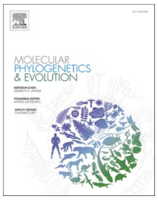
Molecular Phylogenetics and Evolution 206 (2025) 108324

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev



Using phylogenetic data for island biogeography analyses: The DAISIEprep package

Joshua W. Lambert [a], Lizzie Roeble [a,b], Théo Pannetier [a,c], Rampal S. Etienne [a,1], Luis Valente [a,b,1,*]

[a] *Groningen Institute for Evolutionary Life Sciences, University of Groningen, Box 11103, 9700 CC Groningen, the Netherlands*
[b] *Naturalis Biodiversity Center, Darwinweg 2, 2333 CR Leiden, the Netherlands*
[c] *Biological and Environmental Sciences, University of Stirling, Stirling FK9 4LA, United Kingdom*

ARTICLE INFO

Keywords:
DAISIE
Data science
Island biogeography
Macroevolution
Phylogenetics
R package

ABSTRACT

New methodologies to infer past evolutionary, ecological and biogeographical processes from molecular phylogenies are rapidly being developed. However, these often employ unfamiliar data structures that may pose a barrier to their use. DAISIE (Dynamic Assembly of Islands through Speciation, Immigration and Extinction) is an island biogeography model that can estimate rates of colonisation, speciation and extinction from molecular phylogenetic data across insular assemblages. The method uses an unconventional phylogenetic data structure: instead of considering a single island lineage, it focuses on multiple independent lineages descending from different colonisation events of the island. While analysing phylogenies from this perspective has plenty of potential, this comes with challenges for the user. Here we describe software DAISIEprep, an R package to aid the extraction of data from one or many phylogenetic trees to generate and visualise data in a format interpretable by macroevolutionary and biogeographical inference models. DAISIEprep includes simple algorithms to extract data on island colonists and account for biogeographical, topological and taxonomic uncertainty. It also allows flexible incorporation of either missing species or entire insular lineages when molecular data are not available. The software enables reproducible and user-friendly data extraction, formatting and visualisation of phylogenetic data from island lineages, and will facilitate addressing questions about island evolution, community ecology and anthropogenic impacts in insular systems. The tools presented here will also be useful for researchers who do not plan to use DAISIE but are interested in how to interpret, visualise and analyse phylogenetic datasets of islands species or island-like environments.

1. Introduction

Phylogenetic trees are prevalent in many fields of science, including evolution, epidemiology, and linguistics. Well-defined standard tree data structures, particularly the commonly used 'Newick' format (Maddison et al. 1997; Felsenstein 2004; Cardona et al. 2008), are employed within and across programming languages. For example, in the R language, phylogenetic tree formats have been constructed that are compatible with different packages and methodologies (Paradis et al. 2004; FitzJohn 2012; Revell 2012; Pennell et al. 2014; Morlon et al. 2016; R Core Team 2022). However, newly developed methods that use phylogenetic information may require different data structures, potentially creating obstacles for the scientific community to apply those methods. In particular, island biogeography (MacArthur and Wilson 1967) – the field that studies the species richness and diversification of isolated natural communities – is increasingly using phylogenetic information to answer questions on island community assembly and diversification dynamics, but it often requires phylogenetic information in an unconventional format. Instead of relying on a single phylogenetic tree representing a clade of interest, it can require data assembled from multiple phylogenetic trees of different clades and even non-phylogenetic information (Valente et al. 2020). In contrast to the well-established phylogenetic tree formats used across biogeography and evolutionary biology, standards and methods for manipulating island community data are lacking.

DAISIE (Dynamic Assembly of Island biota through Speciation,

[*] Corresponding author at: Groningen Institute for Evolutionary Life Sciences, Box 11103, 9700 CC, Groningen, The Netherlands.
E-mail address: l.m.lima.valente@rug.nl (L. Valente).

[1] These authors contributed jointly to senior authorship.

https://doi.org/10.1016/j.ympev.2025.108324
Received 6 January 2025; Accepted 7 March 2025
Available online 9 March 2025

Immigration and Extinction) is a model of island biogeography and macroevolution that focuses on entire island communities or assemblages (Valente et al. 2015, 2018), complementing a growing body of work that integrates multiple distinct clades within a given community (Webb et al. 2002; Cavender-Bares et al. 2009; Silvestro et al. 2015; Condamine et al. 2019). The model uses times of island colonisation and speciation events (extracted from molecular phylogenies) as well as the endemicity status of the island species to estimate a set of parameters (cladogenesis rate λ^c, extinction rate μ, colonisation rate γ, anagenesis rate λ^a and a carrying capacity K'). The framework is founded on the birth–death model for reconstructed phylogenetic trees (i.e. only extant species) and can incorporate diversity-dependence in colonisation and speciation rates through a carrying capacity. DAISIE has been used for various studies in island evolution and biogeography. For example, it has allowed estimating rates of speciation, colonisation and extinction of the Galápagos terrestrial avifauna (Valente et al. 2015), it enabled the first global analysis of MacArthur and Wilson's (1967) area and isolation model using phylogenetic data (Valente et al. 2020), it has put the impact of human-induced extinction in a macro-evolutionary context (Michielsen et al. 2023), it has shown the presence of colonisation rate shifts of fishes in a Japanese Lake (Hauffe et al. 2020), and it has allowed detecting the phylogenetic level of diversity-dependence of speciation and colonisation (Etienne et al. 2023).

DAISIE uses a novel (and unconventional) data structure: instead of considering a single insular lineage (e.g., the phylogeny of the Darwin's finches of the Galápagos), it focuses on multiple independent lineages descending from different colonisation events of the island, representing various lineages that make up an insular community (e.g. phylogenies for each of the birds that colonised the Galápagos). Some of these lineages may have radiated into a large clade with many species on the island, others may have only a single species or few species on the island. Some of the species may be endemic to the island (found only on the island) and some may be non-endemic (also present outside of the island, i.e. other islands or the mainland). Phylogenetic data of closely related species on the mainland are also needed to obtain an estimate of the colonisation time for each lineage. The data used by DAISIE thus includes colonisation and speciation times for the various lineages (obtained from time-calibrated molecular phylogenies), endemicity status (endemic or non-endemic to the island) and the number of missing species (how many species from each insular lineage are not represented because molecular data is not available). A key feature of the model is that it does not require all the information on the different insular lineages to come from a single phylogeny. Instead, data from different phylogenetic trees can be used (for example, phylogenetic analyses of different genera or families from different publications).

The specific format of the data required to fit DAISIE models (Fig. 1) may constitute a barrier to its use. Here we present an R package, `DAISIEprep`, to provide a comprehensive set of tools to easily create data objects from phylogenetic data and island species checklists to be analysed in the island biogeography model DAISIE (Lambert et al., 2024a). The package facilitates reproducible, fast extraction of community data from phylogenetic data (Fig. 1). It also allows inserting data on species or clades for which phylogenetic data is not available, but for which the number of species and endemicity statuses are known. As software that produces phylogenetic data – for example BEAST2 (Bouckaert et al. 2014) and MrBayes (Huelsenbeck and Ronquist 2001) among others – output phylogenetic tree using standard data structures, a package to convert these types of data into island-specific data is a useful intermediate tool between primary empirical data and model fitting. `DAISIEprep` also facilitates visualising and interpreting the

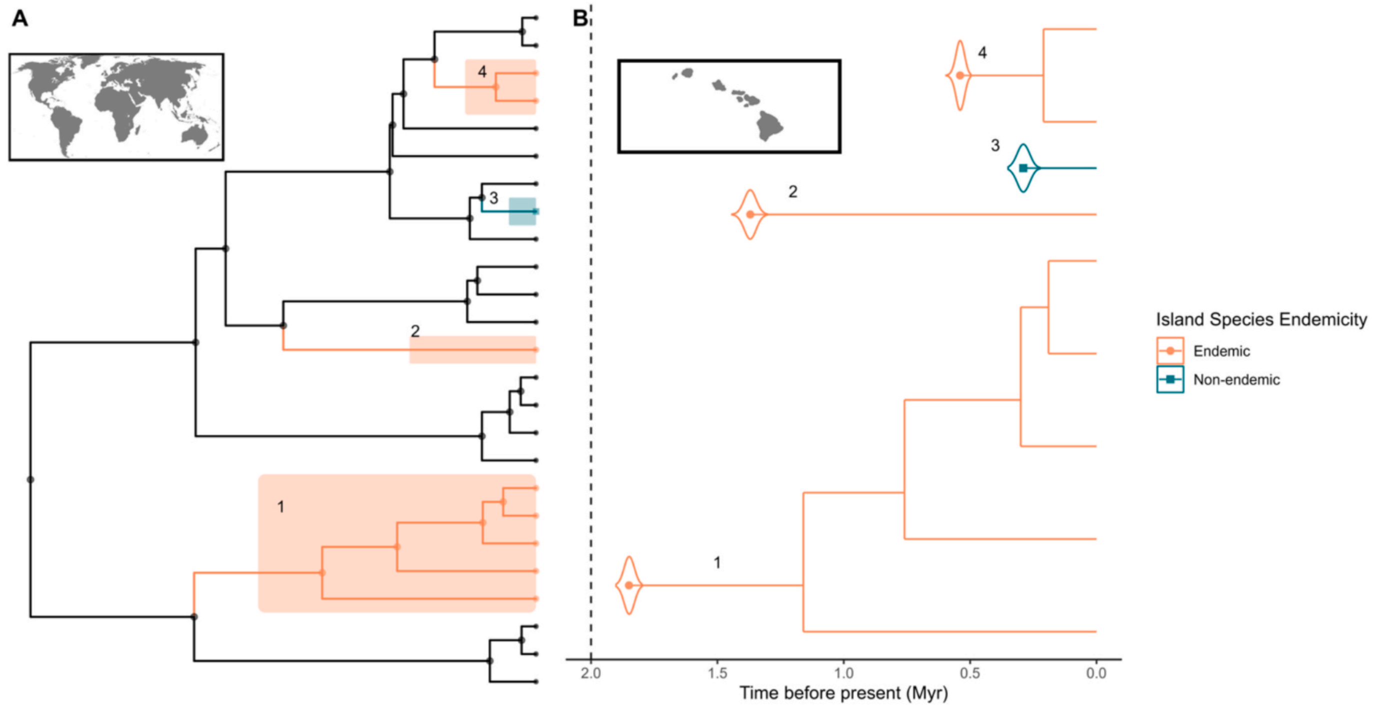


Fig. 1. Visual representation of a typical dataset used in DAISIE analyses. In this hypothetical example, a focal island community on the Hawaiian archipelago has nine species. Panel A shows the wider global phylogeny in which the nine Hawaiian species are embedded (highlighted). Panel B shows the separate phylogenies of those island species from the perspective of the archipelago. Numbers on the plots link the corresponding lineages from each panel. The island community (B) consists of lineages resulting from four colonisation events, which are spread out in different topological locations in the tree of the taxonomic group at the global scale (A). Three of the island lineages contain exclusively species that are endemic to the island and one contains only a single non-endemic species. Two of the endemic lineages have undergone cladogenetic speciation on the island, forming island clades with more than one species. The colonisation time is assumed to be the divergence time from the mainland sister species (stem age). Violin plots around the island colonisation time show the uncertainty in divergence time estimates, representing a range of times from a posterior distribution of trees. The objective of the `DAISIEprep` package is to convert global phylogenetic data (A) to island community data (B).

island phylogenetic data, providing a useful tool for researchers working on islands and island-like environments to address evolutionary and biogeographical questions. Here, we explain the main functionalities of the R package across a range of scenarios, some auxiliary functions, as well as the downstream influence of data extraction choices.

2. Methods

2.1. General data requirements

The DAISIEprep package aims to facilitate and automate extracting and formatting data prior to applying DAISIE (e.g., to estimate rates of colonisation, speciation and extinction for an island). An analysis using the DAISIE model starts with a checklist of the (native) species of the focal group that are found on an island, for example, all ferns, or all squamates. This checklist should include whether each species is endemic to the island or not. The next step is to gather phylogenetic data from which colonisation and branching times of the island species can be inferred. DAISIEprep thus requires as input a checklist of species on the island or archipelago of interest as well as time-calibrated phylogenetic data for these species.

The 'ideal' DAISIE dataset would be a single time-calibrated phylogenetic tree with complete sampling, in which all species of the focal group are included (preferably multiple individuals of the same species) as well as all of their closest non-island relatives. In practice this type of dataset is rare. For most insular communities, there may be a few well-sampled phylogenies available for some of the insular lineages, while other lineages may have poorly sampled phylogenies, and for some lineages there may be no dated phylogenetic data available at all. For some non-endemic species there may be individuals from the mainland sampled in the phylogeny, but not from the island. DAISIE can incorporate information from these heterogeneous data types.

2.2. Data extraction algorithms

The package provides two algorithms to identify island lineages and extract colonisation and island speciation times from phylogenetic data to build an island community dataset: (1) minimum time of colonisation ('min'), and (2) ancestral state reconstruction ('asr'). The 'min' algorithm assumes that the stem node of an island lineage (the split from closest non-island or island non-endemic relatives sampled in the phylogeny) corresponds to the colonisation time of the island, and that there is no back-colonisation from the island to the mainland (i.e. once endemic, a species cannot become non-endemic). The 'min' algorithm's assumption that species do not migrate away from the island is consistent with the assumptions of the original DAISIE model. Under these assumptions, the stem age represents the colonisation time. In the case when non-endemic island species or endemic island species with no sister species on the island are represented by a single tip in the phylogeny, the age of the node leading to the single tip corresponds to the colonisation time. When they are represented by several individuals or tips in the phylogeny, and in the case of endemic island clades with more than one species sampled in the phylogeny, the stem age of the island lineage corresponds to the colonisation time.

The 'min' algorithm is useful when the assumption of no back-colonisation or colonisation to other islands is plausible. However, should such colonisations have occurred, the estimated number of island colonisations may end up be inflated, as island radiations may be artificially broken up into multiple colonisations of the island. In such cases, we recommend using the 'asr' algorithm. To determine when colonisation happened, one requires a probabilistic mapping of island occupancy along the internal branches of the phylogeny, i.e. ancestral state reconstruction (see Joy et al. 2016 for review). The 'asr' algorithm is based on phylogenetic ancestral state reconstruction, and uses the ancestral geographical states at the nodes within the tree inferred with reconstruction methods to determine a likely number of colonisation events of the island. DAISIEprep provides the add_asr_node_states() function to easily estimate ancestral states, using methods implemented in the castor R package (Louca and Doebeli 2018). So far, two options to perform ancestral state reconstruction have been implemented in the package, using either maximum parsimony or a fixed-rates, continuous-time Markov model (see Louca and Doebeli 2018 for details of the methods). However, the most appropriate method of ancestral state reconstruction depends on the group being studied, so in order to offer flexibility and enable the integration of future developments, the DAISIEprep package is also set up to allow users to easily import ancestral range reconstructions obtained from external sources. The a dedicated tutorial/vignette in the DAISIEprep package provides examples of how to generate and extract ancestral ranges using R packages BioGeoBEARS (Matzke 2014), diversitree (Goldberg et al. 2011) and corHMM (Beaulieu et al. 2013).

The 'asr' algorithm prevents clades being artificially split up in the case of a back-colonisation event (Fig. S1). In the extraction process, a non-endemic species resulting from a back-colonisation is represented as an island endemic species in the Island_tbl and when fitting the DAISIE model. Therefore, the differences between the 'min' and 'asr' algorithms are both in the number of colonisation events inferred and in the endemicity of species within island clades (Fig. S1). Furthermore, under both 'min' and 'asr' algorithms, it is possible that the phylogenetic data or state reconstruction indicate a colonisation time older than the island age, for example if the closest relative of the island species has gone extinct or is not sampled in the phylogeny (Lambert et al. 2022). In such cases, the island lineage is assumed to have colonised any time after the island formation, with the DAISIE inference model integrating over this uncertainty.

2.3. General tools

DAISIEprep uses phylogenetic data stored as a phylo4d object from the phylobase R package (Hackathon et al. 2020) which can handle phylogenetic trees with data at the tips and nodes of the tree. We also introduce three novel data structures (Island_colonist, Island_tbl and Multi_island_tbl) to aid in the handling of the specific community phylogenetic data, as well as a set of functions (methods) to easily modify the data, making use of the benefits of object-oriented programming in R (see Chambers 2014). To check if the input data is compatible with DAISIEprep, the function check_phylo_data() can be used.

The package contains a set of utility functions to inspect the data and check for certain characteristics that may influence the choice of extraction algorithm ('min' or 'asr'). The plotting functions plot_phylod() and plot_colonisation() are for plotting the data before and after extraction, using the phylod object (see above) containing phylogenetic and endemicity data (Fig. S2). DAISIEprep uses ggplot2 (Wickham 2016) and ggtree (Yu et al. 2017, 2018) to produce plots, which provides flexibility for plots to be modified using standard ggplot2 layers.

Another utility function is the ability to check for back-or-onward-colonisation (any_back_colonisation()) events within the phylogeny given an ancestral state reconstruction of island presence. Identifying such colonisations can be useful to understand the prevalence of so-called boomerang colonisations (Bellemain and Ricklefs 2008). Finally, a check for any polyphyly (any_polyphyly()) on the tree identifies cases in which multiple samples of conspecific species are not found to be monophyletic in the tree. When polyphyly of a species is identified, the current extraction algorithms cannot correctly extract these data, so the user needs to decide how to assign colonisation events for that species beforehand.

2.4. Installation of DAISIEprep

The DAISIEprep package can be installed from CRAN using:

`install.packages`(“DAISIEprep”), or from Github using the remotes R package (Csárdi et al. 2021).

```
remotes::install_github(“joshwlambert/DAISIEprep”)
```

3. Applied examples

Here we demonstrate the basic usage of the DAISIEprep package through two applications to empirical data. The first is on a species-level “macro-phylogeny” of mammals by Upham et al. (2019), which we use to obtain information on the native mammals of the island of Madagascar. This represents an application that may be common when users want to extract island data from a single global macro-phylogeny containing a whole taxonomic group (e.g. birds (Jetz et al. 2012), or amphibians (Jetz and Pyron, 2018). The second empirical example is a data set of the Asteraceae plant family on the Hawaiian archipelago. Asteraceae form a hyper-diverse group of plants that is globally distributed on islands (Roeble et al. 2024). Unlike for mammals, there is yet no species-level phylogenetic tree for the Asteraceae family (>30,000 species). Thus, the Hawaiian Asteraceae dataset consists of multiple phylogenies for different Hawaiian genera or radiations (Landis et al. 2018; Knope et al. 2020; Keeley et al. 2021), as well as species richness and endemism data for those Hawaiian clades for which no molecular phylogenetic data are available. Together, these two examples illustrate the use of macro-phylogenies, single or multiple phylogenies as well as the addition of island taxa when no phylogenetic data is available. The full scripts to run the examples below as well as all necessary source data are supplied in the `DAISIEprepExtra` R package (Lambert et al., 2024b).

3.1. Single phylogeny and posterior distribution example: Mammals of Madagascar

We used the checklist of all the mammal species that inhabit Madagascar from Michielsen et al. 2023, which includes all the native mammal species found on Madagascar and their endemicity status (endemic or not endemic to Madagascar). As phylogenetic data, we used the time-calibrated maximum clade credibility mammal tree from Upham et al. 2019, which includes all extant species of mammals. We linked the Madagascar checklist table to the phylogenetic data using the `create_endemicity_status()` function and the abovementioned `phylo4d` class (this creates the phylod object in the code example below). For this example we will use the ‘asr’ algorithm. The table (checklist of species with endemicity statuses) and phylogeny are all the data required to extract the island community data using `extract_island_species()`.

```
# extract island community
island_tbl <- DAISIEprep::extract_island_species(
  phylod = phylod,
  extraction_method = “asr”
)
```

This script goes through the entire phylogeny and searches for island species and lineages (in this case the island is Madagascar) and outputs an `Island_tbl` object containing all the information on the colonisation and island speciation times, the endemicity status, the number of missing species (species present on the checklist but not sampled in the phylogeny) and the name of each colonist. By default, the number of missing species for each island colonist is set to zero. To automatically count and assign missing species to the Island_tbl using information stored in the island species checklist, `count_missing_species()`, `unique_island_genera()` and `add_multi_missing_species()` can be used. Conversely, if missing species have been misassigned in this process, they can manually be removed from the Island_tbl using `rm_multi_missing_species()`. This set of functions assists in tabulating and assigning missing species to the island community data that was directly extracted from the phylogenetic tree (using `extract_island_species()`).

To deal with the missing species that remain to be accounted for after automatic assignment, we provide a set of functions to add the missing species manually, in one of two ways, either assigning the missing species to an existing island clade (not to a specific topological location within the clade, but contributing to the total diversity of the clade) in the `Island_tbl using add_missing_species()`, or adding the missing species as a new separate island clade using `add_island_colonist()` (Fig. 2). See the tutorial vignette in the `DAISIEprep` R package for several worked examples of adding missing species under the different scenarios. When using `add_island_colonist()`, a colonisation can either be taken from the literature, extracted from the stem age of the genus if it is present in the phylogeny (i.e. the genus has mainland species in the tree) (Fig. 2B, C, D, E), or can be given as unknown (using `NA`) (Fig. 2F). An example of adding missing species in the case of the Malagasy mammals is the bat genus *Chaerephon*, which is missing from the phylogeny and needs to be assigned, but it has species sampled in the macrophylogeny not present on Madagascar. Thus a stem age can be extracted, which is used as a maximum possible colonisation time. Setting `col_max_age = TRUE` specifies that the colonisation time given (in this case, the stem age) should be considered an upper bound rather than a precise time of colonisation.

```
# extract Chaerephon genus stem age from the phylogeny
Chaerephon_stem_age <- DAISIEprep::extract_stem_age(
  genus_name = “Chaerephon”,
  phylod = phylod,
  stem = “genus”
)
# add Chaerephon island lineage with a maximum colonisation time
as its stem age
island_tbl <- DAISIEprep::add_island_colonist(
  island_tbl = island_tbl,
  clade_name = “Chaerephon_leucogaster”,
  status = “nonendemic”,
  missing_species = 0,
  col_time = Chaerephon_stem_age,
  col_max_age = TRUE,
  branching_times = NA_real_,
  min_age = NA_real_,
  species = “Chaerephon_leucogaster”,
  clade_type = 1
)
```

Once the `Island_tbl` object is finalised it can be converted to the DAISIE data list structure required by the DAISIE likelihood model, using `create_daisie_data()` function.

DAISIEprep offers tools to handle the posterior distribution of phylogenetic trees, which can be useful to explore the effect of phylogenetic uncertainty on the inferred colonisation and diversification history of the island. Here we use the posterior distribution of mammal phylogenies from (Upham et al. 2019). The island data is individually processed for each phylogeny from the posterior, using multi_extract_island_species(), which returns a Multi_island_tbl object.

```
# extract island community using asr algorithm
multi_island_tbl <- DAISIEprep::multi_extract_island_species(
  multi_phylod = multi_phylod,
  extraction_method = “asr”,
  verbose = TRUE
)
```

Thus multiple island data sets that reflect phylogenetic uncertainty are produced, with each providing an evolutionary hypothesis for the colonisation and diversification history of the island. Missing species are handled in the same way as for the single phylogeny example. Finally, `create_daisie_data()` and `DAISIE::DAISIE_ML_CS()` can be looped over each island data set in order to incorporate the phylogenetic uncertainty.

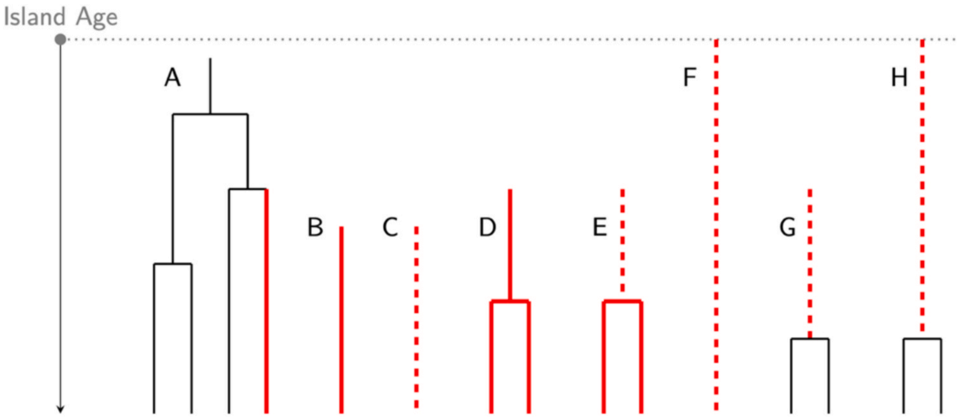


Fig. 2. Scenarios of missing species in island community data and how they can be accounted for in `DAISIEprep`. Black represents species that are sampled and branches that are known based on the phylogenetic data. Red represents missing information (for example, no molecular data available). Dashed line represents uncertainty of colonisation time. Clade A is a case when the phylogeny of an island clade is known (black lines), but one of the species is missing from the phylogeny (red line). In this scenario the add_missing_species() function would be used to add one missing species to the existing island clade in the island data set. The placement of the missing species is not known, so the missing species is not added to a specific topological location within the clade, but contributes to the total diversity of the clade. The lineage B is an island singleton (i.e. not an island radiation/clade) which is known to be on the island, but for which no time-calibrated phylogeny is available. In this case, an estimate of the time of island colonisation is assumed to exist. The lineage C is the same as B except that the time taken from the literature is considered a maximum possible colonisation time rather than an exact time of colonization. Island clade D is the same as B except that there is a clade, in this case two species, instead of a singleton on the island. Island clade E is the same as D, but the colonisation time is a maximum and not an exact time (as in C). Lineage F is the case when a species is known on the island but its time of colonisation is unknown. Island clades G and H represent cases of an island clade with two species where the stem age is known (G) or unknown (H), but for which an island speciation time is known (crown age) and can be used as a minimum time of colonisation. For examples B-H, the function used to add island lineages to the data set is add_island_colonist(). When the DAISIE model is fitted to the cases shown in C, E, F, G and H the model integrates between the maximum and minimum possible colonisation times (dashed lines). The topology of missing species (island clade A) and their branching times (island clades D and E) are not known and are shown here just for illustrative purposes. See the tutorial vignette in the R package for a full explanation of each case with examples. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.2. Multiple phylogenies example: Hawaiian Asteraceae

The case of the Asteraceae species of the archipelago of Hawaii provides a good example of a typical situation when there is a wide (and patchy) spectrum of phylogenetic and species diversity data available for an island community (Fig. 3). At the time of data compilation, the Hawaiian archipelago had 98 described native species (97 endemic) of Asteraceae, resulting from at least 10 colonization events. Each of the Hawaiian Asteraceae clades has received varying degrees of attention, contributing to differing data availability across the community. For example, the Hawaiian Silverswords and Pacific *Bidens* – two classic examples of adaptive radiations – are well-studied with dated phylogenies available. By contrast, *Tetramolopium* – a putative radiation of 11 species – does not have a phylogeny available.

We compiled a checklist of the Asteraceae species native and endemic to Hawaii and gathered all available phylogenetic data. For the 10 assumed colonization events, three clades (representing 60.5 % of the species) have dated phylogenies publicly available: Hawaiian Silversword alliance (Landis et al. 2018); *Bidens* (Knope et al. 2020); *Hesperomannia* (Keeley et al. 2021). Three clades have dated phylogenies but tree files are not publicly available, so we use the colonisation time estimates cited in the text of the publications: *Artemisia* (Hobbs and Baldwin 2013); *Lipochaeta-Melanthera* alliance (Edwards et al. 2018); *Pseudognaphalium* (Nie et al. 2016). Four clades do not have a published phylogeny, and we use the diversity data (number of species thought to belong to the same insular lineage based on taxonomic information, and endemicity status): *Tetramolopium*, *Keysseria*, *Remya*, *Adenostemma*. The resulting community dataset leverages a variety of data sources: multiple published dated phylogenies, colonisation time estimates provided in-text in scientific articles, and island diversity data (number of species and endemicity status) for cases for which no phylogenetic work has been done on the island species or clades (Fig. 3). For the cases where only crown ages were available, these were used as the minimum colonisation times (most recent possible time of colonisation).

The pipeline to extract data from multiple phylogenies is the same as

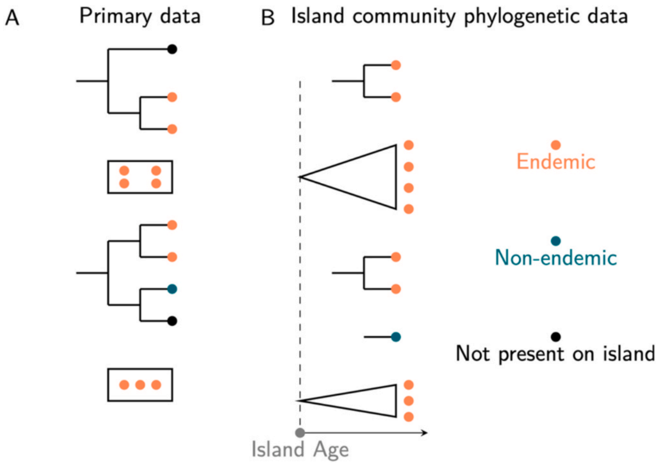


Fig. 3. Example of using different sources and formats of data to obtain island community phylogeny data. In this case, the primary data (A) is composed of two phylogenetic trees that include some of the island species and two cases where phylogenetic data is not available (represented by rectangular boxes). For the latter, diversity data (number of species present on the island assumed to belong to the same insular lineage, and their endemicity status) are available. This example shows how island species or clades that are lacking phylogenetic data but for which the island species diversity and endemism are known (shown in boxes), can be accounted for in the DAISIE framework (B) by considering them as missing species and treating the clade as a polytomy with known species richness and species endemism (shown by the triangles with the dots representing how many species are in those clades). The dashed line represents the age of the island, and those island species that lack phylogenetic data are assumed to have colonised the island any time after island formation. `DAISIEprep` provides tools to go from (A) to (B).

for the single phylogeny example, but now `extract_island_species()` is called for each of the three available phylogenies (i.e. Hawaiian Silversword alliance, *Bidens*, and *Hesperomannia*), with the argument `island_tbl` to bind to the same `Island_tbl object`. Below the example for silverswords:

```
# add silverswords to existing island_tbl
island_tbl <- DAISIEprep::extract_island_species(
  phylod = silversword_phylod,
  extraction_method = “min”,
  island_tbl = island_tbl
)
```

Island colonists that do not have phylogenetic data, but may have a known stem and crown age, are input using the function `add_island_colonist()`. Here the genus *Artemisia* is inserted with a stem and crown age estimate obtained from the literature (Hobbs and Baldwin 2013). The *Lipochaeta-Melanthera* alliance are inserted with a stem age estimated from the literature (Edwards et al. 2018) but unknown crown age (example not shown here).

```
# add Artemisia to existing island_tbl; we have the stem age (3.93 Ma) and crown
# (1.45 Ma) from Hobbs and Baldwin (2013)
island_tbl <- DAISIEprep::add_island_colonist(
  island_tbl = island_tbl,
  clade_name = “Artemisia”,
  status = “endemic”,
  missing_species = 1,
  col_time = 3.93,
  col_max_age = TRUE,
  branching_times = 1.45,
  min_age = NA_real_,
  species = c(
  “Artemisia_kauaiensis”,
  “Artemisia_mauiensis”,
  “Artemisia_australis”
  ),
  clade_type = 1
)
```

Lastly, when no phylogenetic data is available for a taxonomic group and no estimate for the colonisation time exists, a clade can be added with the maximum time of colonisation as the age of the island, and the number of species inserted from diversity data (Fig. 3). This is done for the *Keysseria* clade in our example.

```
# add Keysseria
# diversity data from taxonomic sources
island_tbl <- DAISIEprep::add_island_colonist(
  island_tbl = island_tbl,
  clade_name = “Keysseria”,
  status = “endemic”,
  missing_species = 2,
  col_time = NA_real_,
  col_max_age = TRUE,
  branching_times = NA_real_,
  min_age = NA_real_,
  species = c(
  “Keysseria_maviensis”,
  “Keysseria_erici”,
  “Keysseria_helena”
  ),
  clade_type = 1
)
```

Once a final `Island_tbl` is compiled, `create_daisie_data()` can be used to convert the data for application in DAISIE.

4. Discussion

Until now, the process of compiling phylogenetic island data to fit island biogeography models required manually locating island clades, finding the relevant nodes and extracting the times of colonisation and speciation. Here we have presented the `DAISIEprep` R package, which automates this procedure and thus allows for a more seamless pipeline to go from commonly-used phylogenetic data to model fitting. The examples presented above demonstrate the versatility of the package with a variety of input data that can be handled. The package can be used both on newly inferred phylogenies from sequence data, or previously published phylogenies, for example from a database of phylogenies (e.g. TreeBase, (Piel et al., 2000) or macrophylogenies (Jetz et al. 2012; Jetz and Pyron, 2018; Upham et al. 2019).

Even with this wealth of phylogenetic data there are still many cases for which no phylogenetic data is available or cannot be accessed (Magee et al. 2014). In such cases, the absence of a phylogeny can be mitigated by `DAISIEprep`, using island diversity and endemism data in a flexible framework. Using as much phylogenetic data available as possible is evidently still preferred, as it does improve inference reliability (Valente et al. 2018), and it improves knowledge on the number of colonisation events.

The two phylogenetic data extraction algorithms introduced in `DAISIEprep` (‘min’ and ‘asr’) include ancestral state reconstruction, either implicitly or explicitly. DAISIEprep has been set up to allow for alternative methods of ancestral state reconstruction methods. For example, if it is perceived that species range, either on the island, on the mainland or both, influences the rates of speciation and extinction, then a model that can account for these causes of rate heterogeneity can be used (Maddison 2006; Maddison et al. 2007; Holland et al. 2020). However, it is worth noting the known limitations of ancestral state reconstruction methods. Particularly, they are highly uncertain further back in time (Cunningham 1999; Martins, 2000; O'Meara and Beaulieu 2021). One encouraging aspect is that islands – especially oceanic islands – are usually young (5–10 Mya) and so species will have colonised relatively recently, making the inference of species arrival more accurate than if it were deeper in the past.

By releasing this open-source package we hope to encourage the use, development and scrutiny of phylogenetic inference models, and allow the community of phylogenetic inference to expand and answer more questions in the fields of biogeography and community assembly, while providing high reproducibility by giving clear provenance to all data extraction decisions. This package will continue development in parallel with the `DAISIE` R package.

5. Data archiving statement

The DAISIEprep R package is available in the GitHub repository (https://github.com/joshwlambert/DAISIEprep) and on CRAN (https://CRAN.R-project.org/package=DAISIEprep). All code is open-source. Code is also versioned and stored on Zenodo (Lambert et al., 2024a). The specific data and scripts quoted (i.e. code chunks) or referenced in this paper have been deposited in the DAISIEprepExtra repository: https://github.com/joshwlambert/DAISIEprepExtra (Lambert et al., 2024b).

Authors contributions

JWL formulated the project, drafted the manuscript, created and led development on the `DAISIEprep` package, and ran the analyses. LR compiled the Hawaiian Asteraceae data and assisted in writing the R script and manuscript on the multiple phylogeny example. TP wrote the extending ancestral state reconstruction vignette in the R package. LV assisted in the development of the R package and co-supervised the project. RSE co-supervised the project and gave feedback throughout. All authors revised the manuscript and approved publication.

CRediT authorship contribution statement

Joshua W. Lambert: Conceptualization, Data curation, Formal analysis, Methodology, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Lizzie Roeble:** Data curation, Methodology, Writing – review & editing. **Théo Pannetier:** Software, Writing – review & editing. **Rampal S. Etienne:** Conceptualization, Funding acquisition, Software, Supervision, Writing – review & editing. **Luis Valente:** Conceptualization, Formal analysis, Funding acquisition, Methodology, Project administration, Supervision, Visualization, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank the Theoretical and Evolutionary Community Ecology group at the University of Groningen, particularly Pedro Santos Neves, for discussions and their input. The Center for Information Technology of the University of Groningen for their support and providing access to the Peregrine/Hábrók high performance computing cluster. JWL was funded through a Study Abroad Studentship by the Leverhulme Trust and a NWO VICI grant awarded to RSE who was funded by the same grant. TP was funded by the University of Groningen and the Faculty of Natural Sciences, University of Stirling (UK). LV and LR were funded by an NWO VIDI grant awarded to LV. All authors declare no conflict of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ympev.2025.108324.

References

Beaulieu, J.M., O'Meara, B.C., Donoghue, M.J., 2013. Identifying Hidden Rate Changes in the Evolution of a Binary Morphological Character: The Evolution of Plant Habit in Campanulid Angiosperms. *Systematic Biology* 62 (5), 725–737. https://doi.org/10.1093/sysbio/syt034.

Bellemain, E., Ricklefs, R., 2008. Are Islands the End of the Colonization Road? *Trends in Ecology & Evolution* 23 (8), 461–548. https://doi.org/10.1016/j.tree.2008.05.001.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Chieh-Hsi, Wu., Xie, D., Suchard, M. A., Rambaut, A., Drummond, A.J., 2014. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Computational Biology* 10 (4), e1003537. https://doi.org/10.1371/journal.pcbi.1003537.

Cardona, G., Rosselló, F., Valiente, G., 2008. Extended Newick: It Is Time for a Standard Representation of Phylogenetic Networks. *BMC Bioinformatics* 9 (1), 532. https://doi.org/10.1186/1471-2105-9-532.

Cavender-Bares, J., Kozak, K.H., Fine, P.V.A., Kembel, S.W., 2009. The Merging of Community Ecology and Phylogenetic Biology. *Ecology Letters* 12 (7), 693–715. https://doi.org/10.1111/j.1461-0248.2009.01314.x.

Chambers, J.M., 2014. Object-Oriented Programming, Functional Programming and R. *Statistical Science* 29 (2). https://doi.org/10.1214/13-STS452.

Condamine, F.L., Romieu, J., Guinot, G., 2019. Climate Cooling and Clade Competition Likely Drove the Decline of Lamniform Sharks. *Proceedings of the National Academy of Sciences* 116 (41), 20584–20590. https://doi.org/10.1073/pnas.1902693116.

Csárdi, G., Hester, J., Wickham, H., Chang, W., Morgan, M., Tenenbaum, D. 2021. "Remotes: R Package Installation from Remote Repositories, Including 'GitHub'." https://CRAN.R-project.org/package=remotes.

Cunningham, C.W., 1999. Some Limitations of Ancestral Character State Reconstruction When Testing Evolutionary Hypotheses. *Systematic Biology* 48 (3), 665–674.

Edwards, R.D., Cantley, J.T., Chau, M.M., Keeley, S.C., Funk, V.A., 2018. Biogeography and Relationships Within the *Melanthera* Alliance: A Pan-Tropical Lineage (Compositae: Heliantheae: Ecliptinae). *Taxon* 67 (3), 552–564. https://doi.org/10.12705/673.6.

Etienne, R.S., Haegeman, B., Dugo-Cota, A., Vilà, C., Gonzalez-Voyer, A., Valente, L., 2023. The Phylogenetic Limits to Diversity-Dependent Diversification. *Systematic Biology* 72 (2), 433–445. https://doi.org/10.1093/sysbio/syac074.

Felsenstein, J., 2004. *Inferring Phylogenies*. Sinauer Associates, Sunderland, Mass.

FitzJohn, R.G., 2012. Diversitree : Comparative Phylogenetic Analyses of Diversification in R: *Diversitree*. *Methods in Ecology and Evolution* 3 (6), 1084–1092. https://doi.org/10.1111/j.2041-210X.2012.00234.x.

Goldberg, E.E., Lancaster, L.T., Ree, R.H., 2011. Phylogenetic Inference of Reciprocal Effects Between Geographic Range Evolution and Diversification. *Systematic Biology* 60 (4), 451–465. https://doi.org/10.1093/sysbio/syr046.

Hackathon, R, Ben Bolker, Marguerite Butler, Peter Cowan, Damien de Vienne, Dirk Eddelbuettel, Mark Holder, et al. 2020. "Phylobase: Base Package for Phylogenetic Structures and Comparative Data." https://CRAN.R-project.org/package=phylobase.

Hauffe, T., Delicado, D., Etienne, R.S., Valente, L., 2020. Lake Expansion Elevates Equilibrium Diversity via Increasing Colonization. *Journal of Biogeography* 47 (9), 1849–1860. https://doi.org/10.1111/jbi.13914.

Hobbs, C.R., Baldwin, B.G., 2013. Asian Origin and Upslope Migration of Hawaiian *Artemisia* (Compositae-Anthemideae). *Journal of Biogeography* 40 (3), 442–454. https://doi.org/10.1111/jbi.12046.

Holland, B.R., Ketelaar-Jones, S., O'Mara, A.R., Woodhams, M.D., Jordan, G.J., 2020. Accuracy of Ancestral State Reconstruction for Non-Neutral Traits. *Scientific Reports* 10 (1), 7644. https://doi.org/10.1038/s41598-020-64647-4.

Huelsenbeck, J.P., Ronquist, F., 2001. MRBAYES: Bayesian Inference of Phylogenetic Trees. *Bioinformatics* 17 (8), 754–775. https://doi.org/10.1093/bioinformatics/17.8.754.

Jetz, W., Pyron, R.A., 2018. The Interplay of Past Diversification and Evolutionary Isolation with Present Imperilment Across the Amphibian Tree of Life. Nature Ecology & Evolution 2 (5), 850–888. https://doi.org/10.1038/s41559-018-0515-5.

Jetz, W., Thomas, G.H., Joy, J.B., Hartmann, K., Mooers, A.O., 2012. The Global Diversity of Birds in Space and Time. *Nature* 491 (7424), 444–448. https://doi.org/10.1038/nature11631.

Joy, J.B., Liang, R.H., McCloskey, R.M., Nguyen, T., Poon, A.F.Y., 2016. Ancestral Reconstruction. *PLOS Computational Biology* 12 (7), e1004763. https://doi.org/10.1371/journal.pcbi.1004763.

Keeley, S.C., Cantley, J.T., Gallaher, T.J., 2021. The 'Evil Tribe' Spreads Across the Land: A Dated Molecular Phylogeny Provides Insight into Dispersal, Expansion, and Biogeographic Relationships Within One of the Largest Tribes of the Sunflower Family (Vernonieae: Compositae). *American Journal of Botany* 108 (3), 505–519. https://doi.org/10.1002/ajb2.1614.

Knope, M.L., Renee Bellinger, M., Datlof, E.M., Gallaher, T.J., Johnson, M.A., 2020. Insights into the Evolutionary History of the Hawaiian Bidens (Asteraceae) Adaptive Radiation Revealed Through Phylogenomics. *Journal of Heredity* 111 (1), 119–137. https://doi.org/10.1093/jhered/esz066.

Lambert, Joshua W., Pedro Santos Neves, Richél J. C. Bilderbeek, Luis Valente, and Rampal S. Etienne. 2022. "The Effect of Mainland Dynamics on Data and Parameter Estimates in Island Biogeography." Preprint. Evolutionary Biology. doi: 10.1101/2022.01.13.476210.

Lambert, J.W., Valente, L., Roeble, L., Pannetier, T., Neves, P.S., 2024a. DAISIEprep: Extracts Phylogenetic Island Community Data from Phylogenetic Trees. Zenodo. https://doi.org/10.5281/ZENODO.7653375.

Lambert, J., Valente, L., Roeble, L., 2024b. DAISIEprepExtra: Scripts and Documents to Accompany and Reproduce Lambert Et Al. Zenodo. https://doi.org/10.5281/ZENODO.7654823.

Landis, M.J., Freyman, W.A., Baldwin, B.G., 2018. Retracing the Hawaiian Silversword Radiation Despite Phylogenetic, Biogeographic, and Paleogeographic Uncertainty. *Evolution* 72 (11), 2343–2359. https://doi.org/10.1111/evo.13594.

Louca, S., Doebeli, M., 2018. Efficient Comparative Phylogenetics on Large Trees. *Bioinformatics* 34 (6), 1053–1105. https://doi.org/10.1093/bioinformatics/btx701.

MacArthur, R.H., Wilson, E.O., 1967. *The Theory of Island Biogeography*. Princeton University Press, Princeton.

Maddison, D.R., Swofford, D.L., Maddison, W.P., 1997. Nexus: An Extensible File Format for Systematic Information. *Systematic Biology* 46 (4), 590–621. https://doi.org/10.1093/sysbio/46.4.590.

Maddison, W.P., 2006. Confounding Asymmetries in Evolutionary Diversification and Character Change. *Evolution* 60 (8), 1743–2176. https://doi.org/10.1111/j.0014-3820.2006.tb00517.x.

Maddison, W.P., Midford, P.E., Otto, S.P., 2007. Estimating a Binary Character's Effect on Speciation and Extinction. *Systematic Biology* 56 (5), 701–710. https://doi.org/10.1080/10635150701607033.

Magee, A.F., May, M.R., Moore, B.R., 2014. The Dawn of Open Access to Phylogenetic Data. *PLoS ONE* 9 (10), e110268. https://doi.org/10.1371/journal.pone.0110268.

Martins, E.P., 2000. Adaptation and the Comparative Method. Trends in Ecology & Evolution 15 (7), 296–329. https://doi.org/10.1016/S0169-5347(00)01880-2.

Matzke, N.J., 2014. Model Selection in Historical Biogeography Reveals That Founder-Event Speciation Is a Crucial Process in Island Clades. *Systematic Biology* 63 (6), 951–970. https://doi.org/10.1093/sysbio/syu056.

Michielsen, N.M., Goodman, S.M., Soarimalala, V., van der Geer, A.A.E., Dávalos, L.M., Saville, G.I., Upham, N., Valente, L., 2023. The Macroevolutionary Impact of Recent and Imminent Mammal Extinctions on Madagascar. *Nature Communications* 14 (1), 14. https://doi.org/10.1038/s41467-022-35215-3.

Morlon, H., Lewitus, E., Condamine, F.L., Manceau, M., Clavel, J., Drury, J., 2016. RPANDA: An R Package for Macroevolutionary Analyses on Phylogenetic Trees. *Methods in Ecology and Evolution* 7 (5), 589–597. https://doi.org/10.1111/2041-210X.12526.

Nie, Z.-L., Funk, V.A., Meng, Y., Deng, T., Sun, H., Wen, J., 2016. Recent Assembly of the Global Herbaceous Flora: Evidence from the Paper Daisies (Asteraceae: Gnaphalieae). *New Phytologist* 209 (4), 1795–1806. https://doi.org/10.1111/nph.13740.

O'Meara, B., Beaulieu, J.M., 2021. Potential Survival of Some, but Not All, Diversification Methods. Preprint. EcoEvoRxiv. https://doi.org/10.32942/osf.io/w5nvd.

Paradis, E., Claude, J., Strimmer, K., 2004. APE: Analyses of Phylogenetics and Evolution in R Language. *Bioinformatics* 20 (2), 289–290. https://doi.org/10.1093/bioinformatics/btg412.

Pennell, M.W., Eastman, J.M., Slater, G.J., Brown, J.W., Uyeda, J.C., FitzJohn, R.G., Alfaro, M.E., Harmon, L.J., 2014. Geiger V2.0: An Expanded Suite of Methods for Fitting Macroevolutionary Models to Phylogenetic Trees. *Bioinformatics* 30 (15), 2216–2228. https://doi.org/10.1093/bioinformatics/btu181.

Piel, W.H., Donoghue, M.J., Sanderson M.J., 2000. "TreeBASE: A Database of Phylogenetic Information." *Proceedings of the 2nd International Workshop of Species*.

R Core Team, 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria https://www.R-project.org/.

Revell, L.J., 2012. Phytools: An R Package for Phylogenetic Comparative Biology (and Other Things): *Phytools: R Package*. *Methods in Ecology and Evolution* 3 (2), 217–223. https://doi.org/10.1111/j.2041-210X.2011.00169.x.

Roeble, L., Van Benthem, K.J., Weigelt, P., Kreft, H., Knope, M.L., Mandel, J.R., Vargas, P., Etienne, R.S., Valente, L., 2024. Island Biogeography of the Megadiverse Plant Family Asteraceae. *Nature Communications* 15 (1), 7276. https://doi.org/10.1038/s41467-024-51556-7.

Silvestro, D., Antonelli, A., Salamin, N., Quental, T.B., 2015. The Role of Clade Competition in the Diversification of North American Canids. *Proceedings of the National Academy of Sciences* 112 (28), 8684–8869. https://doi.org/10.1073/pnas.1502803112.

Upham, N.S., Esselstyn, J.A., Jetz, W., 2019. Inferring the Mammal Tree: Species-Level Sets of Phylogenies for Questions in Ecology, Evolution, and Conservation. *PLOS Biology* 17 (12), e3000494. https://doi.org/10.1371/journal.pbio.3000494.

Valente, L., Phillimore, A.B., Etienne, R.S., 2015. Equilibrium and Non-equilibrium Dynamics Simultaneously Operate in the Galápagos Islands. *Ecology Letters* 18 (8), 844–852. https://doi.org/10.1111/ele.12461.

Valente, L., Phillimore, A.B., Etienne, R.S., 2018. Using Molecular Phylogenies in Island Biogeography: It's about Time. *Ecography* 41 (10), 1684–2166. https://doi.org/10.1111/ecog.03503.

Valente, L., Phillimore, A.B., Melo, M., Warren, B.H., Clegg, S.M., Havenstein, K., Tiedemann, R., et al., 2020. A Simple Dynamic Model Explains the Diversity of Island Birds Worldwide. *Nature* 579 (7797), 92–96. https://doi.org/10.1038/s41586-020-2022-5.

Webb, C.O., Ackerly, D.D., McPeek, M.A., Donoghue, M.J., 2002. Phylogenies and Community Ecology. *Annual Review of Ecology and Systematics* 33 (1), 475–505. https://doi.org/10.1146/annurev.ecolsys.33.010802.150448.

Wickham, H., 2016. *ggplot2: Elegant Graphics for Data Analysis*. Second. Use R!. Springer, Switzerland. https://doi.org/10.1007/978-3-319-24277-4.

Yu, G., Lam, T.-Y., Zhu, H., Guan, Y.i., 2018. Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using *Ggtree*. *Molecular Biology and Evolution* 35 (12), 3041–3303. https://doi.org/10.1093/molbev/msy194.

Yu, G., Smith, D.K., Zhu, H., Guan, Y.i., Lam, T.-Y., 2017. Ggtree: An R Package for Visualization and Annotation of Phylogenetic Trees with Their Covariates and Other Associated Data. *Methods in Ecology and Evolution* 8 (1), 28–36. https://doi.org/10.1111/2041-210X.12628.