Research Ideas and Outcomes 12: e187033
doi: 10.3897/rio.12.e187033

Project Report

Integrating the biodiversity genomics continuum: harmonising data from barcodes to reference genomes

Katharina F Heil[‡], Tyler S Alioto[§], Astrid Böhne[|], Tom Brown[¶,#], Eli Chadwick[¤], Physilia Chua[‡], Christian de Guttry[«], Diego De Panis[»,^], Carole Goble[˅], Ivo Gut[§], Marta Gut[§], Nick Juty[¤], Seanna McTaggart[¦], Laura Najera-Cortazar[ˀ], Joana Paupério[˂], Tomasz Rewicz[¢], Felix Shaw[¦], Stian Soiland-Reyes[ℓ], Robert M Waterhouse[«], Peter Woollard[ϑ], Rutger A Vos[P]

‡ ELIXIR Europe, Hinxton, Cambridge, United Kingdom
§ Centro Nacional de Análisis Genómico, Barcelona, Spain
| Leibniz Institute for the Analysis of Biodiversity Change, Museum Koenig Bonn, Bonn, Germany
¶ Leibniz Institute for Zoo and Wildlife Research, Alfred-Kowalke-Straße 17, 10315 Berlin, Germany, Berlin, Germany
# Berlin Center for Genomics in Biodiversity Research (BeGenDiv), Koenigin-Luise-Str 6-8, 14195 Berlin, Germany, Berlin, Germany
¤ University of Manchester, Manchester, United Kingdom
« SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland
» Leibniz Institute for Zoo and Wildlife Research, Alfred-Kowalke-Str 17, 10315, Berlin, Germany
^ Berlin Center for Genomics in Biodiversity Research (BeGenDiv), Königin-Luise-Str 6-8, 14195, Berlin, Germany
˅ School of Computer Science, University of Manchester, Manchester, United Kingdom
¦ Earlham Institute, Norwich, United Kingdom
ˀ BIOPOLIS Association CIBIO, Vairão, Portugal
˂ European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, United Kingdom
¢ University of Lodz, Lodz, Poland
ℓ The University of Manchester, Manchester, United Kingdom
ϑ European Molecular Biology Laboratory, Cambridge, United Kingdom
P Naturalis Biodiversity Center, Leiden, Netherlands

Corresponding author: Rutger A Vos (rutger.vos@naturalis.nl)

Reviewable v 1

Received: 30 Jan 2026 | Published: 17 Mar 2026

Citation: Heil K, Alioto T, Böhne A, Brown T, Chadwick E, Chua P, de Guttry C, De Panis D, Goble C, Gut I, Gut M, Juty N, McTaggart S, Najera-Cortazar L, Paupério J, Rewicz T, Shaw F, Soiland-Reyes S, Waterhouse R, Woollard P, Vos R (2026) Integrating the biodiversity genomics continuum: harmonising data from barcodes to reference genomes. Research Ideas and Outcomes 12: e187033. https://doi.org/10.3897/rio.12.e187033

Abstract

Biodiversity genomics is converging from historically separate approaches — DNA barcoding and reference genome sequencing — into an integrated digital ecosystem

driven by shared data stewardship principles: transparent provenance, persistent identifiers and interoperable repositories. We demonstrate how these workflows can operate within a unified informatics architecture spanning data generation, validation, publication and reuse. We describe coordinated infrastructure, including the European BOLD mirror, ERGA Genome Tracking Console and metadata platforms COPO and PlutoF. These systems employ harmonised validation pipelines, shared metadata standards that bridge the Darwin Core and Genomic Standards Consortium vocabularies and automated data exchange amongst ENA, UNITE and GBIF. Workflows in Galaxy, Nextflow and Snakemake are registered in WorkflowHub as Research Object Crates (RO-Crates), ensuring reproducibility and complete provenance. Key outcomes include comprehensive data flow documentation, automated quality control using BUSCO and ERGA Assembly Reports and robust specimen-to-data linkage. We identify challenges in metadata harmonisation, distributed tracking, collaborative attribution and infrastructure sustainability and provide recommendations for addressing them through existing platforms and emerging RO-Crate standards. This work establishes practical foundations for treating biodiversity molecular data as a continuum, demonstrating how FAIR principles can scale to continental initiatives.

Keywords

biodiversity genomics, DNA barcoding, genome sequencing, FAIR data, metadata standards, reproducible workflows

Introduction

Biodiversity research is entering a new era in which molecular data form a continuous spectrum rather than a set of isolated approaches. At one end lies DNA barcoding (Hebert et al. 2003), which uses short, standardised genetic markers to identify species and quantify community composition through metabarcoding. At the other end lies reference genome sequencing (Supple and Shapiro 2018), which aims to reconstruct full chromosomal assemblies for selected taxa. Although these two approaches differ in scope and scale, their convergence is increasingly evident in how data are managed, analysed, validated and shared. Together, they form a complementary strategy for documenting life, from local inventories to complete genomic blueprints.

DNA barcoding, spearheaded by initiatives such as the International Barcode of Life (iBOL, International Barcode of Life consortium (2025)), has transformed species identification by enabling scalable, sequence-based taxonomy. Meanwhile, large-scale genome sequencing efforts, such as the European Reference Genome Atlas (ERGA, Mazzoni et al. (2023)), have redefined standards for completeness, accuracy and contextual metadata. Both depend on the same principles: transparent provenance, persistent identifiers and interoperable repositories. These principles create fertile ground for alignment. This convergence represents a philosophical shift towards treating biodiversity data as a shared digital ecosystem.

The key commonality lies in data handling. Whether assembling gigabase-scale genomes or processing short barcode reads, both fields require consistent metadata capture, version-controlled analytical workflows and reliable publication in trusted repositories such as the European Nucleotide Archive (ENA, O'Cathail et al. (2025)) and the Barcode of Life Data Systems (BOLD, Ratnasingham and Hebert (2007)). The adoption of community-driven tools and standards, such as Galaxy (Goecks et al. 2010) for workflow execution, WorkflowHub (da Silva et al. 2020) for sharing and documenting pipelines, PlutoF (Abarenkov et al. 2010) and COPO (Shaw et al. 2024) for data management and metadata negotiation and systems such as UNITE (Abarenkov et al. 2024) and BOLD for taxonomically anchored sequence data, creates bridges between communities that once operated in parallel. Through these connections, molecular data can flow seamlessly between generation, processing, validation, publication and application.

This shared informatics foundation enables coordinated scaling of biodiversity genomics. Automated validation pipelines, harmonised metadata vocabularies and FAIR (Findable, Accessible, Interoperable, Reusable, Wilkinson et al. (2016)) data practices ensure that sequence data from diverse origins can be reused across the biodiversity research domain. Crucially, convergence in data handling strengthens the link between specimens, sequences and analyses, preserving provenance even as datasets grow in complexity and size. The result is a knowledge system in which short barcode markers can guide genome sequencing priorities and reference genomes can refine barcode-based identification, with each informing the other through shared digital infrastructure.

This report describes how these conceptual alignments translate into practical workflows and shared infrastructure. They show that the union of barcoding and genome sequencing comprises a synchronisation of data logic: a shift from separate production lines to a common architecture for biodiversity knowledge.

Commonalities in data workflows

The convergence between DNA barcoding and genome sequencing becomes tangible in how the data are handled after laboratory work ends. Despite their differences in scope, with barcodes distilling diversity into compact diagnostic sequences, while genomes capture it at full chromosomal resolution, both approaches pass through strikingly similar stages of digital stewardship. What unites them is not the size of their data, but their logic: structured data processing and validation, transparent provenance and deposition into shared repositories where the information can be discovered, reused and connected across research domains.

Both barcoding and genome sequencing workflows begin with an initial round of pre-processing quality control to verify that sequencing reads meet baseline standards of completeness and accuracy. From that point, they follow parallel analytical paths that differ mainly in technical detail. Barcoding workflows assemble reads into high-fidelity consensus markers for species identification and community profiling, while genome

workflows assemble larger contigs and scaffolds. Then, a consensus is computed, which is annotated. The result is validated and submitted (see Fig. 1). In both cases, every computational step is scripted, version-controlled and logged to ensure reproducibility and to maintain an unbroken record of data provenance.

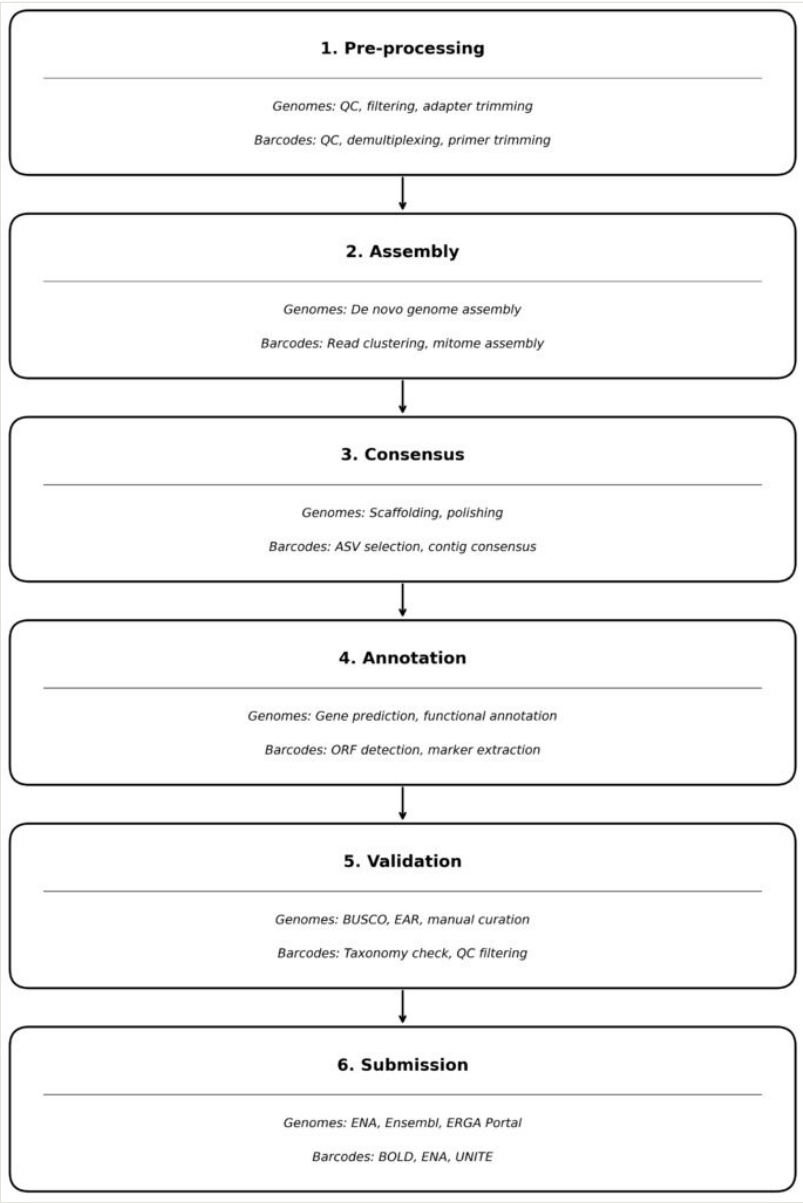

Figure 1. doi

Commonalities in data flows between barcoding and reference genome sequencing for biodiversity.

The key to this shared logic lies in metadata management. Each dataset is accompanied by structured information about its origin, processing and validation, ensuring that specimens, sequences and analyses remain linked throughout their lifecycle. Tools such as COPO, a metadata brokering system that supports FAIR-compliant submission to public repositories and PlutoF, which manages collection of metadata for barcoding and metabarcoding data, illustrate how distinct communities can align around compatible standards. Once validated, data and metadata are deposited in trusted archives, such as ENA for raw sequences and assembled genomes, UNITE for fungal metabarcoding data and BOLD for specimen barcodes, forming an interconnected landscape of biodiversity genomics.

Beyond these repositories, a growing ecosystem of workflow platforms reinforces convergence in analytical practice. The Galaxy environment enables researchers to run, share and reproduce pipelines transparently, while WorkflowHub provides a registry for documenting, crediting or attributing and versioning these pipelines as community assets. Both barcoding and genome sequencing initiatives now adopt these tools to promote openness and interoperability. Increasingly, workflow outputs, metadata and provenance records are bundled into Research Object Crates (RO-Crates, Soiland-Reyes et al. (2022)), i.e. lightweight digital containers that capture every element of a computational process, from raw data to final results, in a machine-readable format. This approach ensures that datasets can travel intact between systems and over time without loss of context.

When viewed through the lens of data handling, the two approaches to biodiversity genomics reveal a single conceptual architecture. Each begins with sequence validation, proceeds through structured assembly and validation pipelines and ends in the publication of self-contained research objects that embody the principles of transparency, reproducibility and reuse. These parallels do more than simplify logistics: they enable interlinking barcode-based surveys with genome assemblies, allowing the same digital infrastructure to support biodiversity discovery, ecological monitoring and evolutionary research. What once were separate data production lines are now strands of the same informatic fabric, woven together by shared standards and a commitment to open, reproducible science.

Implementation of the data-centric work

Translating the shared logic of biodiversity genomics data handling into practice requires infrastructure that supports both the high-throughput demands of DNA barcoding and the high-complexity demands of genome sequencing. The challenge is not only to process large volumes of sequence data, but to do so in a way that preserves provenance, harmonises metadata and ensures interoperability across platforms and repositories. Across European biodiversity genomics efforts, this principle has been realised through coordinated implementation of data systems, validation pipelines and tracking tools that operate on a common digital foundation.

Implementing interoperable systems for barcoding and metabarcoding

For single-specimen and community-level barcoding, implementation has focused on establishing open, synchronised systems that integrate with global biodiversity infrastructures. The European mirror of BOLD exemplifies this approach. It extends the original platform by introducing programmatic interfaces (APIs), enabling automated data exchange with repositories such as ENA, UNITE and the Distributed System of Scientific Collections (DiSSCo, Addink et al. (2019)). This linkage ensures that barcode data and their associated metadata remain discoverable, interconnected and traceable across molecular, ecological and specimen databases.

Complementary systems have been developed for metabarcoding data, where bulk samples or environmental DNA represent complex communities rather than individual specimens. Here, the PlutoF workbench plays a central role: it provides standardised templates for collection metadata and connects to storage solutions for sequence data that are entered into public archives as workflows mature. The emphasis in both cases is on machine-readable metadata, automated validation and FAIR-compliant publication. Together, these systems enable continuity between specimen-level identification and ecosystem-level observation, using shared digital conventions.

Implementing harmonised workflows for genome sequencing

Genome sequencing efforts for biodiversity, whose principal applications are in population genomics, apply the same principles at a per-specimen, more complex scale. Harmonised pipelines, which are typically implemented in Nextflow (Di Tommaso et al. 2017), Snakemake (Köster and Rahmann 2012) and Galaxy, assemble, polish and annotate genomes on distributed compute clusters located across participating institutions. Each step produces quality metrics such as contig N50 (Lander et al. 2001), BUSCO (Simão et al. 2015) completeness and gene-model integrity, which feed into automated validation layers. Assemblies that meet defined benchmarks are deposited in the ENA and annotations are published by Ensembl (Dyer et al. 2025) within days of completion. This process creates a continuous data flow from sequencing to publication, underpinned by transparent provenance and reproducible workflows.

To maintain oversight of these distributed activities, an ERGA Genome Tracking Console (GTC, ERGA (2025a)) has been established as a central coordination hub. Functionally similar to a laboratory information management system, the GTC monitors sample and data progress from field collection and metadata submission through sequencing, assembly and release. It communicates programmatically with COPO, the metadata brokering platform that ensures consistency with community vocabularies before submission to ENA and, by extension, to BioSamples (Courtot et al. 2022). By exposing an open API, the GTC allows sequencing centres to interact directly with the system, whether through automated updates from local databases or manual entries by curators. This architecture keeps genome projects transparent and synchronised across multiple facilities, preventing data silos and duplication of effort.

Linking through shared metadata and interoperability

Both barcoding and genome sequencing systems rely on shared metadata. Harmonised schemas and persistent identifiers connect specimens, sequencing data and analytical results, ensuring that digital and physical records remain linked throughout their lifecycles. The COPO platform has proven particularly effective in this role: it serves as a mediator between laboratory teams and public repositories, validating metadata before brokering submissions to archives such as ENA. By providing a consistent metadata management interface, COPO ensures that data and metadata conform to community standards. For example, a COPO representation of metadata is automatically converted for submission via ENA's Darwin Tree of Life (DToL) checklist to BioSamples. The DToL checklist includes key metadata, based on GSC MIxS, thereby increasing interoperability across INSDC (INSDC 2025) and other systems that map to MIxS, such as GBIF (Telenius 2011).

Interoperability between systems is achieved through open APIs and shared packaging standards. A growing number of workflows and datasets are encapsulated as RO-Crates; those produced by COPO, WorkflowHub or Galaxy maintain the relationships between inputs, analyses and outputs, ensuring that datasets can be moved between repositories without losing context. WorkflowHub, in particular, provides a public registry where analysis pipelines for both barcoding and genome assembly are versioned, described and shared, allowing others to reuse or adapt them. These mechanisms form the connective tissue that integrates separate domains of biodiversity genomics into a coherent digital ecosystem.

Coordination and lessons learned

The success of these implementations has depended as much on coordination as on technology. Regular communication amongst sampling, sequencing and data management teams has been essential for aligning expectations and identifying shared challenges. Developing automated validation and submission pipelines required extensive consultation across institutions to ensure compatibility with existing community standards, while still allowing flexibility for local practices. The reliance on API-driven interoperability proved especially valuable, allowing systems to exchange information programmatically, reducing manual data entry and enabling updates to propagate automatically across the network of repositories.

Collectively, these implementations demonstrate that the convergence between barcoding and genome sequencing is not only conceptual, but operational. Both now participate in a shared informatics architecture grounded in transparency, validation and interoperability. As a result, barcode data can provide information for genome sequencing priorities and genome assemblies can, in turn, refine species identification and taxonomic resolution. The infrastructure now in place shows how the ideals of FAIR data and open science can scale to the continental level, linking every sequence to its biological origin and analytical history through a continuous digital thread.

Outcomes and learnings from the joint activities

The joint activities, including whole-genome sequencing and barcoding, have delivered several key outcomes that have enhanced data management and integration. Central to these achievements are the data flows for both the (meta)barcoding and genome sequencing efforts, which trace data from initial collection through processing to final publication. These process maps represent the first comprehensive documentation of how data move through large-scale biodiversity genomics initiatives in Europe.

(Meta)data publishing and tracking

In barcoding, the data flow centres on the newly-established European instance of BOLD, serving as both a processing pipeline and a public data repository. This implementation includes data validation, standardised quality control procedures and data exchange with other biodiversity and genomics databases, including GBIF, UNITE and ENA.

The genome sequencing data flow involves a more complex network of computational resources and databases, reflecting the larger scale of the data. The established process monitors samples internally via communication with sample collectors and coordinators, sample metadata from COPO, BioSamples, ERGA data portal and ENA, raw sequencing data from ENA and genome assemblies and annotations from ENA and Ensembl. At each stage, standardised quality metrics and criteria are used to evaluate the quality of the collected data, metadata and generated genomes and annotations.

The tracking console discussed above, GTC, facilitates coordination between sampling, sequencing and assembly teams by tracking collection status and sample metadata submission (COPO), sequencing production status, read data submission and genome assembly status. It communicates target lists and project status to Genomes on a Tree (GoaT, Challis et al. (2023), a datastore and search index for genome-relevant metadata and sequencing project plans and statuses) and sends active email notifications to relevant teams when samples or data are ready for processing. Flexible interaction with the console was facilitated by a programmatically accessible RESTful API (CNAG 2025) and by data upload and download capabilities via the user interface.

A significant achievement has been the development of mechanisms for linking data types, metadata and bioinformatic pipelines. By implementing consistent identifier schemes and metadata standards, we have created robust connections between physical specimens, digital specimens, molecular data and analysis results, supported by brokering data platforms, such as COPO and PlutoF, to ensure metadata quality is held in open repositories, such as BioSamples, BioImage Archive, ENA, BOLD, Ensembl and B2DROP.

Furthermore, outputs that go beyond data, extending to pipelines and procedures, have been made available to the community under the aegis of the Biodiversity Genomics Europe (BGE) project. Computational pipelines, released through the WorkflowHub BGE container (WorkflowHub 2025a) and Standard Operating Procedures (published through

the WorkflowHub BGE container and through protocols.io, protocols.io (2025)) were used, including metadata submission guidelines developed for the sampling activities and shared through the WorkflowHub Registry.

Furthermore, to facilitate metadata alignment across both streams, there has been a move towards adopting RO-Crate as a standardised packaging format. This has been effective in maintaining these relationships throughout the (meta-)data lifecycle and in ensuring full provenance for each research output, particularly through the use of specialised RO-Crate profiles to describe workflows and their executions. For example, COPO data records can be exported as RO-Crates via the COPO API, pipelines registered on WorkflowHub can be exported as Workflow RO-Crates (WorkflowHub 2025c) and the Galaxy platform (used for some of the barcode stream tools) can export Workflow Run RO-Crates (Leo et al. 2024, ResearchObject 2025a) describing specific pipeline executions. All of these features require minimal extra effort from researchers to use. Training courses given as part of the BGE project to junior researchers have shown that the project's pipelines, published on WorkflowHub, can be launched directly on Galaxy Europe or deployed on local HPC systems. To further solidify the metadata conventions used throughout the lifecycle and repositories, an RO-Crate Profile (e-Science Lab 2025) is currently being developed for BGE data objects. The Profile will provide clear guidance on describing provenance from sample collection to barcode and genome assembly and analysis, including the people who contributed at each stage. The Profile will be registered in the RO-Crate Profile (ResearchObject 2025b), which is currently under active development and will be officially launched in early 2026.

Validation and quality control

Automated validation procedures verify data quality, coherence and compliance with community standards before data release, ensuring the wider biodiversity genomics community can have high confidence in all data and metadata from the BGE project. These procedures incorporate standardised metrics and algorithms for quality assessment, enabling meaningful cross-comparison of results within and between streams.

For barcode data, we needed to implement automated validation of sequence quality and metadata completeness, while still allowing for expert review of taxonomic assignments. Metadata submitted for the community sampling followed a submission protocol via PlutoF, which requires the submitter to complete the minimum mandatory fields and comply with the standards therein.

A standardised reporting document, the ERGA Assembly Report (EAR), was developed. This structured, transparent and reproducible document consolidates essential standardised metrics, including assembly statistics, visualisation analyses (e.g. Hi-C contact maps and k-mer spectra plots), contamination screening results, detailed notes on the curation process and additional contextual information aligned with Earth BioGenome Project (EBP) criteria (Wilkinson et al. 2016).

Alongside the EAR, a decentralised peer-review process was established to facilitate robust, community-driven evaluation of each genome assembly. Assembly experts were systematically engaged through a transparent GitHub-based review method. Submissions triggered automated checks and notifications managed by a GitHub Actions bot, streamlining reviewer assignments and ensuring a balanced, fair distribution of review responsibilities.

The interactive review process promoted iterative assembly refinement, collaboration and knowledge-sharing between researchers and reviewers, ultimately verifying that assemblies met the high-quality benchmarks required for reference-quality genome data. Approved EAR documents are available in a stable repository, linking comprehensive quality evaluations directly to assemblies deposited in the ENA, adhering closely to the FAIR principles.

Pipeline development and publishing

Significant developments in bioinformatic pipeline management have been achieved. A systematic pipeline for developing and publishing has been implemented, utilising the WorkflowHub platform to ensure reproducibility and reusability. This includes standardised testing procedures and comprehensive documentation requirements for all published pipelines. The workflows developed and published as part of the BGE project can be found within the WorkflowHub Spaces for BGE and ERGA (WorkflowHub 2025b) and cover barcode library curation, building phylogenies from barcode sequences, genome assembly and genome annotation.

These outcomes represent a significant advance in biodiversity genomics data management, establishing practices and infrastructure that will benefit current and future initiatives in this field. The solutions address technical and organisational challenges, demonstrating how large-scale collaborative genomics projects can effectively coordinate their data management activities.

Common challenges

Throughout the implementation of the activities, several significant challenges emerged that are likely to be relevant for future biodiversity genomics initiatives. Many of these challenges stemmed from the inherent complexity of coordinating data management across different scientific communities and diverse technical approaches. The challenges underscore the critical importance of considering the FAIR principles.

The current landscape of biodiversity data management reveals significant fragmentation, with multiple specialised repositories, each serving distinct purposes, making data discovery and linking complex (Fig. 2). This is exemplified by The BioImage Archive, Genomes on a Tree (GoaT), the ERGA (ERGA 2025b) and Darwin Tree of Life (DToL, The Darwin Tree of Life Project Consortium (2022)) data portals, part of the broader EBI Biodiversity Portal, BOLD, GGBN and GBIF. Whilst these platforms are often

interconnected, for example, ENA is linked to both BOLD and GBIF, the lack of a unified data model and standardisation impedes seamless data integration.

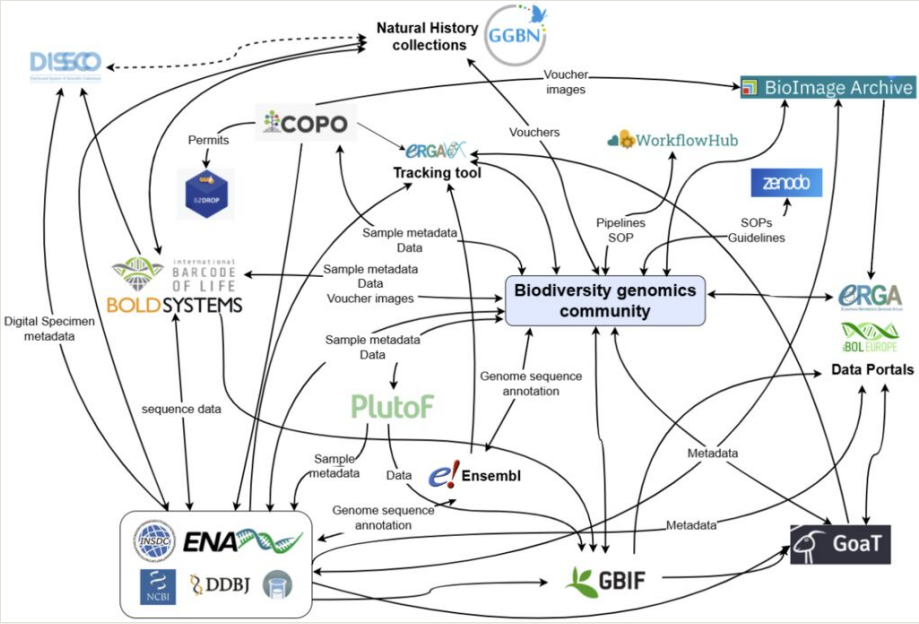

Figure 2. doi

The current (solid line) and near-future (dotted line) complex landscape of the biodiversity data infrastructure from iBOL to ERGA.

A fundamental challenge has been reconciling metadata standards between the barcoding and genome sequencing areas. While both communities require similar core information about specimens and sampling events, they have historically used different terms and structures to represent this information. The barcoding and observational communities have typically used their own standards (e.g. BOLD) or DarwinCore (e.g. GBIF and OBIS) and the genomics community has used GSC MIxS (e.g. INSDC's ENA). Efforts to align these standards, for example, at the 2024 BGE Hackathon (Vos 2025), revealed subtle, but important differences in how sampling events, specimens and derived samples are conceptualised and documented across communities. Bridging these differences required careful negotiation to ensure that neither community's specific needs were compromised, while still achieving sufficient standardisation for interoperability.

Efforts are ongoing to address these challenges by promoting standards like DwC (Wieczorek et al. 2012) and GSC MIxS (Yilmaz et al. 2011), creating mappings between existing standards and ensuring open access and comprehensive documentation to enhance data interoperability and reusability across scientific communities. Both metadata concepts and valid vocabulary 'values and ranges' need to be aligned for easier interoperability. Fortunately, some fundamental values, for example, date

annotation, are handled consistently across most standards conforming to ISO 8602, whereas, for example, country nomenclature can vary.

We note and encourage the gradual increase in the adoption of ontologies rather than flat value lists, as this allows granular mapping. Organism taxonomy is of the highest importance, with many different species taxonomies; for example, the NCBI taxonomy is widely used in the general genomics community (INSDC), while WoRMS is more common in marine observations and the barcoding community publishes data anchored in the BOLD taxonomy.

Even over a few months, standards evolve, as new and emerging concepts crystallise, so mapping is not static. The FAIR-IMPACT project (FAIR-IMPACT 2025) is currently developing guidelines for mapping processes and for generating, storing, sharing and updating them over time (Juty et al. 2025). Even when the same concepts are present in the linked repositories, the requirements for which metadata concepts are mandatory or optional may differ, affecting data reuse opportunities. Efforts, such as Bioschemas, play a role in recommending cardinality and marginality for a core set of metadata properties, assisting users in providing sufficient metadata for reuse; the BGE genome-related activities are involved in the refinement of a ‘BioSample’ (a task within the ‘Sample’ group Bioschemas (2025)) type and profile, in collaboration with a wider community membership. The efforts to increase FAIRness also make it easier to incorporate diverse data, for example, satellite oceanographic data or previous observations from other groups in the same forest.

Tracking samples, data, metadata and outputs across distributed teams and institutions proved challenging. The complexity of modern genomics processes, involving multiple processing steps at different facilities, demanded sophisticated tracking systems. However, different institutions often had existing systems and practices that needed to be accommodated. Balancing the need for standardised tracking with institutional autonomy required flexible solutions, such as the GTC, that could integrate with various local systems, while maintaining consistent project-wide monitoring.

Quality control processes presented another significant challenge, particularly in combining automated and manual assessment steps. For barcode data, we needed to implement automated validation of sequence quality and metadata completeness, while still allowing for expert review of taxonomic assignments. Similarly, genome assembly quality assessment requires automated computation of standard metrics alongside manual inspection and correction of assembly characteristics. The same holds true for the accompanying sample metadata, such as sampling location, sampling methods etc., as demonstrated by the EAR reporting and review system. Coordinating these hybrid automated and manual processes across distributed teams required careful attention to process design and documentation.

The publication of complete data packages - encompassing raw data, metadata and analysis results - posed challenges that existing databases were not fully equipped to handle. While established repositories exist for sequence data and certain metadata

types, there were gaps in the infrastructure for publishing pipeline configurations, intermediate analysis results and links between different data types. This led to the challenge of deciding what auxiliary data to preserve and how to maintain persistent connections between related data objects stored in disparate systems.

Traditionally, this role has been filled by a peer-reviewed publication; however, as the number of biodiversity genomics projects scales into the thousands, this quickly becomes infeasible. Utilising “catch-all” research objects, such as RO-Crate, which collect information on the links and provenance of all data, metadata, individuals and protocols used in each step in a machine-readable format, facilitates monitoring and publishing the full pipeline and details at each step for each project. Assigning credit and authorship for data products proved challenging due to the collaborative nature of biodiversity genomics work. Traditional academic credit systems are not well-suited for recognising the diverse contributions involved in generating a genome assembly or (meta)barcode dataset.

Particular challenges were faced in appropriately crediting field collection work, laboratory processing, bioinformatic analyses and expert curation, especially when these contributions were distributed across multiple institutions. Since this is a general challenge in biodiversity, it was systematically addressed, starting with requirements gathering from the wider biodiversity community. This was initiated at the ELIXIR BioHackathon (Brown et al. 2024) with a survey of key ‘biodiversity’ properties and continued at the German BioHackathon (Juty and Gaignard 2025), focusing on identifying core properties with Biodiversity community representatives (including Biosamples/ENA). This work also brings in a broader community from the Bioschemas working group, which is leading the development of ‘Sample’ and ‘BioSample’ types and profiles, where the ‘BioSample’ type (Bioschemas 2019) is currently being actively revised.

Using those Bioschemas/BioSample-type properties, work is ongoing to build on the basic set to develop an RO-Crate profile (described above) that better captures metadata properties common to the conceptual workflows of both BGE streams. This includes appropriate recognition or accreditation of the various actors involved in sampling and analysis processes, for example, sampling authors, preservation authors, biobanking authors etc.

Encapsulation and implementation of diverse data elements, including sample and specimen descriptions and associated resources (such as permits and material transfer agreements), tied to external links (e.g. images, taxonomic links, publications) and the software/scripts used in processing are essential for ensuring data usability and provenance. Achieving this requires an understanding of the different components and the ability to abstract and link granular provenance information as needed. Such an approach facilitates robust data integration and reuse, while maintaining transparency. Using frameworks like RO-Crate to encapsulate and link these components significantly enhances data interoperability and accessibility.

These challenges have provided valuable lessons for the genomics community and have driven the development of novel solutions, some of which are still evolving. They highlight the importance of flexible, interoperable systems and the need for community-driven standards development. Future biodiversity genomics initiatives would benefit from considering these challenges early in their planning phases.

Recommendations for best practices

Based on the described experiences, several key recommendations for implementing best data management practices in distributed biodiversity genomics projects have been developed. These recommendations focus on practical approaches that have proven effective in coordinating complex data processing flows across multiple institutions and research communities.

Utilising existing infrastructure, such as GBIF, BOLD, PlutoF, ENA, GGBN and WorkflowHub, where possible, is strongly recommended for distributed biodiversity genomics projects rather than developing new solutions. This applies particularly to data repositories, pipeline platforms and metadata catalogues. The key is identifying gaps where new development is necessary and focusing resources there. It was found that extensive use of APIs to link existing systems was more sustainable than creating new, monolithic platforms. When integrating project management tools with data infrastructure, clear guidelines should be established to enable users to perform the required activities and ensure smooth handover between teams and systems.

The following critical risk areas that require careful management were identified:

1. Loss of links between physical specimens and digital data (Groom et al. 2021);
2. Incomplete or inconsistent metadata capture;
3. Silent failures in automated pipelines;
4. Version conflicts in distributed datasets;
5. Loss of provenance information in complex pipelines;
6. Data transfer bottlenecks between institutions;
7. Communication challenges (e.g. language and culture) between specialist teams;
8. Delays in data publication due to incomplete metadata;
9. Loss of credit for contributions to data products;
10. Sustainability of project-specific infrastructure.

Projects should establish monitoring and mitigation strategies for these risks from the outset.

The presented experience has highlighted the value of using RO-Crate as a standardised packaging format for biodiversity genomics data. It is recommended to develop specific RO-Crate profiles for different data types and pipelines within any given project and related work has begun (see above). Relevant profiles should align with emerging community standards, such as the Bioschemas BioSample profile (cf. Bioschemas Sample Group, Bioschemas (2025)), while maintaining the flexibility to accommodate project-specific requirements. The RO-Crate approach has proven valuable for maintaining links between different data objects and preserving pipeline provenance information and these objects are already consumable within Galaxy and WorkflowHub. The next step after refining the object itself is developing methods for integration into existing infrastructures, such as those listed above.

From a technical infrastructure perspective, it is recommended:

- Implementing automated validation pipelines for both data and metadata;
- Establishing clear protocols for data transfer between institutions;
- Maintaining comprehensive audit trails for data processing;
- Using persistent identifiers consistently across all project components;
- Supporting both programmatic and user interfaces for data access;
- Co-create, share and periodically peer review protocols and practices to ensure they are up-to-date and robust;
- Planning for the long-term sustainability of essential infrastructure, publishing and linking (meta)data in permanent repositories with digital identifiers.

These recommendations are based on practical experience rather than theoretical ideals and recognise that their implementation may need to be adapted to specific project contexts. However, they provide a foundation for effective data management in large-scale biodiversity genomics initiatives.

Future directions

As biodiversity genomics continue to scale globally, several key areas emerge in which further development of data management approaches will be crucial. These opportunities span technical, organisational and social aspects of genomics data handling.

The rapid evolution of sequencing technologies presents both opportunities and challenges for data management. Emerging technologies promise higher throughput, longer reads and novel data types, requiring flexible data handling systems that can adapt to these advances. Future infrastructure development should anticipate these changes by emphasising modularity and extensibility in data processing pipelines.

Additionally, the increasing volume of genomic data demands more sophisticated approaches to data compression, transfer and storage.

User interfaces for data access and analysis need significant attention. While current systems serve the needs of specialist users, broader adoption of biodiversity genomics approaches requires more accessible interfaces. Future developments should focus on creating intuitive, well-documented interfaces that lower the barriers to data access while maintaining scientific rigour. This is particularly important for enabling broader participation in biodiversity research and monitoring.

A critical area for future work is the further harmonisation of metadata and data standards across the biodiversity genomics community. While BGE has made progress in aligning barcoding and genome-sequencing standards, significant work remains to create genuinely interoperable data ecosystems. This includes developing standard protocols for new data types and ensuring compatibility with emerging biodiversity data standards. Indeed, other programmes globally have identified similar issues, so it is a shared effort.

An ambitious future goal is the comprehensive integration of BGE outputs through an overarching data model. We envision a system where each component - from field collections to final analyses - has a unique identifier and can be seamlessly referenced through standardised protocols. The proposed approach of using RO-Crate as an integration layer shows promise, allowing individual components to remain in their native repositories, while maintaining machine-readable connections between them.

The role of data management in advancing biodiversity research continues to evolve. Future developments should focus on enabling novel analyses across multiple data types and supporting automated hypothesis generation through integrated data mining. This requires technical infrastructure and new approaches to data modelling and knowledge representation. AI will likely play an increasing role in aiding the capture of metadata, for example, in the lab. The data and metadata management recommendations, when implemented, increasingly facilitate the machine accessibility and interoperability of the biodiversity data and, thus, the provision of “AI-ready” data.

Credit and attribution systems need substantial development to better reflect the collaborative nature of biodiversity genomics work. Future systems should provide finer-grained recognition of contributions, including sampling work, bioinformatic analysis and data curation. Integration with persistent identifiers for people (e.g. ORCID) and institutions should be expanded and new metrics for tracking data reuse and impact should be developed.

Finally, better systems for understanding the impact and utilisation of biodiversity genomics data are needed. This includes tracking not only academic reuse, but also applications in conservation, environmental monitoring and policy development. Understanding these patterns of data use will be crucial for directing future infrastructure development and ensuring the long-term relevance of biodiversity genomics data.

These future directions represent significant challenges, but addressing them is essential to realising the full potential of biodiversity genomics for understanding, monitoring and preserving global biodiversity. The foundations laid by BGE provide a starting point for these developments, but continued coordination and investment will be needed to achieve these ambitious goals.

Abbreviations

- **AI** - Artificial Intelligence
- **API** - Application Programming Interface
- **BeGenDiv** - Berlin Center for Genomics in Biodiversity Research
- **BGE** - Biodiversity Genomics Europe
- **BOLD** - Barcode of Life Data Systems
- **BUSCO** - Benchmarking Universal Single-Copy Orthologs
- **CNAG** - Centro Nacional de Análisis Genómico
- **COPO** - Collaborative Open Plant Omics
- **DiSSCo** - Distributed System of Scientific Collections
- **DNA** - Deoxyribonucleic Acid
- **DToL** - Darwin Tree of Life
- **DwC** - Darwin Core
- **EAR** - ERGA Assembly Report
- **EBP** - Earth BioGenome Project
- **ELIXIR** - European Life Sciences Infrastructure for Biological Information
- **EMBL-EBI** - European Molecular Biology Laboratory - European Bioinformatics Institute
- **ENA** - European Nucleotide Archive
- **ERGA** - European Reference Genome Atlas
- **FAIR** - Findable, Accessible, Interoperable, Reusable
- **GBIF** - Global Biodiversity Information Facility
- **GGBN** - Global Genome Biodiversity Network
- **GoaT** - Genomes on a Tree
- **GSC** - Genomic Standards Consortium
- **GTC** - Genome Tracking Console
- **Hi-C** - High-throughput Chromosome Conformation Capture
- **HiFi** - High Fidelity
- **HPC** - High Performance Computing
- **iBOL** - International Barcode of Life
- **INSDC** - International Nucleotide Sequence Database Collaboration
- **ISO** - International Organization for Standardization
- **MIxS** - Minimum Information about any (x) Sequence
- **N50** - median contig length metric
- **NCBI** - National Center for Biotechnology Information
- **OBIS** - Ocean Biodiversity Information System
- **ORCID** - Open Researcher and Contributor ID

- **RESTful** - Representational State Transfer
- **RNA-seq** - RNA sequencing
- **RO-Crate** - Research Object Crate
- **SIB** - Swiss Institute of Bioinformatics
- **SOP** - Standard Operating Procedure
- **UB** - Universitat de Barcelona
- **UNITE** - Unified system for DNA-based fungal species identification
- **WoRMS** - World Register of Marine Species

Acknowledgements

Funding

This work was supported by Biodiversity Genomics Europe (BGE), funded by the European Union's Horizon Europe Research and Innovation Programme under grant agreement No. 101059492. BGE is co-funded by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract numbers 22.00173 and 24.00054 and by UK Research and Innovation (UKRI) under the UK government's Horizon Europe Guarantee Scheme.

Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Contributors

We thank all members of the BGE consortium for their contributions to the infrastructure, data workflows and collaborative activities described in this work. We are grateful to the broader ERGA and BIOSCAN Europe communities for their ongoing engagement and feedback.

Use of Generative AI

During the preparation of this manuscript, the authors used Claude (Anthropic) for assistance with structuring and refining the text. The authors reviewed, edited and verified all AI-assisted content and take full responsibility for the final publication.

Grant title

Biodiversity Genomics Europe

Hosting institution

Naturalis Biodiversity Center, Leiden, the Netherlands

Ethics and security

This work describes data infrastructure, workflows and metadata standards for biodiversity genomics research. No primary research involving human subjects was conducted. All DNA sequencing activities referenced in this manuscript pertain exclusively to non-human organisms collected for biodiversity documentation purposes.

The data management and coordination activities reported here did not involve the collection, processing or publication of personal data. Where human contributors are acknowledged (e.g. specimen collectors, data curators, analysts), only professional attributions and publicly available identifiers (such as ORCID iDs) are included with the contributors' consent.

Specimen collection activities conducted under the Biodiversity Genomics Europe project adhered to applicable national and international regulations, including permits for access to genetic resources where required under the Nagoya Protocol on Access and Benefit Sharing.

No ethics approval was required for this study.

Author contributions

- **Katharina F. Heil** - Concept, Writing, Review, Editing
- **Tyler S. Alioto** - Writing
- **Astrid Böhne** - Writing
- **Tom Brown** - Writing
- **Eli Chadwick** - Writing, Software development
- **Physilia Chua** - Writing, Editing
- **Christian de Guttry** - Writing, Editing, Supervision
- **Diego De Panis** - Writing
- **Carole Goble** - Supervision
- **Ivo Gut** - Supervision, Editing
- **Marta Gut** - Editing
- **Nick Juty** - Editing, Writing
- **Seanna McTaggart** - Writing, Editing
- **Laura Najera-Cortazar** - Writing, Editing
- **Joana Paupério** - Writing, Editing
- **Tomasz Rewicz** - Editing
- **Felix Shaw** - Writing, Editing, Software development
- **Stian Soiland-Reyes** - Editing
- **Robert M. Waterhouse** - Writing, Review, Editing
- **Peter Woollard** - Writing, Editing
- **Rutger A. Vos** - Concept, Writing, Review, Editing, Supervision

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Abarenkov K, Tedersoo L, Nilsson RH, Vellak K, Saar I, Veldre V, Parmasto E, Prous M, Aan A, Ots M, Kurina O, Ostonen I, Jõgeva J, Halapuu S, Põldmaa K, Toots M, Truu J, Larsson K, Kõljalg U (2010) PlutoF—a Web Based Workbench for Ecological and Taxonomic Research, with an Online Implementation for Fungal ITS Sequences. Evolutionary Bioinformatics 6 https://doi.org/10.4137/EBO.S6271
- Abarenkov K, Nilsson RH, Larsson K, Taylor AS, May T, Frøslev TG, Pawlowska J, Lindahl B, Põldmaa K, Truong C, Vu D, Hosoya T, Niskanen T, Piirmann T, Ivanov F, Zirk A, Peterson M, Cheeke T, Ishigami Y, Jansson AT, Jeppesen TS, Kristiansson E, Mikryukov V, Miller J, Oono R, Ossandon F, Paupério J, Saar I, Schigel D, Suija A, Tedersoo L, Kõljalg U (2024) The UNITE database for molecular identification and taxonomic communication of fungi and other eukaryotes: sequences, taxa and classifications reconsidered. Nucleic Acids Research 52 (D1). https://doi.org/10.1093/nar/gkad1039
- Addink W, Koureas D, Rubio AC (2019) DiSSCo as a New Regional Model for Scientific Collections in Europe. Biodiversity Information Science and Standards https://doi.org/10.3897/biss.3.37502
- Bioschemas (2019) BioSample 0.1 Release 2019_06_19. URL: http://bioschemas.org/types/BioSample/0.1-RELEASE-2019_06_19
- Bioschemas (2025) Samples. URL: http://bioschemas.org/groups/Samples
- Brown T, Collier KA, Cruz F, Gkanogiannis A, Joye-Dind S, Nevers Y, Saenko S, Alioto T, Bretaudeau A, Charleston M, Doan PD, Hahn C, Harrop TW, Herron KE, Kebaso F, Libouban R, Mansueto L, Manu S, Oba A, Swarbreck D, Syme A, Zanarello F, Aury J, Gómez-Garrido J, Dennis AB (2024) Genome Annotation and Other Post-Assembly Workflows for the Tree of Life. BioHackrXiv. https://doi.org/10.37044/osf.io/fy49g
- Challis R, Kumar S, Sotero-Caio C, Brown M, Blaxter M (2023) Genomes on a Tree (GoaT): A versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic tree of life. Wellcome Open Research 8 URL: https://wellcomeopenresearch.org/articles/8-24/v1
- CNAG (2025) API Root – Django REST framework. URL: https://genomes.cnag.cat/erga-stream/api/
- Courtot M, Gupta D, Liyanage I, Xu F, Burdett T (2022) BioSamples database: FAIRer samples metadata to accelerate research data management. Nucleic Acids Research 50 (D1). https://doi.org/10.1093/nar/gkab1046
- da Silva RF, Pottier L, Coleman T, Deelman E, Casanova H (2020) WorkflowHub: Community Framework for Enabling Scientific Workflow Research and Development. 2020 IEEE/ACM Workflows in Support of Large-Scale Science (WORKS). https://doi.org/10.1109/WORKS51914.2020.00012
- Di Tommaso P, Chatzou M, Floden E, Barja PP, Palumbo E, Notredame C (2017) Nextflow enables reproducible computational workflows. Nature Biotechnology 35 (4): 316-319. https://doi.org/10.1038/nbt.3820

- Dyer SC, Austine-Orimoloye O, Azov AG, Barba M, Barnes I, Barrera-Enriquez VP, Becker A, Bennett R, Beracochea M, Berry A, Bhai J, Bhurji SK, Boddu S, Branco Lins PR, Brooks L, Ramaraju SB, Campbell LI, Martinez MC, Charkhchi M, Cortes LA, Davidson C, Denni S, Dodiya K, Donaldson S, El Houdaigui B, El Naboulsi T, Falola O, Fatima R, Genez T, Martinez JG, Gurbich T, Hardy M, Hollis Z, Hunt T, Kay M, Kaykala V, Lemos D, Lodha D, Mathlouthi N, Merino GA, Merritt R, Mirabueno LP, Mushtaq A, Hossain SN, Pérez-Silva JG, Perry M, Piližota I, Poppleton D, Prosovetskaia I, Raj S, Salam AIA, Saraf S, Saraiva-Agostinho N, Sinha S, Sipos B, Sitnik V, Steed E, Suner M, Surapaneni L, Sutinen K, Tricomi FF, Tsang I, Urbina-Gómez D, Veidenberg A, Walsh TA, Willhoft NL, Allen J, Alvarez-Jarreta J, Chakiachvili M, Cheema J, da Rocha JB, De Silva NH, Giorgetti S, Haggerty L, Ilsley GR, Keatley J, Loveland JE, Moore B, Mudge JM, Naamati G, Tate J, Trevanion SJ, Winterbottom A, Flint B, Frankish A, Hunt SE, Finn RD, Freeberg MA, Harrison PW, Martin FJ, Yates AD (2025) Ensembl 2025. Nucleic Acids Research 53 (D1). https://doi.org/10.1093/nar/gkae1071
- ERGA (2025a) Genome Tracking Console. URL: https://genomes.cnag.cat/erga-stream/
- ERGA (2025b) European Reference Genome Atlas - Web Portal. URL: https://portal.erga-biodiversity.eu/home
- e-Science Lab (2025) Biodiversity Genomics RO-Crate Profile. URL: http://esciencelab.org.uk/bge-ro-crate-profile/bge-profile.html
- FAIR-IMPACT (2025) Expanding FAIR solutions across EOSC. URL: https://fair-impact.eu/
- Goecks J, Nekrutenko A, Taylor J, The Galaxy Team (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biology 11 (8). https://doi.org/10.1186/gb-2010-11-8-r86
- Groom Q, Dillen M, Huybrechts P, Johaadien R, Kyriakopoulou N, Fernandez FJQ, Trekels M, Wong WY (2021) Connecting molecular sequences to their voucher specimens. OSF. https://doi.org/10.37044/osf.io/93qf4
- Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. Proceedings of the Royal Society B: Biological Sciences 270 (1512): 313-321. https://doi.org/10.1098/rspb.2002.2218
- INSDC (2025) International Nucleotide Sequence Database Collaboration. URL: https://www.insdc.org/
- International Barcode of Life consortium (2025) Illuminate Biodiversity. URL: https://ibol.org/
- Juty N, Gaignard A (2025) Improving Bioschemas tooling and community support. URL: https://www.denbi.de/de-nbi-events/1618-improving-bioschemas-tooling-and-community-support
- Juty N, Le Franc Y, Goble C, Martínková J (2025) FAIR-IMPACT Task 4.4 Workshop: Developing a Mapping Process Framework. (Zenodo preprint) https://doi.org/10.5281/zenodo.15310718
- Köster J, Rahmann S (2012) Snakemake—a scalable bioinformatics workflow engine. Bioinformatics 28 (19): 2520-2522. https://doi.org/10.1093/bioinformatics/bts480
- Lander E, Linton L, Birren B, Nusbaum C, Zody M, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov J, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N,

Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin J, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston R, Wilson R, Hillier L, McPherson J, Marra M, Mardis E, Fulton L, Chinwalla A, Pepin K, Gish W, Chissoe S, Wendl M, Delehaunty K, Miner T, Delehaunty A, Kramer J, Cook L, Fulton R, Johnson D, Minx P, Clifton S, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng J, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs R, Muzny D, Scherer S, Bouck J, Sodergren E, Worley K, Rives C, Gorrell J, Metzker M, Naylor S, Kucherlapati R, Nelson D, Weinstock G, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Smith D, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis R, Federspiel N, Abola AP, Proctor M, Roe B, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blöcker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey J, Bateman A, Batzoglou S, Birney E, Bork P, Brown D, Burge C, Cerutti L, Chen H, Church D, Clamp M, Copley R, Doerks T, Eddy S, Eichler E, Furey T, Galagan J, Gilbert JR, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones T, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin E, Korf I, Kulp D, Lancet D, Lowe T, McLysaght A, Mikkelsen T, Moran J, Mulder N, Pollara V, Ponting C, Schuler G, Schultz J, Slater G, Smit AA, Stupka E, Szustakowki J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf Y, Wolfe K, Yang S, Yeh R, Collins F, Guyer M, Peterson J, Felsenfeld A, Wetterstrand K, Myers R, Schmutz J, Dickson M, Grimwood J, Cox D, Olson M, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans G, Athanasiou M, Schultz R, Patrinos A, Morgan M, International Human Genome Sequencing Consortium, Whitehead Institute for Biomedical Research CfGR, The Sanger Centre:, Washington University Genome Sequencing Center, US DOE Joint Genome Institute:, Baylor College of Medicine Human Genome Sequencing Center:, RIKEN Genomic Sciences Center:, Genoscope and CNRS UMR-8030:, Department of Genome Analysis IoMB, GTC Sequencing Center:, Beijing Genomics Institute/Human Genome Center:, Multimegabase Sequencing Center TIfSB, Stanford Genome Technology Center:, University of Oklahoma's Advanced Center for Genome Technology:, Max Planck Institute for Molecular Genetics:, Cold Spring Harbor Laboratory LAHGC, GBF—German Research Centre for Biotechnology:, *Genome Analysis Group (listed in alphabetical order aiiluoh, Scientific management: National Human Genome Research Institute UNIoH, Stanford Human Genome Center:, University of Washington Genome Center:, Department of Molecular Biology KUSoM, University of Texas Southwestern Medical Center at Dallas:, Office of Science UDoE, The Wellcome Trust: (2001) Initial sequencing and analysis of the human genome. Nature 409 (6822): 860-921. https://doi.org/10.1038/35057062

- Leo S, Crusoe M, Rodríguez-Navas L, Sirvent R, Kanitz A, Geest PD, Wittner R, Pireddu L, Garijo D, Fernández J, Colonnelli I, Gallo M, Ohta T, Suetake H, Capella-Gutierrez S, Wit Rd, Kinoshita B, Soiland-Reyes S (2024) Recording provenance of workflow runs with RO-Crate. PLOS ONE 19 (9). https://doi.org/10.1371/journal.pone.0309210

- Mazzoni C, Ciofi C, Waterhouse R (2023) Biodiversity: an atlas of European reference genomes. Nature 619 (7969): 252-252. https://doi.org/10.1038/d41586-023-02229-w
- O'Cathail C, Ahamed A, Burgin J, Cummins C, Devaraj R, Gueye K, Gupta D, Gupta V, Haseeb M, Ihsan M, Ivanov E, Jayathilaka S, Kadhirvelu V, Kumar M, Lathi A, Leinonen R, McKinnon J, Meszaros L, Pauperio J, Pesant S, Rahman N, Rinck G, Selvakumar S, Suman S, Sunthornyotin Y, Ventouratou M, Waheed Z, Woollard P, Yuan D, Zyoud A, Burdett T, Cochrane G (2025) The European Nucleotide Archive in 2024. Nucleic Acids Research 53 (D1). https://doi.org/10.1093/nar/gkae975
- protocols.io (2025) Bring structure to your research. URL: https://www.protocols.io
- Ratnasingham S, Hebert PN (2007) bold: The Barcode of Life Data System (http://www.barcodinglife.org). Molecular Ecology Notes 7 (3): 355-364. https://doi.org/10.1111/j.1471-8286.2007.01678.x
- ResearchObject (2025a) Workflow Run RO-Crate. URL: https://www.researchobject.org/workflow-run-crate/profiles/workflow_run_crate/
- ResearchObject (2025b) Research Object Crate (RO-Crate) Profiles. URL: https://www.researchobject.org/ro-crate/profiles
- Shaw F, Minotto A, McTaggart S, Providence A, Harrison P, Paupério J, Rajan J, Burgin J, Cochrane G, Kilias E, Lawniczak M, Davey R (2024) COPO - Managing sample metadata for biodiversity: considerations from the Darwin Tree of Life project [version 3; peer review: 3 approved]. Wellcome Open Research 7 (279). https://doi.org/10.12688/wellcomeopenres.18499.3
- Simão F, Waterhouse R, Ioannidis P, Kriventseva E, Zdobnov E (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31 (19): 3210-3212. https://doi.org/10.1093/bioinformatics/btv351
- Soiland-Reyes S, Sefton P, Crosas M, Castro LJ, Coppens F, Fernández J, Garijo D, Grüning B, La Rosa M, Leo S, Ó Carragáin E, Portier M, Trisovic A, RO-Crate Community, Groth P, Goble C (2022) Packaging research artefacts with RO-Crate. Data Science 5 (2): 97-138. https://doi.org/10.3233/DS-210053
- Supple M, Shapiro B (2018) Conservation of biodiversity in the genomics era. Genome Biology 19 (1). https://doi.org/10.1186/s13059-018-1520-3
- Telenius A (2011) Biodiversity information goes public: GBIF at your service. Nordic Journal of Botany 29 (3): 378-381. https://doi.org/10.1111/j.1756-1051.2011.01167.x
- The Darwin Tree of Life Project Consortium (2022) Sequence locally, think globally: The Darwin Tree of Life Project. Proceedings of the National Academy of Sciences 119 (4). https://doi.org/10.1073/pnas.2115642118
- Vos RA (2025) Enhancing Digital Infrastructures and Data Handling Practices for Single Specimen Barcoding - the 2024 BGE Barcoding Hackathon. DOI: 10.37044/osf.io/xb4ut_v1. https://doi.org/10.37044/osf.io/xb4ut_v1
- Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLOS ONE 7 (1). https://doi.org/10.1371/journal.pone.0029715
- Wilkinson M, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J, da Silva Santos LB, Bourne P, Bouwman J, Brookes A, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo C, Finkers R, Gonzalez-Beltran A, Gray AG, Groth P, Goble C, Grethe J, Heringa J, 't Hoen PC, Hooft R, Kuhn T, Kok R, Kok J, Lusher S, Martone M, Mons A, Packer A, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S, Schultes E, Sengstag T, Slater T, Strawn G, Swertz M, Thompson M, van der

Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (1). https://doi.org/10.1038/sdata.2016.18

- WorkflowHub (2025a) Biodiversity Genomics Europe (BGE). URL: https://workflowhub.eu/projects/202
- WorkflowHub (2025b) European Reference Genome Atlas (ERGA). URL: https://workflowhub.eu/projects/33
- WorkflowHub (2025c) Workflow RO-Crate profile 1.0. URL: https://about.workflowhub.eu/Workflow-RO-Crate/
- Yilmaz P, Kottmann R, Field D, Knight R, Cole J, Amaral-Zettler L, Gilbert J, Karsch-Mizrachi I, Johnston A, Cochrane G, Vaughan R, Hunter C, Park J, Morrison N, Rocca-Serra P, Sterk P, Arumugam M, Bailey M, Baumgartner L, Birren B, Blaser M, Bonazzi V, Booth T, Bork P, Bushman F, Buttigieg PL, Chain PG, Charlson E, Costello E, Huot-Creasy H, Dawyndt P, DeSantis T, Fierer N, Fuhrman J, Gallery R, Gevers D, Gibbs R, Gil IS, Gonzalez A, Gordon J, Guralnick R, Hankeln W, Highlander S, Hugenholtz P, Jansson J, Kau A, Kelley S, Kennedy J, Knights D, Koren O, Kuczynski J, Kyrpides N, Larsen R, Lauber C, Legg T, Ley R, Lozupone C, Ludwig W, Lyons D, Maguire E, Methé B, Meyer F, Muegge B, Nakielny S, Nelson K, Nemergut D, Neufeld J, Newbold L, Oliver A, Pace N, Palanisamy G, Peplies J, Petrosino J, Proctor L, Pruesse E, Quast C, Raes J, Ratnasingham S, Ravel J, Relman D, Assunta-Sansone S, Schloss P, Schriml L, Sinha R, Smith M, Sodergren E, Spor A, Stombaugh J, Tiedje J, Ward D, Weinstock G, Wendel D, White O, Whiteley A, Wilke A, Wortman J, Yatsunenko T, Glöckner FO (2011) Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nature Biotechnology 29 (5): 415-420. https://doi.org/10.1038/nbt.1823