Bomer JG, Jin M, Carlen ET, van den Berg A (2011) Microfluidic DNA fragmentation for on-chip genomic analysis. Nanotechnology 22, 494013. https://doi.org/10.1088/0957-4484/22/49/494013

Simões MF, Antunes A, Ottoni CA, Amini MS, Alam I, Alzubaidy H, Mokhtar N-A, Archer JAC, Bajic VB (2015) Soil and rhizosphere associated fungi in gray mangroves (*Avicennia marina*) from the Red Sea – A metagenomic approach. Genomics Proteomics Bioinformatics 13, 310–320. https://doi.org/10.1016/j.gpb.2015.07.002

Soininen EM, Valentini A, Coissac E, Miquel C, Gielly L, Brochmann C, Brysting AK, Sønstebø JH, Ims RA, Yoccoz NG, Taberlet P (2009) Analysing diet of small herbivores: the efficiency of DNA barcoding coupled with high-throughput pyrosequencing for deciphering the composition of complex plant mixtures. Front. Zool. 6, 16. https://doi.org/10.1186/1742-9994-6-16

Srivathsan A, Ang A, Vogler AP, Meier R (2016) Fecal metagenomics for the simultaneous assessment of diet, parasites, and population genetics of an understudied primate. Front. Zool. 13, 17. https://doi.org/10.1186/s12983-016-0150-4

Srivathsan A, Sha JCM, Vogler AP, Meier R (2015) Comparing the effectiveness of metagenomics and metabarcoding for diet analysis of a leaf-feeding monkey (Pygathrix nemaeus). Mol. Ecol. Resour. 15, 250–261. https://doi.org/10.1111/1755-0998.12302

Stahlschmidt MC, Collin TC, Fernandes DM, Bar-Oz G, Belfer-Cohen A, Gao Z, Jakeli N, Matskevich Z, Meshveliani T, Pritchard JK, McDermott F, Pinhasi R (2019) Ancient mammalian and plant DNA from Late Quaternary stalagmite layers at Solkota Cave, Georgia. Sci. Rep. 9, 6628. https://doi.org/10.1038/s41598-019-43147-0

Stat M, Huggett MJ, Bernasconi R, DiBattista JD, Berry TE, Newman SJ, Harvey ES, Bunce M (2017) Ecosystem biomonitoring with eDNA: metabarcoding across the tree of life in a tropical marine environment. Sci. Rep. 7, 12240. https://doi.org/10.1038/s41598-017-12501-5

Stein JL, Marsh TL, Wu KY, Shizuya H, DeLong EF (1996) Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. J. Bacteriol. 178, 591–599. https://doi.org/10.1128/jb.178.3.591-599.1996

Teeling H, Glöckner FO (2012) Current opportunities and challenges in microbial metagenome analysis – A bioinformatic perspective. Brief. Bioinformatics 13, 728–742. https://doi.org/10.1093/bib/bbs039

Thomas T, Gilbert J, Meyer F (2012) Metagenomics - a guide from sampling to data analysis. Microb. Inform. Exp. 2, 3. https://doi.org/10.1186/2042-5783-2-3

Thorstenson YR, Hunicke-Smith SP, Oefner PJ, Davis RW (1998) An automated hydrodynamic process for controlled, unbiased DNA shearing. Genome Res. 8, 848–855. https://doi.org/10.1101/gr.8.8.848

Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N (2015) MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat. Methods 12, 902–903. https://doi.org/10.1038/nmeth.3589

Tseng Q, Lomonosov AM, Furlong EEM, Merten CA (2012) Fragmentation of DNA in a sub-microliter microfluidic sonication device. Lab Chip 12, 4677–4682. https://doi.org/10.1039/c2lc40595d

Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI (2007) The human microbiome project. Nature 449, 804–810. https://doi.org/10.1038/nature06244

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428, 37–43. https://doi.org/10.1038/nature02340

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Smith HO (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304, 66–74. https://doi.org/10.1126/science.1093857

Vestergaard G, Schulz S, Schöler A, Schloter M (2017) Making big data smart – How to use metagenomics to understand soil quality. Biol. Fertil. Soils 53, 479–484. https://doi.org/10.1007/s00374-017-1191-3

Weyrich LS, Duchene S, Soubrier J, Arriola L, Llamas B, Breen J, Morris AG, Alt KW, Caramelli D, Dresely V, Farrell M, Farrer AG, Francken M, Gully N, Haak W, Hardy K, Harvati K, Held P, Holmes EC, Kaidonis J, Cooper A (2017) Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus. Nature 544, 357–361. https://doi.org/10.1038/nature21674

Wong KK, Markillie LM, Saffer JD (1997) A novel method for producing partial restriction digestion of DNA fragments by PCR with 5-methyl-CTP. Nucleic Acids Res. 25, 4169–4171. https://doi.org/10.1093/nar/25.20.4169

Wood DE, Salzberg SL (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 15, R46. https://doi.org/10.1186/gb-2014-15-3-r46

Xin T, Su C, Lin Y, Wang S, Xu Z, Song J (2018) Precise species detection of traditional Chinese patent medicine by shotgun metagenomic sequencing. Phytomedicine 47, 40–47. https://doi.org/10.1016/j.phymed.2018.04.048

Ye SH, Siddle KJ, Park DJ, Sabeti PC (2019) Benchmarking metagenomics tools for taxonomic classification. Cell 178, 779–794. https://doi.org/10.1016/j.cell.2019.07.010

Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W, Jaroszewski L, Cieplak P, Miller CS, Li H, Mashiyama ST, Joachimiak MP, van Belle C, Chandonia J-M, Soergel DA, Zhai Y, Venter JC (2007) The Sorcerer II global ocean sampling expedition: expanding the universe of protein families. PLoS Biol. 5, e16. https://doi.org/10.1371/journal.pbio.0050016

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18, 821–829. https://doi.org/10.1101/gr.074492.107

Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for aligning DNA sequences. J. Comput. Biol. 7, 203–214. https://doi.org/10.1089/10665270050081478

Ziesemer KA, Mann AE, Sankaranarayanan K, Schroeder H, Ozga AT, Brandt BW, Zaura E, Waters-Rist A, Hoogland M, Salazar-García DC, Aldenderfer M, Speller C, Hendy J, Weston DA, MacDonald SJ, Thomas GH, Collins MJ, Lewis CM, Hofman C, Warinner C (2015) Intrinsic challenges in ancient microbiome reconstruction using 16S rRNA gene amplification. Sci. Rep. 5, 16498. https://doi.org/10.1038/srep16498

Answers

1. DNA fragmentation and library building. Current sequencing technologies are unable to sequence full genomes of most organisms in a single run, so fragmentation is required for downstream procedures. Library preparation prepares the DNA fragments for sequencing, and the addition of adapters allows DNA fragments to be identified.
2. It is bioinformatically more challenging for the assembly process due to the formation of chimeras. Utilising bioinformatics procedures such as assembly after binning and taxonomic assignment, or using long-read sequencing, can overcome these challenges.
3. Samples with unequal abundance can complicate assembly as reads from different taxa do not overlap, reducing the probability of accurate taxonomic or functional assignment.

Chapter 13
DNA Barcoding - High Resolution Melting analysis (Bar-HRM)

Bastien Anthoons[1,2], Andreas D. Drouzas[1], Panagiotis Madesis[3]

1 School of Biology, Aristotle University of Thessaloniki, Thessaloniki, Greece
2 Institute for Applied Biosciences, Centre for Research and Technology, Thessaloniki, Greece
3 Lab of Molecular Biology, Department of Agriculture Crop Production and Rural Environment, University of Thessaly, Volos, Greece

Bastien Anthoons **bastien.anthoons@gmail.com**
Andreas D. Drouzas **drouzas@bio.auth.gr**
Panagiotis Madesis **pmadesis@uth.gr**

Citation: Anthoons B, Drouzas AD, Madesis P (2022) Chapter 13 DNA Barcoding - High Resolution Melting analysis (Bar-HRM). In: de Boer H, Rydmark MO, Verstraete B, Gravendeel B (Eds) Molecular identification of plants: from sequence to species. Advanced Books. https://doi.org/10.3897/ab.e98875

Introduction

Accurate species identification is fundamental for correct assessment of species diversity and for studying the functioning of their communities and ecosystems. Additional applications include the use of species identification in food product authentication and for diagnosing diseases. Species identification can be carried out using morphological, (bio)chemical, or molecular traits. Pertaining to the molecular-based approaches, both PCR and post-PCR analyses have been extensively used for species identification and genotyping. The most widely used PCR-based method is DNA barcoding, which is able to provide species-level identifications using the sequences of short standard DNA regions (Hebert et al. 2003; Raclariu et al. 2018). Similarly, High Resolution Melting (HRM) is a post-PCR technique that enables the detection of genetic variation amongst nucleic acid sequences by measuring the rate at which double-stranded DNA (dsDNA) dissociates into single-stranded DNA (ssDNA) with increasing temperature (Reed and Wittwer 2004). The result of an HRM analysis is a melting curve profile whose shape is specific for a particular organism. The main advantages of HRM are its high sensitivity and accuracy that allows for closed tube reactions, therefore minimising contamination. In addition, carrying out HRM only requires a real-time PCR instrument with appropriate software.

HRM data analysis is straightforward and it does not require advanced bioinformatics skills, in contrast to other genetic analyses used for species identification. Furthermore, HRM is a cost-effective and high-throughput methodology. Due to its ability to discriminate between samples at the resolution of a single nucleotide (allowing for single nucleotide polymorphism (SNP) identification), HRM is commonly used for genotyping, mutation scanning, and DNA methylation analyses. The HRM analysis of DNA barcoding regions, e.g., ITS2, *matK*, *trnL* (see Chapter 10 DNA barcoding), is called Bar-HRM: Barcoding - High Resolution Melting analysis. It has been successfully introduced by Jaakola et al. (2010) for the authentication of berry species and has been largely used in a variety of applications since then (Jaakola et al. 2010; Ganopoulos et al. 2012b, 2012a).

High resolution melting analysis

How does High Resolution Melting work?

PCR amplification of the genetic region of interest is a prerequisite for HRM analysis and is done in the presence of a fluorescent dye that binds dsDNA. Such dyes intercalate into the dsDNA that is produced during a PCR reaction, without affecting PCR efficiency. Asymmetric cyanine dyes fluoresce strongly in the presence of dsDNA and are characterised by low intensity fluorescence in the unbound state (Reed and Wittwer 2004). DNA amplification is followed by a high resolution melting step. The PCR product is gradually heated, which causes the DNA amplicon to denature (dsDNA dissociation), thus releasing the intercalating dye and consequently decreasing fluorescence intensity. This absolute change in fluorescence intensity is measured as a function of temperature at high sensitivity, resolution, and precision. The result is a melt curve profile characteristic of the amplicon (Reed and Wittwer 2004). This specific curve allows rapid discrimination amongst sequences, even if they differ by only one nucleotide (Reed and Wittwer 2004).

Analysis of dsDNA dissociation during HRM

The rate of dsDNA PCR product dissociation, and thus the shape of the HRM curve, depends on (1) the sequence itself and its length, (2) the GC content, (3) the complementarity, and (4) the nearest-neighbour thermodynamics of the amplicon (Reed and Wittwer 2004). The HRM curve is derived from high density of fluorescence data points logged during the analysis and is therefore a highly accurate and specific curve for a particular PCR product. The steps towards creating an HRM curve profile are (1) melting curve normalisation and (2) calculation of derivative curves. Normalisation involves a software-driven numerical recalculation of the individual fluorescence data points. To visualise the melting temperature (T_m) more clearly, derivative curves are often plotted, making the T_ms of the products to appear as peaks (Figure 1B). The T_m of a PCR product is defined as the temperature at which 50% of DNA is dissociated. The T_m is a function of a PCR product's physical characteristics, including GC-content (the T_m is higher in GC-rich PCR products), length, and sequence content. The T_m is highly specific and can be used to most accurately differentiate PCR products (Reed et al. 2007). The derivative curves are created by calculating the first negative derivation of fluorescence with temperature ($-\Delta F/\Delta t$). The distinct peaks of a derivative curve are a characteristic of the melting profile. Both the melting and derivative curves are characterised by three phases (Figure 1). The first phase or pre-melting phase is characterised by a linear, flat appearance. This is because all the PCR products are still double-stranded and all of the dye is bound. Thus, there is no change in relative fluorescence. As the temperature increases, the dsDNA starts to melt and releases the fluorescent dye, resulting in a decrease in fluorescence signal in the raw melting curve (Figure 1A). This is observed as a sharp increase in the derivative curve (Figure 1B). The second phase of decreasing

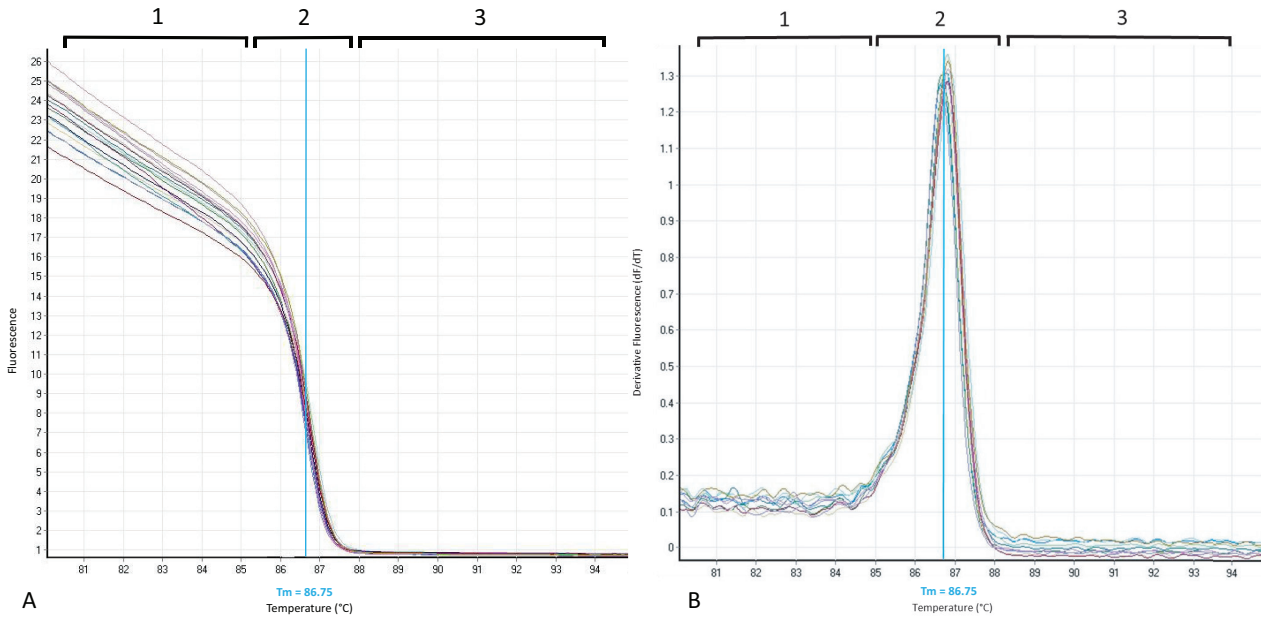

Figure 1. An example of standard output curves of an HRM analysis. In both A and B, the three phases of the DNA melting profile are shown. (1) The pre-melting phase is characterised by an initial fluorescence given in relative fluorescence units (RFUs). Here all PCR products are double-stranded and the maximum amount of dye is bound. (2) In the active melting phase the inflection point (i.e., T_m) is where 50% of the PCR product the samples is denatured. (3) The post-melting phase. As the temperature increases, the PCR products denature, dye is released, and the fluorescent signal drops and plateaus **A**. A normalised melting curve with indication of the inflection point. **B**. A derivative curve, which shows the inflection point on the slope as a melting peak.

fluorescence is called the active melting phase. When the curve reaches the inflection point and the fluorescence continues to decrease, the negative derivative curve reaches a peak value and drops again. Both the inflection point in the sigmoid normalised curve, and the peak in the derivative curve, represent the T_m of the dsDNA molecule. The third or post-melting phase is characterised by a drop in fluorescence as the PCR products denature with increasing temperature. The raw and normalised melting curves show typically sigmoid patterns (Figure 1A). The HRM software automatically clusters samples into groups with similar melting curves. Differences in melting curve shapes are subsequently analysed in detail by subtracting all curves from a reference curve. Subsequently, this clustering of samples is evaluated by comparing it to the reference curve, either visually or by confidence values calculated by the software (Ganopoulos et al. 2011; Wittwer et al. 2003). The genotype confidence percentage (GCP), which is automatically calculated by the HRM software, represents the confidence value of the similarity of a sample to the reference genotype (with a value of 100 indicating an exact match) (Hewson et al. 2009; Sun et al. 2016). Any genotype can be selected as the reference, but typically a wild-type control is used. Specific deviations from the reference are subsequently identified and assigned to alternative alleles/genotypes/samples.

HRM assay and reagent optimization

Identification of species by Bar-HRM relies on small genetic differences between DNA sequences, which will result in different melting curves. However, small differences between melting curve profiles may also arise from sources other than the DNA sequences, thus assay optimisation is a prerequisite for a successful Bar-HRM analysis. An example of a Bar-HRM workflow can be found in the infographic. Factors that could influence the outcome of an HRM analysis include genomic DNA quality, DNA extraction impurities, amplicon length, primer design, dye selection, PCR reagent choice (see Chapter 1 DNA from plant tissue), and the choice of Bar-HRM instruments and software (Montgomery et al. 2007; van der Stoep et al. 2009).

DNA quality

The major factor associated with DNA quality is salt carryover, as this will change the thermodynamics of the DNA melting process. This could lead to lower reproducibility and higher error rates in the Bar-HRM results. A solution is to precipitate and resuspend the DNA extract in a buffer with a low salt concentration such as TE (10 mM Tris, 1 mM EDTA) prior to the PCR (Rogers and Bendich 1989). TE buffer will also neutralise pH extremes, which could also affect the shape of the HRM curve. Low-quality DNA, i.e., degraded or contaminated DNA, may produce non-specific PCR products and therefore may affect the outcome as well. Late amplification of such samples (C_t > 30) or poor quality may result in a PCR reaction that fails to reach a plateau phase and will therefore reduce HRM data resolution. It is therefore recommended to perform the Bar-HRM analysis immediately after the PCR reaction. If this is not possible, PCR products should be stored at -20 °C (Reed et al. 2007).

Amplicon length

Amplicons up to 300 bp are generally preferable for Bar-HRM analysis since they are more suitable for the detection of DNA mutations such as SNPs, inversions, insertions, and deletions. The

larger the fragments, the more likely it is to detect additional mutation sites that may complicate the discrimination between/among different sequence variants. On the other hand, amplicons that are too small (< 50 bp) may produce lower fluorescence signals, due to lower amounts of dye being incorporated into the PCR product (Vossen et al. 2009; Sun et al. 2016).

Primer design

The intercalating dyes used in the Bar-HRM analysis bind generically to any double-stranded DNA product. It is therefore important to design robust PCR primers that are specific to the region of interest, to ensure that this is the only region amplified in the PCR product. Each developed primer pair should be tested for specificity to the region of interest and should not produce any primer-dimers or non-specific products. PCR products from the developed primer pair should be assessed by gel electrophoresis, as the HRM software may not be able to detect all non-specific reaction products if their melting curves are similar (Reed et al. 2007).

Dye selection

HRM uses dsDNA-binding fluorescent dyes that do not interfere with the PCR reaction. The so-called "release-on-demand" dyes are preferred for HRM as they do not inhibit DNA polymerases or alter the T_m of the product (Sun et al. 2016). The amount of dye can also affect the HRM analysis: too little dye can result in low signal and inaccurate results, whereas too much dye can stabilise the double-stranded form and artificially shift the T_m, or inhibit or reduce the efficiency of the PCR reaction. The most commonly used dyes for HRM analysis are SYTO® 9 Green Fluorescent Nucleic Acid Stain, EvaGreen® Dye, LCGreen, and SYBR® Green I (Sun et al. 2016).

PCR reagents

Reagents for HRM analysis and reaction conditions should be optimised to reduce amplification biases as much as possible. Primer dimers and other non-specific products can significantly decrease the performance of the HRM analysis. So, in addition to optimising reactions, one must ensure that variation is not introduced by poor assay design or optimization decisions (see Chapter 1 DNA from plant tissue; Reed et al. 2007). The $MgCl_2$ concentration in particular should also be carefully checked as too low a concentration affects the PCR specificity as well as the amplicon melting properties, creating non-specific products in the PCR reaction. It is recommended to perform a $MgCl_2$ titration to find the best salt concentration for each reaction.

HRM instruments and software

HRM analysis requires a PCR thermal cycler and an instrument with optics capable of detecting fluorescence. This can either be a rotary design in which samples spin past an optical detector or a block-based instrument in which samples are read by a scanning head or stationary camera. This instrument should be coupled with a computer with appropriate HRM analysis software capable of handling the large amounts of data generated during the analysis. A good HRM software package should provide a view of the raw fluorescence data points and a process to both align the data and view melting curve differences between samples (Wittwer 2009).

Advantages and limitations of HRM

The chemical improvement of "release-on-demand" DNA dyes and the increased instrumentation precision has widely expanded the use of Bar-HRM for genotyping (Ganopoulos et al. 2011). Moreover, HRM analysis is more cost effective than other genotyping technologies, (such as sequencing) and it does not require any prior knowledge of the sequence of the organism of study. Although HRM is a very sensitive technique the risk of contamination is reduced compared to other multi-step procedures, because the entire process can be rapidly completed within a single closed tube. Therefore, large numbers of samples can be simultaneously screened for genotypic differences, making HRM a low-cost-per-sample method. In addition, the use of small HRM amplicons allows the analysis of samples containing degraded DNA, such as processed material (e.g., food). Another advantage of the Bar-HRM methodology involves the reduction of researcher biases, given the analysis is mostly performed by the statistical software (Madesis et al. 2014, 2013). HRM can be performed in any type of laboratory as it only requires a thermocycler and a computer with the corresponding HRM software, without the need for expensive or complicated lab equipment. The enormous decrease of post-handling time combined with the exclusion of handling hazardous chemicals such as ethidium bromide, makes this method an excellent alternative for molecular diagnostic approaches (Madesis et al. 2013). The HRM methodology allows for the detection of a single species at a time per product/sample. The species of interest can however be detected in both single-ingredient products and multi-species samples with Bar-HRM (Anthoons et al. 2021). On the other hand, a prerequisite of Bar-HRM is the need to create a melting curve database to serve as a reference when the identification of unknown samples is intended. It is thus of great importance that the different steps of the methods have been standardised to allow the curves produced by the Bar-HRM analysis to be used across different laboratories. In addition, as mentioned earlier, the HRM results can be affected by factors other than the DNA sequence (such as the genomic DNA quality, DNA extraction impurities, amplicon length, primer design, dye selection, and PCR reagent choice).

HRM applications

Taxonomic identification of plant taxa

Since the first description of the HRM methodology in 2003, it has been increasingly used as a research tool (Gundry et al. 2003; Ruskova and Raclavsky 2011). More specifically, Bar-HRM has proven useful for the identification and authentication of plants belonging to both closely related and remote taxa (Osathanunkul et al. 2015). Madesis et al. (2012) used the plant barcoding region *trnL* in combination with HRM analysis to identify legume crops (*Lupinus* spp.) at the genus level. HRM has also been used in combination with the ITS2 region for the identification of *Sideritis* species (Kalivas et al. 2014), *Artemisia* species (Song et al. 2016), as well as for the differentiation between different edible and poisonous ginseng species (Osathanunkul and Madesis 2019). Furthermore, the ITS1 barcode in combination with HRM proved sensitive enough for the discrimination of 12 closely related *Croton* species and for the identification of medicinal plant species of the genus *Kaempferia* (Osathanunkul et al. 2015), (Osathanunkul et al. 2017). Additionally, the detection of simple sequence repeats (SSRs) in combination with HRM has been used to identify landraces and cultivars from *Solanum melongena* (Ganopoulos et al. 2015), as well as *Olea europaea* cultivars (Xanthopoulou et al. 2014; Zhang et al. 2012).

Quality control of food products

Medicinal plants and plant-based food products are often processed and lack the essential parts necessary for morphological identification when sold on the herbal market. In addition, the herbal market is highly competitive and lacks standardised methods for quality assessment. This has contributed to increasing problems with product adulteration and substitution. Numerous studies reported the substitution of costly ingredients in herbal products with plant material of inferior quality or unlabelled plant fillers (Anthoons et al. 2021; Raclariu et al. 2017; Seethapathy et al. 2019; Zhang et al. 2012). HRM can assist with food quality control and the detection of allergens and adulterants in food products. It is noteworthy that extracting and amplifying DNA from processed foods can be challenging. For instance, medicinal plants in herbal products contain secondary metabolites that could inhibit amplification. Additionally, DNA from processed products is often degraded and the plant drying that occurs during processing can complicate DNA extraction and isolation procedures (see Chapter 6 DNA from food and medicine). However, if particular adaptations regarding DNA isolation and amplification are taken into account, HRM is still capable of detecting minute interspecific differences even in processed products.

Bar-HRM has been used for the identification and quantification of the ingredients in plant and animal food products, including Protected Designation of Origin (PDO) products. Olive oil for instance, which is one of the most adulterated vegetable oils on the market, has been successfully authenticated with Bar-HRM (Ganopoulos et al. 2013a). Examples of Bar-HRM studies on PDOs include Fava Santorinis (*Lathyrus clymenum*) adulterants (Ganopoulos et al. 2012a), authentication of Greek PDO Feta cheese by detecting bovine, ovine, and caprine species (Ganopoulos et al. 2013b), and the authentication of Greek saffron PDO (Bosmali et al. 2017). Bar-HRM has also been used as a verification tool for metabarcoding results. Anthoons et al. (2020) detected the presence of two poisonous plant species (*Chelidonium majus* and *Nicotiana tabacum)* and wheat in herbal medicinal products with amplicon metabarcoding. Their presence was further confirmed by Bar-HRM using the ITS2 region.

Quantitative detection in food

Apart from species identification, Bar-HRM can also be used for species quantification, which is also important for quality control, especially for quantifying adulterants in food or other processed products. Serial dilutions of a DNA sample mixed with adulterant DNA are made, corresponding to different known adulterant content percentages. These artificial serial admixtures are then used to create reference curves that can be used to quantify samples of unknown content (Lagiotis et al. 2020; Anthoons et al. 2022). Traces of hazelnuts and almonds, which are common nut allergens, can be quantified with Bar-HRM in processed food products (e.g., *Corylus avellana* in biscuits; Madesis et al. 2013), and amounts of turmeric as low as 0.5% w/w can be detected in saffron PDO products (Bosmali et al. 2017).

Future prospects of HRM

Bar-HRM technology can provide taxonomic identification of plant taxa, the tracking of a wide range of raw and processed herbal products, and the detection of adulterants and poisonous contaminants in food products. As the precision of the "release-on-demand" dyes and HRM instruments further increase, in addition to the development of melting curve reference databases, we can expect that Bar-HRM will be implemented as a routine analytical tool for species

identification and authentication. Finally, the successful application of Bar-HRM as a tool for quality control in the food industry, renders it suitable to be also used in a regulatory framework by the corresponding authorities.

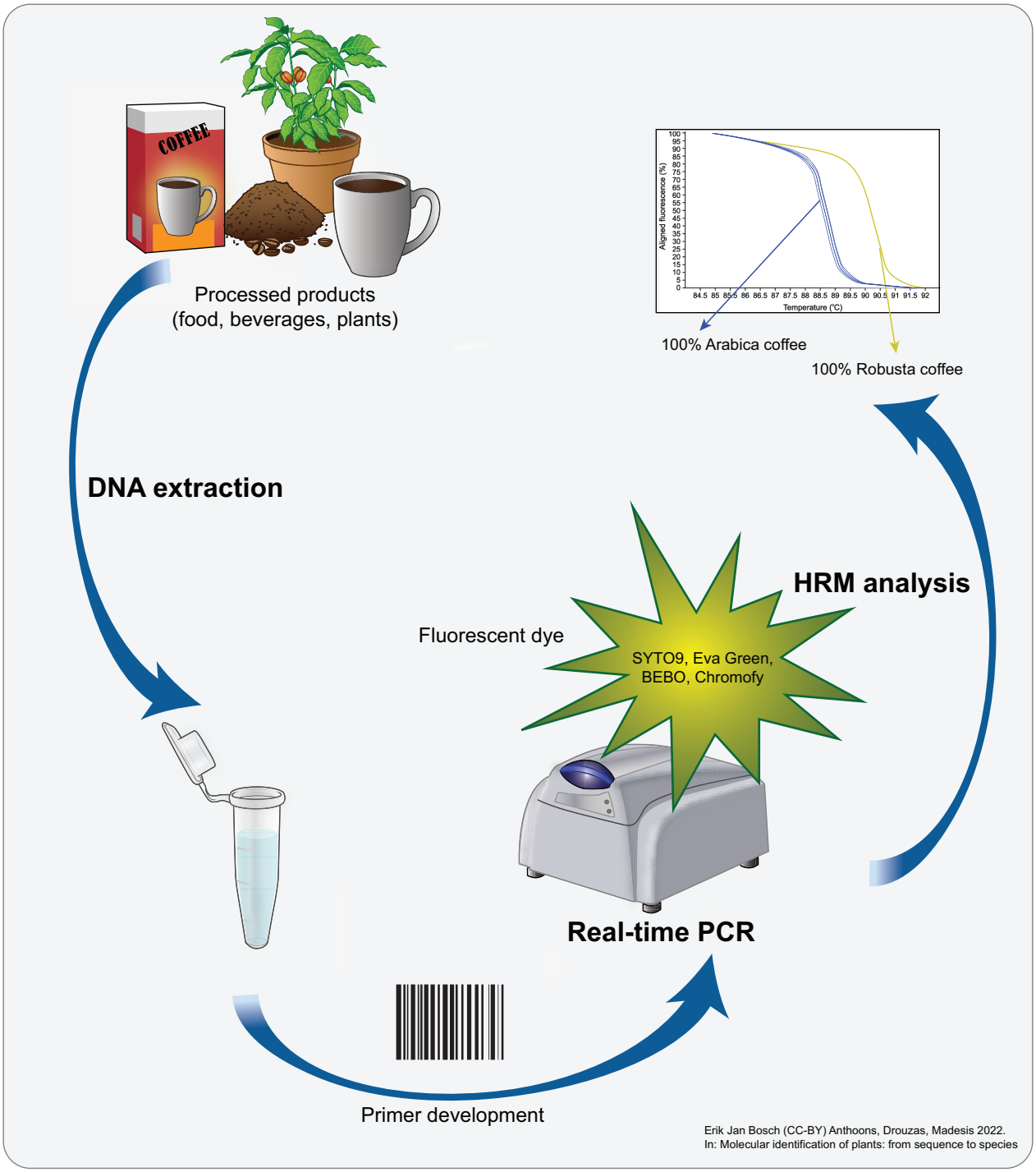

Figure 2. Chapter 13 Infographic: Illustration of a typical HRM methodology workflow. The initial substrate can be a single or multi-ingredient product from raw or processed plant material. Following DNA extraction, taxon-specific primers are developed based on DNA barcodes or other molecular markers and tested in silico. An intercalating fluorescent dye is added to the PCR reaction, which allows the detection of the PCR amplicons by the HRM equipment during the melting process. The output consists of melting graphs and a statistical report including Genotype Confidence Percentages (GCPs) allowing accurate discrimination between the reference and the analysed taxa.

Questions

1. A. Which amplicon length is generally recommended for Bar-HRM analysis? Explain your answer. B. What is the risk for Bar-HRM reactions if the amplicon length is too short?
2. A. In case the HRM analysis detects non-specific products, what could the underlying reasons be? B. How would this issue be verified and resolved?
3. A. Describe the different melting phases of the dsDNA sequence depicted in Figure 1. B. What is the meaning of the melting temperature (T_m) and how can it be used for plant identification?

Glossary

Bar-HRM – DNA Barcoding - High Resolution Melting analysis is an HRM analysis coupled with a barcoding region of interest (such as *trnL-F,* ITS, or *matK* for plants) that is primarily used for the identification of organisms at various taxonomic levels.

GCP – Genotype Confidence Percentage is a parameter calculated by the HRM software and represents the confidence that a sample is the same as the reference genotype, with a value of 100 indicating an exact match.

HRM analysis – High Resolution Melting analysis is a post-PCR analysis that is used to identify variations in nucleic acid sequences. The method is based on detecting small differences in PCR melting (dissociation) curves.

Nearest-neighbour method – The method is based on a model in which the thermodynamic stability of a base pair in a DNA strand is dependent on the identity of the adjacent base pairs. These thermodynamic properties can be used for predicting the melting temperature of the DNA strand.

PDO – Protected Designation of Origin is a registered designation of products that have the strongest links to their area of production and that are protected by intellectual property rights.

qPCR – Quantitative polymerase chain reaction used for quantifying DNA in real-time.

RFU – Relative Fluorescence Unit is a unit of measurement used in real-time PCR analysis, which employs fluorescence detection. The computer software measures the results, determining the quantity or size of the fragments, at each data point, from the level of fluorescence intensity. Samples which contain higher quantities of amplified DNA will have higher corresponding RFU values.

SSR – Simple-sequence repeats (SSR), also known as microsatellites, are short tandem repeated nucleotide motifs that may vary in the number of repeats at a given locus.

References

Anthoons B, Karamichali I, Schrøder-Nielsen A, Drouzas AD, de Boer H, Madesis P (2021) Metabarcoding reveals low fidelity and presence of toxic species in short chain-of-commercialization of herbal products. Journal of Food Composition and Analysis 97, 103767. https://doi.org/10.1016/j.jfca.2020.103767

Anthoons B, Lagiotis G, Drouzas AD, de Boer H, Madesis P (2022) Barcoding High Resolution Melting (Bar-HRM) enables the discrimination between toxic plants and edible vegetables prior to consumption and after digestion. J. Food Sci. 87, 4221–4232. https://doi.org/10.1111/1750-3841.16253

Bosmali I, Ordoudi SA, Tsimidou MZ, Madesis P (2017) Greek PDO saffron authentication studies using species specific molecular markers. Food Res. Int. 100, 899–907. https://doi.org/10.1016/j.foodres.2017.08.001

Ganopoulos I, Argiriou A, Tsaftaris A (2011) Adulterations in Basmati rice detected quantitatively by combined use of microsatellite and fragrance typing with High Resolution Melting (HRM) analysis. Food Chem. 129, 652–659. https://doi.org/10.1016/j.foodchem.2011.04.109

Ganopoulos I, Bazakos C, Madesis P, Kalaitzis P, Tsaftaris A (2013a) Barcode DNA high-resolution melting (Bar-HRM) analysis as a novel close-tubed and accurate tool for olive oil forensic use. J. Sci. Food Agric. 93, 2281–2286. https://doi.org/10.1002/jsfa.6040

Ganopoulos I, Madesis P, Darzentas N, Argiriou A, Tsaftaris A (2012a) Barcode High Resolution Melting (Bar-HRM) analysis for detection and quantification of PDO "Fava Santorinis" (*Lathyrus clymenum*) adulterants. Food Chem. 133, 505–512. https://doi.org/10.1016/j.foodchem.2012.01.015

Ganopoulos I, Madesis P, Tsaftaris A (2012b) Universal ITS2 barcoding DNA region coupled with High-Resolution Melting (HRM) analysis for seed authentication and adulteration testing in leguminous forage and pasture species. Plant Mol. Biol. Rep. 30, 1322–1328. https://doi.org/10.1007/s11105-012-0453-3

Ganopoulos I, Sakaridis I, Argiriou A, Madesis P, Tsaftaris A (2013b) A novel closed-tube method based on high resolution melting (HRM) analysis for authenticity testing and quantitative detection in Greek PDO feta cheese. Food Chem. 141, 835–840. https://doi.org/10.1016/j.foodchem.2013.02.130

Ganopoulos I, Xanthopoulou A, Mastrogianni A, Drouzas A, Kalivas A, Bletsos F, Krommydas SK, Ralli P, Tsaftaris A, Madesis P (2015) High Resolution Melting (HRM) analysis in eggplant (*Solanum melongena* L.): A tool for microsatellite genotyping and molecular characterization of a Greek Genebank collection. Biochem. Syst. Ecol. 58, 64–71. https://doi.org/10.1016/j.bse.2014.11.003

Gundry CN, Vandersteen JG, Reed GH, Pryor RJ, Chen J, Wittwer CT (2003) Amplicon melting analysis with labeled primers: a closed-tube method for differentiating homozygotes and heterozygotes. Clin. Chem. 49, 396–406. https://doi.org/10.1373/49.3.396

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. Proc. Biol. Sci. 270, 313–321. https://doi.org/10.1098/rspb.2002.2218

Hewson K, Noormohammadi AH, Devlin JM, Mardani K, Ignjatovic J (2009) Rapid detection and non-subjective characterisation of infectious bronchitis virus isolates using high-resolution melt curve analysis and a mathematical model. Arch. Virol. 154, 649–660. https://doi.org/10.1007/s00705-009-0357-1

Jaakola L, Suokas M, Häggman H (2010) Novel approaches based on DNA barcoding and high-resolution melting of amplicons for authenticity analyses of berry species. Food Chem. 123, 494–500. https://doi.org/10.1016/j.foodchem.2010.04.069

Kalivas A, Ganopoulos I, Xanthopoulou A, Chatzopoulou P, Tsaftaris A, Madesis P (2014) DNA barcode ITS2 coupled with high resolution melting (HRM) analysis for taxonomic identification of Sideritis species growing in Greece. Mol. Biol. Rep. 41, 5147–5155. https://doi.org/10.1007/s11033-014-3381-5

Madesis P, Ganopoulos I, Anagnostis A, Tsaftaris A (2012) The application of Bar-HRM (Barcode DNA-High Resolution Melting) analysis for authenticity testing and quantitative detection of bean crops (Leguminosae) without prior DNA purification. Food Control 25, 576–582. https://doi.org/10.1016/j.foodcont.2011.11.034

Madesis P, Ganopoulos I, Bosmali I, Tsaftaris A (2013) Barcode High Resolution Melting analysis for forensic uses in nuts: a case study on allergenic hazelnuts (*Corylus avellana*). Food Res. Int 50, 351–360. https://doi.org/10.1016/j.foodres.2012.10.038

Madesis P, Ganopoulos I, Sakaridis I, Argiriou A, Tsaftaris A (2014) Advances of DNA-based methods for tracing the botanical origin of food products. Food Res. Int 60, 163–172. https://doi.org/10.1016/j.foodres.2013.10.042

Montgomery J, Wittwer CT, Palais R, Zhou L (2007) Simultaneous mutation scanning and genotyping by high-resolution DNA melting analysis. Nat. Protoc. 2, 59–66. https://doi.org/10.1038/nprot.2007.10

Osathanunkul M, Madesis P (2019) Bar-HRM: a reliable and fast method for species identification of ginseng (*Panax ginseng*, *Panax notoginseng*, *Talinum paniculatum* and *Phytolacca americana*). PeerJ 7, e7660. https://doi.org/10.7717/peerj.7660

Osathanunkul M, Ounjai S, Osathanunkul R, Madesis P (2017) Evaluation of a DNA-based method for spice/herb authentication, so you do not have to worry about what is in your curry, buon appetito! PLoS ONE 12, e0186283. https://doi.org/10.1371/journal.pone.0186283

Osathanunkul M, Suwannapoom C, Ounjai S, Rora JA, Madesis P, de Boer H (2015) Refining DNA barcoding coupled high resolution melting for discrimination of 12 closely related *Croton* species. PLoS ONE 10, e0138888. https://doi.org/10.1371/journal.pone.0138888

Raclariu AC, Heinrich M, Ichim MC, de Boer H (2018) Benefits and limitations of DNA barcoding and metabarcoding in herbal product authentication. Phytochem. Anal. 29, 123–128. https://doi.org/10.1002/pca.2732

Raclariu AC, Mocan A, Popa MO, Vlase L, Ichim MC, Crisan G, Brysting AK, de Boer H (2017) *Veronica officinalis* product authentication using DNA metabarcoding and HPLC-MS reveals widespread adulteration with *Veronica chamaedrys*. Front. Pharmacol. 8, 378. https://doi.org/10.3389/fphar.2017.00378

Reed GH, Kent JO, Wittwer CT (2007) High-resolution DNA melting analysis for simple and efficient molecular diagnostics. Pharmacogenomics 8, 597–608. https://doi.org/10.2217/14622416.8.6.597

Reed GH, Wittwer CT (2004) Sensitivity and specificity of single-nucleotide polymorphism scanning by high-resolution melting analysis. Clin. Chem. 50, 1748–1754. https://doi.org/10.1373/clinchem.2003.029751

Rogers SO, Bendich AJ (1989) Extraction of DNA from plant tissues, in: Gelvin, S.B., Schilperoort, R.A., Verma, D.P.S. (Eds.), Plant Molecular Biology Manual. Springer Netherlands, Dordrecht, pp. 73–83. https://doi.org/10.1007/978-94-009-0951-9_6

Ruskova L, Raclavsky V (2011) The potential of high resolution melting analysis (hrma) to streamline, facilitate and enrich routine diagnostics in medical microbiology. Biomed Pap Med Fac Univ Palacky Olomouc Czech Repub 155, 239–252. https://doi.org/10.5507/bp.2011.045

Seethapathy GS, Raclariu-Manolica A-C, Anmarkrud JA, Wangensteen H, de Boer HJ (2019) DNA metabarcoding authentication of ayurvedic herbal products on the European market raises concerns of quality and fidelity. Front. Plant Sci. 10, 68. https://doi.org/10.3389/fpls.2019.00068

Song M, Li J, Xiong C, Liu H, Liang J (2016) Applying high-resolution melting (HRM) technology to identify five commonly used *Artemisia* species. Sci. Rep. 6, 34133. https://doi.org/10.1038/srep34133

Sun W, Li J-J, Xiong C, Zhao B, Chen S-L (2016) The potential power of Bar-HRM technology in herbal medicine identification. Front. Plant Sci. 7, 367. https://doi.org/10.3389/fpls.2016.00367

van der Stoep N, van Paridon CDM, Janssens T, Krenkova P, Stambergova A, Macek M, Matthijs G, Bakker E (2009) Diagnostic guidelines for high-resolution melting curve (HRM) analysis: an interlaboratory validation of BRCA1 mutation scanning using the 96-well LightScanner. Hum. Mutat. 30, 899–909. https://doi.org/10.1002/humu.21004

Vossen RHAM, Aten E, Roos A, den Dunnen JT (2009) High-resolution melting analysis (HRMA): more than just sequence variant screening. Hum. Mutat. 30, 860–866. https://doi.org/10.1002/humu.21019

Wittwer CT, Reed GH, Gundry CN, Vandersteen JG, Pryor RJ (2003) High-resolution genotyping by amplicon melting analysis using LCGreen. Clin. Chem. 49, 853–860. https://doi.org/10.1373/49.6.853

Wittwer CT (2009) High-resolution DNA melting analysis: advancements and limitations. Hum. Mutat. 30, 857–859. https://doi.org/10.1002/humu.20951

Xanthopoulou A, Ganopoulos I, Koubouris G, Tsaftaris A, Sergendani C, Kalivas A, Madesis P (2014) Microsatellite high-resolution melting (SSR-HRM) analysis for genotyping and molecular characterization of an *Olea europaea* germplasm collection. Plant Genet. Res. 12, 273–277. https://doi.org/10.1017/S147926211400001X

Zhang J, Wider B, Shang H, Li X, Ernst E (2012) Quality of herbal medicines: challenges and solutions. Complement. Ther. Med. 20, 100–106. https://doi.org/10.1016/j.ctim.2011.09.004

Answers

1. A. The suitable amplicon length for Bar-HRM analysis varies from 50 to 300 bp. The shorter the amplicon length, the more accurate the result.

2. B. Amplicons that are too short (< 50 bp) produce too little fluorescence signal, due to limited dye incorporation in a shorter sequence.
3. A. The underlying reasons for detection of non-specific products by HRM could be (i) low quality DNA, (ii) an increased salt concentration ($MgCl_2$) in the PCR reaction, (iii) insufficient primer specificity, or (iv) possible contaminations.
4. B. This could be verified by (i) checking the DNA quality (in some cases further purifying the sample with a DNA kit or performing DNA extraction again is recommended), (ii) adjusting the $MgCl_2$ amount by performing titration or by using a master mix with standard (known) $MgCl_2$ concentration (diluting the DNA sample in TE buffer can also be attempted), and (iii) ensuring primer specificity with a BLAST search, by running the PCR product on an electrophoresis gel prior to Bar-HRM to check for a single band pattern, or sequencing the amplicon, and (iv) replacement of the working materials
5. A. The pre-melting phase is the stage of initial fluorescence when all products are double-stranded and the maximum amount of dye is bound. The active melting phase includes the inflection point where 50% of the PCR products in the samples are denatured and the post melting phase is characterised by a drop in fluorescence signal (when the PCR products denature) as the temperature increases. In Figure 1A, the pre-melt phase can be clearly distinguished from the active melting phase by the inflection point in the graph, the sudden drop in fluorescence signal with increasing temperature. In Figure 1B, the melting phase is reflected by the slopes of the melting curve with the peak being the melting temperature.
6. B. The melting temperature (T_m) is the temperature at which 50% of the dsDNA has been denatured. It is unique for each sample and therefore of use for species discrimination.

Chapter 14
Target capture

Yannick Woudstra[1,2,3,4], Anne-Sophie Quatela[3,5], Catherine Kidner[6,7], Juan Viruel[1], Alexandre Zuntini[1], Michael D. Martin[8], Thibauld Michel[6,7], Olwen M. Grace[1]

1 Royal Botanic Gardens, Kew, United Kingdom
2 Natural History Museum Denmark, University of Copenhagen, Copenhagen, Denmark
3 Gothenburg Global Biodiversity Center, Department of Biological and Environmental Sciences, University of Gothenburg, Gothenburg, Sweden
4 Department of Plant Sciences, University of Oxford, Oxford, United Kingdom
5 Department of Biological and Environmental Sciences, University of Gothenburg, Gothenburg, Sweden
6 Institute of Molecular Plant Sciences, University of Edinburgh, Edinburgh, United Kingdom
7 Royal Botanic Garden Edinburgh, Scotland, United Kingdom
8 Department of Natural History, Norwegian University of Science and Technology, Trondheim, Norway

Yannick Woudstra **yannickwoudstra@outlook.com**
Anne-Sophie Quatela **anne-sophie.quatela@bioenv.gu.se**
Catherine Kidner **ckidner@rbge.org.uk**
Juan Viruel **j.viruel@kew.org**
Alexandre Zuntini **a.zuntini@kew.org**
Michael D. Martin **mike.martin@ntnu.no**
Thibauld Michel **tmichel@rbge.org.uk**
Olwen M. Grace **o.grace@kew.org**

Citation: Woudstra Y, Quatela A-S, Kidner C, Viruel J, Zuntini A, Martin MD, Michel T, Grace OM (2022) Chapter 14. Target capture. In: de Boer H, Rydmark MO, Verstraete B, Gravendeel B (Eds) Molecular identification of plants: from sequence to species. Advanced Books. https://doi.org/10.3897/ab.e98875

Introduction

Efforts to resolve the plant tree of life have led to the replacement of traditional DNA sequencing markers (Hollingsworth et al. 2011) with vastly greater volumes of high-throughput sequencing (HTS) data (Li et al. 2015). Whole chloroplast genomes have revolutionised DNA barcoding (see Chapter 10 DNA barcoding) but can lack sufficient variability to resolve phylogenetic relationships (Li et al. 2015). Nuclear genes are now favoured in phylogenomics, at least in most clades (Dodsworth et al. 2019), due to their higher evolutionary rates compared to plastid genes (Hollingsworth 2011; Sang 2002), and their independent evolution due to their unlinked nature with higher recombination rates (Shneyer and Rodionov 2019). Unfortunately, many nuclear genes have multiple copies across the genome (paralogs) with slightly different nucleotide compositions. This makes it difficult to use nuclear genes in phylogenetic studies as the copies need to be sorted for each sample. Single-to-low-copy nuclear (SLCN) genes are therefore preferred. Combining sequences from multiple SLCN genes can disentangle complex evolutionary histories, including potential hybridisation events (Sang 2002). Even though ribosomal nuclear DNA might be a good candidate in some cases, SLCN genes are also the best markers that can be used for the identification of genome merging events (i.e., hybridization) in polyploid taxa (Brochmann et al. 1996; Popp et al. 2005; Rauscher et al. 2002; Wendel et al. 1995) since multiple genome copies (haplotypes) further increase the complexity of studying nuclear genes.

Unique challenges exist when trying to obtain sequences from plant nuclear genomes. Plant genomes are often large (Pellicer et al. 2018), e.g., above a mean of 5.13 pg (± 5.02 Gbp) for angiosperms (Pellicer et al. 2018), and are characterised by whole genome duplication events (106 events reported in angiosperms alone (Landis et al. 2018). In addition, plant genomes are often characterised by an abundant number of repetitive sequences and transposons that can constitute over 80% of the genome (Novák et al. 2020). Identifying SLCN genes in these incredibly complex genomes is thus comparable to finding a needle in a haystack.

Thankfully, many nuclear genes have been discovered through an abundance of annotated transcriptomes (Carpenter et al. 2019) and whole genome sequences (Goodstein et al. 2012). SLCN genes have been curated in open access databases for angiosperms (De Smet et al. 2013) and other lineages (Breinholt et al. 2021; Wolf et al. 2018). Still, sequencing large numbers of SLCN genes in numerous samples can quickly become costly and redundant. This is where target capture can save the day. By selectively sequencing hundreds of pre-identified SLCN genes, often chosen for the optimal phylogenomic application (Grover et al. 2012), the complexity of genomic samples can be overcome. With the right background sequence information, target capture makes any given gene accessible to the user with high coverage in a cost-effective manner.

What is target capture?

Analysing SLCN genes requires multiple reads to cover the same genomic region (high coverage) to obtain high-quality assemblies. The goal of target capture, also called bait capture or hybrid capture, is to achieve high coverage on (nuclear) target loci by proportionally increasing (enriching) the target DNA fragments in a genomic library. The workflow is straightforward (see Infographic): DNA is extracted using tissue-specific protocols (see Section 1 Design, sampling, and substrates in this book), sheared to the desirable fragment length (e.g., 300–700 bp for Illumina sequencing, depending on the quality of the extract), processed into genomic libraries using indexing techniques for multiplex sequencing, enriched for the target genomic regions using specific baits (see below), and sequenced on a platform with high sequencing accuracy (e.g., Illumina or PacBio Sequel).

Target capture uses custom-designed short RNA- or DNA-baits (usually between 80 and 120 bp long), also called probes, that hybridise in solution to target loci with complementary sequences (Gnirke et al. 2009). The baits are chemically modified with biotin, which binds to streptavidin. Magnetic beads coated with streptavidin bind to the biotinylated baits, which are hybridised to the target loci (Grover et al. 2012). A magnet retains the beads with target DNA fragments attached, allowing the user to wash away unwanted DNA fragments from the library. This increases the proportion of target DNA fragments in the library, a process called enrichment. The target DNA fragments can subsequently be unloaded from the magnetic beads by denaturing the target DNA-RNA bait complex (e.g., by heating to > 95 °C). Amplification with a small number of PCR cycles is usually needed before sequencing (Kozarewa et al. 2015).

Besides target capture, there are other techniques for reducing genomic complexity in DNA samples (reduced representation sequencing) depending on the target loci, taxon, and/or application (Mamanova et al. 2010). One example is amplicon sequencing (or targeted PCR) where enrichment is achieved through PCR using locus-specific primers (Meuzelaar et al. 2007). Combining this with microfluidics, dozens of nuclear loci are amplified in microdroplets (Tewhey et al. 2009). This technique is especially popular in medical and agricultural studies for variant calling (Mamanova et al. 2010), and has been successfully applied to resolve some phylogenetically challenging plant clades, too (Gostel et al. 2015; Mayland-Quellhorst et al. 2016; Morales-Briones and Tank 2019; Uribe-Convers et al. 2016). Although this technique is less expensive than target capture, specific PCR primers are required, which limits the possible number of loci that can be captured. Amplicon sequencing is therefore less suitable for non-model organisms where less information is available for primer design. The achievable taxonomic breadth is also limited due to sequence variation in the primer-binding sites that reduces primer hybridisation to the target DNA. Furthermore, locus-specific PCR amplification is compromised when working with degraded DNA, such as from herbarium specimens (Staats et al. 2011) or traded plant material for food and medicine (Novak et al. 2007). This is especially problematic for target loci longer than a few hundred base pairs, rendering this method less relevant to plant phylogenetics or population genetics.

Target capture is a robust alternative for these applications as it works independently of specific PCR primers. The small size of the RNA-baits makes this method ideal for degraded DNA and baits do not need to be an exact match for the target to be captured. The hybridisation conditions can be modified for more or less permissive binding between bait and target, and locus capture can still be successful with up to 30% bait mismatch (Johnson et al. 2019). The same bait set can work genus- (Soto Gomez et al. 2019; Woudstra et al. 2021) or even family-wide (de La Harpe et al. 2019), depending on the probe design and the level of sequence variation within the clade. Because of the in-solution hybridisation capture, the technique furthermore works independently of genome size and target copy number (Woudstra et al. 2021). These characteristics make it possible to work on a broad taxonomic scope involving samples of variable DNA degradation level, typical of studies on molecular systematics and plant identification.

How to obtain reference data?

Nuclear sequence data from a single clade member, either a whole-genome or transcriptome, are enough to design efficient target capture baits for application across the clade. As the number of complete plant genomes (e.g., 128 species, Phytozome v.13; Goodstein et al. 2012) and transcriptomes (> 1000 species through 1KP project; Carpenter et al. 2019) increase, the more opportunities become available for the analysis of plant life by target capture. When nuclear sequence data are unavailable for a clade, a transcriptome can be generated with relative ease

using RNA-Seq (Strickler et al. 2012; Van Verk et al. 2013; Chapter 15 Transcriptomics). This is especially relevant when working on a smaller budget since generating a whole-genome reference sequence can be more expensive and time-consuming to assemble. Transcriptomes provide transcribed exon-only sequence data and are therefore essential in discovering intron-exon boundaries in whole-genome sequences.

Annotated whole-genome sequences are preferred for RNA-bait panel design, since: 1) target loci can be more carefully selected with the detailed gene copy number information; 2) annotated whole-genomes provide intronic and intergenic regions (sequences immediately 5′ or 3′ of the gene), which can be included in the panel design.

The advantage of using introns and intergenic regions is the inclusion of highly variable sequences that are useful for phylogenetic inference of recent diversification events (Sang 2002). The trade-off in including these regions is that capturing them with baits is more limited to shorter taxonomic distance from the target reference taxa (e.g., the taxa used to generate the reference transcriptome and/or whole-genome). The threshold of sequence divergence for efficient bait-target hybridization is 30% (Johnson et al. 2019), which is more easily reached for variable sequences. This challenge can be overcome by designing baits that span the intron-exon boundary so that part of the bait will hybridise with a relatively conserved region in the exon (Lesur et al. 2018). Fragments captured by baits designed to only capture exons may also accidentally include parts of introns and intergenic sequences, creating a 'splash zone' (Samuels et al. 2013) of highly variable sequences.

Applications

Target capture is a cost-efficient, high-throughput, and customizable solution for plant phylogenomics and systematics (Dodsworth et al. 2019; Johnson et al. 2019), population genomics, and microevolution studies (de La Harpe et al. 2019; Villaverde et al. 2018). RNA-bait panels for target enrichment of low-copy nuclear exons have been designed for a variety of taxonomic levels (Table 1) ranging from whole phyla to species (Dodsworth et al. 2019). Panels that capture complete genes (including introns) can be designed when necessary (Folk et al. 2015), as well as protocols for targeted long-read sequencing (Lefoulon et al. 2019). Even in highly degraded DNA samples, target capture can retrieve hundreds of nuclear genes (Brewer et al. 2019; Forrest et al. 2019). It is becoming an important tool for molecular identification and characterisation of unknown plant material (Manzanilla et al. 2022), including crop variants (Lesur et al. 2018; Ogutcen et al. 2018). The technique can also be combined with other widely used genomic techniques (Table 2).

Table 1. Examples of available universal and clade-specific target capture panels. Number of exons and introns as well as total number of bases targeted are as reported in the original publications.

Taxonomic level	Number of loci	Exons/introns	Total target size (bp)	Reference
Flagellate plants and gymnosperms	248	451 exons	150,369	Breinholt et al. 2021
Ferns	25	Exons only	39,134	Wolf et al. 2018
Angiosperms	353	Exons only	260,802	Johnson et al. 2019
Order				
Saxifragales	301	Not reported	Not reported	Folk et al. 2019

Taxonomic level	Number of loci	Exons/introns	Total target size (bp)	Reference
Family				
Annonaceae	Not reported	469 exons	364,653	Couvreur et al. 2018
Apocynaceae	853	Exons only	1,545,593	Straub et al. 2020
Arecaceae	4,184	Exons only	4,287,662	de La Harpe et al. 2019
Asparagaceae - Agavoideae	2,473	3,709 exons	Not reported	Heyduk et al. 2016
Asteraceae	1,061	Exons only	Not reported	Mandel et al. 2014
Bromeliaceae	1,776	Exons only	± 2,300,000	Yardeni et al. 2022
Cactaceae	120 (+ A353)	469 exons	136,495	Acha and Majure 2022
Fabaceae	507 (423 SLCN)	Exons only	737,309 (SLCN only)	Vatanparast et al. 2018
Fabaceae - Detarioideae	289	1,021 exons	359,269	Ojeda et al. 2019
Fabaceae - Mimosoideae	964	Exons only	1,134,513	Koenen et al. 2020
Gesneriaceae	830	Exons only	776,754	Ogutcen et al. 2021
Melastomataceae	384 (266 from A353)	Exons only	Not reported	Jantzen et al. 2020
Ochnaceae	275	Exons only	660,000	Schneider et al. 2021
Orchidaceae	963	6,005 exons	Not reported	Eserman et al. 2021
Salicaceae	972 (593 SLCN)	Exons only	Not reported	Sanderson et al. 2020
Sapotaceae	1,241	Exons only	Not reported	Christe et al. 2021
Genus				
Aloe (Asphodelaceae)	189	1,029 exons	353,794	Woudstra et al. 2021
Anacyclus (Asteraceae)	872	Not reported	Not reported	Manzanilla et al. 2022
Begonia (Begoniaceae)	1,239	Exons + introns	Not reported	Forrest et al. 2019
Burmeistera (Campanulaceae)	745	Exons only	Not reported	Bagley et al. 2020
Cyrtandra (Gesneriaceae)	570	Exons only	180,784	Kleinkopf et al. 2019
Dioscorea (Dioscoreaceae)	260	Exons only	441,626	Soto Gomez et al. 2019
Heuchera (Saxifragaceae)	278	Including introns	378,553	Folk et al. 2015
Hibiscus (Malvaceae)	87	Exons only	Not reported	Eriksson et al. 2021
Hosta (Asparagaceae)	283	676 exons	171,365	Yoo et al. 2021
Inga (Fabaceae)	276	907 exons	Not reported	Nicholls et al. 2015
Lens (Fabaceae)	Full exome	Exons only	85 Mbp	Ogutcen et al. 2018
Silene (Caryophyllaceae)	50	Exons (131) + introns	104,374	Cangren et al., unpublished
Rubus (Rosaceae)	926	8,963 exons	Not reported	Carter et al. 2019
Single species				
Euphorbia balsamifera (Euphorbiaceae)	431	Exons only	709 kbp	Villaverde et al. 2018
Quercus robur (Fagaceae)	9,748	Including introns	150 bp per (sub)gene	Lesur et al. 2018

Table 2. Applications for target capture and in combination with other methods.

Technique	Principle	Application	Reference
Target capture	Target enrichment using in-solution hybridisation with specifically designed baits: short oligonucleotides complementary to target loci.	Phylogenomics, population genomics	Gnirke et al. 2009; Grover et al. 2012
RAD-Seq + target capture (Rapture)	Using custom baits to capture selected restriction-site associated DNA (RAD) tags.	Population genomics with museum specimens	Ali et al. 2016; Lang et al. 2020
Target capture + genome skimming (Hyb-Seq)	Adding an unenriched library to the enriched sample before sequencing to obtain low-coverage sequencing results from non-target nuclear regions and organellar genomes.	Phylogenomics, population genomics, comparing chloroplast and nuclear phylogeny.	Weitemier et al. 2014
Target capture + allele phasing	Estimation of ploidy level based on allelic frequency and allelic ratio from the number of reads for each allele.	Estimation of ploidy level from museum specimens.	Viruel et al. 2019
Target capture + molecular identification	Using target capture to obtain high-coverage sequence data for SLCN genes to identify unknown samples of traded plants.	Trade monitoring, authentication of medicinal plants.	Manzanilla et al. 2022
Target capture + repetitive sequence analysis	Using off-target reads to investigate levels of DNA repetition across a taxonomic clade.	Structural evolution of genomes, repetitive DNA analysis.	Costa et al. 2021

Use of target capture for molecular identification

The effective enrichment of degraded DNA samples, wide taxonomic range, and increased availability of custom bait panels make target capture ideally suited for molecular identification of plants (Liu et al. 2017; Manzanilla et al. 2022). Target capture can identify plant material at different taxonomic ranks, from infraspecific variants to species and higher ranks. Intraspecific variation data has been used for population genomic studies (Villaverde et al. 2018) as well as for important agricultural plant species to identify potential crop improvements (Lesur et al. 2018; Ogutcen et al. 2018; Villaverde et al. 2018).

The success of molecular identification depends on a curated database of taxonomically verified reference sequences with a corresponding comprehensive phylogeny (Howard et al. 2020). The taxonomic breadth and density of sampling in the reference sequence database determines the scale and scope of a molecular identification experiment. The more comprehensive the reference database, the greater the confidence with which unidentified material can be defined phylogenetically. With target capture such a comprehensive database can be quickly and cost-effectively obtained, even for very diverse clades (e.g., aloes; Woudstra et al. 2021).

Once sequenced, unknown material can be identified using either genomic distance (Batovska et al. 2016) or phylogenetic approaches (e.g., maximum likelihood (Nguyen et al. 2015), ASTRAL (Zhang et al. 2018), multi-species coalescent (Yang and Rannala 2017)), where the confidence of identification is defined by the node support.

Analysing target capture data from mixed samples, where material from different species is combined into one sample, is complex and requires long-read sequencing and rigorous phasing. Short sequence reads require assembly into longer fragments (contigs), increasing the risk of erroneous assemblies in mixed samples where reads belonging to different species might end up in the same contig. Long sequence reads can be sorted based on variable sites (phasing) and assigned to species directly, circumventing the assembly problem for mixed samples. If traditional markers give sufficient resolution, metabarcoding experiments (see Chapter 11 Amplicon metabarcoding) can be designed for a more cost-effective approach.

Designing a target capture experiment

Universal vs custom panel

The research question will determine whether a customised bait panel is needed for a study, and the choice is a trade-off between cost and detail (Figure 1). Universal bait panels are less expensive (e.g., 50% less than a customised kit on the first order; Daicel Arbor Biosciences, Ann Arbor, USA) but a customised bait panel will provide higher phylogenomic resolution through the capture of more variable sequences, higher on-target read ratios and superior target recovery rates. Universal and customised bait panels can be simultaneously used to increase the number of loci captured for a phylogenomic study (Larridon et al. 2019; Ogutcen et al. 2021; Shah et al. 2021).

Universal bait panels are commonly designed for resolving deeper phylogenetic nodes (e.g., angiosperms; Buddenhagen et al. 2016; Johnson et al. 2019) between orders, families, and sometimes genera, whereas customised bait panels generally target recently evolved clades (e.g., *Aloe* (Woudstra et al. 2021), *Begonia* (Forrest et al. 2019), *Dioscorea* (Soto Gomez

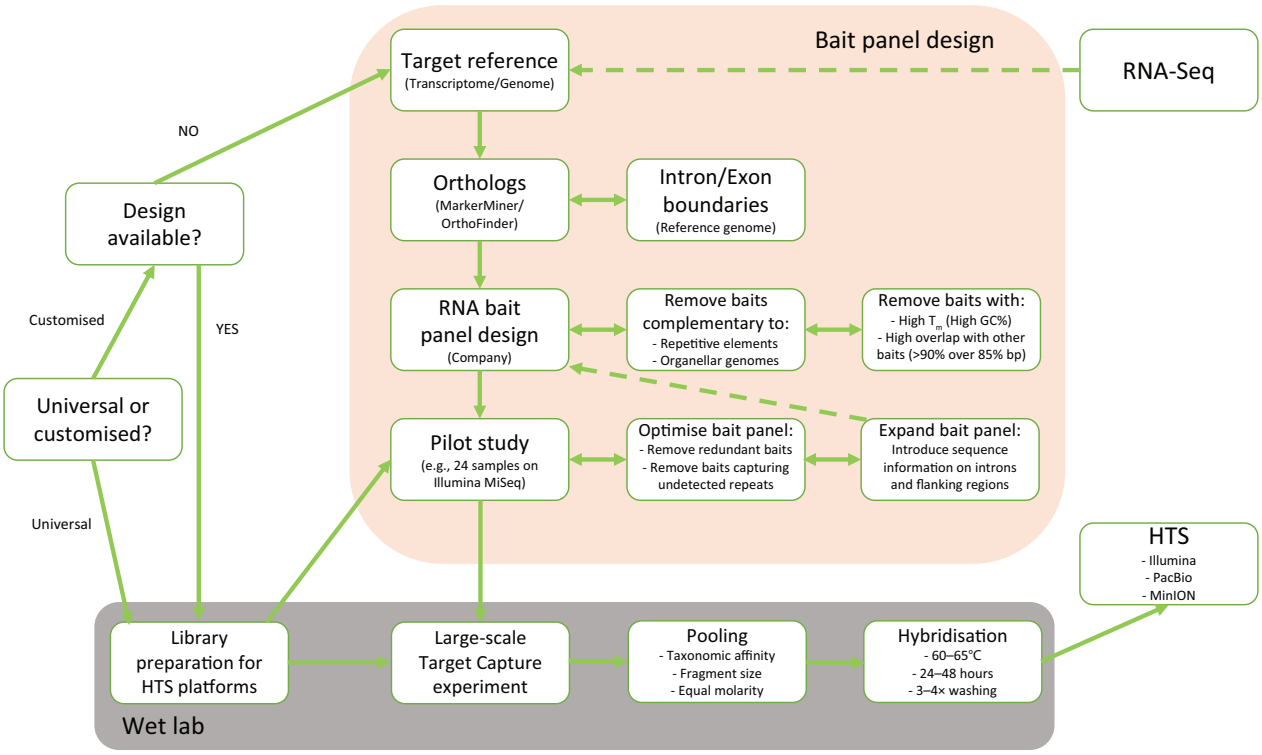

Figure 1. Flow-chart describing the workflow of a target capture experiment.

et al. 2019), *Inga* (Nicholls et al. 2015), *Rubus* (Carter et al. 2019)) at the family and genus rank. It is advisable to check if a clade of interest is already included in a commercially available universal bait panel and test the recovery and sequence divergence in a pilot study. Universal bait panels generally achieve moderate enrichment efficiency (≥ 50%, defined as the proportion of on-target reads) in a wide variety of taxonomic samples. However, they capture highly conserved genes and rarely achieve the recovery rates typical of custom bait panels. For example, although the Angiosperms353 bait panel has been used to resolve infrageneric relationships (Frost et al. 2020; Murphy et al. 2020) and even capture infraspecific variation (Beck et al. 2021) in eudicots, recovery rates in monocot clades are typically low, e.g., < 37% in *Cyperus* (Larridon et al. 2019).

Customised bait panels offer greater recovery and detail but are sensitive to the taxonomic distance from the reference used in the design. For example, target recovery with the *Aloe* bait panel decreased from an average of 93.6% in *Aloe* samples to 74.3% in the sister clade *Bulbine* (Woudstra et al. 2021). If the taxonomic scope of the project is limited (e.g., a single genus or lower rank) customised bait panels are usually better as they offer more adaptability to a specific research question, such as for functional and evolutionary studies. For example, target capture has been used to identify functional genes associated with secondary metabolite biosynthesis in *Inga* (Nicholls et al. 2015) or with shade-growth adaptation in *Begonia* (Forrest et al. 2019).

Input data

Designing a custom bait panel requires transcriptome or whole-genome sequences from at least one, but preferably more, taxon in the clade of interest or from a closely related clade (Kleinkopf et al. 2019). If not available from published resources (e.g., Phytozome (Goodstein et al. 2012), 1KP (Carpenter et al. 2019), NCBI Transcriptome Shotgun Assembly database), transcriptomes can be obtained from RNA sequencing (e.g., Illumina RNA-Seq (Van Verk et al. 2013)) and assembled with a tool such as Trinity (Grabherr et al. 2011). It is very important to use transcriptomes from species that cover the phylogenetic diversity within the group of interest. An appropriate species can be identified using a previously published phylogeny or from taxonomic affinities within the clade. The selected transcripts (see next section) are then aligned to a reference genome and intron/exon boundaries are identified as well as the copy number. A well-annotated model plant genome such as *Arabidopsis thaliana* (Berardini et al. 2015) (for dicots) or *Oryza sativa* (Ouyang et al. 2007) (for monocots) provides a well-studied background.

Ortholog detection

SLCN genes can be retrieved from a set of transcriptomes using software such as Markerminer (Chamala et al. 2015) based on a set of predetermined, angiosperm-wide, SLCN genes (De Smet et al. 2013). The detected SLCN genes are aligned to a user-specified reference genome to determine the intron-exon boundaries. The output includes a list of detected SLCN genes, the corresponding transcript identities with their presence in each transcriptome indicated, and FASTA alignments of all detected genes, both with and without the reference genome attached. Other tools such as OrthoFinder (Emms and Kelly 2019) detect orthologous loci (see Chapter 17 Species delimitation) in a set of diverse sequences, and can increase the number of loci available for bait panel design with downstream advantages for accuracy.

It is important to determine, as much as possible, the copy number of genes identified at this stage to avoid including paralogs in the target design. Annotated whole-genome sequences have an advantage here. When using transcriptomes, Markerminer can indicate the copy

status of identified loci based on the curated dataset from De Smet et al. (2013). Additionally, a reciprocal blast of putative SLCN loci against the transcriptome can be used to identify near-identical matches, providing an indication on the presence of paralogs.

Target gene selection: which criteria to choose

It is usually unnecessary to include all detected loci as they may vary in their phylogenomic value and there is a limit to what a bait panel can efficiently cover. The smallest RNA-bait panels from MYBaits (Arbor Biosciences), for example, include up to 20,000 baits between 80 to 120 bp in length. Larger bait panels are considerably more expensive. It is advisable to use 2–3x tiling in the bait design so that the whole set covers each base of the target loci with 2–3 baits or more (Figure 2). Assuming 3x tiling with 20,000 baits of 80 bp each, up to 530,000 bp of target loci can be covered this way. The coverage generally decreases towards the ends of exons, when introns and intergenic regions are not covered in the target design, due to the removal of identical baits (Figure 2).

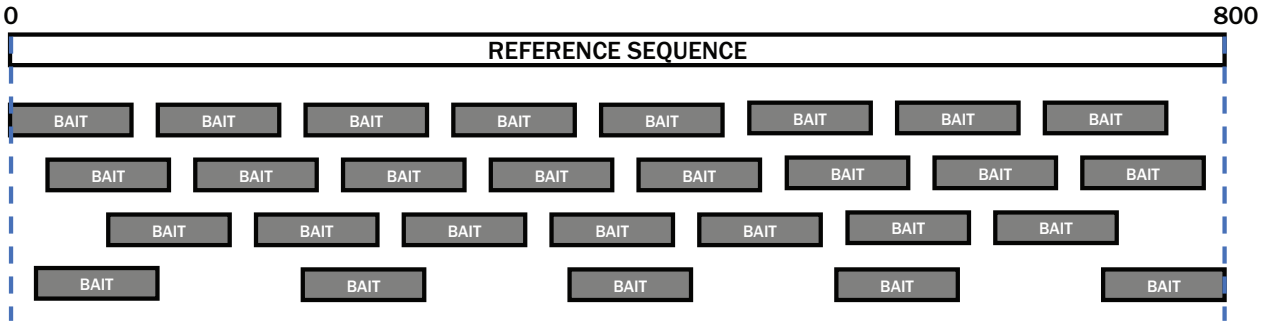

Figure 2. Schematic representation of tiling. The reference sequence at the top represents a hypothetical 800 bp exon with dotted blue lines indicating the intron-exon boundaries. Hypothetical baits are 80 bp long. This example is based on a 3x tiling strategy where each nucleotide is (on average) covered by three unique baits. The bait coverage decreases towards the ends of the exon as the target design of this hypothetical example did not include introns or intergenic regions.

Firstly, prioritising target loci recovered across all taxa in the reference dataset is recommended to ensure consistent recovery and to include as much variation as possible in the bait panel design. If a locus is represented by only one taxon in the design, the resulting capture will be skewed towards samples more closely related to that taxon. This may result in uneven recovery among samples in a pooled capture experiment (see Multiplexing and pooling below).

Secondly, for phylogenomic and molecular identification purposes, loci with low variability should be excluded. The target loci should be variable enough to resolve challenging phylogenetic clades. When designing a bait panel for broader taxonomic applications (e.g., above genus-level), limiting the inclusion of highly variable genes can be considered to keep recovery levels consistent across taxonomic scales. The variability of loci can be assessed based on pairwise sequence identity (ideally < 95% between the reference sequences), phylogenetic resolution on gene trees, and the number of parsimony-informative sites (e.g., ≥ 20 per 1000 bp). Additionally, metrics such as the amount and proportion of missing data can be obtained in a useful summary diagram with the AMAS tool (Borowiec 2016).

Thirdly, it is important to make sure that target loci are long enough to be covered efficiently by the RNA-baits (i.e., > 400–500 bp is recommended). This is especially relevant when targeting exons. Exons shorter than the bait length will not be captured efficiently since bait sequences that span two intron-exon boundaries will have reduced recovery as they only bind partly to one exon.

Finally, target genes should be checked for repetitive regions such as microsatellites or transposons. These can be detected by the presence of short, repetitive sequences of low complexity (e.g., $(ATC)_n$) and should not be included in the bait panel design. Baits spanning these repetitive elements will likely hybridise in many places in the genome, reducing the hybridisation efficiency and recovery for the target gene. Similarly, including highly conserved and high-copy regions common to plants, such as the MYB-domain (Stracke et al. 2001), in bait design leads to promiscuous binding across a large number of loci in the same sample. These can be identified and removed by programmes such as RepeatMasker (Chen 2004) that rely on publicly available genomes (Li et al. 2019).

Designing RNA-baits

Selected SLCN genes should first be aligned to a reference sequence from an annotated genome to indicate putative intron-exon boundaries. These alignments will form the basis of the RNA-bait design. An example pipeline can be found in the publication for the Angiosperms353 panel (Johnson et al. 2019). Companies providing bait panel design services include Arbor Biosciences (MYBaits, Ann Arbor, MI, USA), Agilent (SureSelect, Santa Clara, CA, USA), Roche (KAPA, formerly SeqCap, Basel, Switzerland), and Integrated DNA Technologies (xGen, Coralville, IO, USA). The initial bait designs are often much larger than necessary, and filtering can reduce the number of baits and improve the hybridisation efficiency in the panel.

Firstly, baits need to be checked for potential overlap with high-copy sequences from organellar genomes (plastomes, mitochondrial genomes, and nuclear ribosomes) by mapping them to published sequences, which are typically available in NCBI databases. Secondly, GC-content in baits should be normalised across the panel. The hybridisation temperature governs the specificity of capture and baits with the same melting temperature (T_m) should hybridise evenly. Additionally, a high GC content in baits will lead to more off-target hybridisation as these baits are more likely to bind efficiently to GC-rich areas in the genome. Baits with a GC-content > 75% should therefore be removed, though one might lower the threshold to 60%. Thirdly, identical, or nearly identical, baits should be removed to reduce redundancy in the dataset as well as bias towards regions covered by identical baits. This should be done carefully however to not reduce the desired tiling of the bait panel. For example, removing baits with > 90% identical sequence over 85% of the total bait sequence generally works for 3x tiling. The digital panel design provided by the company should then be checked for accuracy by mapping the designed baits against all reference target sequences (e.g., selected transcripts for the panel design). This is to make sure that the baits align with the target sequence, are not too divergent from the target sequence and are tiled uniformly across the length of all genes.

Setting up a pilot study

A custom-designed bait panel needs to be tested to ensure it efficiently captures the target sequence prior to a large-scale study. A cost- and time-efficient pilot study can include up to 24 samples using an Illumina MiSeq platform (Soto Gomez et al. 2019). The pilot study should mirror the planned target sequence capture study, fragment shearing, and size selection protocols can be optimised here (Soto Gomez et al. 2019; Woudstra et al. 2021). How effectively the bait panel enriches outgroup samples necessary to root the resulting phylogenomic tree should also be tested. The pilot study may also reveal ineffective baits based on the lack of recovery for a certain (part of a) locus. Conversely, it may also reveal baits that capture too many sequences,

such as baits being (partially) complementary to an undiscovered repetitive element. These baits can be removed from subsequent orders of the panel with the supplier to improve the efficacy.

Information on introns and flanking regions may be elucidated from the 'splash zone' in a pilot study that can subsequently be added to the bait panel design (Cangren et al., unpublished). In these cases where the bait panel is expanded, a second pilot study may be required. The results of a pilot study are generally published along with the design of a custom bait panel to the benefit of other researchers who may use the same custom bait panel (e.g., *Dioscorea* (Soto Gomez et al. 2019), *Aloe* (Woudstra et al. 2021), *Asclepia* (Weitemier et al. 2014)).

The target capture procedure

Target capture sequencing uses genomic DNA libraries prepared for sequencing on HTS platforms. These libraries consist of DNA fragments, usually of a controlled size, obtained from source plant material. The source DNA fragment is flanked by standardised identifier sequences (indexes or sometimes also called barcodes) to help identify the sample origin of a sequence read and a standardised adapter sequence to allow binding of the DNA fragment to the flow cell of the sequencer. The number of bp DNA from the source genome in a library fragment (insert size is therefore smaller than the fragment itself:

$$\text{insert size} = \text{average library fragment size} - 2 \times (\text{length of adapter+index sequence})$$

The library preparation procedure is not discussed here, but details can be found in Chapter 9 Sequencing platforms and data types.

Multiplexing and pooling

Sequences from different samples can be distinguished by labelling (indexing) each library with its own unique identifier. Combining differently labelled libraries into one sequencing run (multiplexed sequencing or multiplexing) is a common strategy to reduce per-sample costs. To further reduce the per-sample cost of target capture experiments, libraries from different samples can be combined in one tube for simultaneous enrichment (pooling). This reduces the number of RNA-baits necessary to enrich the same number of samples, and significantly reduces costs. Efficient target enrichment is routinely achieved with up to 48 samples per RNA-baits reaction and even 96-plexing strategies have been successful (Hale et al. 2020). Before beginning an enrichment experiment, it is important to consider pooling strategies.

When deploying a universal bait panel, libraries from different taxonomic groups, particularly at the family rank and above, must be separated. The closer the sample is to the reference taxa the higher the similarity between the bait sequences and the target sequences (Brewer et al. 2019). If the taxonomic distance between the target reference and the samples is non-uniform, samples most closely related to the target reference will hybridise more efficiently to the baits than more distantly related samples and bias the DNA recovery.

In all target capture experiments, libraries in the same pool should contain similar fragment sizes. Short fragments can move around much easier in a solution and will thus encounter the RNA-baits more often, increasing their chances of capture. Mixing short and long fragments in the same target capture reaction can therefore skew the enrichment towards the shorter libraries (Cruz-Dávalos et al. 2017). When pooling very short libraries (those with insert size equal or shorter to the bait size), a small difference in fragment size will proportionally be quite a large difference and it is therefore good practice to pool libraries within a smaller size range. It is also

not advisable to mix libraries from degraded (i.e., herbarium specimens) and high-quality (e.g., fresh/silica) DNA samples. Degraded DNA samples are known to produce libraries with lower nucleotide diversity because a large majority of genomic fragments is too small (e.g., smaller than the bait size) for the final library and are therefore discarded during the library preparation by size selection (see Chapter 9 Sequencing platforms and data types). High molecular weight samples, on the other hand, allow for much more control on fragment size distribution through careful shearing protocols (see Chapter 9 Sequencing platforms and data types). A much higher fraction of the genomic fragments can therefore be retained for library preparation, producing libraries with a higher nucleotide diversity. This increases the probability that fragments complementary to the baits are present and anneal to a greater number of the baits in solution. This can decrease the enrichment success for the less diverse libraries from degraded samples if they pooled.

Pooling libraries for target capture sequencing
- Calculating volumes for equimolar pooling -

Bait panel = MYBaits® custom Hyb-DNA kit (RNA baits)
Desired input DNA for pool = 100 ng
Desired pool volume = 7 µL
Number of indexed libraries = 4
Average library fragment size in pool = 350 bp

Step 1: calculate desired pool concentration

$$C\ (nM) = (C\ (ng/\mu L) * 10^6) / (660^a * \#bp)$$

[a] average molecular weight of 1 DNA bp: 660 g/mol

$$C_{pool} = (14.3 * 10^6) / (660 * 350) = 61.9\ nM$$

Step 2: calculate volume needed from each library

$$V_{input} = (V_{pool} * C_{pool}) / (\#libraries\text{-}in\text{-}pool * C_{input})$$

Library	C_{input} (ng/uL)	Fragment size (bp)	C_{input} (nM)	C_{pool} (nM)	V_{input} (uL)
A	20	350	86.6	61.9	1.25
B	25	350	108.2	61.9	1.00
C	30	350	129.9	61.9	0.83
D	19	350	43.6	61.9	2.48

Step 3: dilute the pool to the final volume (V_{pool} = 7 µL)

Figure 3. Example describing the strategy of pre-target capture pooling.

The number of DNA fragments from each sample in the same pool should be equal, i.e., be present in equimolar quantities. A library with a higher number of DNA fragments than the others in the pool will be overrepresented and potentially bias the DNA sequences that are enriched. Diluting libraries to the same molarity (usually in nM) before pooling is therefore generally advised. An example of how to calculate pooling parameters is shown in Figure 3. Bait reactions work with set amounts of input DNA library in small volumes. The commonly used Daicel MYBaits kit recommends between 100–500 ng of input DNA per 7 µL baits reaction. It is therefore often necessary to concentrate the library pools before starting the hybridisation reaction with a vacuum drying centrifuge optimised to the library solvent, generally a Tris-HCl solution. Alternatively, libraries can be concentrated with silica-column-based PCR purification kits (e.g., Qiagen MinElute), or by ethanol precipitation followed by suspension in the correct volume of buffer. These latter methods are very effective in preventing the concentration of salts in the final sample, though yields may be lower.

Hybridisation and target capture

A target capture wet-lab protocol has three steps: denaturing the DNA libraries, hybridising with target-specific baits, and post-capture washing to remove unwanted DNA fragments. An example protocol using the Daicel Arbor MYBaits kit is detailed here.

In the first step, the genomic libraries are denatured at > 95 °C and 'blocker' oligonucleotides are added that bind to the adapter sequences. This is to keep the single-stranded fragments from hybridising back to their complementary strands. The blockers also reduce any interference of the adapter sequences during hybridisation, in case the baits themselves contain complementary sequences to the adapters and/or index primers used.

In the hybridisation reaction, the target-specific baits are added to each pool and hybridisation will occur at a constant temperature of 60–65 °C (depending on the specifics of the bait panel) for a minimum of 16 hours. These parameters should always be optimised when setting up a target capture protocol. Longer hybridisation times (≥ 24 hours) are needed for enrichment of more complex genomic libraries, such as those from larger genomes and from universal kits. In these reactions, the baits take longer to encounter the target DNA fragments.

For samples that are expected to underperform (short libraries, herbarium samples, or samples taxonomically distant to the target reference), the hybridisation temperature can be dropped to < 60 °C and the hybridisation time extended to 48 hours. To prevent evaporation and any potential loss of target DNA, a small amount of hydrophobic wax can be added on top of the hybridisation reaction. If using a thermocycler, the heated lid should also be on at ± 105 °C to prevent evaporation.

Finally, the magnetic streptavidin beads are added to the reaction mixture to bind the target-bait hybrids. These streptavidin beads need to be washed to remove any storage buffer before they are added to the target capture pools. Once ready, the tube with magnetic beads and bead-bound target DNA can be placed on a magnetic tube rack to concentrate and anchor the beads, allowing the non-bound DNA fragments to be washed away.

PCR amplification

The amount of DNA in enriched pools often needs to be PCR amplified to generate sufficient detectable fragments for sequencing on HTS platforms. This is especially important when capturing loci from large genomes (e.g., Woudstra et al. 2021) as the amount of target DNA in genomic libraries will be proportionally low.

Post-capture amplification can either be done with the DNA still bound to the beads using a specific hot start polymerase or after removing them from the beads and using a standard high-fidelity polymerase (PFU or Q5, various suppliers). In the latter case, the target DNA is released by

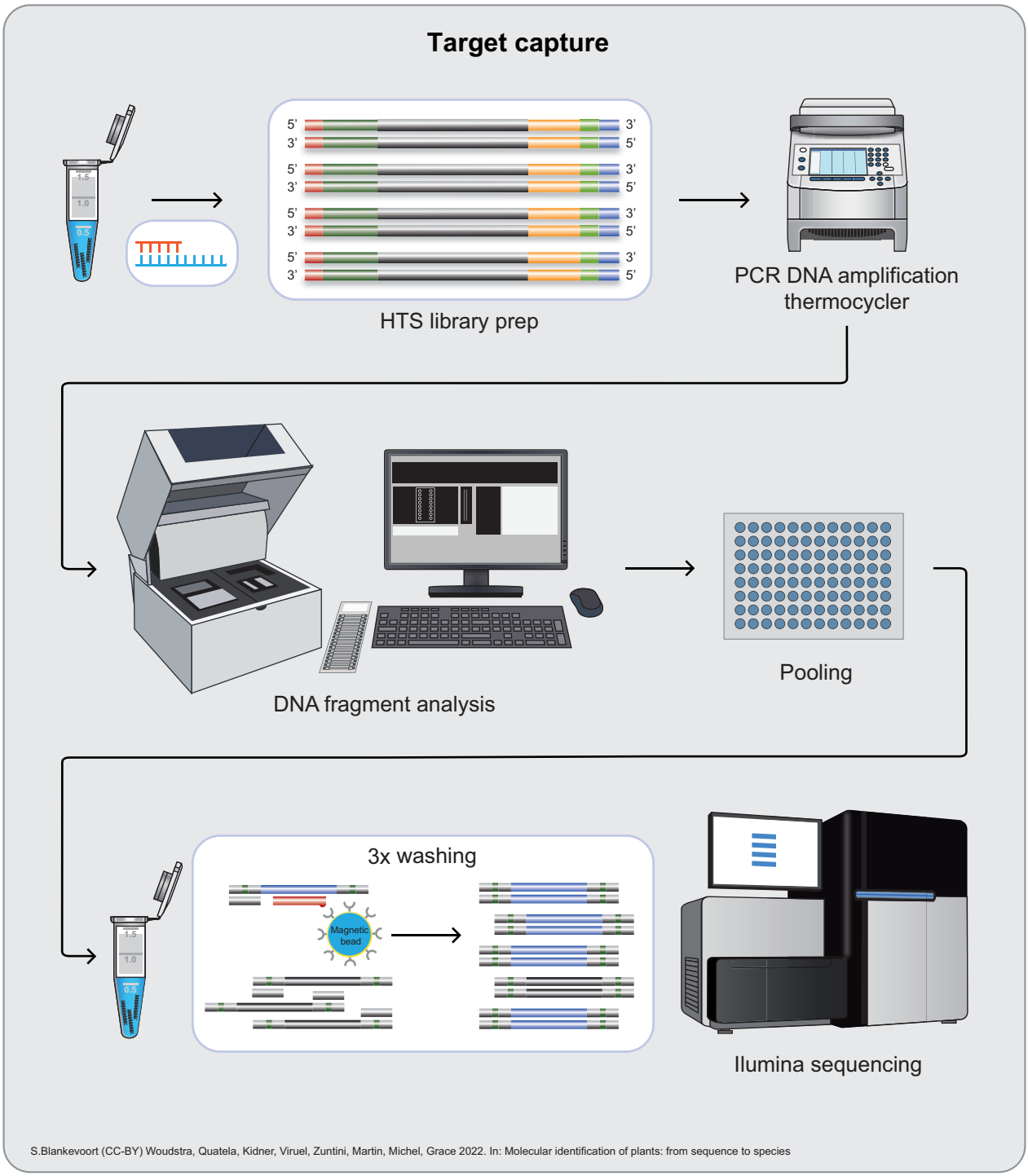

Figure 4. Chapter 14 Infographic: Visual representation of the content of this chapter.

denaturation at > 95 °C in a suitable solvent, i.e., Tris-HCl solution (no EDTA should be present since this will inhibit the polymerase) and immediately transferred to a magnetic tube rack to separate the DNA from the baits, gently removing the solvent containing target DNA and transferring it to clean tubes. A concentrated high-grade detergent such as TWEEN-20 is often added prior to denaturation to enhance the release of target DNA. PCR amplification is then done in-solution on the target DNA using universal PCR primers that bind to the adapter sequences.

Optimising the number of PCR cycles (e.g., via qPCR with a dilution of the captured product) is generally advised since too many PCR cycles can increase the chance of false positives in

the form of random errors in the sequences, which cannot be corrected based on the sequencing data. Furthermore, excessive PCR cycles introduce unnecessary PCR clones or duplicates. Performing just enough PCR cycles to obtain a pool into the desired concentration range for the sequencing platform (generally ≥ 3 nM) is therefore recommended.

As a rough qualitative indication of target capture success, the fragment size distribution can be determined using high-precision electrophoresis instruments such as an Agilent TapeStation. After a successful experiment, there will be a peak in the expected library fragment size range (insert size + adapter and index sequences). While exact target capture success can only be determined from sequencing results, this post-capture fragment distribution analysis acts as an extra quality assurance prior to sequencing.

Raw sequence data

Sequencing output of target capture sequencing experiments is in the same format as for other HTS experiments. Demultiplexing and quality filtering/trimming of the raw reads is required. Demultiplexing is often done automatically with Illumina sequencing data, using the BaseSpace firmware. For Oxford Nanopore and/or PacBio reads, there are freeware options such as PoreChop (https://github.com/rrwick/Porechop). Read filtering/trimming is based on FastQC (Andrews et al. 2010) reports that indicate the average quality of each base position in the reads, as well as various other read quality statistics, and is most commonly performed using freeware such as Trimmomatic (Bolger et al. 2014).

Assembly

High-quality reads are assembled into consensus sequences for target loci to make good sequence alignments between samples. For target capture experiments, reads are aligned to the target reference used in the RNA-bait design, a process called mapping, to reduce the complexity of the assembly. Several tools are available to assemble mapped reads de novo, meaning without further use of the reference sequences. HybPhyloMaker (Fér and Schmickl 2018) is the most complete tool for phylogenomic analysis as it encompasses every step from raw reads to coalescent species tree. SECAPR (Andermann et al. 2018) is useful when captured data needs to be phased into different alleles, while HybPiper (Johnson et al. 2016) is the most widely used pipeline due to its convenient summary statistics output.

HybPiper uses a combination of different mapping and assembly tools to retrieve target sequences from large target capture datasets. Reads are mapped to the reference sequence using a Burrows-Wheeler Aligner (Li and Durbin 2009) per target locus and stored in separate directories. A de novo assembly is performed on each locus individually using SPAdes (Bankevich et al. 2012), aligned back to the reference, and split into exon and intron sequences using Exonerate (Slater and Birney 2005). The sequence length for each target gene in each sample and the number of reads mapped for each sample can be used to determine the on-target ratio (or enrichment efficiency) for each sequenced sample as well as the percentage of target sequences recovered, the criteria used to determine the bait panel efficiency. Intron and intergenic spacer sequences captured from exon-flanking regions can be retrieved as additional information in the target enrichment experiment. HybPiper gives paralog warnings based on the number and frequency of contigs assembled in the same sample for a particular gene. This can be checked using unrooted gene trees in, for example, SplitsTree (Huson and Bryant 2006) where long separation between clades would indicate paralogy. Paralogous sequences can be

problematic in phylogenetic inference if not dealt with correctly (Gabaldón and Koonin 2013) and it is common to remove putative paralogous genes from the phylogenomic dataset.

Concluding remarks

Target capture sequencing achieves reproducible high-quality sequencing results for hundreds of targeted SLCN genes or, in fact, any desired target gene. By reducing the complexity of genomic libraries, high-coverage sequencing results of single-copy genes can be obtained regardless of the organisms' genome size and DNA degradation rate. These characteristics make target capture ideally suited for molecular identification studies (Manzanilla et al. 2022) as well as for taxonomic studies using phylogenetics.

The method is being refined as the underlying molecular techniques (Pel et al. 2018) and computational tools necessary for the sample preparation and analysis continue to develop. The ability to combine dozens of samples in a single enrichment reaction (Hale et al. 2020) drives down the per-sample cost, making target capture a very attractive alternative to more expensive and exhaustive whole genome sequencing experiments.

Questions

1. Explain the difference between universal and customised bait panels. What are the advantages (and drawbacks) of a customised approach?
2. Explain why target capture sequencing is potentially very suitable for obtaining low-copy nuclear genes from herbarium samples.
3. How does in-solution hybridisation-based target capture ensure enrichment of a HTS library?
4. Explain the difference between target capture sequencing and other genomic sequencing protocols, such as genome skimming and whole genome sequencing. What are the potential benefits (and drawbacks) of this technique?

Glossary

Bait – Short oligonucleotide (80–120 bp of RNA or DNA) that is used to capture target sequences in a genomic library. Baits are chemically modified (biotinylated) so they can be bound to magnetic streptavidin beads whilst hybridised to the target DNA for removal from the genomic library.

DNA fragment shearing – The controlled breaking of DNA strands into random smaller fragments by restriction enzymes or, more commonly, by ultrasound shearing (ultrasonication).

Exon – Coding part of a gene. Can be determined from mRNA in RNA-Seq experiments.

Genomic library – A DNA sample containing fragments representative of the different genomes in an organism, e.g., organellar and nuclear genomes, prepared for HTS by addition of platform-specific adapters and sample-specific unique identifiers (in the case of multiplex sequencing).

Infrageneric relationships – Relationships between taxa or taxonomic groups within the same genus.

Infraspecific variation – Variation found within the same species. Can refer to characteristic differences between subspecies or varieties of the same species. In phylogenetics this refers to DNA sequence variation between individuals, populations or subspecies and varieties.

Intron – Non-coding part of a gene, which is often more variable than the exon sequence due to relaxed selective pressures. It is transcribed into RNA in the nucleus but often spliced out of the mRNA that is exported to the ribosomes.

Orthologs – Sequences of the same gene (copy) in different organisms or species. Orthologous loci represent the true evolutionary relationships between organisms and species as their sequences evolved (virtually) independently in different taxa.

Paralogs – Derived copies of a (nuclear) gene in the same organism that arose through either gene or whole genome duplication. Sequences of paralogs can be highly similar, making it difficult to separate them (through phasing of reads) prior to phylogenetic analysis.

Single-to-low copy nuclear (SLCN) gene – A gene located in the nuclear genome that has only one or very few copies per haplotype (a copy of the genome). The limited presence of paralogs for these genes is an advantage, particularly in plant phylogenomics, since it reduces the problem of phasing (sorting out copies of genes from sequencing data).

Tiling (in bait panel design) – A bait panel design approach to ensure consistent coverage across all target loci by more than one bait, without duplicating baits. The principle is to shift the sequence of the second bait slightly downstream of the target sequence so that it overlaps for a large part with the first bait but also captures a new part of the target sequence. See Figure 2 for a visual example.

Unrooted gene trees – A phylogenetic tree calculated from a DNA sequence alignment of a single gene without a rooting point. This is commonly done for quick exploratory analysis of relationships between the organisms studied for a particular gene. In phylogenomics, it is widely used as a tool to identify potential paralogs in a sequence alignment, which will be indicated by a strong bipolar split in the unrooted tree with the sequences of the respective paralogs at each end.

References

Acha S, Majure LC (2022) A new approach using targeted sequence capture for phylogenomic studies across Cactaceae. Genes (Basel) 13, 350. https://doi.org/10.3390/genes13020350

Ali OA, O'Rourke SM, Amish SJ, Meek MH, Luikart G, Jeffres C, Miller MR (2016) RAD Capture (Rapture): flexible and efficient sequence-based genotyping. Genetics 202, 389–400. https://doi.org/10.1534/genetics.115.183665

Andermann T, Cano Á, Zizka A, Bacon C, Antonelli A (2018) SECAPR-a bioinformatics pipeline for the rapid and user-friendly processing of targeted enriched Illumina sequences, from raw reads to alignments. PeerJ 6, e5175. https://doi.org/10.7717/peerj.5175

Andrews S, Krueger F, Segonds-Pichon A, Biggins L, Krueger C, Wingett S (2010) FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics.

Bagley JC, Uribe-Convers S, Carlsen MM, Muchhala N (2020) Utility of targeted sequence capture for phylogenomics in rapid, recent angiosperm radiations: neotropical *Burmeistera* bellflowers as a case study. Mol. Phylogenet. Evol. 152, 106769. https://doi.org/10.1016/j.ympev.2020.106769

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–477. https://doi.org/10.1089/cmb.2012.0021

Batovska J, Blacket MJ, Brown K, Lynch SE (2016) Molecular identification of mosquitoes (Diptera: Culicidae) in southeastern Australia. Ecol. Evol. 6, 3001–3011. https://doi.org/10.1002/ece3.2095

Beck JB, Markley ML, Zielke MG, Thomas JR, Hale HJ, Williams LD, Johnson MG (2021) Is Palmer's elm leaf goldenrod real? The Angiosperms353 kit provides within-species signal in *Solidago ulmifolia* s.l. BioRxiv. https://doi.org/10.1101/2021.01.07.425781

Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, Huala E (2015) The *Arabidopsis* information resource: making and mining the "gold standard" annotated reference plant genome. Genesis 53, 474–485. https://doi.org/10.1002/dvg.22877

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Borowiec ML (2016) AMAS: a fast tool for alignment manipulation and computing of summary statistics. PeerJ 4, e1660. https://doi.org/10.7717/peerj.1660

Breinholt JW, Carey SB, Tiley GP, Davis EC, Endara L, McDaniel SF, Neves LG, Sessa EB, von Konrat M, Chantanaorrapint S, Fawcett S, Ickert-Bond SM, Labiak PH, Larraín J, Lehnert M, Lewis LR, Nagalingum NS, Patel N, Rensing SA, Testo W, Burleigh JG (2021) A target enrichment probe set for resolving the flagellate land plant tree of life. Appl. Plant Sci. 9, e11406. https://doi.org/10.1002/aps3.11406

Brewer GE, Clarkson JJ, Maurin O, Zuntini AR, Barber V, Bellot S, Biggs N, Cowan RS, Davies NMJ, Dodsworth S, Edwards SL, Eiserhardt WL, Epitawalage N, Frisby S, Grall A, Kersey PJ, Pokorny L, Leitch IJ, Forest F, Baker WJ (2019) Factors affecting targeted sequencing of 353 nuclear genes from herbarium specimens spanning the diversity of angiosperms. Front. Plant Sci. 10, 1102. https://doi.org/10.3389/fpls.2019.01102

Brochmann C, Nilsson T, Gabrielsen TM (1996) A classic example of postglacial allopolyploid speciation re-examined using RAPD markers and nucleotide sequences: *Saxifraga osloensis* (Saxifragaceae). Acta Universitatis Upsaliensis Symbolae Botanicae Upsalienses 31, 75–89.

Buddenhagen C, Lemmon AR, Lemmon EM, Bruhl J, Cappa J, Clement WL, Donoghue M, Edwards EJ, Hipp AL, Kortyna M, Mitchell N, Moore A, Prychid CJ, Segovia-Salcedo MC, Simmons MP, Soltis PS, Wanke S, Mast A (2016) Anchored phylogenomics of angiosperms I: assessing the robustness of phylogenetic estimates. BioRxiv. https://doi.org/10.1101/086298

Carpenter EJ, Matasci N, Ayyampalayam S, Wu S, Sun J, Yu J, Jimenez Vieira FR, Bowler C, Dorrell RG, Gitzendanner MA, Li L, Du W, K Ullrich K, Wickett NJ, Barkmann TJ, Barker MS, Leebens-Mack JH, Wong GK-S (2019) Access to RNA-sequencing data from 1,173 plant species: The 1000 plant transcriptomes initiative (1KP). Gigascience 8, giz126. https://doi.org/10.1093/gigascience/giz126

Carter KA, Liston A, Bassil NV, Alice LA, Bushakra JM, Sutherland BL, Mockler TC, Bryant DW, Hummer KE (2019) Target capture sequencing unravels *Rubus* evolution. Front. Plant Sci. 10, 1615. https://doi.org/10.3389/fpls.2019.01615

Chamala S, García N, Godden GT, Krishnakumar V, Jordon-Thaden IE, De Smet R, Barbazuk WB, Soltis DE, Soltis PS (2015) MarkerMiner 1.0: a new application for phylogenetic marker development using angiosperm transcriptomes. Appl. Plant Sci. 3, 1400115. https://doi.org/10.3732/apps.1400115

Chen N (2004) Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Protoc. Bioinformatics 5, 4.10.1-4.10.14. https://doi.org/10.1002/0471250953.bi0410s05

Christe C, Boluda CG, Koubínová D, Gautier L, Naciri Y (2021) New genetic markers for Sapotaceae phylogenomics: more than 600 nuclear genes applicable from family to population levels. Mol. Phylogenet. Evol. 160, 107123. https://doi.org/10.1016/j.ympev.2021.107123

Costa L, Marques A, Buddenhagen C, Thomas WW, Huettel B, Schubert V, Dodsworth S, Houben A, Souza G, Pedrosa-Harand A (2021) Aiming off the target: recycling target capture sequencing reads for investigating repetitive DNA. Ann. Bot. 128, 835–848. https://doi.org/10.1093/aob/mcab063

Couvreur TLP, Helmstetter AJ, Koenen EJM, Bethune K, Brandão RD, Little SA, Sauquet H, Erkens RHJ (2018) Phylogenomics of the major tropical plant family annonaceae using targeted enrichment of nuclear genes. Front. Plant Sci. 9, 1941. https://doi.org/10.3389/fpls.2018.01941

Cruz-Dávalos DI, Llamas B, Gaunitz C, Fages A, Gamba C, Soubrier J, Librado P, Seguin-Orlando A, Pruvost M, Alfarhan AH, Alquraishi SA, Al-Rasheid KAS, Scheu A, Beneke N, Ludwig A, Cooper A, Willerslev E, Orlando L (2017) Experimental conditions improving in-solution target enrichment for ancient DNA. Mol. Ecol. Resour. 17, 508–522. https://doi.org/10.1111/1755-0998.12595

de La Harpe M, Hess J, Loiseau O, Salamin N, Lexer C, Paris M (2019) A dedicated target capture approach reveals variable genetic markers across micro- and macro-evolutionary time scales in palms. Mol. Ecol. Resour. 19, 221–234. https://doi.org/10.1111/1755-0998.12945

De Smet R, Adams KL, Vandepoele K, Van Montagu MCE, Maere S, Van de Peer Y (2013) Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. Proc Natl Acad Sci USA 110, 2898–2903. https://doi.org/10.1073/pnas.1300127110

Dodsworth S, Pokorny L, Johnson MG, Kim JT, Maurin O, Wickett NJ, Forest F, Baker WJ (2019) Hyb-Seq for flowering plant systematics. Trends Plant Sci. 24, 887–891. https://doi.org/10.1016/j.tplants.2019.07.011

Emms DM, Kelly S (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 20, 238. https://doi.org/10.1186/s13059-019-1832-y

Eriksson JS, Bacon CD, Bennett DJ, Pfeil BE, Oxelman B, Antonelli A (2021) Gene count from target sequence capture places three whole genome duplication events in *Hibiscus* L. (Malvaceae). BMC Ecol. Evo. 21, 107. https://doi.org/10.1186/s12862-021-01751-7

Eserman LA, Thomas SK, Coffey EED, Leebens-Mack JH (2021) Target sequence capture in orchids: developing a kit to sequence hundreds of single-copy loci. Appl. Plant Sci. 9, e11416. https://doi.org/10.1002/aps3.11416

Fér T, Schmickl RE (2018) Hybphylomaker: target enrichment data analysis from raw reads to species trees. Evol. Bioinform. Online 14, 1176934317742613. https://doi.org/10.1177/1176934317742613

Folk RA, Mandel JR, Freudenstein JV (2015) A protocol for targeted enrichment of intron-containing sequence markers for recent radiations: a phylogenomic example from *Heuchera* (Saxifragaceae). Appl. Plant Sci. 3, 1500039. https://doi.org/10.3732/apps.1500039

Folk RA, Stubbs RL, Mort ME, Cellinese N, Allen JM, Soltis PS, Soltis DE, Guralnick RP (2019) Rates of niche and phenotype evolution lag behind diversification in a temperate radiation. Proc Natl Acad Sci USA 116, 10874–10882. https://doi.org/10.1073/pnas.1817999116

Forrest LL, Hart ML, Hughes M, Wilson HP, Chung K-F, Tseng Y-H, Kidner CA (2019) The limits of Hyb-Seq for herbarium specimens: impact of preservation techniques. Front. Ecol. Evol. 7. https://doi.org/10.3389/fevo.2019.00439

Frost L, Santamaría-Aguilar DA, Singletary D, Lagomarsino LP (2020) Herbarium-based phylogenomics reveals that the Andes are a biogeographic barrier for *Otoba* (Myristicaceae), an ecologically dominant Neotropical tree genus. BioRxiv. https://doi.org/10.1101/2020.10.02.324368

Gabaldón T, Koonin EV (2013) Functional and evolutionary implications of gene orthology. Nat. Rev. Genet. 14, 360–366. https://doi.org/10.1038/nrg3456

Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat. Biotechnol. 27, 182–189. https://doi.org/10.1038/nbt.1523

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS (2012) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 40, D1178–D1186. https://doi.org/10.1093/nar/gkr944

Gostel MR, Coy KA, Weeks A (2015) Microfluidic PCR-based target enrichment: A case study in two rapid radiations of *Commiphora* (Burseraceae) from Madagascar. J. Syst. Evol. 53, 411–431. https://doi.org/10.1111/jse.12173

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29, 644–652. https://doi.org/10.1038/nbt.1883

Grover CE, Salmon A, Wendel JF (2012) Targeted sequence capture as a powerful tool for evolutionary analysis. Am. J. Bot. 99, 312–319. https://doi.org/10.3732/ajb.1100323

Hale H, Gardner EM, Viruel J, Pokorny L, Johnson MG (2020) Strategies for reducing per-sample costs in target capture sequencing for phylogenomics and population genomics in plants. Appl. Plant Sci. 8, e11337. https://doi.org/10.1002/aps3.11337

Heyduk K, McKain MR, Lalani F, Leebens-Mack J (2016) Evolution of a CAM anatomy predates the origins of Crassulacean acid metabolism in the Agavoideae (Asparagaceae). Mol. Phylogenet. Evol. 105, 102–113. https://doi.org/10.1016/j.ympev.2016.08.018

Hollingsworth PM, Graham SW, Little DP (2011) Choosing and using a plant DNA barcode. PLoS ONE 6, e19254. https://doi.org/10.1371/journal.pone.0019254

Hollingsworth PM (2011) Refining the DNA barcode for land plants. Proc Natl Acad Sci USA 108, 19451–19452. https://doi.org/10.1073/pnas.1116812108

Howard C, Lockie-Williams C, Slater A (2020) Applied barcoding: the practicalities of DNA testing for herbals. Plants 9, 1150. https://doi.org/10.3390/plants9091150

Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23, 254–267. https://doi.org/10.1093/molbev/msj030

Jantzen JR, Amarasinghe P, Folk RA, Reginato M, Michelangeli FA, Soltis DE, Cellinese N, Soltis PS (2020) A two-tier bioinformatic pipeline to develop probes for target capture of nuclear loci with applications in Melastomataceae. Appl. Plant Sci. 8, e11345. https://doi.org/10.1002/aps3.11345

Johnson MG, Gardner EM, Liu Y, Medina R, Goffinet B, Shaw AJ, Zerega NJC, Wickett NJ (2016) HybPiper: extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. Appl. Plant Sci. 4. https://doi.org/10.3732/apps.1600016

Johnson MG, Pokorny L, Dodsworth S, Botigué LR, Cowan RS, Devault A, Eiserhardt WL, Epitawalage N, Forest F, Kim JT, Leebens-Mack JH, Leitch IJ, Maurin O, Soltis DE, Soltis PS, Wong GK-S, Baker WJ, Wickett NJ (2019) A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. Syst. Biol. 68, 594–606. https://doi.org/10.1093/sysbio/syy086

Kleinkopf JA, Roberts WR, Wagner WL, Roalson EH (2019) Diversification of Hawaiian *Cyrtandra* (Gesneriaceae) under the influence of incomplete lineage sorting and hybridization. J. Syst. Evol. 57, 561–578. https://doi.org/10.1111/jse.12519

Koenen EJM, Kidner C, de Souza ÉR, Simon MF, Iganci JR, Nicholls JA, Brown GK, de Queiroz LP, Luckow M, Lewis GP, Pennington RT, Hughes CE (2020) Hybrid capture of 964 nuclear genes resolves evolutionary relationships in the mimosoid legumes and reveals the polytomous origins of a large pantropical radiation. Am. J. Bot. 107, 1710–1735. https://doi.org/10.1002/ajb2.1568

Kozarewa I, Armisen J, Gardner AF, Slatko BE, Hendrickson CL (2015) Overview of target enrichment strategies. Curr. Protoc. Mol. Biol. 112, 7.21.1-7.21.23. https://doi.org/10.1002/0471142727.mb0721s112

Landis JB, Soltis DE, Li Z, Marx HE, Barker MS, Tank DC, Soltis PS (2018) Impact of whole-genome duplication events on diversification rates in angiosperms. Am. J. Bot. 105, 348–363. https://doi.org/10.1002/ajb2.1060

Lang PLM, Weiß CL, Kersten S, Latorre SM, Nagel S, Nickel B, Meyer M, Burbano HA (2020) Hybridization ddRAD-sequencing for population genomics of nonmodel plants using highly degraded historical specimen DNA. Mol. Ecol. Resour. 20, 1228–1247. https://doi.org/10.1111/1755-0998.13168

Larridon I, Villaverde T, Zuntini AR, Pokorny L, Brewer GE, Epitawalage N, Fairlie I, Hahn M, Kim J, Maguilla E, Maurin O, Xanthos M, Hipp AL, Forest F, Baker WJ (2019) Tackling rapid radiations with targeted sequencing. Front. Plant Sci. 10, 1655. https://doi.org/10.3389/fpls.2019.01655

Lefoulon E, Vaisman N, Frydman HM, Sun L, Voland L, Foster JM, Slatko BE (2019) Large enriched fragment targeted sequencing (LEFT-SEQ) applied to capture of wolbachia genomes. Sci. Rep. 9, 5939. https://doi.org/10.1038/s41598-019-42454-w

Lesur I, Alexandre H, Boury C, Chancerel E, Plomion C, Kremer A (2018) Development of target sequence capture and estimation of genomic relatedness in a mixed oak stand. Front. Plant Sci. 9, 996. https://doi.org/10.3389/fpls.2018.00996

Liu J, Jiang J, Song S, Tornabene L, Chabarria R, Naylor GJP, Li C (2017) Multilocus DNA barcoding - species identification with multilocus data. Sci. Rep. 7, 16601. https://doi.org/10.1038/s41598-017-16920-2

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li H-T, Yi T-S, Gao L-M, Ma P-F, Zhang T, Yang J-B, Gitzendanner MA, Fritsch PW, Cai J, Luo Y, Wang H, van der Bank M, Zhang S-D, Wang Q-F, Wang J, Zhang Z-R, Fu C-N, Yang J, Hollingsworth PM, Chase MW, Li D-Z (2019) Origin of angiosperms and the puzzle of the Jurassic gap. Nat. Plants 5, 461–470. https://doi.org/10.1038/s41477-019-0421-0

Li X, Yang Y, Henry RJ, Rossetto M, Wang Y, Chen S (2015) Plant DNA barcoding: from gene to genome. Biol. Rev. Camb. Philos. Soc. 90, 157–166. https://doi.org/10.1111/brv.12104

Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ (2010) Target-enrichment strategies for next-generation sequencing. Nat. Methods 7, 111–118. https://doi.org/10.1038/nmeth.1419

Mandel JR, Dikow RB, Funk VA, Masalia RR, Staton SE, Kozik A, Michelmore RW, Rieseberg LH, Burke JM (2014) A target enrichment method for gathering phylogenetic information from hundreds of loci: an example from the Compositae. Appl. Plant Sci. 2, 1300085. https://doi.org/10.3732/apps.1300085

Manzanilla V, Teixidor-Toneu I, Martin GJ, Hollingsworth PM, de Boer HJ, Kool A (2022) Using target capture to address conservation challenges: population-level tracking of a globally-traded herbal medicine. Mol. Ecol. Resour. 22, 212–224. https://doi.org/10.1111/1755-0998.13472

Mayland-Quellhorst E, Meudt HM, Albach DC (2016) Transcriptomic resources and marker validation for diploid and polyploid *Veronica* (Plantaginaceae) from New Zealand and Europe. Appl. Plant Sci. 4, 1600091. https://doi.org/10.3732/apps.1600091

Meuzelaar LS, Lancaster O, Pasche JP, Kopal G, Brookes AJ (2007) MegaPlex PCR: a strategy for multiplex amplification. Nat. Methods 4, 835–837. https://doi.org/10.1038/nmeth1091

Morales-Briones DF, Tank DC (2019) Extensive allopolyploidy in the neotropical genus *Lachemilla* (Rosaceae) revealed by PCR-based target enrichment of the nuclear ribosomal DNA cistron and plastid phylogenomics. Am. J. Bot. 106, 415–437. https://doi.org/10.1002/ajb2.1253

Murphy B, Forest F, Barraclough T, Rosindell J, Bellot S, Cowan R, Golos M, Jebb M, Cheek M (2020) A phylogenomic analysis of *Nepenthes* (Nepenthaceae). Mol. Phylogenet. Evol. 144, 106668. https://doi.org/10.1016/j.ympev.2019.106668

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32, 268–274. https://doi.org/10.1093/molbev/msu300

Nicholls JA, Pennington RT, Koenen EJM, Hughes CE, Hearn J, Bunnefeld L, Dexter KG, Stone GN, Kidner CA (2015) Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the neotropical rain forest genus *Inga* (Leguminosae: Mimosoideae). Front. Plant Sci. 6, 710. https://doi.org/10.3389/fpls.2015.00710

Novak J, Grausgruber-Gröger S, Lukas B (2007) DNA-based authentication of plant extracts. Food Res. Int 40, 388–392. https://doi.org/10.1016/j.foodres.2006.10.015

Novák P, Guignard MS, Neumann P, Kelly LJ, Mlinarec J, Koblížková A, Dodsworth S, Kovařík A, Pellicer J, Wang W, Macas J, Leitch IJ, Leitch AR (2020) Repeat-sequence turnover shifts fundamentally in species with large genomes. Nat. Plants 6, 1325–1329. https://doi.org/10.1038/s41477-020-00785-x

Ogutcen E, Christe C, Nishii K, Salamin N, Möller M, Perret M (2021) Phylogenomics of Gesneriaceae using targeted capture of nuclear genes. Mol. Phylogenet. Evol. 157, 107068. https://doi.org/10.1016/j.ympev.2021.107068

Ogutcen E, Ramsay L, von Wettberg EB, Bett KE (2018) Capturing variation in *Lens* (Fabaceae): development and utility of an exome capture array for lentil. Appl. Plant Sci. 6, e01165. https://doi.org/10.1002/aps3.1165

Ojeda DI, Koenen E, Cervantes S, de la Estrella M, Banguera-Hinestroza E, Janssens SB, Migliore J, Demenou BB, Bruneau A, Forest F, Hardy OJ (2019) Phylogenomic analyses reveal an exceptionally high number of evolutionary shifts in a florally diverse clade of African legumes. Mol. Phylogenet. Evol. 137, 156–167. https://doi.org/10.1016/j.ympev.2019.05.002

Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, Thibaud-Nissen F, Malek RL, Lee Y, Zheng L, Orvis J, Haas B, Wortman J, Buell CR (2007) The TIGR rice genome annotation resource: improvements and new features. Nucleic Acids Res. 35, D883–D887. https://doi.org/10.1093/nar/gkl976

Pellicer J, Hidalgo O, Dodsworth S, Leitch IJ (2018) Genome size diversity and its impact on the evolution of land plants. Genes (Basel) 9, 88. https://doi.org/10.3390/genes9020088

Pel J, Leung A, Choi WWY, Despotovic M, Ung WL, Shibahara G, Gelinas L, Marziali A (2018) Rapid and highly-specific generation of targeted DNA sequencing libraries enabled by linking capture probes with universal primers. PLoS ONE 13, e0208283. https://doi.org/10.1371/journal.pone.0208283

Popp M, Erixon P, Eggens F, Oxelman B (2005) Origin and evolution of a circumpolar polyploid species complex in *Silene* (Caryophyllaceae) inferred from low copy nuclear RNA polymerase introns, rDNA, and chloroplast DNA. Syst. Bot 30, 302–313. https://doi.org/10.1600/0363644054223648

Rauscher JT, Doyle JJ, Brown AHD (2002) Internal transcribed spacer repeat-specific primers and the analysis of hybridization in the *Glycine tomentella* (Leguminosae) polyploid complex. Mol. Ecol. 11, 2691–2702. https://doi.org/10.1046/j.1365-294x.2002.01640.x

Samuels DC, Han L, Li J, Quanghu S, Clark TA, Shyr Y, Guo Y (2013) Finding the lost treasures in exome sequencing data. Trends Genet. 29, 593–599. https://doi.org/10.1016/j.tig.2013.07.006

Sanderson BJ, DiFazio SP, Cronk QCB, Ma T, Olson MS (2020) A targeted sequence capture array for phylogenetics and population genomics in the Salicaceae. Appl. Plant Sci. 8, e11394. https://doi.org/10.1002/aps3.11394

Sang T (2002) Utility of low-copy nuclear gene sequences in plant phylogenetics. Crit. Rev. Biochem. Mol. Biol. 37, 121–147. https://doi.org/10.1080/10409230290771474

Schneider JV, Jungcurt T, Cardoso D, Amorim AM, Töpel M, Andermann T, Poncy O, Berberich T, Zizka G (2021) Phylogenomics of the tropical plant family Ochnaceae using targeted enrichment of nuclear genes and 250+ taxa. Taxon 70, 48–71. https://doi.org/10.1002/tax.12421

Shah T, Schneider JV, Zizka G, Maurin O, Baker W, Forest F, Brewer GE, Savolainen V, Darbyshire I, Larridon I (2021) Joining forces in Ochnaceae phylogenomics: a tale of two targeted sequencing probe kits. Am. J. Bot. 108, 1201–1216. https://doi.org/10.1002/ajb2.1682

Shneyer VS, Rodionov AV (2019) Plant DNA barcodes. Biol. Bull. Rev. 9, 295–300. https://doi.org/10.1134/S207908641904008X

Slater GSC, Birney E (2005) Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6, 31. https://doi.org/10.1186/1471-2105-6-31

Soto Gomez M, Pokorny L, Kantar MB, Forest F, Leitch IJ, Gravendeel B, Wilkin P, Graham SW, Viruel J (2019) A customized nuclear target enrichment approach for developing a phylogenomic baseline for *Dioscorea* yams (Dioscoreaceae). Appl. Plant Sci. 7, e11254. https://doi.org/10.1002/aps3.11254

Staats M, Cuenca A, Richardson JE, Vrielink-van Ginkel R, Petersen G, Seberg O, Bakker FT (2011) DNA damage in plant herbarium tissue. PLoS ONE 6, e28448. https://doi.org/10.1371/journal.pone.0028448

Stracke R, Werber M, Weisshaar B (2001) The R2R3-MYB gene family in *Arabidopsis thaliana*. Curr. Opin. Plant Biol. 4, 447–456. https://doi.org/10.1016/S1369-5266(00)00199-0

Straub SCK, Boutte J, Fishbein M, Livshultz T (2020) Enabling evolutionary studies at multiple scales in Apocynaceae through Hyb-Seq. Appl. Plant Sci. 8, e11400. https://doi.org/10.1002/aps3.11400

Strickler SR, Bombarely A, Mueller LA (2012) Designing a transcriptome next-generation sequencing project for a nonmodel plant species. Am. J. Bot. 99, 257–266. https://doi.org/10.3732/ajb.1100292

Tewhey R, Warner JB, Nakano M, Libby B, Medkova M, David PH, Kotsopoulos SK, Samuels ML, Hutchison JB, Larson JW, Topol EJ, Weiner MP, Harismendy O, Olson J, Link DR, Frazer KA (2009) Microdroplet-based PCR enrichment for large-scale targeted sequencing. Nat. Biotechnol. 27, 1025–1031. https://doi.org/10.1038/nbt.1583

Uribe-Convers S, Settles ML, Tank DC (2016) A phylogenomic approach based on PCR target enrichment and high throughput sequencing: resolving the diversity within the South American species of *Bartsia* L. (Orobanchaceae). PLoS ONE 11, e0148203. https://doi.org/10.1371/journal.pone.0148203

Van Verk MC, Hickman R, Pieterse CMJ, Van Wees SCM (2013) RNA-Seq: revelation of the messengers. Trends Plant Sci. 18, 175–179. https://doi.org/10.1016/j.tplants.2013.02.001

Vatanparast M, Powell A, Doyle JJ, Egan AN (2018) Targeting legume loci: a comparison of three methods for target enrichment bait design in Leguminosae phylogenomics. Appl. Plant Sci. 6, e1036. https://doi.org/10.1002/aps3.1036

Villaverde T, Pokorny L, Olsson S, Rincón-Barrado M, Johnson MG, Gardner EM, Wickett NJ, Molero J, Riina R, Sanmartín I (2018) Bridging the micro- and macroevolutionary levels in phylogenomics: Hyb-Seq solves relationships from populations to species and above. New Phytol. 220, 636–650. https://doi.org/10.1111/nph.15312

Viruel J, Conejero M, Hidalgo O, Pokorny L, Powell RF, Forest F, Kantar MB, Soto Gomez M, Graham SW, Gravendeel B, Wilkin P, Leitch IJ (2019) A target capture-based method to estimate ploidy from herbarium specimens. Front. Plant Sci. 10, 937. https://doi.org/10.3389/fpls.2019.00937

Weitemier K, Straub SCK, Cronn RC, Fishbein M, Schmickl R, McDonnell A, Liston A (2014) Hyb-Seq: combining target enrichment and genome skimming for plant phylogenomics. Appl. Plant Sci. 2, 1400042. https://doi.org/10.3732/apps.1400042

Wendel JF, Schnabel A, Seelanan T (1995) Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (*Gossypium*). Proc Natl Acad Sci USA 92, 280–284. https://doi.org/10.1073/pnas.92.1.280

Wolf PG, Robison TA, Johnson MG, Sundue MA, Testo WL, Rothfels CJ (2018) Target sequence capture of nuclear-encoded genes for phylogenetic analysis in ferns. Appl. Plant Sci. 6, e01148. https://doi.org/10.1002/aps3.1148

Woudstra Y, Rees P , Rakotoarisoa S, Rønsted N, Howard C, Grace OM (2021) Improved molecular identification of Aloe vera and relatives using low-copy nuclear genes, in: Molecular identification of aloes: the detection, capture and application of low-copy nuclear genes in the giant genomes of Aloe vera and relatives. PhD thesis by Yannick Woudstra, University of Copenhagen.

Woudstra Y, Viruel J, Fritzsche M, Bleazard T, Mate R, Howard C, Rønsted N, Grace OM (2021) A customised target capture sequencing tool for molecular identification of *Aloe vera* and relatives. Sci. Rep. 11, 24347. https://doi.org/10.1038/s41598-021-03300-0

Yang Z, Rannala B (2017) Bayesian species identification under the multispecies coalescent provides significant improvements to DNA barcoding analyses. Mol. Ecol. 26, 3028–3036. https://doi.org/10.1111/mec.14093

Yardeni G, Viruel J, Paris M, Hess J, Groot Crego C, de La Harpe M, Rivera N, Barfuss MHJ, Till W, Guzmán-Jacob V, Krömer T, Lexer C, Paun O, Leroy T (2022) Taxon-specific or universal? Using target capture to study the evolutionary history of rapid radiations. Mol. Ecol. Resour. 22, 927–945. https://doi.org/10.1111/1755-0998.13523

Yoo M-J, Lee B-Y, Kim S, Lim CE (2021) Phylogenomics with Hyb-Seq unravels Korean hosta evolution. Front. Plant Sci. 12, 645735. https://doi.org/10.3389/fpls.2021.645735

Zhang C, Rabiee M, Sayyari E, Mirarab S (2018) ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics 19, 153. https://doi.org/10.1186/s12859-018-2129-y

Answers

1. Universal bait panels are designed for broad taxonomic application, such as all angiosperms. They are cheaper than customised bait panels due to the high consumer demand, but are usually less powerful in phylogenetics of recently diversified clades due to the use of conserved nuclear loci. Customised bait panels can require some investment, especially when no reference transcriptome is available, but offer a high return-on-investment by allowing for the selection of highly informative loci for the clade of interest.
2. Due to the small size of baits used in target enrichment (80–120 bp), the technique is less dependent on DNA fragmentation in (older) historical specimens. Even fragments partially containing a target sequence can be captured by the baits. Compared to traditional Sanger sequencing or PCR-based target enrichment, target capture does not depend on having complete genic fragments present in the DNA samples.
3. Short RNA- (or DNA-)baits containing sequences complementary to target sequences can hybridise in-solution with DNA fragments in any DNA sample. They can effectively hybridise with complementary DNA strands with up to 30% nucleotide mismatches allowing the same bait panel to capture target sequences in even distantly related taxa compared to the original clade used in the design. The baits are chemically modified (biotinylated) so they can be bound to magnetic streptavidin beads with target DNA fragments attached. This allows the user to precipitate the target DNA on a magnetic tube rack to wash away all unwanted non-target DNA fragments.
4. Target capture sequencing is a reduced representation HTS technique that effectively reduces the complexity of a HTS library by increasing the proportion of target DNA fragments in the sample. Compared to genome skimming and other low-coverage HTS techniques, it is much more powerful in obtaining low-copy nuclear loci, which are highly popular in plant phylogenomics. It is, however, much more expensive than genome skimming. On the other hand, it is much more cost-effective than whole genome sequencing if the user knows which genes the study aims to obtain.

Chapter 15
Transcriptomics

Dewi Pramanik[1,2], Ozan Çiftçi[3], Yannick Woudstra[4,5,6,7]

1 Evolutionary Ecology Group, Naturalis Biodiversity Center, Leiden, The Netherlands
2 National Research and Innovation Agency Republic of Indonesia (BRIN), Indonesia
3 Institute of Environmental Sciences, Leiden University, Leiden, The Netherlands
4 Royal Botanic Gardens, Kew, United Kingdom
5 Natural History Museum Denmark, University of Copenhagen, Copenhagen, Denmark
6 Gothenburg Global Biodiversity Center, Department of Biological and Environmental Sciences, University of Gothenburg, Gothenburg, Sweden
7 Department of Plant Sciences, University of Oxford, Oxford, United Kingdom

Dewi Pramanik **dewi.pramanik@naturalis.nl**, **dewi053@brin.go.id**
Ozan Çiftçi **ozancift@gmail.com**
Yannick Woudstra **yannickwoudstra@outlook.com**

Citation: Pramanik D, Çiftçi O, Woudstra Y (2022) Chapter 15. Transcriptomics. In: de Boer H, Rydmark MO, Verstraete B, Gravendeel B (Eds) Molecular identification of plants: from sequence to species. Advanced Books. https://doi.org/10.3897/ab.e98875

Background

Transcriptomics is the study of the transcriptome, which is the complete set of all RNA molecules, including coding and noncoding RNA, that is expressed in a cell, tissue, or organism at a specific spatial, temporal, or developmental stage (Morozova et al. 2009; Piétu et al. 1999). Transcriptomics presents a convenient and cost-effective hybrid approach to study both genomes and biological function at the same time. It is especially useful for making direct links between the genotype and the phenotype as it allows for the accurate detection of gene expression levels in different tissues (Hrdlickova et al. 2017) under different environmental circumstances and even at different spatial scales (Burgess 2019). Sequencing of only the exons of expressed coding genes (Van Verk et al. 2013) represents a simpler alternative to whole genome sequencing that makes the assembly of nuclear coding genes more attainable.

In plant research, transcriptomics is widely used for studying differential expression, identifying novel genes, and general expression patterns (Shakya et al. 2019). It is also widely used to study genetic diversity in environmental samples ranging from the human (and other animal) microbiomes, to microbes found in or on plants, within soil and in aquatic environments which is referred to as metatranscriptomics (Poretsky et al. 2005; Shakya et al. 2019). For instance, metatranscriptomics can be used to understand how plant-microbiome interactions evolve through time and under different environmental conditions (Shakya et al. 2019).

The first publication studying individual transcripts used Northern blotting for RNA detection, which is a hybridization-based method (Alwine et al. 1977) Further developments included gene expression quantification by sequence/sequencing-based methods including the expressed sequence tag (EST) (Adams et al. 1991), serial analysis of gene expression (SAGE) (Velculescu et al. 1997), massively parallel signature sequencing (MPSS) (Brenner et al. 2000), and cap analysis of gene expression (Shiraki et al. 2003).

Microarrays were the first high-throughput method developed for transcriptomics to achieve widespread use due to their affordability and highly sensitive transcript detection (Wang et al. 2019). However, with the introduction of second generation RNA sequencing platforms (RNA-seq), microarrays are no longer widely used (McGettigan 2013). RNA-seq offers many advantages over microarrays in plant studies including higher genomic coverage (Kukurba and Montgomery 2015), better resolution of expression differences amongst paralogs (Sundell et al. 2017) and higher precision in co-expression networks for estimating the exact expression levels for lowly expressed genes (Fu et al. 2014). RNA-seq data can accurately quantify expression levels, is inexpensive to acquire, and does not require highly skilled labour (Wang et al. 2009). The first application of RNA-seq in plant science was with the *Arabidopsis* transcriptome (Weber et al. 2007). Non-model plants can also be sequenced with RNA-seq since it does not require existing genomic data (Wang et al. 2009).

Currently, RNA-seq data is often acquired using technologies that allow for long read data. Long read RNA-seq data allows reading full transcripts, finding new isoforms, identifying fusion transcripts, identifying long noncoding RNA, simplifying the computational analysis, and reducing PCR biases (Depledge et al. 2019). The two technologies that currently dominate long-read sequencing are Pacific Biosciences (PacBio) single-molecule real-time (SMRT) isoform sequencing (Iso-Seq) and Oxford Nanopore Technologies (ONT) (Amarasinghe et al. 2020). Nevertheless, this technology is not without drawbacks, including high experimental cost, low throughput, and higher read error rate (Stark et al. 2019). This chapter thus emphasises short-read RNA-seq, though protocols for long-read RNA seq are included. (Stark et al. 2019).

Experimental design

RNA isolation

Experimental design considerations

Isolating a sufficient quantity of high-quality RNA is critical for conducting transcriptome sequencing experiments and their analyses. When designing a protocol, a number of biological replicates should be considered. Biological replication represents RNA harvested from different plants or different sets of independent samples treated under the same conditions. This biological replication is important for assessing variation between samples, and more biological replicates can increase statistical power during analysis. In general, the minimum number of samples for transcriptomics studies is three biological replicates. Once the minimum number of samples and replications is achieved, the following steps are sample treatment and handling, RNA isolation, and RNA quality and quantity testing.

Tissue preparation and homogenization

RNA to be used in transcriptomic experiments is most commonly isolated from a maximum of 100 mg of fresh plant tissue. If not used immediately, harvested plant tissue should be snap-frozen in liquid nitrogen and stored at -80 °C. If it is not possible to homogenise the fresh material or to snap-freeze it in liquid nitrogen immediately (e.g., in the field), it should be kept in a preservation buffer that maintains a constant pH to preserve proteins and protect the RNA. RNA stabilisation and storage solutions available from manufacturers (e.g., Ambion, Applied Biosystem or RNAlater™, Invitrogen, ThermoFisher Scientific, USA) or other preservatives such as a sulfate salt solution (e.g., ammonium sulfate) preserve tissue samples after harvesting in order to retain the quality and quantity of RNA for long periods (Allewell and Sama 1974). Samples stored in a stabilisation solution can be disrupted and homogenised without the use of liquid nitrogen. A similar treatment can also be carried out for soil samples (metatranscriptomics) with an additional sample screening step by sieving the soil sample to separate the sample from organic debris, roots and rocks (mesh 2 mm) before storing at -80 °C or in an RNA preservative solution (e.g., LifeGuard™ Soil Conservation Solution, MO BIO Laboratories Inc., USA) (Carvalhais and Schenk 2013; Carvalhais et al. 2012, 2013).

A crucial step in tissue preparation is finding the most appropriate method to homogenise the tissue in order to maximise the yield and quality of the RNA. The most common method to homogenise the tissue is snap-freezing in liquid nitrogen and subsequent homogenization/disruption of the tissue by manually grinding with a mortar and pestle or with glass/metal beads and a tissue lyser. However, this is challenging for hard tissue like wood, roots or plant tissues with thick cuticles such as succulent leaves. The combination of snap-freezing in liquid nitrogen,disruption of the tissue by manually grinding, and second grinding with glass/metal beads and a tissue lyser can be a solution to optimise the tissue homogenization of hard tissues. Once the tissue samples are powdered, they can be stored at -80 °C or used immediately for RNA isolation. It is advised to thaw a frozen tissue sample only once and add the lysis buffer immediately to obtain high-quality isolated RNA. It is important that the lytic agent or denaturant comes into contact with the cellular contents when the cells are disrupted. The RNA lysis buffer (e.g., Buffer RLT, Qiagen-USA) is usually composed of phenol and guanidine isothiocyanate. This buffer has two functions as a denaturing agent and stabilises nucleic acid by preventing the activity of the RNase enzyme.

RNA isolation

Compared to DNA, RNA is less stable due to its chemical structure: RNA is single-stranded and can easily be enzymatically degraded by the abundant amounts of ribonuclease (RNAse) that are present in the environment. RNases are secreted through our skin and in the air we breathe out. RNA isolations therefore need to be conducted in RNase-free conditions. Gloves must be worn at all times and the RNA isolation should take place in a fume hood. Designated working spaces and equipment should be cleaned with RNase inhibitors. Common RNase inhibitors to use are strong denaturants such as guanidinium, sodium dodecyl sulfate (SDS), diethyl pyrocarbonate (DEPC), or phenol-based compounds. Additionally, commercially available products include DNase/RNase AWAY™ (Merck BV, The Netherlands) or bleach (sodium hypochlorite). Keep in mind to also use RNase-free plastics and glassware. The main steps for RNA isolation are similar to the DNA isolation protocol (Chapter 1 DNA from Plant Tissues). RNA can be extracted and purified by following protocols described in the literature such as an acidic phenol-chloroform RNA extraction (Chomczynski and Sacchi 2006) or by using commercial kits (e.g., Qiagen Plant RNA kit, Turbo DNA-free kit, etc.). Commercial extraction kits come with the advantage that there is reduced handling of hazardous reagents and less time needed to prepare the reagents. For woody tissues or other tissue types with high phenolic compounds, lysis buffers with high molecular weight polymers (e.g., polyvinylpyrrolidone (PVP)) that can bind and remove polyphenols and polysaccharides may be required (Maceda-López et al. 2021). The purity of the extracted RNA may be improved by digesting genomic DNA (for example with the RNase-Free DNase Set, Qiagen, USA) and by RNA precipitation to enrich RNA over DNA (for example by using ethanol-glycogen or LiCl precipitation). This can be especially important for obtaining pure RNA from plant samples with high amounts of alkaloids (Leh et al. 2019).

Tissue-specific RNA and single-cell RNA isolation

Single-cell RNA-seq (scRNA-seq) is an advanced method to profile transcriptomes from individual cells. scRNA-seq can be used for cell type identification, transcriptome profiling, and inference of gene regulatory networks across the cell (Rich-Griffin et al. 2020). Several methods are available for tissue-specific or single-cell RNA isolations.

One method for tissue-specific isolation is laser microdissection (LMD), which is based on a histological identification that isolates specific cell types by laser capture and laser cutting (Kivivirta et al. 2019). An area of at least 10 μm^2 with a section thickness of 10 µm must be captured to obtain a sufficient amount, which is approximately 10 pg to 1 ng of RNA for cDNA synthesis (Kivivirta et al. 2019).

Single-cell sequencing can further provide high-resolution functional information on an individual cell. In order to capture single cells for scRNA-seq experiments, fluorescence-activated cell sorting (FACS) with the use of protoplasts is commonly used. This is both a high-throughput and highly specific method (Efroni and Birnbaum 2016), but it is also labour intensive. Using protoplasts for FACS is challenging as the enzymatic digestion and cleanup process during protoplast isolation results in a stress response that must be accounted for in subsequent data analysis, and generating protoplast cells from plant tissues remains challenging (Long et al. 2021). Recently, an isolated nuclei approach was developed as an alternative to using plant protoplasts. With this approach, it has been shown that it is possible to design single-cell RNA libraries and obtain meaningful transcriptomic information from plant cells (Thibivilliers et al. 2020).

RNA quality and quantity

The quality and quantity evaluation of RNA is essential to the success of sequencing experiments and the downstream analysis.. The RNA quality and quantity can be evaluated by mea-

suring the UV absorption of a sample. The optical density (OD) ratios at A260/A280 and A260/A230 can be used to determine the RNA purity. Pure RNA has an A260/A280 ratio of 2.1, and an A260/A230 ratio in the range of 2.0-2.2 (Wilfinger et al. 1997). A low A260/A230 ratio may suggest contamination from carbohydrate carry over or residual phenol, while a low A260/280 ratio can indicate contamination from residual phenol, guanidine, or reagents associated with the extraction. These contaminants can affect the downstream application and bias the expression results from qPCR (Carvalhais and Schenk 2013) (e.g., uneven gene coverage or 3′–5′ transcript bias) (Kukurba and Montgomery 2015).

Measuring the RNA integrity in order to determine its degradation level is also recommended. Traditionally, RNA integrity was determined by visualising total RNA using gel electrophoresis and ethidium bromide staining. Intact RNA gives sharp and clear 28S and 18S rRNA bands with an intensity ratio of 28S/18S at 2.0 or higher, in addition to a messenger RNA (mRNA) smear that should be visible between these two distinct bands. A more recent and standardised RNA integrity determination method is determining the RNA integrity number (RIN) with Agilent Bioanalyzer Systems instruments (Agilent Technologies, USA) (Schroeder et al. 2006). The RIN is calculated from total RNA sample characteristics that are based on records of electrophoretic trace data including the ratio of 28S/18S rRNA, the height of the 28S and 18S rRNA peak, and the area between the 18S and 5S rRNA peaks. The RIN software algorithm classifies RNA integrity from 1 to 10, with 1 being the most degraded and 10 being the most intact. A RIN of 7 or higher indicates that the RNA is sufficiently intact for RNA-seq (Jahn et al. 2008). If the isolated RNA is of low quality and quantity, additional precipitation steps may be required to improve the purity of RNA. For large tissue samples, the total amount of harvested RNA is usually between 100 ng and 1 µg. However, for tissue specific RNA, the necessary amounts are between 10 pg to 1 ng. For single-cell RNA sequencing, 1000–8000 cells per single-cell suspension are needed (Rich-Griffin et al. 2020) or 300 ng of total RNA for the SMRT PacBio Iso-Seq platform.

Library preparation

The selection of library preparation methods depends on the fragment size, presence of structural features, and sequencing platform. In the Illumina short-read RNA-seq protocol, the library preparation entails four main steps: (1) RNA molecule selection (mRNA enrichment or rRNA depletion), (2) fragmenting the targeted sequence to the desired length and converting fragmented RNA into cDNA, (3) attaching the adapters and PCR amplification to create the cDNA library, and (4) quantifying the library product for sequencing. The library preparation for long-read sequencing is somewhat simpler than for short-read sequencing. The PacBio Iso-Seq protocol consists of three main steps: (1) cDNA synthesis, (2) cDNA amplification, and (3) library construction. With the Oxford Nanopore platform, the sequencing can be done directly from RNA or by using the amplified (or non-amplified) cDNA input.

mRNA enrichment or rRNA depletion

A total RNA sample after extraction contains ribosomal RNA (rRNA), precursor mRNA (pre-mRNA), mRNA, small noncoding RNA (sRNA/sncRNA), and long ncRNA (transcripts longer than 200 nucleotides), where the majority of material is rRNA (Hrdlickova et al. 2017). Total RNA sequencing or whole transcriptome sequencing refers to the sequencing of all RNA molecules, both coding and noncoding. A selection of the mature polyadenylated (poly(A)) mRNA (mRNA with poly(A) tail) can be made in order to sequence the protein-coding regions only. The addition of a poly(A)tail to the mRNA molecule increases the stability of the molecule and allows the

mRNA to be exported from the nucleus and translated into the protein. Since a high percentage of rRNA (> 80%) can interfere with the analysis of mRNA transcripts, an additional step to enrich mRNA or deplete rRNA may be necessary. rRNA depletion is most commonly used to capture unique transcriptome features. In contrast mRNA enrichment is mainly used to increase exonic coverage (Zhao et al. 2018) or for expression profiling studies. mRNA enrichment, also known as poly(A) enrichment, can be done by selecting only polyadenylated mRNA from total RNA. During this procedure, the total RNA is mixed with oligo (dT) primers and a high-salt binding buffer to promote binding to paramagnetic beads. Oligo dT bound to the bead's surface hybridises to the poly(A) containing mRNA. Precipitate the mRNA bound to beads with a magnet, followed by application of a high-salt washing buffer to discard unbound RNA while retaining oligo (dT) bound poly(A) mRNAs (Green and Sambrook 2019). Similarly, rRNA can be depleted by rRNA hybridization to complementary biotinylated oligo probes followed by extraction with streptavidin-coated paramagnetic beads (Kraus et al. 2019). The selection of a mRNA enrichment or rRNA depletion protocol depends on the aim of the study and other factors such as sample quantity and sample type.

cDNA synthesis

The conversion of RNA into cDNA is an essential step for RNA-seq. This conversion is necessary because DNA is biologically more stable than RNA. PCR amplification can only be done with DNA, and most sequencing protocols are designed for sequencing DNA. The first step in converting RNA to cDNA is the fragmentation of the RNA into an appropriate size for sequencing (i.e., 100–600 bp). Several approaches are available for RNA fragmentation, including physical approaches (e.g., acoustic shearing and sonication), chemical approaches (i.e., heating and divalent metal cation addition), and enzymatic methods (i.e., non-specific endonuclease cocktails and transposase tagmentation reactions) (Marine et al. 2011). The fragmented RNA is then converted to single-stranded cDNA using mRNA as the template, reverse transcriptase, and random primers or oligo (dT) primers (depending on the kit). The first strand of cDNA then can be used as a template for PCR. After this, double-stranded cDNA is produced by second-strand synthesis. The second strand cDNA synthesis is catalysed by *E. coli* DNA polymerase I combined with *E. coli* RNase H and *E. coli* DNA ligase. *E. coli* RNase H degrades the RNA to produce 3′ -hydroxyl and 5′ -phosphate terminated products, which are necessary for the DNA polymerase to function. The *E coli* DNA polymerase I has two activities: the 5′-3′ exonuclease activity removes RNA strands in the direction of synthesis, and in the meantime, it replaces RNA with deoxyribonucleotides. *E. coli* DNA ligase then joins the single strands into double-stranded cDNA.

Adapter ligation and cDNA amplification

Adapters are ligated to one or both ends of the cDNA fragment. Adapters consist of sequences that allow library fragments to bind to the flow cell, sequencing primer binding sites, and index sequences. Index/barcode sequences are sequence identifiers that enable the pooling of several samples (multiplexing) in a single sequencing run or flow cell lane. Products from the ligation reaction are purified using agarose gel electrophoresis prior to PCR amplification to create the cDNA library.

Library preparation kits

Several library preparation kits based on the Illumina platform are available. The "TruSeq Stranded Total RNA with Ribo-Zero Plant" kit is useful for large tissue samples (0.1–1 µg total

RNA). While for low quantities of RNA, the "NEBNext® Ultra™ II Directional RNA Library Prep with Sample Purification Beads" kit (10 ng–1 µg total RNA for polyA mRNA workflow and 5 ng–1 µg total RNA for rRNA depletion workflow) (New England Biolabs Inc., UK) can be used. These kits incorporate Illumina library preparation steps, including bead-based rRNA depletion or mRNA enrichment, cDNA synthesis, adding adaptors, indexing, and PCR. For a tissue sample that yields smaller amounts of RNA, like a single cell (1–25 ng), the "Collibri stranded RNA Library Prep kit" (ThermoFisher Scientific, USA) can be applied.

For the PacBio Iso-Seq platform for long-read RNA-seq, the "NEBNext Single Cell/Low Input cDNA Synthesis & Amplification Module" kit (New England Biolabs Inc., UK) can be used for cDNA synthesis and its amplification from a single cell or ultra-low input RNA (as low as 1 pg–200 ng). The "SMRTbell Express Template Prep Kit 2.0" (Pacific Bioscience, USA) can be used to detect full-length transcripts up to 10 kb.

The ONT platform provides a starter pack for direct RNA-seq, PCR-cDNA sequencing kit, and direct cDNA sequencing kit (Oxford Nanopore Technologies Ltd., UK) with necessary inputs for RNA or Poly-A$^+$(poly(A) on the present of the polyadenylated 3′-ends) 500 ng for direct RNA-seq, 1 ng for PCR-cDNA sequencing, and 100 ng for direct cDNA sequencing.

Quality control of library preparation

A very sensitive method for checking the quantity of a library preparation is with fluorometric methods (i.e., Qubitt™ Fluorometer, ThermoFisher Scientific, USA) or by qPCR. qPCR library quantification is based on the amplification of cDNA fragments with the adapters. The qPCR machine measures the intensity of fluorescence emitted by the probe at each cycle. In this approach, only templates that have both adapter sequences on either end will be measured and subsequently form clusters in a flow cell. Other methods include the use of electrophoresis-based quantification methods such as fragment analyzer systems that use automated parallel capillary electrophoresis to assess the library size distribution (e.g., Tapestation, Agilent Technologies, USA). A critical aspect in the quality check from the fragment analyzer is the library size distribution in the expected range. The peaks near the lower marker on library electrophoresis show contaminants, including primer and adapter dimers. An additional clean-up of the sample is recommended to increase the quality.

RNA sequencing

cDNA sequencing can be performed on several different platforms (see Chapter 9 Sequencing platforms and data types). Overall, RNA sequencing does not differ from the sequencing of genomic DNA. The sequencer reads cDNA fragments in one of two ways: using a single-end or paired-ends. In single-end reading, the sequencer reads the cDNA from the 3′ or 5′ end of only one strand of the insert. This method can produce large volumes of high-quality data especially for differential gene expression studies where an important factor is determining where the reads in transcripts come from (Stark et al. 2019). In the paired-ends reading, the sequencer reads two ends of a cDNA fragment and then combines the forward and reverse reads as reading pairs. Longer overlapping reads are advantageous for detecting splicing variants (Chhangawala et al. 2015). The sequencing length is between 100–600 bp, which will generate at least 20–100 million reads per sample.

The requirements for sequence coverage and depth varies depending on the scientific questions to be answered, with complex studies perhaps needing greater sequencing depth and coverage. For example, a differential expression study using the Illumina platform requires 10-

30 million reads per sample (Stark et al. 2019), while for a more in-depth transcriptome study or de novo transcriptome assembly, the recommended sequencing depth is 100 million reads per sample (Sims et al. 2014). To study isoforms, identify novel transcripts, and detect fusions, long-read sequencing is the appropriate choice. The ONT platform produces 1 to 12+ million reads, with the highest number of reads resulting from amplified cDNA input (7–12+ million reads). At the same time, the PacBio Iso-Seq platform produces up to 3 million full-length reads. Yet, long-read sequencing has some disadvantages compared to the short-read RNA-seq regarding throughput, number of reads, and error reads. In the end, the key in selecting the sequencing platform depends on your budget, research questions, the experimental strategies, technical aspects, data availability on the target organism, and the availability of bioinformatics pipelines.

Bioinformatics

Prior to the development of high-throughput methods, individual transcriptome studies were performed using hybridization-based methods such as Northern blotting and microarrays (see above) or amplification-based methods including Sanger sequencing and RT-qPCR.

Hybridization-based methods require visual inspection or image processing analyses to interpret the output, while in qPCR, it is the amplification that must be monitored. In qPCR, the expression levels are represented by cycle threshold (Ct) values and further normalisation steps and statistical analyses need to be used for the estimation of relative or absolute abundances. Neither hybridization methods nor qPCR require labour-intensive post-processing.

On the other hand, EST/SAGE/MPSS or RNA-seq methods rely on sequence data and require several post-processing steps such as clustering, assembly, and functional annotation. As RNA-seq allows characterization of whole transcriptomes and currently is the most widely used method, we outline the bioinformatic analysis steps for high-throughput RNA-seq data. Long read sequencing methods such as ONT and SMRT allow full-length characterization of transcripts and can be used to study complex transcriptomes. Although one common concern regarding these technologies is high error rates, their accuracy has dramatically increased recently and the development of long-read specific error correction approaches are providing further improvements (Amarasinghe et al. 2020). These technologies have already been used for e.g. isoform identification (Wang et al. 2017) and long noncoding RNA (lncRNA) discovery (Li et al. 2016). Preprocessing of the reads obtained from these platforms requires specific tools, while common short-read analysis tools can be used for downstream analyses (e.g., differential expression) after reconstruction of transcripts. Most long-read isoform detection tools cluster aligned and error-corrected reads and collapse these into isoforms (Amarasinghe et al. 2020). PacBio provides an open source bioinformatics suite "SMRT Analysis" that includes tools for classification of reads, clustering, polishing, alignment, and visualisation of isoforms (Oikonomopoulos et al. 2020). Several tools can be used to align and visualise polished Iso-Seq reads such as MiniMap2 and IsoSeq-Browser (Hu et al. 2017; Li et al. 2017). Similar to PacBio, ONT reads can be analysed with publicly available software and the same tools can be used for these steps.

Quality control and pre-processing

After obtaining raw RNA-seq data, the quality of the reads should be checked and sequencing errors should be corrected in order to improve the accuracy and efficiency of the assembly process. It is also recommended to mask low complexity regions and repetitive sequences that

might generate hits that are artefacts. DUST and SEG modules of BLAST can be used for this purpose on nucleotide and amino acid sequences, respectively. Bacterial and viral contaminants can be removed by running similarity searches against public databases or using tools such as DeconSeq (Schmieder and Edwards 2011a). rRNA sequences can be removed by mapping the reads to an rRNA database (e.g., SILVA, https://www.arb-silva.de) using tools such as bowtie2 (Langmead and Salzberg 2012). FastQC (Andrew 2010) performs an initial assessment of raw high-throughput sequence reads and reports quality metrics that might be useful to determine issues with library preparation or sequencing protocols. There are several other tools that can perform quality control and/or remove artefacts such as sequencing adapters, including Fastx-toolkit (Gordon and Hannon 2009), Prinseq (Schmieder and Edwards 2011b), and Trimmomatic (Bolger et al. 2014).

Most short-read assemblers first divide reads into subsequences of length k (i.e., k-mers) and generate a graph representing the overlap between them (Compeau et al. 2011; Heydari et al. 2017). Sequencing errors introduce problematic k-mers into this process. Several tools with k-mer-based error correction strategies have been developed to overcome this challenge such as Fiona (Schulz et al. 2014) and BFC (Li 2015). Although these tools were originally developed for genomic data, they can correct RNA-seq reads as well. Rcorrector is another tool developed specifically for correcting Illumina RNA-seq reads (Song and Florea 2015). While k-mer-based error correction methods additionally remove reads that originate from rare transcripts, shallow sequencing depth typically does not give accurate assemblies of those transcripts regardless (Martin and Wang 2011).

Transcriptome assembly

Depending on whether a reference genome/transcriptome is available or not, there are different strategies for transcriptome assembly.

De-novo assembly

De-novo assembly is solely based on RNA-seq data and uses the k-mer composition by subdividing the reads into shorter segments of a given length k. This composition and the overlaps between these k-mers are represented on a de Bruijn graph, which is finally resolved to reconstruct transcripts (Pevzner et al. 2001). In general, de-novo assembly requires higher sequencing depths compared to reference-guided assembly; around 30X coverage is needed to reconstruct full length transcripts de-novo, while the same task can be achieved at 10X coverage with a reference (Martin and Wang 2011).

Commonly used de-novo assemblers include Trans-ABySS (Robertson et al. 2010), Trinity (Grabherr et al. 2011), and Oases (Schulz et al. 2012). Trinity is developed specifically for de-novo transcriptome assembly and is widely used (Kerr et al. 2019). Most of the other available tools are simply extensions of de-novo genome assembly tools. Trinity relies on a single k-mer length, while other assemblers can use multiple values of k-mer length (Hölzer and Marz 2019). The choice of the k-mer length can affect the quality of an assembly drastically. The optimal value depends on several factors such as read length, sequencing depth, error rate, and complexity of the target species transcriptome (Góngora-Castillo and Buell 2013). At shorter lengths the possibility of overlap between k-mers is higher, while at longer lengths there are fewer overlaps and reconstruction of transcripts can be comparatively easier. It is also easier to resolve repetitive regions at longer lengths, while the assembly of transcripts with low expression levels becomes more challenging. It should also be noted that memory requirements increase significantly at

longer k-mer lengths. In a case from allopolyploid plants, for example, a k-mer length of 41 produced the highest number of full length transcripts, and researchers suggested to consider a broad range of k-mer lengths and coverages for avoiding chimeric assemblies of homeologous and paralogous gene copies in polyploid taxa (Gruenheit et al. 2012). Some assemblers, such as Trans-ABySS, post-process an assembly to merge contigs and identify isoforms, while other assemblers, such as Trinity, directly use the de Bruijn graph to assemble each isoform (Martin and Wang 2011). Shannon (Kannan et al. 2016) uses a different approach by analysing read abundance information together with the de Bruijn graph in order to resolve complex isoforms and paralogues. After assembly, the longest isoform can be selected as a single representative transcript to simplify the downstream steps (Góngora-Castillo et al. 2012).

There are also combined de-novo assembly approaches such as EvidentialGene (Gilbert 2016) and Oyster River Protocol (MacManes 2018). These tools aim to improve the completeness and accuracy of the assembly by providing an optimised consensus approach that combines several k-mer lengths and transcripts coming from different assemblers.

Reference-guided assembly

Genome-guided assemblers map RNA-seq data to a reference genome and avoid constructing de Bruijn graphs by merging the reads based on their overlapping regions. The quality of the reference genome is critical here, as a high-quality assembly can provide accurate transcript predictions and expression profiles, while using a fragmented or incomplete assembly as reference might aggravate this process. When mapping RNA-seq reads to a reference genome, introns should be accounted for. Therefore genome-guided assemblers allow splitting the reads during mapping. This is achieved by using a splice aware alignment strategy where the downstream regions of a read can map to a downstream exon on the reference. Such splice aware aligners include TopHat2 (Kim et al. 2013) and STAR (Dobin et al. 2013). After mapping, the reads are merged to reconstructed transcripts and isoforms using genome-guided assemblers such as Cufflinks (Trapnell et al. 2012) and StringTie (Pertea et al. 2015). All these tools create a graph representing splice junctions to merge the mapped reads, but they produce different results depending on their different approaches to transcript reconstruction (Hsieh et al. 2019; Voshall and Moriyama 2018). De-novo assemblers can also be used to reconstruct the transcripts at each locus after mapping with splice aware aligners. Trinity offers a genome-guided assembly option using this approach as well. Another genome-guided assembler, RefShannon (Mao et al. 2020), exploits abundance data to reconstruct transcripts with a similar approach to de-novo assembler Shannon.

RNA-seq reads can also be mapped to a transcriptome, if a high-quality assembly is available for the target or a closely related species. This transcriptome-guided approach can improve the contiguity and completeness of the assembly (Garber et al. 2011; Ungaro et al. 2017). Aligners, such as bowtie2 and BWA (Li and Durbin 2009) can be used in this approach, however, it is not possible to identify splicing events in new junctions when using a transcriptome as reference (Garber et al. 2011).

Combined approach

High-sensitivity reference-guided assemblers can be combined with de-novo assemblers in order to detect novel and missing transcripts as well. If the reference genome is incomplete, fragmented, or from a distantly related species, the de-novo assembly should be performed first in order to avoid the potential errors in the reference. This approach can also be useful for extending incomplete transcripts to full-length by merging these based on a reference (Martin and

Wang 2011). On the other hand, if a good quality reference genome is available, the combined approach should start by aligning the reads to a reference, followed by de-novo assembly of the reads that cannot be mapped. Thus, this method can also be used to filter out unwanted sequences before de-novo assembly.

Assembly quality, annotation, and quantification

The average length of assembled contigs in an RNA-seq experiment will vary based on the actual mRNA fragments that are sequenced. Thus, metrics based on assembled contigs do not necessarily indicate the quality of a transcriptome assembly. Transcriptome-specific metrics have been suggested such as ExN50, which computes transcript lengths as expression-weighted means of isoform lengths. Another method to assess the assembly quality is by checking the read percentage that can concordantly align to the final assembly in order to understand if the full complement of paired-end reads are represented in the assembled transcripts. Tools such as bowtie2 or BWA can be used for this type of mapping. Other tools for evaluating the quality of an assembled transcriptome include DETONATE (Li et al. 2014) and TransRate (Smith-Unna et al. 2016).

Transcripts can also be translated into protein sequences and mapped against well annotated databases such as UniProt/Swiss-Prot, Pfam, or NCBI. If the sequenced organism is closely related to a model organism, a high proportion of the contigs should have potential homologs in these databases. Another tool, BUSCO, assesses the completeness of the assembly by comparing it with universal single-copy gene databases specific to different lineages such as bacteria, fungi, or plants.

Expression quantification is a critical step for most RNA-seq experiments. There are two main sources of systematic variability which might introduce errors to this process; (i) longer transcripts generate more reads than shorter transcripts at the same abundance due to RNA fragmentation during library construction, and (ii) the number of fragments mapped across samples are different due to varying number of reads produced for each run. Therefore, read counts need to be normalised in order to obtain accurate gene expression estimates. Inter-sample normalisation methods have been developed for differential expression analysis, such as DeSeq2 (Anders and Huber 2010) and trimmed mean of M-values (TMM) which is implemented in edgeR (Robinson et al. 2010). For next generation sequencing research, reads per kilobase per million mapped reads (RPKM) is the most widely used method. It also accounts for gene lengths and is implemented in Salmon (Patro et al. 2017). Methods based on machine learning algorithms such as RSEM (Li et al. 2014) and Sailfish (Patro et al. 2014) can consider additional variables such as library size. Reference-based quantification approaches need to map the assembled transcripts to the reference genome first and then quantify the annotated genes. Functional annotations of the top expressed transcripts can be quickly examined at this step to check if tissue specific genes are abundantly expressed (e.g., genes known to be important for photosynthesis in leaf tissue).

Assembled transcripts from de-novo or reference-guided assemblies are expected to represent real biological differences such as expression levels, alternative splice forms, and paralogous or allelic transcripts (Schliesky et al. 2012). However, assembling plant transcriptomes with short reads can be more challenging compared to bacteria or lower eukaryotes, due to factors such as polyploidy, diversity in alternative splice isoforms and the heterozygosity of alleles (Góngora-Castillo and Buell 2013; Martin and Wang 2011). Thus, experimental strategies and bioinformatics pipelines should be developed specifically for each individual study and take the target organism and research questions into consideration.

Applications of transcriptomics in species identification

Marker discovery for phylogenomic inference

Transcriptomes have been used for plant phylogenomic inference as they contain abundant information from the nuclear genome. Famously, the generation of > 1000 transcriptomes across the plant kingdom led to new evolutionary insights for land plants (One Thousand Plant Transcriptomes Initiative 2019). However, the application of RNA-seq is limited to fresh tissue with low levels of degradation, making it less applicable to studies with large taxonomic sampling.

An emerging phylogenomic approach that partly relies on transcriptomics uses targeted next-generation sequencing (see Chapter 14 Target capture) to obtain specific genes for high-coverage DNA sequencing in large numbers of samples with varying taxonomic breadth. Target capture is very efficient in recovering hundreds of genes, regardless of the degradation level in the source DNA (Brewer et al. 2019; Hart et al. 2016) making this technique ideally suited for studies of plant systematics. Transcriptomes are the most commonly used source for nuclear marker discovery (Chamala et al. 2015), especially in the absence of a reference genome sequence (see Chapter 14 Target capture). They can be compared against curated databases of low-copy nuclear genes (De Smet et al. 2013) and form the basis for designing the short oligonucleotide 'baits' used to capture the target genes.

Metatranscriptomics

Metatranscriptomics is the application of transcriptome sequencing to environmental samples such as water, soil, or sediments. It gives an overview of the actual metabolic activity and taxonomic diversity within a community. The protocol involves HTS of reverse-transcribed cDNA obtained from an environmental mRNA isolate. While reverse transcriptase PCRs can only detect a single gene at a time, metatranscriptomics gives a whole gene expression profile of a diverse community of organisms playing various functional roles in the ecosystem (Carvalhais et al. 2012; Mason et al. 2012). Coupling these analyses with taxonomically informative rRNA offers the possibility to gather information on the community composition as well.

Some of the main challenges of metatranscriptomics are the presence of PCR inhibitors in environmental samples (e.g., humic acid, polysaccharides; Crump et al. 2018) and the low fraction of mRNA in the total RNA isolates (less than 5%) (Creer et al. 2016). An additional PCR step is often used to increase the total amount of genetic material. However, this might result in biased detection of diversity and quantification estimates in downstream steps (Porter and Hajibabaei 2018). Another challenge is to assign mRNA sequences to a specific function, as existing databases contain the most abundant genes in a limited number of environmental samples, or the genes from cultured species representing a limited proportion of environmental diversity (Prosser 2015).

There are various applications of metatranscriptomics such as revealing the composition of freshwater bacterioplankton communities (Poretsky et al. 2005) and animal/plant microbiomes (Crump et al. 2018; Pérez-Losada et al. 2015), understanding the dynamics of cyanobacterial blooms and viral-host relationships (Berg et al. 2018; Moniruzzaman et al. 2017; Shi et al. 2009), identifying important biochemical pathways (Franzosa et al. 2014; Saminathan et al. 2018), and understanding the mechanisms behind infection and disease (Hayden et al. 2018).

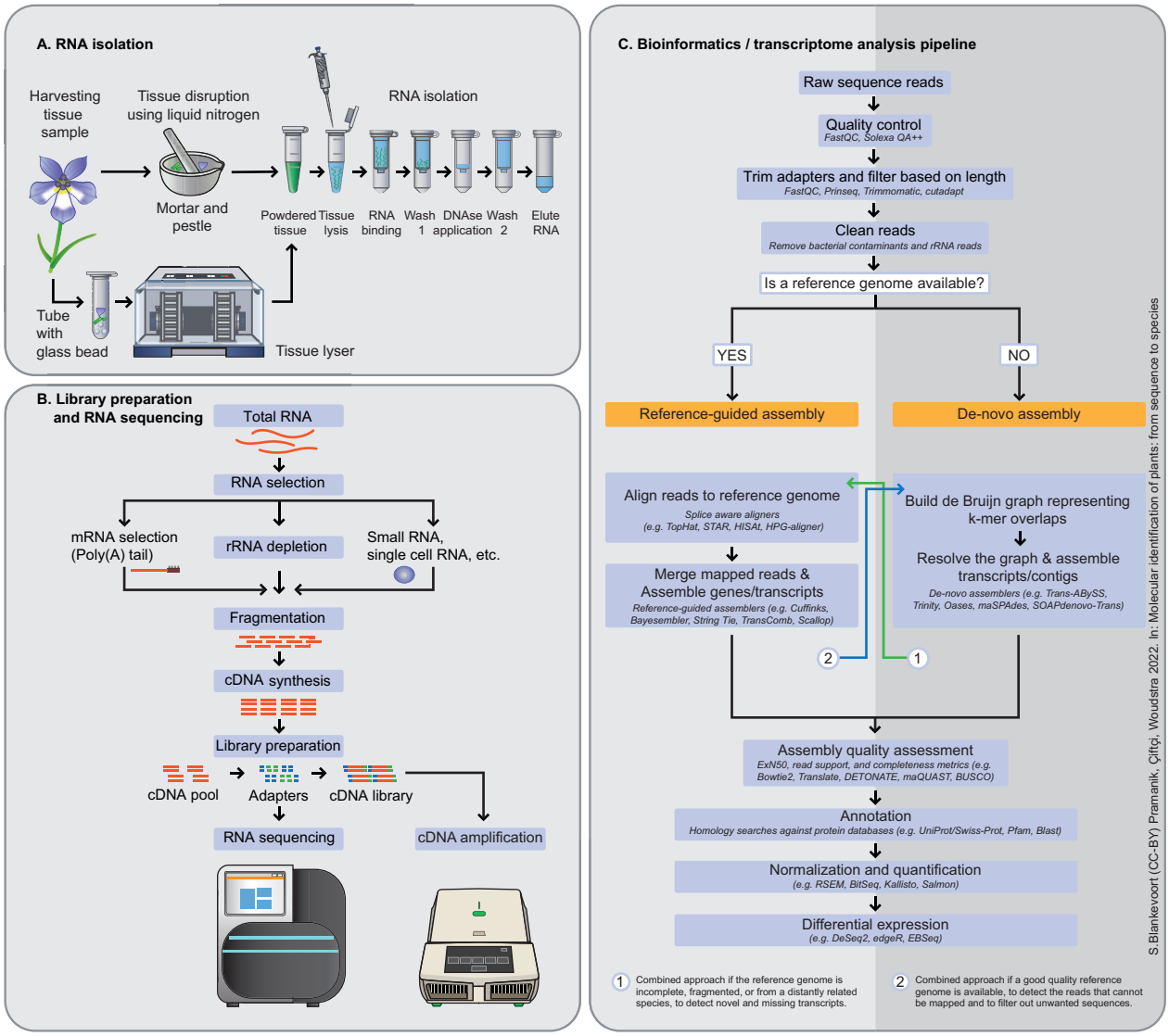

Figure 1. Chapter 15 Infographic: Overview scheme of transcriptomics in plants with emphasis on the RNA-seq method. **(a)** Sample preparation and RNA isolation. **(b)** Library preparation starts by selecting the RNA species from the total RNA, followed by fragmentation of the RNA sequence, cDNA synthesis, adapter ligation, cDNA amplification, and RNA sequencing. **(c)** The first step in transcriptome analysis is assessing the quality and quantity of reads. The clean reads are assembled to the reference genome or through a de-novo assembly or by combining these two approaches. The assemble reads are then annotated, followed by quantification and normalisation of the annotated results. The final step is differential expression analysis to quantify the difference in the expression level of genes between the samples or treatments.

Conclusion and future perspectives

Plant transcriptomics studies have undergone huge advances over the past few years as the costs of the second generation of sequencing, such as Illumina, have declined, third generation sequencing has become more accurate, and a wider range of analysis tools and pipelines have become available and become more accurate (Heather and Chain 2016). It is now possible to sequence large numbers of genes, or even whole genomes, for phylogenomic analyses (Soltis and Soltis 2020). Transcriptomics studies are an integral part of plant research and it's widely

used for marker identification, phylogenetic inference, species diversification, genetic response to abiotic and biotic stresses, evolution and development, metatranscriptomics to reveal the relation between plant and microbiome etc.

Studies using comparative transcriptomics to understand interactions between different organisms (Hayden et al. 2018), as well RNA-seq for single-cell work in particular are at the forefront of transcriptomic applications in functional studies, and open up the possibility of looking into the complex network of gene regulation, with significant implications in both fundamental science as well as in more applied fields such as crop development (Rich-Griffin et al. 2020).

Questions

1. What is the difference between genomics, proteomics, and transcriptomics?
2. What is a more suitable library preparation approach for comparative gene expression study: poly(A) enrichment or rRNA depletion? Motivate your answer.
3. Describe three criteria that are critical for the choice of reference-based vs. de-novo assembly approaches.

Glossary

Adapter – Chemically synthesised single-stranded or double-stranded oligonucleotides to capture a DNA sequence of interest.
Artefacts – Variations in sequences because of non-biological processes. For example, chemical reactions can cause changes in the nucleotides during the sequencing process.
Bait – An oligonucleotide designed for capturing a specific RNA or DNA species for sequencing.
BUSCO – Benchmarking Universal Single-Copy Orthologs.
Contigs – An assembled continuous sequence from overlapping DNA segments.
de Bruijn graph – A graphical representation of overlapping sequences which is used to construct whole length contigs.
De novo assembly – A method for creating a transcriptome assembly without a reference genome.
DNAse (deoxyribonuclease) - An enzyme that cleaves and degrades DNA.
Genome – A haploid set of chromosomes, including genes in microorganisms.
High-throughput sequencing (HTS) – A sequencing technology that enables large massively parallel DNA sequencing.
k-mer – subsequences of length k contained in a nucleotide or amino acid sequence.
Loci (plural locus) – The specific position of a particular gene or marker located on a chromosome.
Oligo (dT) primer – A single-stranded sequence of deoxythymine (dT) that is suitable to use as a primer with reverse transcriptase for first strand cDNA synthesis.
Random primer – Random primers are random hexadeoxynucleotides used for first-strand cDNA synthesis and cloning.
Reverse transcriptase (RT) – A DNA polymerase that enables the synthesis of a double helix DNA (cDNA) from RNA.
RNA transcripts – Single-stranded RNA products (e.g., mRNA, tRNA) synthesised by transcription of DNA.
RNase (ribonuclease) – A nuclease that catalyses the degradation of RNA into small fragments.

Sequencing depth – The coverage that represents the number of unique reads that include a given nucleotide in a final reconstructed sequence.

Splice junction – The site on the mature RNA indicating the position of a former intron which was spliced out after transcription.

Splice variants – A variant form of an mRNA produced by RNA genetic alteration in the DNA sequence that occurs at the splice site or the boundary of an exon and an intron.

Transcriptome assembly – The reconstruction of the RNA sequence composition of a biological sample by computational processing of the raw reads obtained by RNA-seq and subsequent steps for aligning and merging fragments from a longer transcript sequence in order to reconstruct the original sequence.

References

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252, 1651–1656. https://doi.org/10.1126/science.2047873

Allewell NM, Sama A (1974) The effect of ammonium sulfate on the activity of ribonuclease A. Biochimica et Biophysica Acta (BBA) - Enzymology 341, 484–488. https://doi.org/10.1016/0005-2744(74)90240-X

Alwine JC, Kemp DJ, Stark GR (1977) Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. Proc Natl Acad Sci USA 74, 5350–5354. https://doi.org/10.1073/pnas.74.12.5350

Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q (2020) Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 21, 30. https://doi.org/10.1186/s13059-020-1935-5

Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol. 11, R106. https://doi.org/10.1186/gb-2010-11-10-r106

Andrew S (2010) A quality control tool for high throughput sequence data. Fast QC.

Berg C, Dupont CL, Asplund-Samuelsson J, Celepli NA, Eiler A, Allen AE, Ekman M, Bergman B, Ininbergs K (2018) Dissection of microbial community functions during a cyanobacterial bloom in the Baltic Sea via metatranscriptomics. Front. Mar. Sci. 5. https://doi.org/10.3389/fmars.2018.00055

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, Roth R, George D, Eletr S, Albrecht G, Vermaas E, Williams SR, Moon K, Burcham T, Pallas M, DuBridge RB, Corcoran K (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat. Biotechnol. 18, 630–634. https://doi.org/10.1038/76469

Brewer GE, Clarkson JJ, Maurin O, Zuntini AR, Barber V, Bellot S, Biggs N, Cowan RS, Davies NMJ, Dodsworth S, Edwards SL, Eiserhardt WL, Epitawalage N, Frisby S, Grall A, Kersey PJ, Pokorny L, Leitch IJ, Forest F, Baker WJ (2019) Factors affecting targeted sequencing of 353 nuclear genes from herbarium specimens spanning the diversity of angiosperms. Front. Plant Sci. 10, 1102. https://doi.org/10.3389/fpls.2019.01102

Burgess DJ (2019) Spatial transcriptomics coming of age. Nat. Rev. Genet. 20, 317. https://doi.org/10.1038/s41576-019-0129-z

Carvalhais LC, Dennis PG, Tyson GW, Schenk PM (2012) Application of metatranscriptomics to soil environments. J. Microbiol. Methods 91, 246–251. https://doi.org/10.1016/j.mimet.2012.08.011

Carvalhais LC, Schenk PM (2013) Sample processing and cDNA preparation for microbial metatranscriptomics in complex soil communities. Meth. Enzymol. 531, 251–267. https://doi.org/10.1016/B978-0-12-407863-5.00013-7

Carvalhais, V, Delgado-Rastrollo, M, Melo, LDR, Cerca, N (2013) Controlled RNA contamination and degradation and its impact on qPCR gene expression in *S. epidermidis* biofilms. J. Microbiol. Methods 95, 195–200. https://doi.org/10.1016/j.mimet.2013.08.010

Chamala S, García N, Godden GT, Krishnakumar V, Jordon-Thaden IE, De Smet R, Barbazuk WB, Soltis DE, Soltis PS (2015) MarkerMiner 1.0: a new application for phylogenetic marker development using angiosperm transcriptomes. Appl. Plant Sci. 3, 1400115. https://doi.org/10.3732/apps.1400115

Chhangawala S, Rudy G, Mason CE, Rosenfeld JA (2015) The impact of read length on quantification of differentially expressed genes and splice junction detection. Genome Biol. 16, 131. https://doi.org/10.1186/s13059-015-0697-y

Chomczynski P, Sacchi N (2006) The single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction: twenty-something years on. Nat. Protoc. 1, 581–585. https://doi.org/10.1038/nprot.2006.83

Compeau PEC, Pevzner PA, Tesler G (2011) How to apply de Bruijn graphs to genome assembly. Nat. Biotechnol. 29, 987–991. https://doi.org/10.1038/nbt.2023

Creer S, Deiner K, Frey S, Porazinska D, Taberlet P, Thomas WK, Potter C, Bik HM (2016) The ecologist's field guide to sequence-based identification of biodiversity. Methods Ecol. Evol. 7, 1008–1018. https://doi.org/10.1111/2041-210X.12574

Crump BC, Wojahn JM, Tomas F, Mueller RS (2018) Metatranscriptomics and amplicon sequencing reveal mutualisms in seagrass microbiomes. Front. Microbiol. 9, 388. https://doi.org/10.3389/fmicb.2018.00388

Depledge DP, Srinivas KP, Sadaoka T, Bready D, Mori Y, Placantonakis DG, Mohr I, Wilson AC (2019) Direct RNA sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen. Nat. Commun. 10, 754. https://doi.org/10.1038/s41467-019-08734-9

De Smet R, Adams KL, Vandepoele K, Van Montagu MCE, Maere S, Van de Peer Y (2013) Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. Proc Natl Acad Sci USA 110, 2898–2903. https://doi.org/10.1073/pnas.1300127110

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21. https://doi.org/10.1093/bioinformatics/bts635

Efroni I, Birnbaum KD (2016) The potential of single-cell profiling in plants. Genome Biol. 17, 65. https://doi.org/10.1186/s13059-016-0931-2

Franzosa EA, Morgan XC, Segata N, Waldron L, Reyes J, Earl AM, Giannoukos G, Boylan MR, Ciulla D, Gevers D, Izard J, Garrett WS, Chan AT, Huttenhower C (2014) Relating the metatranscriptome and metagenome of the human gut. Proc Natl Acad Sci USA 111, E2329-38. https://doi.org/10.1073/pnas.1319284111

Fu GK, Xu W, Wilhelmy J, Mindrinos MN, Davis RW, Xiao W, Fodor SPA (2014) Molecular indexing enables quantitative targeted RNA sequencing and reveals poor efficiencies in standard library preparations. Proc Natl Acad Sci USA 111, 1891–1896. https://doi.org/10.1073/pnas.1323732111

Garber M, Grabherr MG, Guttman M, Trapnell C (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. Nat. Methods 8, 469–477. https://doi.org/10.1038/nmeth.1613

Gilbert D (2016) Gene-omes built from mRNA seq not genome DNA. F1000Research. https://doi.org/10.7490/f1000research.1112594.1

Góngora-Castillo E, Buell CR (2013) Bioinformatics challenges in de novo transcriptome assembly using short read sequences in the absence of a reference genome sequence. Nat. Prod. Rep. 30, 490–500. https://doi.org/10.1039/c3np20099j

Góngora-Castillo E, Fedewa G, Yeo Y, Chappell J, DellaPenna D, Buell CR (2012) Genomic approaches for interrogating the biochemistry of medicinal plant species. Meth. Enzymol. 517, 139–159. https://doi.org/10.1016/B978-0-12-404634-4.00007-3

Gordon A, Hannon GJ (2009) FASTQ/A short-reads pre-processing tools. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29, 644–652. https://doi.org/10.1038/nbt.1883

Green MR, Sambrook J (2019) Isolation of poly(a)+ messenger RNA using magnetic oligo(dt) beads. Cold Spring Harb. Protoc. 2019. https://doi.org/10.1101/pdb.prot101733

Gruenheit N, Deusch O, Esser C, Becker M, Voelckel C, Lockhart P (2012) Cutoffs and k-mers: implications from a transcriptome study in allopolyploid plants. BMC Genomics 13, 92. https://doi.org/10.1186/1471-2164-13-92

Hart ML, Forrest LL, Nicholls JA, Kidner CA (2016) Retrieval of hundreds of nuclear loci from herbarium specimens. Taxon 65, 1081–1092. https://doi.org/10.12705/655.9

Hayden HL, Savin KW, Wadeson J, Gupta VVSR, Mele PM (2018) Comparative metatranscriptomics of wheat rhizosphere microbiomes in disease suppressive and non-suppressive soils for *Rhizoctonia solani* AG8. Front. Microbiol. 9, 859. https://doi.org/10.3389/fmicb.2018.00859

Heather JM, Chain B (2016) The sequence of sequencers: The history of sequencing DNA. Genomics 107, 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003

Heydari M, Miclotte G, Demeester P, Van de Peer Y, Fostier J (2017) Evaluation of the impact of Illumina error correction tools on de novo genome assembly. BMC Bioinformatics 18, 374. https://doi.org/10.1186/s12859-017-1784-8

Hölzer M, Marz M (2019) De novo transcriptome assembly: a comprehensive cross-species comparison of short-read RNA-Seq assemblers. Gigascience 8, giz039. https://doi.org/10.1093/gigascience/giz039

Hrdlickova R, Toloue M, Tian B (2017) RNA-Seq methods for transcriptome analysis. Wiley Interdiscip. Rev. RNA 8, e1364. https://doi.org/10.1002/wrna.1364

Hsieh P-H, Oyang Y-J, Chen C-Y (2019) Effect of de novo transcriptome assembly on transcript quantification. Sci. Rep. 9, 8304. https://doi.org/10.1038/s41598-019-44499-3

Hu J, Uapinyoying P, Goecks J (2017) Interactive analysis of long-read RNA isoforms with Iso-Seq Browser. BioRxiv. https://doi.org/10.1101/102905

Jahn CE, Charkowski AO, Willis DK (2008) Evaluation of isolation methods and RNA integrity for bacterial RNA quantitation. J. Microbiol. Methods 75, 318–324. https://doi.org/10.1016/j.mimet.2008.07.004

Kannan S, Hui J, Mazooji K, Pachter L, Tse D (2016) Shannon: an information-optimal de novo RNA-Seq assembler. BioRxiv. https://doi.org/10.1101/039230

Kerr SC, Gaiti F, Tanurdzic M (2019) De novo plant transcriptome assembly and annotation using Illumina RNA-Seq reads. Methods Mol. Biol. 1933, 265–275. https://doi.org/10.1007/978-1-4939-9045-0_16

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14, R36. https://doi.org/10.1186/gb-2013-14-4-r36

Kivivirta K, Herbert D, Lange M, Beuerlein K, Altmüller J, Becker A (2019) A protocol for laser microdissection (LMD) followed by transcriptome analysis of plant reproductive tissue in phylogenetically distant angiosperms. Plant Methods 15, 151. https://doi.org/10.1186/s13007-019-0536-3

Kraus AJ, Brink BG, Siegel TN (2019) Efficient and specific oligo-based depletion of rRNA. Sci. Rep. 9, 12281. https://doi.org/10.1038/s41598-019-48692-2

Kukurba KR, Montgomery SB (2015) RNA sequencing and analysis. Cold Spring Harb. Protoc. 2015, 951–969. https://doi.org/10.1101/pdb.top084970

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359. https://doi.org/10.1038/nmeth.1923

Leh TY, Yong CSY, Nulit R, Abdullah JO (2019) Efficient and high-quality RNA isolation from metabolite-rich tissues of *Stevia rebaudiana*, an important commercial crop. Trop. Life Sci. Res. 30, 149–159. https://doi.org/10.21315/tlsr2019.30.1.9

Li B, Fillmore N, Bai Y, Collins M, Thomson JA, Stewart R, Dewey CN (2014) Evaluation of de novo transcriptome assemblies from RNA-Seq data. Genome Biol. 15, 553. https://doi.org/10.1186/s13059-014-0553-5

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li H (2015) BFC: correcting Illumina sequencing errors. Bioinformatics 31, 2885–2887. https://doi.org/10.1093/bioinformatics/btv290

Li J, Harata-Lee Y, Denton MD, Feng Q, Rathjen JR, Qu Z, Adelson DL (2017) Long read reference genome-free reconstruction of a full-length transcriptome from *Astragalus* membranaceus reveals transcript variants involved in bioactive compound biosynthesis. Cell Discov. 3, 17031. https://doi.org/10.1038/celldisc.2017.31

Li S, Yamada M, Han X, Ohler U, Benfey PN (2016) High-resolution expression map of the *Arabidopsis* root reveals alternative splicing and lincRNA regulation. Dev. Cell 39, 508–522. https://doi.org/10.1016/j.devcel.2016.10.012

Long Y, Liu Z, Jia J, Mo W, Fang L, Lu D, Liu B, Zhang H, Chen W, Zhai J (2021) FlsnRNA-seq: protoplasting-free full-length single-nucleus RNA profiling in plants. Genome Biol. 22, 66. https://doi.org/10.1186/s13059-021-02288-0

Maceda-López LF, Villalpando-Aguilar JL, García-Hernández E, Ávila de Dios E, Andrade-Canto SB, Morán-Velázquez DC, Rodríguez-López L, Hernández-Díaz D, Chablé-Vega MA, Trejo L, Góngora-Castillo E, López-Rosas I, Simpson J, Alatorre-Cobos F (2021) Improved method for isolation of high-quality total RNA from *Agave tequilana* Weber roots. 3 Biotech 11, 75. https://doi.org/10.1007/s13205-020-02620-8

MacManes MD (2018) The Oyster River Protocol: a multi-assembler and kmer approach for de novo transcriptome assembly. PeerJ 6, e5428. https://doi.org/10.7717/peerj.5428

Mao S, Pachter L, Tse D, Kannan S (2020) RefShannon: a genome-guided transcriptome assembler using sparse flow decomposition. PLoS ONE 15, e0232946. https://doi.org/10.1371/journal.pone.0232946

Marine R, Polson SW, Ravel J, Hatfull G, Russell D, Sullivan M, Syed F, Dumas M, Wommack KE (2011) Evaluation of a transposase protocol for rapid generation of shotgun high-throughput sequencing libraries from nanogram quantities of DNA. Appl. Environ. Microbiol. 77, 8071–8079. https://doi.org/10.1128/AEM.05610-11

Martin JA, Wang Z (2011) Next-generation transcriptome assembly. Nat. Rev. Genet. 12, 671–682. https://doi.org/10.1038/nrg3068

Mason OU, Hazen TC, Borglin S, Chain PSG, Dubinsky EA, Fortney JL, Han J, Holman H-YN, Hultman J, Lamendella R, Mackelprang R, Malfatti S, Tom LM, Tringe SG, Woyke T, Zhou J, Rubin EM, Jansson JK (2012) Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill. ISME J. 6, 1715–1727. https://doi.org/10.1038/ismej.2012.59

McGettigan PA (2013) Transcriptomics in the RNA-seq era. Curr. Opin. Chem. Biol. 17, 4–11. https://doi.org/10.1016/j.cbpa.2012.12.008

Moniruzzaman M, Wurch LL, Alexander H, Dyhrman ST, Gobler CJ, Wilhelm SW (2017) Virus-host relationships of marine single-celled eukaryotes resolved from metatranscriptomics. Nat. Commun. 8, 16054. https://doi.org/10.1038/ncomms16054

Morozova O, Hirst M, Marra MA (2009) Applications of new sequencing technologies for transcriptome analysis. Annu. Rev. Genomics Hum. Genet. 10, 135–151. https://doi.org/10.1146/annurev-genom-082908-145957

Oikonomopoulos S, Bayega A, Fahiminiya S, Djambazian H, Berube P, Ragoussis J (2020) Methodologies for transcript profiling using long-read technologies. Front. Genet. 11, 606. https://doi.org/10.3389/fgene.2020.00606

One Thousand Plant Transcriptomes Initiative (2019) One thousand plant transcriptomes and the phylogenomics of green plants. Nature 574, 679–685. https://doi.org/10.1038/s41586-019-1693-2

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C (2017) Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods 14, 417–419. https://doi.org/10.1038/nmeth.4197

Patro R, Mount SM, Kingsford C (2014) Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat. Biotechnol. 32, 462–464. https://doi.org/10.1038/nbt.2862

Pérez-Losada M, Castro-Nallar E, Bendall ML, Freishtat RJ, Crandall KA (2015) Dual transcriptomic profiling of host and microbiota during health and disease in pediatric asthma. PLoS ONE 10, e0131819. https://doi.org/10.1371/journal.pone.0131819

Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. 33, 290–295. https://doi.org/10.1038/nbt.3122

Pevzner PA, Tang H, Waterman MS (2001) An Eulerian path approach to DNA fragment assembly. Proc Natl Acad Sci USA 98, 9748–9753. https://doi.org/10.1073/pnas.171285098

Piétu G, Eveno E, Soury-Segurens B, Fayein NA, Mariage-Samson R, Matingou C, Leroy E, Dechesne C, Krieger S, Ansorge W, Reguigne-Arnould I, Cox D, Dehejia A, Polymeropoulos MH, Devignes MD, Auffray C (1999) The genexpress IMAGE knowledge base of the human muscle transcriptome: a resource of structural, functional, and positional candidate genes for muscle physiology and pathologies. Genome Res. 9, 1313–1320. https://doi.org/10.1101/gr.9.12.1313

Poretsky RS, Bano N, Buchan A, LeCleir G, Kleikemper J, Pickering M, Pate WM, Moran MA, Hollibaugh JT (2005) Analysis of microbial gene transcripts in environmental samples. Appl. Environ. Microbiol. 71, 4121–4126. https://doi.org/10.1128/AEM.71.7.4121-4126.2005

Porter TM, Hajibabaei M (2018) Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. Mol. Ecol. 27, 313–338. https://doi.org/10.1111/mec.14478

Prosser JI (2015) Dispersing misconceptions and identifying opportunities for the use of "omics" in soil microbial ecology. Nat. Rev. Microbiol. 13, 439–446. https://doi.org/10.1038/nrmicro3468

Rich-Griffin C, Stechemesser A, Finch J, Lucas E, Ott S, Schäfer P (2020) Single-cell transcriptomics: a high-resolution avenue for plant functional genomics. Trends Plant Sci. 25, 186–197. https://doi.org/10.1016/j.tplants.2019.10.008

Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ, Griffith M, Raymond A, Thiessen N, Cezard T, Butterfield YS, Newsome R, Chan SK, She R, Varhol R, Kamoh B, Birol I (2010) De novo assembly and analysis of RNA-seq data. Nat. Methods 7, 909–912. https://doi.org/10.1038/nmeth.1517

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140. https://doi.org/10.1093/bioinformatics/btp616

Saminathan T, García M, Ghimire B, Lopez C, Bodunrin A, Nimmakayala P, Abburi VL, Levi A, Balagurusamy N, Reddy UK (2018) Metagenomic and metatranscriptomic analyses of diverse watermelon cultivars reveal the role of fruit associated microbiome in carbohydrate metabolism and ripening of mature fruits. Front. Plant Sci. 9, 4. https://doi.org/10.3389/fpls.2018.00004

Schliesky S, Gowik U, Weber APM, Bräutigam A (2012) RNA-Seq assembly - are we there yet? Front. Plant Sci. 3, 220. https://doi.org/10.3389/fpls.2012.00220

Schmieder R, Edwards R (2011a) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27, 863–864. https://doi.org/10.1093/bioinformatics/btr026

Schmieder R, Edwards R (2011b) Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS ONE 6, e17288. https://doi.org/10.1371/journal.pone.0017288

Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M, Lightfoot S, Menzel W, Granzow M, Ragg T (2006) The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Mol. Biol. 7, 3. https://doi.org/10.1186/1471-2199-7-3

Schulz MH, Weese D, Holtgrewe M, Dimitrova V, Niu S, Reinert K, Richard H (2014) Fiona: a parallel and automatic strategy for read error correction. Bioinformatics 30, i356-63. https://doi.org/10.1093/bioinformatics/btu440

Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28, 1086–1092. https://doi.org/10.1093/bioinformatics/bts094

Shakya M, Lo C-C, Chain PSG (2019) Advances and challenges in metatranscriptomic analysis. Front. Genet. 10, 904. https://doi.org/10.3389/fgene.2019.00904

Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T, Fukuda S, Sasaki D, Podhajska A, Harbers M, Kawai J, Carninci P, Hayashizaki Y (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci USA 100, 15776–15781. https://doi.org/10.1073/pnas.2136655100

Shi Y, Tyson GW, DeLong EF (2009) Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. Nature 459, 266–269. https://doi.org/10.1038/nature08055

Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP (2014) Sequencing depth and coverage: key considerations in genomic analyses. Nat. Rev. Genet. 15, 121–132. https://doi.org/10.1038/nrg3642

Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S (2016) TransRate: reference-free quality assessment of de novo transcriptome assemblies. Genome Res. 26, 1134–1144. https://doi.org/10.1101/gr.196469.115

Soltis PS, Soltis DE (2020) Plant genomes: markers of evolutionary history and drivers of evolutionary change. Plants, People, Planet. https://doi.org/10.1002/ppp3.10159

Song L, Florea L (2015) Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. Gigascience 4, 48. https://doi.org/10.1186/s13742-015-0089-y

Stark R, Grzelak M, Hadfield J (2019) RNA sequencing: the teenage years. Nat. Rev. Genet. 20, 631–656. https://doi.org/10.1038/s41576-019-0150-2

Sundell D, Street NR, Kumar M, Mellerowicz EJ, Kucukoglu M, Johnsson C, Kumar V, Mannapperuma C, Delhomme N, Nilsson O, Tuominen H, Pesquet E, Fischer U, Niittylä T, Sundberg B, Hvidsten TR (2017) AspWood: high-spatial-resolution transcriptome profiles reveal uncharacterized modularity of wood formation in *Populus tremula*. Plant Cell 29, 1585–1604. https://doi.org/10.1105/tpc.17.00153

Thibivilliers S, Anderson D, Libault M (2020) Isolation of plant root nuclei for single cell RNA sequencing. Curr. Protoc. Plant Biol. 5, e20120. https://doi.org/10.1002/cppb.20120

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. 7, 562–578. https://doi.org/10.1038/nprot.2012.016

Ungaro A, Pech N, Martin J-F, McCairns RJS, Mévy J-P, Chappaz R, Gilles A (2017) Challenges and advances for transcriptome assembly in non-model species. PLoS ONE 12, e0185020. https://doi.org/10.1371/journal.pone.0185020

Van Verk MC, Hickman R, Pieterse CMJ, Van Wees SCM (2013) RNA-Seq: revelation of the messengers. Trends Plant Sci. 18, 175–179. https://doi.org/10.1016/j.tplants.2013.02.001

Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett DE, Hieter P, Vogelstein B, Kinzler KW (1997) Characterization of the yeast transcriptome. Cell 88, 243–251. https://doi.org/10.1016/s0092-8674(00)81845-0

Voshall A, Moriyama EN (2018) Next-generation transcriptome assembly: strategies and performance analysis. In: Abdurakhmonov, I.Y. (Ed.) Bioinformatics in the Era of Post Genomics and Big Data. InTech. https://doi.org/10.5772/intechopen.73497

Wang B, Kumar V, Olson A, Ware D (2019) Reviving the transcriptome studies: an insight into the emergence of single-molecule transcriptome sequencing. Front. Genet. 10, 384. https://doi.org/10.3389/fgene.2019.00384

Wang T, Wang H, Cai D, Gao Y, Zhang H, Wang Y, Lin C, Ma L, Gu L (2017) Comprehensive profiling of rhizome-associated alternative splicing and alternative polyadenylation in moso bamboo (*Phyllostachys edulis*). Plant J. 91, 684–699. https://doi.org/10.1111/tpj.13597

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10, 57–63. https://doi.org/10.1038/nrg2484

Weber APM, Weber KL, Carr K, Wilkerson C, Ohlrogge JB (2007) Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing. Plant Physiol. 144, 32–42. https://doi.org/10.1104/pp.107.096677

Wilfinger WW, Mackey K, Chomczynski P (1997) Effect of pH and ionic strength on the spectrophotometric assessment of nucleic acid purity. BioTechniques 22, 474–6, 478. https://doi.org/10.2144/97223st01

Zhao S, Zhang Y, Gamini R, Zhang B, von Schack D (2018) Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. Sci. Rep. 8, 4781. https://doi.org/10.1038/s41598-018-23226-4

Answers

1. Genomics is the study of the entire genome from the complete set of DNA in the nucleus, chloroplasts and mitochondria of somatic cells. Proteomics is the study of proteins, protein complexes, their localization, their interactions, and posttranslational modifications. Transcriptomics is the study of genome expression products of the mRNAs that are actively expressed at any given moment in a cell or organism.
2. Poly(A) enrichment is suitable for comparative gene expression studies. Poly(A) can produce sufficient mRNA and separates mRNA from rRNA contaminants. This results in higher exonic coverage for protein-coding genes of a transcriptome. rRNA depletion is mainly applied for comprehensive transcriptome studies. It can capture a wider diversity of unique transcriptome features such as mRNAs lacking the poly(A) tail, long ncRNAs, newly transcribed, and unprocessed transcripts. If rRNA depletion is used for comparative gene expression studies, the results will be biased. The rRNA depletion method results in most reads mapping to intronic regions reducing the number of reads to the exonic region. It also overestimates the expression levels of the genes that overlap with the intronic regions of other genes.
3. (i) Availability, completeness, and fragmentation of reference genomes of target species. (ii) Availability of reference genomes of closely related species. (iii) Detecting novel transcripts.

Chapter 16
Whole genome sequencing

Alex D. Twyford[1,2], Margaretha A. Veltman[3], Phen Garrett[4]

1 University of Edinburgh, Edinburgh, United Kingdom
2 Royal Botanic Garden Edinburgh, Edinburgh, United Kingdom
3 Natural History Museum, University of Oslo, Oslo, Norway
4 Globe Institute, University of Copenhagen, Copenhagen, Denmark

Alex D. Twyford **alex.twyford@ed.ac.uk**
Margaretha A. Veltman **margret.veltman@nhm.uio.no**
Phen Garrett **phengarrett@gmail.com**

Citation: Twyford AD, Veltman MA, Garrett P (2022) Chapter 16. Whole genome sequencing. In: de Boer H, Rydmark MO, Verstraete B, Gravendeel B (Eds) Molecular identification of plants: from sequence to species. Advanced Books. https://doi.org/10.3897/ab.e98875

Introduction

Modern sequencing technologies (see Chapter 9 Sequencing platforms and data types) make it possible to generate large-scale genomic sequencing data for any plant species. This dramatic step-change in genomic data availability, along with improvements in bioinformatic tools, has led to the release of many high-quality plant genomes (Michael and VanBuren 2020), as well as to the proposal of ambitious collaborative projects to produce reference genomes at scale (e.g., the 10,000 Plants Genomes Project (Cheng et al. 2018; Twyford 2018). While whole genome sequencing holds substantial promise for aiding plant identification, there are a number of technical considerations that currently prevent it from routine use. In particular, plant genomes vary 2,400-fold in size, with the largest genome approximately 50-fold larger than the human genome (Pellicer et al. 2018). Since sequencing costs are linearly related to the genome size, large plant genomes in many cases remain prohibitively expensive to sequence. Moreover, the genomic complexity associated with genome duplication events (polyploidy) and the accumulation of repetitive DNA, make the analysis of large plant genomes challenging (Kyriakidou et al. 2018).

In this chapter, we consider best practices for whole genome sequencing as a tool for plant identification, and the relative strengths and weaknesses of different genome sequencing approaches. We start by discussing the overall workflow common to any project using whole genome sequencing, before moving to the specific requirements of three approaches that differ in their sequencing coverage: (1) Genome skimming, which uses low-coverage sequence data to assemble well-represented (high copy number) genomic regions, (2) Genome resequencing, which uses modest-coverage sequence data to investigate genomic diversity relative to an existing nuclear reference genome sequence, (3) De novo whole genome assembly, which uses high-coverage sequence data to generate a nuclear reference genome. We also consider assembly-free approaches for using the nuclear genome.

Sample to sequence: an overview of the workflow for genome sequencing

Genomic sequencing starts with sample collection and DNA extraction, and finishes with a set of sequences or sequence variants that are suitable for analysis. The major stages are as follows.

DNA

Genome sequencing usually uses high-quality DNA extracted from plant tissue (see Chapter 1 DNA from plant tissue), though some approaches can accommodate DNA from degraded specimens (see Chapter 2 DNA from museum collections). The exact requirements depend on the downstream processes, but as a guide:

- Low initial concentrations (500 pg+) of degraded or intact DNA (fragment molecules 100 bp+) can be used for genome skimming (Zeng et al. 2018).
- Modest initial concentrations (100 ng+) of intact DNA without extensive degradation (fragment molecules 400 bp+) are typically used in genome resequencing.
- High initial concentrations (1 µg+) of high molecular weight DNA (fragment molecules 20 kb+) are typically used for de novo genome sequencing.

Most plant identification projects use DNA extracted from individual plant samples. However, metagenomic studies may work on mixed samples such as environmental DNA (see Chapter 12 Metagenomics), while some population genomic studies may choose to pool multiple individuals per population and compare diversity between these sample pools (e.g., Pool-seq; (Rellstab et al. 2013).

Library preparation

The wet lab protocols used to generate sequence-ready DNA libraries (see Chapter 9 Sequencing platforms and data types, Chapter 12 Metagenomics, Chapter 15 Transcriptomics) are varied and depend on the starting DNA quality and the intended downstream sequencing approach. Low amounts of starting input DNA will require amplification via PCR-based library preparation, while higher amounts of input DNA samples can be used in PCR-free libraries, which reduces bioinformatic issues with PCR duplicates. Further, a range of more lab intensive library preparation approaches are available for long read sequencing or to allow users to partition and barcode HMW DNA (e.g., linked read sequencing such as haplotagging; (Meier et al. 2021) or preserve genome-wide organisation in the cell (Hi-C), with these approaches aiding in de novo genome assembly (discussed below, and (Jiao and Schneeberger 2017).

Sequencing

Most plant identification studies using whole genome sequencing rely on short-read data, such as that generated on Illumina platforms or with BGI technologies. Here, the benefits of low per base-pair sequencing costs, high accuracy and throughput, and potential for sample multiplexing make it extremely well-suited for a range of applications. However, recent innovations in long-read sequencing have reduced error profiles and made it more cost-effective (Lang et al. 2020). Notably, PacBio HiFi uses consensus sequencing to generate long (10–25 kb) ultra-high quality reads, while Oxford Nanopore Technologies are consistently improving their raw-read accuracy. Long-read data allow researchers to investigate a broader spectrum of genetic variants including structural genomic variants such as large chromosomal inversions.

Bioinformatic analysis

The computational methods will vary considerably depending on sample type and number, sequence type, and downstream analysis approach. However most projects will involve the following initial stages:

1. De-multiplex samples to give separate sequence files per individual.
2. Data quality control (QC), to check the sequence quality, read number, and other sequence quality metrics.
3. Sequence data post-processing. This may involve trimming or quality filtering raw reads to remove low-quality sequences.
4. Genome assembly. For genome skimming, organellar and ribosome genomes can readily be assembled de novo (see below), while for nuclear genomes the data then goes through a multi-stage genome assembly pipeline (Li and Harkess 2018). This stage is not necessary in genome resequencing studies that rely on an existing reference genome.

5. Alignment. Studies of small genomes, such as plastids, will usually involve whole genome alignment. In resequencing studies, sequence reads are directly mapped to the reference genome (e.g., short-read mapping with bwa-mem). Various additional stages such as marking of duplicates and realignment around indels may be required to produce high quality alignments.
6. Variant calling to produce Single Nucleotide Polymorphism (SNP) datasets for downstream analyses. Many SNP callers are available, with the Genome Analysis Toolkit (GATK; (DePristo et al. 2011) being one of the most popular.
7. Quality filtering to remove low frequency SNPs and sites/individuals with lots of missing data. This may either be by applying 'hard' quality thresholds, or more sophisticated machine-learning approaches for removing sequence artefacts (e.g., variant quality score recalibration in model species with a high quality training set).

Genome skimming

Low-coverage sequencing of genomic DNA, 'genome skimming', is an efficient approach for comparative genomics of diverse species (Dodsworth 2015; Straub et al. 2012). Genome skimming involves sequencing genomic DNA at ca. 0.1–5X coverage. In plant identification studies, a primary use of genome skimming is to assemble high copy regions which will have a suitably high representation even at low average sequencing coverage (Twyford and Ness 2017). Organellar genomes are targeted for broad phylogenetic studies (Palmer et al. 1988), ribosomal DNA for research into recent species relationships, and some studies of genome evolution also investigate mitochondrial DNA diversity (see Box 1). Here, there are considerable benefits over amplicon sequencing (see Chapter 11 Amplicon metabarcoding), as genome skimming typically generates more data and avoids the need to target specific genomic regions with individual primer sets. Genome skimming also provides access to nuclear genomic repeats and as such can be used to characterise repeat content and the abundance of common repeat families. This data can prove useful for stand-alone investigations of genomic diversity, or used to infer phylogenetic relationships based on repeat abundance (Dodsworth et al. 2015).

Genome skimming stands out from many other genomic approaches for its technical ease. It is straightforward at all stages of the workflow, from DNA extraction requirements to easy and comparatively cheap library building options (Liu et al. 2019), to sequencing, to very-well established assembly approaches requiring little computational resources (Jin et al. 2020), and to comparative genomics. Perhaps most noticeably, genome skimming can be successfully used for degraded DNA such as historical museum samples (Zeng et al. 2018) where other whole genome sequencing approaches may fail. Moreover, all analyses can be performed without the need of a reference genome.

The downside of genome skimming is that it fails to reliably sample the bulk of the genome. The regions which are represented at high coverage, particularly organelle genomes, show atypical inheritance and evolutionary patterns, which may yield phylogenetic results that are incongruent with phylogenies of the nuclear genome (e.g., in Orchidaceae; (Pérez-Escobar et al. 2021) or lack the resolution needed to discriminate species, particularly in recent species radiations. Until recently, researchers most interested in recovering specific nuclear loci from degraded DNA, or from species without a reference genome, have been better served by choosing target capture and related methods (see Chapter 14 Target capture). More recently however, the distinction between genome skimming and resequencing have become blurred, with increasing interest in using low-coverage sequencing coupled with pipelines that account for genotype uncertainty to infer variation (reviewed in (Lou et al. 2021). These approaches are

particularly applicable to plant identification questions where confidence in individual genotypes can be sacrificed in exchange for more extensive sampling of individuals. This mostly applies to population genomic analyses where many individuals are sampled per population.

Chapter 16: Box 1. Plastid, ribosomal, and mitochondrial diversity

Plastids are organelles that are responsible for photosynthesis and the synthesis and storage of molecular products. Plastomes are mostly circular and nonrecombinant organellar genomes averaging 120–160 kb in size. Their high copy number per cell means high quality assembly is possible even from low depth nuclear genome sequencing (Coissac et al. 2016). They show a highly conserved gene order and low levels of nucleotide substitution, allowing for comparison over a wide phylogenetic scale (Twyford and Ness 2017). With mostly maternal inheritance, plastomes can also be used to infer the direction of hybridisation in evolutionary studies. Used in plant phylogenetics for over 40 years (Palmer et al. 1988), plastids have been a foundation for telling species apart and there are now thousands of available plastomes (Tonti-Filippini et al. 2017).

Nuclear ribosomal DNA (nrDNA) primarily functions to code for ribosomal RNA. Plant nrDNA has an average size range of 10–15 kb and exists as hundreds to thousands of tandem repeats occurring in high copy numbers throughout cells (Garcia et al. 2017). In contrast to the plastome, ribosomal biparental inheritance can permit the exploration of reticulation, hybrid speciation, and parentage of polyploids. However, nrDNA is subject to concerted evolution—where nrDNA is homogenised to a single copy—resulting in a loss of polymorphism, which can confound inferences of later generation hybridisation. Intragenomic uniformity and intergenomic variability offers the high variation sought after from the nuclear genome, while the high copy number makes nrDNA easy to recover even from degraded material. nrDNA sequences are best used to explore close relationships where fast-evolving, recently occurring relationships are being studied (Álvarez and Wendel 2003).

Mitochondrial genomes have a primary function in respiration. Despite their conserved function and generally conserved gene complement, they show significant structural variation in plants, including in size (100 kb–2.7 Mb), sequence arrangement and repeat content (Kozik et al. 2019). This structural variation makes them more challenging than plastids to assemble, while their high degree of structural change makes them difficult to compare across broad phylogenetic scales. However, their presence in multiple copies per cell, and their potentially different evolutionary histories to plastids, means they are sometimes used in plant phylogenetics (Knoop 2004).

Genome resequencing

Genome resequencing involves sequencing samples at a moderate depth (often 5–30X coverage) and analysing the data in the context of an existing reference genome. Most genome resequencing studies use short-read data and subsequently investigate SNPs and small indels, however, long read sequencing is now becoming more accessible for such resequencing work, thus allowing researchers to investigate longer indels and structural variation(see 'Sequencing' above).

The key benefit of genome resequencing over genome skimming is that it provides reliable and repeatable access to the nuclear genome. This allows researchers to investigate genome-wide diversity and evolutionary relationships both genome-wide and in specific genomic

regions of interest, such as those loci underlying species differences in young taxa (Twyford and Friedman 2015). Another benefit is that resequencing typically involves modest coverage of short-read sequence data and therefore tends to be cost effective, at least for species with small genome sizes.

The key downside to genome resequencing is that it requires a reference genome. As such, genome resequencing has traditionally been restricted to population genomic analyses of model species. However, decreasing sequencing costs and the increasing availability of reference genomes (discussed below) means resequencing is now more widely applicable to a diversity of species. It is also becoming increasingly easy to perform resequencing studies on degraded DNA due to improved laboratory and bioinformatic methods that are able to capture and process short fragments (see Chapter 2 DNA from museum collections). In addition, the increasing usability of genotype likelihoods instead of hard SNP calls, means that sequence variation can be assessed at reduced coverage (and hence, cost). As such genome resequencing is increasingly used to resolve plant identification issues that require population-level sampling or the investigation of closely related species.

The use of a reference genome brings limitations to the analysis of samples of varying quality (exacerbated with fragmentation in degraded ancient DNA; (Günther and Nettelblad 2019) or varying evolutionary distances to the reference, as the most divergent regions of the genome are likely to be misaligned or not aligned at all. This phenomenon, known as 'reference bias', can result in distorted patterns of evolutionary similarity (Sousa and Hey 2013). Resequencing studies should therefore take care in their handling of missing data and regions with low sequencing coverage, where genetic diversity may be underestimated. A solution could be to map to multiple reference genomes, or a collection of genomes resulting in a 'pangenome' (Computational Pan-Genomics Consortium 2018), but such genomic resources are not available in most plant species. Alternatively, recent advances in alignment-free methods mean that genomic analyses can be performed on sequence reads directly, without always requiring a reference genome (see Box 2). Although this circumvents the problem of reference bias, these methods may still require some form of raw read alignment to filter out contaminants.

Chapter 16: Box 2. Reference-free approaches for whole genome analysis

A range of bioinformatic tools are now available for the analysis of short sequence reads without a reference genome. These approaches often rely on the frequency distribution of k-mers (short sequences of length k) across all sequence reads of a given sample (Mapleson et al. 2017). For example, an analysis may look at the frequency of sequences 27 bp in length ('27-mers'). The k-mer profiles can be interpreted for single individuals and used to infer genetic diversity and mode of ploidy (Becher et al. 2020), or compared between samples to estimate genomic distances (Ondov et al. 2016), enabling discrimination of even closely related species. These pairwise distance matrices can be used to construct distance-based trees between samples and assess species identities or relationships, much like sequence alignments can (Bohmann et al. 2020). A downside of this method is that the 'bags of reads' used to generate k-mer profiles cannot easily shed light on individual regions of the genome, and necessarily reduces the complexity of the nuclear genome to a single metric. Further, while k-mer based approaches can be successful at coverages as low as 0.1X (Sarmashghi et al. 2019), they tend to perform best with a modest sequencing coverage where k-mer peaks from the sample can be distinguished from low-coverage contaminants (such as fungi and bacteria; (Jeong et al. 2016).

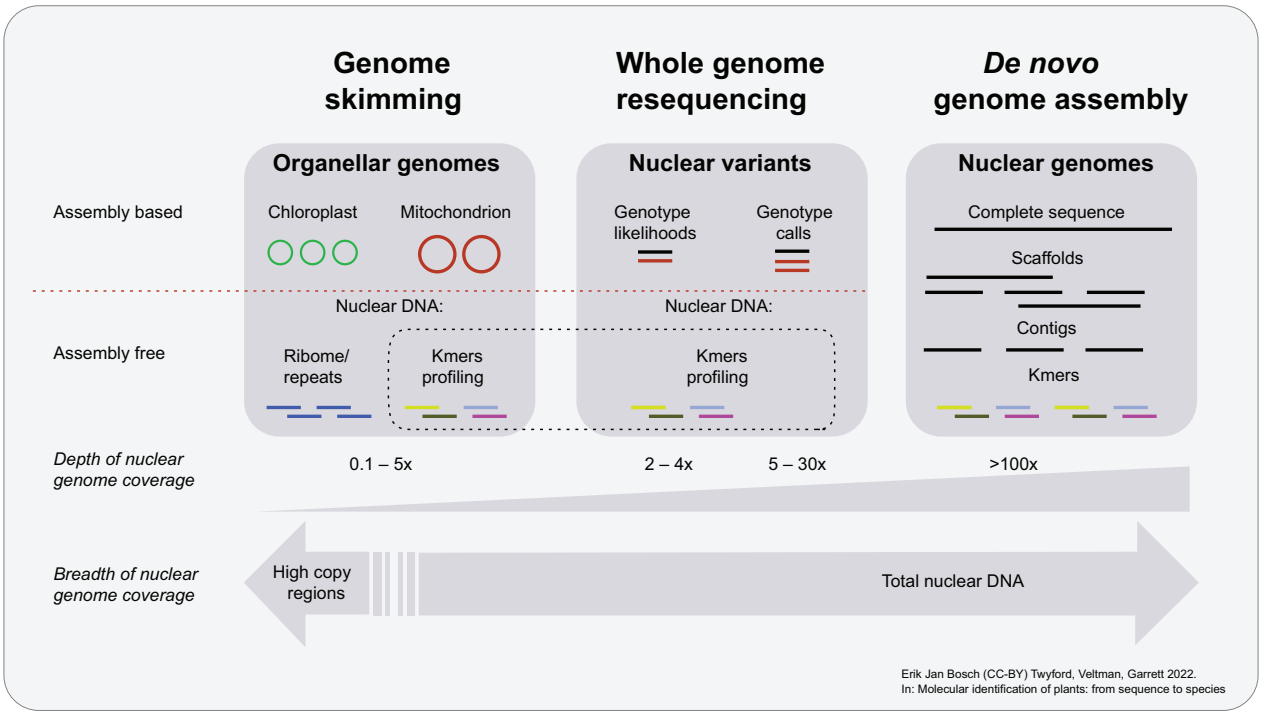

Figure 1. Infographic Chapter 16: Visual representation of the content of this chapter.

De novo whole genome sequencing

De novo whole genome sequencing represents the 'gold standard' in genomics. Here, the aim is to produce a chromosomally contiguous set of sequences that document the complete nuclear genome. To achieve this aim, a complementary range of genomic sequencing technologies (with long read sequencing coupled with long range information, now the standard) are applied to high-quality DNA extracts, producing high coverage sequencing data.

The de novo assembly of plant genomes represents a complex analytical problem. Most assemblers rely on one of two approaches (Li and Harkess 2018). Overlap Layout Consensus (OLC) algorithms identify overlaps between sequence reads, lay out a graph relating these sequences, then generate consensus sequences from the graphs. In contrast, de Bruijn Graph (DBG) approaches first separate out reads into k-mers (see Box 2), build a graph, then identify paths through the graphs to generate contigs. OLC is well suited to modest long-read datasets, where overlaps are long and where the dataset sizes are manageable. In contrast, DBG assemblers are well-placed for large short-read datasets where identifying numerous small sequence overlaps would be computationally prohibitive. After the initial assembly, further work is usually done to 'scaffold' the genome into larger sequences, using long range information such as from HiC sequence libraries.

Published reference genome assemblies vary considerably in quality. Even when contamination and mis-assemblies have been minimised, contig and scaffold size and overall genome completeness can vary widely. These genome properties can be assessed with measures such as N50 (a length-weighted median measure of contig size), or BUSCO completeness (the percentage of fully assembled core plant genes) (Manchanda et al. 2020). Even a genome of relatively low quality, with a contig N50 greater than 10 kb and a BUSCO completeness above 70% may be a useful reference for investigating the genetic structure of natural populations, while a more complete genome will be required for comparative studies of genome evolution. The current best genomes are highly complete, with megabase contigs assigned to chromosomal pseudomolecules (Li et al. 2019).

De novo genome sequencing is likely to play an important role in future studies of plant identification. One can imagine rapidly sequencing long DNA molecules that are directly assembled into chromosomes in real time, that are then compared to existing reference genomes to detect the presence of cryptic species. While this may seem like a fantasy, the dramatic and continued progress in genomic sequencing and bioinformatic algorithms makes this not so far-fetched, as seen with Oxford Nanopore adopting 'adaptive' small genome sequencing where reads are mapped and analysed in real-time. In the meantime, de novo genome sequencing effort is likely to be focused on generating reference genomes for each plant family, and for specific research projects, either as stand-alone research investigating genome evolution, or to facilitate genome resequencing of infraspecific variation. Current barriers to the wider deployment of reference genome production are the cost and bioinformatic complexity of assembling large, repeat-rich, polyploid plant genomes. These challenges are particularly difficult in some evolutionary lineages, such as ferns, which mostly have large polyploid genomes.

Considerations and example applications

Genomic sequencing can aid in numerous aspects of plant identification, discussed in Section 3. Here, we consider a representative set of examples where genome skimming, genome resequencing, and whole genome sequencing may be the preferred approaches.

Genome skimming is particularly suitable for studying a large number of diverse samples (Liu et al. 2019). One such example would be for phylogenetic analyses to understand the plant tree of life. Here, hundreds or thousands of herbarium samples from a wide diversity of plant species (e.g., representatives of all major vascular plant lineages) could be sequenced then the plastid and rDNA used to construct phylogenetic trees. This would reveal the relationship of plant groups, and allow sequence differences to be identified for other applications (e.g., for identifying the presence or absence of a given plant family in a herbal product). However, as the main regions that are recovered are conserved, such as plastid DNA, this approach is best for investigating broad-scale diversity, not the genetic differences defining 'young' species.

Genome resequencing is most appropriate for studying the population genetic structure and/or the relationship of closely related species. For example, a researcher may want to clarify species boundaries and improve species delimitation in a taxonomically complex species group (Becher et al. 2020). Genome resequencing of multiple populations from species in this group could be mapped to the reference genome, and analyses of population structure and phylogenetic relationship performed to help identify discrete entities. This may form essential baseline research for targeted conservation actions or monitoring plant trade, where geographically isolated or genetically distinct populations are subject to specific conservation measures.

The current use of reference genome sequencing is largely to understand the evolution and genome structure of plants, rather than directly being used for plant identification. For example, the production of a reference genome for a medicinal plant species may be a key resource for characterising the evolution of chemical diversity. Here, a reference genome may reveal the genes and genetic pathways involved in secondary compound production (e.g., for medicinal compounds in the orchid genus *Dendrobium* (Zhang et al. 2016)). In parallel to DNA sequencing, RNA samples are useful in genome annotation pipelines, where RNAseq data can aid in the prediction of genes in de novo reference genomes (see Chapter 15 Transcriptomics). This genome may then be used as a reference for resequencing of closely related species, to establish the presence or absence of these genes and pathways in related taxa.

Conclusion

Whole genome sequencing is increasingly used in studies of plant identification. A diverse range of methods are available, from low coverage genome skimming used to recover organelle sequences for reconstructing plant phylogeny, through to high coverage sequencing and de novo nuclear genome assembly used to generate reference genomes for comparative analyses. Future developments in sequencing technologies and bioinformatic tools will make these methods increasingly accessible to the botanical community.

Questions

1. How does sequencing coverage differ between genome skimming, genome resequencing, and reference genome production?
2. Which sequencing approach (mentioned in other chapters) would be a good alternative to genome skimming for sequencing degraded museum specimens?
3. How can genome sequencing approaches be used to facilitate plant identification? Give two examples.

Glossary

BUSCO – Benchmarking Universal Single Copy Orthologous genes used for assessing the completeness of a sequenced genome.
Contig – A single continuous sequence of DNA present in a genome assembly. Contigs in modern genome assemblies are hundreds of kilbases or multiple megabases in length.
DNA barcoding – The sequencing of few standardised DNA regions to aid in plant identification.
Genome resequencing – Low to moderate coverage sequencing of samples that are compared to a reference genome.
Genome skimming – Low coverage sequencing of genomic DNA used to assemble multi-copy regions such as plastids and mitochondria.
HMW DNA – High-molecular-weight DNA (often over 100 kb), which is required for de novo genome sequencing.
Kmer – A sequence of length k. For example, a 27-mer is the collection of all (overlapping) sequences of length 27 base pairs in a given set of sequences.
Reference genome – A high quality genome sequence from a single individual that is used as the foundation for genomic analysis.
Scaffold – An assembly of contigs separated by gaps of known length.

References

Álvarez I, Wendel JF (2003) Ribosomal ITS sequences and plant phylogenetic inference. Mol. Phylogenet. Evol. 29, 417–434. https://doi.org/10.1016/S1055-7903(03)00208-2

Becher H, Brown MR, Powell G, Metherell C, Riddiford NJ, Twyford AD (2020) Maintenance of species differences in closely related tetraploid parasitic *Euphrasia* (Orobanchaceae) on an isolated island. Plant Commun. 1, 100105. https://doi.org/10.1016/j.xplc.2020.100105

Bohmann K, Mirarab S, Bafna V, Gilbert MTP (2020) Beyond DNA barcoding: the unrealized potential of genome skim data in sample identification. Mol. Ecol. 29, 2521–2534. https://doi.org/10.1111/mec.15507

Cheng S, Melkonian M, Smith SA, Brockington S, Archibald JM, Delaux P-M, Li F-W, Melkonian B, Mavrodiev EV, Sun W, Fu Y, Yang H, Soltis DE, Graham SW, Soltis PS, Liu X, Xu X, Wong GK-S (2018) 10KP: a phylodiverse genome sequencing plan. Gigascience 7, 1–9. https://doi.org/10.1093/gigascience/giy013

Coissac E, Hollingsworth PM, Lavergne S, Taberlet P (2016) From barcodes to genomes: extending the concept of DNA barcoding. Mol. Ecol. 25, 1423–1428. https://doi.org/10.1111/mec.13549

Computational Pan-Genomics Consortium (2018) Computational pan-genomics: status, promises and challenges. Brief. Bioinformatics 19, 118–135. https://doi.org/10.1093/bib/bbw089

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43, 491–498. https://doi.org/10.1038/ng.806

Dodsworth S, Chase MW, Kelly LJ, Leitch IJ, Macas J, Novák P, Piednoël M, Weiss-Schneeweiss H, Leitch AR (2015) Genomic repeat abundances contain phylogenetic signal. Syst. Biol. 64, 112–126. https://doi.org/10.1093/sysbio/syu080

Dodsworth S (2015) Genome skimming for next-generation biodiversity analysis. Trends Plant Sci. 20, 525–527. https://doi.org/10.1016/j.tplants.2015.06.012

Garcia S, Kovařík A, Leitch AR, Garnatje T (2017) Cytogenetic features of rRNA genes across land plants: analysis of the plant rDNA database. Plant J. 89, 1020–1030. https://doi.org/10.1111/tpj.13442

Günther T, Nettelblad C (2019) The presence and impact of reference bias on population genomic studies of prehistoric human populations. PLoS Genet. 15, e1008302. https://doi.org/10.1371/journal.pgen.1008302

Jeong H, Pan J-G, Park S-H (2016) Contamination as a major factor in poor Illumina assembly of microbial isolate genomes. BioRxiv. https://doi.org/10.1101/081885

Jiao W-B, Schneeberger K (2017) The impact of third generation genomic technologies on plant genome assembly. Curr. Opin. Plant Biol. 36, 64–70. https://doi.org/10.1016/j.pbi.2017.02.002

Jin J-J, Yu W-B, Yang J-B, Song Y, dePamphilis CW, Yi T-S, Li D-Z (2020) GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biol. 21, 241. https://doi.org/10.1186/s13059-020-02154-5

Knoop V (2004) The mitochondrial DNA of land plants: peculiarities in phylogenetic perspective. Curr. Genet. 46, 123–139. https://doi.org/10.1007/s00294-004-0522-8

Kozik A, Rowan BA, Lavelle D, Berke L, Schranz ME, Michelmore RW, Christensen AC (2019) The alternative reality of plant mitochondrial DNA: One ring does not rule them all. PLoS Genet. 15, e1008373. https://doi.org/10.1371/journal.pgen.1008373

Kyriakidou M, Tai HH, Anglin NL, Ellis D, Strömvik MV (2018) Current strategies of polyploid plant genome sequence assembly. Front. Plant Sci. 9, 1660. https://doi.org/10.3389/fpls.2018.01660

Lang D, Zhang S, Ren P, Liang F, Sun Z, Meng G, Tan Y, Li X, Lai Q, Han L, Wang D, Hu F, Wang W, Liu S (2020) Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of Pacific Biosciences Sequel II system and ultralong reads of Oxford Nanopore. Gigascience 9, 1–7. https://doi.org/10.1093/gigascience/giaa123

Liu H, Wei J, Yang Ting Mu W, Song B, Yang Tuo Fu Y, Wang X, Hu G, Li W, Zhou H, Chang Y, Chen X, Chen H, Cheng L, He X, Cai H, Cai X, Wang M, Li Y, Liu X (2019) Molecular digitization of a botanical garden: high-depth whole-genome sequencing of 689 vascular plant species from the Ruili Botanical Garden. Gigascience 8, 1–9. https://doi.org/10.1093/gigascience/giz007

Li F-W, Harkess A (2018) A guide to sequence your favorite plant genomes. Appl. Plant Sci. 6, e1030. https://doi.org/10.1002/aps3.1030

Li Q, Li H, Huang W, Xu Y, Zhou Q, Wang S, Ruan J, Huang S, Zhang Z (2019) A chromosome-scale genome assembly of cucumber (*Cucumis sativus* L.). Gigascience 8, 1–10. https://doi.org/10.1093/gigascience/giz072

Lou RN, Jacobs A, Wilder AP, Therkildsen NO (2021) A beginner's guide to low-coverage whole genome sequencing for population genomics. Mol. Ecol. 30, 5966–5993. https://doi.org/10.1111/mec.16077

Manchanda N, Portwood JL, Woodhouse MR, Seetharam AS, Lawrence-Dill CJ, Andorf CM, Hufford MB (2020) GenomeQC: a quality assessment tool for genome assemblies and gene structure annotations. BMC Genomics 21, 193. https://doi.org/10.1186/s12864-020-6568-2

Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, Clavijo BJ (2017) KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics 33, 574–576. https://doi.org/10.1093/bioinformatics/btw663

Meier JI, Salazar PA, Kučka M, Davies RW, Dréau A, Aldás I, Box Power O, Nadeau NJ, Bridle JR, Rolian C, Barton NH, McMillan WO, Jiggins CD, Chan YF (2021) Haplotype tagging reveals parallel formation of hybrid races in two butterfly species. Proc Natl Acad Sci USA 118. https://doi.org/10.1073/pnas.2015005118

Michael TP, VanBuren R (2020) Building near-complete plant genomes. Curr. Opin. Plant Biol. 54, 26–33. https://doi.org/10.1016/j.pbi.2019.12.009

Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM (2016) Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 17, 132. https://doi.org/10.1186/s13059-016-0997-x

Palmer JD, Jansen RK, Michaels HJ, Chase MW, Manhart JR (1988) Chloroplast DNA variation and plant phylogeny. Ann Mo Bot Gard 75, 1180–1206. https://doi.org/10.2307/2399279

Pellicer J, Hidalgo O, Dodsworth S, Leitch IJ (2018) Genome size diversity and its impact on the evolution of land plants. Genes (Basel) 9, 88. https://doi.org/10.3390/genes9020088

Pérez-Escobar OA, Dodsworth S, Bogarín D, Bellot S, Balbuena JA, Schley RJ, Kikuchi IA, Morris SK, Epitawalage N, Cowan R, Maurin O, Zuntini A, Arias T, Serna-Sánchez A, Gravendeel B, Torres Jimenez MF, Nargar K, Chomicki G, Chase MW, Leitch IJ, Baker WJ (2021) Hundreds of nuclear and plastid loci yield novel insights into orchid relationships. Am. J. Bot. 108, 1166–1180. https://doi.org/10.1002/ajb2.1702

Rellstab C, Zoller S, Tedder A, Gugerli F, Fischer MC (2013) Validation of SNP allele frequencies determined by pooled next-generation sequencing in natural populations of a non-model plant species. PLoS ONE 8, e80422. https://doi.org/10.1371/journal.pone.0080422

Sarmashghi S, Bohmann K, P Gilbert MT, Bafna V, Mirarab S (2019) Skmer: assembly-free and alignment-free sample identification using genome skims. Genome Biol. 20, 34. https://doi.org/10.1186/s13059-019-1632-4

Sousa V, Hey J (2013) Understanding the origin of species with genome-scale data: modelling gene flow. Nat. Rev. Genet. 14, 404–414. https://doi.org/10.1038/nrg3446

Straub SCK, Parks M, Weitemier K, Fishbein M, Cronn RC, Liston A (2012) Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. Am. J. Bot. 99, 349–364. https://doi.org/10.3732/ajb.1100335

Tonti-Filippini J, Nevill PG, Dixon K, Small I (2017) What can we do with 1000 plastid genomes? Plant J. 90, 808–818. https://doi.org/10.1111/tpj.13491

Twyford AD, Friedman J (2015) Adaptive divergence in the monkey flower *Mimulus guttatus* is maintained by a chromosomal inversion. Evolution 69, 1476–1486. https://doi.org/10.1111/evo.12663

Twyford AD, Ness RW (2017) Strategies for complete plastid genome sequencing. Mol. Ecol. Resour. 17, 858–868. https://doi.org/10.1111/1755-0998.12626

Twyford AD (2018) The road to 10,000 plant genomes. Nat. Plants 4, 312–313. https://doi.org/10.1038/s41477-018-0165-2

Zeng C-X, Hollingsworth PM, Yang J, He Z-S, Zhang Z-R, Li D-Z, Yang J-B (2018) Genome skimming herbarium specimens for DNA barcoding and phylogenomics. Plant Methods 14, 43. https://doi.org/10.1186/s13007-018-0300-0

Zhang G-Q, Xu Q, Bian C, Tsai W-C, Yeh C-M, Liu K-W, Yoshida K, Zhang L-S, Chang S-B, Chen F, Shi Y, Su Y-Y, Zhang Y-Q, Chen L-J, Yin Y, Lin M, Huang H, Deng H, Wang Z-W, Zhu S-L, Liu Z-J (2016) The *Dendrobium catenatum* Lindl. genome sequence provides insights into polysaccharide synthase, floral development and adaptive evolution. Sci. Rep. 6, 19029. https://doi.org/10.1038/srep19029

Answers

1. Genome skimming typically generates low coverage (~0.1–5X), genome resequencing modest coverage (~5–30X), and de novo genome sequencing high sequencing coverage (often > 100X).
2. Target capture would be the obvious alternative (see Chapter 14 Target capture). This enriches for specific target regions which are then sequenced at high coverage.
3. There are many possible uses, two examples: (1) genome skimming could be used as an 'extended barcode' to genetically characterise a degraded, fragmented or processed sample (e.g., a museum specimen, an illegally traded processed sample, or juvenile material) relative to existing DNA barcoding databases, (2) genome resequencing could be used to characterise species limits and species relationships and identify cryptic species.

Chapter 17
Species delimitation

Anne-Sophie Quatela[1,2], Bengt Oxelman[1,2]

1 University of Gothenburg, Gothenburg, Sweden
2 Global Biodiversity Center, Gothenburg, Sweden

Anne-Sophie Quatela **anne-sophie.quatela@bioenv.gu.se**
Bengt Oxelman **bengt.oxelman@bioenv.gu.se**

Citation: Quatela A-S, Oxelman B (2022) Chapter 17. Species delimitation. In: de Boer H, Rydmark MO, Verstraete B, Gravendeel B (Eds) Molecular identification of plants: from sequence to species. Advanced Books. https://doi.org/10.3897/ab.e98875

What is a species?

Species concepts and the reality of the species category

"Species" is often considered to be one of the living world's fundamental categories, having its own ontological status, similar to a gene, cell, population, clade, or organism. Despite its importance, defining the species category is controversial, and many different species concepts have been proposed (de Queiroz 2007). Some emphasize future aspects, such as whether sexual reproduction would be possible if two organisms of different sexual types came into contact (i.e., the Biological Species Concept, BSC; Mayr 1942). Some share a common evolutionary fate (i.e., the Evolutionary Species Concept, ESC; sensu Simpson (1951) and Wiley (1978), see also, e.g., Freudenstein et al. 2017). Others emphasise the present status, for example grouping those organisms that actually mate (i.e., the modified BSC; Mayr 1969), or grouping based on the ecological niche occupied by the organisms (i.e., the Ecological Species Concept; Van Valen 1976). Still other concepts are based on the notion of species as having a single common ancestor (i.e., the Phylogenetic Species Concept; see e.g., Zink and Davis 1999), or some try to include all these aspects while emphasising the difference between grouping and ranking criteria (i.e., the Cohesion Concept; Mishler and Brandon 1987).

In this text, we will focus on concepts that view species as historical individuals. These are composed of the genetic material (i.e., an assembly of alleles), which are reproduced through time, and expressed by ephemeral phenotypes. Historical individuals refer to different ontological kinds than individual organisms. Individuals lack defining properties, they exist restricted in time and space regardless of our ability to recognize them (Ghiselin 1974). Contrary to individual organisms defined by specific features, an individual cannot be defined by its properties, but it can be recognized with some certainty. For example, King George V of Britain and Tsar Nicholas II were apparently so similar that they were mistaken for each other by members of the public at the former's wedding. Thus, their faces, haircut, beard, clothing, etc, did not define them as either George or Nicholas, but nevertheless, they existed irrespectively. By contrast, classes are unrestricted in time and space, and are defined by certain essential properties. Both George and Nicholas belonged to the class "men with beards". Both George and Nicholas were instances (apparently indistinguishable instances) of this class. In contrast, an individual is spatio-temporally restricted, and exists regardless of our ability to recognize it. It follows that there can be no instances of individuals, only parts (Ghiselin 1974). Despite a recognizable appearance, Tsar Nicholas II is not a class but an individual. There cannot be another instance of Tsar Nicholas II, though there may be any number of instances of men with beards.

Pre-evolutionary taxonomy often divided organisms into binary groups based on absence or presence of certain properties, e.g., those with the ability to move versus those that are sedentary. However, we know from phylogenetics that attributes such as the ability to move have evolved many times. Defining animals as organisms having the ability to move results in a class. Phylogenetic taxonomy tries to identify and name monophyletic groups that have a certain spatio-temporal restriction. Thus, phylogenetic taxa are individuals, with no defining properties. In taxonomy, the concept of cryptic species, although not unambiguously defined (Struck et al. 2018), refers to species that cannot readily be recognized on their phenotypic appearances, but they are different because they do not share a genealogical ancestry. The widespread recognition of the concept therefore suggests that taxonomists and other biologists often have a view of species as historical individuals, at least implicitly. In biology, diversity can be sorted in many different, potentially conflicting ways if we choose to view taxa as classes. We therefore argue in favour of a science based taxonomy that identifies hypothetical individuals. Note that this does not mean that defining certain organismal groups as classes is always meaningless. For

example, "trees" are defined as woody plants with a central trunk and a certain height, despite the fact that trees are found in many mutually exclusive clades.

The genetic material constitutes the replicators, i.e., the entity that evolves. The genes express themselves as interactors, the organisms, which are only vehicles for the genes, and everything beyond the information encoded in the DNA is an ephemeral expression of it (Dawkins and Davis 2017). It therefore follows that an evolutionary based species concept that relies on objective individuality must refer to the historical associations of vertical genetic information transmission through time. Species criteria that rely on phenotypic expressions (reproductive, ecological, physiological, etc.) must be assessed otherwise, and need not coincide with the historical individuality of a lineage or a clade. This does not mean that phenotypic traits do not contain historical information, but that these traits serve only as a proxy for tracing the replicator ancestry.

The species concept debate nourishes a fundamental ontological problem: which categories in biological taxonomy are natural or even real? According to the rules of nomenclature (e.g., the International Code of Botanical Nomenclature; Turland et al. 2018), species is one of several hierarchically arranged categories (i.e., "ranks"), along with genus, family, order, etc. We may ask how these different categories differ from one another. In these traditional codes, there is no distinction between the ranks, except that they differ in their inclusiveness and sometimes some formal requirements for names to be valid. In phylogenetic taxonomy, named taxa are clades, which are hypothetically monophyletic. Under the evolutionary paradigm, monophyletic groups exist whether we recognize them or not, and should thus be viewed as individuals, not classes. For example, angiosperms possess a large number of common, unique properties (i.e., synapomorphies) that have strongly corroborated the existence of the angiosperm clade (Judd et al. 1999), but angiosperms are not defined by these properties (e.g., flowers). However, they certainly help us recognize a plant as being an angiosperm. Similarly, our own species, *Homo sapiens*, is paradigmatic and we normally have no problems recognizing phenotypic manifestations of the genetic material connected by genealogical history to a common origin, different from other species. However, despite the alleged reality of species existing universally in nature, and its distinctness from other categories (de Queiroz 1998, 2007, 2020; Hey 2006), the species category is not unambiguously conceptualised and has until recently lacked a quantitative framework in which to view empirical data.

Monophyletic groups, or clades, form a nested hierarchy. A phylogenetic tree is a convenient model to illustrate this. A named clade with family rank will normally include subclades that may be named as genera, and these will include named species. So, if the formal categorical ranks are applied to clades, there is no difference between for example genus and species, except that the latter never can include the former. The International Code of Phylogenetic Nomenclature (PhyloCode; Cantino and de Queiroz 2020) deals with rank-free naming of clades, but not of species. Implicitly, species may here be considered as ontologically different from clades, and some of the proponents of the PhyloCode argue that species should be equated with the branches of the species tree (de Queiroz 2013). Phylogenetic methodology (e.g., Chapter 19 Systematics and evolution; Felsenstein 2004) provides scientists with a rigorous framework to test monophyly of clades with empirical data, and if species are equated with clades, it is applicable to species delimitation, although ranking by necessity will be arbitrary (Mishler and Donoghue 1982). In this chapter, we will discuss how the branches of phylogenetic trees can be conceptualised and identified using empirical genetic data.

Population is a widely used concept among biologists, which usually refers to a geographically confined assemblage of individual organisms of the same species. Strict mathematical definitions enable the parameterization of certain aspects of population genetics if some simplifying assumptions are applied. For example, demographic history can be quantitatively studied using coalescent theory (Kingman 1982) under Wright-Fisher conditions (Fisher 1930; Wright 1931), which assumes random reproduction of alleles and no selection, no migration, and no overlaps between genera-

tions. Clearly, population in this sense has an ontological status different from clades, but can it be equated with branches of a species tree? This question is also relevant if one wants to model reticulation (see "Allopolyploidization and its impact on species delimitation" from this chapter).

de Queiroz (2007) argues that all modern species concepts equate species with "separately evolving metapopulation lineages". The main difference between the concepts remains in their secondary properties, which are used to identify species. In other words, the biological species concept emphasises reproductive criteria, the ecological concept emphasises niche, etc. This implies that these properties actually mirror the genealogical history, which may not necessarily be the case. For example, the development of reproductive barriers is by definition a derived trait, whereas the absence of them is a plesiomorphy, and as such not indicative of monophyly. While we will not further address the species concept debate, or how various secondary criteria can be used to identify lineages, we can recommend a number of excellent reviews for further reading (Mayden 1997; Simpson 1951; de Queiroz 2007, 2005; Hey 2001; Mayden 1999; Stankowski and Ravinet 2021). Instead, we will consider a framework where genetic data can be used to identify clades and branches of phylogenetic trees. We leave to the reader to decide whether clades or branches are best suitable for the species category.

Practical implications of species concepts in the 21st century biodiversity crisis

The implications of species concepts affect essential societal fields such as agronomy and plant breeding, agroforestry, pharmacology and medicine, horticulture, etc. This chapter does not have the ambition nor the goal to provide an exhaustive summary of all those implications. Here, we briefly discuss the implications of different species concepts to taxonomy and how species concepts have consequences on our perception of the current biodiversity crisis.

Some species concepts and their properties (e.g., the biological species concept, the ecological species concept, the phylogenetic species concept, etc.) can be incompatible and lead to the description and naming of differently delimited taxa. To understand the consequences of competing species delimitations, it is essential to acknowledge the central role of taxonomy in many biological studies and societal matters. Traditionally, taxonomy delimits species based on diagnostic morphological differences. However, taxonomists sometimes disagree, and there is a recognition of taxonomists as "lumpers" (favouring broad species delimitations) and "splitters" (favouring narrow). Moreover, morphologically delimited species may be different from those delimited according to other species concepts. The 21st century biodiversity crisis and the conservation efforts that arise from it are in need of a tool for quantitative biodiversity measurements. Species richness is often defined as the number of species per area and/or time, and is central in many biodiversity measurements, for example in Shannon entropy (Shipley et al. 2006). Taxonomic studies contribute heavily to IUCN Red Lists (Rodrigues et al. 2006), which inventory the global conservation status of biological species. These IUCN Red Lists in turn play a significant role in determining conservation policies. This means that the tools and frameworks used for species richness assessments are concrete and assume a unified species concept. In this context, there is no room for ambiguities on what a species is, since conservation efforts and public policies must work with tangible categories. In order to make sense of species richness assessments, species are therefore assumed to be real, discrete, and quantifiable entities. However, if the concept of species does not refer to a real category in nature, Reydon (2019) recently proposed that using populations in biodiversity studies is better suited to meet modern day conservation goals. This of course also requires operational criteria to recognize pop-

ulations, but such methods are in principle available, given some simplifying assumptions and availability of data. Below, we describe some methods and how they relate to concepts typically perceived as populations, though they can also be extended to species concepts.

Essential considerations to implement a species delimitation study

Species delimitation methods described in this chapter are DNA-based phylogenetic approaches investigating the evolutionary history of species. However, their goals and inherent properties should not be confused. While molecular phylogenetics aims to identify and infer the evolutionary relationships among clades, molecular species delimitation aims to estimate parameters identifying species.

There are several practical questions to address before designing and implementing a species delimitation study, we here briefly discuss a few of them.

How to sample?

The sample strategy should reflect sufficient intra-specific variation while mirroring greater interspecific divergence. In this light, the first advice is to sample from the entire known distribution range of the group under study. The second step is to sample the different morphotypes of each taxonomic species. The underlying idea is that as the phenotype is an expression of the genotype, sequencing a wide range of morphotypes per taxonomic species should facilitate a comprehensive study. Another way of putting it is that in this way, you will be able to test taxonomic delimitations based on phenotypic data using genetic data.

How many loci to sample? Single versus multi-locus approaches

Both single and multiple locus approaches have been developed and used in plant species delimitation. In this chapter, we deliberately address multi-locus approaches. Multi-locus approaches present several advantages over single locus methods[1]. A multi-locus approach, with

1 In single-locus approaches, one gene is used to build a gene tree that will be used as an estimate of the species tree. In other words, a single gene genealogy is assumed to accurately represent the species phylogeny. Single-locus phylogenetic methods impose a strict threshold of reciprocal monophyly for delimiting species and aim to detect discontinuity in sequence variation, under the assumption that interspecific divergence exceeds intraspecific variation. Phylogenetic single locus approaches are rooted in the phylogenetic species concept: they aggregate predefined populations with unique nucleotide differences into a single species. These methods rely on the assumption that species are monophyletic (i.e., no ancestral polymorphism and sorting of alleles is complete) for the gene studied. It is assumed that discrete differences in sequence variation are observable within and between species, as a result of allele fixation in species lineages. In other words, reciprocal monophyly of alleles is assumed for the gene under study. However, reciprocal monophyly among lineages is rather improbable, particularly in recent speciation (Rosenberg 2003; Knowles and Carstens 2007). Some statistical methods are specifically designed for single locus data. The GMYC (the Generalised Mixed Yule Coalescent; Pons et al. 2006) identifies the position in a gene tree where the branching process switches from a birth/death process with constant branching probability to a coalescent process. Poisson Tree Processes model Yule coalescent transition points based on the change in substitution rates on the phylogenetic input tree (Zhang et al. 2013). The ABGD method (Puillandre et al. 2012) detects significant differences in intra- and interspecific pairwise distances (i.e., barcoding gaps). Although single locus approaches have some utility in large-scale datasets, serious concerns about its accuracy in delimiting species boundaries have been stressed.

a handful to hundreds of neutral unlinked loci, will highlight gene tree discordance. Gene tree discordance can be due to several biological phenomena that naturally occur in genealogies. Incomplete lineage sorting (ILS) is considered an important one (Edwards 2009), together with reticulation events (i.e., hybridization and introgression), which are the consequences of inter-species gene flow. While multi-locus datasets can be analysed with both concatenation[2] and coalescent-based approaches, i.e., multispecies coalescent (MSC) models, a concatenated matrix of multiple genes masks these mechanisms. All MSC-based methods handle conflicting information from multiple gene trees (such as ILS) and overcome problems associated with concatenation of multiple alignments (Edwards 2009), where all genes are assumed to follow a single common genealogy. One would perhaps think that consensus methods would overcome the problems associated with ILS, but unfortunately, this is not the case. Degnan and Rosenberg (2006) showed that under some combinations of branch lengths and population sizes, the most common gene tree topology does not reflect the species tree.

DNA-based species delimitation methods

Identifying and quantifying all the parameters that influence a biological system is complex, and in stochastic modelling we make simplifying assumptions and approximations. A stochastic model enables quantification of differences between the input data and what the model predicts. Conclusions may therefore be drawn on which processes are responsible for those differences. For example, the linear regression model has two parameters, the slope and the intercept. Given any sampled two-dimensional data, we can estimate the best fitting values for the two parameters. However, the fit will depend on the model's assumptions (e.g., linearity, random and independent sampling, homoscedasticity, etc.). These assumptions can be relaxed by introducing new parameters that will provide a better fit to the data. Although this increases the computational effort, it also reduces the explanatory power, because there will be less data per parameter.

We present two types of parametric multi-locus delimitation approaches, allelic-clustering and coalescent-based methods. Uni-locus approaches and concatenation methods do not take advantage of the information inherent to the discordance among gene trees in a multi-locus dataset. We also note that other approaches than parametric modelling are possible, for example by simply plotting data and analysing the pattern and classifying the data points according to their Euclidean distances to each other (Legendre and Fortin 2010) but these have not been extensively developed for species delimitation.

Clustering of alleles to maximise fit to linkage equilibrium

Species delimitation can be viewed as a process where sampled individuals (i.e., which can be alleles or organisms) are clustered. In population genetics, a class of methods, often referred to

2 Concatenation methods are multi-locus approaches where unlinked loci are concatenated into a supermatrix of genes. The assumption is that all genes have evolved according to the same tree, which is used as an estimate of the species tree. However, this approach oversimplifies the biological processes involved in speciation events. These processes violate the assumption of gene tree congruence across multiple loci. Incomplete lineage sorting (ILS) and gene flow are the two main phenomena responsible for gene tree discordance. Simulations bring mathematical evidence of topologically inconsistent genealogy of concatenation methods in some regions of the tree space (Kubatko and Degnan 2007). This is supported by empirical data (Jiang et al. 2020) showing that there is rarely gene tree congruence in multilocus datasets.

as STRUCTURE-like methods due the original methodology proposed by Pritchard et al. (2000), are extensively used to cluster allelic variation into bins that optimise the fit to linkage equilibrium. Under assumptions including neutrality, random mating, and that the sampled alleles are unlinked, parameters can be estimated that represent the number of groups and proportion of alleles shared between groups of individuals. Variations on this approach have been presented, e.g., to better fit natural conditions, or reduce the assumptions being made, or reduce computational complexity (e.g., ADMIXTURE; Alexander et al. 2009).

Alleles are in linkage equilibrium when they occur randomly and independently in a population, their frequency is the one expected according to the Hardy-Weinberg principle. However, biological processes often violate linkage equilibrium. Linkage disequilibrium (LD) is the non-random association of alleles at two or more loci in a population: they are in LD when they do not occur randomly and are not independent from each other. LD provides information about population genetic phenomena (i.e., migration, mutation, selection, genetic drift). In a population, LD is increased by selection, population structure, and genetic drift, and is eroded by recombination. STRUCTURE (Pritchard et al. 2000) uses clustering algorithms to maximise allele frequency fit to linkage equilibrium. Several approaches are available to assess the fit of the model, e.g., the optimal number of clusters, to identify admixtured individuals, and relax the Hardy-Weinberg assumptions.

An interesting feature of these approaches is that they can directly cluster the genetic material, the alleles, rather than the phenotypic expressions (i.e., the organisms). Thus, they are directly clustering the replicators, and not the interactors, which may be heterogeneous assemblages of such clusters (i.e., hybrids). A shortcoming of allelic clustering methods is that unlike coalescent-based phylogenetic methods they do not assess the phylogenetic divergence of populations.

Coalescent-based approaches: accounting for incomplete lineage sorting

Ideally, species delimitation methods should parametrize gene flow (i.e., migration) and incomplete lineage sorting**,** which happen when the alleles of a certain gene coalesce deeper in the species tree than the species divergence. Indeed, these two phenomena are the primary causes of gene tree discordance when sampling unlinked genes. The eukaryotic nuclear genome usually consists of several chromosomes, and within each chromosome, recombination occurs between linkage groups. By contrast, organellar genomes, which are haploid, are usually considered as non-recombining (but see e.g., Maréchal and Brisson 2010). This means that unlinked regions within individuals in a population of sexually reproducing organisms will have different genealogies. Thus, if we sample a set of unlinked loci from a number of organisms, their gene trees are expected to differ. While this may appear to be a disturbing complication, it provides useful information for both population level phenomena and phylogenetics.

Population genetics aims to understand how and why allelic frequencies vary within and between present populations. Two approaches exist for investigating ancestor-descendant relationships that centre on genetic drift. One approach is prospective/forward where the probability of identity-by-descent for allele copies (i.e., the probability that allele copies are descendants from a single common ancestor) is evaluated. Ancestor-descendant relationships are traced forward in time in order to understand the present pattern of allele copies.

The other approach is a retrospective/backward probabilistic approach called coalescent theory (Figure 1). It traces back the allele ancestry sampled in a population in order to understand present patterns of allelic frequencies, but without prior knowledge about genealogical relationship. In other words, the coalescent theory models how alleles sampled from a population can be traced to their common ancestor, providing a probabilistic framework to model population history and genealogy back in time. When alleles merge (i.e., coalesce) to a single common ancestor, it

is called a coalescence event. Coalescent theory can infer past genetic events in populations (e.g., inbreeding, gene flow, natural selection, bottlenecks) that led to the present populations. Coalescent models thus try to predict the probability of possible patterns for genealogical branching working backwards in time from the present to the most recent common ancestor (MRCA).

The basic, and most simple coalescent model, assumes that the population conforms to the idealised conditions set by the geneticists Wright (1931) and Fisher (1930) (e.g., Wright-Fisher assumptions), where there is no genetic recombination, selection, migration, or population structure, and no overlap between generations. The alleles in the present generation will trace back their parental genes by selecting parents at random. Whenever two or more alleles share a common parent, a coalescent event occurs. This means that the expected distribution of gene trees (i.e., genealogies) can be determined. In small populations, alleles will coalesce rapidly, while in larger populations it will take longer time. This also means that the gene trees on average will have short branch lengths near the tips and longer branch lengths near the root due to the exponential coalescent distribution. By empirically comparing sampled data to the expected genealogical distribution, any deviations from the input assumptions can be identified. If the assumptions are (approximately) fulfilled, the effective population size can be estimated.

Yang and Rannala (2010) showed how to connect several coalescent populations in a bifurcating tree, under what is called the multispecies coalescent (MSC) model. As MSC models coalescent events along branching lineages of a species tree, while dealing with ILS (Heled and Drummond 2010; Liu and Pearl 2007), they also make Wright-Fisher assumptions. Therefore, the probabilistic framework provided by the coalescent theory at the population level is extended to the species level as long as species are equated with Wright-Fisher populations. Nevertheless, this approach provides a mathematical definition for species defined by two parameters: the branch length (i.e., species divergence time) and the branch width (i.e., effective population size).

In the absence of migration between the tree branches, the gene tree splits will always be as old or older than the population branching. As the gene tree branching orders are random in Wright-Fisher populations, the MSC model efficiently handles incomplete lineage sorting (ILS), which is one reason why gene trees are different from species trees. Given the ambiguity of the term "species", it was perhaps unfortunate that it was used for coining the name of the model. Sukumaran et al. (2021) therefore used the term "multipopulation coalescent" instead. This may be more logical, if species are supposed to be able to include more than one lineage (i.e., branch) of an MSC tree, but given the widespread use of the term, and the ambiguous definition also of "population", we here use the MSC term, because it also highlights the different ontological status of species within this framework. It is well established as a proper name for the model. We hope, however, that the reader will be well prepared to understand the meaning of species as implied by the model (i.e., a branch of the MSC tree) and how that may differ from other conceptualisations of species.

Implementation of the multispecies coalescent model for multi-locus data

As with parametric phylogenetic methods in general, parametric phylogenetic species delimitation methods can be based on the Maximum Likelihood (ML) criterion, or on Bayesian approaches. These can be further divided into implementations that use an exact likelihood function, which estimates all parameters of the model, and approximations, where some parameters are fixed. According to Rannala et al. (2020), the only full-likelihood ML implementation for phylogenetic inference under the MSC model that exists is 3s (Yang 2002), which can accommodate three species and sequences per locus only (but thousands of loci).

Approximate likelihood ML phylogenetic methods typically work by dividing the gene tree and species tree estimation into two steps, such that gene trees obtained from phylogenetic

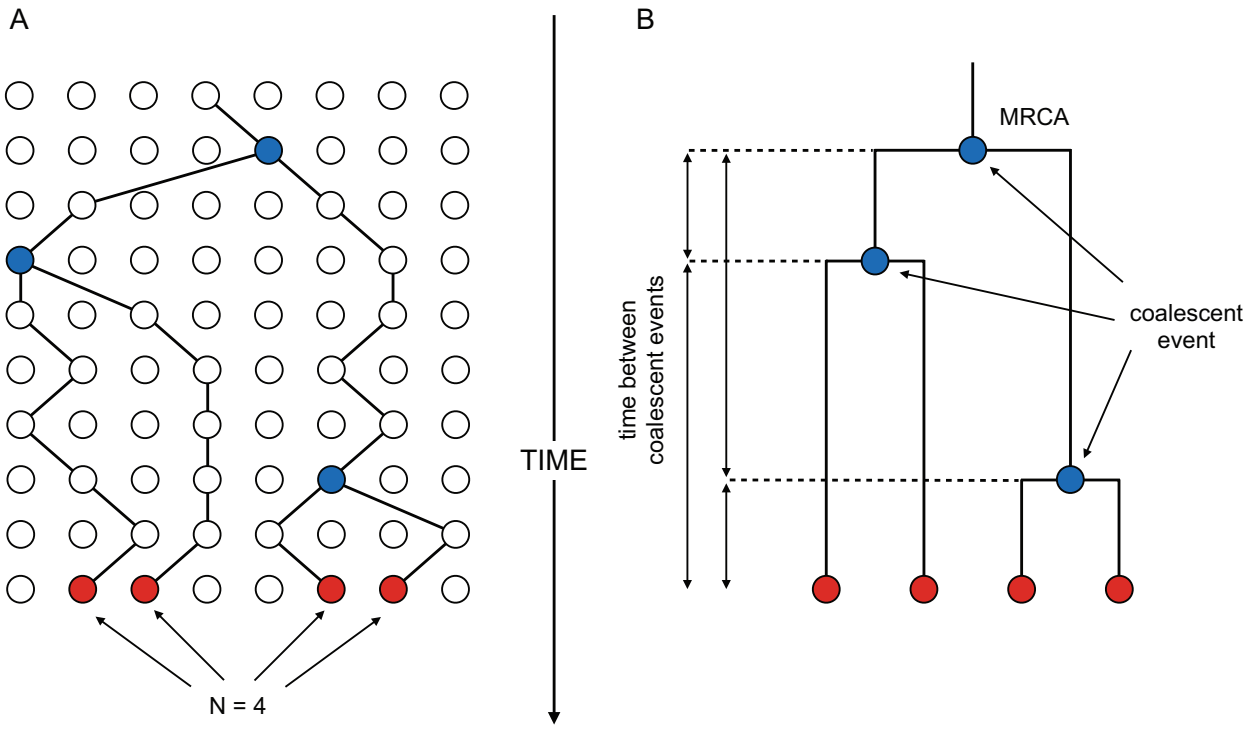

Figure 1. A. Genealogy of a sample of 4 genes (N = 4, in red) in a population of 8 genes at the present time, back to a common ancestor. **B**. Genealogy of the 4 sampled extant alleles (in red) back to the most recent common ancestor, with three coalescent events (in blue).

analyses of each alignment become input data for the species tree estimation. Thus, the gene trees are point estimates for the genealogies at each locus. In addition, the effective population size is a crucial parameter for the MSC model and finding the maximum likelihood value for it is computationally intractable (Ence and Carstens 2011; Knowles and Carstens 2007). Therefore, information about the effective population size must in practice come from other sources. Knowles and Carstens (2007) used the Akaike Information Criterion and likelihood ratio tests to determine whether a collection of gene trees better fits a single-species model or a two-species model, relying on fixed divergence times and population size. O'Meara et al. (2006) and O'Meara (2010) expanded this approach by developing methods that do not require pre-specifying the species tree and therefore simultaneously delimit and infer species under a maximum likelihood framework (Table 1). SpedeSTEM estimates the likelihood of a species tree given a collection of gene trees and uses information theory to generate metrics of comparison (Ence and Carstens 2011).

Bayesian full-likelihood implementations of the MSC model can theoretically accommodate unlimited numbers of sequences per locus, but are dependent on the approximations of the posterior densities that the Markov Chain Monte Carlo (MCMC) techniques provide. Extensive exploration of convergence and mixing are necessary to ensure that the results from MCMC are reliable (Nylander et al. 2008; Rambaut et al. 2018). Several full-likelihood Bayesian implementations exist (see Chapter 19 Systematics and evolution), and DISSECT (Jones et al. 2015), STACEY (Jones 2017a), and BP&P v.4 (Flouri et al. 2018) allow for estimation of posterior probabilities of the species delimitations, simultaneously with estimation of both species and gene trees (and other parameters of the model). These are discovery methods, which means that the user does not have to restrict the exploration to a set of predefined delimitations. By contrast, verification/validation methods only compare delimitations provided by the user. Earlier versions of BP&P (Yang and Rannala 2010) restricted the delimitations that were compatible with collapsing nodes of a user-defined guide tree. In DISSECT (Jones et al. 2015), the usual

birth-death model is replaced by a model that incorporates a spike near zero in the density for node heights, called a "birth-death-collapse model". The "collapse height" is a computational approximation of zero, which means that the dimensionality of the parameter space does not change as the number of species changes, which is a significant computational advantage. STACEY (Jones 2017a) is a computational improvement related to DISSECT, with more efficient species tree proposals, and a simplified version of the MSC model, where the population parameter is integrated out. These computational improvements have later been implemented in StarBeast2 (Ogilvie et al. 2017), and StarBeast3 (Douglas et al. 2022), which however uses a standard birth-death model for the species tree, and therefore needs correct apriori species (i.e., Wright-Fisher population) assignments. In DISSECT, STACEY, and BP&P v.4 (Table 1), there is no a priori assignment of individuals to species, and no need for a guide tree. In BP&P v.4 (Flouri et al. 2018), reversible model jumps (rjMCMC) (Green and Hastie 2009) are used to explore different delimitations (i.e., models), and the posterior probability of each will equate the frequency by which they were visited during the rjMCMC. As far as we know, few comprehensive comparisons between the methods have been made where similar results were obtained (Barley et al. 2018). DISSECT (implemented as part of BEAST1; Drummond and Rambaut 2007) or STACEY (implemented as part of BEAST2; Bouckaert et al. 2014) offer more flexibility when it comes to integration with other models (e.g., substitution models, migration parameters) than BP&P, but the latter is continuously being updated to accommodate more models (Flouri et al. 2018, 2020; https://github.com/bpp).

As the MSC model is ultimately based on a phylogenetic tree, parametric implementations of species delimitation in essence identify extant species as the tip branches of the species tree. Software implementations such as *BEAST and StarBeast2 assume that sequences are assigned to the correct species, which are defined to be Wright-Fisher (WF) populations where the gene trees are distributed according to the coalescent model. DISSECT (Jones et al. 2015), STACEY (Jones 2017a), BP&P v.4 (Flouri et al. 2018; https://github.com/bpp) overcome this problem by letting the user define minimal clusters a priori where sequences are assumed to represent the same species, which in this context means Wright-Fisher (WF) populations. In practice, the smallest minimal cluster possible will be the sequences collected from a single organism. This means that these methods in practice will be clustering organisms which should not be hybrids between WF populations. The assignments to clusters (WF species/populations) are performed simultaneously with the gene and species tree estimations during the MCMC. Sukumaran and Knowles (2017) pointed out that the units delimited by these methods therefore are what most biologists would call populations rather than species. Even if this is true, it is notable that these methods identify species as branches, and thus have different ontological status to clades. Toprak et al. (2016) argued for the utility of DISSECT in identifying minimal clades with strong support as species, assuming that the long, highly supported branches are indicative of absence of significant migration below those branches. However, certain patterns of migration may also result in misleading strong support (Leaché et al. 2014), so under such conditions, models that can accommodate migration are needed (Jones 2019).

Sukumaran et al. (2021) presented a novel MSC-based approach where the speciation completion rate is taken into account. The idea is that there is always a certain time lapse between the onset of speciation and its completion. The completion can for example be determined by the observation that all the secondary criteria (e.g., reproductive isolation, morphological distinctiveness, etc., are fulfilled; see de Queiroz 2007). By including populations, delimited by the WF methods as DISSECT/STACEY/BP&P (Table 1), where some of them can be unambiguously assigned to "good" species (e.g., according to those secondary criteria), and some which are uncertain, the software DELINEATE can take a MSC tree and use the speciation completion rate to estimate the probability that the unassigned populations belong to different partitions

("species") compatible with the MSC tree. A problem with this approach appears to be that the MSC model assumes no migration between branches.

As alternative hierarchical species delimitation models differ with respect to the assignments of sequences to species, this leads to stochastic models having different sets of parameters. To evaluate the fit of the data to different delimitations, model selection criteria are relevant. In a maximum likelihood framework, hierarchical likelihood ratio tests can be applied when models are nested (i.e., for example when the split of A and B is compared to A and B as a single species. However, such methods cannot be applied when classifications are non-nested, e.g., when AB and C is compared to A and BC. In such cases, information-theoretical approaches must be applied (Ence and Carstens 2011). In the Bayesian framework, Bayes Factors (Kass and Raftery 1995) are the natural approach for comparing models. Bayes Factors can be defined as the difference in the marginal likelihoods of competing models. The marginal likelihood is the computationally cumbersome denominator of Bayes' Theorem. Because of the computational difficulties, simplified metrics such as the AIC or BIC are sometimes applied, although they sometimes can be misleading when used in a MCMC phylogenetic framework (Susko and Roger 2020). Also, the relatively simple approximation using harmonic means from the MCMC for calculation of marginal likelihoods has been shown to be grossly misleading under some circumstances (Baele et al. 2012). Some promising improvements have been proposed, including Path Sampling and Stepping Stone methods (Baele et al. 2016). In these methods, several MCMC runs on a path between no data (prior only) to the full data set are performed. It should be noted, though, that each of these MCMCs need to converge to get reliable results. Leaché et al. (2014) proposed to use Bayes Factors to choose between a set of predefined delimitations using the MCMC method SNAPP, which takes two-state allelic data as input, and Grummer et al. (2014) and Aydin et al. (2014) have applied the procedure to multiple sequence alignments using *BEAST. Given the restricted space of delimitations, and the computational efforts needed, it is questionable whether Bayes Factor delimitations are computationally efficient compared to the discovery methods cited above.

The methods cited above assume no migration (hybridization, horizontal gene transfer) between branches, and instantaneous "speciation", i.e., divergence is completed in one generation and no migration is permitted after that. A more flexible, approximate likelihood approach to species delimitation is provided by PHRAPL (Jackson et al. 2017), which estimates the proba-

Table 1. Summary of different species delimitation methods.

Method name	Approach	Statistical framework	Input data	Likelihood function	Example of studies using the method
BP&P (Flouri et al. 2018; Rannala and Yang 2013; Yang and Rannala 2010)	discovery/ validation	Bayesian	MSA Multiple Sequence Alignment	Full likelihood	Košuthová et al. 2020
SpedeSTEM (Ence and Carstens 2011)	validation	Maximum likelihood	gene trees	Approximate likelihood	Lanna et al. 2018
Heuristic method (O'Meara 2010)	discovery	Maximum likelihood	gene trees	Approximate likelihood	O'Meara 2010
STACEY (Jones 2017a)	discovery	Bayesian	Multiple sequence alignment	Full likelihood	Tomasello 2018
DISSECT (Jones et al. 2015)	discovery	Bayesian	Multiple sequence alignment	Full likelihood	Toprak et al. 2016
PHRAPL (Jackson et al. 2017)	validation	model selection	gene trees	Approximate	Jackson et al. 2017

bility of observing a set of gene trees under a model by calculating the frequency at which observed tree topologies occur in a distribution of expected tree topologies. The relative support for a model within a set can be assessed using for example AIC. Because the method uses gene tree topologies only (excluding branch lengths), it can, relatively quickly, compare the fit of a broad range of models that include coalescence times, migration rates, and distinct/fluctuating population sizes, potentially all acting simultaneously.

Allopolyploidization and its impact on species delimitation

All the species delimitation models that we have introduced so far are developed for diploid genomes. However, allopolyploidy is traditionally thought of as being an speciation mechanism, where the allopolyploid hybrid instantaneously becomes reproductively isolated from its parents. Under this view, the problem of species delimitation becomes a problem of tracing the allopolyploidization event, and species delimitation of the descendants will follow the same logic as species delimitation of diploid genetic lineages. The models mentioned below are phylogenetic methods, which potentially can be extended in a similar fashion to the MSC-based methods described above. However, a special complication is the fact that it is usually difficult to assign sequences to subgenomes a priori.

Whole genome duplication (WGD) is ubiquitous in plants

The traditional way to model phylogenetics, and indeed also the MSC model, assumes reproductively isolated species (no migration after divergence) and bifurcating phylogenies. The genetic information is transmitted from ancestors to descendants without modelling gene flow between branches, and with bifurcations representing the speciation events.

However, hybridization and introgression are common natural processes which challenge these assumptions. Hybridization can be followed by whole genome duplication (WGD): this phenomenon is called allopolyploidization and is a significant factor in speciation due to the reproductive isolation of the newly formed polyploid from its diploid parents. Note that WGD may also occur within lineages and is then termed autopolyploidy. Here, we concentrate on the former type.

WGD is characteristic of all major land plant lineages (Clark and Donoghue 2018; Van de Peer et al. 2017).The common ancestor of all seed plants (Spermatophyta) underwent at least one round of WGD (Jiao et al. 2011). In addition, several clades of angiosperms have undergone one or several WGD events, including monocots (Jiao et al. 2011), eudicots (D'Hont et al. 2012; The French-Italian Public Consortium for Grapevine Genome Characterization 2009), Asteraceae (Barker et al. 2016; Huang et al. 2016), Brassicales (Kagale et al. 2014), legumes (Koenen et al. 2021), and grasses (Estep et al. 2014; McKain et al. 2016). This ubiquitous phenomenon shaped the evolution of plants and therefore should be considered in phylogenetic analyses. Phylogenies that incorporate hybridization and introgression can be visualised by species networks or multi-labelled trees (i.e., MUL-tree, which is a bifurcating tree in which more than one tip may be labelled with the same species (Huber and Moulton 2006), in contrast to the tree-like patterns of phylogenetic trees (Reeves and Richards 2007). This "network-like" view of evolution is called reticulate evolution. Note that it may sometimes be appropriate to incorporate hybridization and introgression as a continuous process occurring among branches of

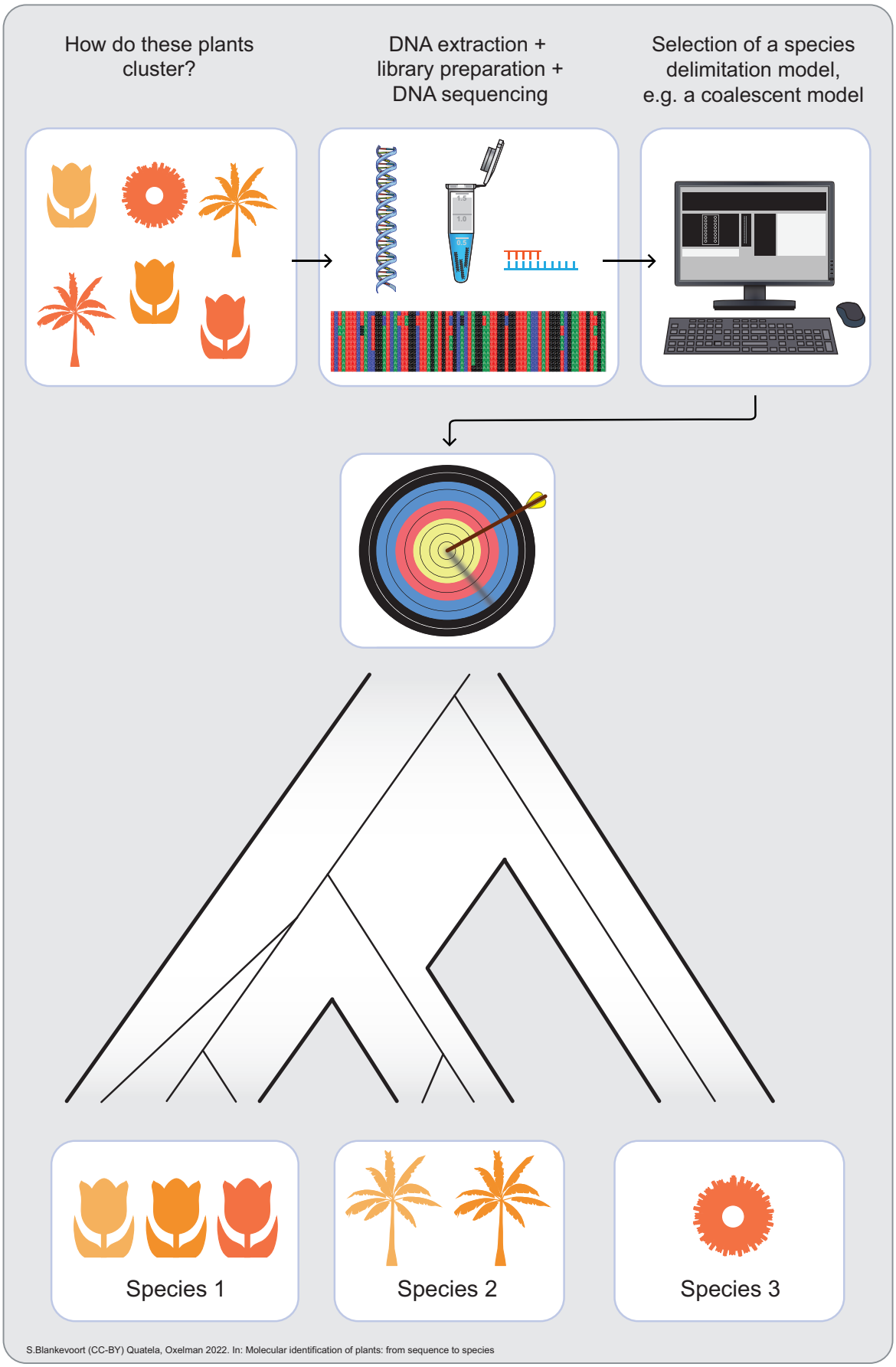

Figure 2. Chapter 17 Infographic: Visual representation of the content of this chapter.

the species trees (see Chapter 19 Systematics and evolution), thus treating the reticulations as a parameter of the model, rather than discrete, merging branches.

Reticulation events challenge the biological species concept, which states that species are different entities that cannot interbreed to produce fertile offspring. This view leads back to the philosophical perception of species and the parameters describing these entities. A hypothetical species genome undergoing several rounds of allopolyploidization will end up with subgenomes carrying genetic information difficult to trace back. In other words, this hypothetical genome would be a mosaic or a melting pot of the parental genomes. With such a changing genome, how can we identify a reticulate entity according to set parameters? Mallet et al. (2016) illustrate how reticulation influences the philosophy of species concept with the Ship of Theseus analogy, a famous metaphysical problem that questions whether an object that has had all his components replaced still remains the same entity. In this view, the Theseus ship is the genome of a reticulate species, and the wood components are DNA "chunks". The replacement of all the wood components keep the whole structure intact, which is still described as a ship. This leads to the idea that despite the chromosome rearrangements, the status of species could still be preserved. Although this analogy is greatly relevant to describe species resulting from introgressions, this is not applicable to those emerging from allopolyploidization. Allopolyploid species do not experience a full replacement of their genome over time, contrary to introgressed genomes, but only parts of it. It is essential to understand this nuance in order to grasp the subtlety of inferring boundaries between allopolyploid species. Even though reticulation can challenge the philosophical perception of species, it does not however challenge the "independently evolving lineages" definition suggested by de Queiroz (2007) that includes hybrids and allopolyploids as species. Regardless, hybridization events are not yet widely implemented in phylogenetic analysis due to both theoretical and practical challenges in resolving full haplotypes, and statistical and modelling challenges for the phylogenetic reconstruction.

Resolving haplotypes: a long-standing challenge, soon belonging to the past?

Inferring species boundaries amongst extant allopolyploid plants requires identification of the two parental sub-genomes to allow accurate inference of allopolyploid ancestral events. The genes used to trace the evolution of a polyploid genome carry information from both parental genomes. Note however, that there may have been "normal" branching speciation events after the polyploidization events, and the parental species may have gone extinct. Therefore, a phylogenetic view is necessary.

Chloroplast and mitochondrial DNA usually carry information only from one parental genome, usually the maternal lineage. Nuclear ribosomal DNA (nrDNA), which in eukaryotes contains many tandem repeats (Gonzalez and Sylvester 1995), has an interesting status since it may or may not carry information from both parents. Because of genomic homogenization processes called concerted evolution (Elder and Turner 1995), the bi-parental information can be lost. Low-copy nuclear genes are therefore a more useful source of information when inferring allopolyploidization.

In order to trace polyploid ancestry, genomes must be resolved at the haplotype level. To recover the full haplotype sequence, the DNA reads must overlap and the overlap should cover informative SNPs. Under these conditions, correct haplotype phasing can be achieved for diploid genomes (Sun et al. 2020; Zhang et al. 2020). However, this task is difficult for polyploid genomes for both theoretical (He et al. 2018) and practical reasons (Zhang et al. 2020). The space of possible haplotypes increases with the rate $O(2^{(k-1)n})$ where k denotes the number of

haplotypes for *n* SNPs (He et al. 2018). Haplotype phasing is directly dependent on sufficient SNPs from diploid genomes to link the reads. Nonetheless, using short read platforms such as Illumina to infer allopolyploidization events, especially ancient ones, often fails due to the inability of sequence reads to span across enough variants. Long reads are therefore preferred (Amarasinghe et al. 2020; Kyriakidou et al. 2018; Schrinner et al. 2020), and in order to accurately phase haplotypes, they also need to be accurate. Thus, while both Oxford Nanopore and PacBio SMRT sequencing technologies provide long reads, PacBio SMRT sequencing technologies currently provide the most accurate reads, because of the SMRTbell technique (Hon et al. 2020). SMRTbell adapters are ligated to the double stranded DNA. Primer and DNA polymerase are bonded to the SMRTbell adapter. The double stranded DNA is circularised and sequenced in repeated passes by the polymerase. Consequently, long read sequencing technologies have paved the way toward haplotype-resolved assemblies but so far tend to focus on model organisms (Kronenberg et al. 2021) and economically important species (Bhat et al. 2021). Nevertheless, as costs decrease, long read sequencing will also be used more routinely to investigate non-model organisms.

Practical implications of allopolyploidization in phylogenetics: multispecies network coalescent models

Phylogenetic methods tracing allopolyploidy aim at assigning homoeologs (i.e., subgenomes) to parental genomes. However, the task is challenging for two reasons: biological phenomena such as recombination and gene loss result in the partial loss of parental genetic information, and secondly, modelling hybridization is computationally challenging. AlloppNET (Jones 2017b; Jones et al. 2013) can account for allopolyploid hybridization, under the assumption that the parental genomes continue to evolve separately (i.e., no recombination between them) but that they are correlated regarding population size and branching events. As it is based on the MSC model, it also accounts for ILS. An alternative approach is based on the principle of Minimum Deep Coalescence (MDC) (Oberprieler et al. 2017), where alleles are assigned to subgenomes by minimising the number of deep coalescent events. The latter is thus a parsimony-based method, which takes gene trees as input data, whereas AlloppNET is a fully parameterized implementation of the MSC model that takes multiple sequence alignments as input data (Oxelman et al. 2017).

AlloppNET is implemented in the BEAST1 framework (Suchard et al. 2018). While the initial model only allowed a single hybridization (Jones et al. 2013), the extended implementation does not restrict the number of hybridization events (Jones 2017b). Since the number of hybridizations is not known a priori, it is a variable that needs to be parameterized, and the parameters are estimated by sampling from the posterior distribution using the reversible-jump MCMC algorithm. The model includes a parameter which assigns the alleles to subgenomes during the MCMC. A limitation is that only diploids and allotetraploids can be considered, while higher ploidy (e.g., hexaploid, octoploid, decaploid, etc.) levels are known across extant plants and their ancestors. Paleohexaploid ancestors have been inferred in Asteraceae (Barker et al. 2008; Truco et al. 2013), in Solanaceae (Tomato Genome Consortium 2012), and in the family Brassicaceae (Lagercrantz and Lydiate 1996; Lysak et al. 2005; Tang et al. 2012). Wild strawberries are the octoploid result of natural hybrids (Edger et al. 2019). The cosmopolitan reeds *Phragmites australis* display intra-specific ploidy variation (i.e., 3x, 4x, 6x, 7x, 8x, 11x, 12x, x = 12; Gorenflot 1976).

The MDC approach by Oberprieler et al. (2017) is based on the parsimony principle and uses a permutation strategy. It builds a MUL-species tree in two steps, the first using permu-

tations to assign alleles to subgenomes, and a second where the species tree is built. In each step the parsimony principle of minimising the number of deep coalescences is employed, and the resulting MUL-tree is converted to a network using the PADRE algorithm (see below). An alternative, faster approach was recently proposed by (Yan et al. 2022) and implemented in the PhyloNet software (Than and Nakhleh 2009; Than et al. 2008; Wen et al. 2018). The first step is performed on all possible parental allele pairs (diploid parents), and runs a MDC tree analysis on those. In the second permutation step, the assignment of polyploid alleles with the lowest number of deep coalescences in the species tree is kept. For each polyploid accession, allele pair combinations across loci are submitted to a species tree based on all gene trees. This step is repeated for all possible allele group combinations across loci. As in the first step, the allele pair combinations across loci that result in a species tree with the minimum number of deep coalescences is kept. These two steps are repeated for all polyploid accessions individually.

PADRE (Package for Analysing and Displaying Reticulate Evolution; Lott et al. 2009) enables transformation of multiple labelled subgenome trees (MUL-trees) to phylogenetic networks. The algorithm used for constructing a phylogenetic network is described in (Huber et al. 2006). The basic idea is to recursively identify maximal isomorphic subtrees within the input MUL-tree, and to merge these until the labels occur only once on the tips of the resulting network. At each iteration, the user has the option to combine or keep separate the identified isomorphic subtrees. This option can be used to incorporate additional biological information as part of the network construction process. Most importantly, the user can accept the merging of isomorphic subtrees identified at every iteration. The resulting phylogenetic network is then guaranteed to have a minimum number of reticulation nodes, where each node represents hybridization of ancestral lineages (see Huber and Moulton (2006) for more details).

Conclusions

Conceptually, various species concepts attempt to accommodate genealogical, phenotypic as well as future aspects, and these need not lead to identical delimitations. There is an emerging view of viewing species as the branches of the phylogenetic tree, and we have focused on species as being historical individuals composed of the vertically transmitted genetic information. The MSC model allows scientists to view genetic data and rigorously test monophyly as well as branch content. However, most current implementations of the MSC model identify the branches as what most biologists would view as populations, and furthermore, they are not capable of including migrations of alleles. STRUCTURE-like methods have the capability to cluster alleles directly, but are dependent on similar assumptions to the MSC model, and lack a phylogenetic component. In principle, the MSC model can be extended to accommodate migrations, and a few recent attempts exist (e.g., DENIM; PHRAPL). A pluridisciplinary approach, involving genomic and evolutionary concepts implemented in a powerful statistical framework is anticipated for future progress. Beyond its importance for biology internally, species delimitation has important societal implications. The current "sixth mass extinction" calls for implementing conservation programs that use appropriate species richness assessments and species definitions in order to accurately measure and limit biodiversity loss. The necessity to agree on a given species definition in a given context (i.e., biodiversity erosion) does not in itself solve the ontological question "what is a species?". The fast-moving next-generation sequencing technologies disclose the necessary genomic information to study virtually any taxonomic group, but there is an urgent need for conceptual development as well as suitable models with sufficient biological realism to view the data in.

Questions

1. According to the latest taxonomic revision (based on morphology) a certain taxon includes 13 species, four species are tetraploid, and the others are diploid. You can afford to sample and generate DNA sequences from multiple loci of 96 individual plants. You want to test whether the taxonomy corresponds to a coalescent-based species delimitation. Given that the assumptions of the MSC model are fairly well met by your data, what properties will the species delimited by STACEY have? Which species concept would be the most relevant, if any, and why?
2. The 13 taxonomic species in the previous example cover a very large area of the Northern Hemisphere. Your MSC-based analysis identified most of your 96 sampled individuals as separate species. In most cases, these are only separated based on the DNA information you have sampled. On the other hand, all four tetraploid species are resolved as having separate allopolyploid origins, with some divergence after these. The diploid taxonomic species form moderately to well supported clades.
   2a. Which assumptions would these results be based on? What could have violated those, in terms of sampling and biological processes not accounted for in your model?
   2b. Given that you trust your results, which taxonomic decisions should be made, if any? Give arguments for different scenarios.
3. You receive a dataset with DNA sequences covering 10 different populations with 50 individuals each. Presumably, they all belong to the same taxonomic species. However, different phenotypes are observed in some populations. You want to know if they belong to the same species based on genetic data but you are also interested in the genetic structure of the population. However, the phylogenetic divergent time is not your first interest. Which class of methods would you use?

Glossary

AIC / BIC – AIC stands for Akaike Information Criterion and BIC for Bayesian Information Criterion. They are estimators of the quality of statistical models for a given set of data, providing a means for model selection.

Allopolyploidy – Inheritable condition of having more than two sets of chromosomes after hybridization. Typically, allopolyploids have disomic inheritance, meaning that there is bivalent pairing of chromosomes during meiosis. Also called "whole genome duplication".

Autopolyploidy – Inheritable condition of having more than two sets of chromosomes received from a single ancestral taxon, by opposition to allopolyploidy. Autopolyploids have polysomic inheritance, meaning that there is multivalent pairing of chromosomes during meiosis.

Ancestral polymorphism – Genetic variation in a species that arose prior to speciation. Synonymous with "deep coalescence".

Anomalous zone of the MSC – Degnan and Rosenberg (2006) define it as a set of short internal branches in species trees that will generate gene trees that are discordant with the species tree more often than concordant.

Bayes Factors – The ratio of the marginal likelihoods of a given parametric model to another one. It can be interpreted as a measure of the weight of an hypothesis compared to another one.

Bayesian approaches – Based on the Bayes theorem, describing the probability of an event based on prior knowledge of conditions related to this event.

Bifurcating tree – A graph where branches (edges) give rise to daughter branches, and never merge.

Biodiversity – Association of the two words "biological" and "diversity". It refers to the variety of life that is found on Earth.

Birth-death model – A continuous-time Markov process with two parameters, births and deaths. In a phylogenetic context, this translates to branching events being births, and extinctions being deaths. A birth-death model for a phylogenetic tree will in its most simple form have constant probabilities for branching events and extinctions.

Clade – A group of taxa that are monophyletic - composed of a common ancestor and all its descendants - on a phylogenetic tree.

Class – A grouping of entities based on defined criteria.

Coalescent theory – Models how alleles in a population have originated from a common ancestor. In its most simple form, it assumes no recombination, no selection, no gene flow, and no population structure. This implies that each allele is equally likely to have been passed on from one generation to the next. The model looks backwards in time, merging alleles into coalescence events according to a random process.

Cryptic species – Term referring to species that cannot readily be distinguished morphologically.

Deep coalescence – When two or more alleles of the same species have their most recent common ancestor in an ancestral species. Synonymous with "ancestral polymorphism".

Discovery methods – Species delimitation methods that do not require pre-defined delimitations to assess, as opposed to verification/validation methods.

Epistemology (in the context of biology) – Concerns the theory of knowledge. For example, how can we know what a species is, is an epistemological question.

Gene flow – The transfer of genetic material from one population to another.

Gene tree – Phylogenetic tree of a gene.

Gene tree discordance – When gene trees from the same set of organisms are different in topology and/or branch lengths.

Genetic drift – Variation in allele frequency in a small population due to random factors

Genotype – The complete set of genetic, inheritable information (DNA) of an organism.

Hardy-Weinberg principle – A population genetics principle, also known as the Hardy-Weinberg equilibrium/model/theorem/law, which states that allele frequencies in a large population will remain constant from generation to generation in the absence of genetic drift and non-random evolutionary factors.

Hierarchical likelihood ratio tests – statistical tests estimating which model is the best fit to a dataset among two models. The competing models must be hierarchically nested. The more complex model must differ from the simpler one by one or more additional parameters.

Historical individual – Refers to an assembly of alleles that are reproduced through time.

Hybridization – Interspecific breeding.

Identity-by-descent – identical nucleotide sequences in two or more individuals inherited by a common ancestor, without recombination. The identical segment has the same origin among these individuals.

Incomplete lineage sorting – A phenomenon in population genetics when ancestral copies of alleles fail to coalesce into a common ancestral copy until deeper/older than previous speciation events. See also ancient polymorphism and deep coalescence, which refer to the same thing.

Interactor – Term defined in an evolutionary context by the biologist Richard Dawkins. Referring to organisms as being ephemeral vehicles for genes (the replicators).

Introgression – Also known as introgressive hybridization. It is the transfer of genetic material from one species into the gene pool of another one by repeated backcrossing of an interspecific hybrid with one of its parent species.
Linkage disequilibrium – Non-random association of alleles at two or more unlinked loci in a population.
Linear regression model – A linear approach to modelling the relationship between a scalar response and one or more explanatory variables.
Markov Chain Monte Carlo – In statistics, methods that sample from a probability distribution by constructing a Markov chain.
Markov chain – A random process describing a sequence of possible states or events where each state/event depends only on the previous one, independently from older ones.
Maximum Likelihood approaches – Estimation of parameters of an expected probability distribution given observed data. The estimated parameters are those that make the observed data most probable.
Metapopulation – A group of spatially separated populations of the same species which interact at some level.
Migration – Gene flow, including phenomena such as hybridization, introgression, horizontal gene transfer.
Monophyly – The condition in which a group of taxa composed only of a common ancestor and all its lineal descendants form a single clade.
Ontology (in the context of biology) – The field that divides living things into categories to better understand them and how they fit into the world. For example, the ontological nature of "species" answers the question "what is a species?".
Phenotype – The expression of the genotype modified by environmental factors.
Plesiomorphy – In phylogenetics, a plesiomorphy is an ancestral state character.
Polyploidization – Event that creates more than two copies of the entire genome of a taxon.
Population – In general, individuals belonging to the same species that live in the same geographic area at the same time. The effective population, by contrast, consists of those involved in the reproduction to the next generation.
Reciprocal monophyly – When two sets of taxa form exclusive clades.
Replicator – Term defined in evolutionary context by the biologist Richard Dawkins. Genetic material that evolves and replicates. Also see interactor.
Reticulation – In phylogenetics, a reticulation is when a lineage originates by the merging of two ancestral lineages.
Speciation completion rate – Parameter describing the transition from a parent to a full independent species. Describes a protracted speciation, in opposition to instantaneous speciation in the birth-death model.
Species richness – The number of different species represented in an ecological community, landscape, or region.
Species tree – Phylogenetic tree representing the evolutionary relationships among species.
Stochastic modelling – A way of describing a certain set of random parameters with their associated probability distributions.
Synapomorphy – Derived traits shared by a group of taxa due to their inheritance from a common ancestor.
Taxon (plural 'taxa') – A set of genotypes and their associated expressed phenotypes that are formally recognized.
Taxonomy – The branch of science where biological taxa are described, named, and identified.
Validation/verification methods – Species delimitation methods that require a subset of predefined delimitations, in contrast to discovery methods, which consider all possibilities.

References

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19, 1655–1664. https://doi.org/10.1101/gr.094052.109

Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q (2020) Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 21, 30. https://doi.org/10.1186/s13059-020-1935-5

Aydin Z, Marcussen T, Ertekin AS, Oxelman B (2014) Marginal likelihood estimate comparisons to obtain optimal species delimitations in *Silene* sect. *Cryptoneurae* (Caryophyllaceae). PLoS ONE 9, e106990. https://doi.org/10.1371/journal.pone.0106990

Baele G, Lemey P, Bedford T, Rambaut A, Suchard MA, Alekseyenko AV (2012) Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. Mol. Biol. Evol. 29, 2157–2167. https://doi.org/10.1093/molbev/mss084

Baele G, Lemey P, Suchard MA (2016) Genealogical working distributions for Bayesian model testing with phylogenetic uncertainty. Syst. Biol. 65, 250–264. https://doi.org/10.1093/sysbio/syv083

Barker MS, Kane NC, Matvienko M, Kozik A, Michelmore RW, Knapp SJ, Rieseberg LH (2008) Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. Mol. Biol. Evol. 25, 2445–2455. https://doi.org/10.1093/molbev/msn187

Barker MS, Li Z, Kidder TI, Reardon CR, Lai Z, Oliveira LO, Scascitelli M, Rieseberg LH (2016) Most Compositae (Asteraceae) are descendants of a paleohexaploid and all share a paleotetraploid ancestor with the Calyceraceae. Am. J. Bot. 103, 1203–1211. https://doi.org/10.3732/ajb.1600113

Barley AJ, Brown JM, Thomson RC (2018) Impact of model violations on the inference of species boundaries under the multispecies coalescent. Syst. Biol. 67, 269–284. https://doi.org/10.1093/sysbio/syx073

Bhat JA, Yu D, Bohra A, Ganie SA, Varshney RK (2021) Features and applications of haplotypes in crop breeding. Commun. Biol. 4, 1266. https://doi.org/10.1038/s42003-021-02782-y

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, Suchard MA, Rambaut A, Drummond AJ (2014) BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput. Biol. 10, e1003537. https://doi.org/10.1371/journal.pcbi.1003537

Cantino PD, de Queiroz K (2020) International Code of Phylogenetic Nomenclature (PhyloCode), 6th ed. CRC Press, Boca Raton. https://doi.org/10.1201/9780429446320

Clark JW, Donoghue PCJ (2018) Whole-Genome Duplication and Plant Macroevolution. Trends Plant Sci. 23, 933–945. https://doi.org/10.1016/j.tplants.2018.07.006

D'Hont A, Denoeud F, Aury J-M, Baurens F-C, Carreel F, Garsmeur O, Noel B, Bocs S, Droc G, Rouard M, Da Silva C, Jabbari K, Cardi C, Poulain J, Souquet M, Labadie K, Jourda C, Lengellé J, Rodier-Goud M, Alberti A, Wincker P (2012) The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. Nature 488, 213–217. https://doi.org/10.1038/nature11241

Dawkins R, Davis N (2017) The selfish gene. Macat Library. https://doi.org/10.4324/9781912281251

Degnan JH, Rosenberg NA (2006) Discordance of species trees with their most likely gene trees. PLoS Genet. 2, e68. https://doi.org/10.1371/journal.pgen.0020068

de Queiroz K (1998) The general lineage concept of species, species criteria, and the process of speciation, in: Howard, D.J., Berlocher, S.H. (Eds.), Endless Forms: Species and Speciation. Oxford University Press, pp. 57–75.

de Queiroz K (2005) A unified concept of species and its consequences for the future of taxonomy. Proceedings of the California Academy of Sciences, ser. 4 56(Suppl. I), 196–215.

de Queiroz K (2007) Species concepts and species delimitation. Syst. Biol. 56, 879–886. https://doi.org/10.1080/10635150701701083

de Queiroz K (2013) Nodes, branches, and phylogenetic definitions. Syst. Biol. 62, 625–632. https://doi.org/10.1093/sysbio/syt027

de Queiroz K (2020) An updated concept of subspecies resolves a dispute about the taxonomy of incompletely separated lineages. Herpetological Review 51, 459–461.

Douglas J, Jiménez-Silva CL, Bouckaert R (2022) StarBeast3: adaptive parallelised Bayesian inference under the multispecies coalescent. Syst. Biol. https://doi.org/10.1093/sysbio/syac010

Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7, 214. https://doi.org/10.1186/1471-2148-7-214

Edger PP, Poorten TJ, VanBuren R, Hardigan MA, Colle M, McKain MR, Smith RD, Teresi SJ, Nelson ADL, Wai CM, Alger EI, Bird KA, Yocca AE, Pumplin N, Ou S, Ben-Zvi G, Brodt A, Baruch K, Swale T, Shiue L, Knapp SJ (2019) Origin and evolution of the octoploid strawberry genome. Nat. Genet. 51, 541–547. https://doi.org/10.1038/s41588-019-0356-4

Edwards SV (2009) Is a new and general theory of molecular systematics emerging? Evolution 63, 1–19. https://doi.org/10.1111/j.1558-5646.2008.00549.x

Elder JF, Turner BJ (1995) Concerted evolution of repetitive DNA sequences in eukaryotes. Q. Rev. Biol. 70, 297–320. https://doi.org/10.1086/419073

Ence DD, Carstens BC (2011) SpedeSTEM: a rapid and accurate method for species delimitation. Mol. Ecol. Resour. 11, 473–480. https://doi.org/10.1111/j.1755-0998.2010.02947.x

Estep MC, McKain MR, Vela Diaz D, Zhong J, Hodge JG, Hodkinson TR, Layton DJ, Malcomber ST, Pasquet R, Kellogg EA (2014) Allopolyploidy, diversification, and the Miocene grassland expansion. Proc Natl Acad Sci USA 111, 15149–15154. https://doi.org/10.1073/pnas.1404177111

Felsenstein J (2004) Inferring phylogenies, 2nd ed. Sinauer Associates, Sunderland, Massachusetts.

Fisher RA (1930) The genetical theory of natural selection. Clarendon Press, Oxford. https://doi.org/10.5962/bhl.title.27468

Flouri T, Jiao X, Rannala B, Yang Z (2018) Species tree inference with BPP using genomic sequences and the multispecies coalescent. Mol. Biol. Evol. 35, 2585–2593. https://doi.org/10.1093/molbev/msy147

Flouri T, Jiao X, Rannala B, Yang Z (2020) A Bayesian implementation of the multispecies coalescent model with introgression for phylogenomic analysis. Mol. Biol. Evol. 37, 1211–1223. https://doi.org/10.1093/molbev/msz296

Freudenstein JV, Broe MB, Folk RA, Sinn BT (2017) Biodiversity and the species concept - Lineages are not enough. Syst. Biol. 66, 644–656. https://doi.org/10.1093/sysbio/syw098

Ghiselin MT (1974) A radical solution to the species problem. Syst. Biol. 23, 536–544. https://doi.org/10.1093/sysbio/23.4.536

Gonzalez IL, Sylvester JE (1995) Complete sequence of the 43-kb human ribosomal DNA repeat: analysis of the intergenic spacer. Genomics 27, 320–328. https://doi.org/10.1006/geno.1995.1049

Gorenflot R (1976) Le complexe polyploïde du *Phragmites australis* (Cav.) Trin. ex Steud. (= *P. communis* Trin.). Bulletin de la Société Botanique de France 123, 261–271. https://doi.org/10.1080/00378941.1976.10835694

Green PJ, Hastie DI (2009) Reversible jump MCMC. Genetics 155.

Grummer JA, Bryson RW, Reeder TW (2014) Species delimitation using Bayes factors: simulations and application to the *Sceloporus scalaris* species group (Squamata: Phrynosomatidae). Syst. Biol. 63, 119–133. https://doi.org/10.1093/sysbio/syt069

Heled J, Drummond AJ (2010) Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27, 570–580. https://doi.org/10.1093/molbev/msp274

Hey J (2001) The mind of the species problem. Trends Ecol. Evol. 16, 326–329. https://doi.org/10.1016/s0169-5347(01)02145-0

Hey J (2006) On the failure of modern species concepts. Trends Ecol. Evol. 21, 447–450. https://doi.org/10.1016/j.tree.2006.05.011

He D, Saha S, Finkers R, Parida L (2018) Efficient algorithms for polyploid haplotype phasing. BMC Genomics 19, 110. https://doi.org/10.1186/s12864-018-4464-9

Hon T, Mars K, Young G, Tsai Y-C, Karalius JW, Landolin JM, Maurer N, Kudrna D, Hardigan MA, Steiner CC, Knapp SJ, Ware D, Shapiro B, Peluso P, Rank DR (2020) Highly accurate long-read HiFi sequencing data for five complex genomes. Sci. Data 7, 399. https://doi.org/10.1038/s41597-020-00743-4

Huang C-H, Zhang C, Liu M, Hu Y, Gao T, Qi J, Ma H (2016) Multiple polyploidization events across Asteraceae with two nested events in the early history revealed by nuclear phylogenomics. Mol. Biol. Evol. 33, 2820–2835. https://doi.org/10.1093/molbev/msw157

Huber KT, Moulton V (2006) Phylogenetic networks from multi-labelled trees. J. Math. Biol. 52, 613–632. https://doi.org/10.1007/s00285-005-0365-z

Huber KT, Oxelman B, Lott M, Moulton V (2006) Reconstructing the evolutionary history of polyploids from multilabeled trees. Mol. Biol. Evol. 23, 1784–1791. https://doi.org/10.1093/molbev/msl045

Jackson ND, Morales AE, Carstens BC, O'Meara BC (2017) PHRAPL: phylogeographic inference using approximate likelihoods. Syst. Biol. 66, 1045–1053. https://doi.org/10.1093/sysbio/syx001

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, Soltis DE, Clifton SW, Schlarbaum SE, Schuster SC, Ma H, Leebens-Mack J, dePamphilis CW (2011) Ancestral polyploidy in seed plants and angiosperms. Nature 473, 97–100. https://doi.org/10.1038/nature09916

Jones G, Aydin Z, Oxelman B (2015) DISSECT: an assignment-free Bayesian discovery method for species delimitation under the multispecies coalescent. Bioinformatics 31, 991–998. https://doi.org/10.1093/bioinformatics/btu770

Jones G, Sagitov S, Oxelman B (2013) Statistical inference of allopolyploid species networks in the presence of incomplete lineage sorting. Syst. Biol. 62, 467–478. https://doi.org/10.1093/sysbio/syt012

Jones G (2017a) Algorithmic improvements to species delimitation and phylogeny estimation under the multispecies coalescent. J. Math. Biol. 74, 447–467. https://doi.org/10.1007/s00285-016-1034-0

Jones G (2017b) Bayesian phylogenetic analysis for diploid and allotetraploid species networks. BioRxiv. https://doi.org/10.1101/129361

Jones G (2019) Divergence estimation in the presence of incomplete lineage sorting and migration. Syst. Biol. 68, 19–31. https://doi.org/10.1093/sysbio/syy041

Judd WS, Campbell CS, Kellogg EA, Teven PF (1999) Plant systematics: a phylogenetic approach. Sinauer Associates.

Kagale S, Robinson SJ, Nixon J, Xiao R, Huebert T, Condie J, Kessler D, Clarke WE, Edger PP, Links MG, Sharpe AG, Parkin IAP (2014) Polyploid evolution of the Brassicaceae during the Cenozoic era. Plant Cell 26, 2777–2791. https://doi.org/10.1105/tpc.114.126391

Kass RE, Raftery AE (1995) Bayes Factors. J. Am. Stat. Assoc. 90, 773–795. https://doi.org/10.1080/01621459.1995.10476572

Kingman JFC (1982) The coalescent. Stoch. Process. Their Appl. 13, 235–248. https://doi.org/10.1016/0304-4149(82)90011-4

Knowles LL, Carstens BC (2007) Delimiting species without monophyletic gene trees. Syst. Biol. 56, 887–895. https://doi.org/10.1080/10635150701701091

Koenen EJM, Ojeda DI, Bakker FT, Wieringa JJ, Kidner C, Hardy OJ, Pennington RT, Herendeen PS, Bruneau A, Hughes CE (2021) The origin of the legumes is a complex paleopolyploid phylogenomic tangle closely associated with the Cretaceous-Paleogene (K-Pg) mass extinction event. Syst. Biol. 70, 508–526. https://doi.org/10.1093/sysbio/syaa041

Košuthová A, Bergsten J, Westberg M, Wedin M (2020) Species delimitation in the cyanolichen genus *Rostania*. BMC Evol. Biol. 20, 115. https://doi.org/10.1186/s12862-020-01681-w

Kronenberg ZN, Rhie A, Koren S, Concepcion GT, Peluso P, Munson KM, Porubsky D, Kuhn K, Mueller KA, Low WY, Hiendleder S, Fedrigo O, Liachko I, Hall RJ, Phillippy AM, Eichler EE, Williams JL, Smith TPL, Jarvis ED, Sullivan ST, Kingan SB (2021) Extended haplotype-phasing of long-read de novo genome assemblies using Hi-C. Nat. Commun. 12, 1935. https://doi.org/10.1038/s41467-020-20536-y

Kyriakidou M, Tai HH, Anglin NL, Ellis D, Strömvik MV (2018) Current strategies of polyploid plant genome sequence assembly. Front. Plant Sci. 9, 1660. https://doi.org/10.3389/fpls.2018.01660

Lagercrantz U, Lydiate DJ (1996) Comparative genome mapping in *Brassica*. Genetics 144, 1903–1910. https://doi.org/10.1093/genetics/144.4.1903

Lanna FM, Werneck FP, Gehara M, Fonseca EM, Colli GR, Sites JW, Rodrigues MT, Garda AA (2018) The evolutionary history of *Lygodactylus* lizards in the South American open diagonal. Mol. Phylogenet. Evol. 127, 638–645. https://doi.org/10.1016/j.ympev.2018.06.010

Leaché AD, Fujita MK, Minin VN, Bouckaert RR (2014) Species delimitation using genome-wide SNP data. Syst. Biol. 63, 534–542. https://doi.org/10.1093/sysbio/syu018

Legendre P, Fortin M-J (2010) Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. Mol. Ecol. Resour. 10, 831–844. https://doi.org/10.1111/j.1755-0998.2010.02866.x

Liu L, Pearl DK (2007) Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst. Biol. 56, 504–514. https://doi.org/10.1080/10635150701429982

Lott M, Spillner A, Huber KT, Moulton V (2009) PADRE: a package for analyzing and displaying reticulate evolution. Bioinformatics 25, 1199–1200. https://doi.org/10.1093/bioinformatics/btp133

Lysak MA, Koch MA, Pecinka A, Schubert I (2005) Chromosome triplication found across the tribe Brassiceae. Genome Res. 15, 516–525. https://doi.org/10.1101/gr.3531105

Mallet J, Besansky N, Hahn MW (2016) How reticulated are species? Bioessays 38, 140–149. https://doi.org/10.1002/bies.201500149

Maréchal A, Brisson N (2010) Recombination and the maintenance of plant organelle genome stability. New Phytol. 186, 299–317. https://doi.org/10.1111/j.1469-8137.2010.03195.x

Mayden RL (1997) A hierarchy of species concepts: the denouement in the saga of the species problem, in: Claridge, M.F., Dawah, H.A., Wilson, M.R. (Eds.), Species: The Units of Diversity. Chapman & Hall, pp. 381–423.

Mayden RL (1999) Consilience and a hierarchy of species concepts: advances toward closure on the species puzzle. J. Nematol. 31, 95–116.

Mayr E (1942) Systematics and the origin of species from the viewpoint of a zoologist. Columbia University Press, New York.

Mayr E (1969) The biological meaning of species*. Biological Journal of the Linnean Society 1, 311–320. https://doi.org/10.1111/j.1095-8312.1969.tb00123.x

McKain MR, Tang H, McNeal JR, Ayyampalayam S, Davis JI, dePamphilis CW, Givnish TJ, Pires JC, Stevenson DW, Leebens-Mack JH (2016) A phylogenomic assessment of ancient polyploidy and genome evolution across the Poales. Genome Biol. Evol. 8, 1150–1164. https://doi.org/10.1093/gbe/evw060

Mishler BD, Brandon RN (1987) Individuality, pluralism, and the phylogenetic species concept. Biol. Philos. 2, 397–414. https://doi.org/10.1007/BF00127698

Mishler BD, Donoghue MJ (1982) Species concepts: a case for pluralism. Syst. Zool. 31, 491–503. https://doi.org/10.2307/2413371

Nylander JAA, Wilgenbusch JC, Warren DL, Swofford DL (2008) AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. Bioinformatics 24, 581–583. https://doi.org/10.1093/bioinformatics/btm388

O'Meara BC, Ané C, Sanderson MJ, Wainwright PC (2006) Testing for different rates of continuous trait evolution using likelihood. Evolution 60, 922–933. https://doi.org/10.1111/j.0014-3820.2006.tb01171.x

O'Meara BC (2010) New heuristic methods for joint species delimitation and species tree inference. Syst. Biol. 59, 59–73. https://doi.org/10.1093/sysbio/syp077

Oberprieler C, Wagner F, Tomasello S, Konowalik K (2017) A permutation approach for inferring species networks from gene trees in polyploid complexes by minimising deep coalescences. Methods Ecol. Evol. 8, 835–849. https://doi.org/10.1111/2041-210X.12694

Ogilvie HA, Bouckaert RR, Drummond AJ (2017) Starbeast2 brings faster species tree inference and accurate estimates of substitution rates. Mol. Biol. Evol. 34, 2101–2114. https://doi.org/10.1093/molbev/msx126

Oxelman B, Brysting AK, Jones GR, Marcussen T, Oberprieler C, Pfeil BE (2017) Phylogenetics of allopolyploids. Annu. Rev. Ecol. Evol. Syst. 48, 543–557. https://doi.org/10.1146/annurev-ecolsys-110316-022729

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155, 945–959. https://doi.org/10.1093/genetics/155.2.945

Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA (2018) Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Syst. Biol. 67, 901–904. https://doi.org/10.1093/sysbio/syy032

Rannala B, Edwards SV, Leaché A, Yang Z (2020) The multi-species coalescent model and species tree inference, in: Scornavacca, C., Delsuc, F., Galtier, N. (Eds.), Phylogenetics in the Genomic Era. p. 3.3:1-3.3:21.

Rannala B, Yang Z (2013) Improved reversible jump algorithms for Bayesian species delimitation. Genetics 194, 245–253. https://doi.org/10.1534/genetics.112.149039

Reeves PA, Richards CM (2007) Distinguishing terminal monophyletic groups from reticulate taxa: performance of phenetic, tree-based, and network procedures. Syst. Biol. 56, 302–320. https://doi.org/10.1080/10635150701324225

Reydon TAC (2019) Are species good units for biodiversity studies and conservation efforts?, in: Casetta, E., Marques da Silva, J., Vecchi, D. (Eds.), From Assessing to Conserving Biodiversity: Conceptual and Practical Challenges, History, Philosophy and Theory of the Life Sciences. Springer International Publishing, Cham, pp. 167–193. https://doi.org/10.1007/978-3-030-10991-2_8

Rodrigues ASL, Pilgrim JD, Lamoreux JF, Hoffmann M, Brooks TM (2006) The value of the IUCN Red List for conservation. Trends Ecol. Evol. 21, 71–76. https://doi.org/10.1016/j.tree.2005.10.010

Schrinner SD, Mari RS, Ebler J, Rautiainen M, Seillier L, Reimer JJ, Usadel B, Marschall T, Klau GW (2020) Haplotype threading: accurate polyploid phasing from long reads. Genome Biol. 21, 252. https://doi.org/10.1186/s13059-020-02158-1

Shipley B, Vile D, Garnier E (2006) From plant traits to plant communities: a statistical mechanistic approach to biodiversity. Science 314, 812–814. https://doi.org/10.1126/science.1131344

Simpson GG (1951) The species concept. Evolution 5, 285–298. https://doi.org/10.2307/2405675

Stankowski S, Ravinet M (2021) Quantifying the use of species concepts. Curr. Biol. 31, R428–R429. https://doi.org/10.1016/j.cub.2021.03.060

Struck TH, Feder JL, Bendiksby M, Birkeland S, Cerca J, Gusarov VI, Kistenich S, Larsson K-H, Liow LH, Nowak MD, Stedje B, Bachmann L, Dimitrov D (2018) Finding evolutionary processes hidden in cryptic species. Trends Ecol. Evol. 33, 153–163. https://doi.org/10.1016/j.tree.2017.11.007

Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A (2018) Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol. 4, vey016. https://doi.org/10.1093/ve/vey016

Sukumaran J, Holder MT, Knowles LL (2021) Incorporating the speciation process into species delimitation. PLoS Comput. Biol. 17, e1008924. https://doi.org/10.1371/journal.pcbi.1008924

Sukumaran J, Knowles LL (2017) Multispecies coalescent delimits structure, not species. Proc Natl Acad Sci USA 114, 1607–1612. https://doi.org/10.1073/pnas.1607921114

Sun X, Jiao C, Schwaninger H, Chao CT, Ma Y, Duan N, Khan A, Ban S, Xu K, Cheng L, Zhong G-Y, Fei Z (2020) Phased diploid genome assemblies and pan-genomes provide insights into the genetic history of apple domestication. Nat. Genet. 52, 1423–1432. https://doi.org/10.1038/s41588-020-00723-9

Susko E, Roger AJ (2020) On the use of information criteria for model selection in phylogenetics. Mol. Biol. Evol. 37, 549–562. https://doi.org/10.1093/molbev/msz228

Tang H, Woodhouse MR, Cheng F, Schnable JC, Pedersen BS, Conant G, Wang X, Freeling M, Pires JC (2012) Altered patterns of fractionation and exon deletions in *Brassica rapa* support a two-step model of paleohexaploidy. Genetics 190, 1563–1574. https://doi.org/10.1534/genetics.111.137349

Than C, Nakhleh L (2009) Species tree inference by minimizing deep coalescences. PLoS Comput. Biol. 5, e1000501. https://doi.org/10.1371/journal.pcbi.1000501

Than C, Ruths D, Nakhleh L (2008) PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinformatics 9, 322. https://doi.org/10.1186/1471-2105-9-322

The French-Italian Public Consortium for Grapevine Genome Characterization (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449, 463–467. https://doi.org/10.1038/nature06148

Tomasello S (2018) How many names for a beloved genus? - Coalescent-based species delimitation in *Xanthium* L. (Ambrosiinae, Asteraceae). Mol. Phylogenet. Evol. 127, 135–145. https://doi.org/10.1016/j.ympev.2018.05.024

Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485, 635–641. https://doi.org/10.1038/nature11119

Toprak Z, Pfeil BE, Jones G, Marcussen T, Ertekin AS, Oxelman B (2016) Species delimitation without prior knowledge: DISSECT reveals extensive cryptic speciation in the *Silene aegyptiaca* complex (Caryophyllaceae). Mol. Phylogenet. Evol. 102, 1–8. https://doi.org/10.1016/j.ympev.2016.05.024

Truco MJ, Ashrafi H, Kozik A, van Leeuwen H, Bowers J, Wo SRC, Stoffel K, Xu H, Hill T, Van Deynze A, Michelmore RW (2013) An ultra-high-density, transcript-based, genetic map of lettuce. G3 (Bethesda) 3, 617–631. https://doi.org/10.1534/g3.112.004929

Turland N, Wiersema J, Barrie F, Greuter W, Hawksworth D, Herendeen P, Knapp S, Kusber W-H, Li D-Z, Marhold K, May T, McNeill J, Monro A, Prado J, Price M, Smith G (Eds) (2018) International Code of Nomenclature for algae, fungi, and plants, Regnum Vegetabile. Koeltz Botanical Books. https://doi.org/10.12705/Code.2018

Van de Peer Y, Mizrachi E, Marchal K (2017) The evolutionary significance of polyploidy. Nat. Rev. Genet. 18, 411–424. https://doi.org/10.1038/nrg.2017.26

Van Valen L (1976) Ecological species, multispecies, and oaks. Taxon 25, 233–239. https://doi.org/10.2307/1219444

Wen D, Yu Y, Zhu J, Nakhleh L (2018) Inferring phylogenetic networks using phylonet. Syst. Biol. 67, 735–740. https://doi.org/10.1093/sysbio/syy015

Wiley EO (1978) The evolutionary species concept reconsidered. Syst. Zool. 27, 17–26. https://doi.org/10.2307/2412809

Wright S (1931) Evolution in Mendelian populations. Genetics 16, 97–159.

Yang Z, Rannala B (2010) Bayesian species delimitation using multilocus sequence data. Proc Natl Acad Sci USA 107, 9264–9269. https://doi.org/10.1073/pnas.0913022107

Yang Z (2002) Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. Genetics 162, 1811–1823. https://doi.org/10.1093/genetics/162.4.1811

Yan Z, Cao Z, Liu Y, Ogilvie HA, Nakhleh L (2022) Maximum parsimony inference of phylogenetic networks in the presence of polyploid complexes. Syst. Biol. 71, 706–720. https://doi.org/10.1093/sysbio/syab081

Zhang X, Wu R, Wang Y, Yu J, Tang H (2020) Unzipping haplotypes in diploid and polyploid genomes. Comput. Struct. Biotechnol. J. 18, 66–72. https://doi.org/10.1016/j.csbj.2019.11.011

Zink RM, Davis JI (1999) New perspectives on the nature of species, in: Adams, N.J., Slotow, R.H. (Eds.), . Presented at the Proceedings of the 22nd International Ornithological Congress, BirdLife South Africa, Johannesburg, pp. 1505–1518.

Answers

1. You could use a phylogenetic method allowing for allopolyploidization, AlloppNET. However, that method assumes that you know the correct species assignments (according to the MSC model). You could therefore try to assign homeologs to subgenomes of the polyploids using an MDC approach, and then run either a full Bayesian full likelihood model (e.g., STACEY) treating the subgenomes as separate diploids, and then finally run AlloppNET using the achieved delimitations. Alternatively, you could run MDC with PADRE network transformation.
2. a. The Wright-FIsher assumptions for the coalescent process, and the sequence evolution model for the gene trees (strict/relaxed clock, Jukes-Cantor, GTR, etc.). Note that if there is a lot of migration among branches, the trees can be grossly misleading.
   b. Clades can be viewed as historical individuals, so the obtained results would not reject the current taxonomy. With a concept of taxonomic species as being branches of the evolutionary tree, you would have to split into many species.
3. Allelic clustering (STRUCTURE-like) methods (population genomics). They are traditionally used in population genetics because alleles are directly clustered into bins and enable testing the fit to Hardy-Weinberg equilibrium. However, those methods are not coalescent-based and do not provide information about divergence time of population/species.

Chapter 18
Sequence to species

Phen Garrett[1], Shyam Gopalakrishnan[2]

1 University of Copenhagen, Copenhagen, Denmark
2 GLOBE Institute, SUND Department, University of Copenhagen, Copenhagen, Denmark

Phen Garrett **phengarrett@gmail.com**
Shyam Gopalakrishnan **shyam.gopalakrishnan@sund.ku.dk**

Citation: Garrett P, Gopalakrishnan S (2022) Chapter 18. Sequence to species. In: de Boer H, Rydmark MO, Verstraete B, Gravendeel B (Eds) Molecular identification of plants: from sequence to species. Advanced Books. https://doi.org/10.3897/ab.e98875

Introduction

Plant DNA can be extracted for species identification from a wide variety of sample types, including fresh, museum or ancient plant tissue collections that represent a single taxon, to highly processed samples that contain multiple individuals or taxa, including food and medicine (Chapter 6 DNA from food and medicine), water (Chapter 3 DNA from water), soil (Chapter 4 DNA from soil), pollen (Chapter 5 DNA from pollen), faeces (Chapter 7 DNA from faeces), or ancient sediments (Chapter 8 aDNA from sediments). Section 2 of this book explores how DNA can be used for plant identification through either targeted (where select regions of the genome are used) or non-targeted (resulting in / producing / allowing for representations of the full genome) approaches. Targeted approaches include barcoding for single taxon samples (Chapter 10 DNA barcoding, Chapter 13 Barcoding - High Resolution Melting, and Chapter 14 Target capture) and metabarcoding for samples representing multiple taxa (Chapter 11 Amplicon metabarcoding). Non-targeted approaches form the field of genomics. For single taxon samples, genome resequencing and whole genome sequencing (Chapter 16 Whole genome sequencing) are used, while in samples containing multiple taxa, metagenomic methods are used (Chapter 12 Metagenomics).

Studies conducting species identification either use known samples to find unknown identifications or use known identifications to assign identity to unknown samples. In the first, labelled samples are used for exploring evolutionary relationships to assign species identity based on some measurement of distance clustering (Chapter 19 Systematics and evolution, Chapter 20 Museomics, and Chapter 21 Palaeobotany). In contrast, the second category of studies utilise databases with predefined species classifications to assign identity to unknown samples (Chapter 22 Healthcare, Chapter 23 Food safety, Chapter 24 Environmental and biodiversity assessments, Chapter 25 Wildlife trade, and Chapter 26 Forensic genetics, botany, and palynology).

The analytical methods used for species identification can be categorised into three groups: i) database alignment analyses, ii) alignment-free methods, and iii) sample alignment analyses (Box 1).

In this chapter, we outline common sequence pre-processing steps used in species identification projects, and then discuss how species identification from sequencing data can be accomplished using the three analytical categories mentioned here.

Chapter 18: Box 1. Analytical methods for species identification

i) Database alignment analyses deal with single or mixed taxa from mostly mixed taxa samples and use targeted or non-targeted molecular methods. These alignment analyses start with unknown samples and attempt to assign known identifications using a pre-existing database of sequences linked to known identities.

ii) Alignment-free analytical methods utilise single taxon sample data from mostly non-targeted molecular methods, and start with known samples and apply distance-based clustering to explore genetic similarity and infer relationships.

iii) Sample alignment analyses also target single taxon samples, can be applied to sequence data from targeted or non-targeted molecular methods, and again start with known samples to explore evolutionary relationships and genetic distance or similarity between samples to create or improve identification understanding. This latter category includes de novo assemblies and the creation of reference genomes.

Sequencing quality control

"Garbage in garbage out" is a phrase that any experimentalist should keep in mind when setting up a species identification project. Obtaining robust and accurate species identities from sequencing data requires that input reads are high-quality and filtered for contamination and sequencing errors. This section outlines the steps necessary for firstly checking that data is of sufficient quality for species identification, as well as the sorts of processing steps that are necessary for sequence data analysis.

Quality check of raw sequence reads

Sequencing reads generated on short and long read platforms contain artefacts that need to be filtered or corrected in order to isolate high-quality reads for use in downstream analyses. Sequencing artefacts include reduction in read end base quality in short read data (common in Illumina sequencing), and amplified rates of homopolymer errors in longer read data generated with Nanopore technologies (Chapter 9 Sequencing platforms and data types). Correcting these errors is a mandatory first step in most bioinformatics analyses as poor quality control of raw sequence reads can result in inconclusive or incorrect species identification. Several quality control software packages including FastQC (Andrews and Others 2010), multiQC (Ewels et al. 2016), LongQC (Fukasawa et al. 2020), and NanoPack (De Coster et al. 2018) facilitate error corrections by visualising the data quality and identifying errors such as low base quality, over-representation of short k-mers, presence of adapter sequences, and GC biases.

Removal of non-biological sequences: adapters, tags, and demultiplexing

Sequence library preparation methods append non-representative, non-biological sequences, such as adapters and tags for multiplexing, to the DNA fragments. These sequences should therefore be removed during sequence processing to avoid failure in species identification or even a false species identification. Tools such as AdapterRemoval (Schubert et al. 2016), cutadapt (Martin 2011), leeHom (Renaud et al. 2014), and trimmomatic (Bolger et al. 2014) remove commonly used adapter sequences, and aid in isolating sample reads.

Removal of PCR artefacts

PCR (polymerase chain reaction) amplification of template DNA can introduce errors including artificial base differences, chimeras, and heteroduplex molecules (Acinas et al. 2005). Such PCR errors affect species identification sensitivity and should be identified and corrected from the data to ensure robust and accurate species identification. Tools such as FastUniq (Xu et al. 2012), fastx toolkit (HannonLab 2021), Kraken (Wood and Salzberg 2014), and BBMap's Clumpify (Bushnell 2021) can be used to identify PCR duplicates from raw sequence data, whereas samtools (Li et al. 2009) and picard ("Piccard Toolkit" 2019) can be used to identify and excise PCR duplicates and errors from reference mapped data.

Processing for targeted sequencing data

Targeted sequencing experiments, where a specific region of the genome or plastome is sequenced, require a few additional quality control steps to remove sequencing artefacts. Tools such as obitools (Boyer et al. 2016), begum and lulu (Frøslev et al. 2017) allow for filtering for singleton errors, chimeras, and other artefacts associated with targeted sequencing studies.

Filtering for DNA damage errors

There are specific challenges to be considered when analysing ancient DNA samples, including archaeological and herbarium samples. DNA damage, primarily driven by chemical changes in the DNA post-mortem, is prevalent in aDNA samples (see Chapter 2 DNA from museum collections). Programs such as mapDamage (Jónsson et al. 2013) and ATLAS (Link et al. 2017) offer statistical models to identify such substitutions and either remove them or assign low qualities. If DNA damage errors are retained, similar to PCR errors, they will affect sensitivity and result in lower species assignment resolution.

Database alignment

Database alignment methods are the most intuitive class of search-based species identification from sequencing data and have been used for the better part of the last three decades to identify species that are the putative sources of sample DNA or protein sequences. These methods compare the sequencing reads, either directly in the form of short reads or in the form of assembled contigs, to a reference database of curated sequences. Widely used alignment tools include BLAST (McGinnis and Madden 2004), BLASTn and MegaBLAST (Chen et al. 2015), DIAMOND (Buchfink et al. 2015), Kraken (Wood and Salzberg 2014), and Kaiju (Menzel et al. 2016). These tools primarily align the reads themselves (DNA) or the translated amino acid sequences to reference datasets. Sequences with high similarity are then identified using a local alignment approach where parts of the credit sequences are aligned with parts of sequences in the database, in many cases using a seed and extend algorithm. Local alignment implies that not all of the query sequence needs to match perfectly with the target sequences and allows finding closely related species in the database even if the evolutionary distance between the target and query sequences is large.

Applicability

In theory, alignment-based approaches using databases can be used for species identification on sequences generated from the entire spectrum of molecular methods detailed previously. However, high computational requirements coupled with logistical issues such as the unavailability of appropriate databases make these methods best suited to targeted sequencing approaches, especially barcoding and metabarcoding. In these approaches, only a limited number of unique sequences are used in the initial data input, making them substantially less computationally expensive methods.

Database choice

The database choice plays an integral role in the sensitivity and specificity of local alignment algorithms and whether the alignment approaches return a species identification. Accurate and positive species identifications are more likely with databases containing high numbers of closely related species. Global databases, such as the NCBI nucleotide database and NCBI non-redundant protein database (Pruitt et al. 2012), provide a large number of DNA and amino acid sequences derived from nuclear, mitochondrial, and plastid genomes for a large number of plants. While this is an excellent resource, the sequences in these databases are not always well curated, and are skewed towards well studied organisms. Another database to consider for targeted sequencing approaches is the "barcode of life database" (BOLD), which contains the curated barcode sequences for more than 9 million barcodes (~232,000 species), including > 69 000 plant species and > 23 000 fungal and other species (Ratnasingham and Hebert 2007).

Alternative options to consider are national or local sequence databases that have been assembled by genetic and genomic researchers to represent the species of a country or region. Prime examples include DNAmark (Margaryan et al. 2020), a pilot database with whole mitochondrial sequences and genome skims of more than 1000 species from across Denmark, as well as NorBOL, a database of Norwegian species' barcode sequences, and R-syst::diatom, the *rbcL* database for diatoms (Rimet et al. 2016). There is a tradeoff however when moving from global to local databases, both in terms of genomic content and geographic region. While the large global databases include more sequences representing more species as well as potentially greater within-species variation, thus offering a better probability of finding a match, they are not always well curated and require higher computational resources. On the other hand, local targeted databases have a smaller number of sequences, but can provide a much finer resolution in terms of species identification.

DNA or protein

Does the choice of using DNA or protein make a difference in the database alignment algorithm? Yes! DNA sequences provide more sensitivity while amino acid sequences are more robust. What do we mean by that? DNA sequences can provide a better resolution in terms of describing the evolutionary relationships between closely related species. Proteins on the other hand can illuminate much older evolutionary relationships, and tend to provide more robust identifications (Wernersson and Pedersen 2003). While more time-consuming, a hybrid approach using a combination of DNA and the translated protein sequences can offer the benefits of both approaches, allowing resolution of recent evolutionary events, while providing enough robust information to place the sequence in the correct phylogenetic context.

Short reads or assembled contigs

Alignment based methods can be used on both raw short reads directly from the sequencing machines and on assembled contigs, where multiple short reads are stitched together into longer stretches of DNA. These approaches come with their own pros and cons. Ease of use is the primary selling point in using short reads directly from the sequencing machine. Using assembled contigs requires additional steps, but the increased length can result in lower error rates and longer read regions, leading to better resolution. The use of assembled contigs additionally takes advantage of databases that allow for alignment of longer regions, including pos-

sibly the entire target region (Bankevich et al. 2012). That being said, longer fragments do not necessarily work better and mini-barcodes are more cost-effective and work well for degraded DNA (Yeo et al. 2020).

Alignment-free methods

The rapid advance and adoption of second and third generation sequencing technologies has led to an exponential increase in the numbers of sequencing studies that employ either whole genome resequencing or genome skimming to characterise sample genomes. With these large genomic datasets, alignment based approaches can be computationally taxing (Elias 2006) and lead to inaccurate results for sequences with low similarity and/or large rearrangements, as is often the case in plants. The limitations of alignment based approaches has led to the development of many classes of alignment-free approaches that maintain accuracy across large evolutionary distances and are computationally feasible given the increasing dataset sizes. These alignment-free methods provide an opportunity to use sequences from multiple samples within a single experiment to compute pairwise dissimilarity metrics and infer evolutionary relationships between them.

Alignment-free approaches come in many flavours, including k-mer based methods, micro-alignments, fourier transformation methods, and information theory methods (Blaisdell 1986; Haubold et al. 2015; Jun et al. 2010; Reinert et al. 2009; Vinga 2014; Yin and Yau 2015; Yi and Jin 2013). A review by Zielenski et al. provides a deep dive into the theory and implementations of these alignment-free methods (Zielezinski et al. 2019). The most commonly used class of alignment-free methods are the k-mer based methods. The k-mer profile of a sequence consists of all possible k-mers in the sequence, where k-mer is a substring of length k embedded in the sequence. Most k-mer based methods work by transforming the sequencing data into the frequencies of the k-mers contained in the sequences. These k-mer frequencies are computed using sequences from different assemblies, and whole genome resequencing or genome skimming experiment data can be used to compute the distances/dissimilarity between sequences, thus providing a proxy for evolutionary distances.

Applicability

Alignment-free methods are primarily restricted for use with non-targeted sequencing approaches. This is due to the short length of targeted regions leading to a limited number of k-mers, which restricts the ability of these approaches to result in meaningful inferences. Although k-mer based methods might look like ideal candidates for use in metagenomics, the fact that metagenomic samples are derived from multiple sources in varying proportions makes it difficult to successfully isolate individual taxa (Pellegrina et al. 2020).

The depth to which samples are sequenced affects the accuracy of the dissimilarity metric estimates computed in alignment-free methods. As the sequencing depth reduces, their variance increases even if the estimates remain unbiased by assembly. This variance is propagated into the downstream analyses. Thus, the robustness of these methods should be verified when using very low coverage sequencing data (Sarmashghi et al. 2019).

Contamination can also be tricky to deal with in alignment-free methods using mixed bags of raw sequencing reads, and therefore filtering for contamination using tools such as BlobTools (Laetsch and Blaxter 2017) and DIAMOND (Buchfink et al. 2015) is often necessary.

From k-mer profiles to distances

K-mer frequency profiles of sequences are used to compute dissimilarity scores between those sequences. There are many distance metric options that can be used to compute the dissimilarity score, e.g., Euclidian, inner product, Kullback-Leibler divergence (relative entropy), and mismatches (Jaccard). The most commonly used distance metric is the Jaccard distance, since it is easy to compute and corresponds to nucleotide changes. Specifically, the Jaccard distance ranges from 0.0 to 1.0, where 0.0 corresponds to identical k-mer profiles, and 1.0 implies no overlap in k-mers. By computing the pairwise Jaccard distances between sequences from an unknown sample and a set of reference sequences with known species identity, we can assign our unknown samples to the closest species among the set of reference sequences. Further, the dissimilarity measures can be used to build a phylogeny of the sequences (Ondov et al. 2016).

Assembled genomes or raw sequencing reads?

An advantage of k-mer based methods is their applicability to different sequencing data types, which allow combining sequence data from different experiment types. For example, one can compute the k-mer frequency profiles directly from the reads or from scaffold sequences. All subsequent steps to compute distances can be applied without regard to potentially different sequence sources.

A few k-mer based programs

Several alignment-free methods have been developed in the last few years, incorporating several of the k-mer algorithms (Zielezinski et al. 2019). One of the most promising tools, which reduces the computational complexity of computing Jaccard distances between pairs of sequences, is Mash (Ondov et al. 2016). Mash reduces large amounts of input sequences to "sketches" consisting of hashed k-mers from the data. Only a subset of the most frequent k-mers are then used to compute the Jaccard distance, thus reducing both the memory and computational time footprint of the program. It also computes the mash (min-hash) distance, which estimates the mutation rates under an evolutionary model. CAFE (Lu et al. 2017) is another popular k-mer based method that can compute several different distance measures based on both k-mer counts and presence/absence of individual k-mers. CAFE also provides background-adjusted dissimilarity measures such as CVTree (Qi et al. 2004), d2star, and d2shepp (Tang et al. 2019). Finally, skmer (Sarmashghi et al. 2019) is a tool that uses mash to estimate k-mer profiles of genome skims and computes genomic distances while modelling sequencing error and correcting for low coverage. This makes it ideal for genome skimming experiments. All these methods can be used to combine de novo assemblies, whole genome resequencing, and genome skimming experiments.

Assembly or mapping for multisequence sample alignments

Sample alignment methods are the foundation of molecular taxonomy, phylogenetic classification, and population genetics, and allow the exploration of evolutionary relationships and ge-

netic distance between samples. These methods include de novo assemblies and the creation of reference genomes, as well as assembly or mapping using a reference. There is inherent bias in terms of reference availability, and inadequate reference mapping can result in skewed representations of genetic similarity in downstream analysis.

Assembly or mapping to a reference

The use of references to inform the assembly of contigs to produce scaffolds and create sample-specific consensus sequences representing genes, gene regions, and genomes is inherently biassed towards the available references. Popular mapping tools include Global BWA (Li 2013), MGMapper assembly (Petersen et al. 2017) and Bowtie2 (Langmead and Salzberg 2012). All require no consensus sequences and are able to deal with millions of reads. The experimentalist should consider the mapping parameters to exclude sequence contamination.

De novo assembly (reference-free)

High quality, in-depth sequencing is required to produce a de novo assembly. A de novo assembly will however avoid any inherent biases introduced by using references for assembly, and in turn can be used as reference in future projects. For an outline of the processes involved in de novo assemblies, please see reviews by (Jiao and Schneeberger 2017) and (Liao et al. 2019).

Alignment

The foundation of any tree-building or comparative gene analysis is multiple sequence alignment (MSA). MSA matches up areas of the genome across samples and allows for comparison. MSA algorithms are based on maximising sum-of-pair scores through heuristic progressive (input-order dependent) alignments (Needleman and Wunsch 1970) refined by global pairwise alignment methods following either polishing (subsetting and iteratively realigning datasets) or consistency (dependent, position-specific substitution scores assigned within pairwise alignments across the dataset) approaches (Chatzou et al. 2016; Wheeler and Kececioglu 2007).

Widely used MSA tools include ClustalW (Thompson et al. 1994), T-Coffee (Notredame et al. 2000), ProbCons (Do et al. 2005), MUSCLE (Edgar 2004), MAFFT (Katoh et al. 2002), PASTA (Mirarab et al. 2015), SaTé (Mirarab et al. 2015), and Clustal Omega (Sievers et al. 2011). These all combine iterative heuristic algorithmic strategies of progressive and global pairwise alignments with incorporated complex schemes to account for different substitution scoring, gap penalties, length divergence, hydrophobicity, and neighbouring gap proximity (Wheeler and Kececioglu 2007).

Other considerations

There are several important factors that can determine or influence which species assignment method is ultimately chosen. The study design and experimental question, as well the DNA

source and extraction methods are important factors. For example, genome skimming and metagenomic studies might be well suited to alignment-free methods (Zielezinski et al. 2019), while targeted sequencing approaches require alignment (Wilson et al. 2019). Additionally the sequencing depth and amount and quality of data can all ultimately determine the chosen assignment method. Sequencing depth is directly related to how reliable species assignments are, with higher sequencing depths providing more reliable results. This is due either to a higher number of raw reads supporting the result, or to more complete assembled contigs for metagenomic experiments and better consensus sequences for targeted sequencing. Similarly, sequence quality, a function of preprocessing and filtering steps performed before the species assignment methods are used, is critical in ensuring that the results of species identification are as accurate and well supported as possible.

It is thus important to be aware of the strengths and limitations of different species assignment methods and to choose the method best suited to the biological questions being posed and the experimental design used to generate the sequencing data. For alignment based methods, it is important to remember that the species identification results are only as good as the databases the sequences are being aligned to, applicable to both targeted sequencing and genomic studies. Further, results from alignment against large databases must be interpreted carefully, since the order of the results are dependent on both the sequence identity and the number of times a certain species is represented in the database. For alignment-free methods, such as k-mer based, sequencing depth and the quality of the k-mer profiles from target species (database) are important factors. Also note that the value of k in the k-mer profile generation is an important parameter to tune. Finally, metagenomic taxonomy assignment tools again depend, in varying degrees, on external databases for identification of taxa.

Questions

1. What are several sequencing artefacts which need consideration and removal?
2. What difference does the choice of using DNA or protein in the database alignment algorithm make?
3. What types of sequencing approaches are alignment-free analytical methods primarily used for?

Glossary

Contamination – DNA from a non-targeted taxa.
Contig – A single continuous sequence of DNA present in a genome assembly. Contigs in modern genome assemblies are hundreds of kilobases or multiple megabases in length.
K-mer – A sequence of length k. For example, a 27-mer is the collection of all (overlapping) sequences of length 27 base pairs in a given set of sequences.
Multiplexing – Combining tagged DNA fragments from multiple samples before sequencing.
Non-targeted approaches (genomics) – Capturing representations of the full genome. See Chapter 16 Whole genome sequencing
Reference genome – A high-quality genome sequence from a single individual that is used as the foundation for genomic analysis.
Scaffold – An assembly of contigs separated by gaps of known length.

Sequencing depth (= coverage) – The number of unique reads including a given nucleotide. This is about the depth of coverage.
Tags – DNA fragment labels for multiplexing.
Targeted approaches (genetics, including amplicon sequencing) – Where the breadth of coverage is defined and a smaller amount than the whole genome is used.

References

Acinas SG, Sarma-Rupavtarm R, Klepac-Ceraj V, Polz MF (2005) PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. Appl. Environ. Microbiol. 71, 8966–8969. https://doi.org/10.1128/AEM.71.12.8966-8969.2005

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–477. https://doi.org/10.1089/cmb.2012.0021

Blaisdell BE (1986) A measure of the similarity of sets of sequences not requiring sequence alignment. Proc Natl Acad Sci USA 83, 5155–5159. https://doi.org/10.1073/pnas.83.14.5155

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Boyer F, Mercier C, Bonin A, Le Bras Y, Taberlet P, Coissac E (2016) obitools: a unix-inspired software package for DNA metabarcoding. Mol. Ecol. Resour. 16, 176–182. https://doi.org/10.1111/1755-0998.12428

Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. Nat. Methods 12, 59–60. https://doi.org/10.1038/nmeth.3176

Chatzou M, Magis C, Chang J-M, Kemena C, Bussotti G, Erb I, Notredame C (2016) Multiple sequence alignment modeling: methods and applications. Brief. Bioinformatics 17, 1009–1023. https://doi.org/10.1093/bib/bbv099

Chen Y, Ye W, Zhang Y, Xu Y (2015) High speed BLASTN: an accelerated MegaBLAST search tool. Nucleic Acids Res. 43, 7762–7768. https://doi.org/10.1093/nar/gkv784

De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C (2018) NanoPack: visualizing and processing long-read sequencing data. Bioinformatics 34, 2666–2669. https://doi.org/10.1093/bioinformatics/bty149

Do CB, Mahabhashyam MSP, Brudno M, Batzoglou S (2005) ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome Res. 15, 330–340. https://doi.org/10.1101/gr.2821705

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797. https://doi.org/10.1093/nar/gkh340

Elias (2006) Settling the Intractability of Multiple Alignment. J Comput Biol 13, 1323–39.

Ewels P, Magnusson M, Lundin S, Käller M (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32, 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Frøslev TG, Kjøller R, Bruun HH, Ejrnæs R, Brunbjerg AK, Pietroni C, Hansen AJ (2017) Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. Nat. Commun. 8, 1188. https://doi.org/10.1038/s41467-017-01312-x

Fukasawa Y, Ermini L, Wang H, Carty K, Cheung M-S (2020) Longqc: A quality control tool for third generation sequencing long read data. G3 (Bethesda) 10, 1193–1196. https://doi.org/10.1534/g3.119.400864

Haubold B, Klötzl F, Pfaffelhuber P (2015) andi: fast and accurate estimation of evolutionary distances between closely related genomes. Bioinformatics 31, 1169–1175. https://doi.org/10.1093/bioinformatics/btu815

Jiao W-B, Schneeberger K (2017) The impact of third generation genomic technologies on plant genome assembly. Curr. Opin. Plant Biol. 36, 64–70. https://doi.org/10.1016/j.pbi.2017.02.002

Jónsson H, Ginolhac A, Schubert M, Johnson PLF, Orlando L (2013) mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics 29, 1682–1684. https://doi.org/10.1093/bioinformatics/btt193

Jun S-R, Sims GE, Wu GA, Kim S-H (2010) Whole-proteome phylogeny of prokaryotes by feature frequency profiles: An alignment-free method with optimal feature resolution. Proc Natl Acad Sci USA 107, 133–138. https://doi.org/10.1073/pnas.0913033107

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30, 3059–3066. https://doi.org/10.1093/nar/gkf436

Laetsch DR, Blaxter ML (2017) BlobTools: Interrogation of genome assemblies [version 1; peer review: 2 approved with reservations]. F1000Res. 6, 1287. https://doi.org/10.12688/f1000research.12232.1

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359. https://doi.org/10.1038/nmeth.1923

Liao X, Li M, Zou Y, Wu F-X, Yi-Pan Wang J (2019) Current challenges and solutions of de novo assembly. Quant. Biol. 7, 90–109. https://doi.org/10.1007/s40484-019-0166-9

Link V, Kousathanas A, Veeramah K, Sell C, Scheu A, Wegmann D (2017) ATLAS: Analysis Tools for Low-depth and Ancient Samples. BioRxiv. https://doi.org/10.1101/105346

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

Lu YY, Tang K, Ren J, Fuhrman JA, Waterman MS, Sun F (2017) CAFE: aCcelerated Alignment-FrEe sequence analysis. Nucleic Acids Res. 45, W554–W559. https://doi.org/10.1093/nar/gkx351

Margaryan A, Noer CL, Richter SR, Restrup ME, Bülow-Hansen JL, Leerhøi F, Langkjær EMR, Gopalakrishnan S, Carøe C, Gilbert MTP, Bohmann K (2020) Mitochondrial genomes of Danish vertebrate species generated for the national DNA reference database, DNAmark. Environmental DNA. https://doi.org/10.1002/edn3.138

McGinnis S, Madden TL (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res. 32, W20-5. https://doi.org/10.1093/nar/gkh435

Menzel P, Ng KL, Krogh A (2016) Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat. Commun. 7, 11257. https://doi.org/10.1038/ncomms11257

Mirarab S, Nguyen N, Guo S, Wang L-S, Kim J, Warnow T (2015) PASTA: ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. J. Comput. Biol. 22, 377–386. https://doi.org/10.1089/cmb.2014.0156

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48, 443–453. https://doi.org/10.1016/0022-2836(70)90057-4

Notredame C, Higgins DG, Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 302, 205–217. https://doi.org/10.1006/jmbi.2000.4042

Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM (2016) Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 17, 132. https://doi.org/10.1186/s13059-016-0997-x

Pellegrina L, Pizzi C, Vandin F (2020) Fast approximation of frequent k-mers and applications to metagenomics. J. Comput. Biol. 27, 534–549. https://doi.org/10.1089/cmb.2019.0314

Pruitt KD, Tatusova T, Brown GR, Maglott DR (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res. 40, D130-5. https://doi.org/10.1093/nar/gkr1079

Qi J, Luo H, Hao B (2004) CVTree: a phylogenetic tree reconstruction tool based on whole genomes. Nucleic Acids Res. 32, W45-7. https://doi.org/10.1093/nar/gkh362

Ratnasingham S, Hebert PDN (2007) BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). Mol. Ecol. Notes 7, 355–364. https://doi.org/10.1111/j.1471-8286.2007.01678.x

Reinert G, Chew D, Sun F, Waterman MS (2009) Alignment-free sequence comparison (I): statistics and power. J. Comput. Biol. 16, 1615–1634. https://doi.org/10.1089/cmb.2009.0198

Renaud G, Stenzel U, Kelso J (2014) leeHom: adaptor trimming and merging for Illumina sequencing reads. Nucleic Acids Res. 42, e141. https://doi.org/10.1093/nar/gku699

Sarmashghi S, Bohmann K, P Gilbert MT, Bafna V, Mirarab S (2019) Skmer: assembly-free and alignment-free sample identification using genome skims. Genome Biol. 20, 34. https://doi.org/10.1186/s13059-019-1632-4

Schubert M, Lindgreen S, Orlando L (2016) AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res. Notes 9, 88. https://doi.org/10.1186/s13104-016-1900-2

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7, 539. https://doi.org/10.1038/msb.2011.75

Tang K, Ren J, Sun F (2019) Afann: bias adjustment for alignment-free sequence comparison based on sequencing data using neural network regression. Genome Biol. 20, 266. https://doi.org/10.1186/s13059-019-1872-3

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680. https://doi.org/10.1093/nar/22.22.4673

Vinga S (2014) Information theory applications for biological sequence analysis. Brief. Bioinformatics 15, 376–389. https://doi.org/10.1093/bib/bbt068

Wernersson R, Pedersen AG (2003) RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. Nucleic Acids Res. 31, 3537–3539. https://doi.org/10.1093/nar/gkg609

Wheeler TJ, Kececioglu JD (2007) Multiple alignment by aligning alignments. Bioinformatics 23, i559-68. https://doi.org/10.1093/bioinformatics/btm226

Wilson J-J, Sing K-W, Jaturas N (2019) DNA barcoding: bioinformatics workflows for beginners, in: Encyclopedia of Bioinformatics and Computational Biology. Elsevier, pp. 985–995. https://doi.org/10.1016/B978-0-12-809633-8.20468-8

Wood DE, Salzberg SL (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 15, R46. https://doi.org/10.1186/gb-2014-15-3-r46

Xu H, Luo X, Qian J, Pang X, Song J, Qian G, Chen J, Chen S (2012) FastUniq: a fast de novo duplicates removal tool for paired short reads. PLoS ONE 7, e52249. https://doi.org/10.1371/journal.pone.0052249

Yeo D, Srivathsan A, Meier R (2020) Longer is not always better: optimizing barcode length for large-scale species discovery and identification. Syst. Biol. 69, 999–1015. https://doi.org/10.1093/sysbio/syaa014

Yin C, Yau SS-T (2015) An improved model for whole genome phylogenetic analysis by Fourier transform. J. Theor. Biol. 382, 99–110. https://doi.org/10.1016/j.jtbi.2015.06.033

Yi H, Jin L (2013) Co-phylog: an assembly-free phylogenomic approach for closely related organisms. Nucleic Acids Res. 41, e75. https://doi.org/10.1093/nar/gkt003

Zielezinski A, Girgis HZ, Bernard G, Leimeister C-A, Tang K, Dencker T, Lau AK, Röhling S, Choi JJ, Waterman MS, Comin M, Kim S-H, Vinga S, Almeida JS, Chan CX, James BT, Sun F, Morgenstern B, Karlowski WM (2019) Benchmarking of alignment-free sequence comparison methods. Genome Biol. 20, 144. https://doi.org/10.1186/s13059-019-1755-7

Answers

1. Sequencing artefacts include: low base quality, over-representation of short k-mers, presence of adapter sequences, GC biases, reduction in read end base quality in short read data, and amplified rates of homopolymer errors in longer read data.
2. DNA sequences provide more sensitivity and can provide a better resolution in terms of describing the evolutionary relationships between closely related species. Proteins on the other hand can illuminate much older evolutionary relationships, and tend to provide more robust identifications.
3. Alignment-free methods are primarily restricted for use with non-targeted sequencing approaches.

SECTION 3

Applications

Chapter 19
Systematics and evolution

Ntwai A. Moiloa[1,2], Meshack N. Dludlu[3], Abubakar Bello[4], Zaynab Shaik[1,2,5], A. Muthama Muasya[6], Bengt Oxelman[1,2]

1 Department of Biological and Environmental Sciences, University of Gothenburg, Gothenburg, Sweden
2 Gothenburg Global Biodiversity Centre, Gothenburg, Sweden
3 Eswatini Institute for Research in Traditional Medicine, Medicinal and Indigenous Food Plants (EIRMIP), University of Eswatini, Kwaluseni, Eswatini
4 Department of Biology, Faculty of Natural and Applied Sciences, Umaru Musa Yar'adua University, Katsina State, Nigeria
5 Department of Botany and Zoology, Natural Sciences Building, Stellenbosch University, Stellenbosch, South Africa
6 Department of Biological Sciences, University of Cape Town, Cape Town, South Africa

Ntwai A. Moiloa **ntwai.moiloa@bioenv.gu.se**
Meshack N. Dludlu **meshack.dludlu@alumni.uct.ac.za**
Abubakar Bello **bello.abubakar@umyu.edu.ng**
Zaynab Shaik **zaynab.shaik@bioenv.gu.se**
A. Muthama Muasya **muthama.muasya@uct.ac.za**
Bengt Oxelman **bengt.oxelman@bioenv.gu.se**

Citation: Moiloa NA, Dludlu MN, Bello A, Shaik Z, Muasya AM, Oxelman B (2022) Chapter 19. Systematics and evolution. In: de Boer H, Rydmark MO, Verstraete B, Gravendeel B (Eds) Molecular identification of plants: from sequence to species. Advanced Books. https://doi.org/10.3897/ab.e98875

Introduction

Systematics is the field of biology that studies biological diversity (or biodiversity) and its evolutionary history (Judd et al. 2007). Systematics comprises three subfields: 1) taxonomy, which focuses on classification, identification, nomenclature of taxa; 2) phylogenetics, which focuses on studying the evolutionary history of taxa; and 3) evolutionary biology, which focuses on studying differentiation processes of populations, speciation, and hybridisation (Stuessy 2009). Thus, systematics broadly seeks to understand how lineages split into two or more lineages, determine the evolutionary changes that occur over time, and how such changes bring about distinct evolutionary entities (Judd et al. 2007). These entities are then classified into taxa (singular: taxon) of various inclusiveness. This classification is based on their evolutionary relationships, such that members of the same taxon are more closely related to each other than to those belonging to other taxa.

An integral part of systematics is taxonomy, which focuses on the identification, description, naming, classification and inventory of taxa (Simpson 2019). In systematics, like in other related disciplines, species is often considered a fundamental biological unit (de Queiroz 2005). Furthermore, species is a categorical rank within the taxonomic hierarchy. In botanical classification, the following major ranks are recognized in descending order: kingdom, division (or phylum), class, order, family, genus, section, series, species, variety, and forma (Turland et al. 2018). Species are often considered to be the key to understanding the origin and evolutionary dynamics of biodiversity (Barraclough 2019). Nonetheless, the definition of what constitutes a species is a highly contentious issue and many different species concepts exist (Mayden 1997; de Queiroz 2007; Zachos 2016) (see Chapter 17 Species delimitation). These include the biological species concept (emphasising reproductive infertility between individuals; Mayr 1942; de Queiroz 2005), the evolutionary species concept (emphasising a common evolutionary history through time; Wiley 1978), and the morphological or phenetic species concept (emphasising shared similarities; Sokal and Crovello 1970).

A major challenge that comes from these different species concepts is that they may be incompatible and often lead to different conclusions on the boundaries of what should be considered the same or different species (de Queiroz 2007). Given the variety of species concepts and their definitions, de Queiroz (2007) suggested two solutions. The first solution identifies commonalities in the different species concepts, resulting in a unified concept where species are defined as separately evolving metapopulation lineages. The second solution emphasises the necessity of separating the problem of species concepts from that of species delimitation. In the present chapter, species delimitation refers to the practice of determining boundaries of species based on empirical data. For a comprehensive discussion of the various species concepts and delimitation approaches the reader is referred to Chapter 17 Species delimitation.

Commonly used methodological approaches in plant systematics include traditional comparative morphological/anatomical systematics, chemosystematics, and molecular systematics, which utilise different sources of data as input for inference. In traditional comparative morphological/anatomical systematics, the grouping of taxa is primarily based on morphological/phenotypic similarity (Bell and Bryan 1991). Thus, morphology (including anatomy) has been emphasised as a basic taxonomic tool where classification has primarily been based on the organism's morphological characteristics (Radford et al. 1974; Singh 2019). Chemosystematics adds various kinds of chemical information as characters for taxonomic purposes, while molecular systematics utilises genomic sequence information to study and understand the evolutionary history of different organisms. The advent of DNA sequencing in the 1970s (Sanger and Coulson 1975; Sanger et al. 1977) revolutionised the field of systematics and reconstruction of phylogenetic relationships. In the current era of molecular systematics phylogenetic reconstruction is based primarily on the

information from genomic data, while other data, such as morphological, anatomical, chemical and ecological data, are considered as auxiliary. Recent advances in high-throughput sequencing technologies (Cotton 2016) enable scientists to generate large-scale data allowing to study evolutionary relationships among members of any of the major domains of life.

Data sources used in systematics

The primary aim of systematics is to recognise evolutionary lineages where the genotypes are reproduced through time. The phenotypes are ephemeral manifestations of these genealogical lineages. Historically, the Aristotelian view of taxa ("natural kinds") having essential features (i.e., to qualify as a vertebrate, the organism must develop vertebrae) has dominated biological systematics. Some philosophers argue that the essence of those natural kinds may exist regardless of humans' abilities to recognise them, but there is no doubt that essentialism has played a great role in recognition of many taxonomic groups where certain phenotypic traits have been used for defining specific taxa. The development of evolutionary theory has provided systematists with the concept of monophyly, which ultimately is based on genealogical relationships. By using phylogenetic methodology, monophyletic groups (clades) sharing a common ancestry can be recognised. Both phenotypic and genotypic data can be useful for this, but the former is considered a proxy for the latter. Thus, while recognising the enormous importance of phenotypic data for the primary identification (i.e., classification and nomenclature) of taxa (and of course of general biology), we will in the following focus on the genetic data.

Anatomy

The use of internal or anatomical features in taxonomy began with the development of microscopes powerful enough to visualise the internal structures of organs and tissues (Dickison 2000). In plant classification, anatomical features such as the positioning of vascular bundles, the form and presence of tissues and cells, and trichomes are important (Simpson 2019).

Chemistry

Most chemotaxonomists recognise three broad categories of chemical compounds as taxonomically important: primary metabolites, secondary metabolites, and semantides (Turner 1969; Cronquist 1977). Although theoretically, many chemical constituents of a plant are of potential taxonomic value, in practice, visible chemical constituents such as crystals, raphides, or starch grains are most commonly used (Cronquist 1977; Reynolds 2007).

Cytology

The number of chromosomes in each cell of all individuals of a species is usually constant and more closely related species are likely to have similar haploid chromosome numbers (Jones 1995). In addition, biological processes such as hybridisation and gene flow can lead to genome duplication where an organism inherits more than two sets of chromosomes. This duplication of the genome, known as polyploidization, results in closely related species having different

numbers of chromosome sets (ploidy levels). Thus, chromosome number is a frequently used taxonomic character (Heslop-Harrison and Schwarzacher 2011). Additionally, the centromere provides information on the relationship between the two chromosome arms, and its position is used in identifying whether chromosomes are metacentric, acrocentric, or telocentric (Sharma 1993). The basic chromosome set in a dividing cell can additionally be analysed to provide information on the chromosome size, volume, and type (Guerra 2008). Studying chromosome behaviour during meiosis provides valuable information on the role of chromosomes in heredity.

Embryology

Embryonic development and structure have historically been used at different levels of classification. For example, the basic division of the plant kingdom into two units, the Thallophyta and the Embryophyta, was based in part on zygotic behaviour. In the same way, embryonic characteristics were an important component in the division of the angiosperms into two major groups, the monocotyledons and the dicotyledons (Johri et al. 1992).

Palynology

Studying plant pollen and spores is useful for determining species relationships in plants (Walker and Doyle 1975; Moore et al. 1994). For example, monosulcate pollen is characterised by a boat shape with one long furrow and a germinal aperture that is associated with some dicots and the majority of monocots, while tricolpate pollen typically has three apertures and is a characteristic feature of eudicots (Radford et al. 1974).

Palaeobotany

Morphological and genetic analysis of fossil material from pollen, leaves, stems, and other plant parts are used to trace evolutionary developments through stratigraphic sequences and also predicting past ecological conditions (Reitsma 1970) (see Chapter 21 Palaeobotany).

Molecular data

Strictly, DNA constitutes the genotype, while RNA, proteins, and associated structures belong to the phenotype. Nevertheless, DNA, RNA, and proteins can all be used to detect basic genotype changes. Very often, nucleotide substitutions are neutral and either do not change the amino acid sequence of the protein that they transcribe for, or result in minimal changes in the amino acid sequence (Kimura 1968). Therefore, the characteristics of the protein structure and resulting function are often conserved across species, reducing the number of associated problems when making homology assertions. The different kinds of molecular data that may be found from DNA/RNA molecules and proteins include: allozymes (allelic variants of enzymes encoded by structural genes). Early methodology developed in the 1970s–1980s includes DNA-DNA hybridization (technique used to determine the genetic similarity between DNA sequences), Restriction Fragment Length Polymorphism (RFLP; a sequence of DNA, restriction sites on each end and a target sequence in between), DNA microsatellites (tandem repeats of 1–6 nucleotides), but at present direct DNA sequencing dominates.

Several methods have been developed to either generate sequence data for whole genomes (whole genome sequencing, WGS), or sample a subset of specific loci from across the genome (Mullis et al. 1986; Elshire et al. 2011; Lemmon et al. 2012). The latter methods allow effective studies on sequence variability at the population level. Explicit statistical models can readily be applied to molecular data, because of the discreteness of the characters in sequence data. Bayesian and maximum likelihood methods for gene-tree estimation relies on likelihood calculations which are based on explicit definitions of parameters in the underlying evolutionary model (Felsenstein 2004). In gene-tree estimation utilising likelihood calculations, the probability (likelihood) of the underlying multiple sequence alignment (MSA) of sequences under a given hypothesis (tree topology with branch lengths) is calculated. It should be noted here that this MSA is usually considered as fixed, despite the computational difficulties and the lack of theoretical justification for many of the methods utilised to derive it (Morrison 2018). Recently, the multispecies coalescent model (MSC; Rannala and Yang 2003) has become commonly applied in phylogenetic studies. The MSC takes gene-trees as data, and efficiently deals with the fact that unlinked genes are expected to have different genealogies.

Advances in molecular phylogenetics

Tree estimation methods

Since the advent of molecular phylogenetics in the late 1980s, several theoretical approaches for reconstructing relationships in the Tree of Life have been developed (Mishler 2013). For brevity, we discuss these approaches in three broad categories: algorithmic, optimality, and Bayesian approaches. Algorithmic (or clustering) approaches to tree inference use an algorithm to derive a phylogenetic tree for a data set. Neighbour-joining (Saitou and Nei 1987) and UPGMA (unweighted pair group method with arithmetic mean) (Sokal and Michener 1958) are both iterative algorithmic approaches that use a pairwise distance matrix to create a tree in which evolutionary divergence is assumed proportional to net pairwise character distance. Although fast and easy to implement for large molecular datasets, algorithmic approaches are often not able to display the distances exactly on the tree and thus they fail to make use of the full sequence information content.

A second set of methods, under the umbrella term "optimality approaches", assess the optimal tree in the full tree-space using predefined criteria. This includes minimum evolution, which optimises the tree that minimises the sum of pairwise distances as expressed on the tree (Kidd and Sgaramella-Zonta 1971; Rzhetsky and Nei 1993), maximum parsimony (or simply parsimony) (Farris 1970; Fitch 1971), and maximum likelihood (ML) (Felsenstein 2004) approaches. Unlike minimum evolution, which uses pairwise distances, the latter two approaches directly display the nucleotide differences in the sequences on the tree. Parsimony methods search for a tree topology that minimises the number of changes required along its branches, while ML identifies the tree that maximises the probability (likelihood) of the sequence data according to an explicitly defined model of sequence evolution (Felsenstein 2004; Wiley and Lieberman 2011). While computationally demanding, optimality approaches for tree inference permit the direct comparison of trees and make better use of the information contained in sequences by differentiating among different types of nucleotide substitutions. Popular parsimony and ML tree estimation programs include Phylogenetic Analysis Using Parsimony and other methods (PAUP*) (Swofford 2002), Tree analysis using New Technology (TNT) (Goloboff et al. 2008), PhyML (Guindon et al. 2010), RAxML (Stamatakis et al. 2005; Stamatakis 2014), IQ-TREE (Nguy-

en et al. 2015), and IQ-TREE2 (Minh et al. 2020). It should be noted that due to the rapid growth of the space of possible trees (Felsenstein 2004), optimality methods must rely on heuristics when the number of sequences are more than 11, depending on the complexity of the data and the method used.

Bayesian methods use Bayes' theorem to estimate the probability of a tree (including topology, branch lengths, and parameters in the underlying model of sequence evolution) given the alignment data (Wiley and Lieberman 2011). Bayesian approaches identify the trees with the highest posterior probability in a landscape of possible trees and model parameter values. Rather than exhaustively computing the posterior probability for all possible tree hypotheses, a memoryless process called Markov Chain Monte Carlo (MCMC) is used to sample the tree-space and simulate the posterior distribution of parameters, including tree topologies and branch lengths. Popular Bayesian tree estimation programs include MrBayes (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003), RevBayes (Höhna et al. 2016), BEAST (Suchard et al. 2018), and BEAST2 (Bouckaert et al. 2019).

Applications of the multispecies coalescent model in systematics

An important finding when sequencing multiple loci across different accessions was that a set of genes for the same group of taxa often supports different branching patterns in the gene trees. A number of phenomena are responsible for this discord among gene trees, including the incomplete sorting of ancestral polymorphisms (incomplete lineage sorting or ILS), gene duplication and loss, horizontal gene transfer, and branch length heterogeneity (Degnan and Rosenberg 2009; Edwards 2009; Heinrich et al. 2009). Incomplete lineage sorting is a ubiquitous source of gene-tree discordance (Carstens and Knowles 2007) and is modelled explicitly in the multispecies coalescent (MSC) model, an extension of the single-population n-coalescent model (Kingman 1982).

The MSC models ILS by assuming that the degree of incongruence among gene-trees is positively related to effective population size and negatively related to the times between lineage divergences (Felsenstein 2004; Hein et al. 2004). In addition to allowing for discord amongst gene trees, the MSC also allows for the estimation of ancestral demographic parameters (Rannala et al. 2020). Broadly, there are two principal approaches for estimating species trees under the MSC model. The first approach, commonly referred to as summary methods, involves an estimation of the individual gene trees which are in turn used as input data for species-tree inference (Liu et al. 2010; Mirarab et al. 2014, 2015). Nute et al. (2018) provide a comparison of the consistency of some of the more popular methods using this approach. Interestingly, Yan et al. (2022) showed that – at least under some circumstances – these methods perform equally well for gene duplication/loss scenarios as for ILS.

The second approach, commonly referred to as co-estimation methods, uses sequence alignments as input data such that gene and species trees can be simultaneously estimated (Liu and Pearl 2007; Heled and Drummond 2010; Ogilvie et al. 2017). The main advantage of co-estimation methods is the accuracy in estimation compared to summary methods. However, they are computationally intensive, especially when analysing more than a handful of genes (Mirarab et al. 2014, 2015). In principle, summary and co-estimation approaches both require that the sequences are a priori assigned to the correct species in the MSC sense. However, two developments of the co-estimation methods, BP&P (Yang and Rannala 2014) and DISSECT/STACEY (Jones et al. 2015; Jones 2017a), have relaxed this requirement such that posterior probabilities for MSC species delimitations are obtained by a priori assigning sequences to minimal clusters (e.g., single individuals) that are assumed to belong to a single MSC species.

Further developments in co-estimation methods under the MSC include tracing polyploidization events that enable the inference of species networks (Jones et al. 2013; Jones 2017b; Oxelman et al. 2017; Wen and Nakhleh 2018; Wen et al. 2018). Users of these methods should be aware that MSC-based tree inference methods assume the conditions of the Wright-Fisher model are met for "species" in the MSC sense, or at least approximately met (Fisher 1930; Wright 1931 Hein et al. 2004;).

While the MSC represents a major advance in modern phylogenetics, it accounts for only one source of gene tree discord, which has a number of alternate causes, collectively summarised under the concept of migration, meaning the transfer of alleles between otherwise discrete lineages of alleles. Thus, migration in this meaning will include processes such as hybridization, introgression, horizontal and lateral gene transfer, and admixture. The classic MSC model assumes that speciation is instantaneous, and that all gene flow ceases directly after two lineages diverge (Hein et al. 2004). In most empirical cases, however, speciation is probably gradual, and gene flow may persist between what we recognise as otherwise "good" species (Nosil 2008). Recent years have seen the development of species tree estimation methods which estimate phylogenetic relationships under gene tree discord caused by both incomplete lineage sorting and low levels of migration. These methods model migration in one of two ways; (i) continuously, as in the MSC-with-migration or isolation-with-migration (IM) models (Hey and Nielsen 2004), and (ii) discretely, as in the MSC with introgression (MSC) (Flouri et al. 2020) or multispecies network coalescent models (Wen et al. 2018; Rannala et al. 2020), where migration is confined to specific branches on the network. In addition to the estimation of population size and speciation times in the classic MSC model, IM models estimate one continuous rate of migration in either direction for each pair of branches in the species tree. The MSCi model is more parameter-rich because, in addition to the usual MSC parameters, it also estimates migration times and migration probabilities across the species tree (Flouri et al. 2020; Rannala et al. 2020). Bayesian MCMC implementations of the IM model include IMa3 (Hey 2010; Hey et al. 2018), AIM (Müller et al. 2018), and DENIM (Divergence estimation notwithstanding ILS and migration) (Jones 2019), and for the MSCi model, PhyloNet (Wen and Nakhleh 2018) and SpeciesNetwork (Zhang et al. 2018).

The larger number of parameters estimated in IM and MSCi models relative to classic MSC methods improves the biological realism with which the evolutionary process is modelled, but also necessitates a larger number of loci for reliable parameter estimation (Chung 2019; Jones 2017a; Rannala et al. 2020). Realistically, it is possible to use up to about 200 loci with these methods (Rannala et al. 2020), and potentially (perhaps especially for MSCi methods) more are needed for robust estimates. Using less complex, classic MSC-based methods may be an attractive option from a computational perspective, but Leaché et al. (2014) and Müller et al. (2018) showed that failing to account for gene flow where migration has occurred in a group's evolutionary history leads to the underestimation of branch lengths, and, under certain conditions, errors in topology estimation. This result is intuitive; migration is a homogenising force between diverging lineages, and where lineages exchange genes post-divergence but this is not modelled explicitly (i.e., in the standard MSC), all incongruence among gene trees is assumed to be the product of ILS alone. Because the coalescence of alleles in the MSC are always modelled as occurring before two lineages diverge, divergence times are biassed toward the present. This divergence time bias becomes progressively worse (i.e., divergence time estimates become shallower) as migration increases (Chung 2019; Müller et al. 2021). Where there is pervasive evidence of historical gene flow among lineages (as evidenced from, for example, admixture-aware STRUCTURE analyses), and where many, long sequence alignments are available, IM and MSCi models should deliver more accurate species-trees estimates than classical MSC-based methods.

A special form of discrete modelling of migration is posed by allopolyploidy, for which the reader is referred to Chapter 17 Species delimitation for a description of available methodologies.

From phylogenetics to taxonomy

The introduction of the concept of monophyly (Hennig 1950) has been an important philosophical development in systematics in general, and in phylogenetics and taxonomy in particular, since it provides a framework for scientists to rigorously test a hypothesis using empirical data. The concept of monophyly represents a testable criterion to discover natural phylogenetic groups (Mishler and Wilkins 2018). In using the monophyly criterion, it is not necessary to formally name all monophyletic groups identified in a particular phylogenetic hypothesis. Rather, the degree of corroboration, as well as other aspects, such as morphological diagnosability, taxonomic conservatism, ecology, geography, and physiology should ideally be optimised when formal taxonomic decisions are made in an integrative taxonomy approach (Dayrat 2005; Padial et al. 2010).

The integrative taxonomy approach attempts to integrate and use information from several different sources (i.e., morphological, chemical, genomic, ecological, etc.) in order to rigorously delineate species and other taxa. However, this approach has received criticism due to the lack of a clear conceptual and methodological framework, particularly with reference to quantitative criteria. Thus, grouping (i.e., recognition of monophyletic groups) precedes ranking (i.e., choice of level for naming and formal ranking in the taxonomic hierarchy).

The taxon-tree contains clades (= monophyletic groups) of various inclusiveness that may be named and given a rank according to the rules of nomenclature. This ranking process is in principle arbitrary, but various auxiliary criteria, such as the different versions of the phylogenetic species concept (Gutiérrez and Garbino 2018) may be used to rank species as clades. Although these criteria may enhance the stability of particular species delimitations, it is unlikely that they will be universally applicable. Thus, there will always be an aspect of subjectivity in the species-as-clades approach, which hampers the usefulness of using the number of species as

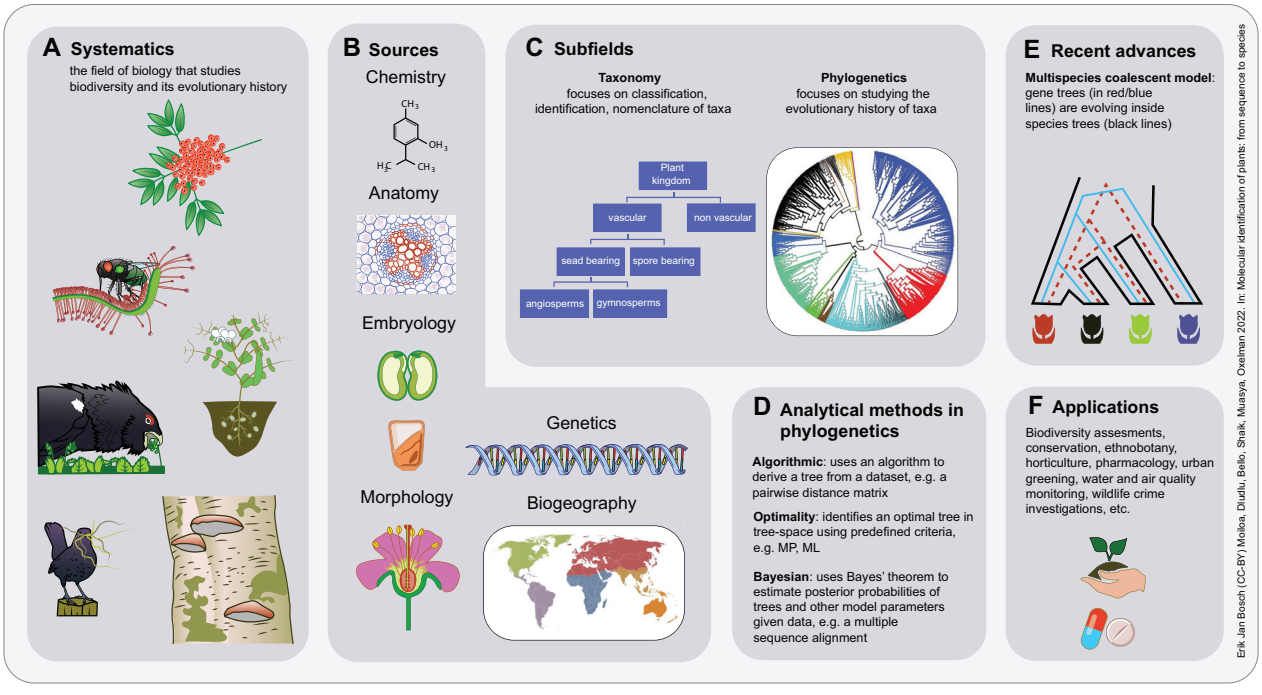

Figure 1. Chapter 19 Infographic: Visual representation of the content of this chapter.

diversity indicators (see Chapter 17 Species delimitation). In contrast, the MSC offers a precise and non-arbitrary definition of species-as-branches (Degnan and Rosenberg 2009). However, the particular discovery of such units are heavily dependent on the parameters included in the MSC model. The most elaborate co-estimation methods to date (e.g., STACEY, BP&P) assume no migration between the branches and instantaneous isolation, and no structure within the branches. Violation of these assumptions will lead to inconsistencies. For example, the addition of more data (i.e., loci) will inflate the number of species recognized (Leaché et al. 2019). On the other hand, lack of information (i.e., no substitutions sampled) in recently diverged groups may lead to underestimations, even if the assumptions are met. Moreover, incomplete sampling is also an important factor to consider.

Application of the MSC model enables rigorous scientific testing of monophyly hypotheses using multi-locus sequence data. Applying species rank to certain clades is valid but will need auxiliary criteria to reduce subjectivity (Gutiérrez and Garbino 2018). This means that the use of species numbers as an objective unit for diversity must be treated with caution, even if an increased application (relative to the vast numbers of species described from morphological diagnosability only) of species-as-clades under MSC may potentially reduce subjectivity. Alternatively, the MSC offers an objective definition of species, given the parameterization of the model. However, for both theoretical and practical reasons, the application of such a concept is premature, and not likely to meaningfully improve biodiversity measures at this point.

Questions

1. Are the terms "systematics" and "taxonomy" synonymous? If not, how do they differ?
2. Discuss the differences between a taxonomy based on hypothetical evolutionary relationships and one based on the possession of certain traits.
3. What is the major difference between discrete and continuous phylogenetic models allowing for migration?

Glossary

Biodiversity – The variety of living organisms encompassed in all forms.
Branch (= edge) – A part of a phylogenetic tree that connects different nodes (= vertices) or terminals (= leaves).
Clade – A part of a phylogenetic tree made up of a common ancestor including all its descendants.
Effective population size – Describes the size of an ideal Wright-Fisher population, containing exactly the equivalent genetic diversity and/or experiencing exactly the same genetic drift as the population surveyed irrespective of its census population size.
Gene flow (= migration) – The transfer of alleles between populations due to various biological processes.
Gene-tree (= genealogy) – A tree representing the evolutionary history of a particular gene.
Gene-tree discordance (= phylogenetic incongruence) – A phenomenon where evolutionary trees from individual genes result in conflicting branching patterns.
Homology – The shared similarity due to descent from a common ancestor.
Horizontal gene transfer (= lateral gene transfer) – The transfer of genetic material through a biological process other than sexual reproduction.

Incomplete lineage sorting (= deep coalescence) – The failure of ancestral gene copies of two or more lineages in a population to coalesce within the population branch.
Lineage – In phylogenetics, a group of populations connected by a single line of descent from a common ancestor.
Metapopulation – A group of spatially separated populations which share the same evolutionary history.
Monophyly – A relationship where descendants of a common ancestor form a single clade.
Phenotypic variation – The variability in the observable expressed and environmentally affected features that exists in a population.
Polymorphism – A phenomenon where a trait has more than one expression.
Polyploidization – A biological process which a single genome undergoes to possess more than two sets of chromosomes.
Posterior probability – An estimation of the probability of a hypothesis given the data, a stochastic model and prior expectations.
Sequence alignment (= alignment) – The process of arranging DNA sequences in order to identify homologous positions.
Species delimitation (= delimitation) – The process of analytically identifying boundaries of species using empirical data.
Species-tree – A tree showing the evolutionary branching history of ancestral to descendant populations.
Species-tree inference – The process of estimating branching history of populations.
Substitution – A change from one nucleotide to another that results in a change in the DNA sequence.
Topology (= tree topology) – The branching pattern and order of nodes on an evolutionary tree.
Tree-space – A collection of all possible trees for a given set of input sequences.
Wright-Fisher model – A model where alleles are sampled from a population which is characterised by random reproduction, no selection, and no overlap between generations.

References

Barraclough TG (2019) The evolutionary biology of species. Oxford University Press, Oxford.

Bell AD, Bryan Alan (1991) Plant form: an illustrated guide to flowering plant morphology. Timber Press, Portland, Oregon.

Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, Heled J, Jones G, Kühnert D, De Maio N, Matschiner M, Mendes FK, Müller NF, Ogilvie HA, du Plessis L, Popinga A, Rambaut A, Rasmussen D, Siveroni I, Suchard MA, Drummond AJ (2019) BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. PLoS Comput. Biol. 15, e1006650. https://doi.org/10.1371/journal.pcbi.1006650

Carstens BC, Knowles LL (2007) Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from *Melanoplus* grasshoppers. Syst. Biol. 56, 400–411. https://doi.org/10.1080/10635150701405560

Chung Y (2019) Recent advances in Bayesian inference of isolation-with-migration models. Genomics Inform. 17, e37. https://doi.org/10.5808/GI.2019.17.4.e37

Cotton JA (2016) From sequence reads to evolutionary inferences, in: Olson, P.D., Hughes, J., Cotton, J.A. (Eds.), The Systematics Association Special: Next Generation Systematics. Cambridge University Press, Cambridge, pp. 305–335. https://doi.org/10.1017/CBO9781139236355.016

Cronquist A (1977) On the taxonomic significance of secondary metabolites in angiosperms, in: Kubitzki, K. (Ed.), Flowering Plants. Springer, Vienna, pp. 179–189. https://doi.org/10.1007/978-3-7091-7076-2_12

Dayrat B (2005) Towards integrative taxonomy. Biological Journal of the Linnean Society 85, 407–415. https://doi.org/10.1111/j.1095-8312.2005.00503.x

Degnan JH, Rosenberg NA (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24, 332–340. https://doi.org/10.1016/j.tree.2009.01.009

de Queiroz K (2005) A unified concept of species and its consequences for the future of taxonomy. Proceedings of the California Academy of Sciences, ser. 4 56(Suppl. I), 196–215.

de Queiroz K (2007) Species concepts and species delimitation. Syst. Biol. 56, 879–886. https://doi.org/10.1080/10635150701701083

Dickison WC (2000) Integrative plant anatomy, 1st ed. Academic Press, San Diego.

Edwards SV (2009) Is a new and general theory of molecular systematics emerging? Evolution 63, 1–19. https://doi.org/10.1111/j.1558-5646.2008.00549.x

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6, e19379. https://doi.org/10.1371/journal.pone.0019379

Farris JS (1970) Methods for computing wagner trees. Syst. Biol. 19, 83–92. https://doi.org/10.1093/sysbio/19.1.83

Felsenstein J (2004) Inferring phylogenies, 2nd ed. Sinauer Associates, Sunderland, Massachusetts.

Fisher RA (1930) The genetical theory of natural selection. Clarendon Press, Oxford. https://doi.org/10.5962/bhl.title.27468

Fitch WM (1971) Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Biol. 20, 406–416. https://doi.org/10.1093/sysbio/20.4.406

Flouri T, Jiao X, Rannala B, Yang Z (2020) A Bayesian implementation of the multispecies coalescent model with introgression for phylogenomic analysis. Mol. Biol. Evol. 37, 1211–1223. https://doi.org/10.1093/molbev/msz296

Goloboff PA, Farris JS, Nixon KC (2008) TNT, a free program for phylogenetic analysis. Cladistics 24, 774–786. https://doi.org/10.1111/j.1096-0031.2008.00217.x

Guerra M (2008) Chromosome numbers in plant cytotaxonomy: concepts and implications. Cytogenet. Genome Res. 120, 339–350. https://doi.org/10.1159/000121083

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59, 307–321. https://doi.org/10.1093/sysbio/syq010

Gutiérrez EE, Garbino GST (2018) Species delimitation based on diagnosis and monophyly, and its importance for advancing mammalian taxonomy. Zool. Res. 39, 301–308. https://doi.org/10.24272/j.issn.2095-8137.2018.037

Heinrich M, Edwards S, Moerman DE, Leonti M (2009) Ethnopharmacological field studies: a critical assessment of their conceptual basis and methods. J. Ethnopharmacol. 124, 1–17. https://doi.org/10.1016/j.jep.2009.03.043

Hein J, Schierup M, Wiuf C (2004) Gene genealogies, variation and evolution: a primer in coalescent theory. Oxford University Press, Oxford.

Heled J, Drummond AJ (2010) Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27, 570–580. https://doi.org/10.1093/molbev/msp274

Hennig W (1950) Grundzüge einer Theorie der phylogenetischen Systematik. Deutscher Zentralverlag, Berlin.

Heslop-Harrison JSP, Schwarzacher T (2011) Organisation of the plant genome in chromosomes. Plant J. 66, 18–33. https://doi.org/10.1111/j.1365-313X.2011.04544.x

Hey J, Chung Y, Sethuraman A, Lachance J, Tishkoff S, Sousa VC, Wang Y (2018) Phylogeny estimation by integration over isolation with migration models. Mol. Biol. Evol. 35, 2805–2818. https://doi.org/10.1093/molbev/msy162

Hey J, Nielsen R (2004) Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. Genetics 167, 747–760. https://doi.org/10.1534/genetics.103.024182

Hey J (2010) Isolation with migration models for more than two populations. Mol. Biol. Evol. 27, 905–920. https://doi.org/10.1093/molbev/msp296

Höhna S, Landis MJ, Heath TA, Boussau B, Lartillot N, Moore BR, Huelsenbeck JP, Ronquist F (2016) RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Syst. Biol. 65, 726–736. https://doi.org/10.1093/sysbio/syw021

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17, 754–755. https://doi.org/10.1093/bioinformatics/17.8.754

Johri BM, Ambegaokar KB, Srivastava PS (1992) Comparative embryology of angiosperms. Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-76395-3

Jones G, Aydin Z, Oxelman B (2015) DISSECT: an assignment-free Bayesian discovery method for species delimitation under the multispecies coalescent. Bioinformatics 31, 991–998. https://doi.org/10.1093/bioinformatics/btu770

Jones G, Sagitov S, Oxelman B (2013) Statistical inference of allopolyploid species networks in the presence of incomplete lineage sorting. Syst. Biol. 62, 467–478. https://doi.org/10.1093/sysbio/syt012

Jones G (2017a) Algorithmic improvements to species delimitation and phylogeny estimation under the multispecies coalescent. J. Math. Biol. 74, 447–467. https://doi.org/10.1007/s00285-016-1034-0

Jones G (2017b) Bayesian phylogenetic analysis for diploid and allotetraploid species networks. BioRxiv. https://doi.org/10.1101/129361

Jones G (2019) Divergence estimation in the presence of incomplete lineage sorting and migration. Syst. Biol. 68, 19–31. https://doi.org/10.1093/sysbio/syy041

Jones RN (1995) B chromosomes in plants. New Phytol. 131, 411–434. https://doi.org/10.1111/j.1469-8137.1995.tb03079.x

Judd WS, Campbell CS, Kellogg EA, Stevens PF, Donoghue MJ (Eds) (2007) Plant systematics: a phylogenetic approach, 3rd ed. Sinauer Associates Inc., Sunderland, MA.

Kidd KK, Sgaramella-Zonta LA (1971) Phylogenetic analysis: concepts and methods. Am. J. Hum. Genet. 23, 235–252.

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217, 624–626. https://doi.org/10.1038/217624a0

Kingman JFC (1982) The coalescent. Stoch. Process. Their Appl. 13, 235–248. https://doi.org/10.1016/0304-4149(82)90011-4

Leaché AD, Fujita MK, Minin VN, Bouckaert RR (2014) Species delimitation using genome-wide SNP data. Syst. Biol. 63, 534–542. https://doi.org/10.1093/sysbio/syu018

Leaché AD, Zhu T, Rannala B, Yang Z (2019) The spectre of too many species. Syst. Biol. 68, 168–181. https://doi.org/10.1093/sysbio/syy051

Lemmon AR, Emme SA, Lemmon EM (2012) Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst. Biol. 61, 727–744. https://doi.org/10.1093/sysbio/sys049

Liu H, Feng C-L, Luo Y-B, Chen B-S, Wang Z-S, Gu H-Y (2010) Potential challenges of climate change to orchid conservation in a wild orchid hotspot in southwestern China. Bot. Rev. 76, 174–192. https://doi.org/10.1007/s12229-010-9044-x

Liu L, Pearl DK (2007) Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst. Biol. 56, 504–514. https://doi.org/10.1080/10635150701429982

Mayden RL (1997) A hierarchy of species concepts: the denouement in the saga of the species problem, in: Claridge, M.F., Dawah, H.A., Wilson, M.R. (Eds.), Species: The Units of Diversity. Chapman & Hall, pp. 381–423.

Mayr E (1942) Systematics and the origin of species from the viewpoint of a zoologist. Columbia University Press, New York.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R (2020) IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. 37, 1530–1534. https://doi.org/10.1093/molbev/msaa015

Mirarab S, Nguyen N, Guo S, Wang L-S, Kim J, Warnow T (2015) PASTA: ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. J. Comput. Biol. 22, 377–386. https://doi.org/10.1089/cmb.2014.0156

Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T (2014) ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30, i541-8. https://doi.org/10.1093/bioinformatics/btu462

Mishler BD, Wilkins JS (2018) The hunting of the SNaRC: a snarky solution to the species problem. Philosophy, Theory, and Practice in Biology 10, 1–18. https://doi.org/10.3998/ptpbio.16039257.0010.001

Mishler BD (2013) History and theory in the development of phylogenetics in botany, in: Hamilton, A. (Ed.), Evolution of Phylogenetic Systematics. University of California Press, pp. 188–210. https://doi.org/10.1525/california/9780520276581.003.0009

Moore PD, Collinson M, Webb JA (1994) Pollen analysis, 2nd ed. Wiley, Oxford.

Morrison DA (2018) Multiple sequence alignment is not a solved problem. arXiv. https://doi.org/10.48550/arXiv.1808.07717

Müller NF, Ogilvie H, Zhang C, Drummond A, Stadler T (2018) Inference of species histories in the presence of gene flow. BioRxiv. https://doi.org/10.1101/348391

Müller NF, Ogilvie H, Zhang C, Fontaine MC, Amaya-Romero JE, Drummond A, Stadler T (2021) Joint inference of species histories and gene flow. bioRxiv. https://doi.org/10.1101/348391

Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H (1986) Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harb. Symp. Quant. Biol. 51 Pt 1, 263–273. https://doi.org/10.1101/SQB.1986.051.01.032

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32, 268–274. https://doi.org/10.1093/molbev/msu300

Nosil P (2008) Speciation with gene flow could be common. Mol. Ecol. 17, 2103–2106. https://doi.org/10.1111/j.1365-294X.2008.03715.x

Nute M, Chou J, Molloy EK, Warnow T (2018) The performance of coalescent-based species tree estimation methods under models of missing data. BMC Genomics 19, 286. https://doi.org/10.1186/s12864-018-4619-8

Ogilvie HA, Bouckaert RR, Drummond AJ (2017) Starbeast2 brings faster species tree inference and accurate estimates of substitution rates. Mol. Biol. Evol. 34, 2101–2114. https://doi.org/10.1093/molbev/msx126

Oxelman B, Brysting AK, Jones GR, Marcussen T, Oberprieler C, Pfeil BE (2017) Phylogenetics of allopolyploids. Annu. Rev. Ecol. Evol. Syst. 48, 543–557. https://doi.org/10.1146/annurev-ecolsys-110316-022729

Padial JM, Miralles A, De la Riva I, Vences M (2010) The integrative future of taxonomy. Front. Zool. 7, 16. https://doi.org/10.1186/1742-9994-7-16

Radford AE, Dickison WC, Massey JR, Bell CR (1974) Vascular plant systematics, 1st ed. HarperCollins.

Rannala B, Edwards SV, Leaché A, Yang Z (2020) The multi-species coalescent model and species tree inference, in: Scornavacca, C., Delsuc, F., Galtier, N. (Eds.), Phylogenetics in the Genomic Era. p. 3.3:1-3.3:21.

Rannala B, Yang Z (2003) Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164, 1645–1656. https://doi.org/10.1093/genetics/164.4.1645

Reitsma Tj (1970) Suggestions towards unification of descriptive terminology of angiosperm pollen grains. Rev. Palaeobot. Palynol. 10, 39–60. https://doi.org/10.1016/0034-6667(70)90021-7

Reynolds T (2007) The evolution of chemosystematics. Phytochemistry 68, 2887–2895. https://doi.org/10.1016/j.phytochem.2007.06.027

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574. https://doi.org/10.1093/bioinformatics/btg180

Rzhetsky A, Nei M (1993) Theoretical foundation of the minimum-evolution method of phylogenetic inference. Mol. Biol. Evol. 10, 1073–1095. https://doi.org/10.1093/oxfordjournals.molbev.a040056

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425. https://doi.org/10.1093/oxfordjournals.molbev.a040454

Sanger F, Coulson AR (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J. Mol. Biol. 94, 441–448. https://doi.org/10.1016/0022-2836(75)90213-2

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74, 5463–5467. https://doi.org/10.1073/pnas.74.12.5463

Sharma O (1993) Plant taxonomy, 2nd ed. McGraw-Hill India.

Simpson MG (2019) Plant systematics, 3rd ed. Academic Press, Burlington, MA.

Singh G (2019) Plant systematics: an integrated approach, 4th ed. CRC Press, Boca Raton. https://doi.org/10.1201/9780429289521

Sokal RR, Crovello TJ (1970) The biological species concept: a critical evaluation. Am. Nat. 104, 127–153. https://doi.org/10.1086/282646

Sokal RR, Michener CD (1958) A statistical method for evaluating systematic relationships. University of Kansas Science Bulletin 28, 1409–1438.

Stamatakis A, Ludwig T, Meier H (2005) RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21, 456–463. https://doi.org/10.1093/bioinformatics/bti191

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Stuessy TF (2009) Plant taxonomy: the systematic evaluation of comparative data, 2nd ed. Columbia University Press, New York.

Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A (2018) Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol. 4, vey016. https://doi.org/10.1093/ve/vey016

Swofford DL (2002) PAUP*: phylogenetic analysis using parsimony (* and other methods). Version. 4. Sinauer Associates, Sunderland, Massachusetts.

Turland N, Wiersema J, Barrie F, Greuter W, Hawksworth D, Herendeen P, Knapp S, Kusber W-H, Li D-Z, Marhold K, May T, McNeill J, Monro A, Prado J, Price M, Smith G (Eds) (2018) International Code of Nomenclature for algae, fungi, and plants, Regnum Vegetabile. Koeltz Botanical Books. https://doi.org/10.12705/Code.2018

Turner BL (1969) Chemosystematics: recent developments. Taxon 18, 134–151. https://doi.org/10.2307/1218672

Walker JW, Doyle JA (1975) The bases of angiosperm phylogeny: palynology. Ann Mo Bot Gard 62, 664. https://doi.org/10.2307/2395271

Wen D, Nakhleh L (2018) Coestimating reticulate phylogenies and gene trees from multilocus sequence data. Syst. Biol. 67, 439–457. https://doi.org/10.1093/sysbio/syx085

Wen D, Yu Y, Zhu J, Nakhleh L (2018) Inferring phylogenetic networks using phylonet. Syst. Biol. 67, 735–740. https://doi.org/10.1093/sysbio/syy015

Wiley EO, Lieberman BS (2011) Phylogenetics: theory and practice of phylogenetic systematics, 2nd ed. John Wiley & Sons, Inc., Hoboken, New Jersey.

Wiley EO (1978) The evolutionary species concept reconsidered. Syst. Zool. 27, 17–26. https://doi.org/10.2307/2412809

Wright S (1931) Evolution in Mendelian populations. Genetics 16, 97–159.

Yang Z, Rannala B (2014) Unguided species delimitation using DNA sequence data from multiple Loci. Mol. Biol. Evol. 31, 3125–3135. https://doi.org/10.1093/molbev/msu279

Yan Z, Smith ML, Du P, Hahn MW, Nakhleh L (2022) Species tree inference methods intended to deal with incomplete lineage sorting are robust to the presence of paralogs. Syst. Biol. 71, 367–381. https://doi.org/10.1093/sysbio/syab056

Zachos FE (2016) Species concepts in biology: historical development, theoretical foundations and practical relevance, 1st ed. Springer, New York, NY.

Zhang C, Ogilvie HA, Drummond AJ, Stadler T (2018) Bayesian inference of species networks from multilocus sequence data. Mol. Biol. Evol. 35, 504–517. https://doi.org/10.1093/molbev/msx307

Answers

1. Although some use the terms interchangeably, they have different meanings. Systematics is the study of diversification of organisms and their relationships through time, whereas taxonomy refers to the theory and practice of identifying, describing, naming, and classifying organisms, which is an integral part of systematics.
2. In the first case, real entities, which form parts of the evolutionary history of the lineages are considered. These hypotheses may be wrong or correct, because there is only one history. In the second approach, taxa can be classified under a multitude of different traits, and each of them may be more or less useful for certain purposes, but there is no unique correct classification.
3. In discrete models, migration only occurs during specific periods of time, resulting in extra, merging branches of the network. In continuous models, migration is a continuous process treated as a parameter in the branching model.

Chapter 20
Museomics

Nataly Allasi Canales[1,2*], Nina Rønsted[3,2], Jazmin Ramos-Madrigal[4,5*]

1 Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark
2 Royal Botanic Gardens, Kew, United Kingdom
3 National Tropical Botanical Garden, Kalaheo, Hawaii, USA
4 Section for Evolutionary Genomics, The GLOBE Institute, University of Copenhagen, Copenhagen, Denmark
5 Center for Evolutionary Hologenomics, The GLOBE Institute, University of Copenhagen, Copenhagen, Denmark

Nataly O. Allasi Canales **allasicanales@gmail.com**
Nina Rønsted **nronsted@ntbg.org**
Jazmin Ramos-Madrigal **jazmingem@gmail.com**

* These authors contributed equally.

Citation: Canales NA, Rønsted N, Ramos-Madrigal J (2022) Chapter 20. Museomics. In: de Boer H, Rydmark MO, Verstraete B, Gravendeel B (Eds) Molecular identification of plants: from sequence to species. Advanced Books. https://doi.org/10.3897/ab.e98875

Museomics in plant research

Have you ever wondered how museum collections can be used for answering fundamental questions about biodiversity and its evolution across space and time? Natural history museums harbour ~3 billion biological specimens that are often linked with a specific collection time and place (Wheeler et al. 2012). These specimens show traits and contain biomolecules including DNA, proteins and lipids that can be analysed to learn more about a specimen's evolutionary history, ecology, and response to environmental change (Cappellini et al. 2018). In this chapter, we explore how the field of museomics can help us in analysing the vast richness of museum collections' metadata, and how current technologies and analyses can unveil past and present biodiversity and its evolution.

Museomics is the study of biological material from museum collections using genomic techniques that allow the reconstruction of partial or complete genomes. In contrast to single-loci PCR-based approaches, these genomic techniques provide information on a genome-wide scale that can, for example, be used to assess evolutionary and ecological processes (Gutaker and Burbano 2017). Thus, genomic methods have increased the potential of natural history collections, making them and their associated information even more accessible and relevant, as they are a valuable source for answering unanticipated questions about public health (Dunnum et al. 2017), invasive species, food security (Schindel and Cook 2018), etc., in addition to their primary use as documentation of biodiversity identification and distribution (Funk 2004).

Although different in age and preservation state, historical DNA from museum collections (typically > 200 years old) can have similar characteristics as ancient DNA (aDNA, typically < 200 years old) such as post-mortem degradation patterns (Raxworthy and Smith 2021). In fact, the DNA decay rate in post-mortem herbarium material is six times higher than in ancient bones (Weiß et al. 2016). As a consequence, laboratory methods and bioinformatic approaches to study these samples are often similar to those used in aDNA studies more broadly (Chapter 2 DNA from museum collections, Chapter 8 aDNA from sediments, and Chapter 21 Palaeobotany).

The plant material available in museum collections is an indispensable source of genetic information for species that are extinct (Van de Paer et al. 2016; Zedane et al. 2016), that are difficult to collect from the wild (Malakasi et al. 2019) , and that have only been collected once (Silva et al. 2017). Over the past decade, museomics studies have explored evolutionary questions that have resulted in resolved phylogenies, assembling genomes from herbarium specimens, a better understanding of crop domestication, ecological processes, and host-pathogen interactions (e.g., Bieker and Martin 2018; Meineke et al. 2018; Rønsted et al. 2020). In addition to their scientific and applied applications, museum collections-based research also provides fascinating stories that continue to inspire and engage different museum audiences through educational programs, newsletters, social media, and visitor experiences.

Plant material in museum collections

Herbaria collections include a variety of sample types including herbarium specimens, seeds, wood or xylarium samples, flowers and fruits in alcohol or desiccated, and biocultural or ethnobotanical collections gathered over hundreds of years. They may (1) originate from general collections seeking to represent the world's biodiversity, (2) have been deposited as vouchers,

or (3) have served as reference material for a specific study (Culley 2013). Some of these collections also have linked archives such as field books, letters, and illustrations. For museomic approaches, three types of collections are particularly relevant: herbarium, xylarium, and economic botany/ethnobotanical collections.

Herbarium specimens generally consist of a pressed plant mounted on acid-free paper. They ideally include leaves, stems, flowers and/or fruits, and roots when possible, and have the necessary plant parts for unambiguous identification. The metadata associated with a specimen should at minimum include the binomial scientific name, who collected it, the collection date, locality, and a unique number. Additional information may include a description of the habitat, and associated plants as well as any other details that cannot be observed from the dried specimen, including specimen's colours and smell at the time of collection, and any observed visiting insects (Liesner 1990).

Xylarium, or wood collections, comprise a collection of different wood parts of a tree and such specimens can inform forensics, timber trade, and conservation efforts. A typical specimen is wood stripped from bark, when present, and has the shape of a book. Some collections can also consist of cross-sections, which can provide valuable ecological and anatomical information than the book-shaped wood (Salick et al. 2014).

Economic botany or biocultural specimens include economically useful plant parts such as fruits, barks, seeds, bark clothes, baskets, and papers for medicine, religious, entertainment, and commercial purposes (Salick et al. 2014). These collections include a vast range of ethnobiological specimens, artefacts, and archives that represent the connection between people and their environment. Biocultural specimens also contain information about their uses and individuals who acquired and documented the items.

Museum collections: fundamental and applied research

The next paragraphs are examples of the applications and impact of museomics research, as well as the potentially negative implications of unethical use of collections on local communities.

Validation of historical identifications

DNA analysis of plant material in museum collections has allowed us to improve their taxonomic annotations and their corresponding scientific value. By analysing their genomes, it is possible to assign taxonomic information to samples that cannot be reliably identified morphologically or that no longer exist. For example, genomics was used for the identification of both endangered and extinct species of Hawaiian endemic mints and the now considered extinct *Hesperelaea palmeri* (Van de Paer et al. 2016; Welch et al. 2016; Zedane et al. 2016). The genomic information from museum specimens can also be used to confirm the taxonomic information already present in the archives, or to identify potential misclassifications (Goodwin et al. 2015). Finally, through museomics it is possible to go beyond a species-level taxonomic classification since genomic data can be used to identify the subspecies, variety, cultivar or population of a given plant. An example of this was the characterisation of the genetic profile of a 90-year-old grapevine specimen present in a herbarium collection at the Natural History Museum of Split,

Croatia (Malenica et al. 2011). By comparing its genetic profile with those of modern grapevine varieties, it was determined that it belonged to the Zinfandel variety.

Unravelling evolutionary processes

Genetic analysis of museum samples has increasingly been used to describe evolutionary processes shaping the genetic diversity, population structure, phylogenetic history, and demography of plants. Both Sanger and high-throughput sequencing have been used to obtain partial and complete genomes of plants in museum collections. Combining the genetic analysis with the information contained in their associated metadata increases the scope of the evolutionary inferences that can be made. Information about the collection date and geographic location can be used to directly measure changes in genetic diversity across time and space, the effect of climate change, domestication, human environmental disturbance, and other natural phenomena (Funk 2018; Funk et al. 2009; James et al. 2018).

One of the principal applications of museomics has been reconstructing plant species' phylogenetic histories. Understanding plants' evolutionary relationships can help refine their taxonomic classification, identify their potential geographic and evolutionary origins, and make predictions on their chemical properties and potential future applications (Ernst et al. 2016). For example, studies based on partial and complete plastid genomes from museum specimens from Malagasy grasses (Besnard et al. 2014) and ragweed (Martin et al. 2018; Sánchez Barreiro et al. 2017) have been used to reconstruct the phylogenetic history of their respective species. For extinct species, using genetic data derived from museum specimens is the only way to reconstruct their phylogenetic history. This was done for two extinct species of Hawaiian mint, where museum specimens were used to complete a comparative chloroplast genome analysis showing present-day mint species are the result of recent genomic radiation (Welch et al. 2016).

Additionally, herbarium material has been used to identify and measure the extent of gene flow (i.e., exchange genetic material through interbreeding) among plant populations such as that occurring between different species of ragweed (Martin et al. 2018). By combining a specimen's metadata with its genetic data, it is possible to measure evolutionary processes through time.

An interesting aspect is the evolution of plants under domestication (i.e., the process through which wild plants became today's crops). Most of the plants (in volume) that we consume today as food or use in the production of plant-based products are the result of domestication. Museomics has made important contributions to the study of the geographic origins, dispersal patterns, and selective evolution of domesticated species. Herbarium specimens have been used to trace the origin of the European potatoes in the Andes (Gutaker et al. 2019) and to identify the dispersion routes of plant species including invasive weeds (Hardion et al. 2014; Martin et al. 2014; Payacan et al. 2017). Seeds from museum collections have been used to identify the timing of genetic changes associated with domestication. As an example, genomic analysis of seeds showed that the genetic mutation responsible for increased grain size in cultivated spelt got fixed during modern crop improvement and not during its early domestication (Asplund et al. 2010). Kistler et al. (2018) also used genomic approaches on ancient and extant South American maize lineages to investigate the genetic changes that accompanied domestication.

Another aspect where museum specimens can be used to provide valuable insights is in the study of genetic erosion, which is the decrease in genetic diversity over time. Samples collected at different points in time are an ideal and reliable way to directly measure changes in genetic diversity through time and in relation with historical, geographic, and climatic changes (Hart et al. 2016). This makes it possible to identify the potential environmental, climatic, or anthropogenic forces that shape plant genetic diversity.

Resolving ecological processes

Genetic analyses of museum collections can also be used for the study of ecological processes (i.e., the interactions between plants, animals, and abiotic components in an ecosystem). Herbarium collections can help in the characterization of the distribution and abundance of plant species through time and in measuring changes in biodiversity. By combining their genetic data and metadata, we can measure the habitat ranges of species through time and identify possible associations between such changes and climatic or anthropogenic events. In one example, the genetic analysis of the grass *Alopecurus myosuroides* from herbarium collections showed that the genetic variants associated with herbicide resistance in this plant pre-dated the use of herbicides, which confirmed that this resistance did not evolve from anthropogenic events (Délye et al. 2013).

DNA-based inventories using collections

Multiple studies have used herbarium and xylarium specimens to develop DNA barcode libraries of entire floras or for more specific applied uses within forensics, authentication, and conservation (see Chapter 26 Forensic genetics, botany, and palynology; Chapter 23 Food safety; Chapter 22 Healthcare; Chapter 13 Barcoding - High Resolution Melting). Kuzmina et al. (2017) generated DNA barcodes for 95% of all vascular plant species in Canada. Such an extensive reference DNA database allows for linking the genetic information of a species to its species identity (i.e., its name) and the location of the specimens of that species. Methodologically, this study also showed that gene recovery could vary according to the family studied and the age of specimens. In another directly applied example, barcodes of *Dalbergia* (rosewood) xylarium specimens have been developed to help in monitoring illegal logging of this species group (Hassold et al. 2016).

Facilitating crop improvement

Developing plant cultivars with desirable characteristics is essential to guarantee food security in the future. One of the initial stages in improving crops is identifying plants that already have certain beneficial traits that can be used in the breeding process (Swarup et al. 2021). Museum samples can contribute towards identifying and uncovering lost genetic diversity and to understanding how such diversity contributes to phenotypic variation. This can ultimately guide preservation efforts by seed banks. These seed banks can then be used in plant breeding as a genetically diverse resource to improve crops and improve their yield, pathogen resistance, nutritional properties, flavour, or resilience to climate change. Moreover, genetic analysis of museum specimens can be used to pinpoint potential geographic locations of plants with desirable characteristics that can then be used for crop improvement.

Elucidating pathogen-host interactions

Plant pathogens cause diseases and losses at different levels from hunger and famines to the extinction of entire species (Agrios 2009). Plant pathogens may evolve from previously benign organisms, have not been previously described, or have the potential to re-emerge in the future. Modern genomic technologies can help us to accurately understand how ge-

nomes have evolved over time, reducing the need for speculation by inference from modern samples (Yoshida et al. 2015). Species characterization is also possible using museum collections. An early study that analysed 90-year-old asteraceous crops in a herbarium collection was able to isolate a taxon that was unique in morphology, pathogenic specificity, and phylogenetic relationships from other taxa in the *Colletotrichum acutatum* complex (Uematsu et al. 2012).

Previous studies have shown that museomics can answer questions about the evolution and origin of plant pathogens as in the case of the potato late blight, *Phytophthora infestans* (Yoshida et al. 2014) and *Xanthomonas axonopodis* pv. *citri* (Li et al. 2007). Additionally, herbarium genomics research can be used to help in pointing out the origin of an outbreak (Yoshida et al. 2013). Understanding the genomes' evolutionary history and the population structure of major lineages can help to shed light on the ancestry of fast-evolving species like pathogens. This will ultimately allow us to study host-specific adaptation and the evolution of resistance loci in better detail to face future pandemics (Martin et al. 2016).

Bolstering conservation genetics

Studies on endangered species often lack past diversity estimations or rely on inferring past genetic diversity from modern populations. Museum specimens can provide an important perspective for past population evolutionary events and eventually contribute to the conservation of the species through scientifically supported conservation and management recommendations (Hofman et al. 2015; Leonard 2008; Roy et al. 1994).

The conservation of rare and endangered plants relies on a sound understanding of their genetic diversity to ensure the health of both wild and ex situ collections, and to avoid or overcome genetic bottlenecks (Hoban et al. 2020; Wood et al. 2020). For endangered plant species, genomic profiling of museum specimens can help understand past diversity through inclusion of samples from lost or dwindling subpopulations.

Other studies have documented genetic changes of endangered species in response to human disturbance. Cozzolino et al. (2007) assessed genetic variation of an endangered orchid *Anacamptis palustris* from collections before the Second World War and compared them with extant populations. They were able to map once wide-spread haplotypes that are now extinct and highlight the presence of a much wider genetic diversity showing how human-induced habitat changes reduced the genetic diversity of this species (Cozzolino et al. 2007). While museomics can provide scientific background information for conservation planning, translating genomic research into conservation practice remains challenging and requires close dialogue between policy-makers, scientists, and managers to implement the benefits of the research (Hofman et al. 2015).

Ethical museomics, fair and equitable sharing, and co-design of research

International conventions govern accessing, researching, and moving plant material between institutions and countries. Specifically, researchers need to understand and adhere to the Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilisation to the Convention on Biological Diversity.

Additionally, some museums have destructive sampling policies and committees that weigh the pros and cons of destroying precious and unique samples against the possibility of advancing scientific knowledge (Austin et al. 2019). Methods are being developed for non-destructive sampling to support the ethical use of critical samples such as type specimens and other unique and invaluable samples (Shepherd 2017).

Ideally, researchers aiming to use biocultural or other culturally sensitive collections of high human interest should involve and consult with the Indigenous peoples and Local Communities from an early research stage to ensure the fair and equal use of the collections and their associated information, as well as to take advantage of the knowledge related to the cultural value and uses of the specific plants and artefacts. This is becoming a common practice for human paleogenomic studies, but unfortunately not yet for plant material. From the researcher perspective, knowing Indigenous communities' practices and concerns minimises potential unintended cultural harm in the future by paleogenomic studies and can also provide additional advice on relevant research questions to consider (Bardill et al. 2018). Due to historical and contemporary collecting practices, colonialism, and trade, collections are often held in Western institutions and are more or less inaccessible for researchers and other users from the country of origin (Friis and Balslev 2017). In addition, it is crucial to co-design research at the initial stage when possible (more in Chapter 2 DNA from museum collections). Fair and equitable sharing can also involve access and sharing of genomic data and training with the wider scientific and public community, as well as education, outreach, and communication to communities and the general public in local languages.

Finally, museums have experienced continuous reductions in funding and staff resulting in the lack of curatorial expertise and capacity leading to increasingly orphaned collections (Kemp 2015). Researchers should therefore also consider any benefits their research may provide to the curation of the specimens and the importance of helping to communicate and promote the importance of funding and supporting museum collections for the future.

Questions

1. What types of plant materials can be found in museum collections?
2. What are the advantages of using museum collections for genomics compared to fresh specimens?
3. What challenges are faced when using museum collections for genomic studies?

Glossary

Biocultural – The combination of biological and cultural factors.
Crop improvement – Genetic improvement of crops in terms of quality and/or quality to satisfy human needs.
Extant – Species, lineage, or specimen that still exists today.
Gene flow – Allele exchange between populations, one of the main forces that drives evolution.
Haplotype – DNA sequences that are closely located on the chromosomes and thus likely to be inherited together.
Host-pathogen coevolution – The constant competition between hosts and pathogens to infect and spread and to avoid death from infection, respectively. Results in genetic innovations from both sides.

Metadata – Description of a collection event including, among other details, the possible species identification, collection locality, and collector. Usually found on a label, but can also be tracked from associated databases or archives.
Museomics – The study of museum collections using genomic techniques that allow the reconstruction of partial or complete genomes.
Paleogenomics – A field in evolution that attempts to reconstruct and analyse the genetics of specimens that no longer exist.
Xylarium (xylotheque) – Museum collection consisting of authenticated wood samples.
Zebra chip disease – A disease affecting potatoes caused by *Candidatus* Liberibacter solanacearum.

References

Agrios GN (2009) Plant pathogens and disease: general introduction, in: Schaechter, M. (Ed.), Encyclopedia of Microbiology. Academic Press, pp. 613–646. https://doi.org/10.1016/B978-012373944-5.00344-8

Asplund L, Hagenblad J, Leino MW (2010) Re-evaluating the history of the wheat domestication gene NAM-B1 using historical plant material. J. Archaeol. Sci. 37, 2303–2307. https://doi.org/10.1016/j.jas.2010.04.003

Austin RM, Sholts SB, Williams L, Kistler L, Hofman CA (2019) Opinion: To curate the molecular past, museums need a carefully considered set of best practices. Proc Natl Acad Sci USA 116, 1471–1474. https://doi.org/10.1073/pnas.1822038116

Bardill J, Bader AC, Garrison NA, Bolnick DA, Raff JA, Walker A, Malhi RS, Summer internship for INdigenous peoples in Genomics (SING) Consortium (2018) Advancing the ethics of paleogenomics. Science 360, 384–385. https://doi.org/10.1126/science.aaq1131

Besnard G, Christin P-A, Malé P-JG, Lhuillier E, Lauzeral C, Coissac E, Vorontsova MS (2014) From museums to genomics: old herbarium specimens shed light on a C3 to C4 transition. J. Exp. Bot. 65, 6711–6721. https://doi.org/10.1093/jxb/eru395

Bieker VC, Martin MD (2018) Implications and future prospects for evolutionary analyses of DNA in historical herbarium collections. Botany Letters 165, 1–10. https://doi.org/10.1080/23818107.2018.1458651

Cappellini E, Prohaska A, Racimo F, Welker F, Pedersen MW, Allentoft ME, de Barros Damgaard P, Gutenbrunner P, Dunne J, Hammann S, Roffet-Salque M, Ilardo M, Moreno-Mayar JV, Wang Y, Sikora M, Vinner L, Cox J, Evershed RP, Willerslev E (2018) Ancient biomolecules and evolutionary inference. Annu. Rev. Biochem. 87, 1029–1060. https://doi.org/10.1146/annurev-biochem-062917-012002

Cozzolino S, Cafasso D, Pellegrino G, Musacchio A, Widmer A (2007) Genetic variation in time and space: the use of herbarium specimens to reconstruct patterns of genetic variation in the endangered orchid *Anacamptis palustris*. Conserv. Genet. 8, 629–639. https://doi.org/10.1007/s10592-006-9209-7

Culley TM (2013) Why vouchers matter in botanical research. Appl. Plant Sci. 1, 1300076. https://doi.org/10.3732/apps.1300076

Délye C, Deulvot C, Chauvel B (2013) DNA analysis of herbarium Specimens of the grass weed Alopecurus myosuroides reveals herbicide resistance pre-dated herbicides. PLoS ONE 8, e75117. https://doi.org/10.1371/journal.pone.0075117

Dunnum JL, Yanagihara R, Johnson KM, Armien B, Batsaikhan N, Morgan L, Cook JA (2017) Biospecimen repositories and integrated databases as critical infrastructure for pathogen discovery and pathobiology research. PLoS Negl. Trop. Dis. 11, e0005133. https://doi.org/10.1371/journal.pntd.0005133

Ernst M, Saslis-Lagoudakis CH, Grace OM, Nilsson N, Toft Simonsen H, Horn JW, Stærk D, Rønsted N (2016) Molecular phylogenetics as a predictive tool in plant-based drug discovery in the genus *Euphorbia* L. Planta Med. 81, S1–S381. https://doi.org/10.1055/s-0036-1596164

Friis I, Balslev H (2017) Tropical plant collections: legacies from the past? Essential tools for the future? Det Kongelige Danske Videnskabernes Selskab, Copenhagen.

Funk VA, Susanna A, Stuessy TF, Bayer RJ (Eds) (2009) Systematics, evolution, and biogeography of compositae. IAPT, Vienna.

Funk VA (2018) Collections-based science in the 21st Century. J. Syst. Evol. 56, 175–193. https://doi.org/10.1111/jse.12315

Funk V (2004) 100 uses for an herbarium (well at least 72). The Yale University Herbarium.

Goodwin ZA, Harris DJ, Filer D, Wood JRI, Scotland RW (2015) Widespread mistaken identity in tropical plant collections. Curr. Biol. 25, R1066–R1067. https://doi.org/10.1016/j.cub.2015.10.002

Gutaker RM, Burbano HA (2017) Reinforcing plant evolutionary genomics using ancient DNA. Curr. Opin. Plant Biol. 36, 38–45. https://doi.org/10.1016/j.pbi.2017.01.002

Gutaker RM, Weiß CL, Ellis D, Anglin NL, Knapp S, Luis Fernández-Alonso J, Prat S, Burbano HA (2019) The origins and adaptation of European potatoes reconstructed from historical genomes. Nat. Ecol. Evol. 3, 1093–1101. https://doi.org/10.1038/s41559-019-0921-3

Hardion L, Verlaque R, Saltonstall K, Leriche A, Vila B (2014) Origin of the invasive *Arundo donax* (Poaceae): a trans-Asian expedition in herbaria. Ann. Bot. 114, 455–462. https://doi.org/10.1093/aob/mcu143

Hart ML, Forrest LL, Nicholls JA, Kidner CA (2016) Retrieval of hundreds of nuclear loci from herbarium specimens. Taxon 65, 1081–1092. https://doi.org/10.12705/655.9

Hassold S, Lowry PP, Bauert MR, Razafintsalama A, Ramamonjisoa L, Widmer A (2016) DNA barcoding of Malagasy rosewoods: towards a molecular identification of CITES-listed *Dalbergia* species. PLoS ONE 11, e0157881. https://doi.org/10.1371/journal.pone.0157881

Hoban S, Callicrate T, Clark J, Deans S, Dosmann M, Fant J, Gailing O, Havens K, Hipp AL, Kadav P, Kramer AT, Lobdell M, Magellan T, Meerow AW, Meyer A, Pooler M, Sanchez V, Spence E, Thompson P, Toppila R, Griffith MP (2020) Taxonomic similarity does not predict necessary sample size for ex situ conservation: a comparison among five genera. Proc. Biol. Sci. 287, 20200102. https://doi.org/10.1098/rspb.2020.0102

Hofman CA, Rick TC, Fleischer RC, Maldonado JE (2015) Conservation archaeogenomics: ancient DNA and biodiversity in the Anthropocene. Trends Ecol. Evol. 30, 540–549. https://doi.org/10.1016/j.tree.2015.06.008

James SA, Soltis PS, Belbin L, Chapman AD, Nelson G, Paul DL, Collins M (2018) Herbarium data: global biodiversity and societal botanical needs for novel research. Appl. Plant Sci. 6, e1024. https://doi.org/10.1002/aps3.1024

Kemp C (2015) Museums: the endangered dead. Nature 518, 292–294. https://doi.org/10.1038/518292a

Kistler L, Maezumi SY, Gregorio de Souza J, Przelomska NAS, Malaquias Costa F, Smith O, Loiselle H, Ramos-Madrigal J, Wales N, Ribeiro ER, Morrison RR, Grimaldo C, Prous AP, Arriaza B, Gilbert MTP, de Oliveira Freitas F, Allaby RG (2018) Multiproxy evidence highlights a complex evolutionary legacy of maize in South America. Science 362, 1309–1313. https://doi.org/10.1126/science.aav0207

Kuzmina ML, Braukmann TWA, Fazekas AJ, Graham SW, Dewaard SL, Rodrigues A, Bennett BA, Dickinson TA, Saarela JM, Catling PM, Newmaster SG, Percy DM, Fenneman E, Lauron-Moreau A, Ford B, Gillespie L, Subramanyam R, Whitton J, Jennings L, Metsger D, Hebert PDN (2017) Using herbarium-derived DNAs to assemble a large-scale DNA barcode library for the vascular plants of Canada. Appl. Plant Sci. 5, 1700079. https://doi.org/10.3732/apps.1700079

Leonard JA (2008) Ancient DNA applications for wildlife conservation. Mol. Ecol. 17, 4186–4196. https://doi.org/10.1111/j.1365-294X.2008.03891.x

Liesner R (1990) Field techniques used by Missouri Botanical Garden. Missouri Botanical Garden.

Li W, Song Q, Brlansky RH, Hartung JS (2007) Genetic diversity of citrus bacterial canker pathogens preserved in herbarium specimens. Proc Natl Acad Sci USA 104, 18427–18432. https://doi.org/10.1073/pnas.0705590104

Malakasi P, Bellot S, Dee R, Grace OM (2019) Museomics clarifies the classification of *Aloidendron* (Asphodelaceae), the iconic African tree aloes. Front. Plant Sci. 10, 1227. https://doi.org/10.3389/fpls.2019.01227

Malenica N, Simon S, Besendorfer V, Maletić E, Kontić JK, Pejić I (2011) Whole genome amplification and microsatellite genotyping of herbarium DNA revealed the identity of an ancient grapevine cultivar. Naturwissenschaften 98, 763–772. https://doi.org/10.1007/s00114-011-0826-8

Martin MD, Quiroz-Claros E, Brush GS, Zimmer EA (2018) Herbarium collection-based phylogenetics of the ragweeds (*Ambrosia*, Asteraceae). Mol. Phylogenet. Evol. 120, 335–341. https://doi.org/10.1016/j.ympev.2017.12.023

Martin MD, Vieira FG, Ho SYW, Wales N, Schubert M, Seguin-Orlando A, Ristaino JB, Gilbert MTP (2016) Genomic characterization of a South American *Phytophthora* hybrid mandates reassessment of the geographic origins of *Phytophthora infestans*. Mol. Biol. Evol. 33, 478–491. https://doi.org/10.1093/molbev/msv241

Martin MD, Zimmer EA, Olsen MT, Foote AD, Gilbert MTP, Brush GS (2014) Herbarium specimens reveal a historical shift in phylogeographic structure of common ragweed during native range disturbance. Mol. Ecol. 23, 1701–1716. https://doi.org/10.1111/mec.12675

Meineke EK, Davies TJ, Daru BH, Davis CC (2018) Biological collections for understanding biodiversity in the Anthropocene. Philos. Trans. R. Soc. Lond. B Biol. Sci. 374, 20170386. https://doi.org/10.1098/rstb.2017.0386

Payacan C, Moncada X, Rojas G, Clarke A, Chung K-F, Allaby R, Seelenfreund D, Seelenfreund A (2017) Phylogeography of herbarium specimens of asexually propagated paper mulberry [*Broussonetia papyrifera* (L.) L'Hér. ex Vent. (Moraceae)] reveals genetic diversity across the Pacific. Ann. Bot. 120, 387–404. https://doi.org/10.1093/aob/mcx062

Raxworthy CJ, Smith BT (2021) Mining museums for historical DNA: advances and challenges in museomics. Trends Ecol. Evol. 36, 1049–1060. https://doi.org/10.1016/j.tree.2021.07.009

Rønsted N, Grace OM, Carine MA (2020) Editorial: integrative and translational uses of herbarium collections across time, space, and species. Front. Plant Sci. 11, 1319. https://doi.org/10.3389/fpls.2020.01319

Roy MS, Girman DJ, Taylor AC, Wayne RK (1994) The use of museum specimens to reconstruct the genetic variability and relationships of extinct populations. Experientia 50, 551–557. https://doi.org/10.1007/BF01921724

Salick J, Konchar K, Nesbitt M (2014) Curating biocultural collections: a handbook. Royal Botanic Gardens, Kew, Kew.

Sánchez Barreiro F, Vieira FG, Martin MD, Haile J, Gilbert MTP, Wales N (2017) Characterizing restriction enzyme-associated loci in historic ragweed (*Ambrosia artemisiifolia*) voucher specimens using custom-designed RNA probes. Mol. Ecol. Resour. 17, 209–220. https://doi.org/10.1111/1755-0998.12610

Schindel DE, Cook JA (2018) The next generation of natural history collections. PLoS Biol. 16, e2006125. https://doi.org/10.1371/journal.pbio.2006125

Shepherd LD (2017) A non-destructive DNA sampling technique for herbarium specimens. PLoS ONE 12, e0183555. https://doi.org/10.1371/journal.pone.0183555

Silva C, Besnard G, Piot A, Razanatsoa J, Oliveira RP, Vorontsova MS (2017) Museomics resolve the systematics of an endangered grass lineage endemic to north-western Madagascar. Ann. Bot. 119, 339–351. https://doi.org/10.1093/aob/mcw208

Swarup S, Cargill EJ, Crosby K, Flagel L, Kniskern J, Glenn KC (2021) Genetic diversity is indispensable for plant breeding to improve crops. Crop Sci. 61, 839–852. https://doi.org/10.1002/csc2.20377

Uematsu S, Kageyama K, Moriwaki J, Sato T (2012) *Colletotrichum carthami* comb. nov., an anthracnose pathogen of safflower, garland chrysanthemum and pot marigold, revived by molecular phylogeny with authentic herbarium specimens. J. Gen. Plant Pathol. 78, 316–330. https://doi.org/10.1007/s10327-012-0397-3

Van de Paer C, Hong-Wa C, Jeziorski C, Besnard G (2016) Mitogenomics of Hesperelaea, an extinct genus of Oleaceae. Gene 594, 197–202. https://doi.org/10.1016/j.gene.2016.09.007

Weiß CL, Schuenemann VJ, Devos J, Shirsekar G, Reiter E, Gould BA, Stinchcombe JR, Krause J, Burbano HA (2016) Temporal patterns of damage and decay kinetics of DNA retrieved from plant herbarium specimens. R. Soc. Open Sci. 3, 160239. https://doi.org/10.1098/rsos.160239

Welch AJ, Collins K, Ratan A, Drautz-Moses DI, Schuster SC, Lindqvist C (2016) Data characterizing the chloroplast genomes of extinct and endangered Hawaiian endemic mints (Lamiaceae) and their close relatives. Data Brief 7, 900–922. https://doi.org/10.1016/j.dib.2016.03.037

Wheeler QD, Knapp S, Stevenson DW, Stevenson J, Blum SD, Boom BM, Borisy GG, Buizer JL, De Carvalho MR, Cibrian A, Donoghue MJ, Doyle V, Gerson EM, Graham CH, Graves P, Graves SJ, Guralnick RP, Hamilton AL, Hanken J, Law W, Woolley JB (2012) Mapping the biosphere: exploring species to understand the origin, organization and sustainability of biodiversity. Systematics and Biodiversity 10, 1–20. https://doi.org/10.1080/14772000.2012.665095

Wood J, Ballou JD, Callicrate T, Fant JB, Griffith MP, Kramer AT, Lacy RC, Meyer A, Sullivan S, Traylor-Holzer K, Walsh SK, Havens K (2020) Applying the zoo model to conservation of threatened exceptional plant species. Conserv. Biol. 34, 1416–1425. https://doi.org/10.1111/cobi.13503

Yoshida K, Burbano HA, Krause J, Thines M, Weigel D, Kamoun S (2014) Mining herbaria for plant pathogen genomes: back to the future. PLoS Pathog. 10, e1004028. https://doi.org/10.1371/journal.ppat.1004028

Yoshida K, Sasaki E, Kamoun S (2015) Computational analyses of ancient pathogen DNA from herbarium samples: challenges and prospects. Front. Plant Sci. 6, 771. https://doi.org/10.3389/fpls.2015.00771

Yoshida K, Schuenemann VJ, Cano LM, Pais M, Mishra B, Sharma R, Lanz C, Martin FN, Kamoun S, Krause J, Thines M, Weigel D, Burbano HA (2013) The rise and fall of the Phytophthora infestans lineage that triggered the Irish potato famine. eLife 2, e00731. https://doi.org/10.7554/eLife.00731

Zedane L, Hong-Wa C, Murienne J, Jeziorski C, Baldwin BG, Besnard G (2016) Museomics illuminate the history of an extinct, paleoendemic plant lineage (*Hesperelaea* , Oleaceae) known from an 1875 collection from Guadalupe Island, Mexico. Biological Journal of the Linnean Society 117, 44–57. https://doi.org/10.1111/bij.12509

Answers

1. Herbarium, xylarium and other wood specimens, seed collections, and economic botany collections are examples of museum collections that may be useful sources of DNA.
2. Inference using fossils or other ancient materials adds a reliable temporal dimension to the analyses, which can, for example, be used to measure the evolutionary processes plants have gone through in response to environmental and/or evolutionary changes.
3. Challenges include physical and permit availability of sufficient amounts of material for destructive sampling from the outset. Museum specimens may not always contain sufficient characters for unambiguous identification or information about origin. In the wet lab and during the analysis, challenges are posed by traits of aDNA that are highly fragmented and low content of endogenous DNA.

Chapter 21
Palaeobotany

Thibauld Michel[1], Michael D. Martin[2], Catherine Kidner[1]

1 Royal Botanic Garden of Edinburgh, University of Edinburgh, Edinburgh, United Kingdom
2 NTNU University Museum, Norwegian University of Science and Technology, Trondheim, Norway

Thibauld Michel **tmichel@rbge.org.uk**
Michael D. Martin **mike.martin@ntnu.no**
Catherine Kidner **ckidner@rbge.org.uk**

Citation: Michel T, Martin MD, Kidner C (2022) Chapter 21. Palaeobotany. In: de Boer H, Rydmark MO, Verstraete B, Gravendeel B (Eds) Molecular identification of plants: from sequence to species. Advanced Books. https://doi.org/10.3897/ab.e98875

Introduction

The evolution of ancient plant DNA analysis across time

The study of ancient plant remains was historically limited to morphological studies, palaeontology being the primary field of study of past organisms. However, since the 1980s, genetic analysis of biological matter within fossils has become increasingly informative, thanks to the development of new sequencing technologies such as polymerase chain reaction (PCR) and high-throughput sequencing (HTS). Since the first identification of aDNA from extinct species in 1984 (Higuchi et al. 1984), it is now possible to identify many organisms, including plant taxa from only microscopic plant remnants or even short fragments of DNA bound to a substrate using modern molecular tools. These new methods allow us to explore their recent evolutionary history, study the ecology of paleoenvironments, and understand population relationships, migration, and domestication processes. While the domain of paleogenetics is limited to the study of a few genetic markers, the establishment of new DNA isolation techniques, high-throughput sequencing (HTS), and more robust computational methods have enabled the analysis of longer DNA fragments and in the last couple of decades shifted the domain towards the field of paleogenomics with the analysis of full plant genomes (Mitchell and Rawlence 2021).

Ancient plant DNA in historical remains

In the context of paleogenetics, ancient DNA (aDNA) is DNA from long-deceased tissues preserved by conditions allowing DNA survival. Despite appropriate preservation conditions, aDNA is usually degraded by biotic or abiotic processes. Though often damaged, it can carry valuable historical information (Schlumbaum et al. 2008) (see Chapter 2 DNA from museum collections). Ancient DNA from wild plants can be used to reconstruct the evolutionary and demographic histories of populations to trace ecological and climatic changes (Hofreiter et al. 2001). Ancient plant DNA from anthropogenic sources can be used for studying the processes of plant domestication, generating insights into past plant usage, agricultural techniques, and the migration patterns of ancient human societies (Kistler et al. 2014; Trucchi et al. 2021).

The sequences used for most plant aDNA studies are derived from the nuclear and organellar genomes and are quite often the same markers typically used for plant identification or studies of evolutionary history. Markers from plastids are usually favoured for their high copy number and short length, despite reported problems resulting from their high propensity for genetic rearrangements. Furthermore, horizontal transfer from the plastome to nuclear and mitochondrial genomes complicates the analysis as the mutational rate differs between in the nucleus and in other organelles (Kistler et al. 2014; Wales et al. 2016) Compared to chloroplasts markers, mitochondrial markers are considered less informative, and have rarely been used in plant aDNA due to their slow rate of mutation (Schlumbaum et al. 2008). However, they are not the only target to consider as traces of ancient RNA have also been amplified in cress seeds by hybridization and later sequenced in maize (Fordyce et al. 2013; Gnirke et al. 2009; Rollo 1985). Other targets previously detected include epigenetics patterns such as methylation in response to pathogen infection, and small plant RNA (miRNA) as a response to environmental stress (Smith et al. 2017, 2014) in barley. Bacterial and viral DNA can also be amplified from ancient plant material (Bieker et al. 2020; Smith et al. 2014).

Historical and archeological aDNA challenges

Several difficulties are inherent to working with aDNA from plant specimens: the complexity and variability of the genome, aDNA damages, and potential contamination increase downstream analytical difficulties. The combination of often very low aDNA concentrations with the amplification power of PCR dramatically increases the probability of amplifying contaminating modern DNA. Specialised methods and laboratory procedures have been established to reduce the risk of contamination. These include: the use of positively pressurised clean laboratory facilities dedicated to aDNA work, the replication of experimental works in different institutions, and the use of biomarkers for prediction of DNA survival such as mitochondrial DNA (mtDNA) detection, aDNA damage patterns, and detection of associated remains (Capo et al. 2021; Cooper and Poinar 2000) (see Chapter 2 DNA from museum collections).

Another complication in the analysis of plant DNA is its variability. The presence of different organelle genomes (plastid and mitochondrial) as well as the interspecific differences in ploidy level and chromosome size can complicate the alignment of sequencing reads to a reference sequence (Kapusta et al. 2017; Kistler et al. 2020). Target capture is one strategy that can be used to reduce this complexity, even in sequences that are heavily degraded and/or contaminated (Parducci et al. 2019) (see Chapter 14 Target capture).

Sources of plant DNA

Macrofossils

Macrofossils are defined as fossils that are observable without magnification, and in the case of plant-based studies, they are ancient preserved tissues found in archaeological or sedimentological contexts. aDNA can be extracted from macroscopic plant remains such as leaves, needles, bud scales, wood, or seeds. However, individually-based approaches on plant macrofossils are scarce and most of the studies focusing on plant DNA are based on metabarcoding using sedimentary DNA (sedaDNA) material (Jaenicke-Després et al. 2003; Rollo et al. 2002; Schwörer et al. 2022). The scarcity of macrofossils in plant studies can be explained by the difficulty of aDNA recovery in preserved plants, which can be due to the low-level of endogenous DNA in plant remains, high amounts of contaminant microbial DNA, and aDNA specific damages (Green and Speller 2017). Regardless of this limitation, macrofossil DNA studies have some advantages: they can be directly dated without the use of proxies, they represent local species in contrast to pollen studies where pollen grains could be dispersed over large distances, and DNA from a single analysis can be authentified from its aDNA specific damage patterns (Schwörer et al. 2022).

Charred and desiccated

A very common archaeological plant material is charred remains. One example is superficially burnt seeds in hearth remains found in ancient settlements. Molecular identification of even lightly charred remains is however challenging since the DNA is often very fragmented and contaminated (Palmer et al. 2012). Target enrichment has not yet been able to overcome this issue (Nistelberger et al. 2016). Thus, charred plant remains are primarily identified using morphological analysis.

In contrast, desiccated samples are often suitable for molecular analysis. Desiccated samples are typically found in dry environments such as caves, shelters formed by rock features (well suited for long-term food storage), or deserts. Desiccation can limit DNA degradation, and plastid and

mitochondrial DNA from sunflower seeds as old as 3,100 years old has been successfully recovered (Kistler and Shapiro 2011; Mascher et al. 2016; Swarts et al. 2017; Wales and Kistler 2019).

Waterlogged

Biological remains preserved under waterlogged anaerobic conditions may also contain sufficient aDNA for molecular identification. Lakes and marine sediments can provide sedimentary DNA (sedaDNA) from plant remains and pollen grains found in different strata of core samples. They can be used to reconstruct past ecological diversity. Microorganism communities can as well be a source of aDNA. For example, diatoms are commonly used bioindicators for assessing the biological composition (trophic state) of a lake since their morphology is highly sensitive to the surrounding environment (Ibrahim et al. 2021). The taxonomic diversity of diatoms found in the sediments of glacial and thermokarst lakes has for instance been linked to lake type and age, environmental changes, and surrounding vegetation (Huang et al. 2020). Cyanobacteria, which are sensitive to temperature, can be used as a biomarker for detecting the effects of climate change by studying their population diversity. The microbial communities of Lake Constance (Central Europe) for instance, including microbial eukaryotes, diatoms, and cyanobacteria, have been used as bioindicators for both biotic and abiotic changes due to warming by studying the phylogenetic distance of microbial communities, and their geographic and temporal change of diversity (Monchamp et al. 2019).

Waterlogged remains can be found in the context of archaeological studies. Wells, latrines, ditches, and pits can result in anaerobic conditions. DNA from grape seeds from the Iron Age have been sequenced successfully with Hyb-Seq, and it was shown that the grapes are related to present-day West European cultivars, which provides evidence that there has been 900 years of uninterrupted vegetative propagation of the crop (Ramos-Madrigal et al. 2019). Gourd rinds, squash seeds, and oak wood thousands of years old have provided high-quality aDNA using target-capture methods, or using plastid or mitochondrial DNA (Wagner et al. 2018). This has led to a correction on the view of how gourd domestication happened by showing that the pre-Columbian bottle gourds originated from Africa and reached Latin America via the Atlantic by ocean drift (Kistler et al. 2014). Other studies have shown a link between the Holocene megafauna extinction and the decline of wild Cucurbita, while domestic lineages thrived because of cultivation (Kistler et al. 2015).

Mineralized and embedded

Mineralized samples or those embedded in resin or fossilised in amber are both potential sources for aDNA, though the high probability of contamination, extreme fragmentation of the material, and non-reproducibility of the results have led some authors to strongly discourage aDNA analysis from amber-preserved fossils (Modi et al. 2021). However, partially mineralized remains (subfossils) less than 10,000 years old can still contain biological material and are potentially a source of biomolecules including DNA (Wagner et al. 2018). Recently developed methodologies for specimen extraction from amber that reduces contamination have enabled the recovery of insect DNA up to 3900 years old from copal, a precursor to amber (Peris et al. 2020). This leads to the possibility that these sample types may be sources for plant aDNA in the future.

Microfossils

Microfossils can be found in any environment, including in humid conditions and tropical zones where macrofossil preservation is rare. These include pollen, starch grains, and phytoliths. Plastid aDNA obtained from pollen grains is very often endogenous, and its amplification has

previously established the first genetic link between extant and fossilised Scots Pine specimens from post glacial lake sediments in Sweden (Parducci et al. 2005). Phytoliths enable radiocarbon dating, even though no aDNA has been isolated from them so far (Elbaum et al. 2009). Yet, they are hypothesised to be a potential source of aDNA (Grass et al. 2015).

Sedimentary DNA

Sediments found in lakes, temperate caves, permafrost, and ice cores can retain plant aDNA for thousands, and in some cases, millions of years (Kirkpatrick et al. 2016). Sedimentary DNA may be used as a proxy for the reconstruction of the paleoenvironment, even though other plant structures have been destroyed (Willerslev et al. 2003). Metabarcoding to amplify short amplicons of cpDNA is by far the most commonly used approach (Capo et al. 2021; Parducci et al. 2017; Rijal et al. 2021). Shotgun sequencing has only been used sparsely because of the lack of reference libraries (Slon et al. 2017), but as full genome reference databases are being built, this method could improve the ability to investigate lake sedaDNA (Parducci et al. 2019). More recently, shotgun metagenomics was used for retrieval of whole plant genomes from archeological settlements and marine deposits (Parducci et al. 2019; Pedersen et al. 2013; Slon et al. 2017). However, sedaDNA taphonomy for sedimentary material is still a subject to explore, as the conditions that lead to its preservation are not yet clear (Kistler et al. 2020) (see Chapter 8 aDNA from sediments).

sedaDNA provides a broad understanding of the past environment, climate, and ecology of the paleosol studied. It can also provide insights on the movement and cultivation of plants by Neolithic populations and their social network in absence of other archeological evidence (Brown et al. 2021; Smith et al. 2015).

sedaDNA from lake sediments has been used to reconstruct ancient plant vegetation and to assess the impact of anthropogenic activities on the paleoenvironment. For example, the impact of cattle grazing on deforestation dynamics during the Late Iron Age and Roman period has been demonstrated by using a metabarcoding approach on sediment samples from a subalpine lake (Giguet-Covex et al. 2014).

sedaDNA can also be used to study the impact of climatic changes on plant biodiversity and help prioritise conservation management. A research project using metabarcoding of lake sediments was able to show that a heterogeneous mountain landscape served as a refugium for arctic-alpine plants in a warm climate (Clarke et al. 2019).

Another study on Arctic Canada lake sediments gave clues about the effect of the rise in temperature during the Last Interglacial period (LIG) on plant population dynamics. Previous attempts to reconstruct the LIG paleoclimate with climate modelling based on the simulation of atmosphere, sea, and ice circulation have yielded inconsistent results (Otto-Bliesner et al. 2013). Comparison of the model results with sedaDNA vegetation reconstruction suggests that models underestimated the magnitude of Arctic warming during the LIG. This discrepancy could be due to the lack of vegetation-related feedback such as arctic greening in the models, but are observable in sedaDNA records (Crump et al. 2021).

We can improve modelling of future climate change effects on plant diversity based on these studies that inform how plant richness has evolved in reaction to previous episodes of climate warming. Several environmental changes that might have been overlooked such as arctic amplification or arctic greening can be studied with sedaDNA (Clarke et al. 2019; Crump et al. 2021; Liu et al. 2021). The impact of sea ice on plant colonisation of Iceland during different periods of the Holocene suggests that the melting of the ice sheet due to future warming might limit plant distribution rather than favour it (Alsos et al. 2021).

SedaDNA studies are furthermore more robust than pollen-based methods for detecting plant richness and deliver taxa diversity with more resolution (Crump et al. 2021). As an example,

a study based on multiple-sites lake sedaDNA analysis and pollen records shows the steep increase of plant richness in the early Holocene in northern Fennoscandia (Rijal et al. 2021). The causes of this increase are the higher level of available soil nutrients and the lower level competition just after deglaciation. However, the pollen records did not match the sedaDNA findings that taxonomic richness has continued to increase even after climate stabilisation. These discrepancies are due to problems affecting pollen records such as overabundance of a few taxa and underrepresentation of others (swamping). In contrast, sedaDNA provides higher taxonomic resolution, lower swamping effect, and represents local plant groups.

The same observations can be done using sedaDNA extracted from permafrost, as presented in a study encompassing 50,000 years of megafauna diet and arctic vegetation history from samples collected across the Arctic. While pollen-based reconstruction showed a majority of graminoids in unglaciated Arctic during the Late Glacial Maximum, the metabarcoding approach has revealed a forb-dominated vegetation (Willerslev et al. 2014).

Palaeofaeces

Ancient faeces, though relatively uncommon, are a rich source of biomolecules and paleodietary information that can be related to demographic, ecological, and climatic changes in the locations in which they are found (Green and Speller 2017). Genetic identification from plastome barcoding can also provide evidence missing in classic macroscopic morphological analysis (Poinar et al. 1998; Rasmussen et al. 2009; Rollo et al. 2002). Recent approaches using shotgun metagenomic methods provided identification of plants in ancient faeces as well as information on the gut microbiome, parasitic worms, and the actual identification of the defecator (Boast et al. 2018; Wood et al. 2016) (see Chapter 7 DNA from faeces).

Bioinformatic tools and challenges

The analysis of an aDNA dataset is complicated by post-mortem DNA degradation that leads to short fragments, specific nucleotide substitution patterns, and overall low DNA yields (Briggs et al. 2007). These difficulties will affect subsequent evolutionary inferences and population genetics studies. Consequently, numerous tools have been developed to detect and quantify nucleotide substitution, deletion, and DNA fragmentation (see Chapter 2 DNA from museum collections).

The initial alignment step with a reference genome during bioinformatic analyses is already affected by aDNA chemical damage, which can increase the apparent error rate and lower the alignment accuracy. Subsequent steps in variant calling of genetic markers can be complicated by the high mapping error rate and low coverage (Bilinski et al. 2018). Strategies have been developed to prevent bias resulting from low coverage. This can include random sampling of a single read at each locus of interest (Bakker et al. 2016; Kistler et al. 2018) and genotype likelihood estimation (Korneliussen et al. 2014). More specific tools have also been designed to solve the issue of identifying the ancestry of unknown samples with a low coverage dataset using multidimensional scaling (MDS) methods (Malaspinas et al. 2014; Ramos-Madrigal et al. 2019). Issues related to aDNA specific damage patterns can be prevented using strategies such as only considering transversion polymorphisms, using statistical algorithms to rescale the base quality scores before variant calling (Jónsson et al. 2013), or soft-clipping fragment ends to avoid deamination sites (Kistler et al. 2018). Tools for rescaling base quality scores have also been implemented into bioinformatic pipelines that are dedicated to aDNA alignment (Schubert et al. 2014).

Applications

Evolutionary studies

The evolutionary history of a species or a population can be established based on genomic inference from modern samples, providing clues about the evolutionary processes that form the basis for present genomic variation. However, allelic patterns in contemporary specimens are shaped by a range of demographic events, including changes in population size, gene flow, and hybridization events. These may be due to very recent events, and do not necessarily represent the lineage's deeper evolutionary history. A time series of samples can provide greater resolution in a genomic analysis and resolve phylogenetic questions. It can also detect recent demographic events such as population bottlenecks and provide chronological estimates for these events without using a molecular clock. Allele frequencies can be directly estimated for each time point and used to estimate the strength of selection pressure during that period (Malaspinas 2016). This approach can be used to distinguish between different selection processes and to establish their tempo across time (Dehasque et al. 2020).

The Dramatic global warming and extinction events that occurred during the later Anthropocene coincided with the active collection of specimens for museums and herbaria (Bieker and Martin 2018). Genetic analysis of collections provides a detailed understanding on how human activity has shaped the evolutionary fate of many organisms. Modern techniques also allow us to recover information on extinct species. One example is the genus *Hesperelaea* from the Oleaceae family, which was collected once 140 years ago in Mexico, and is now extinct. A genomic analysis of this *H. palmeri* specimen traced its American lineage, the date of its divergence, and helped to characterise its endemism (Zedane et al. 2016).

Positive selection can also be detected in contemporary specimens using statistical tools such as coalescence, population differentiation (F_{ST}), and linkage disequilibrium. Selection pressure, however, can be conflated with demographic change or background selection. Specific methods have been developed to detect positive selection on a polygenic trait using an admixture graph to represent the admixture events relating different populations through time (Racimo et al. 2018).

Purifying selection or negative selection can be detected in present-day specimens as signals of reduced genetic diversity. However, similar signals can be caused by demographic events such as population bottlenecks or background selection (Henn et al. 2015). Again, using a sample time series, these signatures can be disentangled by considering regions with lower recombination rate where selection has more impact (Murray et al. 2017). Therefore, loci located in regions with low recombination rates and lower genetic diversity are likely to be a signature for selection rather than past demographic events. A good understanding of genome structure and dynamics in the target species is thus key to accurate inference of selection.

Balancing selection is more difficult to detect since it affects narrow genomic regions on a short timescale. This can be mistaken for positive selection, demographic events, or introgression (Fijarczyk and Babik 2015). For these reasons, methods using contemporary specimens have low statistical power. A time series of samples can help detect alleles under balancing selection as their frequencies are maintained over time at frequencies higher than expected by the Hardy-Weinberg law.

Tracing domestication

All current crops are the products of single or repeated domestication events starting less than 12,000 years ago from the ancestral wild species (Kistler et al. 2015; Larson et al. 2014). Understanding the geographical origin and the ancestral lineages of domesticated species during the Holocene and the subsequent spread of the cultivars are central questions for different domains such as archaeology, anthropology, and ecology.

Archaeobotanical remains can be arranged in a time series to study the evolution of domestication over time and space. They can indicate the number of times that domestication events occurred and their location, the pace and stringency of anthropogenic selection, introgression with wild relatives and between different cultivars and be used to determine the date of these events (Brown 1999).

Molecular methods have made an increasingly large contribution to the field of archaeobotany. Starting with simple genetic analysis for taxonomic identification to supplement morphological examination, the field has rapidly progressed following advances in high-throughput technologies in archaeogenomics. Methods such as shotgun sequencing have enabled genome-wide studies, exploring in detail the genome of domesticated plants and analysing the genome-wide rearrangements that occurred during this process (Palmer et al. 2012).

As both a key crop and a genetic model organism deeply studied for over 100 years, a wealth of domestication studies have been conducted on maize, revealing a detailed picture of evolution. Molecular analysis of palaeobotanical remains continues to provide new information on maize evolution, and PCR-based studies have identified the likely geographic region of its original domestication in Mexico and traced its dispersal across Central America and South America (Kistler et al. 2018).

The target capture method, or Hyb-Seq (see Chapter 14 Target capture) has been used to confirm and refine models for maize domestication over time mediated with progressive introgression from wild relatives (da Fonseca et al. 2015). A recent study on maize domestication and diversification in South America based on the genomes of present-day and ancient American maize cobs has shown that maize had a stratified mode of domestication that started with a large Mesoamerican gene pool that was partially domesticated. This was followed by a dispersal to different locations in which the sub populations become reproductively isolated by different selection pressures (Kistler et al. 2018).

Wheat domestication has not been studied as extensively as maize, but modern genome-wide studies on emmer wheat chaff found shared haplotypes between 3,000-year-old Egyptian emmer wheat from museum collection and modern emmer wheat, including domestication loci as two QTLs related to grain size and seed dormancy. Although several haplotypes present in historical specimens are absent from modern emmer, similarities between museum specimens and Arabian and Indian emmer landraces suggest an early South-Eastern dispersal of ancient Egyptian emmer (Scott et al. 2019).

Bottlenecks are a common feature in the domestication process and have also been revealed from ancient plant material in beans. One of the symptoms of a bottleneck event in the demographic history of a lineage is genetic erosion, the loss of allele diversity in a population due to genetic drift and inbreeding caused by the bottleneck event. This effect was found in the case of the Andean bean domestication, which was likely triggered by stringent varietal selection (Trucchi et al. 2021). In this study, ancient bean genomes dated between 600 and 2,500 years ago showed ten times more heterozygosity than modern genomes, despite that the set of genes that characterise the domestication had already been selected. It is likely that initial improvements in common beans occurred via soft sweeps rather than under strong selection

pressure, while selection strategies in recent centuries produced further improvement at the cost of genetic erosion (Trucchi et al. 2021).

Phylogeography

Climatic and environmental changes can be responsible for major shifts in species' geographic distributions. For example, the glaciation cycles over the past 2.4 million years have restricted some species in separate refugia, often resulting in a loss of allelic variation that persists after the species' expansion out of the refugium. Phylogeography allows studying the history of geographic distribution of genealogical lineages using population genetic tools to detect the changes in genetic variation caused by historical events such as migration and dispersal (Cruzan and Templeton 2000). In contrast to studies of modern populations using selection inference from a single time point, aDNA studies including multiple time points can show the shift of alleles before and after periods of environmental or demographic change, providing information about the selection coefficient of the event (Bank et al. 2014).

Early plant phylogeography studies were based on plastid DNA (pDNA) sequencing methods, as a study of the distribution and circumpolar migration of saxifrage, suggesting the possibility that plant refugia were located in the Arctic (Abbott et al. 2000). Later studies used DNA fingerprinting, such as amplified fragment length polymorphism (AFLP), in addition to pDNA to disentangle signatures of hybridization due to isolation in a refugia and postglacial migration for two species of Birches (Eidesen et al. 2015). More recently, a target capture method has been used on lake sediments to recover the complete larix chloroplast genome and study its dynamics at population level (Schulte et al. 2021). Another recent study has used shotgun sequencing to analyse Ice Age algal populations from lake sediments. It has enabled the mapping of chloroplast and mitochondrial genomes to reconstruct the genomic variation of the lake populations (Lammers et al. 2021).

Paleoecology

Ancient DNA studies can unravel the ecological past and temporally explore the adaptation mechanism and interactions between organisms. This can include processes such as convergent evolution of different species in a similar environment, present plant adaptations due to standing or de novo mutation in the evolutionary history of a species, or metagenomics of an aDNA specimen to reveal the dynamics of plant pathogens (Bieker et al. 2020; Kistler et al. 2020).

Innovations in shotgun metagenomics have increased the possibilities for using sedDNA analysis for reconstruction of past vegetation with higher taxonomic resolution than with pollen DNA barcoding (Bjune et al. 2021; Clarke et al. 2020), and they can detect more taxa in a single sample than macrofossils (Alsos et al. 2016). Provided that an appropriate reference library is available, minimal sampling can enable the identification of hundreds of different taxa in a few samples, giving an estimation of species diversity. This information allows reconstruction of the paleoenvironment and its biodiversity change over time (Anderson-Carpenter et al. 2011).

Some limitations do however remain. SedaDNA is preserved in lake environments since the stable temperature conditions can conserve DNA. However, sampling can be challenging in these areas. There are also major challenges in detecting species that are rare or have a low biomass. Additionally, the taxonomic resolution provided by sedaDNA is variable in function of the method used. While metabarcoding sedaDNA almost always provides higher resolution than direct pollen analysis (Clarke et al. 2020; Sønstebø et al. 2010), the reference library to match the dataset must match the method used and the flora of the region (Parducci et al. 2019)).

Conservation archaeogenomics

The Anthropocene presents major global challenges, including climate change, loss of biodiversity through extinction, and emerging zoonotic infectious diseases. An understanding of previous human interactions with the environment can guide conservation management during this era of massive environmental change and rapid loss of biodiversity. The field of conservation archaeogenomics involves analysing aDNA with the goal of guiding present-day biological conservation (Hofman et al. 2015).

Genomic archaeological data can also reveal details about the time and potential reasons for local or global extinction events, and help to understand the resulting consequences on ecosystems and human societies. Studies that use these data may also contribute to better understanding how human activities and behaviours may have contributed to past extinction events. Studying the distribution of species and how they colonise new areas can also help us to anticipate how ecosystems may respond to future climate change (Alsos et al. 2021).

A theoretical application of the recent progress in molecular biology and sequencing techniques follows from the concept of "de-extinction" or "species revivalism". The possibility of de-extinction is controversial and still debated on both technical and ethical levels, as it is difficult to justify the ecological need for reviving extinct species rather than supporting current conservation efforts for endangered species (Orlando and Cooper 2014).

Future perspectives on plant aDNA analysis

Over the last several decades, paleogenetics has made substantial contributions towards our understanding of ancient plant science, ecology, and archeology. In contrast, paleogenomics is just in its infancy and sequencing and analysis techniques are constantly improving. The study of full genome datasets has allowed to accurately characterise taxonomic diversity (Wagner et al. 2018), to study changes in distribution and demography over time, including changes in population size and measurement of genetic diversity on a population scale (Schwörer et al. 2022; Zimmermann et al. 2017), to investigate the origin of ancient domesticated plant cultivars with high resolution (Ramos-Madrigal et al. 2019; Scott et al. 2019; Trucchi et al. 2021), and to reconstruct entire palaeoenvironments (Capo et al. 2021).

The race to understand biological diversity before it is lost is, to some degree, mitigated by the presence of valuable genomic information in archaeological and natural history collections that include extinct and endangered species. As this field of research provides information about common species and their ecological background, it provides a framework in which to study and understand how the past 200 years of human activity have impacted patterns of genetic diversity in the natural world. It is essential that we use insights from the study of ancient plant genomics to help us reduce biodiversity loss over the next 200 years.

Questions

1. Human faecal material recovered from the latrines of an ancient settlement were analysed with a shotgun sequencing approach, yielding puzzling results. The plants identified from this archaeological site were not domesticated at the time of its occupation and are not supposed to be present at this location. How can you explain this discrepancy? What protocols can be used to verify this result?

2. A study of the Holocene glacial retreat will be designed to assess the time and zone affected by deglaciation using plant aDNA as a proxy. What aDNA specimens can be used to assess the changes in plant diversity over time at each sampling point, and identify the species involved?

Glossary

Amber – Fossilised tree resin, may contain animal or plant material as inclusion.
Paleogenetics – The study of the past using genetic material from ancient specimens.
Palaeogenomics – Genome-scale sequencing studies of genetic material from ancient specimens.
Balancing selection – Different selective processes which maintain genetic diversity at a frequency superior to that expected under neutral genetic drift.
Coprolite (or coprolith) – Fossilised human or animal faeces. Contrary to paleofaeces, most of their original composition has been replaced by mineral deposit.
cpDNA – Chloroplast DNA, or plastome.
De-extinction – Theoretical possibility to rebuild extinct species using aDNA sequences.
Ice core – Long cylinder of ice recovered by drilling through ice sheets or glaciers.
mtDNA – Mitochondrial DNA.
Palaeoecology (or paleoecology) – The study of interactions between organisms and their environment across geologic timescales.
Palaeofaeces (or paleofeces) – Ancient animal or human faeces. Contrary to coprolites, they retain some parts of their original biological composition, although in practice the terms are used interchangeably.
Permafrost – Ground continuously frozen (below 0 °C) for two or more years.
Phytoliths – Silica microstructures found in some plant tissues.
Plant domestication – Human selection of desirable traits in plants that has taken place in the last 12,000 years.
Positive selection (or directional selection) – Process by which one phenotype is selected preferentially to others, causing allele frequency to shift over time towards this phenotype.
Purifying selection (or negative selection) – The removal of deleterious alleles from a population genome.
SedDNA – Sedimentary DNA, younger and better preserved sedimentary DNA.
SedaDNA – Sedimentary ancient DNA, older, more poorly preserved.
Subfossil – Organism partially fossilised still containing biological matter such as bone, skin, or faecal deposit, while a fossil is completely mineralized.
Taphonomy – Study of how organic remains pass from the biosphere to the lithosphere, including processes affecting remains from the time of death of an organism through decomposition, burial, and preservation as mineralized fossils or other stable biomaterials.

References

Abbott RJ, Smith LC, Milne RI, Crawford RM, Wolff K, Balfour J (2000) Molecular analysis of plant migration and refugia in the Arctic. Science 289, 1343–1346. https://doi.org/10.1126/science.289.5483.1343

Alsos IG, Ehrich D, Seidenkrantz M-S, Bennike O, Kirchhefer AJ, Geirsdottir A (2016) The role of sea ice for vascular plant dispersal in the Arctic. Biol. Lett. 12. https://doi.org/10.1098/rsbl.2016.0264

Alsos IG, Lammers Y, Kjellman SE, Merkel MKF, Bender EM, Rouillard A, Erlendsson E, Guðmundsdóttir ER, Benediktsson ÍÖ, Farnsworth WR, Brynjólfsson S, Gísladóttir G, Eddudóttir SD, Schomacker A (2021) Ancient sedimentary DNA shows rapid post-glacial colonisation of Iceland followed by relatively stable vegetation until the Norse settlement (Landnám) AD 870. Quat. Sci. Rev. 259, 106903. https://doi.org/10.1016/j.quascirev.2021.106903

Anderson-Carpenter LL, McLachlan JS, Jackson ST, Kuch M, Lumibao CY, Poinar HN (2011) Ancient DNA from lake sediments: bridging the gap between paleoecology and genetics. BMC Evol. Biol. 11, 30. https://doi.org/10.1186/1471-2148-11-30

Bakker FT, Lei D, Yu J, Mohammadin S, Wei Z, van de Kerke S, Gravendeel B, Nieuwenhuis M, Staats M, Alquezar-Planas DE, Holmer R (2016) Herbarium genomics: plastome sequence assembly from a range of herbarium specimens using an iterative organelle genome assembly pipeline. Biological Journal of the Linnean Society 117, 33–43. https://doi.org/10.1111/bij.12642

Bank C, Ewing GB, Ferrer-Admettla A, Foll M, Jensen JD (2014) Thinking too positive? Revisiting current methods of population genetic selection inference. Trends Genet. 30, 540–546. https://doi.org/10.1016/j.tig.2014.09.010

Bieker VC, Martin MD (2018) Implications and future prospects for evolutionary analyses of DNA in historical herbarium collections. Botany Letters 165, 1–10. https://doi.org/10.1080/23818107.2018.1458651

Bieker VC, Sánchez Barreiro F, Rasmussen JA, Brunier M, Wales N, Martin MD (2020) Metagenomic analysis of historical herbarium specimens reveals a postmortem microbial community. Mol. Ecol. Resour. 20, 1206–1219. https://doi.org/10.1111/1755-0998.13174

Bilinski P, Albert PS, Berg JJ, Birchler JA, Grote MN, Lorant A, Quezada J, Swarts K, Yang J, Ross-Ibarra J (2018) Parallel altitudinal clines reveal trends in adaptive evolution of genome size in Zea mays. PLoS Genet. 14, e1007162. https://doi.org/10.1371/journal.pgen.1007162

Bjune AE, Greve Alsos I, Brendryen J, Edwards ME, Haflidason H, Johansen MS, Mangerud J, Paus A, Regnéll C, Svendsen J, Clarke CL (2021) Rapid climate changes during the Lateglacial and the early Holocene as seen from plant community dynamics in the Polar Urals, Russia. J. Quaternary Sci. https://doi.org/10.1002/jqs.3352

Boast AP, Weyrich LS, Wood JR, Metcalf JL, Knight R, Cooper A (2018) Coprolites reveal ecological interactions lost with the extinction of New Zealand birds. Proc Natl Acad Sci USA 115, 1546–1551. https://doi.org/10.1073/pnas.1712337115

Briggs AW, Stenzel U, Johnson PLF, Green RE, Kelso J, Prüfer K, Meyer M, Krause J, Ronan MT, Lachmann M, Pääbo S (2007) Patterns of damage in genomic DNA sequences from a Neandertal. Proc Natl Acad Sci USA 104, 14616–14621. https://doi.org/10.1073/pnas.0704665104

Brown AG, Van Hardenbroek M, Fonville T, Davies K, Mackay H, Murray E, Head K, Barratt P, McCormick F, Ficetola GF, Gielly L, Henderson ACG, Crone A, Cavers G, Langdon PG, Whitehouse NJ, Pirrie D, Alsos IG (2021) Ancient DNA, lipid biomarkers and palaeoecological evidence reveals construction and life on early medieval lake settlements. Sci. Rep. 11, 11807. https://doi.org/10.1038/s41598-021-91057-x

Brown TA (1999) How ancient DNA may help in understanding the origin and spread of agriculture. Phil. Trans. R. Soc. Lond. B 354, 89–98. https://doi.org/10.1098/rstb.1999.0362

Capo E, Giguet-Covex C, Rouillard A, Nota K, Heintzman PD, Vuillemin A, Ariztegui D, Arnaud F, Belle S, Bertilsson S, Bigler C, Bindler R, Brown AG, Clarke CL, Crump SE, Debroas D, Englund G, Ficetola GF, Garner RE, Gauthier J, Parducci L (2021) Lake sedimentary DNA research on past terrestrial and aquatic biodiversity: overview and recommendations. Quaternary 4, 6. https://doi.org/10.3390/quat4010006

Clarke CL, Alsos IG, Edwards ME, Paus A, Gielly L, Haflidason H, Mangerud J, Regnéll C, Hughes PDM, Svendsen JI, Bjune AE (2020) A 24,000-year ancient DNA and pollen record from the Polar Urals reveals temporal dynamics of arctic and boreal plant communities. Quat. Sci. Rev. 247, 106564. https://doi.org/10.1016/j.quascirev.2020.106564

Clarke CL, Edwards ME, Gielly L, Ehrich D, Hughes PDM, Morozova LM, Haflidason H, Mangerud J, Svendsen JI, Alsos IG (2019) Persistence of arctic-alpine flora during 24,000 years of environmental change in the Polar Urals. Sci. Rep. 9, 19613. https://doi.org/10.1038/s41598-019-55989-9

Cooper A, Poinar HN (2000) Ancient DNA: do it right or not at all. Science 289, 1139. https://doi.org/10.1126/science.289.5482.1139b

Crump SE, Fréchette B, Power M, Cutler S, de Wet G, Raynolds MK, Raberg JH, Briner JP, Thomas EK, Sepúlveda J, Shapiro B, Bunce M, Miller GH (2021) Ancient plant DNA reveals High Arctic greening during the Last Interglacial. Proc Natl Acad Sci USA 118. https://doi.org/10.1073/pnas.2019069118

Cruzan MB, Templeton AR (2000) Paleoecology and coalescence: phylogeographic analysis of hypotheses from the fossil record. Trends Ecol. Evol. 15, 491–496. https://doi.org/10.1016/s0169-5347(00)01998-4

da Fonseca RR, Smith BD, Wales N, Cappellini E, Skoglund P, Fumagalli M, Samaniego JA, Carøe C, Ávila-Arcos MC, Hufnagel DE, Korneliussen TS, Vieira FG, Jakobsson M, Arriaza B, Willerslev E, Nielsen R, Hufford MB, Albrechtsen A, Ross-Ibarra J, Gilbert MTP (2015) The origin and evolution of maize in the Southwestern United States. Nat. Plants 1, 14003. https://doi.org/10.1038/nplants.2014.3

Dehasque M, Ávila-Arcos MC, Díez-Del-Molino D, Fumagalli M, Guschanski K, Lorenzen ED, Malaspinas A-S, Marques-Bonet T, Martin MD, Murray GGR, Papadopulos AST, Therkildsen NO, Wegmann D, Dalén L, Foote AD (2020) Inference of natural selection from ancient DNA. Evol. Lett. 4, 94–108. https://doi.org/10.1002/evl3.165

Eidesen PB, Alsos IG, Brochmann C (2015) Comparative analyses of plastid and AFLP data suggest different colonization history and asymmetric hybridization between *Betula pubescens* and *B. nana*. Mol. Ecol. 24, 3993–4009. https://doi.org/10.1111/mec.13289

Elbaum R, Melamed-Bessudo C, Tuross N, Levy AA, Weiner S (2009) New methods to isolate organic materials from silicified phytoliths reveal fragmented glycoproteins but no DNA. Quaternary International 193, 11–19. https://doi.org/10.1016/j.quaint.2007.07.006

Epp LS, Gussarova G, Boessenkool S, Olsen J, Haile J, Schrøder-Nielsen A, Ludikova A, Hassel K, Stenøien HK, Funder S, Willerslev E, Kjær K, Brochmann C (2015) Lake sediment multi-taxon DNA from North Greenland records early post-glacial appearance of vascular plants and accurately tracks environmental changes. Quat. Sci. Rev. 117, 152–163. https://doi.org/10.1016/j.quascirev.2015.03.027

Fijarczyk A, Babik W (2015) Detecting balancing selection in genomes: limits and prospects. Mol. Ecol. 24, 3529–3545. https://doi.org/10.1111/mec.13226

Fordyce SL, Ávila-Arcos MC, Rasmussen M, Cappellini E, Romero-Navarro JA, Wales N, Alquezar-Planas DE, Penfield S, Brown TA, Vielle-Calzada J-P, Montiel R, Jørgensen T, Odegaard N, Jacobs M, Arriaza B, Higham TFG, Ramsey CB, Willerslev E, Gilbert MTP (2013) Deep sequencing of RNA from ancient maize kernels. PLoS ONE 8, e50961. https://doi.org/10.1371/journal.pone.0050961

Giguet-Covex C, Pansu J, Arnaud F, Rey P-J, Griggo C, Gielly L, Domaizon I, Coissac E, David F, Choler P, Poulenard J, Taberlet P (2014) Long livestock farming history and human landscape shaping revealed by lake sediment DNA. Nat. Commun. 5, 3211. https://doi.org/10.1038/ncomms4211

Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat. Biotechnol. 27, 182–189. https://doi.org/10.1038/nbt.1523

Grass RN, Heckel R, Puddu M, Paunescu D, Stark WJ (2015) Robust chemical preservation of digital information on DNA in silica with error-correcting codes. Angew. Chem. Int. Ed 54, 2552–2555. https://doi.org/10.1002/anie.201411378

Green EJ, Speller CF (2017) Novel substrates as sources of ancient DNA: prospects and hurdles. Genes (Basel) 8, 180. https://doi.org/10.3390/genes8070180

Henn BM, Botigué LR, Bustamante CD, Clark AG, Gravel S (2015) Estimating the mutation load in human genomes. Nat. Rev. Genet. 16, 333–343. https://doi.org/10.1038/nrg3931

Higuchi R, Bowman B, Freiberger M, Ryder OA, Wilson AC (1984) DNA sequences from the quagga, an extinct member of the horse family. Nature 312, 282–284. https://doi.org/10.1038/312282a0

Hofman CA, Rick TC, Fleischer RC, Maldonado JE (2015) Conservation archaeogenomics: ancient DNA and biodiversity in the Anthropocene. Trends Ecol. Evol. 30, 540–549. https://doi.org/10.1016/j.tree.2015.06.008

Hofreiter M, Serre D, Poinar HN, Kuch M, Pääbo S (2001) Ancient DNA. Nat. Rev. Genet. 2, 353–359. https://doi.org/10.1038/35072071

Huang S, Herzschuh U, Pestryakova LA, Zimmermann HH, Davydova P, Biskaborn BK, Shevtsova I, Stoof-Leichsenring KR (2020) Genetic and morphologic determination of diatom community composition in surface sediments

from glacial and thermokarst lakes in the Siberian Arctic. J. Paleolimnol. 64, 225–242. https://doi.org/10.1007/s10933-020-00133-1

Ibrahim A, Capo E, Wessels M, Martin I, Meyer A, Schleheck D, Epp LS (2021) Anthropogenic impact on the historical phytoplankton community of Lake Constance reconstructed by multimarker analysis of sediment-core environmental DNA. Mol. Ecol. 30, 3040–3056. https://doi.org/10.1111/mec.15696

Jaenicke-Després V, Buckler ES, Smith BD, Gilbert MTP, Cooper A, Doebley J, Pääbo S (2003) Early allelic selection in maize as revealed by ancient DNA. Science 302, 1206–1208. https://doi.org/10.1126/science.1089056

Jónsson H, Ginolhac A, Schubert M, Johnson PLF, Orlando L (2013) mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics 29, 1682–1684. https://doi.org/10.1093/bioinformatics/btt193

Kapusta A, Suh A, Feschotte C (2017) Dynamics of genome size evolution in birds and mammals. Proc Natl Acad Sci USA 114, E1460–E1469. https://doi.org/10.1073/pnas.1616702114

Kirkpatrick JB, Walsh EA, D'Hondt S (2016) Fossil DNA persistence and decay in marine sediment over hundred-thousand-year to million-year time scales. Geology 44, 615–618. https://doi.org/10.1130/G37933.1

Kistler L, Bieker VC, Martin MD, Pedersen MW, Ramos Madrigal J, Wales N (2020) Ancient plant genomics in archaeology, herbaria, and the environment. Annu. Rev. Plant Biol. 71, 605–629. https://doi.org/10.1146/annurev-arplant-081519-035837

Kistler L, Maezumi SY, Gregorio de Souza J, Przelomska NAS, Malaquias Costa F, Smith O, Loiselle H, Ramos-Madrigal J, Wales N, Ribeiro ER, Morrison RR, Grimaldo C, Prous AP, Arriaza B, Gilbert MTP, de Oliveira Freitas F, Allaby RG (2018) Multiproxy evidence highlights a complex evolutionary legacy of maize in South America. Science 362, 1309–1313. https://doi.org/10.1126/science.aav0207

Kistler L, Montenegro A, Smith BD, Gifford JA, Green RE, Newsom LA, Shapiro B (2014) Transoceanic drift and the domestication of African bottle gourds in the Americas. Proc Natl Acad Sci USA 111, 2937–2941. https://doi.org/10.1073/pnas.1318678111

Kistler L, Newsom LA, Ryan TM, Clarke AC, Smith BD, Perry GH (2015) Gourds and squashes (Cucurbita spp.) adapted to megafaunal extinction and ecological anachronism through domestication. Proc Natl Acad Sci USA 112, 15107–15112. https://doi.org/10.1073/pnas.1516109112

Kistler L, Shapiro B (2011) Ancient DNA confirms a local origin of domesticated chenopod in eastern North America. J. Archaeol. Sci. 38, 3549–3554. https://doi.org/10.1016/j.jas.2011.08.023

Korneliussen TS, Albrechtsen A, Nielsen R (2014) ANGSD: analysis of next generation sequencing data. BMC Bioinformatics 15, 356. https://doi.org/10.1186/s12859-014-0356-4

Lammers Y, Heintzman PD, Alsos IG (2021) Environmental palaeogenomic reconstruction of an Ice Age algal population. Commun. Biol. 4, 220. https://doi.org/10.1038/s42003-021-01710-4

Larson G, Piperno DR, Allaby RG, Purugganan MD, Andersson L, Arroyo-Kalin M, Barton L, Climer Vigueira C, Denham T, Dobney K, Doust AN, Gepts P, Gilbert MTP, Gremillion KJ, Lucas L, Lukens L, Marshall FB, Olsen KM, Pires JC, Richerson PJ, Fuller DQ (2014) Current perspectives and the future of domestication studies. Proc Natl Acad Sci USA 111, 6139–6146. https://doi.org/10.1073/pnas.1323964111

Liu S, Kruse S, Scherler D, Ree RH, Zimmermann HH, Stoof-Leichsenring KR, Epp LS, Mischke S, Herzschuh U (2021) Sedimentary ancient DNA reveals a threat of warming-induced alpine habitat loss to Tibetan Plateau plant diversity. Nat. Commun. 12, 2995. https://doi.org/10.1038/s41467-021-22986-4

Malaspinas A-S, Tange O, Moreno-Mayar JV, Rasmussen M, DeGiorgio M, Wang Y, Valdiosera CE, Politis G, Willerslev E, Nielsen R (2014) bammds: a tool for assessing the ancestry of low-depth whole-genome data using multidimensional scaling (MDS). Bioinformatics 30, 2962–2964. https://doi.org/10.1093/bioinformatics/btu410

Malaspinas A-S (2016) Methods to characterize selective sweeps using time serial samples: an ancient DNA perspective. Mol. Ecol. 25, 24–41. https://doi.org/10.1111/mec.13492

Mascher M, Schuenemann VJ, Davidovich U, Marom N, Himmelbach A, Hübner S, Korol A, David M, Reiter E, Riehl S, Schreiber M, Vohr SH, Green RE, Dawson IK, Russell J, Kilian B, Muehlbauer GJ, Waugh R, Fahima T, Krause J, Stein N (2016) Genomic analysis of 6,000-year-old cultivated grain illuminates the domestication history of barley. Nat. Genet. 48, 1089–1093. https://doi.org/10.1038/ng.3611

Mitchell KJ, Rawlence NJ (2021) Examining natural history through the lens of palaeogenomics. Trends Ecol. Evol. 36, 258–267. https://doi.org/10.1016/j.tree.2020.10.005

Modi A, Vergata C, Zilli C, Vischioni C, Vai S, Tagliazucchi GM, Lari M, Caramelli D, Taccioli C (2021) Successful extraction of insect DNA from recent copal inclusions: limits and perspectives. Sci. Rep. 11, 6851. https://doi.org/10.1038/s41598-021-86058-9

Monchamp M-E, Spaak P, Pomati F (2019) High dispersal levels and lake warming are emergent drivers of cyanobacterial community assembly in peri-Alpine lakes. Sci. Rep. 9, 7366. https://doi.org/10.1038/s41598-019-43814-2

Murray GGR, Soares AER, Novak BJ, Schaefer NK, Cahill JA, Baker AJ, Demboski JR, Doll A, Da Fonseca RR, Fulton TL, Gilbert MTP, Heintzman PD, Letts B, McIntosh G, O'Connell BL, Peck M, Pipes M-L, Rice ES, Santos KM, Sohrweide AG, Shapiro B (2017) Natural selection shaped the rise and fall of passenger pigeon genomic diversity. Science 358, 951–954. https://doi.org/10.1126/science.aao0960

Nistelberger HM, Smith O, Wales N, Star B, Boessenkool S (2016) The efficacy of high-throughput sequencing and target enrichment on charred archaeobotanical remains. Sci. Rep. 6, 37347. https://doi.org/10.1038/srep37347

Orlando L, Cooper A (2014) Using ancient DNA to understand evolutionary and ecological processes. Annu. Rev. Ecol. Evol. Syst. 45, 573–598. https://doi.org/10.1146/annurev-ecolsys-120213-091712

Otto-Bliesner BL, Rosenbloom N, Stone EJ, McKay NP, Lunt DJ, Brady EC, Overpeck JT (2013) How warm was the last interglacial? New model-data comparisons. Philos. Transact. A Math. Phys. Eng. Sci. 371, 20130097. https://doi.org/10.1098/rsta.2013.0097

Palmer SA, Clapham AJ, Rose P, Freitas FO, Owen BD, Beresford-Jones D, Moore JD, Kitchen JL, Allaby RG (2012) Archaeogenomic evidence of punctuated genome evolution in *Gossypium*. Mol. Biol. Evol. 29, 2031–2038. https://doi.org/10.1093/molbev/mss070

Parducci L, Alsos IG, Unneberg P, Pedersen MW, Han L, Lammers Y, Salonen JS, Väliranta MM, Slotte T, Wohlfarth B (2019) Shotgun environmental DNA, pollen, and macrofossil analysis of lateglacial lake sediments from southern Sweden. Front. Ecol. Evol. 7. https://doi.org/10.3389/fevo.2019.00189

Parducci L, Bennett KD, Ficetola GF, Alsos IG, Suyama Y, Wood JR, Pedersen MW (2017) Ancient plant DNA in lake sediments. New Phytol. 214, 924–942. https://doi.org/10.1111/nph.14470

Parducci L, Suyama Y, Lascoux M, Bennett KD (2005) Ancient DNA from pollen: a genetic record of population history in Scots pine. Mol. Ecol. 14, 2873–2882. https://doi.org/10.1111/j.1365-294X.2005.02644.x

Pedersen MW, Ginolhac A, Orlando L, Olsen J, Andersen K, Holm J, Funder S, Willerslev E, Kjær KH (2013) A comparative study of ancient environmental DNA to pollen and macrofossils from lake sediments reveals taxonomic overlap and additional plant taxa. Quat. Sci. Rev. 75, 161–168. https://doi.org/10.1016/j.quascirev.2013.06.006

Peris D, Janssen K, Barthel HJ, Bierbaum G, Delclòs X, Peñalver E, Solórzano-Kraemer MM, Jordal BH, Rust J (2020) DNA from resin-embedded organisms: past, present and future. PLoS ONE 15, e0239521. https://doi.org/10.1371/journal.pone.0239521

Poinar HN, Hofreiter M, Spaulding WG, Martin PS, Stankiewicz BA, Bland H, Evershed RP, Possnert G, Pääbo S (1998) Molecular coproscopy: dung and diet of the extinct ground sloth *Nothrotheriops shastensis*. Science 281, 402–406. https://doi.org/10.1126/science.281.5375.402

Racimo F, Berg JJ, Pickrell JK (2018) Detecting polygenic adaptation in admixture graphs. Genetics 208, 1565–1584. https://doi.org/10.1534/genetics.117.300489

Ramos-Madrigal J, Runge AKW, Bouby L, Lacombe T, Samaniego Castruita JA, Adam-Blondon A-F, Figueiral I, Hallavant C, Martínez-Zapater JM, Schaal C, Töpfer R, Petersen B, Sicheritz-Pontén T, This P, Bacilieri R, Gilbert MTP, Wales N (2019) Palaeogenomic insights into the origins of French grapevine diversity. Nat. Plants 5, 595–603. https://doi.org/10.1038/s41477-019-0437-5

Rasmussen M, Cummings LS, Gilbert MTP, Bryant V, Smith C, Jenkins DL, Willerslev E (2009) Response to comment by Goldberg et al. on "DNA from pre-Clovis human coprolites in Oregon, North America." Science 325, 148–148. https://doi.org/10.1126/science.1167672

Rijal DP, Heintzman PD, Lammers Y, Yoccoz NG, Lorberau KE, Pitelkova I, Goslar T, Murguzur FJA, Salonen JS, Helmens KF, Bakke J, Edwards ME, Alm T, Bråthen KA, Brown AG, Alsos IG (2021) Sedimentary ancient DNA shows terres-

trial plant richness continuously increased over the Holocene in northern Fennoscandia. Sci. Adv. 7, eabf9557. https://doi.org/10.1126/sciadv.abf9557

Rollo F, Ubaldi M, Ermini L, Marota I (2002) Otzi's last meals: DNA analysis of the intestinal content of the Neolithic glacier mummy from the Alps. Proc Natl Acad Sci USA 99, 12594–12599. https://doi.org/10.1073/pnas.192184599

Rollo F (1985) Characterisation by molecular hybridization of RNA fragments isolated from ancient (1400 B.C.) seeds. Theor. Appl. Genet. 71, 330–333. https://doi.org/10.1007/BF00252076

Schlumbaum A, Tensen M, Jaenicke-Després V (2008) Ancient plant DNA in archaeobotany. Veg. Hist. Archaeobot. 17, 233–244. https://doi.org/10.1007/s00334-007-0125-7

Schubert M, Ermini L, Der Sarkissian C, Jónsson H, Ginolhac A, Schaefer R, Martin MD, Fernández R, Kircher M, McCue M, Willerslev E, Orlando L (2014) Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. Nat. Protoc. 9, 1056–1082. https://doi.org/10.1038/nprot.2014.063

Schulte L, Bernhardt N, Stoof-Leichsenring K, Zimmermann HH, Pestryakova LA, Epp LS, Herzschuh U (2021) Hybridization capture of larch (*Larix* Mill.) chloroplast genomes from sedimentary ancient DNA reveals past changes of Siberian forest. Mol. Ecol. Resour. 21, 801–815. https://doi.org/10.1111/1755-0998.13311

Schwörer C, Leunda M, Alvarez N, Gugerli F, Sperisen C (2022) The untapped potential of macrofossils in ancient plant DNA research. New Phytol. https://doi.org/10.1111/nph.18108

Scott MF, Botigué LR, Brace S, Stevens CJ, Mullin VE, Stevenson A, Thomas MG, Fuller DQ, Mott R (2019) A 3,000-year-old Egyptian emmer wheat genome reveals dispersal and domestication history. Nat. Plants 5, 1120–1128. https://doi.org/10.1038/s41477-019-0534-5

Slon V, Hopfe C, Weiß CL, Mafessoni F, de la Rasilla M, Lalueza-Fox C, Rosas A, Soressi M, Knul MV, Miller R, Stewart JR, Derevianko AP, Jacobs Z, Li B, Roberts RG, Shunkov MV, de Lumley H, Perrenoud C, Gušić I, Kućan Ž, Meyer M (2017) Neandertal and Denisovan DNA from Pleistocene sediments. Science 356, 605–608. https://doi.org/10.1126/science.aam9695

Smith O, Clapham AJ, Rose P, Liu Y, Wang J, Allaby RG (2014) Genomic methylation patterns in archaeological barley show de-methylation as a time-dependent diagenetic process. Sci. Rep. 4, 5559. https://doi.org/10.1038/srep05559

Smith O, Momber G, Bates R, Garwood P, Fitch S, Pallen M, Gaffney V, Allaby RG (2015) Archaeology. Sedimentary DNA from a submerged site reveals wheat in the British Isles 8000 years ago. Science 347, 998–1001. https://doi.org/10.1126/science.1261278

Smith O, Palmer SA, Clapham AJ, Rose P, Liu Y, Wang J, Allaby RG (2017) Small RNA activity in archeological barley shows novel germination inhibition in response to environment. Mol. Biol. Evol. 34, 2555–2562. https://doi.org/10.1093/molbev/msx175

Søe MJ, Nejsum P, Seersholm FV, Fredensborg BL, Habraken R, Haase K, Hald MM, Simonsen R, Højlund F, Blanke L, Merkyte I, Willerslev E, Kapel CMO (2018) Ancient DNA from latrines in Northern Europe and the Middle East (500 BC-1700 AD) reveals past parasites and diet. PLoS ONE 13, e0195481. https://doi.org/10.1371/journal.pone.0195481

Sønstebø JH, Gielly L, Brysting AK, Elven R, Edwards M, Haile J, Willerslev E, Coissac E, Rioux D, Sannier J, Taberlet P, Brochmann C (2010) Using next-generation sequencing for molecular reconstruction of past Arctic vegetation and climate. Mol. Ecol. Resour. 10, 1009–1018. https://doi.org/10.1111/j.1755-0998.2010.02855.x

Swarts K, Gutaker RM, Benz B, Blake M, Bukowski R, Holland J, Kruse-Peeples M, Lepak N, Prim L, Romay MC, Ross-Ibarra J, Sanchez-Gonzalez J de J, Schmidt C, Schuenemann VJ, Krause J, Matson RG, Weigel D, Buckler ES, Burbano HA (2017) Genomic estimation of complex traits reveals ancient maize adaptation to temperate North America. Science 357, 512–515. https://doi.org/10.1126/science.aam9425

Trucchi E, Benazzo A, Lari M, Iob A, Vai S, Nanni L, Bellucci E, Bitocchi E, Raffini F, Xu C, Jackson SA, Lema V, Babot P, Oliszewski N, Gil A, Neme G, Michieli CT, De Lorenzi M, Calcagnile L, Caramelli D, Bertorelle G (2021) Ancient genomes reveal early Andean farmers selected common beans while preserving diversity. Nat. Plants 7, 123–128. https://doi.org/10.1038/s41477-021-00848-7

Wagner S, Lagane F, Seguin-Orlando A, Schubert M, Leroy T, Guichoux E, Chancerel E, Bech-Hebelstrup I, Bernard V, Billard C, Billaud Y, Bolliger M, Croutsch C, Čufar K, Eynaud F, Heussner KU, Köninger J, Langenegger F, Leroy F, Lima C, Orlando L (2018) High-Throughput DNA sequencing of ancient wood. Mol. Ecol. 27, 1138–1154. https://doi.org/10.1111/mec.14514

Wales N, Kistler L (2019) Extraction of ancient DNA from plant remains. Methods Mol. Biol. 1963, 45–55. https://doi.org/10.1007/978-1-4939-9176-1_6

Wales N, Ramos Madrigal J, Cappellini E, Carmona Baez A, Samaniego Castruita JA, Romero-Navarro JA, Carøe C, Ávila-Arcos MC, Peñaloza F, Moreno-Mayar JV, Gasparyan B, Zardaryan D, Bagoyan T, Smith A, Pinhasi R, Bosi G, Fiorentino G, Grasso AM, Celant A, Bar-Oz G, Gilbert MTP (2016) The limits and potential of paleogenomic techniques for reconstructing grapevine domestication. J. Archaeol. Sci. 72, 57–70. https://doi.org/10.1016/j.jas.2016.05.014

Willerslev E, Davison J, Moora M, Zobel M, Coissac E, Edwards ME, Lorenzen ED, Vestergård M, Gussarova G, Haile J, Craine J, Gielly L, Boessenkool S, Epp LS, Pearman PB, Cheddadi R, Murray D, Bråthen KA, Yoccoz N, Binney H, Taberlet P (2014) Fifty thousand years of Arctic vegetation and megafaunal diet. Nature 506, 47–51. https://doi.org/10.1038/nature12921

Willerslev E, Hansen AJ, Binladen J, Brand TB, Gilbert MTP, Shapiro B, Bunce M, Wiuf C, Gilichinsky DA, Cooper A (2003) Diverse plant and animal genetic records from Holocene and Pleistocene sediments. Science 300, 791–795. https://doi.org/10.1126/science.1084114

Wood JR, Crown A, Cole TL, Wilmshurst JM (2016) Microscopic and ancient DNA profiling of Polynesian dog (kurī) coprolites from northern New Zealand. Journal of Archaeological Science: Reports 6, 496–505. https://doi.org/10.1016/j.jasrep.2016.03.020

Zedane L, Hong-Wa C, Murienne J, Jeziorski C, Baldwin BG, Besnard G (2016) Museomics illuminate the history of an extinct, paleoendemic plant lineage (*Hesperelaea* , Oleaceae) known from an 1875 collection from Guadalupe Island, Mexico. Biological Journal of the Linnean Society 117, 44–57. https://doi.org/10.1111/bij.12509

Zimmermann HH, Raschke E, Epp LS, Stoof-Leichsenring KR, Schirrmeister L, Schwamborn G, Herzschuh U (2017) The history of tree and shrub taxa on Bol'shoy Lyakhovsky Island (New Siberian Archipelago) since the last interglacial uncovered by sedimentary ancient DNA and pollen data. Genes (Basel) 8, 273. https://doi.org/10.3390/genes8100273

Answers

1. Several biases specific to aDNA analysis lead to an incorrect identification of the specimen species. The low quantity of aDNA in historical specimens can increase the effect of cross-contamination between samples and differential amplification of the DNA fragments during the PCR process of making genomic libraries. Different replicates of the samples can be analysed in separate facilities to test reproducibility of the results, and a negative control devoid of DNA can be used to check contamination. Ancient DNA damages such as substitution or deletion can affect the DNA sequence itself and lead to incorrect identification. Software assessing aDNA damage and recalibrating the alignment file can be used to minimise this bias. Another source of error can be the incompleteness of the plastid reference database used to match the sequencing reads. If many species are missing from the reference database, the detection might occur at genus level instead of species level. For more information, this question is based on a study that characterised the diet and intestinal parasites of ancient communities in Northern Europe and Middle East from latrines remains aDNA (Søe et al. 2018).
2. To study the evolution of plant richness over time, we can use a time series of samples to reconstruct the evolution of vegetation diversity at the sampling point. A range of datasets

from several sampling points can be used to model the Holocene glacial retreat over time. Lake sedaDNA can be an adequate source of aDNA to study climatic change via taxonomic plant diversity detection. SedaDNA is extracted from lake sediment cores, each sediment layer of the core corresponds to a different era. This kind of sampling might provide a measure of plant vegetation richness before and after deglaciation, and might be used to confirm models of the Holocene glacial retreat. For more information about lake sedaDNA cores used to reconstruct changes in plant diversity over time and geographically, have a look at a study using sedaDNA to characterise the emergence of vascular plants after glaciation in Greenland (Epp et al. 2015) or another study reconstructing the post-glacial plant colonisation of Iceland (Alsos et al. 2021; Epp et al. 2015).

Chapter 22
Healthcare

Felicitas Mück[1], Carlos A. Vásquez-Londoño[2]

1 Department of Pharmacy, Faculty of Mathematics and Natural Sciences, University of Oslo, Oslo, Norway
2 Department of Pharmacy, Faculty of Sciences, National University of Colombia, Bogotá, Colombia

Felicitas Mück **felicitas.mueck@farmasi.uio.no**
Carlos A. Vásquez-Londoño **cavasquezl@unal.edu.co**

Citation: Mück F, Vásquez-Londoño CA (2022) Chapter 22. Healthcare. In: de Boer H, Rydmark MO, Verstraete B, Gravendeel B (Eds) Molecular identification of plants: from sequence to species. Advanced Books. https://doi.org/10.3897/ab.e98875

Plants as medicine

Plants have been used as medicines for millennia in diverse geographical and cultural contexts. They continue to play an essential role as therapeutic and prophylactic agents in traditional and complementary medicine (Palhares et al. 2015; Schmidt et al. 2008). Scientific evidence supports traditional applications of some medicinal plants, and current research and discovery of plant-derived natural compounds offers promising alternatives for local and global health remedies (Howes et al. 2020). Adulteration of crude drugs and phytotherapeutic products however makes authentication challenging (see Chapter 23 Food safety). Routine authentication techniques are based on a long tradition in pharmacovigilance and set the current industry standards. DNA-based methods are however becoming increasingly popular for species identification in quality control measures for herbal medicines and phytotherapeutic products as they target the biological species rather than the bioactive compounds (Anthoons et al. 2020; Ichim 2019; Seethapathy et al. 2019). DNA-based authentication of herbal medicines can detect adulterant species and can thus serve as an important safeguard in the quality control of raw or processed plant products in healthcare (Ernst et al. 2016; Hao and Xiao 2020; Saslis-Lagoudakis et al. 2012).

Quality control and authentication of herbal medicines

Accurate medicinal plant identification is very often a challenging task. Products can either be from single species or from mixtures, can be in a dried, fragmented, or powdered form, and can originate from plant leaves, flowers, stems, barks, roots, fruit, and seeds. They may also come in the form of phytopharmaceutical products, including oral, topical, parenteral, ophthalmic, or inhaled forms. Most regulatory guidelines and pharmacopoeias for conventional plant authentication are based on diagnostic morphological or chemical features.

Macroscopic botanical analysis for herbal drug authentication is complicated in many cases, since it requires an experienced taxonomist and the plant samples should include flowers or fruits along with a segment of the stem with leaves enough to observe branching patterns (Ichim et al. 2020). These conditions are usually not available in fragmented, powdered, or processed samples (Palhares et al. 2015). Microscopic pharmacognostic methods require experts, are time consuming, and imply the identification of specific cell types or diagnostic histological features, which can be absent in processed products or may not provide enough data to allow the discrimination between closely related species (Shen et al. 2019).

Chemical authentication of herbal medicines includes metabolite detection, quantification, profiling, and elucidation through analytical methods such as thin layer chromatography (TLC), high-performance thin layer chromatography (HPTLC), high-performance liquid chromatography (HPLC), liquid chromatography coupled to mass spectrometry (LC-MS), gas chromatography coupled to mass spectrometry (GC-MS), or nuclear magnetic resonance (NMR) (Cuadros-Rodríguez et al. 2016). Phytochemical analyses are useful to verify if the levels of active compounds are consistent in herbal products and allow the detection of contaminants originating from cultivation, harvesting, or manufacturing steps. It is also possible to differentiate between identified plant organs according to metabolite screening of well characterised medicinal plant species (Ichim and Booker 2021). However, there are different plant metabolites present within different parts of the plant. Additionally, even within the same plant species the metabolic profile can vary. Thus, using chemical authentication methods only for species identification is not always adequate.

Molecular plant identification techniques such as DNA barcoding have proven to be cost-effective procedures useful in pharmacovigilance to authenticate herbal medicines at species level and to detect adulterants (de Boer et al. 2015a). Molecular identification of medicinal plants has been suggested for routine market surveillance and for screening the quality of raw materials in early stages of the herbal supply chain (see Chapter 23 Food safety). DNA-based methods come with several advantages, including that DNA is present in all plant organs, its presence is less sensitive to external factors than metabolites, and it can be used for the identification of dried and powdered products where morphological characteristics are absent (Grazina et al. 2020; Ichim 2019). Thus, DNA-based methods for herbal product identification are increasingly accepted.

DNA-based identification of adulterants in herbal medicines

The increase in herbal medicine adulteration is of growing concern due to the expansion of the global market for natural products (Lee and Hxiao 2019). Adulteration comprises substitutions of genuine ingredients by other species or substances that may have scarce or zero medicinal property and might even induce adverse reactions. Substituents and contaminants are commonly reported in herbal remedies that are traded on the international market, increasing the risk of harm for consumers of traditional and complementary medicine (Posadzki et al. 2013). Herbal medicine adulteration may be caused unintentionally by misidentification of closely related plant species, or might be fraudulent through substituting authentic ingredients with less expensive, prohibited, ineffective, or harmful species or substances. Discrepancies between the vernacular names and biological species have been reported as a cause of substitution in herbal products as well. Natural product adulteration is more commonly reported in powdered raw drugs than in intact plant materials. Triturated samples are more susceptible to substitution since they are more difficult to differentiate morphologically by consumers, traders, and taxonomists (Vassou et al. 2016). The high market value for some commercialised medicines, such as ginseng products, can incentivise their fraudulent adulteration (Ichim and de Boer 2020). The adulteration of medicinal plants has provoked serious health problems for consumers, such as progressive renal failure requiring dialysis and kidney transplantation, as well as urothelial carcinoma cases that were reported to occur in a group of women taking slimming pills containing nephrotoxic alkaloids from *Aristolochia fangchi*, a substituent in the herbal medicine *Stephania tetrandra* (Nortier et al. 2000). DNA-based plant identification techniques for the rapid screening of herbal medicines and the reliable detection of adulterant species on the global market can thus serve to minimise health risks involved in the worldwide trade of medicinal herbs (Howard et al. 2020; Ichim and de Boer 2020).

A number of methods for identifying medicinal plants using DNA-based methods have been previously described. Several barcoding regions have been shown to be effective for identifying adulterants and physiologically difficult-to-discriminate plant species from a variety of sample forms (Amritha et al. 2020; Anthoons et al. 2020; Bansal et al. 2018; Ghorbani et al. 2020; Selvaraj et al. 2012; Sheidai et al. 2019). While many of these studies were performed with a single barcode, a combination of barcodes may also be necessary for an incremental identification approach. For example, a combination of *matK*, *rbcL*, and nrITS was tested for the authentication of 27 toxic plant species in herbal products from China (Xie et al. 2014). Additionally, metabarcoding can be used for single or multi-ingredient products in dried, powdered, and highly processed products, and is useful for identifying adulterants and confirming product label information. Metabarcoding approaches generally use shorter barcodes, i.e. shorter segments of normal markers, such

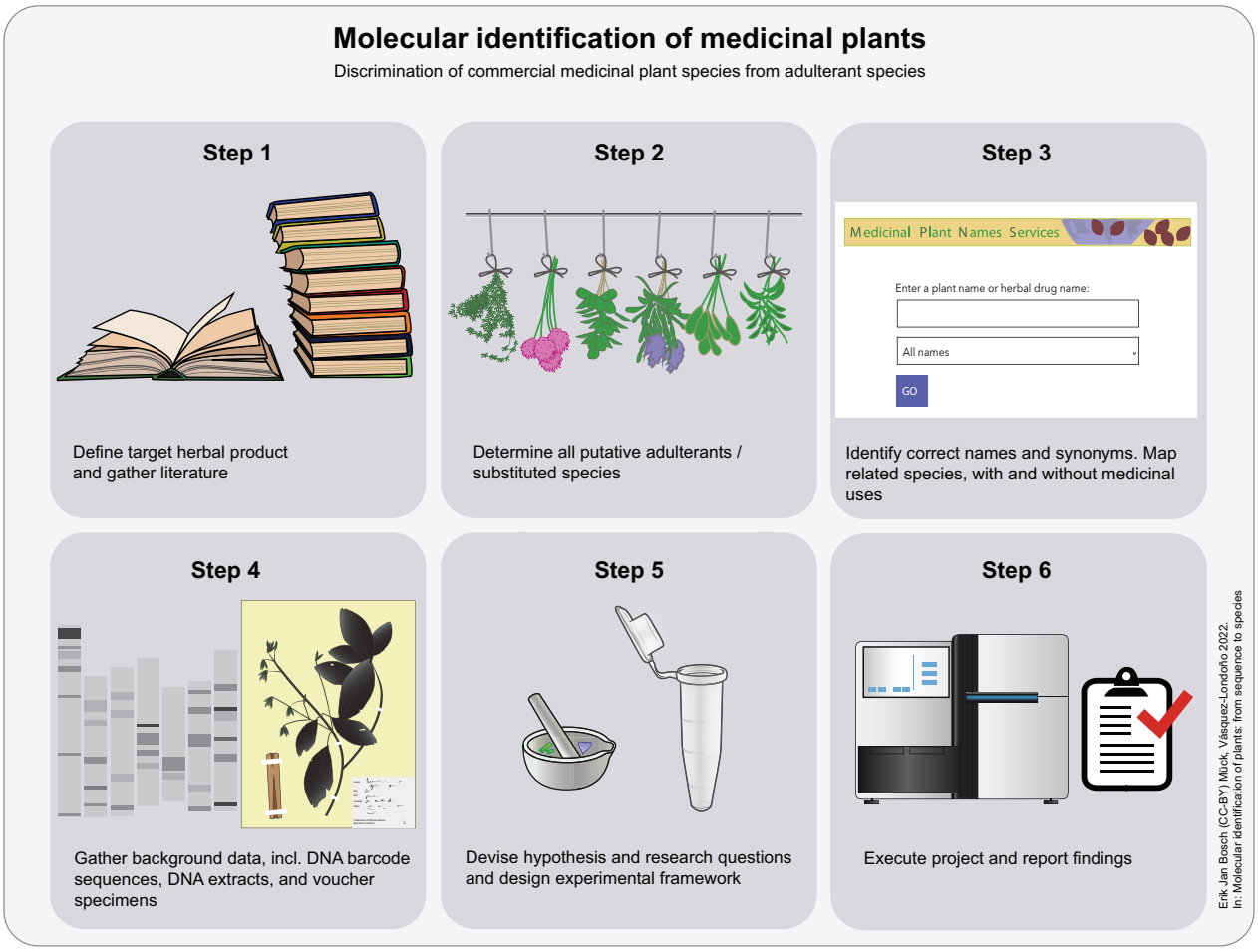

Figure 1. Chapter 22 Infographic: Visual representation of methodological steps for molecular identification of medicinal plants.

as the trnL intron P6 loop from the *trnL* intron (Taberlet et al. 2007); the *rbcL* mini-barcode from *rbcL* (Little 2014) or nrITS1 or nrITS2 instead of the full nrITS (Raclariu et al. 2018, 2017a, 2017b). These DNA-based methods may also be combined with other analytical methods, such as NMR, TLC, and LC-MS methods for identification as well detection and quantification of metabolites (Palhares et al. 2015; Raclariu et al. 2017b; Seethapathy et al. 2019; Urumarudappa et al. 2016; Vassou et al. 2015). For example, high levels of substitutions were detected and identified in eight medicinal plant (MP) products approved by WHO and sold in Brazilian markets by combining *matK*, *rbcL*, and nrITS2 sequences with TLC (Palhares et al. 2015). Interestingly, these studies highlight that some samples recognised as the correct species using DNA-based methods did not have detectable levels of the chemical compounds typically used for their identification, while some of the substitute species showed low concentrations of the expected metabolites.

DNA-based identification in ethnopharmacology

Ethnopharmacology studies the use of drugs made by humans, and integrates anthropological, pharmacological, toxicological, and chemical approaches (Heinrich 2014). Molecular

authentication techniques are useful in ethnopharmacological research to identify plants used in traditional healthcare, both in herbal markets, rural communities, or urban contexts. DNA barcoding techniques with a single barcode or a combination of markers have been successfully used for commercialised medicinal products in a number of internationally commercialised herbal products, from genus to species level (Coghlan et al. 2012; Jia et al. 2017; Osathanunkul et al. 2015; Seethapathy et al. 2019). Similarly to the internationally traded products discussed above, these techniques can be used in combination with other analytical techniques or morphological methods to identify plant species and the potential presence of adulterants. A number of case studies exist where DNA barcoding for plant identification was utilised in ethnopharmacological studies (Costa et al. 2016; Posthouwer et al. 2018). Particularly relevant examples include the identification of medicinal plant species commercialised in Moroccan herbal markets (de Boer et al. 2014; Kool et al. 2012; Manzanilla et al. 2022), the authentication of *Bupleurum* species utilised in traditional Chinese medicine (TCM) (Chao et al. 2014), and the identification of Thai medicinal plants (Rungpragayphan et al. 2014). DNA barcoding in combination with high resolution melting (Bar-HRM; see Chapter 13 Barcoding - High Resolution Melting) has enabled the authentication of several medicinal plant products with increased discriminatory power when using multi-locus combinations rather than single-locus barcoding (Osathanunkul et al. 2016). Finally, authentication of herbal medicines containing a mixture of different species is often challenging. Here High Throughput Sequencing (HTS)-based metabarcoding methods are a useful complement to existing chemical analytical methods to assess species diversity in commercial products (de Boer et al. 2017; Raclariu et al. 2018, 2017a, 2017b; Seethapathy et al. 2019).

DNA-based identification in natural product research and bioprospecting

DNA-based identification of plant ingredients in herbal medicines is important for resolving taxonomic controversies, assessing the genetic variability and evolutionary traits of medicinal plants, as well as enabling the detection and further conservation of endangered, illegally traded species (de Boer et al. 2017). Molecular identification of medicinal plants is an important force in driving taxonomic research on medicinal species, guiding forensic DNA and toxicological research. Some examples of applications within the field include nrITS2 barcoding of 90 Fabaceae species from China, including 24 species approved in the Chinese Pharmacopoeia (Gao et al. 2010); Five Chinese *Stephania* species, with morphologically and chemically similar features, were also successfully distinguished by DNA barcoding using the nrITS2 sequence combined with LC-MS and HPLC (Zhao et al. 2020). In evolutionarily complex groups, conventional DNA barcoding using single barcodes may fail to resolve relationships at species level. For this reason, it is preferred to establish the best association of markers to optimise species delimitation (Kreuzer et al. 2019; Li et al. 2020; Manzanilla et al. 2018; Shen et al. 2019; Wang et al. 2013), and there are several cases where plastid regions or even plastomes and large single-copy analyses have been used for species discrimination (Kreuzer et al. 2019; Li et al. 2020; Manzanilla et al. 2018). For instance, high-throughput sequencing of MBD2-Fc fractionated *Panax* DNA sampled in Vietnam enabled the creation of a phylogenetic tree to understand the relationships between *Panax* species (Manzanilla et al. 2018). DNA techniques have also shown their utility in analysing the genetic variability of micropropagated plants. DNA barcoding using *rbcL* gene primers together with start codon targeted (SCoT), inter simple sequence repeats (ISSR) markers and foliar micromorphological analysis allowed the confirmation of the clonal nature of in

vitro generated and mother plants of *Artemisia vulgaris* (Jogam et al. 2020). The genetic uniformity of regenerated plantlets of *Andrographis echioides*, an important medicinal plant in South Asia, was corroborated by DNA barcoding using *rbcL* and ISSR (Savitikadi et al. 2020).

The vast majority of commercialised medicinal plants are collected from wild resources, and in many cases they are overexploited and some are becoming increasingly scarce. These factors threaten the conservation of endangered plants, endemic species, or species with limited distributions. DNA analysis is also important for the detection of endangered species through screening marketed natural products. For instance, DNA barcoding using *rbcL*, *matK*, *psbA-trnH*, and nrITS allowed the identification of species of the cycad genus *Encephalartos*, which are catalogued as threatened and are illegally traded in South African herbal markets (Williamson et al. 2016). The identity and geographic origin of samples traded as *Anacyclus pyrethrum*, a red-listed medicinal plant, was evaluated using target-capture genomic DNA barcoding (see Chapter 14 Target capture) with 443 markers, constituting a promising approach for endangered species monitoring, research, and conservation (Manzanilla et al. 2022).

Legal framework

According to the WHO, accurate identification of medicinal plants is an essential measure for the assurance of the quality, safety, and effectiveness of natural medicines (Sheidai et al. 2019). Most of the pharmacopoeia standards available for the authentication of herbal medicines are based on chemical and morphological methods, and molecular identification of medicinal plants were only recently included for regulatory purposes into the British and Chinese Pharmacopoeias (Kreuzer et al. 2019). Regulatory policies for the manufacturing, utilisation, and commercialisation of herbal medicines differ among countries, however, raising concerns about protocols for authentication, quality, and safety, and hindering uninterrupted international trade. It has therefore been proposed to strengthen pharmacovigilance strategies regarding the authentication of raw materials, adulterant detection, and manufacturing practices (Bansal et al. 2018; de Boer et al. 2015b; Tnah et al. 2019). The Nagoya protocol regulates international access to genetic resources and guarantees that benefits derived from their commercialisation and use are fairly and equitably shared among countries. Thus, molecular plant identification also constitutes a strategy to ensure compliance to the Nagoya protocol and other regulations regarding intellectual property rights and patenting (Campanaro et al. 2019).

Future perspectives of DNA-based identification for healthcare

Authentication of medicinal plants and the detection of adulterants is a crucial concern for regulatory agencies and phytopharmaceutical industries in order to guarantee optimal quality, safety, and efficacy of herbal products for consumers. Challenges associated with conventional pharmacognostic procedures to authenticate processed or multi-ingredient herbal products can be mitigated with DNA technologies, enabling the accurate identification of medicinal species and substituents in complex samples. It is recommended to combine molecular, chemical, and morphological plant identification methods to increase the discriminatory capacity of authentication approaches (Hao and Xiao 2020). Development of validated methods using DNA metabarcoding

will increase reproducibility of results and allow for the setting of standards for exclusion of false positives (Arulandhu et al. 2019, 2017). A more standardised and coordinated regulation of natural products by countries, increasing efforts by manufacturers to authenticate botanical raw materials will increase consumer trust in traditional medicines, minimise health risks derived from adulteration and enable a more fluent and trustworthy global trade (Tnah et al. 2019).

Questions

1. List three categories of herbal medicine adulteration.
2. Why are genetic methods an important tool for plant identification in ethnopharmacology in addition to analytical chemical methods?
3. Molecular identification has various applications in healthcare. How could molecular identification aid bioprospecting?

Glossary

Herbal medicines – Plant(s) or plant part(s) or extract(s) used to improve health, and to prevent and treat disease.

Phytopharmaceutical – Pharmaceutical agents derived from plants or plant parts, for which the active compounds are known.

Traditional medicines – Knowledge, skills, and practices based on traditional cultures, aimed to promote health and to diagnose, prevent, and treat disease.

Trituration – Trituration refers to different methods used to reduce the particle size of a substance and to produce homogeneous material from various components.

Complementary medicines – Health practices not belonging to the tradition of a country or to conventional medicine.

Ethnopharmacology – Scientific study of drugs traditionally used by people.

Pharmacovigilance – A pharmacological science relating to the collection, detection, assessment, monitoring, and prevention of adverse effects of pharmaceutical products.

Phylogeny – History of the evolution of a species or group, especially in reference to lines of descent and relationships among broad groups of organisms.

Bioprospecting – The exploration of natural sources for small molecules, macromolecules, and biochemical and genetic information that could be developed into commercially valuable products.

References

Amritha N, Bhooma V, Parani M (2020) Authentication of the market samples of Ashwagandha by DNA barcoding reveals that powders are significantly more adulterated than roots. J. Ethnopharmacol. 256, 112725. https://doi.org/10.1016/j.jep.2020.112725

Anthoons B, Karamichali I, Schrøder-Nielsen A, Drouzas AD, de Boer H, Madesis P (2020) Metabarcoding reveals low fidelity and presence of toxic species in short chain-of-commercialization of herbal products. Journal of Food Composition and Analysis 103767. https://doi.org/10.1016/j.jfca.2020.103767

Arulandhu AJ, Staats M, Hagelaar R, Peelen T, Kok EJ (2019) The application of multi-locus DNA metabarcoding in traditional medicines. Journal of Food Composition and Analysis 79, 87–94. https://doi.org/10.1016/j.jfca.2019.03.007

Arulandhu AJ, Staats M, Hagelaar R, Voorhuijzen MM, Prins TW, Scholtens I, Costessi A, Duijsings D, Rechenmann F, Gaspar FB, Barreto Crespo MT, Holst-Jensen A, Birck M, Burns M, Haynes E, Hochegger R, Klingl A, Lundberg L, Natale C, Niekamp H, Kok E (2017) Development and validation of a multi-locus DNA metabarcoding method to identify endangered species in complex samples. Gigascience 6, 1–18. https://doi.org/10.1093/gigascience/gix080

Bansal S, Thakur S, Mangal M, Mangal AK, Gupta RK (2018) DNA barcoding for specific and sensitive detection of Cuminum cyminum adulteration in *Bunium persicum*. Phytomedicine 50, 178–183. https://doi.org/10.1016/j.phymed.2018.04.023

Campanaro A, Tommasi N, Guzzetti L, Galimberti A, Bruni I, Labra M (2019) DNA barcoding to promote social awareness and identity of neglected, underutilized plant species having valuable nutritional properties. Food Res. Int. 115, 1–9. https://doi.org/10.1016/j.foodres.2018.07.031

Chao Z, Zeng W, Liao J, Liu L, Liang Z, Li X (2014) DNA barcoding Chinese medicinal Bupleurum. Phytomedicine 21, 1767–1773. https://doi.org/10.1016/j.phymed.2014.09.001

Coghlan ML, Haile J, Houston J, Murray DC, White NE, Moolhuijzen P, Bellgard MI, Bunce M (2012) Deep sequencing of plant and animal DNA contained within traditional Chinese medicines reveals legality issues and health safety concerns. PLoS Genet. 8, e1002657. https://doi.org/10.1371/journal.pgen.1002657

Costa J, Campos B, Amaral JS, Nunes ME, Oliveira MBPP, Mafra I (2016) HRM analysis targeting ITS1 and *matK* loci as potential DNA mini-barcodes for the authentication of Hypericum perforatum and *Hypericum androsaemum* in herbal infusions. Food Control 61, 105–114. https://doi.org/10.1016/j.foodcont.2015.09.035

Cuadros-Rodríguez L, Ruiz-Samblás C, Valverde-Som L, Pérez-Castaño E, González-Casado A (2016) Chromatographic fingerprinting: an innovative approach for food "identitation" and food authentication - A tutorial. Anal. Chim. Acta 909, 9–23. https://doi.org/10.1016/j.aca.2015.12.042

de Boer HJ, Cross HB, de Wilde WJJO, Duyfjes-de Wilde BEE, Gravendeel B (2015a) Molecular phylogenetic analyses of Cucurbitaceae tribe Benincaseae urge for merging of *Pilogyne* with *Zehneria*. Phytotaxa 236, 173–183. https://doi.org/10.11646/phytotaxa.236.2.6

de Boer HJ, Ghorbani A, Manzanilla V, Raclariu A-C, Kreziou A, Ounjai S, Osathanunkul M, Gravendeel B (2017) DNA metabarcoding of orchid-derived products reveals widespread illegal orchid trade. Proc. Biol. Sci. 284, 20171182. https://doi.org/10.1098/rspb.2017.1182

de Boer HJ, Ichim MC, Newmaster SG (2015b) DNA barcoding and pharmacovigilance of herbal medicines. Drug Saf. 38, 611–620. https://doi.org/10.1007/s40264-015-0306-8

de Boer HJ, Ouarghidi A, Martin G, Abbad A, Kool A (2014) DNA barcoding reveals limited accuracy of identifications based on folk taxonomy. PLoS ONE 9, e84291. https://doi.org/10.1371/journal.pone.0084291

Ernst M, Saslis-Lagoudakis CH, Grace OM, Nilsson N, Toft Simonsen H, Horn JW, Stærk D, Rønsted N (2016) Molecular phylogenetics as a predictive tool in plant-based drug discovery in the genus *Euphorbia* L. Planta Med. 81, S1–S381. https://doi.org/10.1055/s-0036-1596164

Gao T, Yao H, Song J, Liu C, Zhu Y, Ma X, Pang X, Xu H, Chen S (2010) Identification of medicinal plants in the family Fabaceae using a potential DNA barcode ITS2. J. Ethnopharmacol. 130, 116–121. https://doi.org/10.1016/j.jep.2010.04.026

Ghorbani A, Mosaddegh M, Esmaeili S (2020) Molecular authentication of Radix Behen Albi ("Bahman Sefid") commercial products reveals widespread adulteration. Research Journal of Pharmacognosy, 7 7, 57–64.

Grazina L, Amaral JS, Mafra I (2020) Botanical origin authentication of dietary supplements by DNA-based approaches. Comp. Rev. Food Sci. Food Safety 19, 1080–1109. https://doi.org/10.1111/1541-4337.12551

Hao D, Xiao P (2020) Pharmaceutical resource discovery from traditional medicinal plants: pharmacophylogeny and pharmacophylogenomics. Chinese Herbal Medicines 12, 104–117. https://doi.org/10.1016/j.chmed.2020.03.002

Heinrich M (2014) Ethnopharmacology: quo vadis? Challenges for the future. Revista Brasileira de Farmacognosia 24, 99–102. https://doi.org/10.1016/j.bjp.2013.11.019

Howard C, Lockie-Williams C, Slater A (2020) Applied barcoding: the practicalities of DNA testing for herbals. Plants 9, 1150. https://doi.org/10.3390/plants9091150

Howes MR, Quave CL, Collemare J, Tatsis EC, Twilley D, Lulekal E, Farlow A, Li L, Cazar M, Leaman DJ, Prescott TAK, Milliken W, Martin C, De Canha MN, Lall N, Qin H, Walker BE, Vásquez-Londoño C, Allkin B, Rivers M, Nic Lughadha E (2020) Molecules from nature: reconciling biodiversity conservation and global healthcare imperatives for sustainable use of medicinal plants and fungi. Plants, People, Planet 2, 463–481. https://doi.org/10.1002/ppp3.10138

Ichim MC, Booker A (2021) Chemical authentication of botanical ingredients: a review of commercial herbal products. Front. Pharmacol. 12, 666850. https://doi.org/10.3389/fphar.2021.666850

Ichim MC, de Boer HJ (2020) A review of authenticity and authentication of commercial ginseng herbal medicines and food supplements. Front. Pharmacol. 11, 612071. https://doi.org/10.3389/fphar.2020.612071

Ichim MC, Häser A, Nick P (2020) Microscopic authentication of commercial herbal products in the globalized market: potential and limitations. Front. Pharmacol. 11, 876. https://doi.org/10.3389/fphar.2020.00876

Ichim MC (2019) The DNA-based authentication of commercial herbal products reveals their globally widespread adulteration. Front. Pharmacol. 10, 1227. https://doi.org/10.3389/fphar.2019.01227

Jia J, Xu Z, Xin T, Shi L, Song J (2017) Quality control of the traditional patent medicine yimu wan based on SMRT sequencing and DNA barcoding. Front. Plant Sci. 8, 926. https://doi.org/10.3389/fpls.2017.00926

Jogam P, Sandhya D, Shekhawat MS, Alok A, M, M, Abbagani S, Allini VR (2020) Genetic stability analysis using DNA barcoding and molecular markers and foliar micro-morphological analysis of in vitro regenerated and in vivo grown plants of *Artemisia vulgaris* L. Industrial Crops and Products 151, 112476. https://doi.org/10.1016/j.indcrop.2020.112476

Kool A, de Boer HJ, Krüger A, Rydberg A, Abbad A, Björk L, Martin G (2012) Molecular identification of commercialized medicinal plants in southern Morocco. PLoS ONE 7, e39459. https://doi.org/10.1371/journal.pone.0039459

Kreuzer M, Howard C, Adhikari B, Pendry CA, Hawkins JA (2019) Phylogenomic approaches to DNA barcoding of herbal medicines: developing clade-specific diagnostic characters for *Berberis*. Front. Plant Sci. 10, 586. https://doi.org/10.3389/fpls.2019.00586

Lee M-S, Hxiao H-J (2019) Rapid and sensitive authentication of *Polygonum multiflorum* (He-Shou-Wu) of Chinese medicinal crop using specific isothermal nucleic acid amplification. Industrial Crops and Products 129, 281–289. https://doi.org/10.1016/j.indcrop.2018.12.014

Little DP (2014) A DNA mini-barcode for land plants. Mol. Ecol. Resour. 14, 437–446. https://doi.org/10.1111/1755-0998.12194

Li L, Jiang Y, Liu Y, Niu Z, Xue Q, Liu W, Ding X (2020) The large single-copy (LSC) region functions as a highly effective and efficient molecular marker for accurate authentication of medicinal *Dendrobium* species. Acta Pharm. Sin. B 10, 1989–2001. https://doi.org/10.1016/j.apsb.2020.01.012

Manzanilla V, Kool A, Nguyen Nhat L, Nong Van H, Le Thi Thu H, de Boer HJ (2018) Phylogenomics and barcoding of *Panax*: toward the identification of ginseng species. BMC Evol. Biol. 18, 44. https://doi.org/10.1186/s12862-018-1160-y

Manzanilla V, Teixidor-Toneu I, Martin GJ, Hollingsworth PM, de Boer HJ, Kool A (2022) Using target capture to address conservation challenges: population-level tracking of a globally-traded herbal medicine. Mol. Ecol. Resour. 22, 212–224. https://doi.org/10.1111/1755-0998.13472

Nortier JL, Martinez MC, Schmeiser HH, Arlt VM, Bieler CA, Petein M, Depierreux MF, De Pauw L, Abramowicz D, Vereerstraeten P, Vanherweghem JL (2000) Urothelial carcinoma associated with the use of a Chinese herb (*Aristolochia fangchi*). N. Engl. J. Med. 342, 1686–1692. https://doi.org/10.1056/NEJM200006083422301

Osathanunkul M, Madesis P, de Boer H (2015) Bar-HRM for authentication of plant-based medicines: evaluation of three medicinal products derived from Acanthaceae species. PLoS ONE 10, e0128476. https://doi.org/10.1371/journal.pone.0128476

Osathanunkul M, Suwannapoom C, Osathanunkul K, Madesis P, de Boer H (2016) Evaluation of DNA barcoding coupled high resolution melting for discrimination of closely related species in phytopharmaceuticals. Phytomedicine 23, 156–165. https://doi.org/10.1016/j.phymed.2015.11.018

Palhares RM, Gonçalves Drummond M, Dos Santos Alves Figueiredo Brasil B, Pereira Cosenza G, das Graças Lins Brandão M, Oliveira G (2015) Medicinal plants recommended by the world health organization: DNA barcode identification associated with chemical analyses guarantees their quality. PLoS ONE 10, e0127866. https://doi.org/10.1371/journal.pone.0127866

Posadzki P, Watson L, Ernst E (2013) Contamination and adulteration of herbal medicinal products (HMPs): an overview of systematic reviews. Eur. J. Clin. Pharmacol. 69, 295–307. https://doi.org/10.1007/s00228-012-1353-z

Posthouwer C, Veldman S, Abihudi S, Otieno JN, van Andel TR, de Boer HJ (2018) Quantitative market survey of non-woody plants sold at Kariakoo Market in Dar es Salaam, Tanzania. Journal of ethnopharmacology 222, 280–287.

Raclariu AC, Mocan A, Popa MO, Vlase L, Ichim MC, Crisan G, Brysting AK, de Boer H (2017a) *Veronica officinalis* product authentication using DNA metabarcoding and HPLC-MS reveals widespread adulteration with *Veronica chamaedrys*. Front. Pharmacol. 8, 378. https://doi.org/10.3389/fphar.2017.00378

Raclariu AC, Paltinean R, Vlase L, Labarre A, Manzanilla V, Ichim MC, Crisan G, Brysting AK, de Boer H (2017b) Comparative authentication of *Hypericum perforatum* herbal products using DNA metabarcoding, TLC and HPLC-MS. Sci. Rep. 7, 1291. https://doi.org/10.1038/s41598-017-01389-w

Raclariu AC, Ţebrencu CE, Ichim MC, Ciupercă OT, Brysting AK, de Boer H (2018) What's in the box? Authentication of *Echinacea* herbal products using DNA metabarcoding and HPTLC. Phytomedicine 44, 32–38. https://doi.org/10.1016/j.phymed.2018.03.058

Rungpragayphan S, Pamonsinlapatham P, Powthongchin B, Prommanee W, Wongakson P (2014) Exploring DNA barcode information of selected thai medicinal plants. Adv. Mat. Res. 1060, 219–222. https://doi.org/10.4028/www.scientific.net/AMR.1060.219

Saslis-Lagoudakis CH, Savolainen V, Williamson EM, Forest F, Wagstaff SJ, Baral SR, Watson MF, Pendry CA, Hawkins JA (2012) Phylogenies reveal predictive power of traditional medicine in bioprospecting. Proc Natl Acad Sci USA 109, 15835–15840. https://doi.org/10.1073/pnas.1202242109

Savitikadi P, Jogam P, Rohela GK, Ellendula R, Sandhya D, Allini VR, Abbagani S (2020) Direct regeneration and genetic fidelity analysis of regenerated plants of *Andrographis echioides* (L.) - An important medicinal plant. Industrial Crops and Products 155, 112766. https://doi.org/10.1016/j.indcrop.2020.112766

Schmidt B, Ribnicky DM, Poulev A, Logendra S, Cefalu WT, Raskin I (2008) A natural history of botanical therapeutics. Metab. Clin. Exp. 57, S3-9. https://doi.org/10.1016/j.metabol.2008.03.001

Seethapathy GS, Raclariu-Manolica A-C, Anmarkrud JA, Wangensteen H, de Boer HJ (2019) DNA metabarcoding authentication of ayurvedic herbal products on the European market raises concerns of quality and fidelity. Front. Plant Sci. 10, 68. https://doi.org/10.3389/fpls.2019.00068

Selvaraj D, Shanmughanandhan D, Sarma RK, Joseph JC, Srinivasan RV, Ramalingam S (2012) DNA barcode ITS effectively distinguishes the medicinal plant *Boerhavia diffusa* from its adulterants. Genomics Proteomics Bioinformatics 10, 364–367. https://doi.org/10.1016/j.gpb.2012.03.002

Sheidai M, Tabaripour R, Talebi SM, Noormohammadi Z, Koohdar F (2019) Adulteration in medicinally important plant species of *Ziziphora* in Iran market: DNA barcoding approach. Industrial Crops and Products 130, 627–633. https://doi.org/10.1016/j.indcrop.2019.01.025

Shen Z, Lu T, Zhang Z, Cai C, Yang J, Tian B (2019) Authentication of traditional Chinese medicinal herb "Gusuibu" by DNA-based molecular methods. Industrial Crops and Products 141, 111756. https://doi.org/10.1016/j.indcrop.2019.111756

Taberlet P, Coissac E, Pompanon F, Gielly L, Miquel C, Valentini A, Vermat T, Corthier G, Brochmann C, Willerslev E (2007) Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding. Nucleic Acids Res. 35, e14. https://doi.org/10.1093/nar/gkl938

Tnah LH, Lee SL, Tan AL, Lee CT, Ng KKS, Ng CH, Nurul Farhanah Z (2019) DNA barcode database of common herbal plants in the tropics: a resource for herbal product authentication. Food Control 95, 318–326. https://doi.org/10.1016/j.foodcont.2018.08.022

Urumarudappa SKJ, Gogna N, Newmaster SG, Venkatarangaiah K, Subramanyam R, Saroja SG, Gudasalamani R, Dorai K, Ramanan US (2016) DNA barcoding and NMR spectroscopy-based assessment of species adultera-

tion in the raw herbal trade of *Saraca asoca* (Roxb.) Willd, an important medicinal plant. Int. J. Legal Med. 130, 1457–1470. https://doi.org/10.1007/s00414-016-1436-y

Vassou SL, Kusuma G, Parani M (2015) DNA barcoding for species identification from dried and powdered plant parts: a case study with authentication of the raw drug market samples of *Sida cordifolia*. Gene 559, 86–93. https://doi.org/10.1016/j.gene.2015.01.025

Vassou SL, Nithaniyal S, Raju B, Parani M (2016) Creation of reference DNA barcode library and authentication of medicinal plant raw drugs used in Ayurvedic medicine. BMC Complement. Altern. Med. 16 Suppl 1, 186. https://doi.org/10.1186/s12906-016-1086-0

Wang M, Zhao H-X, Wang L, Wang T, Yang R-W, Wang X-L, Zhou Y-H, Ding C-B, Zhang L (2013) Potential use of DNA barcoding for the identification of *Salvia* based on cpDNA and nrDNA sequences. Gene 528, 206–215. https://doi.org/10.1016/j.gene.2013.07.009

Williamson J, Maurin O, Shiba SNS, van der Bank H, Pfab M, Pilusa M, Kabongo RM, van der Bank M (2016) Exposing the illegal trade in cycad species (Cycadophyta: *Encephalartos*) at two traditional medicine markets in South Africa using DNA barcoding. Genome 59, 771–781. https://doi.org/10.1139/gen-2016-0032

Xie L, Wang YW, Guan SY, Xie LJ, Long X, Sun CY (2014) Prospects and problems for identification of poisonous plants in China using DNA barcodes. Biomed. Environ. Sci. 27, 794–806. https://doi.org/10.3967/bes2014.115

Zhao W, Liu M, Shen C, Liu H, Zhang Z, Dai W, Liu X, Liu J (2020) Differentiation, chemical profiles and quality evaluation of five medicinal *Stephania* species (Menispermaceae) through integrated DNA barcoding, HPLC-QTOF-MS/MS and UHPLC-DAD. Fitoterapia 141, 104453. https://doi.org/10.1016/j.fitote.2019.104453

Answers

1. Categories that could be mentioned include: unconscious misidentification by collectors, intentional fraudulent substitution, discrepancies between vernacular names and scientific species names, high market value of medicines incentivise adulteration, and lack of regulation in some countries.
2. Genetic methods in plant identification can be used in combination with chemical analytical methods to identify plants used as medicines in traditional health systems during ethnobotanical or ethnopharmacological research since chemically based methods alone often cannot correctly identify a plant species or its origin.
3. Molecular plant identification, potentially in combination with chemical analytical methods, can be used to systematically identify plants with potential medicinal properties.

Chapter 23
Food safety

Bastien Anthoons[1,2], Panagiotis Madesis[3]

1 School of Biology, Aristotle University of Thessaloniki, Thessaloniki, Greece
2 Institute for Applied Biosciences, Centre for Research and Technology, Thessaloniki, Greece
3 Lab of Molecular Biology, Department of Agriculture Crop Production and Rural Environment, University of Thessaly, Volos, Greece

Bastien Anthoons **bastien.anthoons@gmail.com**
Panagiotis Madesis **pmadesis@uth.gr**

Citation: Anthoons B,Madesis P (2022) Chapter 23. Food safety. In: de Boer H, Rydmark MO, Verstraete B, Gravendeel B (Eds) Molecular identification of plants: from sequence to species. Advanced Books. https://doi.org/10.3897/ab.e98875

Introduction

Food safety is defined as the routines used in food handling, preparation, and storage to reduce the risk of individuals becoming sick from foodborne illnesses. Food safety draws from the expertise in a wide range of academic fields, including chemistry, microbiology, molecular biology, and engineering. Although advances in science and technology have led to a substantial improvement in food quality, food can still be a source for public health issues (Borchers et al. 2010). From farm to factory to fork, food products may encounter various hazards during their journey through the supply chain. Harmful microorganisms may accidently enter the food chain due to bad hygiene practices or irregularities in production steps. Food products might get contaminated with allergens such as peanuts, tree nuts, or wheat. Unauthorised genetically modified organisms (GMOs), if not properly regulated, can transfer allergenic genes to other organisms. Access to sufficient amounts of safe and nutritious food is key to sustaining life and promoting good health. Unsafe food containing harmful bacteria, viruses, contaminants, adulterants, or chemical substances results in more than 200 different adverse reactions, from cancer to acute intoxication (World Health Organization 2020). Around the world, an estimated 600 million people, approximately 1 in 10, develop severe symptoms of illness after eating unsafe food. This results in 420,000 deaths and the loss of 33 million healthy life years (Fung et al. 2018; World Health Organization 2020). In Europe, more than 23 million people per year become ill from foodborne contamination. This means that approximately 44 people per minute are affected by issues with food safety in Europe (EIT Food 2019). The globalisation of food trade, a growing world population, climate change, and rapidly changing food systems all impact food safety (World Health Organization 2020). Plants are the major source of human food worldwide, including various plant grains, vegetables, and fruits. Plants are also used as spices and in traditional medicines, where their widespread use has been documented through the ages. This chapter focuses on plant-based food products and their regulation within the global supply chain.

Chapter 23: Box 1. Dietary supplement or herbal medicine?

Numerous botanical products with considerable differences in their classification can be bought throughout the world. These include foodstuffs, herbal medicinal products, and cosmetics. Foodstuffs include dietary supplements, food ingredients, functional foods, and foods for particular nutritional use including various botanical extracts. Herbal medicinal products can only be sold in pharmacies, under the supervision of a pharmacist, and are marketed after registration procedures according to their classification (see Chapter 22 Healthcare). Dietary supplements and herbal medicines are usually considered as two different regulatory categories, but for each of them, the consensus for regulation is also lacking across countries (Low et al. 2017; Thakkar et al. 2020). For botanical foodstuffs national legislations may foresee: (1) positive lists of botanicals which may be used, (2) negative lists of botanicals which may not be used, (3) restrictions/modalities for their use (e.g., maximum limits, labelling requirements) (European Commision 2006). Moreover, the distinction between medicinal products and food supplements has generated borderline botanical-sourced products in many cases, which generally causes confusion among consumers (Bilia 2015).

Food hazards and impacts

Food safety hazards

Food hazards refer to any agents with the potential to cause adverse health consequences for consumers. Food safety hazards occur when food is exposed to and contaminated by hazardous agents. Food hazards may be biological, chemical, physical, allergenic, nutritional, and/or biotechnology-related (Bouxin 2014). Food hazards enter the supply chain in various ways and can pose threats to human health, the economy, or biodiversity (Bouxin 2014). Foodborne illnesses can be caused by a wide range of organisms, including bacteria, viruses, parasitic worms, and plants. Plants may fit in all food hazard categories based on their taxon-specific properties. Poisonous plants are considered biological food safety hazards since they are considered living organisms that can accidentally end up in a food product (Anthoons et al. 2020). However, it is plant phytotoxins that result in harm if ingested. Plant toxins are naturally produced secondary metabolites that protect the plant from natural threats. The main groups of plant toxins are alkaloids, terpenes, glycosides, proteinaceous compounds, organic acids, and resinoid compounds (Speijers and van Egmond 1999). Poisonous plants are therefore also considered chemical hazards. It is worth emphasising that the plant's chemistry varies greatly and that many plant species that produce secondary compounds contain individuals with distinct chemical phenotypes, called chemotypes (Keefover-Ring et al. 2009). *Thymus vulgaris* provides an example of such a chemically polymorphic species, with 6 different chemotypes described within the species (Keefover-Ring et al. 2009) Physical hazards or extraneous material covers all (non-toxic) material associated with unsanitary production conditions, processing, handling, storage, and distribution of food. Plant-based materials may include leaves and wood chips (Edwards 2014). Plant-derived allergens may either be present in a wide range of plant food products causing food allergic reactions but also as pollen, taken in from the respiratory tract (Hoffmann-Sommergruber 2000). Plant allergens commonly possess a defence-related function against invading pathogens. Various plant foods including peanuts, tree nuts, seeds, celery, fresh fruits, and legumes may trigger severe anaphylactic reactions when ingested by allergic patients (Shahali and Dadar 2018).

A final category of food hazards are biotechnology-related hazards such as genetically modified organisms (GMOs). GMOs are the products of genetic engineering where new genes are transferred from one species into another. The resulting properties may lead to better optimised agricultural performance or the new or increased production of valuable pharmaceutical substances (Hefferon 2015). Plants and crops are often engineered to provide higher yields and more resistance to herbicides or specific pests. Although the benefits of GMOs are well accepted, it is important to note that some risks are associated with them that are not always fully understood (Bawa and Anilakumar 2013). These potential risks include adverse effects on biodiversity and potentially reduced nutritional quality (Gatew and Mengistu 2019). The introduction of a modified organism into the environment may cause the displacement of indigenous fauna and flora and impact the entire food chain and the predator-prey relationships (Wong and Candolin 2015). The genetically modified food may cause a hazard in developing allergenicity, transfer of genes from GM food to cells of the body or to bacteria in the gastrointestinal tract (Bakshi 2003; Keese 2008). The genetic-engineering process may cause "unnatural" changes in a plant's own metabolic pathways and proteome and result in an unexpected production of toxins or allergens in food (Fagan et al. 2014).

It is also important to point out that the toxicity of any substance, including plant-based food and medicinal plants, is largely dependent on the dose or amount used. A (harmless) plant may be toxic at high doses, and a highly toxic plant could be considered safe at low

dose (Deshpande 2002). For instance, the Chinese medicine Radix Bupleuri (*Bupleurum chinense*) becomes toxic when the dose ingested exceeds 21 times the common clinical dose of 9 g/60 kg used in commercialised products (Lv et al. 2009).

Food fraud

Food fraud is a collective term used to encompass the intentional substitution, addition, tampering, or misrepresentation of food, food ingredients, or food packaging with the aim of increased economic gain (Spink and Moyer 2011). It is an issue that affects all food supply chains and thus the entire food industry including consumers. The impact of food fraud is a real public health vulnerability (Spink and Moyer 2011). However, the European Union has not set a legal definition for what food fraud is and this has led to misconceptions amongst both researchers and the food industry (Wisniewski and Buschulte 2019). For example, a food quality risk is an economic threat and can be intentional, and thus considered food fraud, or unintentional, such as proliferation of fungi on corn, which would not necessarily be considered fraudulent.

Adulteration

Adulteration is the failure of a product to meet legal quality standards. According to the US Federal Food, Drug and Cosmetic Act (FFDCA), food can be declared adulterated if (1) a substance is added which is injurious to health, (2) a cheaper or inferior quality item is added to the food, (3) any valuable constituent is extracted from the main food article, (4) the quality of food is below the standards, (5) a substance is added to increase bulk or weight, and (6) a substance is added to make it appear more valuable. Adulterated food can be dangerous since it may be toxic to human or animal health, it may lead to the deprivation of nutrients required for health, and it may cause intoxication or allergic reactions in sensitised individuals. Adulterants in food can be categorised as follows (Bansal et al. 2017).

Intentional adulteration is the inclusion of inferior substances having properties similar to the foods to which they are added. The adulterant can be physical, chemical, or biological. An example of intentional adulteration is the addition of wheat or other grains as an inexpensive filler to increase profit margins (Spink and Moyer 2011).

Unintentional adulteration is the inclusion of unwanted substances due to ignorance, carelessness, or lack of proper facilities and hygiene during food processing. This includes contamination of foods by bacteria and fungi, or harmful residues from packing material, or even inherent adulteration including the presence of certain chemicals, organic compounds, or radicals that naturally occur in foods such as toxic varieties of plants, mushrooms, etc.

Metallic contamination is the intentional or unintentional inclusion of different types of metals and metal compounds in food. Arsenic, cadmium, lead and mercury are amongst the most toxic ones.

Microbial contamination is the spoilage of food due to infusion of different microbes through various sources.

Agrobioterrorism

Agrobioterrorism can be defined as the use of pathogens or toxins against agricultural products or facilities usually with the purpose of causing casualties or fatalities from contaminated

agricultural resources or food (Roberge 2019). Acts of bioterrorism on the food supply can include biological, physical, chemical, or even radiological agents. Many potential agents are highly toxic and are not prevented or inactivated by conventional food safety interventions. Most of these potential agents are difficult to detect, or at least difficult to detect when in a variety of foods (Mitenius et al. 2014). Agrobioterrorist attacks can be directed at many different targets in the food supply chain, including crops, livestock, food products in the processing and distribution chain, wholesale and retail facilities, storage facilities, transportation, and food and agricultural laboratories (Dyckman 2003). The most common agents used to destroy crops are plant pathogens (Suffert et al. 2009). While biological warfare programs stopped after most countries signed the Biological and Toxin Weapons Convention (BTWC) in 1972, new concerns over the possible use of biological anti-crop weapons arose in the late 1980s. For instance, Iraq, after the First Gulf War, began developing weaponized strains of the fungi *Tilletia caries* and *T. tritici* (wheat smut) and aflatoxin-producing *Aspergillus* in order to destroy Iranian crops (Whitby 2002).

Chapter 23: Box 2. Poisonous plants: incidence of misidentification

Toxic plants are occasionally eaten due to their misidentification. The fruits of toxic plants such as the very poisonous deadly nightshade (*Atropa belladonna*) appears similar to edible fruits such as blueberries (*Vaccinium* sp.) or black nightshade (*Solanum nigrum*) (Colombo et al. 2010), and the very toxic poisonous hemlock (*Conium maculatum*) resembles wild species of carrot (*Daucus carota*) and chervil (*Anthriscus cerefolium*) (Biberci et al. 2002; Vetter 2004). Some common bulbs are mildly toxic such as daffodils (*Narcissus* sp.), tulip (*Tulipa* sp.), and autumn crocus (*Colchicum autumnale*), and can be mistaken for onions (*Allium* sp.) (Klintschar et al. 1999). Poisonous plants can be consumed raw or after processing by drying or cooking. These activities affect the toxicity of most poisonous plants. Most toxins can be eliminated by cooking, but some are concentrated by drying or extraction into a tea-like beverage. In recent years, natural herbal remedies have become more popular, but can also lead to poisoning. This is most often caused by misidentification or unawareness (or even disbelief) of any potential toxins and their necessary treatments to reduce their dangerous properties, and overdoses. Harmful toxins can be transferred from plants to human foods in animal products such as milk, bird eggs, and honey produced by bees foraging on toxic plants. Poisonous substances can be present in some or all parts of a plant including the roots, leaves, fruits, and seeds, and the toxins can be dangerous either by oral ingestion, inhalation, or skin contact (Bruneton 2000; Schilter et al. 2014).

Authentication of plant-based food stuffs

Food authentication is the process by which food is verified as complying with its label description. According to Gizaw (2019) food adulteration and mislabelling are amongst the most important public health risks associated with food safety on the food market. Food ingredient authentication enables the manufacturer to detect adulterations so that the consumer receives a product that matches the written product specifications, is free of contaminants, and is safe to consume (Gizaw 2019); see Chapter 22 Healthcare). In Europe, detected food adulterations are judged as a violation of EU food law (EC General Food Law Regulation 178/2002) (Euro-

pean Commission 2002), which includes specific matters including GMOs, allergens, and food hygiene (European Commission 2003).

Plant-based oils and fats dominate food applications. A balanced intake of oils and fatty acids are essential for human health (van Duijn 2014). Extra virgin olive oil is a high-priced product with high nutritional value. Due to the high market prices and its increasing demand, olive oil is one of the most adulterated products on the global food market. Usually olive oil is substituted with less expensive edible vegetable oils (Ganopoulos et al. 2013; Song et al. 2020). In some cases, this could lead to serious health problems such as the Spanish toxic oil syndrome or Spanish olive oil syndrome due to substitution of non-edible rapeseed oil (*Brassica napus* subsp. *napus)* for edible rapeseed oil or even olive oil (Azadmard-Damirchi and Torbati 2015; Clemente and Cahoon 2009). The low availability of high-quality extra virgin olive oil and inadequate screening from regulatory agencies are primary reasons why adulteration of this product remains prevalent (Mailer and Gafner 2020).

The supply chains for herbs and spices tend to be long and complex and pass through many countries. These complexities and the increase in crushed and ground herbs and spices render those products more prone to intentional adulteration (Galvin-King et al. 2018). Saffron, oregano, vanilla, turmeric, and paprika are some of the most adulterated spices and herbs. Substitution with cheap fillers such as wheat (Mishra et al. 2016) and the addition of colouring agents to spices are common practice (Galvin-King et al. 2018).

The adulteration of dietary supplements has been reported fairly frequently, as a result of their rising popularity. Fraudulent practices may result in reduced therapeutic potential of the original drug, posing a serious risk to the health of the consumers (Newmaster et al. 2013); see Chapter 22 Healthcare). Coffee is one of the most used food products globally and of great economic importance in countries involved in its production and export. Common additives in adulterated coffee include chicory, coffee stems instead of beans, and soybeans. Additionally, the coffee bean species and growing location is often misrepresented (Toci et al. 2018).

From farm to fork - traceability

Food hazards may enter the food chain in various ways and have large impacts on human health. It is therefore important that products "moving" along the food supply chain (FSC) are both tracked and traced. Traceability, under EU law, means the ability to track any food, feed, food-producing animal, or substance that will be used for consumption, through all stages of production, processing, and distribution. Traceability applies to both upstream (where the product comes from) as downstream (where the product is delivered to) tracking (Overbosch and Blanchard 2014). Tracking and tracing involve important decisions in the value chain to improve processing organisation and risk management, as well as a good level of buyer-supplier strategy (Rábade and Alfaro 2006). Increased global food trade has led to an increase in imports and exports and the need for joint efforts to apply traceability strategies at the international level.

This issue was debated in the UN's joint Food and Agriculture Organization (FAO) and World Health Organization (WHO), leading to the Codex Alimentarius or "Food Code", a collection of standards, guidelines and codes of practice adopted by the Codex Alimentarius Commission (Dabbene et al. 2014). This commission recommends the HACCP method (Hazard Analysis Critical Control Points) for inspection of finished food products but also

to find, correct, and prevent hazards throughout the production process (Overbosch and Blanchard 2014). Seven universally accepted principles define the HACCP methodology: hazard analysis, identifying critical control points, establishing critical limits for each critical control point, establishing monitoring procedures for critical control points, establishing corrective actions and verification procedures, and record keeping (U.S. Food and Drug Administration 1997). The critical control points in the production process are those steps where an action should be taken to prevent, eliminate, or reduce a food safety hazard to an acceptable level.

In Europe, risks along the supply chain are assessed by the European Food Safety Authority (EFSA). EFSA monitors and analyses information and data on biological hazards, chemical contaminants, food consumption, and emerging risks (European Food Safety Authority (EFSA), 2012). It is important to note that the principles of the universal HACCP method depend on the origin and nature of the food products as well as the type of end-product. Hazards and their subsequent risk assessment will therefore differ between the olive oil supply chain and the herbal tea supply chain. Olive oil production involves specific processing steps and uses industrial settings such as extraction mills (for pressing or centrifugation), which are absent in the processing phase of dry plant material such as herbal teas and spices (Vlachos and Malindretos 2012).

Fraudulent practices can happen at any step of the supply chain. The most effective way to eliminate illegal practices in the food sector is food chain transparency and full raw material traceability. For example, food companies that implement a digital traceability system using unique product identifiers increase their transparency since they have supply chain visibility in real-time (Tayal et al. 2020).

In the following example, risk assessment in a chain of commercialization of plant-based products based on dry plant material (e.g., herbal tea, spices, medicinal mixtures) is discussed.

Plant cultivation

Plant cultivation is the first step in the supply chain for a herbal product, from seed(ling) to adult plant. During growing periods in agricultural fields or greenhouses, different hazardous sources may affect downstream processing and production. These hazards can include: faeces, contaminated soil, irrigation water, water used to apply pesticides, foliar treatments, growth hormones, dust, wild and domestic animals, insects, and human handling. Automated and regular monitoring as well as personal hygiene are therefore essential.

Harvesting

Harvesting can be performed by hand or mechanically, and involves several important commercial steps including pre-sorting and removal of foliage and other non-edible parts. Personal hygiene is particularly important during manual harvesting. Contamination of the herbal product with other plants such as weeds can result from insufficient quality control during harvesting (Speranskaya et al. 2018). Some common undesirable plants that are detected in herbal products are *Convolvulus arvensis* (bindweed), *Urtica dioica* (nettle), and *Triticum aestivum* (wheat) (Anthoons et al. 2020). Additionally, the timing and environmental conditions during harvesting and storage can reduce the growth of microorganisms and plant pathogens (Brackett 1992). After a plant has been cultivated/collected, it should be authenticated at the species or vari-

ety-level to assure product quality. This is common practice in the herbal medicinal industry and should be applied to plants used in food products (such as herbal teas or spices) (Govindaraghavan and Sucher 2015) (see Chapter 22 Healthcare).

Authentication

Authentication or verification of raw plant material can be done by traditional morphological analysis or by DNA-based methodologies (see paragraph "Methodologies for identification of plant food hazards"). In the case of products with a protected designation of origin (PDO), the label originates from a certain region or area and the product quality and/or characteristics are due to the particular geographical environment, e.g., Greek extra virgin olive oil or PDO saffron (Bosmali et al. 2017; Ganopoulos et al. 2013). PDO product verification is usually performed after processing and before commercialisation at the end of the supply chain. After harvesting, the plant parts will be processed at an industrial facility following transportation.

Transportation and processing

During transportation, the raw plant material might be damaged due to poor handling, cross-contamination with other materials in the vehicle, or contaminated with vehicle exhaust from petrol and diesel (Nerín et al. 2016). Processing of raw plant material for the production of herbal teas and medicinal mixtures includes classifying, sorting, and drying. The most common risks during processing include cross-contamination and unstable environmental conditions. Also poor handling can damage fresh material, rendering the product susceptible to the growth of microorganisms that cause spoilage or are pathogenic (Francis et al. 1999).

Storage

The plant material might be stored for an extended period of time before packing. Storage requirements depend on the state of the plant-based product (i.e., fresh, raw, processed). Raw plant material needs to be stored in a cool and dry place since fungi can grow if the humidity is too high (Piližota 2014).

Packing and packaging

The purpose of packing is to protect against food pathogens, spoilage-causing organisms, pests, damage, etc. Good hygiene practices should be followed in handling containers and improved packing materials to prevent product contamination (Piližota 2014). Incorrect labelling, either related to the mandatory information (e.g., allergen not listed in the ingredient list despite being added deliberately) or to precautions, is the most common issue in food safety in the food market in developed countries (Gizaw 2019). Not only can mislabelling result in adverse effects on human health, it can also lead to cross-contamination, poor food quality, degradation of nutrients, and thus has serious financial and legal consequences (Armani et al. 2015).

Methodological approaches for the detection of plant food hazards

Analytical methods for detecting adulterated food are traditionally seen as a first line of defence against food fraud (Ulberth 2020). It is important to note that the effectiveness of targeted regulatory analyses in case of a food safety incident is highly dependent on the quality and performance of the laboratory methods needed to support regulatory compliance, investigations and enforcement actions. Evaluation and validation of the methods employed for food analysis are therefore necessary to ensure that they meet the highest analytical performance standards appropriate for their intended purposes (U.S. Food and Drug Administration 2019). Analytical techniques used to identify food fraud involve sensory, physicochemical, DNA-based, chromatographic, and spectroscopic methods (Fritsche 2018). The most commonly detected food hazards are of chemical origin, and chemical hazard analysis is therefore highly desirable in food safety and integrity fields to ensure consumer health. Naturally occurring plant toxins can readily be detected with methods based on liquid chromatography coupled to mass spectrometry (LC-MS) (Picardo et al. 2019). Mycotoxins, produced by fungi including the genera *Fusarium*, *Aspergillus*, and *Penicillium,* under certain temperature and humidity conditions are commonly detected and quantified in food using chromatographic techniques and antibody-based assays (Kharayat and Singh 2018; Picardo et al. 2019). DNA-based technologies are used for identification of the specific biological contaminants, i.e., the specific fungus or plant.

Detection of food-microbial contamination

Several techniques are used in the food industry to detect food-microbial contamination. Omics-based techniques (i.e., genomics, transcriptomics, proteomics, and metabolomics; see Chapter 12 Metagenomics and Chapter 15 Transcriptomics) are robust tools to gain insight into microbial communities along the food chain and can detect pathogens, the origin of a foodborne illness, microbial source tracking investigations, and antimicrobial resistance (Cook and Nightingale 2018). Near-infrared (NIR) spectroscopy can also distinguish infected from non-infected food products, as well as to perform qualitative and quantitative determination of available bacteria, or as an indicator of freshness or spoilage (Atanassova et al. 2018). Other techniques for microbial detection in food include multiple-locus variable-number of tandem-repeats analysis (MLVA), and biosensors (Starodub et al. 2018).

Molecular methods for plant product authentication

Several common molecular techniques for plant-based food authentication are available.

PCR-based techniques are useful for the detection and identification of animal and plant species in foods because of their high sensitivity and specificity, in addition to being relatively fast and inexpensive. Multiplex PCR assays simultaneously identify several species by using species-specific primers, and they are being extensively applied to the detection and differentiation of species present in food products (Fairchild et al. 2006). qPCR assays estimate rather than exactly quantify the contents and ratios of different animal or plant species, for example in fruit juices containing different ratios of mandarin and orange juices in the samples. qPCR-based approaches are also commonly applied in food authentication processes due to their

high sensitivity and specificity (Aldeguer et al. 2014). GMOs are routinely detected and quantified using microarrays or qPCR assays based on the screening of genetic elements like p35S, tNos, pat, or bar or event specific markers for official GMOs like Mon810, Bt11, or GT73 (Bawa and Anilakumar 2013).

Like PCR and qPCR, Loop-Mediated Isothermal Amplification (LAMP) detects specific DNA sequences, but can target up to eight different sequences. The LAMP method uses self-recurring strand-displacement DNA synthesis to replicate a target DNA at a constant temperature and avoids any PCR amplification steps, saving time and avoiding PCR bias. LAMP has been applied for the detection of foodborne pathogens, the screening of pesticide residues, the assessment of adulterations in meat and various food allergens as well as the authentication of GM crops (Huang et al. 2020).

High resolution melting (HRM) is a post-PCR analysis method that monitors the rate of double stranded DNA dissociation to single stranded DNA with increasing temperature and is used to identify variations in nucleic acid sequences. HRM, especially in combination with DNA barcoding, has proven successful for species discrimination, adulterant and allergen detection and product authentication on a wide range of complex food materials of plant as well as animal origin (Ganopoulos et al. 2013, 2012; Bosmali et al. 2017; Lagiotis et al. 2020; Anthoons et al. 2022); see Chapter 13 DNA Barcoding - High Resolution Melting).

Next generation sequencing (NGS) combined with powerful bioinformatics tools are advancing food microbiology and authentication of products of botanical origin (Ivanova et al. 2016; Jagadeesan et al. 2019). DNA barcoding, consisting of sequencing and comparing orthologous DNA regions for taxonomic identification, has been proposed as a standardised method for species (and taxa) authentication. DNA barcoding, either by Sanger sequencing or HTS technologies, is important in both food species identification and traceability (Galimberti et al. 2013); see Chapter 10 DNA barcoding). High-throughput sequencing and DNA barcoding can be combined for DNA metabarcoding studies, which enables simultaneous multi-taxa identification by using DNA extracted from complex samples with DNA of different origins (Staats et al. 2016). DNA metabarcoding has been used to assess the composition of multi-ingredient marketed herbal products (Raclariu et al. 2018, 2017; Seethapathy et al. 2019). Authentication studies using DNA metabarcoding investigate the level of discrepancy between the expected and detected plant species based on the label claims of the products (Anthoons et al. 2020); see Chapter 11 Amplicon metabarcoding).

More advanced molecular methods such as shotgun metagenomic and whole genome sequencing are becoming more widely adopted in the food industry. These approaches provide deeper information in one single analysis and provide more complete sequence information (Ripp et al. 2014). A bioinformatics pipeline for food authentication from shotgun sequencing data (FASER) has been developed by Haiminen et al. (2019). FASER uses a comprehensive database with > 6000 plant and animal sequences and is able to identify and quantify food matrix components of mixed plant origin using short sequencing reads (100s to 1000s range). This is the first step towards implementing shotgun sequencing in standardised food safety testing procedures (Haiminen et al. 2019; Zhang et al. 2017).

Conclusions and future prospects

There is an urgent need to combat food safety issues in plant-based products. Further methodological improvements in food hazard detection and digitalization of food safety protocols are necessary for quality assurance of food products. Mislabelling and fraudulent practices such as

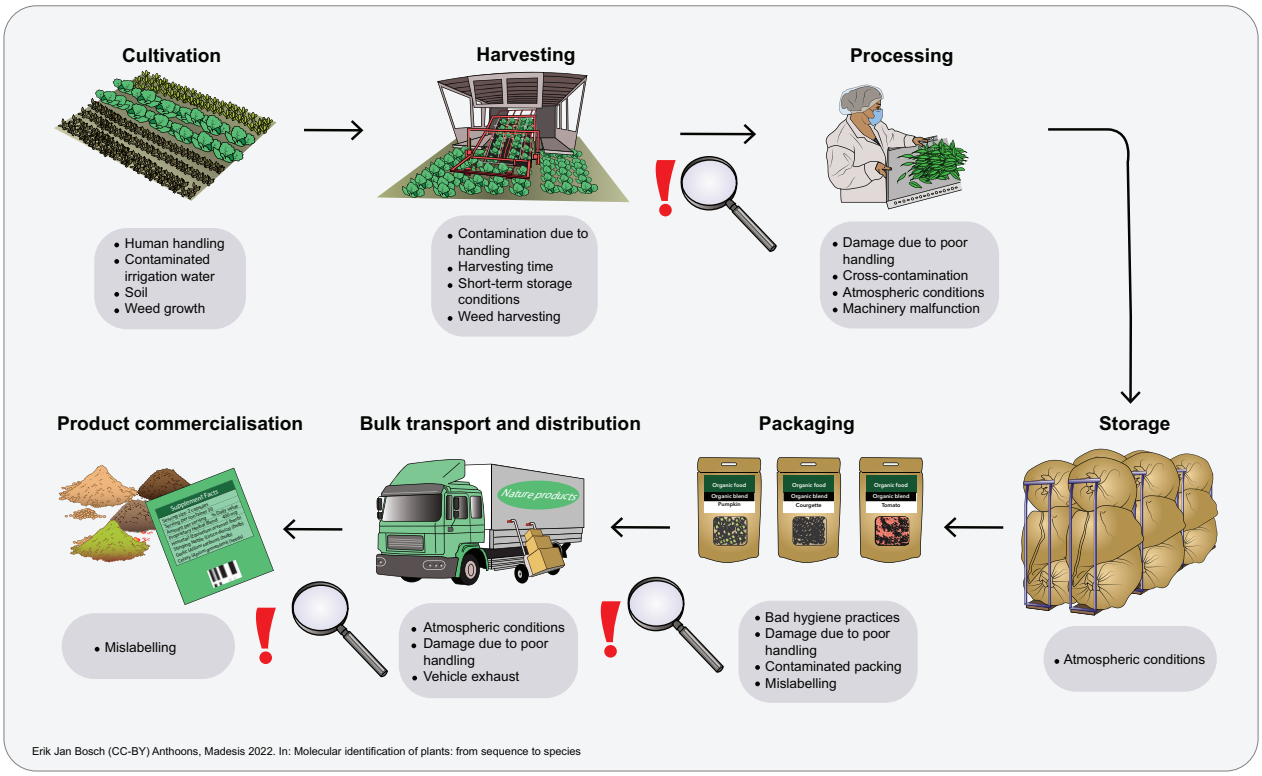

Figure 1. Chapter 23 Infographic: From farm to fork: risks along the herbal product supply chain. An example of potential risks and their sources in the supply chain chain for commercially-based dried plant products (e.g., herbal tea, spices, medicinal mixtures). The most common food safety risks associated with every step in the supply chain are highlighted in boxes, and the critical risk assessment steps are highlighted in the process (indicated by a magnifying glass and an exclamation mark).

adulteration require special attention as they are the most common issues in the global food supply chain. Most food hazard detection techniques are chemistry-based and used for detecting chemical food hazards or focus on microbial contamination, issues which have important repercussions on human health. DNA based methodological advances for plant-based foods should focus more on the creation and curation of reference databases and the use of innovative bioinformatics tools for fast and accurate food authentication. Standardisation of DNA based methodologies is a prerequisite for the successful implementation of hazard risk assessment protocols at the national and international level.

Questions

1. A. Food analysts found a considerable amount of wheat (*Triticum aestivum*) in several spice mixtures on a local market. The wheat is not mentioned on the labels of the products. How would you define this food safety issue? B. How could undesirable traces of wheat enter the supply chain? Provide at least two possible causes. C. What are the risks associated with this food safety issue?
2. What is the food safety risk of undeclared GMOs in food products?
3. Extra virgin olive oil is one of the most adulterated plant foods, either intentionally or unintentionally. State three reasons why this product is more prone to end up in an adulterated state compared to other products.

Glossary

Adulteration – A food product that fails to meet the legal standards set by the government is said to have been adulterated. Food adulteration is a legal offence and occurs when substances that lower the quality of food are present, either intentionally or unintentionally.

Bioterrorism – Bioterrorism is defined as a release of biological agents or toxins that affect humans, animals, or plants with the intent to harm or intimidate.

Codex Alimentarius – Also known as "Food Code", this is a collection of standards, guidelines and codes of practice adopted by the Codex Alimentarius Commission. The Commission, also known as CAC, is the central part of the Joint FAO/WHO Food Standards Programme and was established by the FAO and WHO to protect consumer health and promote fair practices in food trade.

EFSA – European Food Safety Authority. EFSA provides independent scientific advice on food-related risks. This advice informs European laws, rules, and policymaking, and thus helps protect consumers from risks in the food chain.

FAO – Food and Agriculture Organisation. The FAO is a neutral intergovernmental organisation established by the United Nations. It strives to provide information and supports sustainable agriculture through legislation and national strategies, with the goal of alleviating hunger.

FDA – The United States Food and Drug Administration (also known as USFDA) is a federal agency of the Department of Health and Human Services. The FDA is responsible for protecting and promoting public health through the control and supervision of food safety.

FFDCA – The Food, Drug, and Cosmetic Act is the primary food safety law in the US. The FFDCA authorises the FDA to monitor and regulate the safety of food, drugs, and cosmetics.

Food authentication – The process of irrefutably proving that a food or food ingredient is in its original, genuine, verifiable, and intended form as declared and represented.

Food fraud – A collective term used to encompass the intentional substitution, addition, tampering, or misrepresentation of food, food ingredients, or food packaging with the aim of increased economic gain.

Food hazard – Food safety hazards occur when food is exposed to hazardous agents which result in contamination of that food. Food hazards may be biological, chemical, physical, allergenic, nutritional, and/or biotechnology-related.

General Food Law – The General Food Law Regulation is the foundation of food and feed law of the European Union. It sets out an overarching and coherent framework for the development of food and feed legislation both at Union and national levels. To this end, it lays down general principles, requirements and procedures that underpin decision making in matters of food and feed safety, covering all stages of food and feed production and distribution.

GMO – Genetically modified organism. An organism whose genome has been engineered in the laboratory in order to favour the expression of desired physiological traits or the generation of desired biological products.

HACCP – Hazard Analysis Critical Control Point. A management system in which food safety is addressed through the analysis and control of biological, chemical, and physical hazards from raw material production, procurement and handling, to manufacturing, distribution and consumption of the finished product.

Mycotoxins – Naturally occurring toxins produced by certain moulds (fungi) that are chemical food hazards. The moulds can grow on a variety of different crops and foodstuffs including cereals, nuts, spices, dried fruits, apples, and coffee beans, often in warm and humid conditions.

PDO – Protected Designation of Origin. Registered designation of products that have the strongest links to their area of production and protected by intellectual property rights.

Phytotoxins – Plant toxins are naturally produced as secondary metabolites, and play a central role in the organism from natural threats. The main groups of plant toxins are alkaloids, terpenes, glycosides, proteinaceous compounds, organic acids, and resinoid compounds.

Plant allergen – A plant derived substance that causes an allergic reaction in humans.

Supply chain – The network of all individuals, organisations, resources, activities, and technologies involved in the creation and sale of a product. A supply chain encompasses everything from the delivery of source materials from the supplier to the manufacturer through to its eventual delivery to the end user.

Traceability – The ability to track any food, feed, food-producing animal, or substance that will be used for consumption, through all stages of production, processing, and distribution.

WHO – World Health Organisation, a part of the United Nations that deals with major health issues around the world. The World Health Organization sets standards for disease control, health care, and medicines, conducts education and research programs, and publishes scientific papers and reports.

References

Aldeguer M, López-Andreo M, Gabaldón JA, Puyet A (2014) Detection of mandarin in orange juice by single-nucleotide polymorphism qPCR assay. Food Chem. 145, 1086–1091. https://doi.org/10.1016/j.foodchem.2013.09.002

Anthoons B, Karamichali I, Schrøder-Nielsen A, Drouzas AD, de Boer H, Madesis P (2020) Metabarcoding reveals low fidelity and presence of toxic species in short chain-of-commercialization of herbal products. Journal of Food Composition and Analysis 103767. https://doi.org/10.1016/j.jfca.2020.103767

Anthoons B, Lagiotis G, Drouzas AD, de Boer H, Madesis P (2022) Barcoding High Resolution Melting (Bar-HRM) enables the discrimination between toxic plants and edible vegetables prior to consumption and after digestion. J. Food Sci. 87, 4221–4232. https://doi.org/10.1111/1750-3841.16253

Armani A, Guardone L, La Castellana R, Gianfaldoni D, Guidi A, Castigliego L (2015) DNA barcoding reveals commercial and health issues in ethnic seafood sold on the Italian market. Food Control 55, 206–214. https://doi.org/10.1016/j.foodcont.2015.02.030

Atanassova S, Veleva P, Stoyanchev T (2018) Near-infrared spectral informative indicators for meat and dairy products, bacterial contamination, and freshness evaluation, in: Microbial Contamination and Food Degradation. Elsevier, pp. 315–340. https://doi.org/10.1016/B978-0-12-811515-2.00010-X

Azadmard-Damirchi S, Torbati M (2015) Adulterations in some edible oils and fats and their detection methods. Journal of Food Quality and Hazards Control 2, 38–44.

Bakshi A (2003) Potential adverse health effects of genetically modified crops. J. Toxicol. Environ. Health B Crit. Rev. 6, 211–225. https://doi.org/10.1080/10937400306469

Bansal S, Singh A, Mangal M, Mangal AK, Kumar S (2017) Food adulteration: sources, health risks, and detection methods. Crit. Rev. Food Sci. Nutr. 57, 1174–1189. https://doi.org/10.1080/10408398.2014.967834

Bawa AS, Anilakumar KR (2013) Genetically modified foods: safety, risks and public concerns-a review. J. Food Sci. Technol. 50, 1035–1046. https://doi.org/10.1007/s13197-012-0899-1

Biberci E, Altuntas Y, Cobanoglu A, Alpinar A (2002) Acute respiratory arrest following hemlock (*Conium maculatum*) intoxication. J. Toxicol. Clin. Toxicol. 40, 517–518. https://doi.org/10.1080/14773996.2002.11681088

Bilia AR (2015) Herbal medicinal products versus botanical-food supplements in the European market: state of art and perspectives. Nat. Prod. Commun. 10, 125–131. https://doi.org/10.1177/1934578X1501000130

Borchers A, Teuber SS, Keen CL, Gershwin ME (2010) Food safety. Clin. Rev. Allergy Immunol. 39, 95–141. https://doi.org/10.1007/s12016-009-8176-4

Bosmali I, Ordoudi SA, Tsimidou MZ, Madesis P (2017) Greek PDO saffron authentication studies using species specific molecular markers. Food Res. Int. 100, 899–907. https://doi.org/10.1016/j.foodres.2017.08.001

Bouxin A (2014) Management of safety in the feed chain, in: Food Safety Management. Elsevier, pp. 23–43. https://doi.org/10.1016/B978-0-12-381504-0.00002-0

Brackett RE (1992) Shelf stability and safety of fresh produce as influenced by sanitation and disinfection. J. Food Prot. 55, 808–814. https://doi.org/10.4315/0362-028X-55.10.808

Bruneton J (2000) Toxic plants: dangerous to humans and animals. New Phytologist 148, 57–58. https://doi.org/10.1046/j.1469-8137.2000.00735b.x

Clemente TE, Cahoon EB (2009) Soybean oil: genetic approaches for modification of functionality and total content. Plant Physiol. 151, 1030–1040. https://doi.org/10.1104/pp.109.146282

Colombo ML, Francesca A, Puppa TD, Paola M, Sesana F, Maurizio B, Rossana B, Sandro P, Gabriele G, Enrico B, Franca D (2010) Most commonly plant exposures and intoxications from outdoor toxic plants. J Pharm Sci & Res 2, 417–425.

Cook PW, Nightingale KK (2018) Use of omics methods for the advancement of food quality and food safety. Anim. Front. 8, 33–41. https://doi.org/10.1093/af/vfy024

Dabbene F, Gay P, Tortia C (2014) Traceability issues in food supply chain management: a review. Biosystems Engineering 120, 65–80. https://doi.org/10.1016/j.biosystemseng.2013.09.006

Deshpande SS (2002) Handbook of food toxicology. CRC Press. https://doi.org/10.1201/9780203908969

Dyckman LJ (2003) Bioterrorism: a threat to agriculture and the food supply. United Stated General Accounting Office.

Edwards M (2014) Food hygiene and foreign bodies, in: Hygiene in Food Processing. Elsevier, pp. 441–464. https://doi.org/10.1533/9780857098634.3.441

EIT Food (2019) Food safety eurobarometer: 50% of Europeans rank food safety among their top three food-buying priorities. EIT Food.

European Commision (2006) Corrigendum to Regulation (EC) No 1924/2006 of the European Parliament and of the Council of 20 December 2006 on nutrition and health claims made on foods, 1924/2006.

European Commission (2002) Regulation (EC) No 178/2002 of the European Parliament and of the Council.

European Commission (2003) Regulation (EC) No 1830/2003 of the European Parliament and of the Council.

European Food Safety Authority (EFSA) (2012) Science protecting consumers from field to fork.

Fagan J, Antoniou M, Robinson C (2014) GMO myths and truths. Earth Open Source, Great Britain.

Fairchild A, Lee MD, Maurer JJ (2006) PCR basics, in: Maurer, J. (Ed.), PCR Methods in Foods. Springer, pp. 1–25. https://doi.org/10.1007/0-387-31702-3_1

Francis GA, Thomas C, O'beirne D (1999) The microbiological safety of minimally processed vegetables. Int. J. Food Sci. Technol. 34, 1–22. https://doi.org/10.1046/j.1365-2621.1999.00253.x

Fritsche J (2018) Recent developments and digital perspectives in food safety and authenticity. J. Agric. Food Chem. 66, 7562–7567. https://doi.org/10.1021/acs.jafc.8b00843

Fung F, Wang H-S, Menon S (2018) Food safety in the 21st century. Biomed. J. 41, 88–95. https://doi.org/10.1016/j.bj.2018.03.003

Galimberti A, De Mattia F, Losa A, Bruni I, Federici S, Casiraghi M, Martellos S, Labra M (2013) DNA barcoding as a new tool for food traceability. Food Res. Int 50, 55–63. https://doi.org/10.1016/j.foodres.2012.09.036

Galvin-King P, Haughey SA, Elliott CT (2018) Herb and spice fraud; the drivers, challenges and detection. Food Control 88, 85–97. https://doi.org/10.1016/j.foodcont.2017.12.031

Ganopoulos I, Bazakos C, Madesis P, Kalaitzis P, Tsaftaris A (2013) Barcode DNA high-resolution melting (Bar-HRM) analysis as a novel close-tubed and accurate tool for olive oil forensic use. J. Sci. Food Agric. 93, 2281–2286. https://doi.org/10.1002/jsfa.6040

Ganopoulos I, Madesis P, Darzentas N, Argiriou A, Tsaftaris A (2012) Barcode High Resolution Melting (Bar-HRM) analysis for detection and quantification of PDO "Fava Santorinis" (*Lathyrus clymenum*) adulterants. Food Chem. 133, 505–512. https://doi.org/10.1016/j.foodchem.2012.01.015

Gatew H, Mengistu K (2019) Genetically modified foods (GMOs); a review of genetic engineering. jwpr 9, 157–163. https://doi.org/10.36380/scil.2019.jlsb25

Gizaw Z (2019) Public health risks related to food safety issues in the food market: a systematic literature review. Environ. Health Prev. Med. 24, 68. https://doi.org/10.1186/s12199-019-0825-5

Govindaraghavan S, Sucher NJ (2015) Quality assessment of medicinal herbs and their extracts: criteria and prerequisites for consistent safety and efficacy of herbal medicines. Epilepsy Behav. 52, 363–371. https://doi.org/10.1016/j.yebeh.2015.03.004

Haiminen N, Edlund S, Chambliss D, Kunitomi M, Weimer BC, Ganesan B, Baker R, Markwell P, Davis M, Huang BC, Kong N, Prill RJ, Marlowe CH, Quintanar A, Pierre S, Dubois G, Kaufman JH, Parida L, Beck KL (2019) Food authentication from shotgun sequencing reads with an application on high protein powders. npj Sci. Food 3, 24. https://doi.org/10.1038/s41538-019-0056-6

Hefferon KL (2015) Nutritionally enhanced food crops; progress and perspectives. Int. J. Mol. Sci. 16, 3895–3914. https://doi.org/10.3390/ijms16023895

Hoffmann-Sommergruber K (2000) Plant allergens and pathogenesis-related proteins. What do they have in common? Int. Arch. Allergy Immunol. 122, 155–166. https://doi.org/10.1159/000024392

Huang T, Li L, Liu X, Chen Q, Fang X, Kong J, Draz MS, Cao H (2020) Loop-mediated isothermal amplification technique: principle, development and wide application in food safety. Anal. Methods 12, 5551–5561. https://doi.org/10.1039/d0ay01768j

Ivanova NV, Kuzmina ML, Braukmann TWA, Borisenko AV, Zakharov EV (2016) Authentication of herbal supplements using next-generation sequencing. PLoS ONE 11, e0156426. https://doi.org/10.1371/journal.pone.0156426

Jagadeesan B, Gerner-Smidt P, Allard MW, Leuillet S, Winkler A, Xiao Y, Chaffron S, Van Der Vossen J, Tang S, Katase M, McClure P, Kimura B, Ching Chai L, Chapman J, Grant K (2019) The use of next generation sequencing for improving food safety: translation into practice. Food Microbiol. 79, 96–115. https://doi.org/10.1016/j.fm.2018.11.005

Keefover-Ring K, Thompson JD, Linhart YB (2009) Beyond six scents: defining a seventh *Thymus vulgaris* chemotype new to southern France by ethanol extraction. Flavour Fragr. J. 24, 117–122. https://doi.org/10.1002/ffj.1921

Keese P (2008) Risks from GMOs due to horizontal gene transfer. Environ. Biosafety Res. 7, 123–149. https://doi.org/10.1051/ebr:2008014

Kharayat BS, Singh Y (2018) Mycotoxins in foods: mycotoxicoses, detection, and management, in: Microbial Contamination and Food Degradation. Elsevier, pp. 395–421. https://doi.org/10.1016/B978-0-12-811515-2.00013-5

Klintschar M, Beham-Schmidt C, Radner H, Henning G, Roll P (1999) Colchicine poisoning by accidental ingestion of meadow saffron (*Colchicum autumnale*): pathological and medicolegal aspects. Forensic Sci. Int. 106, 191–200. https://doi.org/10.1016/S0379-0738(99)00191-7

Lagiotis G, Stavridou E, Bosmali I, Osathanunkul M, Haider N, Madesis P (2020) Detection and quantification of cashew in commercial tea products using High Resolution Melting (HRM) analysis. J. Food Sci. 85, 1629–1634. https://doi.org/10.1111/1750-3841.15138

Low TY, Wong KO, Yap ALL, De Haan LHJ, Rietjens IMCM (2017) The regulatory framework across international jurisdictions for risks associated with consumption of botanical food supplements. Comp. Rev. Food Sci. Food Safety 16, 821–834. https://doi.org/10.1111/1541-4337.12289

Lv L, Huang W, Yu X, Ren H, Sun R (2009) Comparative research of different Bupleurum chinense composition to influence of hepatotoxicity of rats and oxidative damage mechanism. Zhongguo Zhong Yao Za Zhi 34, 2364–2368.

Mailer RJ, Gafner SG (2020) Adulteration of olive (*Olea europaea*) oil. Botanical Adulterants Prevention Bulletin 19, 1–14.

Mishra P, Kumar A, Nagireddy A, Mani DN, Shukla AK, Tiwari R, Sundaresan V (2016) DNA barcoding: an efficient tool to overcome authentication challenges in the herbal market. Plant Biotechnol. J. 14, 8–21. https://doi.org/10.1111/pbi.12419

Mitenius N, Kennedy SP, Busta FF (2014) Food defense, in: Food Safety Management. Elsevier, pp. 937–958. https://doi.org/10.1016/B978-0-12-381504-0.00035-4

Nerín C, Aznar M, Carrizo D (2016) Food contamination during food process. Trends Food Sci. Technol. 48, 63–68. https://doi.org/10.1016/j.tifs.2015.12.004

Newmaster SG, Grguric M, Shanmughanandhan D, Ramalingam S, Ragupathy S (2013) DNA barcoding detects contamination and substitution in North American herbal products. BMC Med. 11, 222. https://doi.org/10.1186/1741-7015-11-222

Overbosch P, Blanchard S (2014) Principles and systems for quality and food safety management, in: Food Safety Management. Elsevier, pp. 537–558. https://doi.org/10.1016/B978-0-12-381504-0.00022-6

Picardo M, Filatova D, Nuñez O, Farré M (2019) Recent advances in the detection of natural toxins in freshwater environments. TrAC Trends in Analytical Chemistry 112, 75–86. https://doi.org/10.1016/j.trac.2018.12.017

Piližota V (2014) Fruits and vegetables (including herbs), in: Food Safety Management. Elsevier, pp. 213–249. https://doi.org/10.1016/B978-0-12-381504-0.00009-3

Rábade LA, Alfaro JA (2006) Buyer–supplier relationship's influence on traceability implementation in the vegetable industry. Journal of Purchasing and Supply Management 12, 39–50. https://doi.org/10.1016/j.pursup.2006.02.003

Raclariu AC, Heinrich M, Ichim MC, de Boer H (2018) Benefits and limitations of DNA barcoding and metabarcoding in herbal product authentication. Phytochem. Anal. 29, 123–128. https://doi.org/10.1002/pca.2732

Raclariu AC, Mocan A, Popa MO, Vlase L, Ichim MC, Crisan G, Brysting AK, de Boer H (2017) *Veronica officinalis* product authentication using DNA metabarcoding and HPLC-MS reveals widespread adulteration with *Veronica chamaedrys*. Front. Pharmacol. 8, 378. https://doi.org/10.3389/fphar.2017.00378

Ripp F, Krombholz CF, Liu Y, Weber M, Schäfer A, Schmidt B, Köppel R, Hankeln T (2014) All-Food-Seq (AFS): a quantifiable screen for species in biological samples by deep DNA sequencing. BMC Genomics 15, 639. https://doi.org/10.1186/1471-2164-15-639

Roberge LF (2019) Agrobioterrorism, in: Singh, S.K., Kuhn, J.H. (Eds.), Defense against Biological Attacks. Springer International Publishing, Cham, pp. 359–383. https://doi.org/10.1007/978-3-030-03071-1_16

Schilter B, Constable A, Perrin I (2014) Naturally occurring toxicants of plant origin, in: Food Safety Management. Elsevier, pp. 45–57. https://doi.org/10.1016/B978-0-12-381504-0.00003-2

Seethapathy GS, Raclariu-Manolica A-C, Anmarkrud JA, Wangensteen H, de Boer HJ (2019) DNA metabarcoding authentication of ayurvedic herbal products on the European market raises concerns of quality and fidelity. Front. Plant Sci. 10, 68. https://doi.org/10.3389/fpls.2019.00068

Shahali Y, Dadar M (2018) Plant food allergy: influence of chemicals on plant allergens. Food Chem. Toxicol. 115, 365–374. https://doi.org/10.1016/j.fct.2018.03.032

Song W, Song Z, Vincent J, Wang H, Wang Z (2020) Quantification of extra virgin olive oil adulteration using smartphone videos. Talanta 216, 120920. https://doi.org/10.1016/j.talanta.2020.120920

Speijers G, van Egmond H (1999) Natural toxins III: inherent plant toxins, in: van der Heijden, K., Younes, M., Fishbein, L., Miller, S. (Eds.), International Food Safety Handbook. pp. 369–380.

Speranskaya AS, Khafizov K, Ayginin AA, Krinitsina AA, Omelchenko DO, Nilova MV, Severova EE, Samokhina EN, Shipulin GA, Logacheva MD (2018) Comparative analysis of Illumina and Ion Torrent high-throughput sequencing platforms for identification of plant components in herbal teas. Food Control 93, 315–324. https://doi.org/10.1016/j.foodcont.2018.04.040

Spink J, Moyer DC (2011) Defining the public health threat of food fraud. J. Food Sci. 76, R157-63. https://doi.org/10.1111/j.1750-3841.2011.02417.x

Staats M, Arulandhu AJ, Gravendeel B, Holst-Jensen A, Scholtens I, Peelen T, Prins TW, Kok E (2016) Advances in DNA metabarcoding for food and wildlife forensic species identification. Anal. Bioanal. Chem. 408, 4615–4630. https://doi.org/10.1007/s00216-016-9595-8

Starodub NF, Novgorodova O, Ogorodnijchuk Y (2018) Biosensors and express control of bacterial contamination of different environmental objects, in: Microbial Contamination and Food Degradation. Elsevier, pp. 367–394. https://doi.org/10.1016/B978-0-12-811515-2.00012-3

Suffert F, Latxague É, Sache I (2009) Plant pathogens as agroterrorist weapons: assessment of the threat for European agriculture and forestry. Food Sec. 1, 221–232. https://doi.org/10.1007/s12571-009-0014-2

Tayal A, Solanki A, Kondal R, Nayyar A, Tanwar S, Kumar N (2020) Blockchain-based efficient communication for food supply chain industry: transparency and traceability analysis for sustainable business. Int. J. Commun. Syst. 34, e4696. https://doi.org/10.1002/dac.4696

Thakkar S, Anklam E, Xu A, Ulberth F, Li J, Li B, Hugas M, Sarma N, Crerar S, Swift S, Hakamatsuka T, Curtui V, Yan W, Geng X, Slikker W, Tong W (2020) Regulatory landscape of dietary supplements and herbal medicines from a global perspective. Regul. Toxicol. Pharmacol. 114, 104647. https://doi.org/10.1016/j.yrtph.2020.104647

Toci AT, de Moura Ribeiro MV, de Toledo PRAB, Boralle N, Pezza HR, Pezza L (2018) Fingerprint and authenticity roasted coffees by 1H-NMR: the Brazilian coffee case. Food Sci. Biotechnol. 27, 19–26. https://doi.org/10.1007/s10068-017-0243-7

U.S. Food and Drug Administration (1997) HACCP principles & application guidelines.

U.S. Food and Drug Administration (2019) Guidelines for the validation of chemical methods for the FDA foods program.

Ulberth F (2020) Tools to combat food fraud - A gap analysis. Food Chem. 330, 127044. https://doi.org/10.1016/j.foodchem.2020.127044

van Duijn G (2014) Oils and fats, in: Food Safety Management. Elsevier, pp. 325–345. https://doi.org/10.1016/B978-0-12-381504-0.00013-5

Vetter J (2004) Poison hemlock (*Conium maculatum* L.). Food Chem. Toxicol. 42, 1373–1382. https://doi.org/10.1016/j.fct.2004.04.009

Vlachos I, Malindretos GP (2012) Farm SMEs sustainability assessment based on Bellagio principles. The case of the Messinian region, Greece. Regional Science Inquiry Journal 4, 137–153.

Whitby SM (2002) Biological warfare against crops. Palgrave Macmillan UK, London. https://doi.org/10.1057/9780230514645

Wisniewski A, Buschulte A (2019) How to tackle food fraud in official food control authorities in Germany. J. Consum. Prot. Food Saf. 14, 319–328. https://doi.org/10.1007/s00003-019-01228-2

Wong BBM, Candolin U (2015) Behavioral responses to changing environments. Behavioral Ecology 26, 665–673. https://doi.org/10.1093/beheco/aru183

World Health Organization (2020) Food safety. World Health Organization.

Zhang N, Erickson DL, Ramachandran P, Ottesen AR, Timme RE, Funk VA, Luo Y, Handy SM (2017) An analysis of *Echinacea* chloroplast genomes: Implications for future botanical identification. Sci. Rep. 7, 216. https://doi.org/10.1038/s41598-017-00321-6

Answers

1. A. Since a considerable amount of the wheat was detected, intentional adulteration would be the most likely food safety issue. Unfortunately, this is common practice in herbal products and spices. Wheat and other grain based components are being used as fillers to increase profits due to cost differential. B. Intentional adulteration will likely happen during processing or packaging. In case of unintentional adulteration, the harvesting step in the supply chain would be the most probable step that might have caused this issue. Wheat is a very common plant in agricultural fields and is very likely to get harvested accidently with the cultivated herb/ vegetable, either by machines or manually. In addition, like many other agricultural crops, wheat is wind-pollinated. Wheat pollen can therefore contaminate the product at various steps in the supply chain, from cultivation to the packaging step. C. Adding wheat, a plant allergen, as an adulterant in food products could create serious health issues especially for those with gluten intolerance. Also, adding cheap substitutes to expensive spices has economic consequences as well, since you pay a high price for low quality products.
2. Although the benefits of GMOs are vast, it is important to note that some health risks are associated with them that are not always fully understood. Genetically modified plants may cause hazards related to increased allergenicity, transfer of genes from GM food to cells of the body, or to bacteria in the gastrointestinal tract.
3. The significant financial rewards, low availability of high-quality extra virgin olive oil as a result of the increasing demand, and inadequate screening from regulatory agencies are the three main reasons for extra virgin oil adulteration.

Chapter 24
Environmental and biodiversity assessments

Maria Ariza[1], Sandra Garcés-Pastor[2], Hugo J. de Boer[1]

1 Natural History Museum, University of Oslo, Norway
2 The Arctic University Museum of Norway, UiT - The Arctic University of Norway, Tromsø, Norway

Maria Ariza **mariadelosangelesariza@gmail.com**
Sandra Garcés-Pastor **sandra.garces-pastor@uit.no**
Hugo J. de Boer **h.de.boer@nhm.uio.no**

Citation: Ariza M, Garcés-Pastor S, de Boer HJ (2022) Chapter 24. Environmental and biodiversity assessments. In: de Boer H, Rydmark MO, Verstraete B, Gravendeel B (Eds) Molecular identification of plants: from sequence to species. Advanced Books. https://doi.org/10.3897/ab.e98875

Introduction

Being the world's most abundant life kingdom, plants are virtually everywhere: in terrestrial, freshwater, and marine ecosystems and even in the air in the form of pollen and spores (Bar-On et al. 2018). They can survive in extreme environments such as the arctic, deserts, and even concrete (Antonelli et al. 2020). Plants are crucial to nearly all ecosystems, and sustain primary production, nutrient cycling, food chains, and multi-scale networks (Corlett 2020). These characteristics make them good indicators of associated biodiversity, surrounding abiotic features, anthropogenic activities (Brunbjerg et al. 2018; Kier et al. 2005; Terwayet Bayouli et al. 2021; Uuemaa et al. 2013) and suitable organisms for environmental and total biodiversity assessments. Since many plants are sessile and perennial, their spatial distribution is not restricted to temporal fluctuations as with organisms, i.e., animals, and thus diversity can be easily quantified, leveraging the accuracy and efficiency of its assessment. Indeed, plant biodiversity assessments are often used to describe biome and landscape changes, to map habitats, and to monitor environmental quality, pollution, and responses to climate change (Halvorsen et al. 2020; Mucina 2019; Steinbauer et al. 2018; Terwayet Bayouli et al. 2021).

However, plant biodiversity assessments are impeded by problems associated with species detection, taxonomic assignment, abundance quantification, and sample bias given the unknown spatial and temporal distribution of target species (Beng and Corlett 2020). Traditional plant assessments have relied on plant morphological characters to identify and inventory diversity, these processes are also often limited to seasonal or life-history stages and require skilled botanists (Scott and Hallam 2003). Additionally, morphology-based assessments are labour intensive, invasive, and prone to observer-bias (Milberg et al. 2008). Although plant identification through organismal or extra-organismal DNA traces extracted from environmental samples (namely environmental DNA or eDNA) has enabled multiple and simultaneous detections at any season, including detection of rare taxa and those that are challenging to collect, complete and reliable plant biodiversity assessments remain challenging (Deiner et al. 2017; Hartvig et al. 2021; Taberlet et al. 2012). Hence, the complementary strength and knowledge of both traditional and eDNA-based assessments and from botanists and molecular ecologists is still required for better estimations of total plant diversity.

Improving plant biodiversity assessments is one of the century's greatest challenges as less than 10% of the world's plant diversity is currently known, and its loss outpaces the rate at which is discovered, inventoried, and protected (Corlett 2016). Furthermore, current global pressures on biodiversity, e.g., invasive species, climate change, environmental pollution and habitat loss, highlight the necessity of biodiversity data to mitigate these impacts (Corlett 2020). Molecular inventorying of plant diversity through eDNA-based assessments show great potential to meet these needs and offers novel opportunities to register the dynamics of species, populations, and communities over long time periods and across large spatial scales (Kersey et al. 2020; Rodríguez-Ezpeleta et al. 2021). This chapter focuses on plant biodiversity (green algae, liverworts, hornworts, mosses, and vascular plants) and environmental characteristics that can be assessed using both eDNA substrates and organismal DNA, and their applications to conservation, ecology, monitoring both diversity and invasive species.

Assessing plant DNA from the environment: power, precautions, and limitations

While many plants are sessile and their biomass is mainly located below or above anchoring surfaces, some vegetative and reproductive plant parts (i.e. flowers, leaf debris, pollen, seeds) detach and are transported on short or great distances from the main organismal body until they are finally deposited onto substrates (i.e., ground, water, and more). Hence, plant DNA can be found in environmental substrates as organismal and extra-organismal DNA at various proportions, with each substrate potentially tracking different spatial and temporal signatures of biodiversity (Rodríguez-Ezpeleta et al. 2021). The detection of these plant DNA sources can also be associated with both DNA status (intracellular or extracellular) and environmental conditions that may enhance or diminish DNA permanence and degradation, i.e., organic particles that bind DNA support its environmental persistence or UV light exposure that results in degradation (Nagler et al. 2018; Pietramellara et al. 2009). Nevertheless, DNA from environmental substrates degrades and decays over time and thus, its assessment can be facilitated by targeting short informative DNA fragments (Shogren et al. 2018). Indeed, eDNA-based plant assessments commonly employ metabarcoding analysis of the chloroplast *trnL* (UUA) intron p6 loop which has a short sequence ranging from 10–143 bp and primer binding sites that are well conserved in vascular and nonvascular plants (Taberlet et al. 2007). As DNA degrades over time, it is easier to target a short fragment for amplification for eDNA, sedaDNA and aDNA applications. Additionally, the p6 loop has a secondary structure that provides extra stability and resistance to degradation (Taberlet et al. 2007). However, low species resolution, particularly for bryophytes, and misidentification due to PCR bias hinders the use of this region to perform complete biodiversity assessments (Ariza et al. 2022).

As no single marker provides resolution for all taxa, eDNA-based assessments often employ metabarcoding of different nuclear and chloroplast regions such as ITS, *rbcL*, and *matK* to harvest their complementary resolution power (see Chapter 11 Amplicon metabarcoding for information about these regions and their suitable applicability; CBOL Plant Working Group 2009; Hollingsworth et al. 2011). Targeted capture of multiple informative genes and shotgun sequencing of environmental samples have recently gained attention as alternative approaches for assessment of plant diversity from eDNA samples as amplification-free methods (see Chapter 12 Metagenomics and Chapter 14 Target capture for more on these methods and the markers used; Chua et al. 2021a; Foster et al. 2021).

Despite the major recent advances in detection, eDNA-based assessments remain limited to reliably quantify abundance, which in turn makes it hard to assess population status and take management actions (Deiner et al. 2017). Although correlations between plant biomass and DNA concentration in the environmental samples are poorly understood, the use of sequence counts of identified taxa is becoming widely accepted in eDNA studies as a proxy of relative abundance (Deagle et al. 2019, 2013; Deiner et al. 2021). Particularly for plants, the assessment of eDNA from root communities has been shown to provide robust abundance estimations (Matesanz et al. 2019).

Furthermore, presence/absence estimations provided by eDNA-based assessments can be misleading as DNA may remain in the environment after the organism is no longer present (Harrison et al. 2019). Thus, plant eDNA-based assessments should be interpreted as merely detections of organismal DNA until evaluations of false occurrence estimations are investigated. Site occupancy-detection models have recently gained attention for this purpose, though false detections of plant DNA remain largely unexplored (Ficetola et al. 2016; Guillera-Arroita et al. 2017).

Substrates for eDNA in environmental and biodiversity assessments

About a decade after the term eDNA was introduced, the eDNA scientific community has adopted different terminology in reference to the state, source, or substrate from which eDNA is isolated (Pawlowski et al. 2020; Rodríguez-Ezpeleta et al. 2021). This chapter focuses on the substrates from which plant eDNA can be isolated, including air, faeces, pollen, soil, sediments, and water, as well as bulk samples such as flowers, leaves, or roots from which organismal DNA can be isolated. Each of these substrates harbours different plant eDNA sources and spatio-temporal signals from the environment. Careful consideration of the study questions and/or applications are required when selecting an eDNA substrate as this will impact the conclusions that can be derived from the assessments. More details on sampling and DNA extraction from eDNA substrates can be found in section 1 of this book.

Airborne samples

Pollen DNA is most commonly the main source of plant eDNA present in airborne samples, although single-cell algae, leaf and flower fragments may also be present (Eaton et al. 2018; Johnson et al. 2019; Núñez et al. 2019, 2017; Sherwood et al. 2017). Pollen from anemophilous terrestrial plants is especially abundant in airborne samples. Since airborne pollen can be transported over long distances it can provide information on regional vegetation (Eaton et al. 2018; Johnson et al. 2019; Núñez et al. 2019, 2017; Sherwood et al. 2017). Using dust traps, pollen from insect-pollinated plants can also be detected but its relation to local plant biomass and the effect of climatic conditions such as wind and temperature on detectability are poorly understood. Nevertheless, plant assessments through pollen metabarcoding from airborne samples have successfully characterised spatial and temporal heterogeneity (Leontidou et al. 2021; Polling et al. 2022), airborne communities (Craine et al. 2017; Núñez et al. 2017), and have been applied to pollen allergen monitoring (Kraaijeveld et al. 2015; Polling et al. 2022; Rowney et al. 2021). eDNA-based airborne monitoring in particular leverages the identification resolution of common plant-allergen families, i.e., Urticaceae, Taxaceae, Poaceae, and abundance estimations (Campbell et al. 2020; Polling et al. 2022; Rowney et al. 2021).

Faecal substrates

Faeces, mucus, and saliva contain DNA from the host and from the organisms that were ingested or that have been in contact with the host (Valentini 2007). Here, we follow Yoccoz (2012) and Pawlowski et al. (2020) and include faeces and other bodily substances as eDNA. Other authors have excluded these sources of DNA as host-associated and distinct from environmentally distributed DNA. It is important to consider that although such DNA transported in faeces and other materials associated with animals can become environmental DNA, it is not yet the case when faeces is collected for dietary assessments. Faecal samples are the most common excrement source of eDNA used for plant assessments and provide a snapshot of vegetation implicated in trophic interactions. Faeces from herbivorous animals are most commonly used, as droppings are easy to collect and represent a viable option to detect the diet of elusive animals (Holechek et al. 1982). Compared to

morphological assessments of plant remains in faeces, faecal DNA metabarcoding and metagenomics have leveraged the taxonomic resolution of plant dietary items from extinct megafauna, mammals, birds, reptiles, insects, and molluscs (Chua et al. 2021b; Koizumi et al. 2016; Polling et al. 2021; Valentini et al. 2009), revealing in turn more diverse diets than previously conceived (Chua et al. 2021b). Simultaneously, eDNA-inventorying of plant diet items have provided new ecological information to support habitat protection efforts (Chua et al. 2021b; Yamamoto and Uchida 2018), and the monitoring of invasive species (Mori et al. 2017), overgrazing (Craine et al. 2015; Fløjgaard et al. 2017), and dietary niche dynamics (Jorns et al. 2020; Kartzinel et al. 2015; Schure et al. 2021). Furthermore, parallel eDNA assessments of scats from communities of large herbivores has allowed the reconstruction of present and past landscape mosaics of the dominant vegetation (Polling et al. 2021; Schure et al. 2021). Moreover, the collection of residual saliva or mucus directly from plant organs can guide the evaluation of niche specialism and competition for plant resources (Harrer and Levi 2018).

Soil and sedimentary substrates

Soil and sediments, from both terrestrial and aquatic environments, are presumably the substrates where most plant DNA is present, as extra-organismal and organismal DNA from both active and dormant tissues including, roots, debris, fallen vegetative parts, seeds, and pollen are gathered or ultimately deposited in these substrates. Because of the major presence of plant eDNA and the ubiquity of these substrates in both aquatic and terrestrial ecosystems, soil and sedimentary eDNA samples are advantageously appropriate for plant assessments. Differences between soil and sediments can be ambiguous, as both are products of the earth's crusts weathering (Wood 1987). However, in soils the deposition of these products happens in situ and remains on the surface, while in sediments these products are transported and redeposited elsewhere in layers over time. As a consequence, these substrates have different spatio-temporal signals when it comes to the reconstruction of the environment (Deiner et al. 2017; Thomsen and Willerslev 2015). Plant eDNA from soil has been shown to signal local and contemporary vegetation (Ariza et al. 2022; Edwards et al. 2018; Yoccoz et al. 2012), whereas sedimentary samples from marine, lake, or terrestrial cores can combine local, regional, contemporary, and past vegetation signals (Alsos et al. 2018; Thomsen and Willerslev 2015; Willerslev et al. 2003).

Soil eDNA plant assessments have successfully characterised diversity in tropical (Osathanunkul et al. 2021; Yoccoz 2012; Zinger et al. 2019), temperate (Fahner et al. 2016; Yang et al. 2014; Yoccoz et al. 2012), and boreal ecosystems (Edwards et al. 2018; Yoccoz et al. 2012), including the hidden diversity from extreme environments such as deserts (Carrasco-Puga et al. 2021; Palacios Mejia et al. 2021), Antarctica (Carvalho-Silva et al. 2021), geothermal sites (Fraser et al. 2017), and permafrost (Willerslev et al. 2014). Soil eDNA plant inventories have been used to assess both natural and cultivated landscapes (Foucher et al. 2020; Yoccoz et al. 2012), woody encroachment in grasslands (Sepp et al. 2021), habitat from crime scenes (Fløjgaard et al. 2019), and rare terrestrial orchids (Hartvig et al. 2021).

As sediments are deposited throughout time and form distinguishable layers, the eDNA present in these layers (namely sedaDNA) can signal organisms that were likely locally present in ancient environments (Thomsen and Willerslev 2015). The assessment of plant eDNA present in terrestrial ancient sediments has been used to reconstruct the vegetation of the Pleistocene and Holocene in Siberia (Liu et al. 2021; Willerslev et al. 2003), and glacial and interglacial periods in the Arctic (Sønstebø et al. 2010). Further, plant eDNA from sedimentary samples collected in freshwater ecosystems, i.e., lake or riverine sediments, can signal contemporary

and surrounding terrestrial vegetation (Alsos et al. 2018; Giguet-Covex et al. 2019). However, ancient plant DNA present in these samples has been purposely targeted to infer past vegetations including paleo floras (Parducci et al. 2017; Thomsen and Willerslev 2015). Plant eDNA collected from lake sediments has revealed vegetation growing in the arctic during the last interglacial (Crump et al. 2021; Parducci et al. 2012) and post-glacial migration of trees from southern Europe (Epp et al. 2015), human-induced landscape changes and the biological invasions that followed (Ficetola et al. 2018; Giguet-Covex et al. 2014), and even a 5000 year timeline of tropical diversity (Bremond et al. 2017). eDNA metabarcoding of sediments from ancient water reservoirs at the centre of major Maya cities was used to reconstruct the forest types in these ancient cities (Lentz et al. 2021). Finally, eDNA present in coastal marine sediments has been used to monitor seagrasses, salt marshes, and mangrove communities (Foster et al. 2020; Ortega et al. 2020a).

Water samples

eDNA-based biodiversity assessments have proliferated in marine and freshwater environments in recent years, and our knowledge on the persistence, decay rates, and states of eDNA in water samples and its resolution compared to traditional assessments has in parallel increased (Ji et al. 2021; Mauvisseau et al. 2022). However, assessments of plant biodiversity from aquatic environments have been widely overlooked compared to assessments of other organisms across the tree of life. Presumably, plant eDNA present in water samples is mostly composed of extra-organismal DNA bound to suspended small particles derived either from terrestrial or aquatic habitats (Deiner et al. 2016; Drummond et al. 2021; Lacoursière-Roussel and Deiner 2021; Turner et al. 2014). In addition, DNA presence can be vertically stratified, influencing the signals that are retrieved with either shallow or deep water samples (Canals et al. 2021). However, comparisons of assessed diversity with water samples collected at different vertical and horizontal positions in small lakes revealed similar aquatic and terrestrial plant signals, suggesting that eDNA is evenly distributed in freshwater environments and that shore-based sampling can successfully capture beta diversity (Drummond et al. 2021). The latter study in addition showed that read abundances are heavily weighted toward aquatic macrophytes, while taxon richness was greatest in algae and other nonvascular plants. Similar detection patterns were registered in rivers (Ji et al. 2021). Furthermore, aquarium experiments suggest that eDNA concentration and submerged biomass are positively correlated (Matsuhashi et al. 2016).

The assessment of aquatic plant eDNA in freshwater ecosystems has simultaneously enabled the early detection of invasive species (Coghlan et al. 2021; Doi et al. 2021; Fujiwara et al. 2016; Gantz et al. 2018; Scriver et al. 2015), endangered species (Tsukamoto et al. 2021), and water quality indicator species (Gao et al. 2018; Kuzmina et al. 2018; Stoeck et al. 2018). Assessing plant diversity from eDNA in marine systems is harder due to salinity and the massive volumes of water in which plant DNA is diluted. Several studies have still shown the feasibility of using marine plant eDNA to study marine macrophytes (Foster et al. 2020) as well as blue carbon cycling (Ortega et al. 2019, 2020b).

Plant DNA can also be isolated from water samples in the form of snow, firn, and ice (Pedersen et al. 2015). In particular, glacier ice can be advantageous for plant assessments as it gathers plant remains from surrounding environments while maintaining freezing temperatures that preserve DNA naturally for long periods and thus allows the reconstruction of past environments (Varotto et al. 2021). Plant assessments from glacier ice cores have allowed the reconstruction of the conifer communities that once inhabited Greenland (Willerslev et al. 2007) and vegetation transitions during the Last Glacial Maximum throughout Beringia (Pedersen et al. 2016).

Bulk samples

Bulk samples from plants are distinctly different from pitfall or Malaise traps filled with insects. In bulk samples of plants, one can distinguish natural bulk samples such as pollen samples from pollen samplers, or those scraped or washed from pollinating vectors, and those that are artificially assembled such as collected roots, leaves, or flowers. Nevertheless, all bulk samples constitute organismal DNA from plant communities that can be used either to assess plant or other diversity (Deiner et al. 2017).

Flower bulk samples have been assembled to assess arthropod communities that leave DNA traces after either visitation or pollination (Thomsen and Sigsgaard 2019). Leaf bulk samples can be easily collected from the leaf litter. The latter has been particularly used to assess soil fauna and arthropod communities as it can reveal differences in habitat and beta diversity (Lopes et al. 2021; Ritter et al. 2018; Yang et al. 2014). However, the potential of leaf litter metabarcoding to assess vegetation remains unexplored. Artificially assembled leaf bulk samples have been used to assess plant diversity in tropical forests in the Brazilian canga (Vasconcelos et al. 2021). Natural pollen bulk samples are often collected from pollinators or flower visitors, particularly from their pollen baskets (Sookhan et al. 2021). Plant signals from these samples mainly correspond to vegetation involved in ecological interactions of pollination and parasitism and thus are valuable to reconstruct food webs (McFrederick and Rehan 2016; Sookhan et al. 2021). DNA metabarcoding of pollen bulk samples can be used to assess more diverse pollination networks from insects and bats as well as the seasonal availability of food resources (Koyama et al. 2018; Lim et al. 2018; Smart et al. 2017). Furthermore, pollen present in honey can be regarded as a bulk sample as it signals floral sources that bees have visited. Melissopalynology metabarcoding studies have focused either on identification of floral composition of honey, regional provenance, or identification of entomological sources of the honey (Chiara et al. 2021; Hawkins et al. 2015; Milla et al. 2021; Prosser and Hebert 2017; Richardson et al. 2015). Artificially assembled pollen samples such as pollen collected using Burkard samplers for allergenic pollen prognoses can be used to identify airborne pollen as well. Root bulk samples can be assembled to signal hidden belowground plant diversity and its abundance (Lamb et al. 2016; Matesanz et al. 2019). Metabarcoding root diversity in grasslands has revealed a larger fraction of diversity that cannot be detected with traditional assessments of aboveground diversity (Rucińska et al. 2022; Sepp et al. 2021). In addition, the assessment of root bulk samples has elucidated mycorrhizal and parasitic plant associations (Holá et al. 2017; Marčiulynienė et al. 2021).

Beyond eDNA samples: assessing biodiversity through eDNA biotic samplers

A recent development in eDNA metabarcoding is the use of organisms as natural samplers of DNA (coined nsDNA; Mariani et al. 2019). Siegenthaler et al. (2019) show how DNA assessment of gut contents from the European brown shrimp can recover the same number of taxa as using water or sediment eDNA samples from the same area where the shrimps were collected. Similarly, sponges have been shown to be robust natural samplers as they filter high volumes of water and simultaneously trap and concentrate DNA traces from faunal assemblages (Mariani et al. 2019; M. Turon et al. 2020). In terrestrial ecosystems, insectivorous bats have proven to be useful for assessing natural and invasive insect pests (Kemp et

al. 2019; Montauban et al. 2021). Most valuations of biotic samplers have focused on their potential to assess fauna whilst for flora this remains rather unexplored. Hence, we will highlight a few examples of potential biotic samplers that can characterise local floras and other environmental characteristics.

In aquatic ecosystems, macroinvertebrates (Chironomidae, Coleoptera, Hemiptera, Ephemeroptera) that feed both on aquatic vascular plants and plant fragments leached to the environment hold great potential to signal overall vegetation implicated in freshwater trophic relationships. Likewise, filtering organisms or animals that use specialised structures to filter fine particles from the water in lakes and rivers harbour the same potential, i.e., sponges (*Ephydatia*), Simuliidae, Ephemeroptera, Chironomidae, and Trichoptera.

For the assessment of terrestrial vegetation in tropical areas, bats hold great potential as biotic samplers of plant DNA since omnivorous and frugivorous communities are abundant and thus easy to collect (Kalko et al. 1996). For example, seed rains from leaf-nosed bats (Phyllostomidae) can signal understory vegetation that is presently abundant and part of secondary forest succession (Andrade et al. 2013; Charles-Dominique and Cockle 2001). Moreover, assessment of seed rain over time can help track phenological adaptations resulting from recent community turnover and reveal competition avoidance mechanisms of plant coexistence (Thies and Kalko 2004). The DNA assessment of seed rains may also overcome the low taxonomic resolution that traditional morphological identification of seeds yields. Furthermore, specialist organisms for pollination, nectarivore, and seed dispersal harbour the potential to detect elusive plants and reveal other plants that are visited or potentially pollinated. For example, 600 neotropical orchids are specifically pollinated by *Euglossa* bees, which in turn can visit other floral sources (Pemberton and Wheeler 2006; Ramírez et al. 2011).

Finally, amplifying hypervariable markers from biotic DNA samplers, i.e., COI for animals, has recently gained attention as it can assess diversity below the species level, and thus signals ecosystem population assemblages in space and time (metaphylogeography; X. Turon et al. 2020). Metaphylogeography datasets have the potential to provide novel insights that can be applied to conservation genetics, biodiversity management and assessment of protected areas.

While the exploration of eDNA samples and methods for plant assessments is still at its infancy, eDNA has already revolutionised the way and speed in which biodiversity can be inventoried. Plant detection via eDNA has enabled the discovery of plants living in extreme and/or ancient environments and yielded myriad applications with societal relevance. A decade after the rise of eDNA-based assessments, the limitations of this method across different eDNA samples are still being recognised while in parallel different strategies are being developed to overcome and mitigate these. In this rapidly developing field, it is essential to combine the basics of eDNA metabarcoding with the most recent insights and developments in the field to devise the most robust study design to answer your research questions.

Questions

1. You want to assess the floral resources available in summer for a butterfly species and identify potential food competitors. Describe your experimental design and the eDNA substrate(s) that you would use and why.
2. You are hired to conduct a vegetation assessment of a landscape mosaic composed of several small lakes and grasslands, however, you only have the time and budget to collect samples from a single eDNA substrate. Which eDNA substrate would you choose and why?

3. You use soil eDNA to detect the spread of an invasive alien gymnosperm tree species (Sitka spruce, *Picea sitchensis*). Though this species is conspicuously visible, you have not seen it nor has been reported around the sampling area.You detect OTUs in nearly every possible sample, and after a bout of cold sweat realise how this might be explained. What would explain this finding?

Glossary

Organismal DNA – The DNA that is isolated from bulk-extracted mixtures of organisms that are separated from the environmental sample. Also named community DNA.

Extra-organismal DNA – DNA originated (i) from biological material shed from an organism as part of tissue replacement or metabolic waste; (ii) as biologically active propagules such as gametes, pollen, seeds or spores; or (iii) as a result of cell lysis or cell extrusion (Rodriguez-Ezpeleta et al. 2021).

Environmental DNA – DNA captured from modern environments, i.e., seawater, freshwater, soil, or air; or ancient environments, i.e., cores from sediment, ice or permafrost (Thomsen and Willerslev 2015) that have originated from both organismal and extra organismal DNA (Rodriguez-Ezpeleta et al. 2021).

Intracellular DNA – DNA that is located within cell membranes.

Extracellular DNA – DNA that is located free in the environment after cell lysis or cell extrusion.

Anemophily – Plant pollination where pollen is distributed by wind, i.e. wind pollination.

Firn – Crystalline or granular snow, especially on the upper part of a glacier, where it has not yet been compressed into ice.

Melissopalynology – The study of pollen contained in honey and, in particular, the pollen's source.

References

Alsos IG, Lammers Y, Yoccoz NG, Jørgensen T, Sjögren P, Gielly L, Edwards ME (2018) Plant DNA metabarcoding of lake sediments: how does it represent the contemporary vegetation. PLoS ONE 13, e0195403. https://doi.org/10.1371/journal.pone.0195403

Andrade TY, Thies W, Rogeri PK, Kalko EKV, Mello MAR (2013) Hierarchical fruit selection by Neotropical leaf-nosed bats (Chiroptera: Phyllostomidae). J. Mammal. 94, 1094–1101. https://doi.org/10.1644/12-MAMM-A-244.1

Antonelli A, Fry C, Smith RJ, Simmonds MSJ, Kersey PJ, Pritchard HW, Abbo MS, Acedo C, Adams J, Ainsworth AM, Allkin B, Annecke W, Bachman SP, Bacon K, Bárrios S, Barstow C, Battison A, Bell E, Bensusan K, Bidartondo MI, et al. (2020) State of the World's Plants and Fungi 2020. Royal Botanic Gardens, Kew. https://doi.org/10.34885/172

Ariza M, Fouks B, Mauvisseau Q, Halvorsen R, Alsos IG, de Boer H (2022) Plant biodiversity assessment through soil eDNA reflects temporal and local diversity. Methods Ecol. Evol. https://doi.org/10.1111/2041-210X.13865

Bar-On YM, Phillips R, Milo R (2018) The biomass distribution on Earth. Proc Natl Acad Sci USA 115, 6506–6511. https://doi.org/10.1073/pnas.1711842115

Beng KC, Corlett RT (2020) Applications of environmental DNA (eDNA) in ecology and conservation: opportunities, challenges and prospects. Biodivers. Conserv. 29, 2089–2121. https://doi.org/10.1007/s10531-020-01980-0

Bremond L, Favier C, Ficetola GF, Tossou MG, Akouégninou A, Gielly L, Giguet-Covex C, Oslisly R, Salzmann U (2017) Five thousand years of tropical lake sediment DNA records from Benin. Quat. Sci. Rev. 170, 203–211. https://doi.org/10.1016/j.quascirev.2017.06.025

Brunbjerg AK, Bruun HH, Dalby L, Fløjgaard C, Frøslev TG, Høye TT, Goldberg I, Læssøe T, Hansen MDD, Brøndum L, Skipper L, Fog K, Ejrnæs R (2018) Vascular plant species richness and bioindication predict multi-taxon species richness. Methods Ecol. Evol. 9, 2372–2382. https://doi.org/10.1111/2041-210X.13087

Campbell BC, Al Kouba J, Timbrell V, Noor MJ, Massel K, Gilding EK, Angel N, Kemish B, Hugenholtz P, Godwin ID, Davies JM (2020) Tracking seasonal changes in diversity of pollen allergen exposure: targeted metabarcoding of a subtropical aerobiome. Sci. Total Environ. 747, 141189. https://doi.org/10.1016/j.scitotenv.2020.141189

Canals O, Mendibil I, Santos M, Irigoien X, Rodríguez-Ezpeleta N (2021) Vertical stratification of environmentalDNA in the open ocean captures ecological patterns and behavior of deep-sea fishes. Limnol. Oceanogr. 6, 339–347. https://doi.org/10.1002/lol2.10213

Carrasco-Puga G, Díaz FP, Soto DC, Hernández-Castro C, Contreras-López O, Maldonado A, Latorre C, Gutiérrez RA (2021) Revealing hidden plant diversity in arid environments. Ecography 44, 98–111. https://doi.org/10.1111/ecog.05100

Carvalho-Silva M, Rosa LH, Pinto OHB, Da Silva TH, Henriques DK, Convey P, Câmara PEAS (2021) Exploring the plant environmental DNA diversity in soil from two sites on Deception Island (Antarctica, South Shetland Islands) using metabarcoding. Antarctic Science 33, 469–478. https://doi.org/10.1017/S0954102021000274

CBOL Plant Working Group (2009) A DNA barcode for land plants. Proc Natl Acad Sci USA 106, 12794–12797. https://doi.org/10.1073/pnas.0905845106

Charles-Dominique P, Cockle A (2001) Frugivory and seed dispersal by bats, in: Bongers, F., Charles-Dominique, P., Forget, P.-M., Théry, M. (Eds.), Nouragues, Monographiae Biologicae. Springer Netherlands, Dordrecht, pp. 207–216. https://doi.org/10.1007/978-94-015-9821-7_19

Chiara B, Francesco C, Fulvio B, Paola M, Annalisa G, Stefania S, Luigi AP, Simone P (2021) Exploring the botanical composition of polyfloral and monofloral honeys through DNA metabarcoding. Food Control 128, 108175. https://doi.org/10.1016/j.foodcont.2021.108175

Chua PYS, Crampton-Platt A, Lammers Y, Alsos IG, Boessenkool S, Bohmann K (2021a) Metagenomics: a viable tool for reconstructing herbivore diet. Mol. Ecol. Resour. 21, 2249–2263. https://doi.org/10.1111/1755-0998.13425

Chua PYS, Lammers Y, Menoni E, Ekrem T, Bohmann K, Boessenkool S, Alsos IG (2021b) Molecular dietary analyses of western capercaillies (*Tetrao urogallus*) reveal a diverse diet. Environmental DNA 3, 1156–1171. https://doi.org/10.1002/edn3.237

Coghlan SA, Shafer ABA, Freeland JR (2021) Development of an environmental DNA metabarcoding assay for aquatic vascular plant communities. Environmental DNA 3, 372–387. https://doi.org/10.1002/edn3.120

Corlett RT (2016) Plant diversity in a changing world: status, trends, and conservation needs. Plant Diversity 38, 10–16. https://doi.org/10.1016/j.pld.2016.01.001

Corlett RT (2020) Safeguarding our future by protecting biodiversity. Plant Diversity 42, 221–228. https://doi.org/10.1016/j.pld.2020.04.002

Craine JM, Barberán A, Lynch RC, Menninger HL, Dunn RR, Fierer N (2017) Molecular analysis of environmental plant DNA in house dust across the United States. Aerobiologia (Bologna) 33, 71–86. https://doi.org/10.1007/s10453-016-9451-5

Craine JM, Towne EG, Miller M, Fierer N (2015) Climatic warming and the future of bison as grazers. Sci. Rep. 5, 16738. https://doi.org/10.1038/srep16738

Crump SE, Fréchette B, Power M, Cutler S, de Wet G, Raynolds MK, Raberg JH, Briner JP, Thomas EK, Sepúlveda J, Shapiro B, Bunce M, Miller GH (2021) Ancient plant DNA reveals High Arctic greening during the Last Interglacial. Proc Natl Acad Sci USA 118, e2019069118. https://doi.org/10.1073/pnas.2019069118

Deagle BE, Thomas AC, McInnes JC, Clarke LJ, Vesterinen EJ, Clare EL, Kartzinel TR, Eveson JP (2019) Counting with DNA in metabarcoding studies: how should we convert sequence reads to dietary data? Mol. Ecol. 28, 391–406. https://doi.org/10.1111/mec.14734

Deagle BE, Thomas AC, Shaffer AK, Trites AW, Jarman SN (2013) Quantifying sequence proportions in a DNA-based diet study using Ion Torrent amplicon sequencing: which counts count? Mol. Ecol. Resour. 13, 620–633. https://doi.org/10.1111/1755-0998.12103

Deiner K, Bik HM, Mächler E, Seymour M, Lacoursière-Roussel A, Altermatt F, Creer S, Bista I, Lodge DM, de Vere N, Pfrender ME, Bernatchez L (2017) Environmental DNA metabarcoding: transforming how we survey animal and plant communities. Mol. Ecol. 26, 5872–5895. https://doi.org/10.1111/mec.14350

Deiner K, Fronhofer EA, Mächler E, Walser J-C, Altermatt F (2016) Environmental DNA reveals that rivers are conveyer belts of biodiversity information. Nat. Commun. 7, 12544. https://doi.org/10.1038/ncomms12544

Deiner K, Yamanaka H, Bernatchez L (2021) The future of biodiversity monitoring and conservation utilizing environmental DNA. Environmental DNA 3, 3–7. https://doi.org/10.1002/edn3.178

Doi H, Akamatsu Y, Goto M, Inui R, Komuro T, Nagano M, Minamoto T (2021) Broad-scale detection of environmental DNA for an invasive macrophyte and the relationship between DNA concentration and coverage in rivers. Biol. Invasions 23, 507–520. https://doi.org/10.1007/s10530-020-02380-9

Drummond JA, Larson ER, Li Y, Lodge DM, Gantz CA, Pfrender ME, Renshaw MA, Correa AMS, Egan SP (2021) Diversity metrics are robust to differences in sampling location and depth for environmental DNA of plants in small temperate lakes. Front. Environ. Sci. 9, 617924. https://doi.org/10.3389/fenvs.2021.617924

Eaton S, Zúñiga C, Czyzewski J, Ellis C, Genney DR, Haydon D, Mirzai N, Yahr R (2018) A method for the direct detection of airborne dispersal in lichens. Mol. Ecol. Resour. 18, 240–250. https://doi.org/10.1111/1755-0998.12731

Edwards ME, Alsos IG, Yoccoz N, Coissac E, Goslar T, Gielly L, Haile J, Langdon CT, Tribsch A, Binney HA, von Stedingk H, Taberlet P (2018) Metabarcoding of modern soil DNA gives a highly local vegetation signal in Svalbard tundra. The Holocene 28, 2006–2016. https://doi.org/10.1177/0959683618798095

Epp LS, Gussarova G, Boessenkool S, Olsen J, Haile J, Schrøder-Nielsen A, Ludikova A, Hassel K, Stenøien HK, Funder S, Willerslev E, Kjær K, Brochmann C (2015) Lake sediment multi-taxon DNA from North Greenland records early post-glacial appearance of vascular plants and accurately tracks environmental changes. Quat. Sci. Rev. 117, 152–163. https://doi.org/10.1016/j.quascirev.2015.03.027

Fahner NA, Shokralla S, Baird DJ, Hajibabaei M (2016) Large-scale monitoring of plants through environmental DNA metabarcoding of soil: recovery, resolution, and annotation of four DNA markers. PLoS ONE 11, e0157505. https://doi.org/10.1371/journal.pone.0157505

Ficetola GF, Poulenard J, Sabatier P, Messager E, Gielly L, Leloup A, Etienne D, Bakke J, Malet E, Fanget B, Støren E, Reyss J-L, Taberlet P, Arnaud F (2018) DNA from lake sediments reveals long-term ecosystem changes after a biological invasion. Sci. Adv. 4, eaar4292. https://doi.org/10.1126/sciadv.aar4292

Ficetola GF, Taberlet P, Coissac E (2016) How to limit false positives in environmental DNA and metabarcoding? Mol. Ecol. Resour. 16, 604–607. https://doi.org/10.1111/1755-0998.12508

Fløjgaard C, De Barba M, Taberlet P, Ejrnæs R (2017) Body condition, diet and ecosystem function of red deer (*Cervus elaphus*) in a fenced nature reserve. Glob. Ecol. Conserv. 11, 312–323. https://doi.org/10.1016/j.gecco.2017.07.003

Fløjgaard C, Frøslev TG, Brunbjerg AK, Bruun HH, Moeslund J, Hansen AJ, Ejrnæs R (2019) Predicting provenance of forensic soil samples: linking soil to ecological habitats by metabarcoding and supervised classification. PLoS ONE 14, e0202844. https://doi.org/10.1371/journal.pone.0202844

Foster NR, Gillanders BM, Jones AR, Young JM, Waycott M (2020) A muddy time capsule: using sediment environmental DNA for the long-term monitoring of coastal vegetated ecosystems. Mar. Freshwater Res. 71, 869. https://doi.org/10.1071/MF19175

Foster NR, van Dijk K, Biffin E, Young JM, Thomson VA, Gillanders BM, Jones AR, Waycott M (2021) A multi-gene region targeted capture approach to detect plant DNA in environmental samples: a case study from coastal environments. Front. Ecol. Evol. 9, 735744. https://doi.org/10.3389/fevo.2021.735744

Foucher A, Evrard O, Ficetola GF, Gielly L, Poulain J, Giguet-Covex C, Laceby JP, Salvador-Blanes S, Cerdan O, Poulenard J (2020) Persistence of environmental DNA in cultivated soils: implication of this memory effect for reconstructing the dynamics of land use and cover changes. Sci. Rep. 10, 10502. https://doi.org/10.1038/s41598-020-67452-1

Fraser CI, Connell L, Lee CK, Cary SC (2017) Evidence of plant and animal communities at exposed and subglacial (cave) geothermal sites in Antarctica. Polar Biol. 41, 1–5. https://doi.org/10.1007/s00300-017-2198-9

Fujiwara A, Matsuhashi S, Doi H, Yamamoto S, Minamoto T (2016) Use of environmental DNA to survey the distribution of an invasive submerged plant in ponds. Freshwater Science 35, 748–754. https://doi.org/10.1086/685882

Gantz CA, Renshaw MA, Erickson D, Lodge DM, Egan SP (2018) Environmental DNA detection of aquatic invasive plants in lab mesocosm and natural field conditions. Biol. Invasions 20, 1–18. https://doi.org/10.1007/s10530-018-1718-z

Gao W, Chen Z, Li Y, Pan Y, Zhu J, Guo S, Hu L, Huang J (2018) Bioassessment of a drinking water reservoir using plankton: high throughput sequencing vs. traditional morphological method. Water (Basel) 10, 82. https://doi.org/10.3390/w10010082

Giguet-Covex C, Ficetola GF, Walsh K, Poulenard J, Bajard M, Fouinat L, Sabatier P, Gielly L, Messager E, Develle AL, David F, Taberlet P, Brisset E, Guiter F, Sinet R, Arnaud F (2019) New insights on lake sediment DNA from the catchment: importance of taphonomic and analytical issues on the record quality. Sci. Rep. 9, 14676. https://doi.org/10.1038/s41598-019-50339-1

Giguet-Covex C, Pansu J, Arnaud F, Rey P-J, Griggo C, Gielly L, Domaizon I, Coissac E, David F, Choler P, Poulenard J, Taberlet P (2014) Long livestock farming history and human landscape shaping revealed by lake sediment DNA. Nat. Commun. 5, 3211. https://doi.org/10.1038/ncomms4211

Guillera-Arroita G, Lahoz-Monfort JJ, van Rooyen AR, Weeks AR, Tingley R (2017) Dealing with false-positive and false-negative errors about species occurrence at multiple levels. Methods Ecol. Evol. 8, 1081–1091. https://doi.org/10.1111/2041-210X.12743

Halvorsen R, Skarpaas O, Bryn A, Bratli H, Erikstad L, Simensen T, Lieungh E (2020) Towards a systematics of ecodiversity: the EcoSyst framework. Global Ecol. Biogeogr. 29, 1887–1906. https://doi.org/10.1111/geb.13164

Harrer LEF, Levi T (2018) The primacy of bears as seed dispersers in salmon-bearing ecosystems. Ecosphere 9, e02076. https://doi.org/10.1002/ecs2.2076

Harrison JB, Sunday JM, Rogers SM (2019) Predicting the fate of eDNA in the environment and implications for studying biodiversity. Proc. Biol. Sci. 286, 20191409. https://doi.org/10.1098/rspb.2019.1409

Hartvig I, Kosawang C, Kjær ED, Nielsen LR (2021) Detecting rare terrestrial orchids and associated plant communities from soil samples with eDNA methods. Biodivers. Conserv. 30, 3879–3901. https://doi.org/10.1007/s10531-021-02279-4

Hawkins J, de Vere N, Griffith A, Ford CR, Allainguillaume J, Hegarty MJ, Baillie L, Adams-Groom B (2015) Using DNA metabarcoding to identify the floral composition of honey: A new tool for investigating honey bee foraging preferences. PLoS ONE 10, e0134735. https://doi.org/10.1371/journal.pone.0134735

Holá E, Kocková J, Těšitel J (2017) DNA barcoding as a tool for identification of host association of root-hemiparasitic plants. Folia Geobot. 52, 227–238. https://doi.org/10.1007/s12224-017-9286-z

Holechek JL, Vavra M, Pieper RD (1982) Botanical compostion determination of range herbivore diets. a review. Journal of Range Management 35, 309–315.

Hollingsworth PM, Graham SW, Little DP (2011) Choosing and using a plant DNA barcode. PLoS ONE 6, e19254. https://doi.org/10.1371/journal.pone.0019254

Ji F, Yan L, Yan S, Qin T, Shen J, Zha J (2021) Estimating aquatic plant diversity and distribution in rivers from Jingjinji region, China, using environmental DNA metabarcoding and a traditional survey method. Environ. Res. 199, 111348. https://doi.org/10.1016/j.envres.2021.111348

Johnson MD, Cox RD, Barnes MA (2019) Analyzing airborne environmental DNA: A comparison of extraction methods, primer type, and trap type on the ability to detect airborne eDNA from terrestrial plant communities. Environmental DNA 1, 176–185. https://doi.org/10.1002/edn3.19

Jorns T, Craine J, Towne EG, Knox M (2020) Climate structures bison dietary quality and composition at the continental scale. Environmental DNA 2, 77–90. https://doi.org/10.1002/edn3.47

Kalko EKV, Handley CO, Handley D (1996) Organization, diversity, and long-term dynamics of a neotropical bat community, in: Cody, M.L., Smallwood, J.A. (Eds.), Long-Term Studies of Vertebrate Communities. Elsevier, pp. 503–553. https://doi.org/10.1016/B978-012178075-3/50017-9

Kartzinel TR, Chen PA, Coverdale TC, Erickson DL, Kress WJ, Kuzmina ML, Rubenstein DI, Wang W, Pringle RM (2015) DNA metabarcoding illuminates dietary niche partitioning by African large herbivores. Proc Natl Acad Sci USA 112, 8019–8024. https://doi.org/10.1073/pnas.1503283112

Kemp J, López-Baucells A, Rocha R, Wangensteen OS, Andriatafika Z, Nair A, Cabeza M (2019) Bats as potential suppressors of multiple agricultural pests: a case study from Madagascar. Agriculture, Ecosystems & Environment 269, 88–96. https://doi.org/10.1016/j.agee.2018.09.027

Kersey PJ, Collemare J, Cockel C, Das D, Dulloo EM, Kelly LJ, Lettice E, Malécot V, Maxted N, Metheringham C, Thormann I, Leitch IJ (2020) Selecting for useful properties of plants and fungi – Novel approaches, opportunities, and challenges. Plants, People, Planet 2, 409–420. https://doi.org/10.1002/ppp3.10136

Kier G, Mutke J, Dinerstein E, Ricketts TH, Küper W, Kreft H, Barthlott W (2005) Global patterns of plant diversity and floristic knowledge. J. Biogeogr. 32, 1107–1116. https://doi.org/10.1111/j.1365-2699.2005.01272.x

Koizumi N, Mori A, Mineta T, Sawada E, Watabe K, Takemura T (2016) Exploratory environmental DNA analysis for investigating plant-feeding habit of the red-eared turtle using their feces samples. JT 78, 9–13. https://doi.org/10.11113/jt.v78.7253

Koyama A, Egawa C, Taki H, Yasuda M, Kanzaki N, Ide T, Okabe K (2018) Non-native plants are a seasonal pollen source for native honeybees in suburban ecosystems. Urban Ecosyst. 21, 1113–1122. https://doi.org/10.1007/s11252-018-0793-3

Kraaijeveld K, de Weger LA, Ventayol García M, Buermans H, Frank J, Hiemstra PS, den Dunnen JT (2015) Efficient and sensitive identification and quantification of airborne pollen using next-generation DNA sequencing. Mol. Ecol. Resour. 15, 8–16. https://doi.org/10.1111/1755-0998.12288

Kuzmina ML, Braukmann TWA, Zakharov EV (2018) Finding the pond through the weeds: eDNA reveals underestimated diversity of pondweeds. Appl. Plant Sci. 6, e01155. https://doi.org/10.1002/aps3.1155

Lacoursière-Roussel A, Deiner K (2021) Environmental DNA is not the tool by itself. J. Fish Biol. 98, 383–386. https://doi.org/10.1111/jfb.14177

Lamb EG, Winsley T, Piper CL, Freidrich SA, Siciliano SD (2016) A high-throughput belowground plant diversity assay using next-generation sequencing of the trnL intron. Plant Soil 404, 361–372. https://doi.org/10.1007/s11104-016-2852-y

Lentz DL, Hamilton TL, Dunning NP, Tepe EJ, Scarborough VL, Meyers SA, Grazioso L, Weiss AA (2021) Environmental DNA reveals arboreal cityscapes at the Ancient Maya Center of Tikal. Sci. Rep. 11, 12725. https://doi.org/10.1038/s41598-021-91620-6

Leontidou K, Vokou D, Sandionigi A, Bruno A, Lazarina M, De Groeve J, Li M, Varotto C, Girardi M, Casiraghi M, Cristofori A (2021) Plant biodiversity assessment through pollen DNA metabarcoding in Natura 2000 habitats (Italian Alps). Sci. Rep. 11, 18226. https://doi.org/10.1038/s41598-021-97619-3

Lim V-C, Ramli R, Bhassu S, Wilson J-J (2018) Pollination implications of the diverse diet of tropical nectar-feeding bats roosting in an urban cave. PeerJ 6, e4572. https://doi.org/10.7717/peerj.4572

Liu S, Li K, Jia W, Stoof-Leichsenring KR, Liu X, Cao X, Herzschuh U (2021) Vegetation reconstruction from Siberia and the Tibetan plateau using modern analogue technique–comparing sedimentary (ancient) DNA and pollen data. Front. Ecol. Evol. 9, 668611. https://doi.org/10.3389/fevo.2021.668611

Lopes CM, Baêta D, Sasso T, Vanzetti A, Raquel Zamudio K, Taberlet P, Haddad CFB (2021) Power and limitations of environmental DNA metabarcoding for surveying leaf litter eukaryotic communities. Environmental DNA 3, 528–540. https://doi.org/10.1002/edn3.142

Marčiulynienė D, Marčiulynas A, Lynikienė J, Vaičiukynė M, Gedminas A, Menkis A (2021) DNA-metabarcoding of belowground fungal communities in bare-root forest nurseries: focus on different tree species. Microorganisms 9, 150. https://doi.org/10.3390/microorganisms9010150

Mariani S, Baillie C, Colosimo G, Riesgo A (2019) Sponges as natural environmental DNA samplers. Curr. Biol. 29, R401–R402. https://doi.org/10.1016/j.cub.2019.04.031

Matesanz S, Pescador DS, Pías B, Sánchez AM, Chacón-Labella J, Illuminati A, de la Cruz M, López-Angulo J, Marí-Mena N, Vizcaíno A, Escudero A (2019) Estimating belowground plant abundance with DNA metabarcoding. Mol. Ecol. Resour. 19, 1265–1277. https://doi.org/10.1111/1755-0998.13049

Matsuhashi S, Doi H, Fujiwara A, Watanabe S, Minamoto T (2016) Evaluation of the environmental DNA method for estimating distribution and biomass of submerged aquatic plants. PLoS ONE 11, e0156217. https://doi.org/10.1371/journal.pone.0156217

Mauvisseau Q, Harper LR, Sander M, Hanner RH, Kleyer H, Deiner K (2022) The multiple states of environmental DNA and what is known about their persistence in aquatic environments. Environ. Sci. Technol. 56, 5322–5333. https://doi.org/10.1021/acs.est.1c07638

McFrederick QS, Rehan SM (2016) Characterization of pollen and bacterial community composition in brood provisions of a small carpenter bee. Mol. Ecol. 25, 2302–2311. https://doi.org/10.1111/mec.13608

Milberg P, Bergstedt J, Fridman J, Odell G, Westerberg L (2008) Observer bias and random variation in vegetation monitoring data. Journal of Vegetation Science 19, 633–644. https://doi.org/10.3170/2008-8-18423

Milla L, Sniderman K, Lines R, Mousavi-Derazmahalleh M, Encinas-Viso F (2021) Pollen DNA metabarcoding identifies regional provenance and high plant diversity in Australian honey. Ecol. Evol. 11, 8683–8698. https://doi.org/10.1002/ece3.7679

Montauban C, Mas M, Wangensteen OS, Sarto i Monteys V, Fornós DG, Mola XF, López-Baucells A (2021) Bats as natural samplers: first record of the invasive pest rice water weevil *Lissorhoptrus oryzophilus* in the Iberian Peninsula. Crop Prot. 141, 105427. https://doi.org/10.1016/j.cropro.2020.105427

Mori E, Mazza G, Galimberti A, Angiolini C, Bonari G (2017) The porcupine as "little thumbling": the role of *Hystrix cristata* in the spread of *Helianthus tuberosus*. Biologia 72, 1211–1216. https://doi.org/10.1515/biolog-2017-0136

Mucina L (2019) Biome: evolution of a crucial ecological and biogeographical concept. New Phytol. 222, 97–114. https://doi.org/10.1111/nph.15609

Nagler M, Insam H, Pietramellara G, Ascher-Jenull J (2018) Extracellular DNA in natural environments: features, relevance and applications. Appl. Microbiol. Biotechnol. 102, 6343–6356. https://doi.org/10.1007/s00253-018-9120-4

Núñez A, Amo de Paz G, Ferencova Z, Rastrojo A, Guantes R, García AM, Alcamí A, Gutiérrez-Bustillo AM, Moreno DA (2017) Validation of the Hirst-type spore trap for simultaneous monitoring of prokaryotic and eukaryotic biodiversities in urban air samples by next-generation sequencing. Appl. Environ. Microbiol. 83, e00472-17. https://doi.org/10.1128/AEM.00472-17

Núñez A, Amo de Paz G, Rastrojo A, Ferencova Z, Gutiérrez-Bustillo AM, Alcamí A, Moreno DA, Guantes R (2019) Temporal patterns of variability for prokaryotic and eukaryotic diversity in the urban air of Madrid (Spain). Atmos. Environ. 217, 116972. https://doi.org/10.1016/j.atmosenv.2019.116972

Ortega A, Geraldi NR, Alam I, Kamau AA, Acinas SG, Logares R, Gasol JM, Massana R, Krause-Jensen D, Duarte CM (2019) Important contribution of macroalgae to oceanic carbon sequestration. Nat. Geosci. 12, 748–754. https://doi.org/10.1038/s41561-019-0421-8

Ortega A, Geraldi NR, Díaz-Rúa R, Ørberg SB, Wesselmann M, Krause-Jensen D, Duarte CM (2020a) A DNA mini-barcode for marine macrophytes. Mol. Ecol. Resour. 20, 920–935. https://doi.org/10.1111/1755-0998.13164

Ortega A, Geraldi NR, Duarte CM (2020b) EnvironmentalDNA identifies marine macrophyte contributions to Blue Carbon sediments. Limnol. Oceanogr. 65, 3139–3149. https://doi.org/10.1002/lno.11579

Osathanunkul M, Sawongta N, Pheera W, Pechlivanis N, Psomopoulos F, Madesis P (2021) Exploring plant diversity through soil DNA in Thai national parks for influencing land reform and agriculture planning. PeerJ 9, e11753. https://doi.org/10.7717/peerj.11753

Palacios Mejia M, Curd E, Edalati K, Renshaw MA, Dunn R, Potter D, Fraga N, Moore J, Saiz J, Wayne R, Parker SS (2021) The utility of environmental DNA from sediment and water samples for recovery of observed plant and animal species from four Mojave Desert springs. Environmental DNA 3, 214–230. https://doi.org/10.1002/edn3.161

Parducci L, Bennett KD, Ficetola GF, Alsos IG, Suyama Y, Wood JR, Pedersen MW (2017) Ancient plant DNA in lake sediments. New Phytol. 214, 924–942. https://doi.org/10.1111/nph.14470

Parducci L, Jørgensen T, Tollefsrud MM, Elverland E, Alm T, Fontana SL, Bennett KD, Haile J, Matetovici I, Suyama Y, Edwards ME, Andersen K, Rasmussen M, Boessenkool S, Coissac E, Brochmann C, Taberlet P, Houmark-Nielsen M, Larsen NK, Orlando L, Willerslev E (2012) Glacial survival of boreal trees in northern Scandinavia. Science 335, 1083–1086. https://doi.org/10.1126/science.1216043

Pawlowski J, Apothéloz-Perret-Gentil L, Altermatt F (2020) Environmental DNA: What's behind the term? Clarifying the terminology and recommendations for its future use in biomonitoring. Mol. Ecol. 29, 4258–4264. https://doi.org/10.1111/mec.15643

Pedersen MW, Overballe-Petersen S, Ermini L, Sarkissian CD, Haile J, Hellstrom M, Spens J, Thomsen PF, Bohmann K, Cappellini E, Schnell IB, Wales NA, Carøe C, Campos PF, Schmidt AMZ, Gilbert MTP, Hansen AJ, Orlando L, Willerslev E (2015) Ancient and modern environmental DNA. Philos. Trans. R. Soc. Lond. B Biol. Sci. 370, 20130383. https://doi.org/10.1098/rstb.2013.0383

Pedersen MW, Ruter A, Schweger C, Friebe H, Staff RA, Kjeldsen KK, Mendoza MLZ, Beaudoin AB, Zutter C, Larsen NK, Potter BA, Nielsen R, Rainville RA, Orlando L, Meltzer DJ, Kjær KH, Willerslev E (2016) Postglacial viability and colonization in North America's ice-free corridor. Nature 537, 45–49. https://doi.org/10.1038/nature19085

Pemberton RW, Wheeler GS (2006) Orchid bees don't need orchids: evidence from the naturalization of an orchid bee in Florida. Ecology 87, 1995–2001. https://doi.org/10.1890/0012-9658(2006)87[1995:obdnoe]2.0.co;2

Pietramellara G, Ascher J, Borgogni F, Ceccherini MT, Guerri G, Nannipieri P (2009) Extracellular DNA in soil and sediment: fate and ecological relevance. Biol. Fertil. Soils 45, 219–235. https://doi.org/10.1007/s00374-008-0345-8

Polling M, Sin M, de Weger LA, Speksnijder AGCL, Koenders MJF, de Boer H, Gravendeel B (2022) DNA metabarcoding using nrITS2 provides highly qualitative and quantitative results for airborne pollen monitoring. Sci. Total Environ. 806, 150468. https://doi.org/10.1016/j.scitotenv.2021.150468

Polling M, ter Schure ATM, van Geel B, van Bokhoven T, Boessenkool S, MacKay G, Langeveld BW, Ariza M, van der Plicht H, Protopopov AV, Tikhonov A, de Boer H, Gravendeel B (2021) Multiproxy analysis of permafrost preserved faeces provides an unprecedented insight into the diets and habitats of extinct and extant megafauna. Quat. Sci. Rev. 267, 107084. https://doi.org/10.1016/j.quascirev.2021.107084

Prosser SWJ, Hebert PDN (2017) Rapid identification of the botanical and entomological sources of honey using DNA metabarcoding. Food Chem. 214, 183–191. https://doi.org/10.1016/j.foodchem.2016.07.077

Ramírez SR, Eltz T, Fujiwara MK, Gerlach G, Goldman-Huertas B, Tsutsui ND, Pierce NE (2011) Asynchronous diversification in a specialized plant-pollinator mutualism. Science 333, 1742–1746. https://doi.org/10.1126/science.1209175

Richardson RT, Lin C-H, Sponsler DB, Quijia JO, Goodell K, Johnson RM (2015) Application of ITS2 metabarcoding to determine the provenance of pollen collected by honey bees in an agroecosystem. Appl. Plant Sci. 3, 1400066. https://doi.org/10.3732/apps.1400066

Ritter CD, Zizka A, Roger F, Tuomisto H, Barnes C, Nilsson RH, Antonelli A (2018) High-throughput metabarcoding reveals the effect of physicochemical soil properties on soil and litter biodiversity and community turnover across Amazonia. PeerJ 6, e5661. https://doi.org/10.7717/peerj.5661

Rodriguez-Ezpeleta N, Morissette O, Bean CW, Manu S, Banerjee P, Lacoursière-Roussel A, Beng KC, Alter SE, Roger F, Holman LE, Stewart KA, Monaghan MT, Mauvisseau Q, Mirimin L, Wangensteen OS, Antognazza CM, Helyar SJ, de Boer H, Monchamp M-E, Nijland R, Deiner K (2021) Trade-offs between reducing complex terminology and producing accurate interpretations from environmental DNA: comment on "Environmental DNA: what's behind the term?" by Pawlowski et al., (2020). Mol. Ecol. 30, 4601–4605. https://doi.org/10.1111/mec.15942

Rodríguez-Ezpeleta N, Zinger L, Kinziger A, Bik HM, Bonin A, Coissac E, Emerson BC, Lopes CM, Pelletier TA, Taberlet P, Narum S (2021) Biodiversity monitoring using environmental DNA. Mol. Ecol. Resour. 21, 1405–1409. https://doi.org/10.1111/1755-0998.13399

Rowney FM, Brennan GL, Skjøth CA, Griffith GW, McInnes RN, Clewlow Y, Adams-Groom B, Barber A, de Vere N, Economou T, Hegarty M, Hanlon HM, Jones L, Kurganskiy A, Petch GM, Potter C, Rafiq AM, Warner A, PollerGEN Consortium Wheeler B, Creer S (2021) Environmental DNA reveals links between abundance and composition of airborne grass pollen and respiratory health. Curr. Biol. 31, 1995-2003.e4. https://doi.org/10.1016/j.cub.2021.02.019

Rucińska A, Świerszcz S, Nobis M, Zubek S, Boczkowska M, Olszak M, Kosiński JG, Nowak S, Nowak A (2022) Is it possible to understand a book missing a quarter of the letters? Unveiling the belowground species richness of grasslands. Agriculture, Ecosystems & Environment 324, 107683. https://doi.org/10.1016/j.agee.2021.107683

Schure ATM, Pillai AAS, Thorbek L, Bhavani Shankar M, Puri R, Ravikanth G, Boer HJ, Boessenkool S (2021) eDNA metabarcoding reveals dietary niche overlap among herbivores in an Indian wildlife sanctuary. Environmental DNA 3, 681–696. https://doi.org/10.1002/edn3.168

Scott WA, Hallam CJ (2003) Assessing species misidentification rates through quality assurance of vegetation monitoring. Plant Ecology 165, 101–115.

Scriver M, Marinich A, Wilson C, Freeland J (2015) Development of species-specific environmental DNA (eDNA) markers for invasive aquatic plants. Aquatic Botany 122, 27–31. https://doi.org/10.1016/j.aquabot.2015.01.003

Sepp S, Davison J, Moora M, Neuenkamp L, Oja J, Roslin T, Vasar M, Öpik M, Zobel M (2021) Woody encroachment in grassland elicits complex changes in the functional structure of above- and belowground biota. Ecosphere 12, e03512. https://doi.org/10.1002/ecs2.3512

Sherwood AR, Dittbern MN, Johnston ET, Conklin KY (2017) A metabarcoding comparison of windward and leeward airborne algal diversity across the Ko'olau mountain range on the island of O'ahu, Hawai'i1. J. Phycol. 53, 437–445. https://doi.org/10.1111/jpy.12502

Shogren AJ, Tank JL, Egan SP, August O, Rosi EJ, Hanrahan BR, Renshaw MA, Gantz CA, Bolster D (2018) Water flow and biofilm cover influence environmental DNA detection in recirculating streams. Environ. Sci. Technol. 52, 8530–8537. https://doi.org/10.1021/acs.est.8b01822

Siegenthaler A, Wangensteen OS, Soto AZ, Benvenuto C, Corrigan L, Mariani S (2019) Metabarcoding of shrimp stomach content: Harnessing a natural sampler for fish biodiversity monitoring. Mol. Ecol. Resour. 19, 206–220. https://doi.org/10.1111/1755-0998.12956

Smart MD, Cornman RS, Iwanowicz DD, McDermott-Kubeczko M, Pettis JS, Spivak MS, Otto CRV (2017) A comparison of honey bee-collected pollen from working agricultural lands using light microscopy and ITS metabarcoding. Environ. Entomol. 46, 38–49. https://doi.org/10.1093/ee/nvw159

Sønstebø JH, Gielly L, Brysting AK, Elven R, Edwards M, Haile J, Willerslev E, Coissac E, Rioux D, Sannier J, Taberlet P, Brochmann C (2010) Using next-generation sequencing for molecular reconstruction of past Arctic vegetation and climate. Mol. Ecol. Resour. 10, 1009–1018. https://doi.org/10.1111/j.1755-0998.2010.02855.x

Sookhan N, Lorenzo A, Tatsumi S, Yuen M, MacIvor JS (2021) Linking bacterial diversity to floral identity in the bumble bee pollen basket. Environmental DNA 3, 669–680. https://doi.org/10.1002/edn3.165

Steinbauer MJ, Grytnes J-A, Jurasinski G, Kulonen A, Lenoir J, Pauli H, Rixen C, Winkler M, Bardy-Durchhalter M, Barni E, Bjorkman AD, Breiner FT, Burg S, Czortek P, Dawes MA, Delimat A, Dullinger S, Erschbamer B, Felde VA, Fernández-Arberas O, Wipf S (2018) Accelerated increase in plant species richness on mountain summits is linked to warming. Nature 556, 231–234. https://doi.org/10.1038/s41586-018-0005-6

Stoeck T, Pan H, Dully V, Forster D, Jung T (2018) Towards an eDNA metabarcode-based performance indicator for full-scale municipal wastewater treatment plants. Water Res. 144, 322–331. https://doi.org/10.1016/j.watres.2018.07.051

Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. Mol. Ecol. 21, 2045–2050. https://doi.org/10.1111/j.1365-294X.2012.05470.x

Taberlet P, Coissac E, Pompanon F, Gielly L, Miquel C, Valentini A, Vermat T, Corthier G, Brochmann C, Willerslev E (2007) Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding. Nucleic Acids Res. 35, e14. https://doi.org/10.1093/nar/gkl938

Terwayet Bayouli I, Terwayet Bayouli H, Dell'Oca A, Meers E, Sun J (2021) Ecological indicators and bioindicator plant species for biomonitoring industrial pollution: eco-based environmental assessment. Ecological Indicators 125, 107508. https://doi.org/10.1016/j.ecolind.2021.107508

Thies W, Kalko EKV (2004) Phenology of neotropical pepper plants (Piperaceae) and their association with their main dispersers, two short-tailed fruit bats, *Carollia perspicillata* and *C. castanea* (Phyllostomidae). Oikos 104, 362–376. https://doi.org/10.1111/j.0030-1299.2004.12747.x

Thomsen PF, Sigsgaard EE (2019) Environmental DNA metabarcoding of wild flowers reveals diverse communities of terrestrial arthropods. Ecol. Evol. 9, 1665–1679. https://doi.org/10.1002/ece3.4809

Thomsen PF, Willerslev E (2015) Environmental DNA – An emerging tool in conservation for monitoring past and present biodiversity. Biological Conservation 183, 4–18. https://doi.org/10.1016/j.biocon.2014.11.019

Tsukamoto Y, Yonezawa S, Katayama N, Isagi Y (2021) Detection of endangered aquatic plants in rapid streams using environmental DNA. Front. Ecol. Evol. 8. https://doi.org/10.3389/fevo.2020.622291

Turner CR, Barnes MA, Xu CCY, Jones SE, Jerde CL, Lodge DM (2014) Particle size distribution and optimal capture of aqueous macrobial eDNA. Methods Ecol. Evol. 5, 676–684. https://doi.org/10.1111/2041-210X.12206

Turon M, Angulo-Preckler C, Antich A, Præbel K, Wangensteen OS (2020) More than expected from old sponge samples: a natural sampler DNA metabarcoding assessment of marine fish diversity in Nha Trang bay (Vietnam). Front. Mar. Sci. 7, 605148. https://doi.org/10.3389/fmars.2020.605148

Turon X, Antich A, Palacín C, Praebel K, Wangensteen OS (2020) From metabarcoding to metaphylogeography: separating the wheat from the chaff. Ecol. Appl. 30, e02036. https://doi.org/10.1002/eap.2036

Uuemaa E, Mander Ü, Marja R (2013) Trends in the use of landscape spatial metrics as landscape indicators: a review. Ecological Indicators 28, 100–106. https://doi.org/10.1016/j.ecolind.2012.07.018

Valentini A, Miquel C, Nawaz MA, Bellemain E, Coissac E, Pompanon F, Gielly L, Cruaud C, Nascetti G, Wincker P, Swenson JE, Taberlet P (2009) New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the trnL approach. Mol. Ecol. Resour. 9, 51–60. https://doi.org/10.1111/j.1755-0998.2008.02352.x

Valentini A (2007) Non-invasive diet analysis based on DNA barcoding: the Himalayan brown bears (*Ursus arctos isabellinus*) as a case study (Doctoral dissertation).

Varotto C, Pindo M, Bertoni E, Casarotto C, Camin F, Girardi M, Maggi V, Cristofori A (2021) A pilot study of eDNA metabarcoding to estimate plant biodiversity by an alpine glacier core (Adamello glacier, North Italy). Sci. Rep. 11, 1208. https://doi.org/10.1038/s41598-020-79738-5

Vasconcelos S, Nunes GL, Dias MC, Lorena J, Oliveira RRM, Lima TGL, Pires ES, Valadares RBS, Alves R, Watanabe MTC, Zappi DC, Hiura AL, Pastore M, Vasconcelos LV, Mota NFO, Viana PL, Gil ASB, Simões AO, Imperatriz-Fonseca VL, Harley RM, Oliveira G (2021) Unraveling the plant diversity of the Amazonian canga through DNA barcoding. Ecol. Evol. 11, 13348–13362. https://doi.org/10.1002/ece3.8057

Willerslev E, Cappellini E, Boomsma W, Nielsen R, Hebsgaard MB, Brand TB, Hofreiter M, Bunce M, Poinar HN, Dahl-Jensen D, Johnsen S, Steffensen JP, Bennike O, Schwenninger J-L, Nathan R, Armitage S, de Hoog C-J, Alfimov V, Christl M, Beer J, Collins MJ (2007) Ancient biomolecules from deep ice cores reveal a forested southern Greenland. Science 317, 111–114. https://doi.org/10.1126/science.1141758

Willerslev E, Davison J, Moora M, Zobel M, Coissac E, Edwards ME, Lorenzen ED, Vestergård M, Gussarova G, Haile J, Craine J, Gielly L, Boessenkool S, Epp LS, Pearman PB, Cheddadi R, Murray D, Bråthen KA, Yoccoz N, Binney H, Taberlet P (2014) Fifty thousand years of Arctic vegetation and megafaunal diet. Nature 506, 47–51. https://doi.org/10.1038/nature12921

Willerslev E, Hansen AJ, Binladen J, Brand TB, Gilbert MTP, Shapiro B, Bunce M, Wiuf C, Gilichinsky DA, Cooper A (2003) Diverse plant and animal genetic records from Holocene and Pleistocene sediments. Science 300, 791–795. https://doi.org/10.1126/science.1084114

Wood JM (1987) Biological processes involved in the cycling of elements between soil or sediments and the aqueous environment. Hydrobiologia 149, 31–42. https://doi.org/10.1007/BF00048644

Yamamoto S, Uchida K (2018) A generalist herbivore requires a wide array of plant species to maintain its populations. Biological Conservation 228, 167–174. https://doi.org/10.1016/j.biocon.2018.10.018

Yang Chenxue Wang X, Miller JA, de Blécourt M, Ji Y, Yang Chunyan Harrison RD, Yu DW (2014) Using metabarcoding to ask if easily collected soil and leaf-litter samples can be used as a general biodiversity indicator. Ecological Indicators 46, 379–389. https://doi.org/10.1016/j.ecolind.2014.06.028

Yoccoz NG, Bråthen KA, Gielly L, Haile J, Edwards ME, Goslar T, Von Stedingk H, Brysting AK, Coissac E, Pompanon F, Sønstebø JH, Miquel C, Valentini A, De Bello F, Chave J, Thuiller W, Wincker P, Cruaud C, Gavory F, Rasmussen M, Taberlet P (2012) DNA from soil mirrors plant taxonomic and growth form diversity. Mol. Ecol. 21, 3647–3655. https://doi.org/10.1111/j.1365-294X.2012.05545.x

Yoccoz NG (2012) The future of environmental DNA in ecology. Mol. Ecol. 21, 2031–2038. https://doi.org/10.1111/j.1365-294X.2012.05505.x

Zinger L, Taberlet P, Schimann H, Bonin A, Boyer F, De Barba M, Gaucher P, Gielly L, Giguet-Covex C, Iribar A, Réjou-Méchain M, Rayé G, Rioux D, Schilling V, Tymen B, Viers J, Zouiten C, Thuiller W, Coissac E, Chave J (2019) Body size determines soil community assembly in a tropical forest. Mol. Ecol. 28, 528–543. https://doi.org/10.1111/mec.14919

Answers

1. The analysis of eDNA from gut contents, faeces, or eDNA traces from the butterfly's body (vegetation fragments or pollen grains) would reveal the floral resources available and visited. To reveal other organisms that are using the same floral resources (other pollinators competitors), one could target insect eDNA present in flowers that have been visited.
2. eDNA water samples from near-shore sites would optimise the vegetation assessment as they are both easy to collect and signal terrestrial and aquatic diversity. Though airborne DNA could be also considered for this purpose, it may miss dormant DNA or non wind-dispersed plants. In addition, sedimentary eDNA may also signal nearby diversity.
3. Spruce and pine spores are tiny, light, and spread by wind. These have a tendency to show up anywhere, and are not a good indication for local presence. Invasive species monitoring needs approaches that provide a clear link between detected species and specific environments.

Chapter 25
Wildlife trade

Mehrdad Jahanbanifard[1,2], Margaretha A. Veltman[3], Sarina Veldman[1], Ida Hartvig[4], Carly Cowell[5], Frederic Lens[1], Steven Janssens[6], Erik Smets[1]

1 Naturalis Biodiversity Center, Leiden, The Netherlands
2 Leiden Institute of Advanced Computer Science (LIACS), Leiden University, Leiden, The Netherlands
3 Natural History Museum, University of Oslo, Oslo, Norway
4 University of Copenhagen, Copenhagen, Denmark
5 Royal Botanical Garden, Kew, United Kingdom
6 Meise Botanic Garden, Meise, Belgium

Mehrdad Jahanbanifard **mehrdad.jahanbanifard@naturalis.nl**
Margaretha A. Veltman **margret.veltman@nhm.uio.no**
Sarina Veldman **sarina.veldman@naturalis.nl**
Ida Hartvig **ihla@ign.ku.dk**
Carly Cowell **c.cowell@kew.org**
Frederic Lens **frederic.lens@naturalis.nl**
Steven Janssens **steven.janssens@plantentuinmeise.be**
Erik Smets **erik.smets@naturalis.nl**

Citation: Jahanbanifard M, Veltman MA, Veldman S, Hartvig I, Cowell C, Lens F, Janssens S, Smets E (2022) Chapter 25. Wildlife trade. In: de Boer H, Rydmark MO, Verstraete B, Gravendeel B (Eds) Molecular identification of plants: from sequence to species. Advanced Books. https://doi.org/10.3897/ab.e98875

Introduction

Wildlife trade and its effects

Wildlife trade is the trading of living or dead wild plants, fungi, or animals, either as whole organisms or as parts and the products derived from them. This varies from rare animal and plant species for collectors, to ingredients made of wild organisms for medicinal or cosmetic purposes, to wood for timber, paper, craftwork, and construction, and various animals, plants, and mushrooms for nutritional purposes. Although conservation concerns about the unsustainable use of wildlife became more prominent from the 1960s onward, evidence shows that large-scale wildlife trade is older than the Roman Empire and ancient Greek civilisations ('t Sas-Rolfes et al. 2019). International wildlife trade is a billion-dollar industry, and together with illegal wildlife trafficking, it has become a substantial threat to global biodiversity and the preservation of endangered species (Smith et al. 2017). In addition, the overall impact of wildlife trade on national economies as well as public health is largely underestimated (Kurland et al. 2017; Rosen and Smith 2010).

The impacts of wildlife trade are substantial with both conservation and socio-economic importance. Unsustainable trade could lead to (local) extinction of populations or even entire species. For plants that occupy a specialised niche, it can destabilise interactions with other species, with potential consequences for the entire ecosystem. Therefore, after habitat loss, wildlife trade is the second-biggest threat to species survival (WWF, 2020). Not only does illegal wildlife trade threaten biodiversity due to consistent overexploitation, it also competes with legal use of natural resources and results in a substantial loss of income for both local communities and governments (Cooney et al. 2015). Many source countries rely on the products and/or income generated from wildlife trade, meaning that the livelihoods of the people that depend on it would be compromised if these species go extinct or if trade would be banned. In some areas in Tanzania, for example, illegal chikanda orchid gathering is the primary economic activity for vulnerable HIV/AIDS-affected households (Challe and Price 2009), although resellers further down the supply chain actually profit the most from this trade (Veldman et al. 2014). The best-known examples of wildlife trade in plants can be found in timber commerce (e.g., rosewood and ebony wood), for which the legal market has an annual value of around $200 billion and the illegal market an estimated annual $30–$157 billion (Jenkins et al. 2018; World Bank 2019). Furthermore, it is estimated that 60–90% of medicinal and aromatic plants are harvested from the wild, among which several high-value species, such as sandalwood (*Santalum* spp.), agarwood (*Aquilaria* spp.), African cherry (*Prunus africana*), and American and Chinese ginseng (*Panax* spp.) (Jenkins et al. 2018). Moreover, several groups of plants are traded for ornamental purposes, including species from threatened taxa such as cycads, cacti, aloes, conifers, euphorbs, and orchids. An overview of the global hotspots for wildlife trade, with some examples of plant groups targeted, is given in Figure 1.

Regulating wildlife trade

In order to regulate the trade in vulnerable wildlife, the Convention on International Trade of Endangered Species of Wild Fauna and Flora (CITES) was established in 1975. Species at risk of overexploitation due to international trade are listed on one of three appendices depending on how much they are threatened by unrestricted trade. Appendix I lists the most endangered species, for which commercial trade is not permitted - except for pre-convention material - and

for which non-commercial trade is strictly regulated. Appendix II lists the species that may become extinct if trade is not carefully controlled, which therefore requires a proper permit. Finally, Appendix III lists species that are protected in at least one country and other CITES Parties assistance is required to control the trade. Listing species on Appendix III helps to establish international cooperation in order to control trade in the species according to the laws and regulations of that country. Species can be added to Appendix I and II or removed from them, or shifted from Appendix I to II and vice versa only by voting at a Conference of the Parties (CoP), which is a meeting of the CITES Parties to review the implementation of the Convention. Species can be added to Appendix III or removed from it at any time and by any Party unilaterally (CITES, n.d.).

At the moment, roughly 39,000 species, including ca. 6000 species of animals and ca. 33,000 species of plants (395 species in Appendix I, 32,364 species in Appendix II, and 9 species in Appendix III) are protected by CITES (CITES, n.d.). In countries that are signatories to the convention, import and export permits must be issued for international trade of plants and animals listed in these appendices. Some countries set annual export quotas for certain species to ensure that they will not be traded beyond the sustainable limits for species survival. Non-compliance with CITES regulations can lead to confiscation of the material as well as fines and prison sentences, and in some cases trade sanctions against a country (CITES, n.d.). Since 2017, CITES has also facilitated the Wildlife Cybercrime Working Group that has coordinated national responses to the threat posed by online trade (Sajeva et al. 2013).

Other international and national regulations have been put into place to support the implementation of and in some cases expand on CITES regulations. Examples are the EU Action Plan Against Wildlife Trafficking (European Commission 2016), the EU Wildlife Trade Regulations (European Commission 2010), European Union Timber Regulation (EUTR), United States LEMIS wildlife trade data (Eskew et al. 2020), and the United States Lacey Act (Anderson 1995). Under the National Legislation Project (NLP), various domestic measures need to be implemented in order to meet the four CITES criteria, without which the CITES regulations are not in force at the national level: countries need to designate at least one Management Authority and one Scientific Authority; prohibit trade in specimens in violation of the Convention; penalise such trade; or confiscate specimens illegally traded or possessed. Diverse governmental and non-governmental programmes exist that implement enforcement in source, transit, and consumer countries, and are used to increase the risks of being involved in illegal wildlife trade as well as to decrease the rewards. In terms of global law enforcement, INTERPOL examines websites and social media posts offering wildlife products for sale. This happens annually and a number of seizures and arrests take place every year.

Challenges in combating wildlife trade

Despite the fact that plant species far outnumber animal species on the CITES appendices, in the public discourse on wildlife trade and conservation, charismatic mammals such as elephants, rhinos, tigers, and lions usually take centre stage. Smaller animals (e.g., insects, molluscs), but also most plant groups, receive less attention and generate less funding in discussions regarding wildlife trade and conservation. And although plants appear frequently in national and international regulations, regulatory enforcement and additional conservation measures still primarily target iconic megafauna (Margulies et al. 2019). The relative 'invisibility' of plants as organisms of importance for our lives and worthy of conservation is called "plant blindness", and is one of the biggest challenges in combating illegal plant trade (Box 1).

Chapter 25: Box 1. Example of a challenge in depth: plant blindness

Plant blindness is a psychological bias that leads us to notice (large) animals, and take plants largely for granted, reducing them to background vegetation for other organisms. The term was coined by Wandersee and Schussler (1999) and refers to a number of common problems in the perception of plants: not noticing plants in one's environment; ignoring plants' aesthetic and unique biological features; not recognising the importance of plants (e.g., food production, absorbing carbon dioxide and releasing oxygen, etc.); and considering plants as inferior to animals. Plant blindness has both a physical and a psychological component. The human eye picks-up the colour green more easily than other colours, and hence does not focus on it quite as much (Knapp 2019). Green is also experienced as safe and therefore warrants limited attention. Furthermore, our eyes perceive movements more readily than static objects, which probably stems from an evolutionary function in spotting (attacking) predators and (fleeing) prey.

Plant blindness has been institutionalised throughout society, from (higher) education to governance and wildlife management (Margulies et al. 2019; Wandersee and Schussler 1999), leading to a focus on animals in biology courses, natural history museums, research funding, and conservation policies. Plant blindness is therefore one of the biggest challenges in combating illegal wildlife trade.

Apart from the limited attention that plants receive in research, education, and conservation, effective control of trade in plant species is hampered because some of the traded goods are difficult to recognise, either because they are processed or because they contain only parts of the organism, which lack the morphological characters needed for identification (Lavorgna et al. 2018). Plant products are therefore often harder to identify than living animals or animal parts, and to identify them routinely requires standardised and scalable technologies, many of which are still being developed (for more details, see Methods).

Other challenges are posed by the growing use of the internet for transactions, which makes wildlife material more readily accessible and at lower costs, while preserving anonymity. The internet is not only increasingly used to sell and obtain specimens, but even to organise poaching events (Lavorgna 2014). Rare and exotic plant species can be ordered with ease from a range of online retailers, shipping of plants in the postal system is relatively easy and the search for plant material in these systems is limited. In addition, the scale of the internet and speed at which online marketplaces proliferate make the monitoring of online criminal activities costly and time consuming (Lavorgna et al. 2020; TRAFFIC, 2019). The online market thus facilitates participation in illegal wildlife trade, making it more attractive due to potentially high sales and profits and reduced detection rate (TRAFFIC, 2019). The challenges for curbing illegal online trade are therefore manifold, and only exacerbate existing challenges with law enforcement by enabling covert activities and thereby increasing the volume of illegally traded goods. Distinguishing legal from illegal trade is difficult even with specialist knowledge or extensive training (Vaglica et al. 2017). Mixing legal and illegal shipments, nontransparent supply chains and lack of institutional monitoring capacity in biodiversity rich countries are some of the practical challenges underpinning this difficulty (Engler and Parry-Jones 2007). International conventions such as CITES can also have unintended loopholes that allow wildlife traffickers to circumvent restrictions or to present their information in a way that gives the impression of legal trade. For example, newly discovered rare species that have not yet made their way onto one of the CITES appendices can often be traded freely, despite detrimental effects, if there is no national legislation in place to protect the species. Another commonly observed practice is the export of wild harvested or poached wildlife as captive bred (in the case of animals) or artificially propagated

(in the case of plants) organisms. Verification of legal acquisition can be challenging without sufficient documentation, opening up space for laundering of illegally obtained specimens.

Lastly, since international wildlife trade per definition transcends borders, enforcement of legal trade requires coordinated action between multiple countries to address the whole supply chain. While there are already many institutional collaborations that work across international borders to help track and catch illegal wildlife trafficking syndicates - including financial institutions, NGOs, customs and police forces and online tech platforms - one of the main bottlenecks to combating wildlife trade will be to sustain sufficient international attention to allow the detection and prevention, not just of single illegal transactions, but of organised trade networks operating at larger scales.

The importance of wildlife and the impacts of unsustainable trade on biodiversity are undeniable, which highlights the urgency of developing high-throughput methods that are widely applicable. The next section presents some of the most commonly used methods in illegal trade identification today. In the final section, we provide recommendations on which techniques to use for the identification and tracking of illegally traded plants, and discuss future developments that could improve global wildlife trade monitoring and control.

Methods for identification of plants in trade

Traded plant materials come in all shapes and sizes and in different stages of processing, ranging from complete living plants to raw timber logs and to engineered wood products. There is a wide variety of molecular and non-molecular methods for illegal wildlife trade monitoring, from DNA (meta) barcoding and genetic methods, to chemical identification, and computer vision and pattern recognition tools. Each of these methods is applicable to certain types of materials and requires knowledge about different aspects of the traded product that determines its legality, including species identity, geographic origin, source population (wild or cultivated), and the sample age. Here we describe the most commonly used methods to identify each of these aspects, and why they are important.

Species identity

Methods for species identification are used to ascertain whether the organism being traded is CITES-listed or not. Depending on the taxonomic rank that is listed, it may be necessary to identify the exact species (e.g., *Panax ginseng*), genus (e.g., *Aloe* spp.), or family (e.g., Orchidaceae) to which an organism belongs. Species identification methods include genetic based methods (based on DNA sequencing information), chemical methods (based on molecular mass spectra), and computational methods (based on image recognition). Each of these methods require suitable reference data against which to query an unknown sample. The availability of reference data and the nature of the sample will dictate which method is most suitable for species identification.

Mass spectrometry

The main chemical method used to identify species is Direct Analysis in Real Time (DART) coupled with time-of-flight (TOF) mass spectrometry (DART-TOF MS). DART-TOF MS consists of two parts: DART is an ionisation source that ionises ambient atmospheric molecules by using electronically excited-state helium which reacts with the molecules in the investigated sample to produce analyte ions (Gross 2014). These ions are then sucked into the AccuTOF mass spec-

trometer. Spectral data on molecular masses and their relative intensities (so called chemical fingerprint) can be analysed to identify timbers (Deklerck et al. 2020; Evans et al. 2017; Lancaster and Espinoza 2012), keratin fibres of camelids (Price et al. 2020), rhinoceros keratin (Price et al. 2018), explosives (Lennert and Bridge 2018), and narcotics (Lian et al. 2017). DART-TOF MS is fast and has a simple sample preparation procedure. The accuracy of the result is however dependent on the reference database - as is the case for all other species identification methods - and whether the investigated samples have enough variation in molecular composition to be distinguished with their chemotype (Deklerck et al. 2017).

Computer vision and pattern recognition

Thanks to machine learning and computer vision, expert systems are playing an increasingly important role in identification of a wide variety of wildlife related objects, such as medicinal leaves (Sabu et al. 2017), herbarium specimens (Lorieul et al. 2019; Pearson et al. 2020), wood identification (Lens et al. 2020), mulberry ripeness detection (Ashtiani et al. 2021), pollen grains (Polling et al. 2021), corn seed varieties detection (Javanmardi et al. 2021) and wildlife monitoring (Di Minin et al. 2019, 2018). The concept of this method is pretty simple: train a model using a reliable database (usually an image database) to recognise specific objects such as humans, cars, trees, etc, in an image that the model has not seen before. Not only images (e.g., light microscopic images) can be used as input data, but also Near infrared (NIR) spectroscopy and X-ray micro computed tomography (CT) data can be used for automated pattern recognition. These are nondestructive alternative methods that can be useful when the conventional methods (such as light microscopy or DNA-based methods) are not acceptable or difficult to use, as is often the case in the investigation of registered cultural objects (Kobayashi et al. 2019). The main advantage of using computer vision methods is that it is accurate and applicable on a wide range of materials, such as wood, leaves, flowers, and pollen grains. The main drawback of computer vision, apart from a general lack of reliable databases, is the insufficient resolution of many morphological traits for species recognition, especially amongst closely related species. In some cases, better algorithms, more powerful machines, and high-quality reference databases can mitigate this challenge. However, in the cases where morphological traits do not provide distinctive features, pattern recognition cannot be used.

DNA barcoding and metabarcoding

DNA-based identification methods can use different genomic markers that offer different levels of identification, from universal loci such as conserved genes or intergenic spacers, to neutrally evolving markers with sufficient variation to resolve specific taxa, such as microsatellites and genome-wide Single Nucleotide Polymorphisms (SNPs). In addition to these markers, which require information about genomic context, it is also possible to identify species and populations using alignment-free shotgun data (see Chapter 17 Species delimitation).

For species identification, DNA barcoding (see Chapter 10 DNA barcoding) is often the method of choice. It can effectively identify traded plant species in a number of cases, including the identification of rosewood (*Dalbergia* spp.), species used in Ayurvedic medicine (*Decalepis* spp.), and cycads (*Encephalartos* spp.) (Hartvig et al. 2015; Mishra et al. 2017; Williamson et al. 2016). In addition, DNA metabarcoding (see Chapter 11 Amplicon metabarcoding) detects multiple species in mixed products such as traditional medicine and processed foods (Arulandhu et al. 2017; Veldman et al. 2017). An advantage of DNA barcoding is that, for the core land plant barcodes such as *rbcL*, *matK*, and nrITS, reference data is readily and freely available in public databases such as NCBI's GenBank or BOLD (barcodinglife.org). Tropical species are generally under-represented in these databases, and NCBI GenBank is known to contain er-

roneous sequences due to limited quality control. Species-level discrimination using standard barcodes has proven to be difficult among closely related and hybridising species, as well as taxa with low rates of evolution (Hassold et al. 2016; Veldman et al. 2017). An alternative in these cases is to develop custom barcodes. This provides researchers with more control over choosing genomic features that are informative for their plant group, but requires generating novel reference data, raising both the financial costs and time investment.

Source population and geographic origin

Neutral genetic markers

An advantage of DNA barcoding is that the sequence data is universally comparable among labs and large numbers of species. But since DNA barcoding was originally meant to distinguish between species and not within species, this method often falls short when higher resolution is needed. Identification below the species level may be useful if the legality of trade is determined by the source population. In some cases, the country of origin determines the legal status of traded plants, which requires population level data for a collection of reference samples spanning the species range. Cost-effective traditional population genetic methods use a number of species-specific variable markers, typically simple sequence repeats (SSRs) or inter simple sequence repeats (ISSRs), which can be highly variable and show fine-grained population structure. More recently developed high-throughput sequencing methods cover larger sections of the genome, such as reduced representation sequencing methods (RAD-seq, target capture, or low coverage whole genome shotgun sequencing (also known as genome skimming, see Chapter 16 Whole genome sequencing).

These methods can generate large numbers of SNPs that allow inference of geographic origins at various scales. Although the increased costs for library preparation and sequencing means that these methods are not economically feasible in all cases, they offer the added advantage that functional analyses of genes or markers linked to genes with adaptive significance is possible.

Geographic origins have even been identified at the level of continents using genome skimming (Schroeder et al. 2016), at the level of countries with SNPs generated by target enrichment of nuclear loci (Manzanilla et al. 2022) and RAD-seq (Blanc-Jolivet et al. 2017; Pakull et al. 2020), and even at the level of individual forest concessions with microsatellites (Vlam et al. 2018). Population genetic methods could potentially also be useful in detecting laundering of illegally harvested plants that are claimed to be cultivated. Genetic diversity analysis of the same neutral markers that are used to infer geographic origin, could then point out whether the plants were indeed sourced from a particular plantation or rather from the wild - in which case their genetic composition would be much more diverse than expected from artificially propagated material.

Stable isotope analysis

While population genetic markers can offer unmatched resolution of spatial variation, a general disadvantage is that many of them (with the exception of those used in RAD-seq and shotgun sequencing) need to be tested or developed specifically for each species, and reference data must be generated for populations across the distribution range to be tested. Stable isotope analysis can also infer geographic origin of samples, and does not depend on species-specific reference data to the same extent as genetic methods do. Stable isotope analysis is based on the principle that the presence of stable isotopes in the environment depends on both climate and geography. This creates a correlation between the stable isotope profile and its geographic location (Hermes et al. 2018). Since plants generally incorporate the stable isotopes into their

tissue at the same ratios as they occur in their environment, stable isotope analysis of plant material can be used to infer its geographic origin and be a tool in wildlife forensics (Matos and Jackson 2019). Stable isotope analysis however does not have a geographic resolution as high as population genetic methods have (Gori et al. 2015; Horacek et al. 2009). Georeferenced data is also required for stable isotope analysis, and global isotope databases are currently not freely available yet (Camin et al. 2017), limiting broad application of this method.

Harvesting pre- or post CITES legislation

Radiocarbon dating

There are two methods to measure radiocarbon abundance: radiometric dating and accelerator mass spectrometry (AMS). These methods can be used to date samples based on the decay of carbon isotopes. The estimated age gives an indication of whether or not the traded sample is a pre-convention material, meaning that the traded material predates the convention or listing of the species (e.g., Kalt-O'Bannon 1994; Uno et al. 2013; Cerling et al. 2016). While both radiometric dating and AMS provide high quality results, they are fundamentally different. AMS quantifies the number of carbon 14 (^{14}C) atoms in the investigated samples, while radiometric dating methods are based on the detection of beta particles resulting from the ^{14}C decay. AMS requires a much smaller sample size (20–500 mg) compared with radiometric methods (10–100 g). AMS is also faster and usually gains higher precision results than radiometric methods. Samples can be analysed in a few hours with AMS, while it can take one or two days with radiometric methods.

Recommendations to improve wildlife trade monitoring

Currently, no genetic methods for inferring sample age can compete with radiocarbon dating, and while DNA fragment sizes tend to be shorter for older and more degraded plant tissues, this alone cannot be used to determine the plant age (see Chapter 2 DNA from museum collections). For other purposes, genetic markers are the method of choice to infer species identity and geographic origin, whenever DNA extraction is a realistic option. Any genetic method will however be limited by the quality and quantity of DNA that can be extracted, which can be notoriously difficult for some materials, especially timber and processed products (Jiao et al. 2020; Lo and Shaw 2018). The obtained DNA quality and quantity will influence the range of techniques that can be applied downstream. High-copy regions such as chloroplast markers or nuclear ITS, for example, are easier to retrieve from samples with highly degraded DNA than low copy nuclear markers. For applications that require broader genomic coverage, amplification of low copy nuclear target regions can be achieved even with highly fragmented DNA, making target capture preferable over untargeted RAD-seq or genome-wide shotgun sequencing for degraded samples. However, for fresher material RAD-seq or WGS libraries may be easier to prepare and require less time for the bioinformatic analyses needed to develop markers prior to sequencing.

Despite significant progress in methods and computational analyses, applications for most methods are still limited by the lack or incompleteness of suitable reference data. As shown in Table 1, reference databases are currently under development or need further development for nearly all the methods currently used. The ForeST database for CITES protected timbers, the U.S. Fish & Wildlife Service Forensics Laboratory (Ashland, Oregon, USA), CITESwoodID by the

Thünen Institute (Hamburg, Germany), and the ebony wood microscopic database (Jahanbanifard et al. 2020, 2019) are examples of ongoing projects that are developing databases for the identification of CITES protected species.

When one method lacks sufficient reference data or is not sensitive enough to infer species identity or population of origin, multiple identification techniques tools (e.g., DNA barcoding, machine learning, and DART-TOF MS) can be combined to improve identification accuracy. Developing an integrated identification framework, which links reference databases and connects multiple sources of data for taxa of interest, is expected to play a major role in the future of regulating wildlife trade, though this would rely on standardisation and equitable distribution to enforcement agencies around the world. Coupled with new technologies that ensure quality control and compliance across the supply chain of wildlife products, the tools available for wildlife trade monitoring can aid not just the detection and confiscation of illegally traded goods, but also the transparency and traceability of legally traded commodities.

With blockchain for example, it may eventually be possible to develop a secure and robust infrastructure to register and track wildlife-related products from source to destination (Chang et al. 2020; Pournader et al. 2020). A blockchain is a database, consisting of several distributed nodes called blocks that are connected to one another using cryptography. Each block contains a cryptographic hash of the previous block, a timestamp, and transaction data (Narayanan et al. 2016). Blockchain provides an immutable and decentralised network which increases its reliability and security as no single party has full control of the system and no one can manipulate the transactions (Aimin and Yunfeng 2019; Saurabh and Dey 2021; Zheng et al. 2020).

The technology has already proven its relevance in agriculture and fisheries, where the WWF Blockchain Tuna Project demonstrates it is possible to track the history of a fishing product from ocean to plate with just a QR Code (WWF, 2018). The customisable and scalable features of blockchain make it a promising technology for application to traded timber and other wildlife-related products (MoonX, 2019). Once it is possible to keep track of all steps taken throughout the commercialisation of wild harvested plants, the checkpoints for identification will no longer be restricted to points of entry or sales, enabling monitoring of wildlife trade from the source.

Table 1. A comparison of the methods used for identifying plants in trade with an indication of their applications and limitations.

	DNA (meta) barcoding	Population genetic markers	Computer vision and pattern recognition	DART-TOF MS	AMS/ ^{14}C dating	Stable isotope
Material input	Whole plants, organs, tissues, powder	Whole plants, organs, tissues, powder	Timber, leaves, flowers, pollen	All	Anything containing organic matter	Anything containing organic matter
Purpose of application	Determine taxonomic identity from genus to species level	Determine population or region of origin	Determine taxonomic identify, from genus to (sometimes) species level	Determine taxonomic identity at species level	Determine age of material	Determine the region of origin
Availability of reference data	Well-developed for temperate species, less for tropical species and regions	Needs to be developed and referenced for each species separately	Being developed for CITES protected timber and plants	Being developed for CITES protected timber	Calibration might be required depending on the sample	Needs to be developed for each region separately

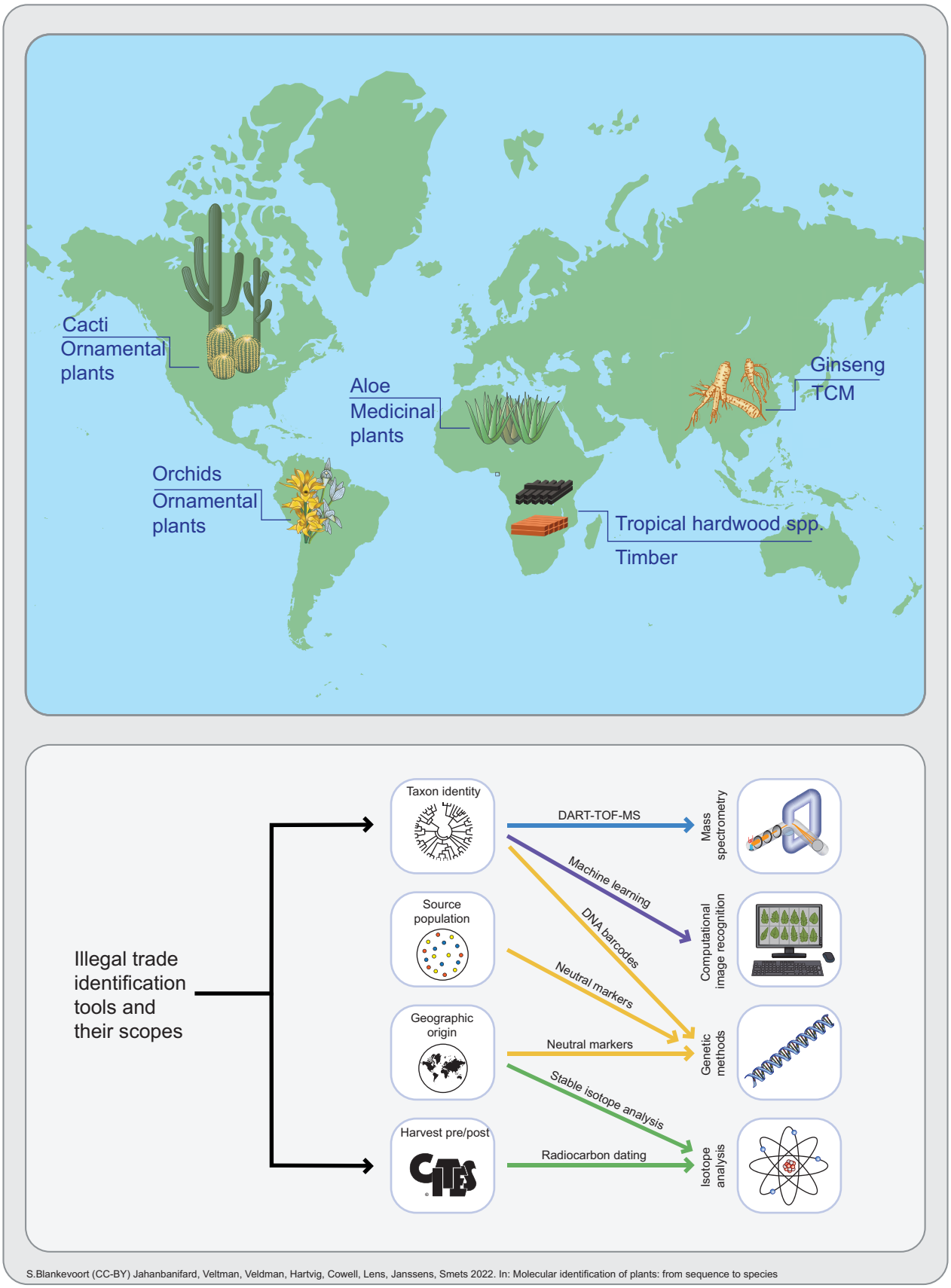

Figure 1. Chapter 25 Infographic: Global wildlife trade hotspots and some examples of traded plants from those areas, and their respective uses (ornamental, medicinal, or timber).

Questions

1. Customs officers often come across cultural heritage such as sculptures made from economically costly, legally protected wood (such as Brazilian rosewood). Which method could they use to find out whether the sculpture is made from CITES-listed species? Motivate your answer.
2. What is "plant blindness" and why is it hampering the battle against illegal plant trade?
3. Provide two advantages of AMS over radiometric dating when investigating illegal wildlife trade. Motivate your answer.

Glossary

Accelerator Mass Spectrometry (AMS) – A form of mass spectrometry that accelerates ions to extraordinarily high kinetic energies before mass analysis.

Ayurvedic medicine – A medical system from India that aims to cleanse the body and to restore balance to the body, mind, and spirit by using diet, herbal medicines, exercise, meditation, breathing, physical therapy, and other methods.

Blockchain – A decentralised and distributed network that is used to record transactions across many computers.

Computer vision – An interdisciplinary scientific field that deals with how computers can gain high-level understanding from digital images or videos.

Expert systems – In artificial intelligence, an expert system is a computer system emulating the decision-making ability of a human expert.

Inter-simple sequence repeats (ISSRs) – ISSRs are regions in the genome flanked by microsatellite sequences. PCR amplification of these regions using a single primer yields multiple amplification products that can be used as a dominant multilocus marker system for the study of genetic variation in various organisms.

Near infrared spectroscopy – A spectroscopic method that uses a certain range of the electromagnetic spectrum from 780 nm to 2500 nm which is called the near infrared region.

Pattern recognition – The automated recognition of patterns and regularities in data.

Restriction site Associated DNA Sequencing (RAD-Seq) – A fractional genome sequencing strategy, designed to interrogate anywhere from 0.1% to 10% of a selected genome.

Simple sequence repeats (SSRs) – SSRs are DNA tracts in which a short base-pair motif is repeated several to many times in tandem. These sequences experience frequent mutations that alter the number of repeats.

Spectroscopy – The study of the interaction between matter and electromagnetic radiation as a function of the wavelength or frequency of the radiation.

X-ray microtomography – A 3D modelling method uses X-rays to create cross-sections of a physical object that can be used to recreate a virtual model without destroying the original object.

References

't Sas-Rolfes M, Challender DWS, Hinsley A, Veríssimo D, Milner-Gulland EJ (2019) Illegal wildlife trade: patterns, processes, and governance. Annu. Rev. Environ. Resour. 44, 201–228. https://doi.org/10.1146/annurev-environ-101718-033253

Aimin D, Yunfeng L (2019) "Intelligent factoring" business model and game analysis in the supply chain based on block chain. Management Review 31, 231–240.

Anderson RS (1995) The Lacey Act: America's premier weapon in the fight against unlawful wildlife trafficking. Public Land Law Review, United States.

Arulandhu AJ, Staats M, Hagelaar R, Voorhuijzen MM, Prins TW, Scholtens I, Costessi A, Duijsings D, Rechenmann F, Gaspar FB, Barreto Crespo MT, Holst-Jensen A, Birck M, Burns M, Haynes E, Hochegger R, Klingl A, Lundberg L, Natale C, Niekamp H, Kok E (2017) Development and validation of a multi-locus DNA metabarcoding method to identify endangered species in complex samples. Gigascience 6, 1–18. https://doi.org/10.1093/gigascience/gix080

Ashtiani S-HM, Javanmardi S, Jahanbanifard M, Martynenko A, Verbeek FJ (2021) Detection of mulberry ripeness stages using deep learning models. IEEE Access 1–1. https://doi.org/10.1109/ACCESS.2021.3096550

Blanc-Jolivet C, Kersten B, Daïnou K, Hardy O, Guichoux E, Delcamp A, Degen B (2017) Development of nuclear SNP markers for genetic tracking of Iroko, *Milicia excelsa* and *Milicia regia*. Conserv. Genet. Resour. 1–3. https://doi.org/10.1007/s12686-017-0716-2

Camin F, Boner M, Bontempo L, Fauhl-Hassek C, Kelly SD, Riedl J, Rossmann A (2017) Stable isotope techniques for verifying the declared geographical origin of food in legal cases. Trends Food Sci. Technol. 61, 176–187. https://doi.org/10.1016/j.tifs.2016.12.007

Cerling TE, Barnette JE, Chesson LA, Douglas-Hamilton I, Gobush KS, Uno KT, Wasser SK, Xu X (2016) Radiocarbon dating of seized ivory confirms rapid decline in African elephant populations and provides insight into illegal trade. Proc Natl Acad Sci USA 113, 13330–13335. https://doi.org/10.1073/pnas.1614938113

Challe JFX, Price LL (2009) Endangered edible orchids and vulnerable gatherers in the context of HIV/AIDS in the southern highlands of Tanzania. J. Ethnobiol. Ethnomed. 5, 41. https://doi.org/10.1186/1746-4269-5-41

Chang Y, Iakovou E, Shi W (2020) Blockchain in global supply chains and cross border trade: a critical synthesis of the state-of-the-art, challenges and opportunities. International Journal of Production Research 58, 2082–2099. https://doi.org/10.1080/00207543.2019.1651946

CITES ([n.d.] 2022a) The CITES species [WWW Document]. URL https://cites.org/eng/disc/species.php (accessed 5.25.22a).

CITES ([n.d.] 2022b) Convention on International Trade in Endangered Species of Wild Fauna and Flora [WWW Document]. URL https://cites.org/eng/disc/text.php (accessed 5.25.22b).

Cooney R, Kasterine A, Macmillan DC, Milledge SAH, Nossal K, Roe D, 't Sas-Rolfes M (2015) The trade in wildlife: a framework to improve biodiversity and livelihood outcomes. IUCN, Geneva.

Deklerck V, Finch K, Gasson P, Van den Bulcke J, Van Acker J, Beeckman H, Espinoza E (2017) Comparison of species classification models of mass spectrometry data: kernel discriminant analysis vs random forest; A case study of Afrormosia (*Pericopsis elata* (Harms) Meeuwen). Rapid Commun. Mass Spectrom. 31, 1582–1588. https://doi.org/10.1002/rcm.7939

Deklerck V, Lancaster CA, Van Acker J, Espinoza EO, Van den Bulcke J, Beeckman H (2020) Chemical fingerprinting of wood sampled along a pith-to-bark gradient for individual comparison and provenance identification. Forests 11, 107. https://doi.org/10.3390/f11010107

Di Minin E, Fink C, Hiippala T, Tenkanen H (2019) A framework for investigating illegal wildlife trade on social media with machine learning. Conserv. Biol. 33, 210–213. https://doi.org/10.1111/cobi.13104

Di Minin E, Fink C, Tenkanen H, Hiippala T (2018) Machine learning for tracking illegal wildlife trade on social media. Nat. Ecol. Evol. 2, 406–407. https://doi.org/10.1038/s41559-018-0466-x

Engler, M, Parry-Jones, R (2007) Opportunity or threat: the role of the European Union in global wildlife trade. TRAFFIC.

Eskew EA, White AM, Ross N, Smith KM, Smith KF, Rodríguez JP, Zambrana-Torrelio C, Karesh WB, Daszak P (2020) United States wildlife and wildlife product imports from 2000-2014. Sci. Data 7, 22. https://doi.org/10.1038/s41597-020-0354-5

European Commission (2010) Wildlife trade regulations in the European Union. European Union.

European Commission (2016) EU action plan against wildlife trafficking, COM/2016/087.

Evans PD, Mundo IA, Wiemann MC, Chavarria GD, McClure PJ, Voin D, Espinoza EO (2017) Identification of selected CITES-protected Araucariaceae using DART TOFMS. IAWA J. 38, 266–281. https://doi.org/10.1163/22941932-20170171

Gori Y, Wehrens R, La Porta N, Camin F (2015) Oxygen and hydrogen stable isotope ratios of bulk needles reveal the geographic origin of Norway spruce in the European Alps. PLoS ONE 10, e0118941. https://doi.org/10.1371/journal.pone.0118941

Gross JH (2014) Direct analysis in real time--a critical review on DART-MS. Anal. Bioanal. Chem. 406, 63–80. https://doi.org/10.1007/s00216-013-7316-0

Hartvig I, Czako M, Kjær ED, Nielsen LR, Theilade I (2015) The use of DNA barcoding in identification and conservation of rosewood (*Dalbergia* spp.). PLoS ONE 10, e0138231. https://doi.org/10.1371/journal.pone.0138231

Hassold S, Lowry PP, Bauert MR, Razafintsalama A, Ramamonjisoa L, Widmer A (2016) DNA barcoding of Malagasy rosewoods: towards a molecular identification of CITES-listed *Dalbergia* species. PLoS ONE 11, e0157881. https://doi.org/10.1371/journal.pone.0157881

Hermes TR, Frachetti MD, Bullion EA, Maksudov F, Mustafokulov S, Makarewicz CA (2018) Urban and nomadic isotopic niches reveal dietary connectivities along Central Asia's Silk Roads. Sci. Rep. 8, 5177. https://doi.org/10.1038/s41598-018-22995-2

Horacek M, Jakusch M, Krehan H (2009) Control of origin of larch wood: discrimination between European (Austrian) and Siberian origin by stable isotope analysis. Rapid Commun. Mass Spectrom. 23, 3688–3692. https://doi.org/10.1002/rcm.4309

Jahanbanifard M, Beckers V, Koch G, Beeckman H, Gravendeel B, Verbeek F, Baas P, Priester C, Lens F (2020) Description and evolution of wood anatomical characters in the ebony wood genus *Diospyros* and its close relatives (Ebenaceae): a first step towards combatting illegal logging. IAWA 41, 577–619. https://doi.org/10.1163/22941932-bja10040

Jahanbanifard M, Gravendeel B, Lens F, Verbeek F (2019) Ebony wood identification to battle illegal trade. BISS 3. https://doi.org/10.3897/biss.3.37084

Javanmardi S, Miraei Ashtiani S-H, Verbeek FJ, Martynenko A (2021) Computer-vision classification of corn seed varieties using deep convolutional neural network. J. Stored Prod. Res. 92, 101800. https://doi.org/10.1016/j.jspr.2021.101800

Jenkins M, Timoshyna A, Cornthwaite M (2018) Wild at home: an overview of the harvest and trade in wild plant ingredients. TRAFFIC, Cambridge, UK.

Jiao L, Lu Y, He T, Guo J, Yin Y (2020) DNA barcoding for wood identification: global review of the last decade and future perspective. IAWA 41, 620–643. https://doi.org/10.1163/22941932-bja10041

Kalt-O'Bannon J (1994) Scientific dating and the law: establishing the age of old objects for legal purposes. UMKC L. Rev. 63, 93–132.

Knapp S (2019) Are humans really blind to plants? Plants, People, Planet 1, 164–168. https://doi.org/10.1002/ppp3.36

Kobayashi K, Hwang S-W, Okochi T, Lee W-H, Sugiyama J (2019) Non-destructive method for wood identification using conventional X-ray computed tomography data. Journal of Cultural Heritage 1. https://doi.org/10.1016/j.culher.2019.02.001

Kurland J, Pires SF, McFann SC, Moreto WD (2017) Wildlife crime: a conceptual integration, literature review, and methodological critique. Crime Sci. 6, 4. https://doi.org/10.1186/s40163-017-0066-0

Lancaster C, Espinoza E (2012) Analysis of select *Dalbergia* and trade timber using direct analysis in real time and time-of-flight mass spectrometry for CITES enforcement. Rapid Commun. Mass Spectrom. 26, 1147–1156. https://doi.org/10.1002/rcm.6215

Lavorgna A, Middleton SE, Pickering B, Neumann G (2020) FloraGuard: tackling the online illegal trade in endangered plants through a cross-disciplinary ICT-enabled methodology. J. Contemp. Crim. Justice 36, 428–450. https://doi.org/10.1177/1043986220910297

Lavorgna A, Rutherford C, Vaglica V, Smith MJ, Sajeva M (2018) CITES, wild plants, and opportunities for crime. Eur. J. Crim. Pol. Res. 24, 269–288. https://doi.org/10.1007/s10610-017-9354-1

Lavorgna A (2014) Wildlife trafficking in the Internet age. Crime Sci. 3, 5. https://doi.org/10.1186/s40163-014-0005-2

Lennert E, Bridge C (2018) Analysis and classification of smokeless powders by GC-MS and DART-TOFMS. Forensic Sci. Int. 292, 11–22. https://doi.org/10.1016/j.forsciint.2018.09.003

Lens F, Liang C, Guo Y, Tang X, Jahanbanifard M, da Silva FSC, Ceccantini G, Verbeek FJ (2020) Computer-assisted timber identification based on features extracted from microscopic wood sections. IAWA 1–21. https://doi.org/10.1163/22941932-bja10029

Lian R, Wu Z, Lv X, Rao Y, Li H, Li J, Wang R, Ni C, Zhang Y (2017) Rapid screening of abused drugs by direct analysis in real time (DART) coupled to time-of-flight mass spectrometry (TOF-MS) combined with ion mobility spectrometry (IMS). Forensic Sci. Int. 279, 268–280. https://doi.org/10.1016/j.forsciint.2017.07.010

Lo Y-T, Shaw P-C (2018) DNA-based techniques for authentication of processed food and food supplements. Food Chem. 240, 767–774. https://doi.org/10.1016/j.foodchem.2017.08.022

Lorieul T, Pearson KD, Ellwood ER, Goëau H, Molino J-F, Sweeney PW, Yost JM, Sachs J, Mata-Montero E, Nelson G, Soltis PS, Bonnet P, Joly A (2019) Toward a large-scale and deep phenological stage annotation of herbarium specimens: case studies from temperate, tropical, and equatorial floras. Appl. Plant Sci. 7, e01233. https://doi.org/10.1002/aps3.1233

Manzanilla V, Teixidor-Toneu I, Martin GJ, Hollingsworth PM, de Boer HJ, Kool A (2022) Using target capture to address conservation challenges: population-level tracking of a globally-traded herbal medicine. Mol. Ecol. Resour. 22, 212–224. https://doi.org/10.1111/1755-0998.13472

Margulies JD, Bullough L, Hinsley A, Ingram DJ, Cowell C, Goettsch B, Klitgård BB, Lavorgna A, Sinovas P, Phelps J (2019) Illegal wildlife trade and the persistence of "plant blindness." Plants, People, Planet 1, 173–182. https://doi.org/10.1002/ppp3.10053

Matos MPV, Jackson GP (2019) Isotope ratio mass spectrometry in forensic science applications. Forensic Chemistry 13, 100154. https://doi.org/10.1016/j.forc.2019.100154

Mishra P, Kumar A, Sivaraman G, Shukla AK, Kaliamoorthy R, Slater A, Velusamy S (2017) Character-based DNA barcoding for authentication and conservation of IUCN Red listed threatened species of genus *Decalepis* (Apocynaceae). Sci. Rep. 7, 14910. https://doi.org/10.1038/s41598-017-14887-8

MoonX (2019) Blockchain – Future of wildlife conservation. Medium.

Narayanan A, Bonneau J, Felten E, Miller A, Goldfeder S (2016) Bitcoin and cryptocurrency technologies. Network Security 2016, 4. https://doi.org/10.1016/S1353-4858(16)30074-5

Pakull B, Schindler L, Mader M, Kersten B, Blanc-Jolivet C, Paulini M, Lemes MR, Ward SE, Navarro CM, Cavers S, Sebbenn AM, di Dio O, Guichoux E, Degen B (2020) Development of nuclear SNP markers for mahogany (*Swietenia* spp.). Conserv. Genet. Resour. https://doi.org/10.1007/s12686-020-01162-8

Pearson KD, Nelson G, Aronson MFJ, Bonnet P, Brenskelle L, Davis CC, Denny EG, Ellwood ER, Goëau H, Heberling JM, Joly A, Lorieul T, Mazer SJ, Meineke EK, Stucky BJ, Sweeney P, White AE, Soltis PS (2020) Machine learning using digitized herbarium specimens to advance phenological research. Bioscience 70, 610–620. https://doi.org/10.1093/biosci/biaa044

Polling M, Li C, Cao L, Verbeek F, de Weger LA, Belmonte J, De Linares C, Willemse J, de Boer H, Gravendeel B (2021) Neural networks for increased accuracy of allergenic pollen monitoring. Sci. Rep. 11, 11357. https://doi.org/10.1038/s41598-021-90433-x

Pournader M, Shi Y, Seuring S, Koh SCL (2020) Blockchain applications in supply chains, transport and logistics: a systematic review of the literature. International Journal of Production Research 58, 2063–2081. https://doi.org/10.1080/00207543.2019.1650976

Price E, Larrabure D, Gonzales B, McClure P, Espinoza E (2020) Forensic identification of the keratin fibers of South American camelids by ambient ionization mass spectrometry: Vicuña, alpaca and guanaco. Rapid Commun. Mass Spectrom. 34, e8916. https://doi.org/10.1002/rcm.8916

Price ER, McClure PJ, Jacobs RL, Espinoza EO (2018) Identification of rhinoceros keratin using direct analysis in real time time-of-flight mass spectrometry and multivariate statistical analysis. Rapid Commun. Mass Spectrom. 32, 2106–2112. https://doi.org/10.1002/rcm.8285

Rosen GE, Smith KF (2010) Summarizing the evidence on the international trade in illegal wildlife. Ecohealth 7, 24–32. https://doi.org/10.1007/s10393-010-0317-y

Sabu A, Sreekumar K, Nair RR (2017) Recognition of ayurvedic medicinal plants from leaves: a computer vision approach, in: 2017 Fourth International Conference on Image Information Processing (ICIIP). Presented at the 2017 Fourth International Conference on Image Information Processing (ICIIP), IEEE, pp. 1–5. https://doi.org/10.1109/ICIIP.2017.8313782

Sajeva M, Augugliaro C, Smith MJ, Oddo E (2013) Regulating internet trade in CITES species. Conserv. Biol. 27, 429–430. https://doi.org/10.1111/cobi.12019

Saurabh S, Dey K (2021) Blockchain technology adoption, architecture, and sustainable agri-food supply chains. J. Clean. Prod. 284, 124731. https://doi.org/10.1016/j.jclepro.2020.124731

Schroeder H, Cronn R, Yanbaev Y, Jennings T, Mader M, Degen B, Kersten B (2016) Development of molecular markers for determining continental origin of wood from white oaks (*Quercus* L. sect. *Quercus*). PLoS ONE 11, e0158221. https://doi.org/10.1371/journal.pone.0158221

Smith KM, Zambrana-Torrelio C, White A, Asmussen M, Machalaba C, Kennedy S, Lopez K, Wolf TM, Daszak P, Travis DA, Karesh WB (2017) Summarizing US wildlife trade with an eye toward assessing the risk of infectious disease introduction. Ecohealth 14, 29–39. https://doi.org/10.1007/s10393-017-1211-7

TRAFFIC (2019) Combating wildlife crime linked to the internet - Global trends and China's experiences. TRAFFIC.

Uno KT, Quade J, Fisher DC, Wittemyer G, Douglas-Hamilton I, Andanje S, Omondi P, Litoroh M, Cerling TE (2013) Bomb-curve radiocarbon measurement of recent biologic tissues and applications to wildlife forensics and stable isotope (paleo)ecology. Proc Natl Acad Sci USA 110, 11736–11741. https://doi.org/10.1073/pnas.1302226110

Vaglica V, Sajeva M, McGough HN, Hutchison D, Russo C, Gordon AD, Ramarosandratana AV, Stuppy W, Smith MJ (2017) Monitoring internet trade to inform species conservation actions. Endanger. Species Res. 32, 223–235. https://doi.org/10.3354/esr00803

Veldman S, Gravendeel B, Otieno JN, Lammers Y, Duijm E, Nieman A, Bytebier B, Ngugi G, Martos F, van Andel TR, de Boer HJ (2017) High-throughput sequencing of African chikanda cake highlights conservation challenges in orchids. Biodivers. Conserv. 26, 2029–2046. https://doi.org/10.1007/s10531-017-1343-7

Veldman S, Otieno J, van Andel T, Gravendeel B, de Boer HJ (2014) Efforts urged to tackle thriving illegal orchid trade in Tanzania and Zambia for chikanda production. Traffic Bulletin 26, 47–50.

Vlam M, de Groot GA, Boom A, Copini P, Laros I, Veldhuijzen K, Zakamdi D, Zuidema PA (2018) Developing forensic tools for an African timber: regional origin is revealed by genetic characteristics, but not by isotopic signature. Biological Conservation 220, 262–271. https://doi.org/10.1016/j.biocon.2018.01.031

Wandersee JH, Schussler EE (1999) Preventing plant blindness. Am. Biol. Teach. 61, 82–86. https://doi.org/10.2307/4450624

Williamson J, Maurin O, Shiba SNS, van der Bank H, Pfab M, Pilusa M, Kabongo RM, van der Bank M (2016) Exposing the illegal trade in cycad species (Cycadophyta: *Encephalartos*) at two traditional medicine markets in South Africa using DNA barcoding. Genome 59, 771–781. https://doi.org/10.1139/gen-2016-0032

World Bank (2019) Illegal logging, fishing, and wildlife trade : the costs and how to combat it. World Bank.

WWF (2018) New blockchain project has potential to revolutionise seafood industry. WWF.

WWF (2020) Wildlife crime [WWW Document]. URL https://www.wwf.org.la/what_we_do/wildlife_crime/ (accessed 5.25.22).

Zheng K, Zhang Z, Gauthier J, 2020 Blockchain-based intelligent contract for factoring business in supply chains. Ann. Oper. Res. https://doi.org/10.1007/s10479-020-03601-z

Answers

1. Any non destructive method would be potentially usable such as near infrared spectroscopy or X-ray micro CT, to preserve the samples in their original form.
2. Plant blindness is the bias towards animals, and taking-for-granted plants, which are not recognised as anything but background. The downside of plant blindness is that illegal plant trade is considered as relatively harmless as compared with illegal animal trade.
3. AMS requires a much smaller sample size (20–500 mg) compared to radiometric methods (10–100 g). It is also faster and usually produces higher precision results than radiometric methods. Samples can be analysed in a few hours with AMS, while it can take one or two days with radiometric methods. In case confiscated organisms are still alive, a fast verdict increases the chances of survival as rescued animals or plants can quickly be transferred back to the wild before they die.

Chapter 26
Forensic genetics, botany, and palynology

Panagiotis Madesis[1]

1 Lab of Molecular Biology, Department of Agriculture Crop Production and Rural Environment, University of Thessaly, Volos, Greece

Panagiotis Madesis **pmadesis@uth.gr**

Citation: Madesis P (2022) Chapter 26. Forensic genetics, botany, and palynology. In: de Boer H, Rydmark MO, Verstraete B, Gravendeel B (Eds) Molecular identification of plants: from sequence to species. Advanced Books. https://doi.org/10.3897/ab.e98875

Introduction

Forensic science is the use of science in criminal cases. Many scientific disciplines can be involved, among others chemistry, botany, entomology, and physics. In many trials, the presence and identification of physical evidence can be a critical factor in determining the final verdict. Physical evidence may include among other plant material such as leaves, flowers, fruits, or pollen. In this sense, forensic botany is the study of plants during criminal investigations, as botanical samples can be critical evidence in crimes (Coyle 2004). Distinct areas of specialisation within these broad scientific disciplines can be further recognized. In this regard, specialisation in plant morphology and DNA analysis is highly relevant.

Plant material that is usually found at crime scenes may include leaves, stems, seeds, pollen, flowers, or any other plant parts (Aquila et al. 2014; Coyle et al. 2001, 2005; Ward et al. 2009, 2005).

Plant seeds can be caught and carried in a pant cuff or on a shoe, and plant leaves and stems can be found in a victim's and/or suspect's car. Plant parts can also be identified in the victim's stomach, nose or lungs, under fingernails, on skin, clothes, or hair. However, data generated from recovered botanical material is often not fully exploited since forensic agents may lack the appropriate know-how. The role of the forensic botanist within the investigative process is to compare samples recovered from crime scenes, macroscopically and microscopically examine the biological material. Although botanists might be able to identify a species phenotypically, many times this is not possible and DNA analysis with molecular techniques must be used. New tools, and especially new molecular tools, are being developed in forensic botany to aid in both criminal and civil cases. Although chemical analysis of plant material can serve as evidence when a relevant reference database is available, DNA is much more stable than many macromolecules and metabolites and can persist for long periods, even if broken into smaller fragments. It is therefore very often the preferred method for identifying plants in forensics (Butler 2015).

Examination of DNA is a powerful technique allowing the identification of an individual. A suspect's DNA and a crime scene sample are matched to reference databases containing the profiles of large numbers of individuals generated over time (Werrett 1997). In order to do this, forensic laboratories receive recovered material from crime scenes and reference samples from both suspects and victims or DNA data banks. Historically, DNA was analysed through restriction endonucleases and RFLP analysis of polymorphic DNA regions producing DNA fingerprints (Jeffreys et al. 1985). This technique was successfully used for the first time in 1986 in Leicestershire (UK), to identify the culprit of the murder of two young women (Seton 1988). The successful use of DNA-based evidence in solving a murder resulted in the development of forensic genetics.

Since then, the development and application of DNA-based methods and genetics have revolutionised forensic science. Nowadays the use of DNA as forensic evidence is routine, with a major impact on the criminal justice system and society (Young et al. 2019). Genetic material from plants can also be used in forensic sciences, and many of the same methods that are used to identify individuals can also be used to identify plants. As early as the 1970s, the DNA-based restriction fragment length polymorphism (RFLP) technique allowed plant scientists to analyze samples collected from any plant species taken from almost anywhere, though species discrimination was still limited due to the unavailability of suitable loci and inadequate polymorphism. The invention of the polymerase chain reaction (PCR) led to the development of PCR-based approaches for the production of DNA fingerprints. PCR-based fingerprint methods use single oligonucleotide primers with random sequences for the generation of PCR fragments from genomic DNA, (random amplified polymorphic DNA - RAPD). Morevere, the genome of higher organisms including plants contains three types of simple repetitive DNA sequences, satellite

DNAs, minisatellites, and microsatellites, that are organised in clusters of differing sizes. Microsatellites can be classified as simple sequence repeats (SSRs), simple repetitive sequences (SRS), or simple tandem repeats (STRs) and are common regions used for developing markers that can discriminate between plant species, though this is difficult and requires comprehensive databases (Rakoczy-Trojanowska and Bolibok 2004). Single nucleotide polymorphisms (SNPs) are also useful for plant typing as they possess a low mutation rate and are highly abundant in the plant genome (one SNP per 100 to 500 bp). High-throughput sequencing allows the mining of SNPs suitable for species discrimination and fingerprinting, and SNP markers are already well established in all major crop species (Useche et al. 2001).

DNA technology for forensic plant analysis

Forensic genetics is progressing rapidly, as highly sensitive methods for DNA recovery and new sequencing technologies are being developed (Montelius and Lindblom 2012) (see Chapter 1 DNA from plant tissue and Chapter 9 Sequencing platforms and data types). It is also rapidly becoming an important tool in tackling wildlife crimes (Johnson et al. 2014). There are however several prerequisites before a forensic genetics-led approach can be successful. This includes a high success rate for PCR amplification, even when low amounts or degraded organic material is recovered, small size amplicon products, high primer universality among different plant families (see Chapter 10 DNA barcoding), and robust and reliable reference databases.

DNA barcoding is the most commonly used method for genetic identification in forensic genetics (Hebert et al. 2003; von Cräutlein et al. 2011) (see Chapter 10 DNA barcoding). DNA is isolated from a single specimen and specific agreed-upon regions in the genome that have evolved fast enough to differentiate between closely related species are sequenced (Figure 1A infographic). A sequence produced from an unidentified specimen under investigation is compared to a reference database (such as NCBI) that contains sequences generated from known species. A match is found when the sequence of the unknown sample is homologous to the sequence of a known plant species, thus accurately identifying the plant material found on a crime scene.

In plants, DNA barcodes are mainly derived from the chloroplast (Hollingsworth et al. 2009) and include the *rbcL*, *rpoB*, *rpoC*, and *matK* genes, the *psbK-psbI*, *trnL-F* and *trnL* (UAA) introns, and the non-coding spacers *atpF-atpH* and *trnH-psbA*. A region of the nuclear ribosomal DNA, i.e., internal transcribed spacer 2 (ITS2), is also commonly used (Kress et al. 2005). There is however not a single standard barcode region defined for plants, in contrast to for example animals (Edwards et al. 2008; Hollingsworth et al. 2009; Kress et al. 2005; Taberlet et al. 2007; Yao et al. 2010). The Consortium for the Barcode of Life (CBOL) proposes using *rbcL* and *matK* as the standard two-marker pair for plant identification but it may be necessary to combine this pair with more DNA regions for robust species-level identification of an unknown plant sample (Hollingsworth et al. 2011). The *rbcL* gene is easily recovered across land plants but it is not the most variable region in a plant genome. The *matK* gene has greater sequence variability but its recovery success is lower. Although efforts are being made to improve amplification success for *matK*, limitations in its use remain, especially for forensic genetics where the recovered plant material is often limited or of low quality. Both *rbcL* and *matK* are therefore not always useful for plant identification in forensic cases. Alternative genetic regions for plant identification could be the P6 loop of *trnL* (Taberlet et al. 2007) and *trnH-psbA*.

Recent developments in DNA analysis now allow for the wider use of biological materials, for example, mixtures of samples such as soil or stomach contents (Figure 1B infographic). Techniques to process mixtures are metabarcoding and metagenomics (Habtom et al. 2017; Kho-

dakova et al. 2014; Mills et al. 2017; Shokralla et al. 2012; Taberlet et al. 2012; Yan et al. 2018; Young et al. 2015, 2014). Both methods however require that a number of technical criteria (see Chapter 11 Amplicon metabarcoding and Chapter 12 Metagenomics) are met for the identification of the unknown material to be successful. If these criteria cannot be met, other molecular methods are available, for example, DNA barcoding combined with high-resolution melting (i.e., Bar-HRM). This method is based on the analysis of DNA melting curves, which is an analytical molecular technique that automatically measures the dissociation rate of double-stranded DNA into single-stranded DNA with increasing temperature (see Chapter 13 Barcoding - High resolution melting). This curve is unique for a plant species as it depends on the DNA sequence which is selected to be unique for this species and therefore allows for the identification of the plants in a crime scene sample. However, access to a comprehensive database is a prerequisite for DNA barcoding. Additionally, for closely related species intra and interspecies variation must be carefully considered when using this approach to ensure that results are correctly interpreted.

Use of palynology in forensics

Palynology is the study of palynomorphs, including pollen, spores, dinocysts, etc. Pollen grains are however the most studied palynomorph, and especially in forensics it can be an important piece of evidence if it can be associated with a crime scene or be retrieved from the suspect or equipment used at the crime scene. Pollen is of microscopic dimensions and can very easily be retained in clothes, home objects, and soil. Crime scenes limited to a few square meters, like a rape scene or the entry point of a burglary, are very often the best choices for the use of forensic palynology (Bever and Cimino 2000; Bever et al. 2000).The use of pollen in criminal investigations is referred to as forensic palynology. Vascular plants, like flowering plants and conifers, produce large quantities of pollen, while ferns produce spores (Bever and Cimino 2000). Plant species belonging to different families often have unique pollen morphology, a characteristic that allows pollen to be used for plant identification in forensics (Bock and Norris 1997). However, pollen originating from plant species from the same family might be very similar and hard to distinguish based on morphology alone (see Chapter 5 DNA from pollen), though it is often possible for an unknown pollen grain's morphology to be analysed and compared to large databases to determine the plant species from which it originates (Bever and Cimino 2000). Thus the application of DNA techniques (DNA barcoding and DNA metabarcoding) could help towards the attribution of the retrieved pollen to certain species.

Forensic case studies

Case study 1

In Auckland (New Zealand), a prostitute claimed she was attacked in a passageway by a suspect, around seven meters away from the suspect's car (Horrocks and Walsh 1999). The offender claimed that he never entered the passageway or moved away from his car for more than one metre. Examination of the area and the crime scene revealed no footprints. However, pollen that was retrieved from the soil in and around the crime scene revealed that the types of pollen between the passageway and driveway (where the car was parked) were similar, but the quantities were different between both areas. The passageway contained 76% *Coprosma* pollen, but

the driveway sample contained only 8%. The offender's clothing contained approximately 80% of *Coprosma* pollen and only minor amounts of other pollen species. These results suggested that the victim's claims about being assaulted in the passageway were true.

Case study 2

In Taipei (Taiwan), the body of a young woman was found lying by a drain in an urban area. It was unknown whether she was a homicide or suicide victim. Her body showed no obvious bone fractures and it was suspected that she was involved in a hit and run by a car. By the time investigators arrived at the scene, the body had already been transferred to a hospital, where a tiny berry and stem was found in the victim's hair. This berry was however not commonly found in the area where the victim lived or where the body was found. The investigators discovered the same plant on the edge of a railing above a drain attached to a building directly next to where the body was found, suggesting that the woman fell from the building, and the plant piece became tangled in her hair during the fall (Coyle 2004).

Case study 3

A murder case in 1992 in Arizona (USA) revealed the power of forensic botany. Seed pods of a Palo Verde tree (*Cercidium* sp.) were retrieved from a suspect's pickup truck (Yoon 1993). In the forensic analysis, 11 trees from around the crime scene were compared to 18 trees from different areas further away from the crime scene. DNA analysis using RAPD markers demonstrated that the pods from the pickup truck genetically matched to a tree near where the dead body was found, suggesting that the suspect was the culprit (Yoon 1993).

Case study 4

In a Finnish study, RAPD and SSR molecular markers were used on mosses to connect three suspects to a murder scene (Korpelainen and Virtanen 2003). The suspects had been spotted leaving a cafe with the victim. Moss samples were retrieved from the suspects' cars, clothes, and shoes. Although the analysis did not provide an exact match between moss types found on the suspect's properties versus what was found on the victim, it suggested that the moss from the victim and the crime scene belonged to the same plant population. In addition, the samples collected from the suspects were found to be genetically closer to moss samples found on the victim than to other moss samples in the region.

Case study 5

Forensic botany also helped to resolve a case of theft that occurred at a Catholic church in Florence (Italy). In this case, the thief made a mistake, leaving faecal material at the crime scene, as, unfortunately for him, he suffered from diarrhoea. Although a priest at the church had previously cleaned the crime scene of faecal matter, there was still enough material left to be collected by the police. The police suspected a local man with a police record who suffered from Crohn's disease. The suspect denied the accusations and presented an alibi. The police, who had retrieved his blue jeans from the time of the robbery, found them stained with faeces, yet

the suspect still denied being guilty and challenged the police to "prove it". The comparison of the two samples revealed 14 dietary items of botanic origin that matched and none that did not, forcing the suspect to confess the crime (Bock and Norris 2016).

Case study 6

In the early 1980s, a young girls' body was found whose last known meal was with her boyfriend at a local fast-food restaurant. An autopsy however revealed the presence of vegetables in her stomach that were not on the fast food restaurant's menu. A botanical investigation confirmed the autopsy results, suggesting she had another meal before her death, which helped her boyfriend to be cleared of any charges. The case was solved a few years later when a serial killer confessed to the murder (Bock and Norris 2016).

Case study 7

In the Black Widow case, in 1993, a domestic homicide was solved with the help of forensic botany. The victim Gerry was married to Jill who had 7 previous marriages. When Gerry found out that Jill had not actually divorced her 7th husband before marrying him, he went to court to annul the marriage and freeze his assets. On the day of his death, Gerry had a breakfast of coffee, hash browns, eggs, and toast, and Jill and her then-boyfriend were spotted near his house. Forensic botanists examined the contents of his stomach and found starch and onion, concluding that the only meal he had was his breakfast and that he did not go out to have another meal. That coincided with the time that Jill was seen at his house and allowed the police to issue a search warrant for her property. The police found a gun and other evidence which led the court to find her and her boyfriend guilty (Bock and Norris 2016).

Case study 8

In a homicide case, a body was found in a stream near a roadside covered with the knotgrass *Polygonum aviculare*. Seeds of knotgrass were recovered from the wheels of the suspect's car. Additional knotgrass samples were collected from different sites and locations. The investigators used AFLP molecular markers to demonstrate that the origin of the seeds found in the suspect's car came from the crime site (Koopman et al. 2012).

Case study 9

Metagenomic analysis for human DNA was used in a sexual assault case that took place in the Netherlands in 2015 and involved a 28-year-old woman. The woman preserved her clothes after the assault and also took intimate samples from herself. Initially, the samples were analysed using capillary electrophoresis (CE) analysis. A year later, these CE results produced a hit in the Dutch convicted criminal database. However, the analysis was challenged, and the ambiguous results made the suspect go free. Only after the use of massively parallel sequencing, it was possible to match the suspect's environmental DNA with the assault evidence which finally led to his conviction in 2018 (de Knijf 2020).

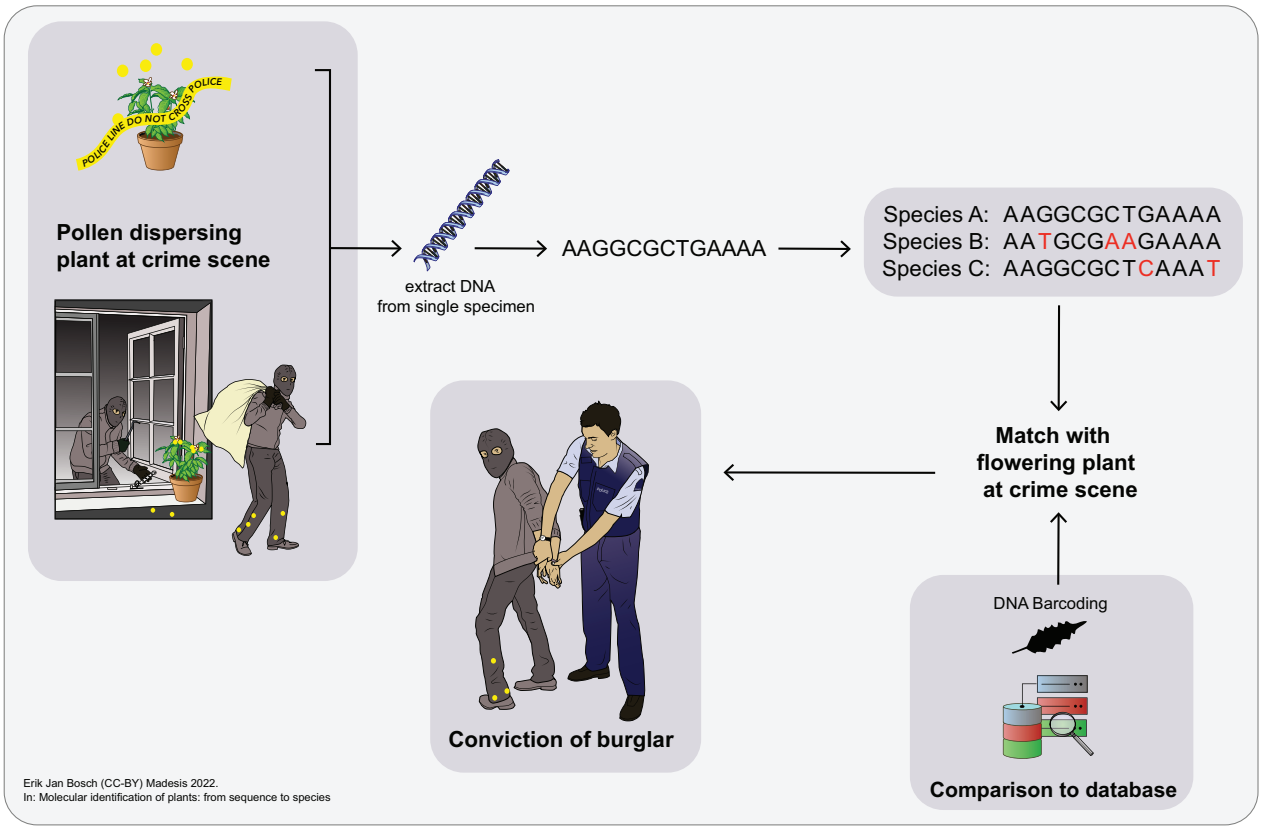

Figure 1. Infographic Chapter 26: Visual representation of the content of this chapter.

Questions

1. What is the advantage of plant DNA over other plant metabolites as forensic evidence?
2. Is DNA barcoding a suitable approach for plant forensics? Motivate your answer.
3. Can Bar-HRM be used in plant forensics? Motivate your answer.
4. Why is palynology a suitable method for plant forensics? Motivate your answer.

Glossary

AFLP – Amplified fragment length polymorphism is a PCR-based technique that uses selective amplification of a subset of digested DNA fragments to generate and compare unique fingerprints for genomes of interest.

RAPD – Random amplification of polymorphic DNA is PCR-based technique in which DNA fragments are amplified at random using primers with arbitrary nucleotide sequences.

RFLP – Restriction fragment length polymorphism is a technique that utilises variations in DNA, i.e. polymorphisms, to differentiate between individuals.

RFLP analysis – A DNA sample is fragmented with restriction enzymes, which selectively cleave the DNA. The produced fragments are separated with agarose gel electrophoresis and since different individuals have fragments of different length, it is possible to distinguish between them.

References

Aquila I, Ausania F, Di Nunzio C, Serra A, Boca S, Capelli A, Magni P, Ricci P (2014) The role of forensic botany in crime scene investigation: case report and review of literature. J. Forensic Sci. 59, 820–824. https://doi.org/10.1111/1556-4029.12401

Bever R, Cimino M (2000) Analysis of botanical trace evidence. Presented at the 11th International Symposium on Human Identification, Promega.

Bever R, Golenberg E, Barnes L, Brinkac L, Jones E, Yoshida K (2000) Molecular analysis of botanical trace evidence: development of techniques. Presented at the American Academy of Forensic Sciences Annual Meeting, American Academy of Forensic Sciences.

Bock JH, Norris DO (1997) Forensic botany: an under-utilized resource. J. Forensic Sci. 42, 14130J. https://doi.org/10.1520/JFS14130J

Bock JH, Norris DO (2016) Cases using evidence from plant anatomy, in: Forensic Plant Science. Elsevier, pp. 85–94. https://doi.org/10.1016/B978-0-12-801475-2.00005-1

Butler JM (2015) The future of forensic DNA analysis. Philos. Trans. R. Soc. Lond. B Biol. Sci. 370. https://doi.org/10.1098/rstb.2014.0252

Coyle HM, Ladd C, Palmbach T, Lee HC (2001) The Green Revolution: botanical contributions to forensics and drug enforcement. Croat. Med. J. 42, 340–345.

Coyle HM (2004) Forensic botany: principles and applications to criminal casework, 1st ed. CRC Press, Boca Raton, FL.

Coyle ME, Smith CA, Peat B (2005) Cephalic version by moxibustion for breech presentation. Cochrane Database Syst. Rev. CD003928. https://doi.org/10.1002/14651858.CD003928.pub2

de Knijf P (2020) How next generation sequencing resolved a difficult case, leading to the first criminal conviction of its kind. Verogen.

Edwards D, Horn A, Taylor D, Savolainen V, Hawkins JA (2008) DNA barcoding of a large genus, *Aspalathus* L. (Fabaceae). Taxon 57, 1317–1327.

Habtom H, Demanèche S, Dawson L, Azulay C, Matan O, Robe P, Gafny R, Simonet P, Jurkevitch E, Pasternak Z (2017) Soil characterisation by bacterial community analysis for forensic applications: A quantitative comparison of environmental technologies. Forensic Sci. Int. Genet. 26, 21–29. https://doi.org/10.1016/j.fsigen.2016.10.005

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. Proc. Biol. Sci. 270, 313–321. https://doi.org/10.1098/rspb.2002.2218

Hollingsworth ML, Andra Clark A, Forrest LL, Richardson J, Pennington RT, Long DG, Cowan R, Chase MW, Gaudeul M, Hollingsworth PM (2009) Selecting barcoding loci for plants: evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants. Mol. Ecol. Resour. 9, 439–457. https://doi.org/10.1111/j.1755-0998.2008.02439.x

Hollingsworth PM, Graham SW, Little DP (2011) Choosing and using a plant DNA barcode. PLoS ONE 6, e19254. https://doi.org/10.1371/journal.pone.0019254

Horrocks M, Walsh KAJ (1999) Fine resolution of pollen patterns in limited space: differentiating a crime scene and alibi scene seven meters apart. J. Forensic Sci. 44, 14477J. https://doi.org/10.1520/JFS14477J

Jeffreys AJ, Wilson V, Thein SL (1985) Individual-specific "fingerprints" of human DNA. Nature 316, 76–79. https://doi.org/10.1038/316076a0

Johnson RN, Wilson-Wilde L, Linacre A (2014) Current and future directions of DNA in wildlife forensic science. Forensic Sci. Int. Genet. 10, 1–11. https://doi.org/10.1016/j.fsigen.2013.12.007

Khodakova AS, Smith RJ, Burgoyne L, Abarno D, Linacre A (2014) Random whole metagenomic sequencing for forensic discrimination of soils. PLoS ONE 9, e104996. https://doi.org/10.1371/journal.pone.0104996

Koopman WJM, Kuiper I, Klein-Geltink DJA, Sabatino GJH, Smulders MJM (2012) Botanical DNA evidence in criminal cases: knotgrass (*Polygonum aviculare* L.) as a model species. Forensic Sci. Int. Genet. 6, 366–374. https://doi.org/10.1016/j.fsigen.2011.07.013

Korpelainen H, Virtanen V (2003) DNA fingerprinting of mosses. J. Forensic Sci. 48, 804–807.

Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH (2005) Use of DNA barcodes to identify flowering plants. Proc Natl Acad Sci USA 102, 8369–8374. https://doi.org/10.1073/pnas.0503123102

Mills JG, Weinstein P, Gellie NJC, Weyrich LS, Lowe AJ, Breed MF (2017) Urban habitat restoration provides a human health benefit through microbiome rewilding: the Microbiome Rewilding Hypothesis. Restor. Ecol. 25, 866–872. https://doi.org/10.1111/rec.12610

Montelius K, Lindblom B (2012) DNA analysis in disaster victim identification. Forensic Sci. Med. Pathol. 8, 140–147. https://doi.org/10.1007/s12024-011-9276-z

Rakoczy-Trojanowska M, Bolibok H (2004) Characteristics and a comparison of three classes of microsatellite-based markers and their application in plants. Cell. Mol. Biol. Lett. 9, 221–238.

Seton C (1988) Life for sex killer who sent decoy to take genetic test. The Times (London) 3.

Shokralla S, Spall JL, Gibson JF, Hajibabaei M (2012) Next-generation sequencing technologies for environmental DNA research. Mol. Ecol. 21, 1794–1805. https://doi.org/10.1111/j.1365-294X.2012.05538.x

Taberlet P, Coissac E, Hajibabaei M, Rieseberg LH (2012) Environmental DNA. Mol. Ecol. 21, 1789–1793. https://doi.org/10.1111/j.1365-294X.2012.05542.x

Taberlet P, Coissac E, Pompanon F, Gielly L, Miquel C, Valentini A, Vermat T, Corthier G, Brochmann C, Willerslev E (2007) Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding. Nucleic Acids Res. 35, e14. https://doi.org/10.1093/nar/gkl938

Useche FJ, Gao G, Harafey M, Rafalski A (2001) High-throughput identification, database storage and analysis of SNPs in EST sequences. Genome Inform. 12, 194–203.

von Cräutlein M, Korpelainen H, Pietiläinen M, Rikkinen J (2011) DNA barcoding: a tool for improved taxon identification and detection of species diversity. Biodivers. Conserv. 20, 373–389. https://doi.org/10.1007/s10531-010-9964-0

Ward J, Gilmore SR, Robertson J, Peakall R (2009) A grass molecular identification system for forensic botany: a critical evaluation of the strengths and limitations. J. Forensic Sci. 54, 1254–1260. https://doi.org/10.1111/j.1556-4029.2009.01196.x

Ward J, Peakall R, Gilmore SR, Robertson J (2005) A molecular identification system for grasses: a novel technology for forensic botany. Forensic Sci. Int. 152, 121–131. https://doi.org/10.1016/j.forsciint.2004.07.015

Werrett DJ (1997) The national DNA database. Forensic Sci. Int. 88, 33–42. https://doi.org/10.1016/S0379-0738(97)00081-9

Yan D, Mills JG, Gellie NJC, Bissett A, Lowe AJ, Breed MF (2018) High-throughput eDNA monitoring of fungi to track functional recovery in ecological restoration. Biological Conservation 217, 113–120. https://doi.org/10.1016/j.biocon.2017.10.035

Yao H, Song J, Liu C, Luo K, Han J, Li Y, Pang X, Xu H, Zhu Y, Xiao P, Chen S (2010) Use of ITS2 region as the universal DNA barcode for plants and animals. PLoS ONE 5. https://doi.org/10.1371/journal.pone.0013102

Yoon CK (1993) Forensic science. Botanical witness for the prosecution. Science 260, 894–895. https://doi.org/10.1126/science.8493521

Young JM, Higgins D, Austin JJ (2019) Soil DNA: advances in DNA technology offer a powerful new tool for forensic science. Geological Society, London, Special Publications SP492-2017–351. https://doi.org/10.1144/SP492-2017-351

Young JM, Weyrich LS, Breen J, Macdonald LM, Cooper A (2015) Predicting the origin of soil evidence: High throughput eukaryote sequencing and MIR spectroscopy applied to a crime scene scenario. Forensic Sci. Int. 251, 22–31. https://doi.org/10.1016/j.forsciint.2015.03.008

Young JM, Weyrich LS, Cooper A (2014) Forensic soil DNA analysis using high-throughput sequencing: a comparison of four molecular markers. Forensic Sci. Int. Genet. 13, 176–184. https://doi.org/10.1016/j.fsigen.2014.07.014

Answers

1. DNA is more stable over time and persists over a longer period of time, so it is more useful for identifying unknown plant material than other plant metabolites.

2. DNA barcoding allows the identification of plants and the development of a suitable database so it is the appropriate solution for forensic use when we want to identify plant species and match species in a given area with plant material identified on a suspect. If matches are sought on plant population level, though, fingerprint methods such as microsatellites might be more appropriate. However, this requires that the population of plants refers exclusively to individuals from the same species.
3. Bar-HRM is a method that combines DNA barcoding and High Resolution Melting Analysis. The method could be an alternative to DNA sequencing which allows rapid results should this be necessary, however, the use of sequencing is probably indispensable for forensic use.
4. Palynology is the use of pollen for the identification of a species, which becomes a powerful tool when combined with DNA barcoding. Pollen can stay intact for thousands of years, protecting the DNA it contains.

This book seeks to provide a practical overview of all aspects of relevance in the field of molecular identification of plants. The first section, "From sample to DNA", provides information on how to set up an experiment, how to design the best sampling protocol, and how to extract DNA from different substrates. The second section, "From DNA to sequence or identification", gives an overview of the methods that can be used. The final section, "From identification to application", shows what kind of scientific questions that can be addressed or which applications with relevance for society are possible with plant identification. This book is meant for people with previous experience who want to bring themselves up to date with the latest techniques, but also for early stage researchers who need a first overview of the available options. We hope that this book will be a useful tool for both science and education.

Plant.ID