Post-Processing of Phylogenetic Trees

On Islands, Clumps and (Non-)Effective Overlap

By

ANA LUÍSA DE ALMEIDA SERRA JORGE DA SILVA

ORCID: 0000-0001-8020-3227

School of Earth Sciences
UNIVERSITY OF BRISTOL

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of DOCTOR OF PHILOSOPHY in the Faculty of Science.

JUNE 2022

Word count: c.56285

Abstract

Taxonomic instability in (multi)sets of phylogenetic trees is often caused by missing data, analytical artefacts and/or data incongruence due to homoplasy or loci with different evolutionary histories. This thesis focuses, primarily, on methods to subset and summarise heterogeneous (multi)sets of trees, and on an approach to mitigate the effects of non-effective overlap caused by non-random patterns of missing data.

A generalised definition of tree islands to any tree-to-tree distance metric is provided, which allows these heterogeneous tree subsets to be easily identified from any tree distribution, and not just as a byproduct of heuristic parsimony tree searches. Expanding on earlier studies, partitioned-by-island, weighted- and rarefied-by-island-size consensus methods are proposed, and the effect of islands on topology-based taxonomic instability tests explored. An R package to extract islands from trees on the same leaf set, *islandNeighbours*, is described and applied to a Bayesian tree distribution. For trees on non-identical leaf sets, a new subsetting strategy based on tree-to-supertree distances, clumps of trees, is proposed and applied to multiple tree (multi)sets with the newly developed *clumpy* Python pipeline.

An approach combining (gene-)tree jackknifing on matrix representation of splits with Concatabominations (a heuristic compatibility-based taxonomic instability test, Siu-Ting et al. 2015) is proposed to identify instances of non-effective overlap on a newly inferred caecilian Tree of Life, and also candidate loci for targeted taxon sampling with the aim of ameliorating taxonomic overlap. This approach is also compared to the mathematical gene sampling sufficiency approach.

Lastly, a morphological dataset used to illustrate the presence and effects of islands, and the effects of focal tree choice on clumps, is thoroughly reanalysed and an easily implementable tool for comparison of branch support measures across trees with identical leaf sets described and illustrated with trees inferred from a hypothetical dataset.

Dedication and acknowledgements

Todo o Mundo é feito de mudança,
Tomando sempre novas qualidades.
Luís Vaz de Camões

First and foremost, I would like to thank my supervisors Dr. Mark Wilkinson, Prof. Davide Pisani and Prof. Mike Benton for their help and feedback. Mark for his indefatigable support, patience, and ability to recognise and redirect 'ooh, shiny!'-driven headlong dives into topics that could have sustained their own four-year projects. Davide and Mike for the complete freedom to pursue increasingly non-empirical work, willingness to untangle abstruse admin queries, and perfectly timed reading suggestions.

Secondly, thank you to everyone at the NHM who helped in big and small ways. Particularly, Dr. Peter Foster, Dr. Natalie Cooper and Dr. Gustavo Burin for their helpful discussions about and input on wrangling Python, R and LaTeX; Dr. Jeff Streicher for always being willing to provide datasets and a rapt audience for clustering-based ramblings; and regular attendees of the macroevolution group and PhD student Zoom quizzes for giving me something to look forward to during lockdown.

Thirdly, I am very grateful to Dr. Karen Siu-Ting and Prof. Chris Creevey for giving me access to an updated version of the Concatabominations pipeline, and for their help troubleshooting its installation.

Also, a very big thank you to Tom Trapman, proofreader extraordinaire, occasional Maths sounding board, and, with the remainder of the Musings-crew, dispenser of laughter and coffee.

Last, but most certainly not least, thank you to my parents and family, who always recognised and encouraged my love of science.

This work was funded by a NERC GW4+ Doctoral Training Partnership grant, NE/L002434/1.

AUTHOR'S DECLARATION

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: ANA LUÍSA DE ALMEIDA SERRA JORGE DA SILVA DATE: 29/06/2022

TABLE OF CONTENTS

List of Tables

List of Figures

CHAPTER 1

INTRODUCTION

The use of trees, whether graphical depictions or written similes, to illustrate changes and relationships between objects has a long history (Pietsch 2012). However, one of, if not, the first explicit uses of a branching diagram to depict evolutionary history, which clearly resembles a modern phylogenetic tree, can be traced to Charles Darwin and his theory of descent with modification, with the only illustration featured in *On the Origin of Species* (Darwin 1859) being a tree-like diagram of descent (Fig.1.1). While there are other tree-like depictions of Life on Earth that predate Darwin's, none do so in quite the same vein. For example, Edward Hitchcock's "Paleontological Chart' resembles modern chronograms incorporating stratigraphic information (e.g., Kamei et al. 2012), but it also harkens to Aristotle's *scala naturae* and to God, not evolution, as the agent of change (Hitchcock 1840, Pietsch 2012); and, while Jean-Baptist Lamarck did link a tree-like diagram with his theory of evolution through inheritance of acquired traits, he also believed in spontaneous generation and multiple origins of life (de Lamarck 1809). Following from Darwin's diagram, tree-like depictions of evolution became ubiquitous, but they still consisted of untestable classifications—even if some where later confirmed (e.g., dinosaurs as ancestors of modern birds Agnolin et al. 2019, Fürbringer 1888)—with many retaining a view that increased complexity equalled 'more evolved' forms and that humans were the pinnacle of evolution (e.g., Haeckel 1874). It was not until the advent of systematic phylogenetics, the approach based on shared derived characters still in use today, that these branching diagrams went beyond the graphical depiction of classifications and became testable hypotheses of relatedness (Hennig 1966) that could, in turn, be used to test processes and patterns of evolution.

However, as methods for phylogenetic analysis improved (along with the inescapable advances in computing technologies), and the amount, and types, of available data increased, it became clear that there was not one, but many evolutionary histories (e.g., gene tree-species tree problem, reviewed in Degnan & Rosenberg 2009; morphology *vs.* molecules, Swofford 1991). This recogni-

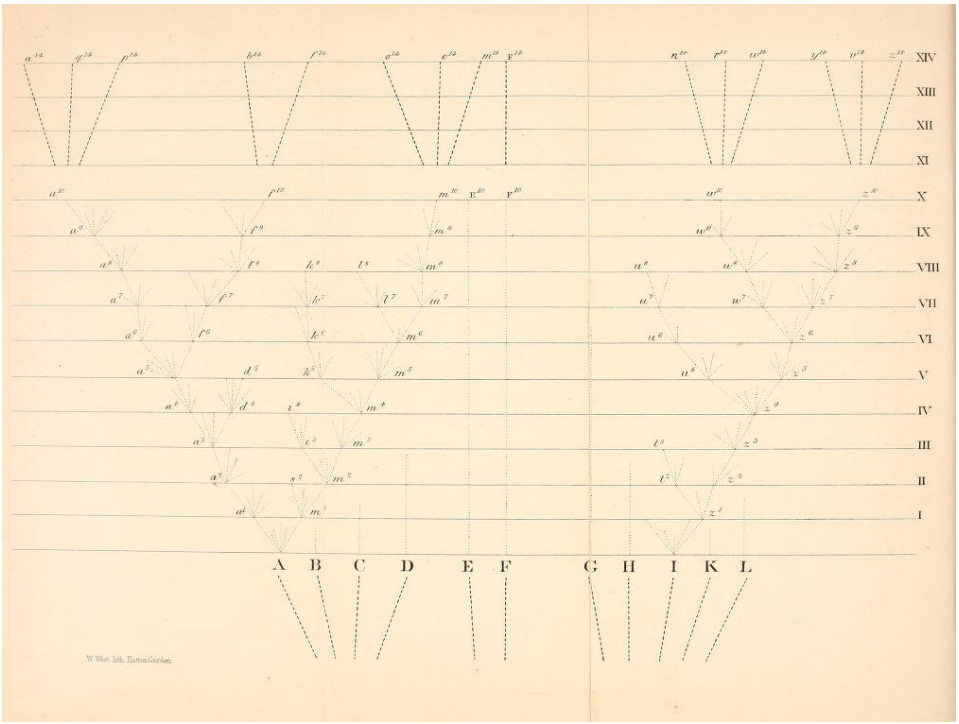


FIGURE 1.1. Tree-like diagram foldout from Charles Darwin's (1859) *On the Origin of Species*.

tion of inferred tree set heterogeneity then led to attempts to identify and summarise subsets of similar trees from heterogeneous tree distributions (e.g., Hendy et al. 1988, Maddison 1991), along with efforts to understand the causes of topological incongruence. The latter include evolutionary processes—gene paralogy, incomplete lineage sorting (ILS), reticulate events, etc. (e.g., Alexander et al. 2017, Martin & Burg 2002, Pollard et al. 2006)—missing data (e.g., Huelsenbeck 1991, Lemmon et al. 2009, Simmons 2012), and analytical artefacts (Léveillé-Bourret et al. 2017, Simmons et al. 2022). As for the former, the question of whether a single or multiple consensus trees should be used to summarise heterogeneous tree sets remains, as does the question of how to subset these tree sets (e.g., Nixon & Carpenter 1996, Tahiri et al. 2018). This thesis is an attempt to generalise some of the previously proposed approaches to subset distributions of heterogeneous trees on two fronts: extend existing subsetting strategies to easy to compute tree-to-tree distances, and to the context of trees with non-identical leaf sets. It is also an attempt to identify and mitigate the effects of non-random missing data, caused by sparse sampling, on tree inference.

1.1 Concepts and terminology

While, whenever possible, empirical data is used to test and illustrate the newly generalised and/or proposed approaches to test for, and mitigate the effects of, the presence of real and

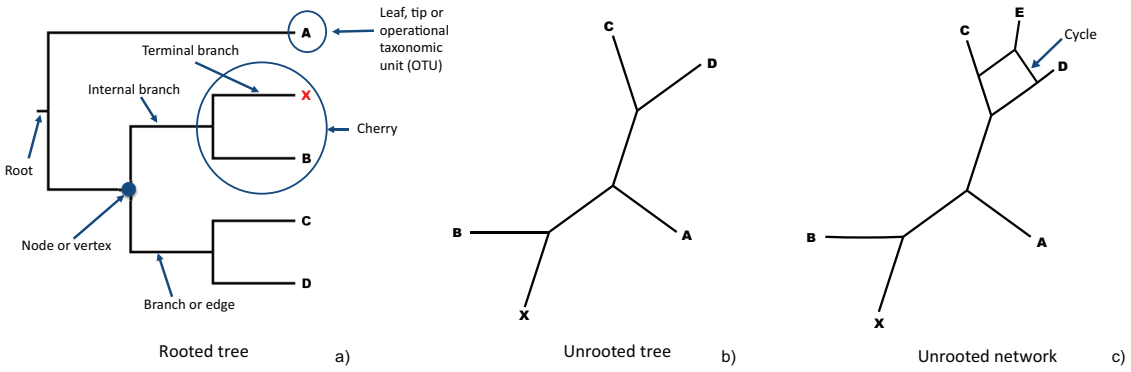


FIGURE 1.2. Basic tree terminology with a) showing a rooted and b) an unrooted tree, and c) an unrooted network.

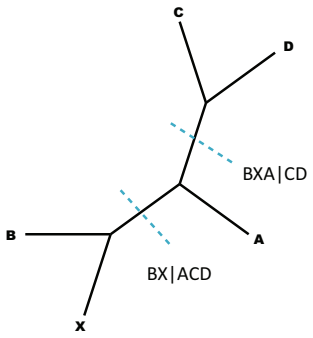
artefactual topological incongruence, this is also a piece of work that spans the biological, computational and mathematical facets of phylogenetic analyses. Because of this, the terminology used throughout this thesis switches between that commonly used by biologists and that used by mathematicians and computer scientists to refer to phylogenetic trees. I will, thus, use this section to introduce the terminology and concepts used throughout the next chapters. I will also give an overview of amphibian taxonomy, since it is the only biological model used in all chapters to follow.

1.1.1 Phylogenetic trees and networks

The mathematically-minded reader may want to refer to Semple & Steel (2003) and Steel (2016) for in depth explanations of phylogenetic trees and related concepts, and to Huson et al. (2010) for an overview of phylogenetic networks methods and concepts. While the biologically-minded reader may want to brush up on graph and set theory concepts (e.g., Balakrishnan 1997, Lipschutz 1998).

1.1.1.1 Basic terminology

Phylogenetic trees are acyclic connected graphs consisting of branches (edges) connected to nodes (vertices) and the more general phylogenetic networks are (a)cyclic connected graphs, see figure 1.2. These graphs can be directed (rooted) or undirected (unrooted), and trees are the special case of networks where no cycles are present. In unrooted phylogenetic trees, all nodes with degree = 1 are leaves (tips) and all others are internal nodes. Leaves are on terminal branches, while internal nodes are connected by internal branches. For rooted trees, the root has indegree of zero, all leaves have outdegree = 0 and any node with outdegree > 0 is an internal node. In rooted phylogenetic trees, all nodes have indegree = 1, if any node has indegree > 1



X	A	B	C	D
*	.	*	.	.
*	*	*	.	.

FIGURE 1.3. Tree splits and their notations, including a table of splits.

the acyclic graph is a network. Additionally, all leaves are uniquely labelled in both trees and networks (for the case of multilabelled trees see Huber et al. 2006).

Other useful terminology deals with tree resolution (ρ): a fully resolved tree ($\rho = 1$) can also be referred to as a binary or bifurcating tree; whereas, a partially resolved tree ($\rho < 1$) can also be referred to as unresolved or polytomous. While only briefly mentioned in Chapters 5 and 6, polytomies can be soft or hard (Maddison 1989). The latter correspond to multifurcations due to concurrent divergence events (if the tips are species then the multifurcations are concurrent speciation events), while the former denotes topological uncertainty.

1.1.1.2 Splits

Beyond graphs and parenthetical notation—e.g., the tree in figure 1.3 can be written as (A,((X,B),(C,D)))—trees can also be described as collections of splits. Splits are partitions of the leaves in a tree into two non-empty and disjoint sets. For example, the tree in figure 1.3 is made up of two splits: BX|ACD and BXA|CD. It may be useful to equate branches and splits, with the splits showing the subsets of tips on either side of a branch. Because terminal branches will be present in any tree, their corresponding splits are called trivial splits. Internal branches, on the other hand, correspond to non-trivial splits and these are the ones we are interested in when exploring topological incongruence. Lastly, full splits show the full set of tips on the input tree(s), while partial splits show a subset of all tips.

1.1.2 Consensus trees and supertrees

Supertrees are phylogenetic trees that summarise a set of input trees, and a consensus tree is the special case of a summary tree whose input set is made up of trees with identical leaf sets (Fig. 1.4). While there is a very rich literature about consensus (e.g., Adams 1986, Bonnard et al. 2006, Bryant 2003, Wilkinson 1994) and supertree methods (e.g., Baum & Ragan 2004, Gordon 1986, Wilkinson & Cotton 2006, Wilkinson et al. 2005), for this thesis only an understanding of the strict (Sokal & Rohlf 1981) and majority-rule (Margush & McMorris 1981) consensus methods

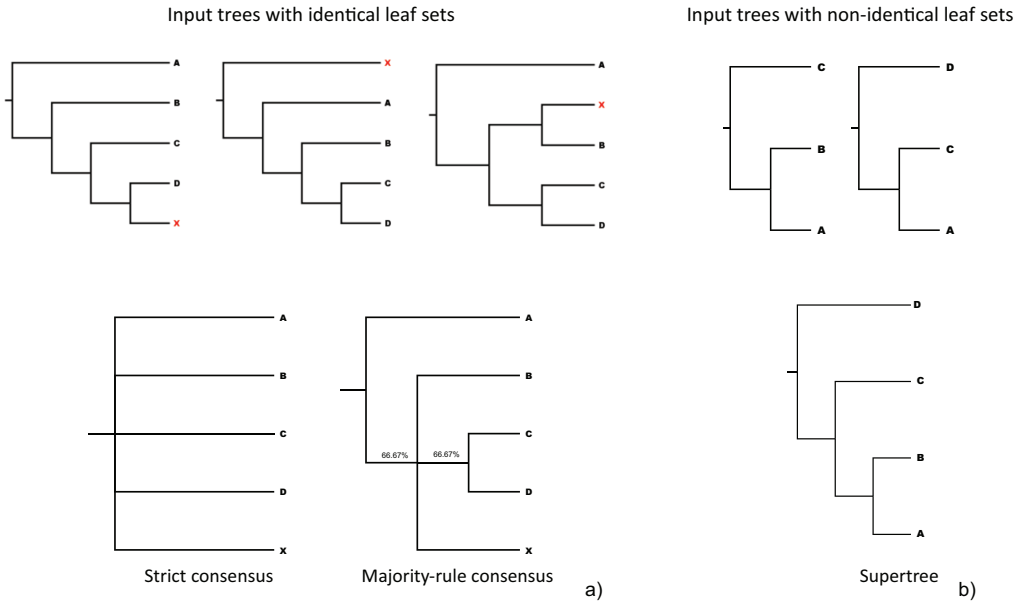


FIGURE 1.4. Tree summaries, a) shows the case where consensus trees are used (trees with identical leaf sets), and b) shows the supertree case (trees with non-identical leaf sets).

is required. The strict consensus tree displays only those splits that are common to all trees in the input set, whereas the majority-rule consensus displays all (non-conflicting) splits present in at least 50% of the input trees (Fig. 1.4a).

1.1.3 Amphibian taxonomy

Given that amphibian phylogenies are used in all chapters of this thesis, a short summary of Lissamphibia taxonomy is warranted. Crown group Lissamphibia includes all descendants of the last common ancestor of the extant amphibian orders: Gymnophiona (caecilians), Caudata (salamanders) and Anura (frogs), figure 1.5. Currently, the best supported hypothesis of extant lissamphibian relationships (based on molecular and morphological data) is the Batrachia hypothesis, which places caecilians as sister to frogs and salamanders (e.g., Frost et al. 2006, Pyron & Wiens 2011, Trueb & Cloutier 1991). But there is evidence of multiple gene trees supporting the Procera (frogs sister to other lissamphibians) or Acauda (salamanders sister to caecilans and frogs) hypotheses (e.g., Hime et al. 2021, Siu-Ting et al. 2019). With the Procera hypothesis also being identified from morphological data (e.g., Vallin & Laurin 2004). However, despite some evidence for the Procera and Acauda hypotheses, most workers accept the Batrachia hypothesis of extant amphibian relationships.

The relationships between fossil amphibians, on the other hand, remain uncertain, an example being the salamander-like albanerpetontids. Phylogenetic analyses have placed them both crownward (e.g., Gardner 2001, McGowan 2002) and stemward of Gymnophiona (e.g., Mar-

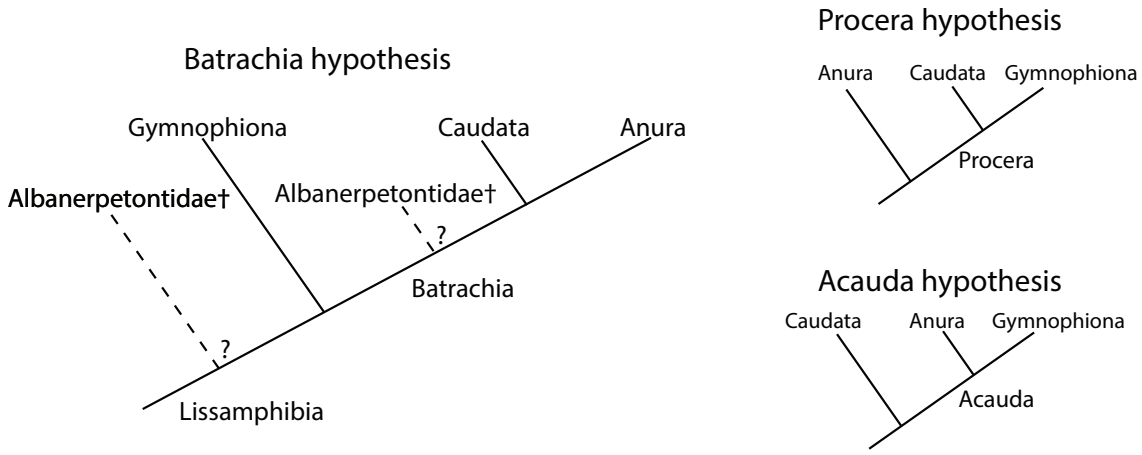


FIGURE 1.5. Traditional crown group Lissamphibia and their hypothesised relationships.

janović & Laurin 2019). In fact, Marjanović & Laurin's (2019) analyses find both placements for albanerpetontids, meaning that their phylogenetic placement remains elusive. In addition to this, some studies also raise questions of the content, potential non-monophyly of and origins of Lissamphibia (e.g., Anderson et al. 2008, Pardo et al. 2017). For a recent, in depth, review of lissamphibia (and early tetrapod) relationships see Marjanović & Laurin (2019).

1.2 Aims

In this thesis, I aim to provide approaches/tools that can be used to explore and summarise heterogeneity in (multi)sets of inferred phylogenetic trees, in both the consensus and supertree contexts. I also attempt to tackle the problem of mitigating non-effective overlap between sampled loci with inferred gene trees, and discuss strategies to compare branch support values across multiple trees. The common thread to all chapters is the recognition that the results of phylogenetic inference analyses are only as good as their underlying data, and that attempts should be made to explore and understand any inferred topological uncertainty.

In Chapter 2, I generalise the definition of islands of trees (Maddison 1991) to any tree-to-tree distance metric, describe an R (R Core Team 2019) package to extract islands from any (multi)set of trees on the same leaf set and compare it to existing approaches to subset and summarise heterogeneous (multi)sets of trees. I also address the single *vs.* multiple consensus debate by comparing three island-based consensus methods. Throughout the chapter, tree sets inferred from Pardo et al.'s (2017) data matrix are used to explore and illustrate the concepts and analyses introduced.

In the third chapter, I explore whether the presence of islands in a (multi)set of trees can be detected by topology-based tests of taxonomic instability, and, if the island structure is already known, whether there are recognisable patterns linking taxa identified as unstable and causes of

instability. Additionally, I reanalyse the Pardo et al. (2017) dataset under multiple parsimony and Bayesian inference analytical settings to better understand what is causing the extensive taxonomic instability encoded in this dataset. As part of this reanalysis, I also review the scoring of logically and/or biologically dependent characters and compare the tree sets inferred from the original Pardo et al. (2017) dataset to a modified version of it (Schoch et al. 2020) .

In Chapter 4, I define and describe the analytical pipeline for a novel approach, clumps of trees, to subsetting (multi)sets of trees with non-identical leaf sets, based on tree-to-supertree distances. I then run multiple phylogenomic datasets through this pipeline and explore the biological information relayed by the extracted clumps of trees. I also use the set of most parsimonious trees (MPTs) from Chapters 2 and 3 to compare the outputs of the clumping and island extraction analyses, and to test whether changing the supertree method alters the identified clump structure.

Chapter 5 is a departure from tree clustering analyses, dealing instead with non-effective overlap due to missing data. Here, I propose a method to inform targeted taxon/locus sampling aimed at decreasing taxonomic instability caused by non-effective overlap between sampled loci based on gene tree jackknifing, and compare it to the gene sampling sufficiency mathematical approach (Steel 2016). I then apply the proposed approach to a phylogenetic tree of caecilians, and to a selection of datasets previously analysed with gene sampling sufficiency (Dobrin et al. 2018).

Lastly, in Chapter 6, I introduce two Python (Van Rossum & Drake 2009) scripts: one to test for the non-random distribution of poor, or inflated, branch support values in a fully resolved phylogenetic tree; and another to compare probabilistic support measures for sets of trees on the same leaf set. I also discuss potential extensions to the case of (multi)sets of trees with non-identical leaf sets, and close this thesis with suggestions for further research on the topics covered.

Note: Chapter 2 is written in the first person plural, we, because it consists of a co-authored publication. The preamble to the chapter includes a short description of each author's contribution to the final text. All other chapters are written in the first person singular, I.

CHAPTER 2

ON DEFINING AND FINDING ISLANDS OF TREES AND MITIGATING LARGE ISLAND BIAS

A version of this chapter was published as:

Serra Silva, A. and Wilkinson, M. (2021), 'On defining and finding islands of trees and mitigating large island bias', *Systematic Biology* **70** (6), 1282–1294. https://doi.org/10.1093/sysbio/syab015.

ASS ran the analyses, implemented the R code, designed the x-NNI extraction algorithms, made all figures and tables, and wrote the first draft. MW designed the property algorithm and implemented the Pascal code. ASS and MW wrote the accepted manuscript.

2.1 Introduction

Phylogenetic analyses may recover multiple trees, either by design (e.g., Bayesian inference, resampling techniques) or because the data support multiple sufficiently optimal solutions. Typically, in such cases, a consensus tree is used to provide a graphical summary of the multiple trees. There are many consensus methods, but the strict (Sokal & Rohlf 1981) and the majority-rule (Margush & McMorris 1981) are among the most commonly used, and easiest to interpret. Majority-rule consensus trees display only those splits present in a majority of the input trees, and, decorated with the frequencies of occurrence of the displayed splits, are routinely used to summarise bootstrap and Bayesian analyses. Strict consensus trees display just those splits that are present in all input trees, a subset of those displayed by majority-rule trees, and are mainly used to summarise sets of most parsimonious trees (MPTs). Despite concerns that summarising MPTs with the majority-rule consensus is potentially misleading (e.g., Sharkey & Leathers 2001, Sharkey et al. 2013, Sumrall et al. 2001, Wilkinson & Benton 1996) some workers still use the

majority-rule method as if it were unproblematic (e.g., Coiffard et al. 2013, Gunnell et al. 2017, Pardo et al. 2017, Taylor et al. 2008).

It has long been appreciated that sets of inferred trees may comprise (may be partitioned into) distinct subsets (families or islands) of trees, which it may be useful to summarise separately (Hendy et al. 1988, Maddison 1991). These subsets have mostly been defined based on tree-to-tree distances (e.g., families of trees, Hendy et al. 1988), including those based on branch rearrangement metrics (e.g., tree islands, Maddison 1991). As originally defined, islands are sets of trees such that any pair is connected by a series of included trees, each of which is sufficiently similar to the adjacent members of the series. Islands were discovered through application of heuristic branch swapping to different starting trees. Initially, the main concerns were that different islands could have major implications for character evolution (Maddison 1991) and that heuristic searches could get trapped in sub-optimal islands (e.g., Olmstead et al. 1993, Olmstead & Palmer 1994), so that large numbers of starting trees should be used to improve the chances of finding all islands. In a parsimony context, it has been shown that if most of the overall variation in tree topology is between islands of trees and the islands contain very disparate numbers of trees, then the majority-rule consensus will be dominated by the largest islands. The concern is that such 'large island bias' will conceal important variation in tree topology (Sumrall et al. 2001). In the extreme, if the size of one tree subset sufficiently outnumbers all others, then the majority-rule consensus will show only those relationships found in the largest island, thus losing the information in smaller islands. Beyond parsimony, the number of such subsets, the distances between them, and their posterior probabilities can affect chain convergence in Bayesian analyses (Höhna & Drummond 2011, Lakner et al. 2008), and the presence of multiple sets of equally-optimal trees (terraces) can negatively affect tree search in maximum likelihood analyses of concatenated alignments (Sanderson et al. 2011, 2015). However, in model-based phylogenetics, subsets of trees are seldom explored outside of the tree search context, and it is thus unknown how prevalent the issue of large island bias is when summarising tree distributions obtained by Bayesian and likelihood analyses.

Here we revisit the problem of large island bias, illustrate it with a recent empirical example, investigate its cause in this case, and consider how it may be mitigated. We briefly review the use of tree-to-tree distance metrics in defining subsets of trees. We extend the concept of islands of trees to encompass multisets (weighted sets) of trees, as may result from resampling methods and Bayesian analyses, and to allow them to be based on any tree-to-tree distance. We consider how islands can be discovered *a posteriori*, and identify islands in a tree distribution recovered by Bayesian inference. We compare islands and some alternative approaches to partitioning these empirical sets of trees. We seek to highlight the potential importance of subsets of trees, such as islands, and motivate further work on their discovery and interpretation.

2.2 Defining islands of trees

The existence of distinct subsets of similar trees and implications for consensus were first considered by Hendy et al. (1988). They conceived of families of trees as subsets such that all members are only a small distance from all other members, defined a family more formally as all trees within a fixed distance from a tree T, and employed clustering based on a pairwise tree-to-tree distance to identify families in their examples. Their definition and clustering used the symmetric difference on full splits (Hendy et al. 1984), also known as the partition metric or Robinson-Foulds distance (RF, Robinson & Foulds 1981), as the tree-to-tree distance. Several similar heuristic clusterings of trees have been developed subsequently (e.g., Guénoche 2013, Stockham et al. 2002). Somewhat differently, Maddison (1991) defined the mutually exclusive subsets he denoted as tree islands, based on branch/tree rearrangement operations, with each island being the set of all trees of parsimony length $\leq L$, connected to each other through a series of included trees that differ by no more than one branch rearrangement.

Maddison (1991) focused on the branch rearrangement operations commonly used in heuristic searches of tree space, nearest-neighbour interchange (NNI), subtree prune-regrafting (SPR) and tree bisection-reconnection (TBR), which are also the bases for corresponding tree-to-tree distances (the minimum numbers of each such operation needed to convert one tree into the other). This facilitates the discovery of islands during tree search without any need for computing tree-to-tree distances and clustering. Otherwise, branch rearrangement operation metric calculations are NP-hard problems (e.g., Allen & Steel 2001, Bordewich & Semple 2005, DasGupta et al. 2000) and, despite attempts to develop efficient algorithms for calculating or approximating branch rearrangement metrics (e.g., Brown & Day 1984, DasGupta et al. 2000, Goloboff 2008, Whidden & Matsen IV 2018), *a posteriori* identification of islands defined by branch rearrangement operation metrics from sets of trees remains computationally expensive. Given that NNI rearrangements are special cases of SPR rearrangements, which are special cases of TBR rearrangements (Bryant 2004, Chernomor et al. 2015), the number of TBR islands will be less than or equal to the number of SPR islands, which in turn will be less than or equal to the number of NNI islands (Maddison 1991).

We can usefully extend Maddison's (1991) tree island definition in three ways. Firstly, we allow for islands to be defined using any tree-to-tree distance, not just those based on branch rearrangement operations. This has important consequences for the discovery of islands. Secondly, we define subsets of sufficiently optimal trees (i.e. all trees of length $\leq L$ in parsimony, or all trees with likelihood l or better in model-based inference methods), where all adjacent trees differ by less than some threshold distance (such as a maximum of x branch rearrangements rather than a single branch rearrangement, or some chosen RF value). This leads us to recognise, 1-NNI islands, which are contained in 2-NNI islands, 3-NNI islands to x-NNI islands (Fig. 2.1), and the same follows for SPR, TBR and RF islands. In this formulation, it follows that for any set of trees there will exist some categorisation under which the set comprises a single island. A similar idea,

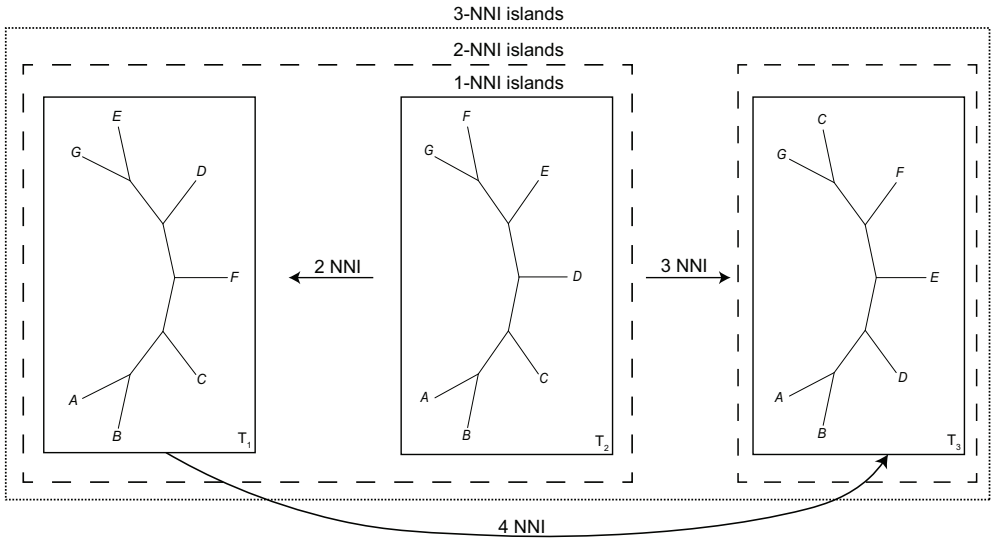


FIGURE 2.1. x-NNI islands. Three trees separated by more than one nearest neighbour interchange (NNI) branch rearrangement, and thus occupying individual 1-NNI islands (solid line boxes). With the NNI threshold increased to two, T_1 and T_2 comprise one 2-NNI island, but T_3 being more than two NNIs away from $T_{1,2}$ is in its own 2-NNI island (dashed line boxes). When the NNI threshold is three, because all trees are three or fewer NNIs from T_2, there is a single 3-NNI island comprising all trees (dotted line box). If T_2 were not present, the NNI threshold required for T_1 and T_3 to be in the same x-NNI island is four, since that is the NNI distance between T_1 and T_3.

"water level", was explored by Morrison (2007), but it dealt exclusively with the relaxation of tree optimality, not with the connectivity distance of trees within an island. Insofar as more or less substantial incongruences between trees may be better reflected by TBR or NNI distances, comparing tree islands defined using different measures and thresholds may help clarify the nature of any incongruence. Thirdly, we can remove the restriction to sufficiently optimal trees and allow islands to be defined for any given set or multiset (i.e. a weighted set) of trees. This is useful for extending the notion of tree islands to the potential multisets (where elements may be repeated) found through Bayesian inference and bootstrap resampling. Lakner et al. (2008) and Höhna et al. (2011) have both employed the notion of islands in a Bayesian context as areas of tree space with a high probability density. In the case of multisets of trees, we use island size to denote the number of distinct tree topologies in an island, island mass to denote the total number of trees in an island, and island density to denote the ratio between island size and mass.

Formally, given i) a set $\mathcal{T}$ of trees, ii) a pairwise tree distance function $d : \mathcal{T} \times \mathcal{T} \to \mathbb{R}_0^+$, and iii) a threshold $x \in \mathbb{R}_0^+$, we define an undirected, edge-weighted graph $G = (V, E)$, where $V = \mathcal{T}$ and there is an edge $(T, T') \in E$ if and only if $d(T, T') \leq x$. The *tree islands* of $(\mathcal{T}, d, x)$ correspond to connected components of graph G. Allowing the tree-to-tree distance function to take on all non-negative real numbers ($x \in \mathbb{R}_0^+$) means that this definition encompasses non-binary trees and

can be applied to tree distance metrics that take branch lengths into account.

The definition of islands, irrespective of the tree-to-tree distance metric used, leads to natural partitionings of a set of trees into mutually exclusive and exhaustive subsets. By contrast, families of trees as defined by Hendy et al. (1988) do not yield mutually exclusive subsets of trees, since a tree may be x RF units away from multiple trees ($T_1, T_2, ..., T_n$), and thus belong to multiple x-RF families. Note that in the case of binary trees, an NNI of 1 corresponds to an RF of 2 (Chernomor et al. 2015), so that a 2-RF island and a 1-NNI island are equivalent. In contrast, there is no one-to-one correspondence between an SPR (or TBR) and any RF value, so that, for example, from the equations in Chernomor et al. (2015), a 4-RF island might contain trees that are in the same 1-TBR, but not the same 1-SPR islands. Unlike branch rearrangement metric calculations, RF calculations are not an NP-hard problem and RF calculators are widely available. Thus, identifying RF islands *a posteriori* is more tractable than *a posteriori* discovery of islands based on NNIs, SPRs or TBRs. Of course, other metrics could be used to define and identify islands or other subsets of trees, and what metrics are most helpful under what circumstances remains an open question.

Distinct subsets of trees may provide insights into real biological processes and/or into our attempts to infer relationships, and thus some attention has been paid to the identification of optimal partitions and the associated question of whether a single or multiple consensus trees are required to adequately, or best, represent a set of trees (e.g., Bonnard et al. 2006, Guénoche 2013, Stockham et al. 2002). In this context, interest in islands is justified by their potential to produce natural partitions of tree space without heuristic clustering (or concern for any theoretical 'best' clustering). Islands based on branch rearrangement operations are a virtually cost-free byproduct of some searches of tree space, and, as we show, finding islands based on more readily calculated tree-to-tree distances is not intractable. Islands may have a role in investigating what number of consensus trees best summarises a tree distribution, but our primary practical purpose here is to illustrate the potential negative impact of large island bias and how this may be mitigated.

2.3 Islands and consensus

2.3.1 Example

Pardo et al. (2017) described the fossil amphibian *Chinlestegophis jenkinsi* from the Triassic of North America, and sought to infer its relationships to extant and fossil amphibians through Bayesian and parsimony analyses of a dataset comprising 76 taxa and 345 morphological characters, both summarised using the majority-rule consensus. Their Bayesian analysis provided high posterior probabilities for a close relationship of *Chinlestegophis* with extant Gymnophiona, and, in contrast to many other studies (e.g., Maddin et al. 2012, Ruta & Coates 2007), only a distant relationship between these and the other living amphibians (Anura and Caudata, collectively Batrachia). This is a surprising and potentially paradigm shifting result with major implications

for the meaning, content, age and evolutionary history of the Lissamphibia (the least inclusive clade including all living amphibians). Congruence between the majority-rule consensus from their parsimony and Bayesian analyses was used to bolster their phylogenetic conclusion.

Pardo et al.'s (2017) parsimony analysis yielded 882 equally optimal trees. Although the majority-rule consensus of these trees is highly congruent with their Bayesian analysis, it is noteworthy that none of the approximately 25 internal branches separating the Gymnophiona from their more traditional placement with Batrachia occur in every MPT and that, in a bootstrap analysis (Felsenstein 1985), none garnered support of more than 50%. These observations suggest that the parsimonious interpretation of the data offers little or no support for their novel interpretation of Lissamphibia.

Repeating the parsimony analysis, the "Tree-island profiles" in PAUP* v.4a165 (Swofford 2003) reveals that the 882 MPTs are distributed in five 1-TBR islands (that also correspond to five 1-SPR, 1-NNI and 2-RF islands) and that we refer to simply as islands. The largest island contains more than half of the trees (486), a condition under which we expect the majority-rule consensus of all the trees to be dominated by the largest island. For example, any splits that are common to all the trees in the largest island will necessarily be in the majority-rule consensus of all the MPTs. Thus, in this example, the majority-rule consensus tree of all MPTs (Fig. 2.2a) and that of the subset of these trees in the largest island (Fig. 2.2b) share 69 splits, most (64) of which are present in every tree in the largest island, with just one branch in each consensus tree that is unresolved in the other (RF = 2).

The key issue is whether island size can reasonably be taken as a proxy for support. Are inferred relationships in larger islands better supported than those in smaller islands by virtue of the relative sizes of the islands? Sumrall et al. (2001) showed, with examples of bimodal distributions of labile taxa, that subsets of trees may be larger simply because of their having greater local instability. This is not a good reason for preferring relationships in one subset over another, hence Sumrall et al.'s (2001) recommendation that palaeontologists should not use the majority-rule consensus to summarise MPTs. As we shall see, the present example illustrates this problem in a multimodal (multiple island) context, demonstrates how Sumrall et al.'s (2001) sensible advice is sometimes ignored or overlooked, and leads to the potential solutions or ameliorations we consider below.

2.3.2 Partitioned-by-island consensus

Recognising that islands may "form sets of trees that might be profitably studied separately" (Maddison 1991, p.325) and that "[c]hoosing just a single consensus tree may ignore information in the data" (Hendy et al. 1988, p.358), we can instead generate a consensus of each tree island. If we consider topological variants within islands to be minor, then computing a consensus of each island will help reveal the major variants (Maddison 1991). Applied to our example we obtain one well resolved consensus per island (Figs. 2.2b, 2.3a-d). Note that whereas we have

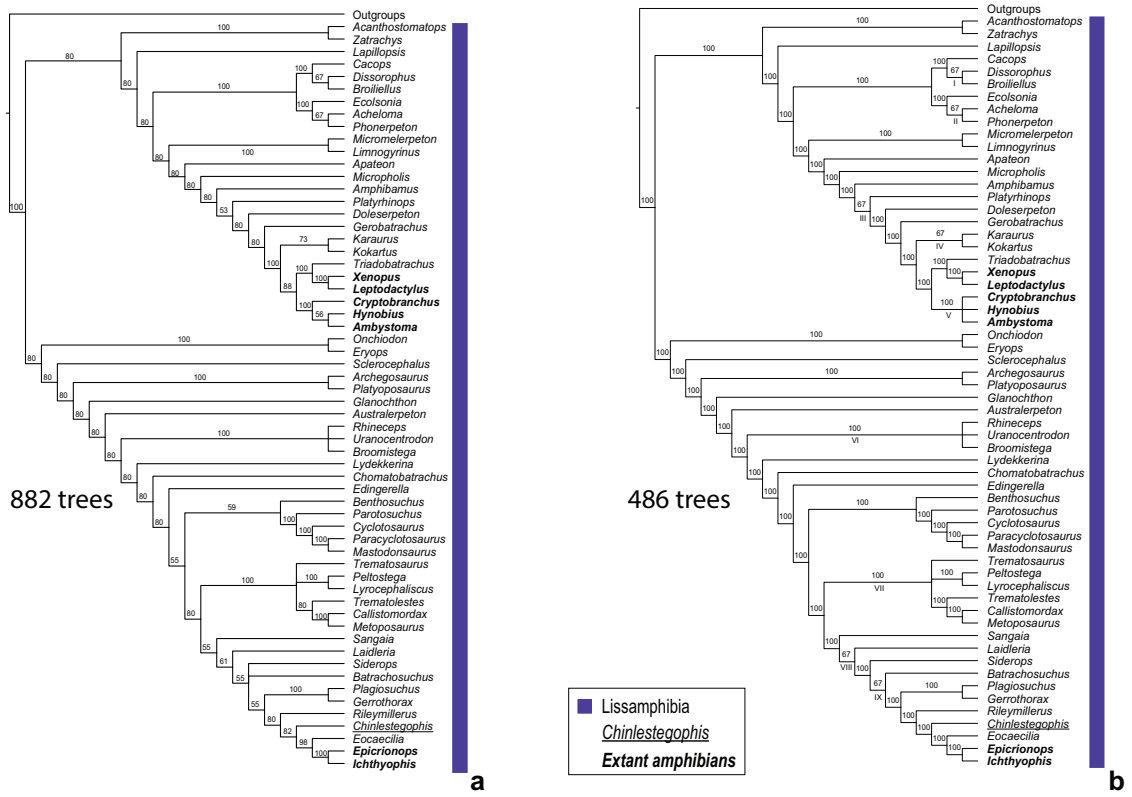


FIGURE 2.2. Majority-rule consensus trees of all MPTs from Pardo et al.'s (2017) amphibian dataset (a), and of the largest MPT island (b), with the number of MPTs present in each set. All taxa whose placement is the same in all tree islands were collapsed under the label Outgroups. Extant taxa are highlighted in bold, *Chinlestegophis jenkinsi* is underlined and the implied membership of Lissamphibia is highlighted in blue. Arabic numerals above branches correspond to branch frequencies and Roman numerals below branches indicate areas of local instability.

used the majority-rule method to produce the partitioned-by-island consensus trees, most of the displayed branches are common to all the relevant input trees and the strict consensus would have been just as useful as it has been in other studies of incongruence (e.g., Hibbett & Donoghue 2001, Soltis & Kuzoff 1995). Our results reveal that a majority, the three smallest of the five islands, feature a more traditional Lissamphibia in which Gymnophiona is closely related to Batrachia. Although an important caveat to their phylogenetic conclusions, Pardo et al. (2017) do not mention that the more traditional Lissamphibia is as parsimonious as the novel relationship inferred using Bayesian inference.

Comparison of the partitioned-by-island consensus trees sheds light on the likely cause of the substantial size disparity between islands. Each consensus shows areas of local instability (indicated either by polytomies or by branches labelled with occurrences of less than 100%). Across the five partitioned-by-island consensus trees, there are 12 areas of local instability (all reflecting up to three possible alternative resolutions corresponding to NNIs). Of these, nine

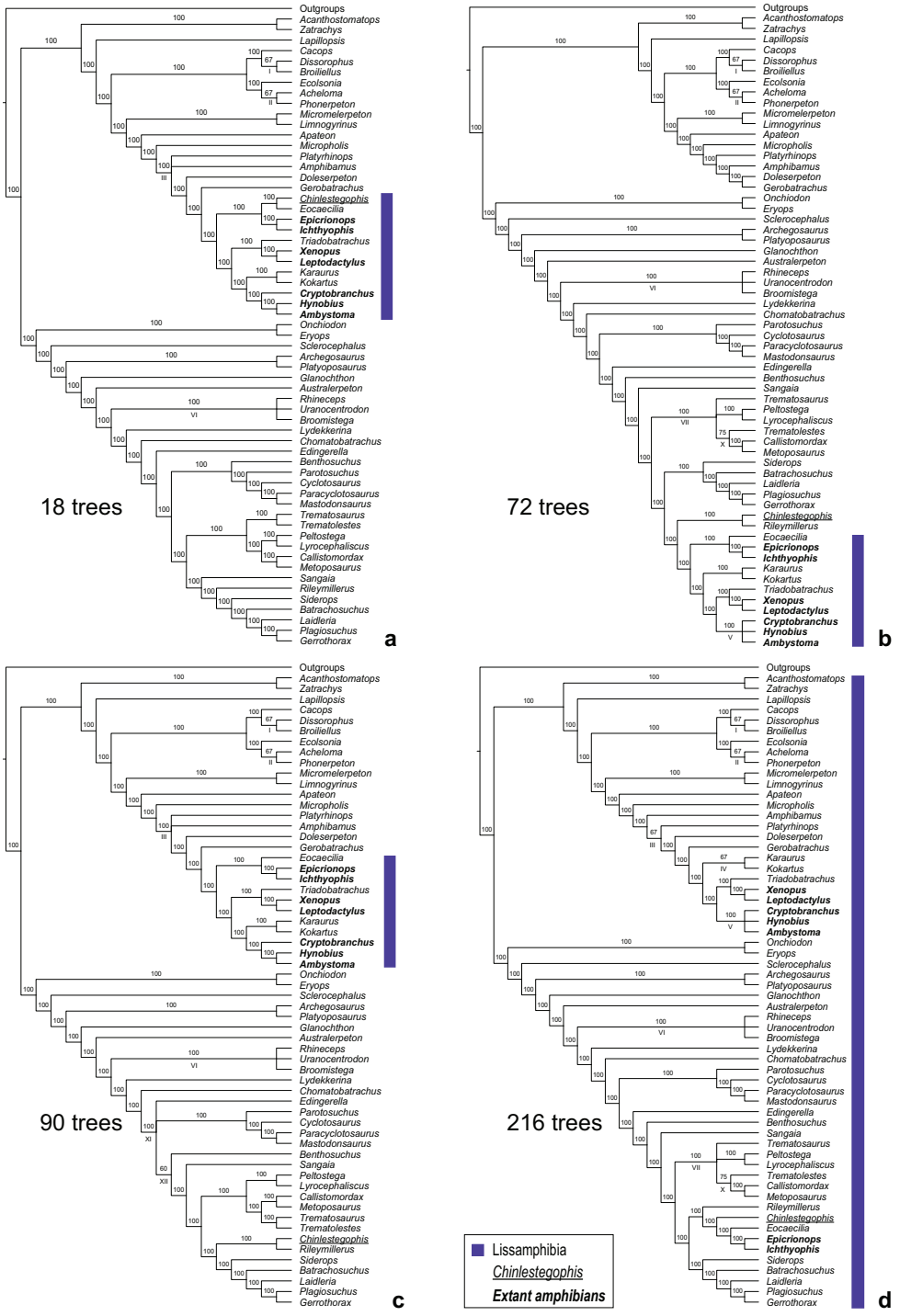


FIGURE 2.3. Majority-rule consensus trees for the four smaller tree islands recovered from the parsimony analyses of Pardo et al.'s (2017) dataset, with the number of MPTs present in each island (a-d). All taxa whose placement is the same in all tree islands were collapsed under the label Outgroups. Extant taxa are highlighted in bold, *Chinlestegophis jenkinsi* is underlined and the implied members of Lissamphibia is highlighted in blue. Arabic numerals above branches correspond to branch frequencies and Roman numerals below branches indicate areas of local instability.

are present in the largest island, while only four are present in the smallest island, and three occur in every island (Fig. 2.2b, Fig. 2.3a-d, table 2.1). Differences in island size are mostly explained by the number of instances of local instability that the island includes (Pearson's correlation test: r = 0.9053, p-value = 0.0345), each of which provides an independent source of alternative relationships. Linear increases in the number of instances of local instability i produce exponential increases (bounded by 3^i) in the number of trees as a result of their possible combinations. The link between island size and local instability is confirmed when the bounded maximum size of each island is taken as an alternate measure of instability (r = 0.9925, p-value = 0.0008). Another line of evidence that links large island size with high local instability, in this dataset, is the average degree of each island. If islands result from rearrangments around a small number of poorly resolved nodes then they would be densely connected, with lots of NNI edges connecting trees. We used NetworkAnalyzer (Assenov et al. 2007), available through Cytoscape v.3.7.1 (Shannon et al. 2003), to calculate the average degree of trees within each island, which corresponds to an NNI graph, and found that the smallest and largest islands had, respectively, the lowest (4.33) and highest (9.00) average vertex degrees (table 2.1). Furthermore, the highly positive significant correlations between local instability and island sizes were also found between number of local instabilities in and average vertex degree of an island (r = 0.9936, p-value = 0.0006), and island size and average vertex degree (r = 0.9138, p-value = 0.0290). Should island size be unrelated to the amount of local instability present in an island, we would not expect this pattern.

Consider two conflicting relationships, one present in all the trees in a small island and the other present in all trees in a larger island. We contend that confidence in any such branch should be considered independent of the combinatoric effects on island size of regions of instability in other parts of the trees. From that point of view the effect is a bias toward stable relationships in larger islands.

In this example, the partioned by island consensus approach also allows us to distinguish between major conflicts reflecting alternative placements of Gymnophiona, Batrachia and *Chinlestegophis* and more minor patterns of local instability. Among the latter it enables us to distinguish those that are contingent on, and those that are entirely independent (I, II and VI which are present in all islands) of these major conflicts.

2.3.3 Weighted-by-island-size majority-rule consensus

If we are interested in finding which relationships are supported across multiple islands (and those that are not), island size bias can be avoided by giving all islands equal weight. One means of achieving this is by, under the assumption of island equiprobability, assigning weights inversely proportional to the size of the island to which the input trees belong, so that trees in larger islands will contribute less to the consensus. Here then, trees are assigned a weight of $\frac{1}{n_i}$, where n_i = size of the i^{th} island. To implement this, tree weights can be added to Nexus format

TABLE 2.1. Areas of local instability present in each island. Presence of each area of local instability is denoted by a plus sign (+), and the islands are identified by their size. The areas identified by Roman numerals correspond to the areas of local instability labelled in figures 2.2b and 2.3a-d. The average degree for each island (NNI graph) is also provided.

Area	**18**	**72**	**90**	**216**	**486**
I	+	+	+	+	+
II	+	+	+	+	+
III	+	-	+	+	+
IV	-	-	-	+	+
V	-	+	-	+	+
VI	+	+	+	+	+
VII	-	+	-	+	+
VIII	-	-	-	-	+
IX	-	-	-	-	+
X	-	+	-	+	-
XI	-	-	+	-	-
XII	-	-	+	-	-
Total	4	6	6	8	9
Average degree	4.33	6.33	6.33	7.67	9.00

tree files by inserting the expression "[&W $\frac{1}{n_i}$]", with n_i replaced by the corresponding island size, before each tree string, and the option "usetreewts" must be set to "yes" in PAUP*. When dealing with multisets, the weight of each unique topology will be $\frac{m_t}{n_i}$, where m_t corresponds to the number of times a unique topology is present in the tree distribution. As such, in multisets the sum of tree weights for any island with repeat topologies might exceed one.

In our example, trees in the smallest island have a weight of $\frac{1}{18}$, while those in the largest island have a weight of $\frac{1}{486}$, but each island has a weight of 1. The resulting weighted-by-island-size majority-rule consensus (Fig. 2.4) is, unsurprisingly, less resolved than each partitioned-by-island consensus. It resembles a strict consensus of all the MPTs with some additional information on splits that occur (with sufficient frequency) in a majority of islands. Interestingly, the weighted-by-island-size consensus recovers a subtree, (*Siderops*,(*Batrachosuchus*,(*Laidleria*,(*Plagiosuchus*, *Gerrothorax*)))), including three splits that are present in all trees of all but the largest island but are not present in the standard majority-rule consensus. To some extent this approach mitigates against failure to acknowledge alternatives that may follow from the uncritical use and unwarranted acceptance of the majority-rule consensus highlighted by Sumrall et al. (2001).

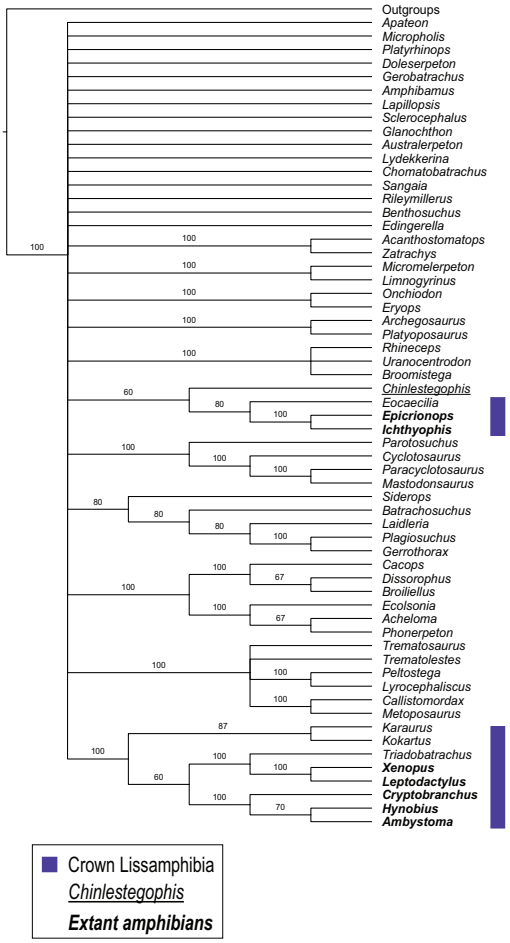


FIGURE 2.4. Weighted majority-rule consensus of all MPTs from Pardo et al.'s (2017) amphibian dataset. All taxa whose placement is the same in all tree islands were collapsed under the label Outgroups. Extant taxa are highlighted in bold, *Chinlestegophis jenkinsi* is underlined and members of the traditional Lissamphibia crown-group are highlighted in blue.

2.3.4 Rarefied-by-island-size majority-rule consensus

Alternatives to differential weighting of trees are rarefaction and oversampling. One such strategy is to randomly select n tree topologies from each island, without replacement, resulting in equal representation of each island in any consensus. The number of trees included per island can be formulated as $n = s \times p$, where s is the size of the smallest island and p the proportion of trees from the smallest island to be included, with $p \in [\frac{1}{s}, 1]$. Using $p = 1$ will minimise stochastic loss of information from the larger islands, resulting from the random draw of source trees, which may be particularly important if s is small (e.g., ≤ 20). For multisets, resampled topologies that appear more than once in the tree distribution can be weighted by the number of times they are present in the multiset.

With Pardo et al.'s (2017) MPTs, s is 18, so we set $p = 1$, giving a total of 90 trees as input to

the majority-rule consensus. To further ensure results were not unduly affected by stochastic errors introduced by the random draw of input trees, we repeated the random selection and consensus computations 1000 times (see below for implementation, for single and multiple replicates). Both the strict and standard majority-rule consensus trees of these 1000 rarefied majority-rule consensus have the same topology as the weighted-by-island-size majority-rule consensus (Fig. 2.4), emphasising that the two approaches are both attempting to remove large island bias by giving islands equal weight.

2.3.4.1 Implementation of rarefied-by-island-size consensus

The rarefied-by-island-size majority rule consensus was implemented as follows, the randomised tree sets were obtained in R v.3.5.3 (R Core Team 2019) with the following command, *t(replicate(1000, unlist(list(sort(sample.int(18, 18)), sort(sample.int(72,18) + 18), sort(sample.int(90, 18) + 90), sort(sample.int(216, 18) + 180), sort(sample.int(486, 18) + 396))))*, where the values following the addition sign correspond to the sum of the length of all islands smaller than the island being sampled. The resulting table was edited by substituting all "^" (start of line) by "\t*contree* ", and all "*$*" (end of line) by "*/ majrule=yes strict=no treefile=RarefiedM50-#.tre;*\n", the edited text was then inserted into a Nexus file containing the PAUP* block and the output file numbers edited manually. If a single replicate is desired the R command should be *c(sort(sample.int(18, 18)), sort(sample.int(72,18) + 18), sort(sample.int(90, 18) + 90), sort(sample.int(216, 18) + 180), sort(sample.int(486, 18) + 396))*, and the regular expressions for text manipulation do not need to be modified beyond changing the desired treefile name.

2.4 Extracting islands of trees and other tree clustering methods

2.4.1 *islandNeighbours* R package

To facilitate the extraction of islands under our generalised definition, we wrote an R package that makes use of two algorithms (see below) to identify and extract islands from any distribution of trees with the same leaf set. The non-tree generating functions in the *islandNeighbours* package can be applied directly to a tree distribution or to a pre-calculated tree-to-tree distance matrix. The tree-generating functions are currently restricted to the x-NNI case, but may in time be extended to the SPR and TBR cases. Additionally, all functions in the package work with unrooted trees, but there is the capacity to modify the functions to allow rooted trees to be used directly for island identification. The package is available from the GitHub repository: `https://github.com/anaserrasilva/islandNeighbours`.

2.4.1.1 *x*-NNI islands

Exhaustive search for x*-NNI island extraction:*
Step 1:
Given a tree distribution
For each tree in the distribution:
Generate 1-NNI neighbourhood of tree:
If $x > 1$:
For each tree in neighbourhood generate its 1-NNI neighbourhood:
Repeat $x - 2$ times
Filter the x-neighbourhood for those trees shared with the tree distribution
Step 2:
When all filtered neighbourhoods have been identified:
Compare and merge filtered neighbourhoods with shared trees:
Recurse until only x-NNI islands are left

This algorithm requires two functions to be implemented: one to generate the filtered neighbourhoods, and another to compare and merge them. The algorithm is implemented in the *islandNeighbours* R package for extraction of 1- and 2-NNI islands. The current exhaustive search implementation for 2-NNI islands is bounded by $(2(n-3))^x \times t$ non-unique trees, where n is the length of the leaf set, $2(n-3)$ is the size of a tree's 1-NNI neighbourhood, $x = 2$ is the x-NNI threshold, and t is the size of the tree distribution. Thus, this implementation is too slow and not suitable for anything beyond small datasets (e.g., the tree set in Fig. 2.1).

2.4.1.2 *x*-Distance islands

Island extraction from a matrix of any pairwise tree-to-tree distances (D):
Choose a threshold x-D
Create a vector with the tree distribution's length and all values set to property a
(Indices in the tree distribution and property vector have 1:1 correspondence)
Set the first instance of a to b
Find the trees within x-D of first tree and change them to property b in vector
For all but the first tree:
Find the trees within x-D of trees with property b:
Set corresponding vector indices to b
For all trees in the distribution:
Find the trees within x-D of trees with property b:
Set corresponding vector indices to b
Remove all trees with property b from tree distribution
Recurse until all x-D islands have been extracted

This approach is analogous to graph colouring (properties *a*, *b*). It is implemented in the *islandNeighbours* R package as a stand-alone function, and combined with the calculation of an RF distance matrix, and that of a quartet distance (QD, Estabrook et al. 1985) matrix.

2.4.2 Finding NNI islands in a Bayesian tree distribution

The presence of 'tree islands' in Bayesian analyses can affect chain convergence (Lakner et al. 2008, Höhna & Drummond 2011), but, to our knowledge, their effects on summarising the resulting tree distribution have not been explored, perhaps due to the computational expense of calculating branch rearrangement metrics on typically large samples of often large trees *a posteriori* (Allen & Steel 2001, Bordewich & Semple 2005, DasGupta et al. 2000). We developed a small R (R Core Team 2019) package that uses the *nni* function in *phangorn* (Schliep 2011) to iteratively generate the 1-NNI neighbourhood of each tree in a distribution and filter the neighbourhood to retain only those trees that are also present in the tree distribution. Filtered neighbourhoods are then recursively checked for shared trees, if these are present tree neighbourhoods are merged until only the 1-NNI islands remain (see subsection 2.4.1.1 above).

Re-analysing the Pardo et al. (2017) dataset in MrBayes v.3.2.6 (Ronquist et al. 2012), under the Mk+I+G model, with two independent runs of four chains for 10 million generations, sampled every 10000, and a relative burn-in of 25%, yielded a tree distribution containing 1502 unique trees. The majority-rule consensus tree is identical to that reported by Pardo et al. (2017). Applied to this distribution, our R script yielded 1489 1-NNI islands, 1480 comprise a single tree (mass = 1, density = 1), five of the islands contained two trees (mass = 2, density = 1), and four islands were made up of three trees (mass = 3, density = 1), see supplementary materials for the tree files. Given that the distribution is composed exclusively of very small islands we can conclude that, unlike in the parsimony analysis, the majority-rule consensus of the Bayesian tree distribution has not been substantially affected by any 1-NNI large island bias.

2.4.3 More on finding islands *a posteriori*

The discovery of tree islands, both as a general phenomenon and in specific instances, was associated with heuristic tree searches using the branch rearrangement operations (NNI, SPR, TBR) that are the bases of the tree-to-tree distances used in Maddison's (1991) original definition of tree islands. Although convenient and helpful for islands to be found as a byproduct of tree searches, the original definition of islands in terms of tree-to-tree distances that are particularly hard to compute *a posteriori* probably has limited subsequent application of the concept of tree islands to investigations of tree distributions more generally. Indeed, to our knowledge, our example above is the first. However, our NNI-island finder R script is very slow, and extended to find 2-NNI islands (see subsection 2.4.1.1) it is too slow to be used with anything other than small toy datasets.

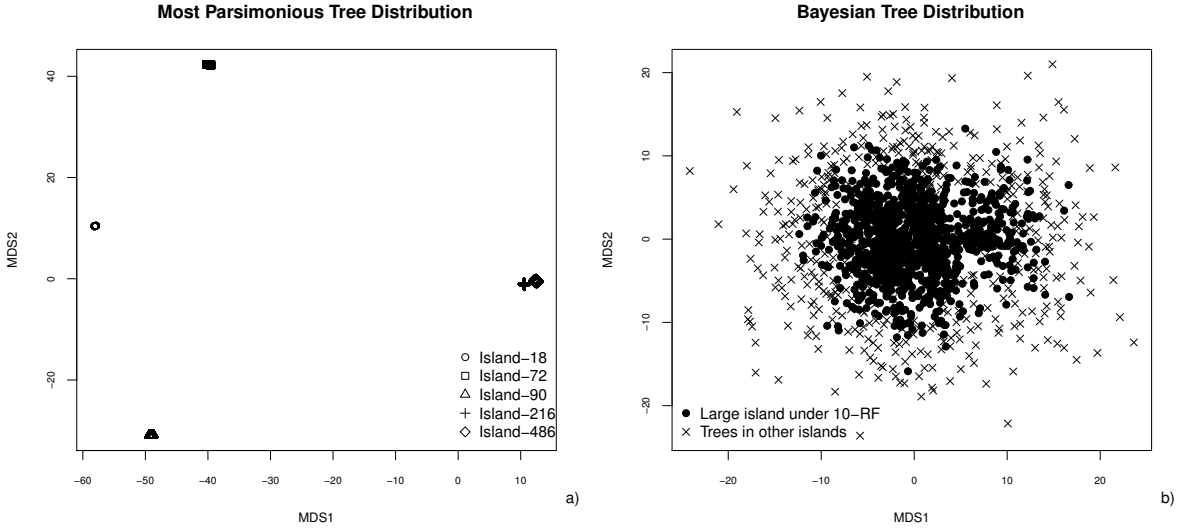


FIGURE 2.5. Multidimensional scaling plots of the a) set of MPTs and b) Bayesian posterior distribution of trees obtained from Pardo et al.'s (2017) amphibian dataset.

However, our revised definition of islands, usefully extends the notion to any tree-to-tree distance, including those whose calculation is not NP-hard, such as RF. Finding islands *a posteriori* based on such metrics is expected to be more efficient and tractable than finding islands based on branch rearrangement metrics. Indeed, given the pairwise distances of a set of trees, finding islands is not difficult (they are the disconnected components of the corresponding graph after removal of edges that are below the threshold value). We have implemented a simple exact algorithm that finds islands from a tree-to-tree distance matrix (see subsection 2.4.1.2).

As expected, 2-RF islands found using this algorithm in our example parsimony and Bayesian tree distributions are identical to the corresponding 1-NNI islands. Applying increasingly higher RF thresholds, we found that the island structure of the MPT distribution is robust, with the first change in island structure at $x = 12$ (when the two largest islands merged into one). In contrast, the Bayesian island structure is less stable, with a single large island forming at $x = 6$ that steadily increases in size with each increment in the RF threshold, while still identifying large numbers of single tree islands and without finding any alternative islands of substantial size (table 2.2). This pattern is as expected if we partition a homogeneous tree distribution, with increasing thresholds adding more trees to a single island in the centre of the distribution that excludes progressively fewer outliers. By $x = 10$ the largest island encompasses over 50% of the trees in the Bayesian tree distribution, and the partitioning of outliers from trees toward the centre of the distribution is apparent in a multidimensional scaling based on pairwise RF distances (Fig. 2.5, see next subsection). At this point the majority-rule consensus of the single large island is, as expected, identical to that of the full Bayesian tree distribution.

TABLE 2.2. Number and size of x-RF islands found in the Bayesian tree distribution at different RF thresholds.

Island size	2-RF	4-RF	6-RF	8-RF	10-RF	12-RF
1	1480	1373	1125	725	397	190
2	5	28	38	22	9	1
3	4	6	5	2	1	0
4	0	2	0	2	0	0
5	0	0	0	2	0	0
6	0	1	0	0	0	0
7	0	1	0	0	0	0
9	0	1	0	0	0	0
25	0	1	0	0	0	0
286	0	0	1	0	0	0
709	0	0	0	1	0	0
1084	0	0	0	0	1	0
1310	0	0	0	0	0	1
Total islands	1489	1413	1169	754	408	192

2.4.4 Other approaches to partitioning sets of trees

In this section we briefly consider other methods that have been developed for partitioning sets of trees and, where possible, compare their results with the partitioning by islands of our examples. Of the various clustering approaches that have been proposed few have been compared to tree islands and all are as yet rarely used in practice.

Similar to Hendy et al. (1988) and Höhna and Drummond (2011), we visualised our example empirical tree distributions through multidimensional scaling (MDS). Using the function *metaMDS* in the R package *vegan* v2.5-7 (Oksanen et al. 2020) with the RF distance matrix as input, MDS produces a clear separation of the five islands of MPTs (Fig. 2.5a). In contrast, MDS produces no partitioning of the Bayesian tree distribution but does reveal that trees in the large 10-RF island are concentrated at the centre of the sampled tree space.

Stockham et al. (2002) framed the question of how many consensus trees should be used to summarise a set of trees as a bicriterion problem of complexity (numbers of consensus trees) versus information loss (the distance between the tree distribution being summarised and the tree distribution induced by the strict consensus trees of the proposed clusters). While this remains an important open problem, it motivated and informed their comparison of 1-TBR islands (a byproduct of heuristic parsimony searches) and partitionings of several empirical sets of trees based on two families of clustering algorithms applied to RF distances, k-means

and agglomerative clustering (sometimes referred to as linkage clustering). Of the investigated methods they preferred complete linkage clustering. For our MPTs the complete linkage algorithm, computed using the R function *hclust*, yielded five well defined clusters, corresponding to the five islands from our example, and no discernible clusters in the Bayesian tree distribution. Results of the k-means analyses, however, change depending on the parameters used, making them harder to interpret. The function *find.clusters*, from the R package *adegenet* v.2.1.3 (Jombart 2008), using the Bayesian Information Criterion (BIC), 'goodfit', and 'choose.n.clust=FALSE' settings, identified k=8 as the optimal number of clusters for the MPT distribution and k=10 for the Bayesian distribution. However, if 'goodfit' is exchanged for the 'diffNgroup' setting, while the optimal number of clusters does not change for the MPT distribution, for the Bayesian distribution the optimal k becomes 2.

More recently, Tahiri et al. (2018) modified the k-medoids algorithm to work with RF distances, and compared its performance under squared and non-squared RF and the Silhouette and Caliński-Harabasz validity indices. Differently from Stockham et al. (2002), they sought to provide an optimal set of majority-rule (rather than strict) consensus trees and they made no comparison to islands. Tahiri et al. (2018) found that their k-medoids algorithm performs best with the non-squared RF distance and the Silhouette validity index, and that it cannot deal with single cluster datasets. We ran their k-medoids implementation under the recommended setting, which yielded two clusters for the MPT distribution, corresponding to the extended and restricted Lissamphibia groups of trees, so one cluster includes the three smallest islands, while the other is made up of the two largest islands. For the Bayesian distribution the application accepted the input file but did not generate any output files, suggesting that this corresponds to the k=1 instance, which Tahiri et al.'s (2018) algorithm cannot deal with. Another clustering algorithm is the basis of Guénoche's (2013) multiple consensus tree method, but we were unable to find an implementation that would allow any investigation of this method.

An alternative to distance-based clustering algorithms are graph-based methods. Bonnard et al. (2006) introduced the multipolar consensus which defines a minimal set of consensus trees (poles) that display all the splits that occur above some minimum frequency (α) in the input trees, and implemented a greedy heuristic graph-colouring algorithm based on split compatibility to approximate the multipolar consensus. At $\alpha = 0.5$ the method yields the majority-rule consensus. Applied to our example datasets, the multipolar consensus, with the default $\alpha = 0$ (i.e. including all splits), identified eight poles from the MPT distribution, four of which are relatively well resolved and are most similar to the partitioned-by-island consensus trees of each of the three smallest and the largest islands, the other four poles are mostly unresolved. For the Bayesian tree distribution, 13 poles were identified, but only one is well resolved. At $\alpha = 0.4$, the multipolar consensus identifies two poles in the MPT distribution, which are most similar to the partitioned consensus trees of each of the two largest islands.

Finally, an approach originally designed to partition sets of gene trees is based on the notion

of trees of trees (Darlu & Guénoche 2011, Nye 2008), where each tree topology becomes a leaf in a tree, and the nodes represent intermediate topologies between the input trees. We were unable to find an implementation of Darlu and Guénoche's (2011) TreeOfTrees and our tree distributions are too large for Nye's (2008) metaTree (both the web-based and stand-alone implementations). As a rough proxy (in which nodes do not represent intermediate tree topologies) we computed neighbour-joining (NJ) and unweighted pair group method with arithmetic mean (UPGMA) trees from the RF distance matrices (e.g., Graham et al. 1998). These methods also found the five islands of MPTs and no clear partitioning of the Bayesian tree distribution.

While unsurprising that the various methods considered sometimes yield somewhat different partitionings it is noteworthy that several produce partitionings that are either identical to or otherwise similar to islands.

2.5 Discussion

Historically, most users of parsimony were interested in discovering relationships that are present in every most parsimonious tree, leading to a strong preference for the exclusive use of the strict consensus to summarise most parsimonious trees, and a focus on the unambiguously or strictly supported relationships they display (Nixon & Carpenter 1996). Other relationships might be considered unsupported but there is a difference between a relationship that is present in some MPTs and a relationship that is not present in any MPT. The former seems somewhat better, if ambiguously, supported by the parsimonious interpretation of the data. It is tempting to go further and interpret the frequencies of occurrence of alternative ambiguously supported relationships in a set of MPTs as a measure of their relative support. This would follow from an assumption that, for example, all MPTs are equally probable. However, if there are multiple islands of MPTs, alternative measures of support would follow from an assumption that all islands are equally probable.

Each of the potential assumptions (equiprobable trees or equiprobable islands) might be justified by appeal to the principle of indifference (Keynes 1921). However, both assumptions can be met simultaneously only if all MPTs belong to a single island or all islands have exactly the same size. Otherwise, as our example shows, they can lead to very different and contradictory conclusions. In general, we expect such disagreements to be more likely and profound the greater the asymmetries in island size. Clearly, when confronted with multiple islands of disparate sizes, it is sensible to try to understand the causes of the differences.

In our parsimony example, island size is highly correlated with the number of areas of local instability, and is explained by their combinatorial consequences. This argues against island size being correlated with the probability of the island containing the correct tree, against equiprobability of trees, and against the use of the majority-rule consensus, because of the unjustifiable large island bias that ensues. The partitioned-by-island consensus reveals five major variants

involving alternative interrelationships of Gymnophiona, Batrachia, and *Chinlestegophis*, and helps in understanding the substantial large island bias due to different combinations of more local instability. In contrast to the majority of trees, the majority of islands recovered the traditional Lissamphibia (Figs. 2.3a-c). Both the majority of trees and the majority of islands have *Chinlestegophis* closely related to Gymnophiona (Figs. 2.2b, 2.3a, b, d). Interestingly, Schoch et al.'s (2020) recent analyses of a dataset slightly modified from that of Pardo et al. (2017) recovered the restricted Lissamphibia crown group, with *Chinlestegophis* only distantly related (similar to Fig. 2.3c), emphasising that robust inferences of amphibian interrelationships, i.e. inferences that are insensitive to minor variations in the underlying data and/or to method of analysis, may be hard to obtain. More detailed exploration of the phylogenetic relationships seemingly supported by the data is facilitated by the variant consensus approaches and may be especially intriguing for groups whose evolutionary history is still being debated, including amphibians (reviewed in Marjanović & Laurin 2019).

Sumrall et al. (2001) recognised the large island bias issue in bimodal tree distributions, warned against using the majority-rule consensus to summarise MPTs, and advocated the sole use of the strict consensus. However, that strict consensus trees can be very poorly resolved has seemingly motivated the use of less strict methods and, over time, Sumrall et al.'s (2001) findings and recommendations have been increasingly overlooked or forgotten. Revisiting the issue, we also urge caution against uncritical use of the majority-rule consensus. If the strict consensus is poorly resolved, then the partitioned-by-island consensus, where islands are summarised individually, can be particularly useful in distinguishing major alternatives and local instabilities. Reduced consensus methods (Wilkinson 1994) may also be helpful in this context. If island sizes are disparate, then simple modifications to the majority-rule consensus, through weighting or rarefaction can remove any large island bias from a unitary consensus summary, if such is needed. Our preference is for exploration and flexible use of multiple consensus methods. Discovery of islands should motivate interest in their biological or methodological significance, and discovery of disparate sizes raises the possibility of large-island bias, and should motivate further assessment of the cause of the size disparity and whether it should impinge on our assumptions of equiprobability of, for example, trees or islands. Note that while we have focused on the majority-rule consensus and the attendant issue of large island bias, researchers may choose to investigate or summarise islands with whichever approaches they prefer, including construction of any form of consensus tree or network (explored in the next chapter).

As we have shown, the notion of islands is extendable to methods that can produce multisets of trees or where the sampled trees are not optimal *per se*, but are due to resampling methods or come from regions of tree space with sufficiently high probability densities. Bayesian and resampling analyses can provide direct evidence that trees are not equiprobable, because each topology can be sampled multiple times, such that island size (number of unique topologies) is less than the sampling of trees from an island (island mass). Whereas island mass should be

driven by the posterior probabilities/frequency of sampling of the included trees given the data, island size differences may result in large island biases, which would be of concern. However, the Bayesian distribution from our example revealed no great disparities in 1-NNI/2-RF island sizes, with the number of islands very close to the number of trees. Thus both equiprobable tree and equiprobable island assumptions seem reasonable enough. Furthermore, that increasing the RF threshold only partitions the Bayesian tree distribution into a single large 'central' island, and many small islands of outliers is consistent with the tree distribution being homogeneous. Thus, island structure provides no basis for questioning the use of the majority-rule consensus to summarise the results of the Bayesian analysis.

Unfortunately, given that computing branch rearrangement metrics *a posteriori* is an NP-hard problem (Allen & Steel 2001, Bordewich & Semple 2005, DasGupta et al. 2000), finding islands based on these metrics in tree distributions from Bayesian and resampling analyses will be intractable in many cases. This is particularly unfortunate if topological differences that are readily described by such branch rearrangements are potentially linkable to or suggestive of specific evolutionary processes and/or analytical artefacts. For example, we suggest that whereas single NNI moves might represent stochastic error, SPR moves (that are not NNIs) might indicate instances of horizontal gene transfer (HGT), and TBR moves (that are not SPRs) might indicate local rooting problems such as might result from long branch attraction.

Our method of finding 1-NNI islands *a posteriori* directly (i.e.without computing NNI distances), is of limited use for large tree distributions, because it replaces distance metric calculations with very large numbers of pairwise comparisons of trees. However, our extended definition of tree islands allows for the use of any tree-to-tree distance metrics to define islands and makes it possible for islands to be identified *a posteriori* and for their causes and consequences to be explored. It reflects our point of view that islands are interesting more for the natural way in which they partition a set of trees than for any specific tree-to-tree distance that they were originally based upon. As such tree islands are complementary to other means of data exploration that involve attempts at partitioning sets of trees in order to provide better summaries and promote better understanding.

Other extensions to the notion of islands might be helpful. Allowing for trees with partially overlapping leaf sets might be achievable through generalised tree-to-tree distances (see Cotton and Wilkinson 2007) and allow clustering of gene trees without having to prune/regraft taxa, and might also help shed light on the phenomenon of tree terraces (Sanderson et al. 2011, 2015). Another possible extension might be to node-labelled trees, this would be particularly interesting given the recent drive to solve the single *vs*. multiple consensus problem in cancer phylogenetics (Aguse et al. 2019, Govek et al. 2018). Current methods are based exclusively on graph-based clustering, using a variety of distances for rooted trees (Aguse et al. 2019, Govek et al. 2018) that could conceivably be used to define and find islands.

2.6 Data and software availability

The data, scripts and results for this chapter are available on the following repository:

- Dryad, `https://doi.org/10.5068/D14X10`.

The *islandNeighbours* R package is available from:

- GitHub, `https://github.com/anaserrasilva/islandNeighbours`.

CHAPTER 3

MORE ON ISLANDS AND THE *Chinlestegophis jenkinsi* CONUNDRUM

3.1 Introduction

Large (multi)sets of phylogenetic trees with poorly resolved consensus trees are often diagnostic of the presence of unstable taxa. To identify which, and sometimes why, taxa are unstable multiple data- and topology-based methods for taxonomic instability identification have been developed (e.g., Aberer et al. 2012, Thorley & Wilkinson 1999, Wilkinson 1995*b*). In addition to these methods, some non-parametric resampling techniques, primarily used to test tree robustness to data perturbations (e.g., first order taxon-jackknifing Lanyon 1985), can also be used to identify rogue taxa. However, these resampling approaches often require multiple rounds of tree inference, thus limiting their widespread use. While the effects of isolated or grouped unstable taxa have been tested for some topology-based methods (Wilkinson & Crotti 2017), to my knowledge the same has not been done for the presence of multiple tree subsets, e.g. tree islands (Maddison 1991, Serra Silva & Wilkinson 2021). Given that the presence of multiple subsets of equally-optimal trees can substantially influence both tree search (e.g., Höhna & Drummond 2011, Lakner et al. 2008, Olmstead et al. 1993, Sanderson et al. 2011) and the summary of inferred trees (see Chapter 2 and citations therein), it is not unreasonable to expect that topology-based taxonomic instability identification analyses, such as leaf stability (LS, Thorley & Wilkinson 1999, Wilkinson 2006), are also affected by large island bias.

Another alternative to topology-based instability tests is the use of (non-tree) phylogenetic networks. These are being increasingly used to summarise (multi)sets of phylogenetic trees where incomplete lineage sorting (ILS), hybridization or other reticulation events are suspected (reviewed in Elworth et al. 2019). While phylogenetic networks have been primarily used to

explore incongruences in molecular datasets and/or trees inferred from molecular data, the latter are often called consensus networks (Holland & Moulton 2003), they have also started to be used with morphological data. Using a set of craniodental characters, Caparros and Prat (2021) have recently applied phylogenetic networks to the study of hominin evolution. However, despite the large size of the tree set they used for the network analyses (9639 trees), all trees belonged to the same 1-TBR island (PAUP*'s (Swofford 2003) default branch-rearrangement setting for maximum parsimony heuristic searches), and the network analyses would thus have been immune from large island bias. As such, it remains unclear whether the presence of multiple 1-TBR islands affects the visualisation of consensus networks.

In this chapter, using Pardo et al.'s (2017) set of most parsimonious trees (MPTs), I explore how two topology-based instability tests, LS and relative bipartition information content (RBIC, Aberer et al. 2012), behave in the presence of islands and how island structure can be taken into account when using these methods. I also use two (non-tree) phylogenetic network methods to show that large island bias extends beyond the majority-rule consensus (MRC, Margush & McMorris 1981). And, lastly, I reanalyse the Pardo et al. (2017) data matrix to identify some potential causes of instability, and explore the robustness of their conclusions on lissamphibian relationships.

3.2 Islands and taxonomic instability

Methods to explore taxonomic instability can be broadly divided into data (e.g., safe taxonomic reduction (STR, Wilkinson 1995*b*) and its heuristic extension Concatabominations (Siu-Ting et al. 2015), and *a priori* taxon jackknifing) and topology-based methods (e.g., LS, RBIC and *a posteriori* taxon jackknifing). While the presence of multiple subsets of trees is not expected to negatively influence data-based methods (even if we may expect some of, or all, the taxa driving island structure to be identified as unstable), topology-based methods may be prone to some of the problems highlighted for the MRC. In other words, if there is more topological variation between subsets of trees than within, could the set of identified unstable taxa change between analysing the full set of trees and analysing each subset separately? For ease of interpretation and discussion, I will focus on the exhaustive and mutually exclusive islands of trees (previous chapter, Maddison 1991) and will again use the Pardo et al. (2017) set of MPTs and its 1-TBR (also 1-SPR, 1-NNI and 10-RF) islands for the analyses below. If we do expect topological variation to be greater between islands, we may sensibly expect four outcomes: a taxon is i) stable within and between tree islands; ii) stable within islands, but unstable between; iii) unstable within islands, but stable between; or iv) unstable within and between islands, table 3.1. These outcomes correspond to interactions between causes of instability and relationships between taxa in a (sub)set of trees. Specifically, taxa that are stable only within islands correspond to the taxa responsible for island structure (in other words, globally unstable taxa) and their instability is

TABLE 3.1. Instability patterns possible for any taxon present in a partitionable (multi)set of trees.

Between islands \ Within islands	stable	unstable
stable	+ +	+ -
unstable	- +	- -

caused, primarily, by homoplasy, i.e. independently evolved similarity. Taxa that are unstable only within islands, on the other hand, correspond to locally unstable taxa and are likely caused by missing data (potentially ILS in molecular datasets), although less severe levels of homoplasy (than seen in the globally unstable taxa) cannot be discarded. Finally, taxa that are unstable both within and between islands likely have very large amounts of missing data and/or homoplasy, and the inference analyses may benefit from their removal. Understanding if, and how, topology-based methods are influenced by the presence of multiple islands is especially important when dealing with datasets, like Pardo et al.'s (2017), where the strict consensus (SC, Sokal & Rohlf 1981) is mostly unresolved, hinting at extensive taxon instability, but *a priori* data-based approaches, like STR (PerlEQ v1.0, Jeffery and Wilkinson (2003), available from `https://uol.de/en/biology/research-groups/systematics-and-evolutionary-biology/programs`), do not identify any taxa for safe removal, i.e. taxa whose removal does not affect the relationships between the remaining taxa in a matrix.

To explore whether large islands influence topology-based taxonomic instability analyses, I ran the full set of Pardo et al.'s (2017) MPTs, and each tree island, through LS analyses, as implemented in RogueNaRok (RNR, Aberer et al. 2012). All LS measures were used for the analyses: i) LS_{max}, the normalised maximum (bootstrap) proportion for a quartet/triplet, ii) LS_{diff}, the normalised difference between the proportions of the two most supported quartet/triplet resolutions, and iii) LS_{ent}, the entropic LS or the normalised negative sum of the product of the frequency of each quartet/triplet resolution and the log of that frequency (f),

$$LS_{ent} = 1 - \frac{-\sum(f \times log(f))}{log(f)}$$

(Thorley 2000, Thorley & Wilkinson 1999, Wilkinson 2006). The MPTs were also run through RNR's RBIC optimality criterion, set to optimise the SC, with dropsets of 1, 5, 10 and 15 taxa. The RBIC is defined by Aberer et al. (2012) as the ratio between the sum of all support values (S) and the theoretical maximum support for a binary tree prior to taxon removal,

$$RBIC = \frac{\sum S}{100 \times (n-2)},$$

TABLE 3.2. Leaf stability (LS) metrics for all most parsimonious trees and for each of the five tree islands recovered from Pardo et al.'s (2017) dataset. LS metrics are normalised, with 1 corresponding to maximal support and 0 to no support. Islands are identified by their sizes. Results for all taxa can be found in the Figshare Repository: `https://doi.org/10.6084/m9.figshare.c.6050033`.

All trees				**Island-18**				**Island-72**			
Taxa	LS_{max}	LS_{diff}	LS_{ent}	Taxa	LS_{max}	LS_{diff}	LS_{ent}	Taxa	LS_{max}	LS_{diff}	LS_{ent}
Eocaecilia	0.828	0.772	0.715	*Platyrhinops*	0.991	0.988	0.992	*Trematosaurus*	0.994	0.994	0.993
Epicrionops	0.828	0.772	0.715	*Amphibamus*	0.991	0.988	0.992	*Trematolestes*	0.998	0.998	0.997
Ichthyophis	0.828	0.772	0.715	*Rhineceps*	0.999	0.999	0.999	*Peltostega*	0.998	0.998	0.997
Karaurus	0.905	0.874	0.822	*Uranocentron*	0.999	0.999	0.999	*Lyrocephaliscus*	0.998	0.998	0.997
Kokartus	0.905	0.874	0.822	*Broomistega*	0.999	0.999	0.999	*Callistomordax*	0.998	0.998	0.997
Cryptobranchus	0.907	0.877	0.825	*Dissorophus*	0.999	0.999	0.999	*Metoposaurus*	0.998	0.998	0.997
Hynobius	0.907	0.877	0.825	*Broiliellus*	0.999	0.999	0.999	*Rhineceps*	0.999	0.999	0.999
Ambystoma	0.908	0.877	0.825	*Cacops*	0.999	0.999	0.999	*Uranocentrodon*	0.999	0.999	0.999
Triadobatrachus	0.908	0.878	0.825	*Acheloma*	0.999	0.999	0.999	*Broomistega*	0.999	0.999	0.999
Xenopus	0.908	0.878	0.825	*Phonerpeton*	0.999	0.999	0.999	*Cryptobranchus*	0.999	0.999	0.999
Leptodactylus	0.908	0.878	0.825	*Ecolsonia*	0.999	0.999	0.999	*Hynobius*	0.999	0.999	0.999
Sangaia	0.913	0.887	0.864	*Doleserpeton*	0.999	0.999	0.999	*Ambystoma*	0.999	0.999	0.999
Edingerella	0.917	0.893	0.862	*Gerobatrachus*	0.999	0.999	0.999	*Dissorophus*	0.999	0.999	0.999
Bentosuchus	0.919	0.894	0.868	*Chinlestegophis*	0.999	0.999	0.999	*Cacops*	0.999	0.999	0.999
Chinlestegophis	0.920	0.894	0.844	*Eocaecilia*	0.999	0.999	0.999	*Broiliellus*	0.999	0.999	0.999
Laidleria	0.930	0.907	0.888	*Epicriopnops*	0.999	0.999	0.999	*Acheloma*	0.999	0.999	0.999
Average	0.940	0.920	0.891	Average	0.999	0.999	0.999	Average	≈ 1	≈ 1	0.999

Continued on the next page

TABLE 3.2. *(cont.)* Leaf stability (LS) metrics for all most parsimonious trees and for each of the five tree islands recovered from Pardo et al.'s (2017) dataset. LS metrics are normalised, with 1 corresponding to maximal support and 0 to no support. Islands are identified by their sizes. Results for all taxa can be found in the Figshare Repository: `https://doi.org/10.6084/m9.figshare.c.6050033`.

Island-90				**Island-216**				**Island-486**			
Taxa	LS_{max}	LS_{diff}	LS_{ent}	Taxa	LS_{max}	LS_{diff}	LS_{ent}	Taxa	LS_{max}	LS_{diff}	LS_{ent}
Edingerella	0.951	0.945	0.946	*Trematosaurus*	0.994	0.994	0.993	*Trematosaurus*	0.994	0.993	0.993
Bentosuchus	0.965	0.964	0.951	*Platyrhinops*	0.995	0.993	0.994	*Platyrhinops*	0.995	0.993	0.994
Parotosuchus	0.982	0.980	0.978	*Amphibamus*	0.995	0.993	0.994	*Amphibamus*	0.995	0.993	0.994
Paracyclotosaurus	0.982	0.980	0.978	*Karaurus*	0.996	0.995	0.996	*Sangaia*	0.995	0.994	0.995
Mastodonsaurus	0.982	0.980	0.978	*Kokartus*	0.996	0.995	0.996	*Laidleria*	0.995	0.994	0.995
Cyclotosaurus	0.982	0.980	0.978	*Trematolestes*	0.997	0.997	0.997	*Siderops*	0.996	0.994	0.995
Platyrhinops	0.990	0.987	0.991	*Peltostega*	0.997	0.997	0.997	*Batrachosuchus*	0.996	0.994	0.995
Amphibamus	0.990	0.987	0.991	*Lyrocephaliscus*	0.997	0.997	0.997	*Karaurus*	0.996	0.995	0.996
Sangaia	0.995	0.994	0.993	*Callistomordax*	0.998	0.998	0.997	*Kokartus*	0.996	0.995	0.996
Peltostega	0.995	0.994	0.993	*Metoposaurus*	0.998	0.998	0.997	*Peltostega*	0.997	0.997	0.997
Lyrocephaliscus	0.995	0.994	0.993	*Cryptobranchus*	0.998	0.998	0.998	*Lyrocephaliscus*	0.997	0.997	0.997
Trematosaurus	0.995	0.994	0.993	*Ambystoma*	0.998	0.998	0.998	*Trematolestes*	0.998	0.998	0.998
Trematolestes	0.995	0.994	0.993	*Hynobius*	0.998	0.998	0.998	*Callistomordax*	0.998	0.998	0.998
Callistomordax	0.995	0.994	0.993	*Rhineceps*	0.999	0.999	0.999	*Metoposaurus*	0.998	0.998	0.998
Metoposaurus	0.995	0.994	0.993	*Broomistega*	0.999	0.999	0.999	*Cryptobranchus*	0.998	0.997	0.998
Laidleria	0.995	0.994	0.993	*Uranocentrodon*	0.999	0.999	0.999	*Hynobius*	0.998	0.997	0.998
Average	0.995	0.995	0.995	Average	0.999	0.999	0.999	Average	0.999	0.998	0.999

where $n-2$ corresponds to the number of internal branches in a fully resolved unrooted tree with n tips ($n-1$ for rooted trees), and the multiplication factor of 100 assumes the use of percentual support values (e.g., bootstrap proportions, Felsenstein 1985). The SC was chosen due to the MRC biases detailed in the previous chapter, and under the SC the RBIC is equivalent to Pattengale et al.'s (2011) relative information content optimisation criterion, which consists of optimising the number of bipartitions in the SC. Interestingly, this measure was also one of the parameters (the other parameter being number of taxa) used by Wilkinson (1995*c*) to distinguish primary reduced consensus (RC) trees from other RC trees. Following Wilkinson and Crotti (2017), several dropset sizes were used in the RBIC, since unstable taxa may be missed if they are part of an unstable group, i.e. where the taxa in a clade are stable in relation to each other but the group is unstable in relation to the rest of the tree.

Starting with all 882 MPTs, LS analyses identified crown lissamphibians as the most unstable taxa (table 3.2), with *Chinlestegophis* within the 15 most unstable taxa (taxon instability ranking varies slightly between LS measures). The RBIC results for the set of MPTs changed with dropset size, with RBIC-1, 5 and 10 recovering *Chinlestegophis* as the single most unstable taxon, and its removal increasing the number of bipartitions in the SC from 35 to 36. RBIC-15, however, recovered two sequential sets of unstable taxa, the first included all crown lissamphibians (both extant and extinct) and *Chinlestegophis*, removal of which drives the number of bipartitions in the SC to 48, and the second set made up of *Laidleria*, whose placement differs between the four smallest islands and the largest island (Fig.2.3a-d and Fig.2.2b, respectively), removal of this taxon adds another bipartition to the SC. Thus, removal of the taxa identified by RBIC-15 on the full set of trees increases the resolution of the SC from 35 to 49 bipartitions.

When the islands were individually tested for the presence of unstable taxa, however, all LS measures exceed 0.94 (table 3.2), and the RBIC analyses identify *Trematosaurus* as unstable in island-72 and -216 (Fig.2.3b and Fig.2.3d). Thus, just like in the MRC scenario, identification of unstable taxa in the presence of multiple tree subsets yields different results if we look at those sets individually, or if we look at the full (multi)set. In this example, for the within island analyses, the taxa with the lowest stability scores can be found in the areas of local instability identified in the previous chapter (Fig.2.2b, Fig.2.3a-d and Table 2.1). In short, with Pardo et al.'s (2017) trees, testing all MPTs identified the group(s) of taxa whose position changed between, but not within, islands of MPTs as unstable (crown Lissamphibia and *Chinlestegophis*), but the within islands analyses found only "locally" unstable taxa (e.g., *Trematosaurus*). Additionally, RBIC analyses whith dropsets > 1 identifying groups of unstable taxa may also be indicative of an underlying island structure in the (multi)set of trees being analysed.

To summarise, the taxonomic instability analyses on the MPTs recovered three of the four predicted outcomes: i) most taxa are stable within and between islands; ii) crown Lissamphibia and *Chinlestegophis* are stable within islands, but unstable between them; iii) *Trematosaurus* is unstable within some islands, but stable between them; iv) no taxa were identified as unstable

within and between islands. Following from the links between instability patterns and causes made earlier in the section, Lissamphibia and *Chinlestegophis*' instability is caused by homoplasy, particularly between *Chinlestegophis*, *Eocaecilia* and *Rileymillerus* (see subsection 3.4.1 below for details), while the instability of taxa like *Trematosaurus* is likely due to missing data, although low levels of homoplasy cannot be discarded. Thus, understanding how unstable taxa behave in the presence of multiple islands of trees is not essential to interpreting taxonomic instability analyses, but understanding the relationship between island structure and causes of instability might help mitigate the latter.

3.3 Phylogenetic networks

From Chapter 2 and the topology-based taxonomic instability analyses, we know that the topological incongruences in Pardo et al.'s (2017) MPTs are not restricted to a single unstable taxon, rather to a group of unstable taxa (crown Lissamphibia+*Chinlestegophis*). Thus far, I have focused on the effects of large island bias in the computation of consensus trees and on how to mitigate these effects within the single or multiple consensus tree debate. However, there is an alternative to consensus trees and clustering analyses that can display at least some of the topological incongruences present in a (multi)set of trees in the same graphical summary: phylogenetic networks. While there are multiple types of phylogenetic networks (summarised in Fig. 4.1 of Huson et al. 2010), including phylogenetic trees, I will focus only on those used in Caparros & Prat (2021): consensus (Holland & Moulton 2003) and reticulation networks (Huson et al. 2005). Consensus and reticulation networks differ in that, for a given multi(set) of trees, the former display all splits over a selected frequency threshold (setting this to 50% is equivalent to computing the MRC), and thus the split incompatibilities present in the tree (multi)set (Holland & Moulton 2003). Reticulation networks, on the other hand, use those split incompatibilities to identify and display the presence, but not type, of reticulation events (Huson et al. 2005). In short, consensus networks summarise split incompatibilities, while reticulation networks display evolutionary histories. These approaches, particularly reticulation networks, have generally been applied to molecular datasets (including, but not restricted to, sequence data, inferred trees, matrix representation of splits, etc.), partly because they were developed with the aim of understanding the topological conflicts in (multi)sets of gene trees (e.g., Huson et al. 2005), and because reticulation events are often investigated at the molecular not morphological level (e.g., Cai et al. 2021, Suvorov et al. 2022).

Using SplitsTree v.4.14.8 (Huson & Bryant 2006) I computed the consensus and reticulation networks for the Pardo et al. (2017) MPTs. The consensus networks were computed for split frequency thresholds of 33% (default), 20%, 10% and 0%, all displayed with 'EqualAngle'. The reticulation networks were computed from the consensus networks, using the RECOMB2007 algorithm (Huson & Kloepper 2007), rooted on *Proterogyrinus*, and displayed with 'EqualAngle120'.

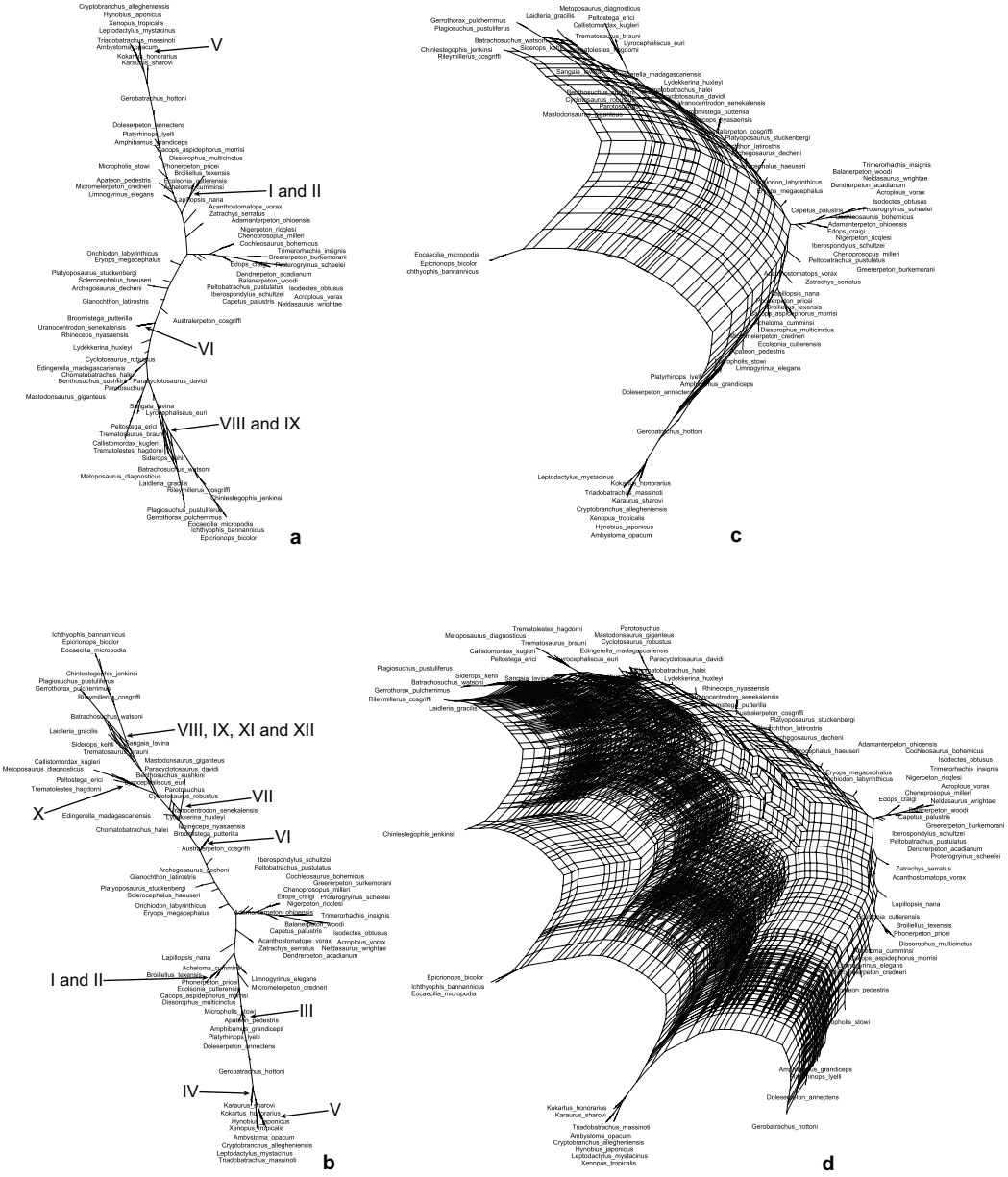


FIGURE 3.1. Consensus networks of Pardo et al.'s (2017) set of most parsimonious trees for a) 33%, b) 20%, c) 10% and d) 0% split frequency thresholds. Roman numerals correspond to the areas of instability identified in figures 2.2b and 2.3a-d. High resolution image files for each network are available from `https://doi.org/10.6084/m9.figshare.c.6050033`.

For ease of visualisation, Nexus formatted network files and EPS image files are available from `https://doi.org/10.6084/m9.figshare.c.6050033`. The consensus network for the default split frequency threshold shows most of the split incompatibilities centred on the areas of local instability VIII and IX (Fig. 3.1a and Fig.2.2b), along with the possible solutions for the polytomies corresponding to areas I, II, V and VI (Fig. 3.1a, Fig.2.2b and Fig.2.3). The 20% threshold displays

split incompatibilities from areas VII to XII, which share taxa with areas VIII and IX, and the solutions to polytomies I to VI (Fig. 3.1b, Fig.2.2b and Fig.2.3). As in the two largest islands, *Chinslestegophis* is placed with *Rileymillerus* and Gymnophiona within Stereospondyli in the 33% and 20% threshold consensus networks. At thresholds 10% and 0% the consensus networks are (near) uninterpretable (Fig.3.1c-d), due to the very large number of split incompatibilities present in the input tree set. Although, in the 10% threshold network *Chinlestegophis* is found with *Rileymillerus*, while Gymnophiona and Batrachia are each in well-defined clades, and in the 0% threshold network the relationship between *Chinlestegophis* and *Rileymillerus* breaks down, but Gymnophiona and Batrachia are still recovered as clades. These results are consistent with what we know of the island structure of Pardo et al.'s (2017) MPTs. First, areas of instability VII–XII are each present in at least one of the four largest islands, share taxa (e.g., *Sangaia* contributes to areas VIII and XII) and/or are in close proximity in at least two islands. Additionally, many of the taxa in these local instability areas were also recovered by most island-specific LS analyses as some of the least stable taxa (e.g., *Trematosaurus*, *Trematolestes*, *Bentosuchus*). That the 33% threshold consensus network displays split incompatibilities that match up with areas of local instability (VIII and IX) restricted to the largest island (~ 55% of all trees), and the 20% threshold network starts recovering split incompatibilities present in the next largest islands (~ 25% and ~ 10% of all trees), suggests that, much like the MRC, consensus networks are also affected by large island bias. This is further supported by the amount of split incompatibilities found between *Chinlestegophis*, *Rileymillerus*, Gymnophiona and Batrachia.

The reticulation networks yielded very similar results to the consensus networks, as expected from the latter being the input for the identification of reticulation events. For the default threshold, two reticulation events were identified between taxa contributing to areas VIII and IX, and single events were identified for the isolated polytomies corresponding to areas I, II, V and VI (Fig.3.2a). The 20% threshold reticulation network finds single reticulation events for the localised polytomies (areas I to VI) and a minimum of eight reticulation events in areas VII–XII, including the two found under the 33% threshold (Fig.3.2b). Unfortunately, the large number of taxa and high level of split incompatibilities in this dataset makes the networks hard to interpret and the reticulation nodes in area VII/X hard to parse. It is, however, still possible to recognise the placement of *Chinlestegophis* with *Rileymillerus* and Gymnophiona in Stereospondyli, i.e. the extended Lissamphibia hypothesis. The 10% network displays a very complex history of reticulations with all Dissorophoidea and Stereospondyli taxa (see Fig.3.4 for higher order amphibian taxonomy) being involved in, or descended from, at least one reticulation event (Fig.3.2c). This analysis still finds *Chinlestegophis* nested in Stereospondyli, but neither the extended nor restricted Lissamphibia are found. This reflects the drastic topological shifts between islands-18 and -90 (restricted Lissamphibia in Dissorophoidea), island-72 (restricted Lissamphibia in Stereospondyli), and islands-216 and -486 (extended Lissamphibia). The computation of the reticulation network for the 0% threshold consensus network was cancelled after running for 36

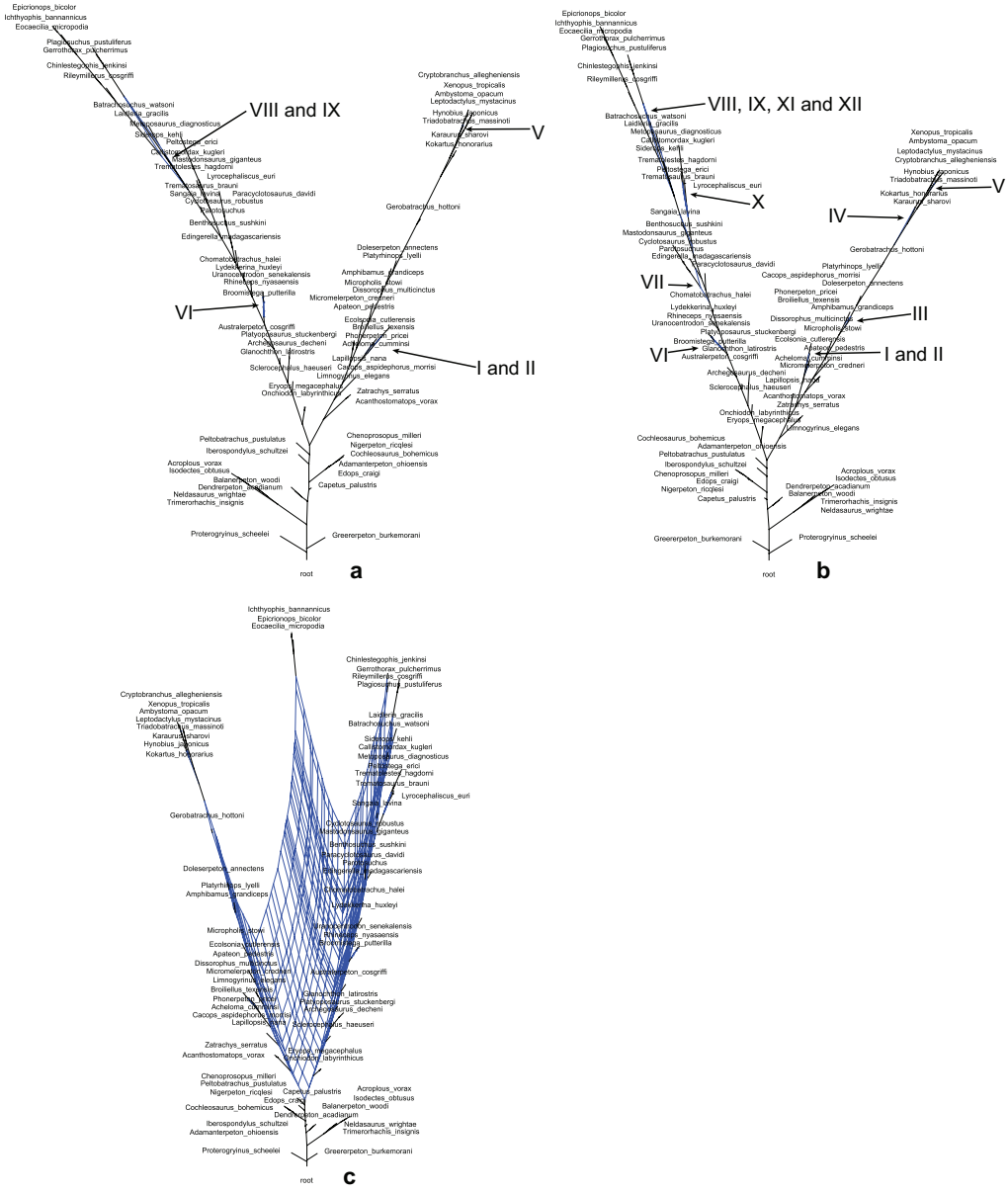


FIGURE 3.2. Reticulation networks of Pardo et al.'s (2017) set of most parsimonious trees for a) 33%, b) 20% and c) 10% split frequency thresholds. Reticulate nodes are in dark blue, as per SplitsTree default drawing settings. High resolution image files for each network are available from `https://doi.org/10.6084/m9.figshare.c.6050033`.

hours, as the next slowest analysis (10% threshold) took less than five minutes. However, given the consensus network for this threshold (Fig.3.1d), we might expect a reticulation network at least as complex as the one inferred for the 10% threshold. Thus, consensus and reticulation networks, much like the MRC, are prone to large island bias, with reticulation networks not always computable in the presence of high levels of split incompatibility.

3.4 Reanalysis of Pardo et al.'s (2017) data matrix

Congruence between the tree summaries of multiple phylogenetic analyses of the same dataset using distinct inference methods is often used as a sign of data robustness and used to bolster confidence on the inferred tree topologies (e.g., Pardo et al. 2017, San Mauro et al. 2014, Taboada et al. 2020). However, when using the MRC to summarise inferred tree (multi)sets, this congruence can be positively misleading, particularly when large island bias is present. As briefly mentioned in Chapter 2, the MRCs of the parsimony and Bayesian inference analyses of Pardo et al.'s (2017) data matrix are highly congruent (see their Fig.S7). Yet, the island structures of the tree sets yielded by the two analyses are highly disparate, with the parsimony analyses recovering a set that can be partitioned into five well-defined islands that support distinct Lissamphibia make-ups and placements, whereas the island structure of Bayesian tree distribution changes every time the distance threshold changes (see fig.2.5 and table 2.2), and consists of a large central island surrounded by large numbers of singleton or small islands that uniformly support the extended Lissamphibia hypothesis.

Moreover, topology-based instability analyses further emphasise the difference between the tree sets recovered by the parsimony and Bayesian analyses. While the RBIC analyses of the MPTs recovered different sets of unstable taxa for different dropset sizes (see section 3.2), the same analyses for the Bayesian tree distribution recovered *Apateon*, *Iberospondylus*, *Lapillopsis* and *Sangaia* as the most unstable taxa for all dropsets. The only variations between the results of the latter analyses are whether *Iberospondylus* and *Lapillopsis* should be removed separately or as a unit, and the order in which the four taxa should be removed. The results obtained by the inference and taxonomic instability analyses show that the Pardo et al. (2017) data matrix is not robust to changes in inference method, which is further supported by Schoch et al.'s (2020) analyses of a slightly modified version of the Pardo et al. (2017) data, see subsection 3.4.3. Thus, to understand the *Chinlestegophis* conundrum, we must first understand the data matrix.

3.4.1 Resampling analyses

A fundamental aspect of phylogenetic analyses that I have only briefly mentioned in regards to the Pardo et al. (2017) data matrix is branch support. In the previous chapter, I mentioned that, under parsimony, none of the splits that separate Gymnophiona and Batrachia had a bootstrap support (Felsenstein 1985) greater than 50%. In fact, a bootstrap analysis in PAUP* v.4a165, with 1000 replicates under the same settings as the heuristic equal-weights parsimony search and with the 'MulTrees' option selected, weakly supported (53%) the restricted Lissamphia and was uniformative in regards to *Chinlestegophis*'s placement. This result might lead us to think that despite the islands that display the restricted Lissamphibia making up only c.20% of the inferred MPTs, they may in fact be better supported by the underlying data. However, if jackknife analyses are used to calculate branch support, neither the restricted nor the extended Lissamphibia

hypotheses are supported. With similar settings to the bootstrap analysis (number of replicates set to 100 due to memory constraints), delete-half (Lanyon 1985) and Farris jackknife (delete $\frac{1}{e} \approx 36.8\%$, Farris et al. 1996) both weakly support *Gerobatrachus*+Batrachia (54% and 57%) and *Rileymillerus*+*Chinlestegophis*+Gymnophiona (55% and 54%), but are uninformative regarding the relationship between these two clades, and thus restricted *vs.* extended Lissamphibia. With only 11 out of 345 characters not having taxa with missing ('?') and/or inapplicable ('-') character states (see next subsection), it is not surprising that the trees summarising the resampling analyses are poorly resolved (e.g., dos Santos & Falaschi 2007, Wilkinson 2003). Unfortunately, memory constraints prevented saving the trees inferred for each resampling replicate, and I was unable to check whether the large island bias present in the heuristic parsimony search also affected the bootstrap and/or jackknife analyses, commonly summarised with their MRC. That different support measures find generally low branch support across the inferred relationships and that they yield conflicting results regarding the relationship between Batrachia and Gymnophiona, not only confirm that the MRC is a poor summary of the MPTs but also that the Pardo et al. (2017) data matrix is not robust to perturbations.

In addition to serving as measures of branch support, resampling techniques can also be used to identify unstable taxa. While the character-based jackknifing analyses showed only that there is no overwhelming support for either the restricted or extended Lissamphibia hypotheses, taxon jackknifing might help us identify the taxa whose data underpin the taxonomic instability that characterises this dataset. Because taxon jackknife requires a tree inference analysis each time a taxon is removed from the matrix, these resampling analyses can be quite time and resource intensive. A way of optimising first order taxon jackknifing is to use the output from (other) taxonomic instability analyses as a guide. Unfortunately, as mentioned above, the STR analysis (and its heuristic extension, Concatabominations (Siu-Ting et al. 2019)) did not find any candidate taxa for safe removal. However, the topology-based dropset-1 RBIC analysis did identify *Chinlestegophis* as the most unstable taxon, offering a starting point for the first order taxon jackknifing analyses. Additionally, from the partitioned-by-island consensuses computed in Chapter 2 (Fig.2.2b and Fig.2.3), we know that *Chinlestegophis* always places with *Eocaecilia* and/or *Rileymillerus*. We thus have a trio of taxa for which we can explore whether removing any one of them from the matrix affects the number of trees and islands inferred from the resampled matrices, and whether, removed taxa aside, the topologies inferred are similar between taxon jackknifing runs.

It is worth noting that we can also jackknife taxa from the inferred trees *a posteriori*. However, the removal of *Chinlestegophis* results only in removing it from the set of MPTs, without affecting the number of unique MPTs nor of tree islands. In fact, *a posteriori* removal of *Chinlestegophis*, *Eocaecilia* and *Rileymillerus*—whether we remove one, two, or all three of these taxa—affects only the number of tips on the MPTs, not their numbers nor island structure. Using LS_{max} as the guide for taxon removal order, we need to remove *Eocaecilia*, *Epicrionops*, *Ichthyopis* and *Karaurus*

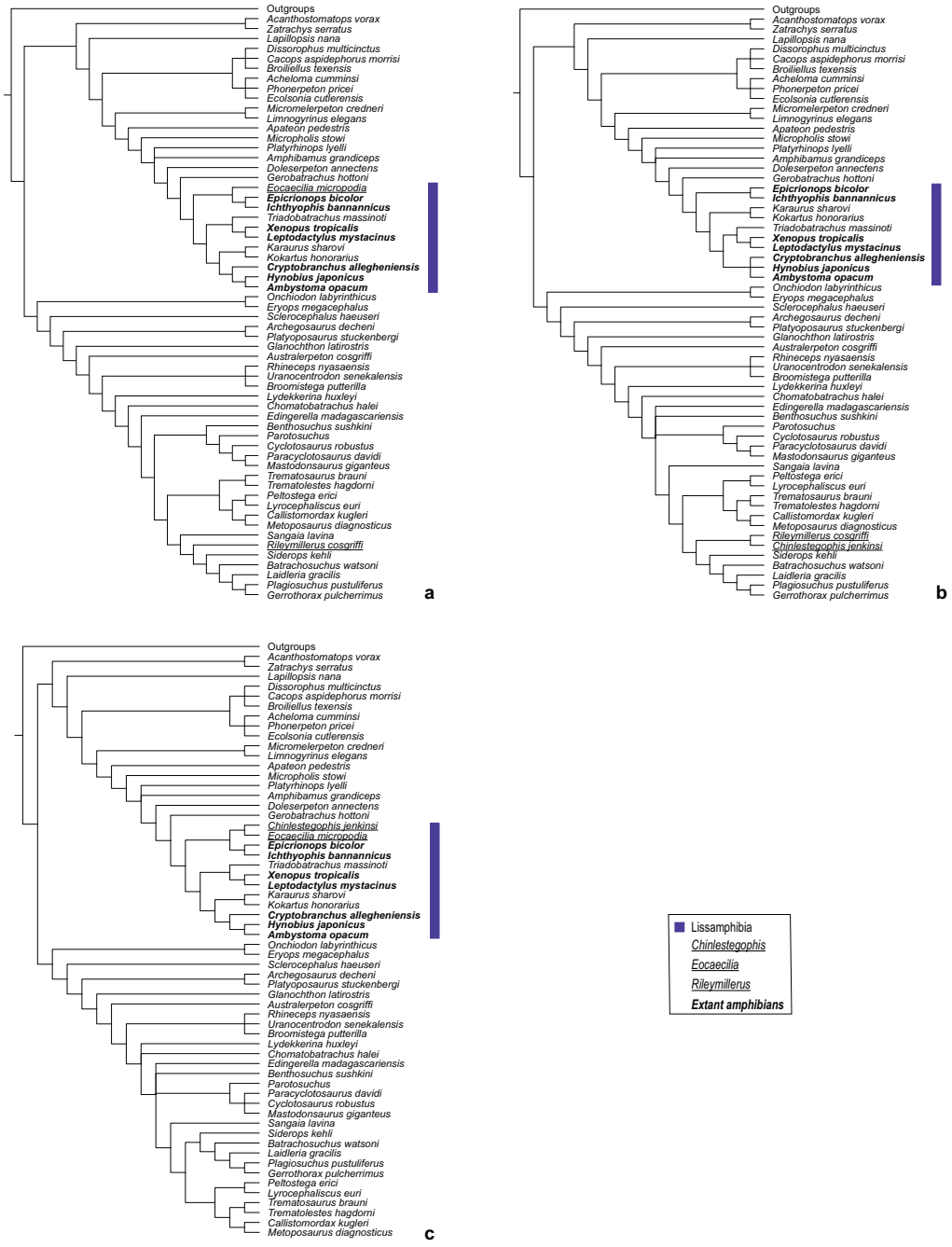


FIGURE 3.3. Strict consensus trees of the first-order taxon jackknife analyses omitting a) *Chinlestegophis*, b) *Eocaecilia* and c) *Rileymillerus*.

before the number of trees drops to 648 and that of 2-Robinson-Foulds (RF, Robinson & Foulds 1981) islands to four. If any external information is ignored and all taxa are tested, the taxa whose removal affects tree number the most are *Broomistega*, *Rhineceps* and *Uranocentrodon*, which correspond to polytomy VI (one of the three areas of instability present in all islands, table

TABLE 3.3. Post-inference first order taxon jackknife. List of taxa whose removal from the set of inferred MPTs resulted in decrease in MPT number, the change in MPT set size (ΔTrees) and the size of the ensuing set of MPTs. Island number was not affected by taxon removal.

Taxon	**ΔTrees**	**MPT set**
Broomistega, *Rhineceps*, *Uranocentrodon*	588	294
Trematosaurus	540	342
Ambystoma, *Cryptobranchus*, *Hynobius*	387	495
Acheloma, *Broiliellus*, *Cacops*, *Dissorophus*, *Ecolsonia*, *Phonerpeton*	294	588
Amphibamus, *Platyrhinops*	288	594
Karaurus, *Kokartus*	234	648
Batrachosuchus, *Laidleria*, *Sangaia*, *Siderops*	162	720
Trematolestes	72	810
Benthosuchus, *Edingerella*	36	846

2.1 and Figs.2.2b and 2.3). The removal of any one of these taxa decreases the size of the MPT set to 294 trees, removing exactly two thirds of the original trees, by virtue of breaking the polytomy. Thus, with post-inference taxon jackknife we can identify locally unstable taxa but not the taxa driving island structure.

Returning to data-based jackknifing, under the same settings as the original parsimony tree search, all three first order jackknife analyses recovered a single 1-TBR island, with restricted Lissamphia nested within Dissorophoidea (Fig.3.3a-c and table 3.4). The run without *Chinlestegophis* recovers 18 MPTs with three unstable subtrees: (*Dissorophus*, *Broiliellus*, *Cacops*), (*Acheloma*, *Phonerpeton*, *Ecolsonia*) and (*Rhineceps*, *Uranocentrodon*, *Broomistega*). These subtrees correspond to the three areas of instability present in all 1-TBR islands identified from the MPTs inferred from the the original Pardo et al. (2017), areas I, II and VI (Fig.2.2b and table 2.1). However, for the two runs where *Chinlestegophis* is retained the number of MPTs recovered increases comparatively to the no-*Chinlestegophis* analysis, and *Chinlestegophis* is placed as the sister taxon to the remaining taxon (*Rileymillerus* or *Eocaecilia*), table 3.4 and figures 3.3b-c. The increase in number of MPTs is also reflected by an increase in taxonomic instability. In addition to the three subtrees above, the relationship between and placement of *Edingerella* and *Benthosuchus* also become uncertain, and, with the removal of *Rileymillerus*, so does the subtree (*Paracyclotosaurus*, *Mastodonsaurus*, *Cyclotosaurus*). Interestingly, these taxa all correspond to those identified as least stable in the partitioned-by-islands LS analyses (table 3.2). Thus, from the taxon-based resampling analyses, it is clear that the uncertainty surrounding the lissamphibian relationships is anchored by the *Chinlestegophis-Eocaecilia-Rileymillerus* triad. However, how characters have been described and scored in the Pardo et al. (2017) matrix may

also influence whether the extended or restricted Lissamphibia hypotheses are recovered by the phylogenetic inference analyses.

3.4.2 Treatment of polymorphic taxa, missing and inapplicable data

Unlike aligned molecular datamatrices, where characters consist of positions in a sequence and the states are dictated by the nucleotide or amino acid residue identified at each position, in morphological datasets characters and character states are delineated and identified by systematists (Wilkinson 1995*a*). As such, morphological datamatrices carry with them a certain degree of subjectivity, which can lead to different workers scoring the same character(s) differently for the same taxa (e.g., Pardo et al. 2017, Schoch et al. 2020). While these rescorings alone can lead to inferring trees displaying different evolutionary hypotheses, the Pardo et al. (2017) data matrix also includes two aspects of character scoring, polymorphic taxa and inapplicable data, whose presence and choice of how to analyse them can lead to the recovery of different (multi)sets of inferred trees (e.g., Brazeau et al. 2019, Nixon & Davis 1991, Platnick et al. 1991).

Starting with polymorphic taxa, which can at best lead to changes in tree length (Nixon & Davis 1991), the Pardo et al. (2017) matrix contains 10 characters for which at least one taxon was scored as polymorphic (characters 10, 17, 106, 160, 246, 248, 253, 267, 321 and 324). In PAUP*, the user can choose to treat these character codings as uncertainty (the taxon might have any of the selected character states) or polymorphism (individuals in the taxon display one character state or the other(s)), which may lead to distinct character optimisations (see pages 103–106 in the PAUP* manual for details, `https://phylosolutions.com/paup-documentation/paupmanual.pdf`). Due to how the optimisation for polymorphic taxa is implemented in PAUP*, choosing the 'msTaxa=polymorph' setting affects only tree length, not topology. This was the pattern found for the Pardo et al. (2017) data matrix, with MPTs resulting from analyses where polymorphic taxa were treated as such being 18 steps longer than analyses set to uncertainty (table 3.4). MrBayes (Ronquist et al. 2012), on the other hand, always treats polymorphism as uncertainty (pages 11–12 and 85–87 in the manual, `http://mrbayes.sourceforge.net/mb3.2_manual.pdf`), and thus no comparison to analysing polymorphism as variation is possible. Thus, while some software packages offer multiple settings for the analysis of polymorphic taxa, these settings do not affect topology and can be ignored when comparing analytical settings, the only exception being if polymorphic taxa are coded as missing data (Nixon & Davis 1991).

A side note about analysing morphological/standard datasets in MrBayes is that, because the programme implicitly partitions datasets by each characters' highest possible score—i.e. if a character's highest score in the matrix is '2', the character is assumed to have three possible states: 0, 1 or 2—it occasionally makes assumptions about character states (page 148 in the manual). Also, when confronted with characters where the only numerical score present is zero, MrBayes assumes that those characters are binary (0,1). This was the case for characters 61 and 244 in Pardo et al.'s (2017) matrix, because all taxa were coded as zero, missing or inapplicable

TABLE 3.4. Summary of equal-weights maximum parsimony settings and results of Pardo et al.'s (2017) and Schoch et al.'s (2020) re-analyses

Matrix	Polymorphism	Inapplicable data	Trees	Tree length	Islands	Lissamphibia
Pardo original	Uncertainty	Missing	882	1514	5	Extended and restricted†
Pardo, no *Chinlestegophis*	Uncertainty	Missing	18	1479	1	Restricted
Pardo, no *Eocaecilia*	Uncertainty	Missing	180	1471	1	Restricted
Pardo, no *Rileymillerus*	Uncertainty	Missing	162	1494	1	Restricted
Pardo original	Polymorphism	Missing	882	1532	5	Extended and restricted†
Pardo original	Uncertainty	Additional state	351	1699	2	Restricted‡
Pardo original	Polymorhism	Additional state	351	1717	2	Restricted‡
Pardo rescored*	Uncertainty	Missing	108	1506	2	Restricted
Pardo rescored*	Polymorphism	Missing	108	1524	2	Restricted
Pardo rescored*	Uncertainty	Additional state	2376	1735	1	Restricted†
Pardo rescored*	Polymorphism	Additional state	2376	1753	1	Restricted†
Schoch original	NA	NA	2	1392	1	Restricted
Schoch, 345 characters	NA	NA	3	1366	1	Extended
Schoch rescored*	NA	NA	2	1377	1	Restricted
Schoch rescored*, 345 characters	NA	NA	3	1352	1	Restricted

*†Lissamphibia within Stereospondyli in one island. ‡Lissamphibia within Stereospondyli in two islands. *See table 3.4 for character recoding.*

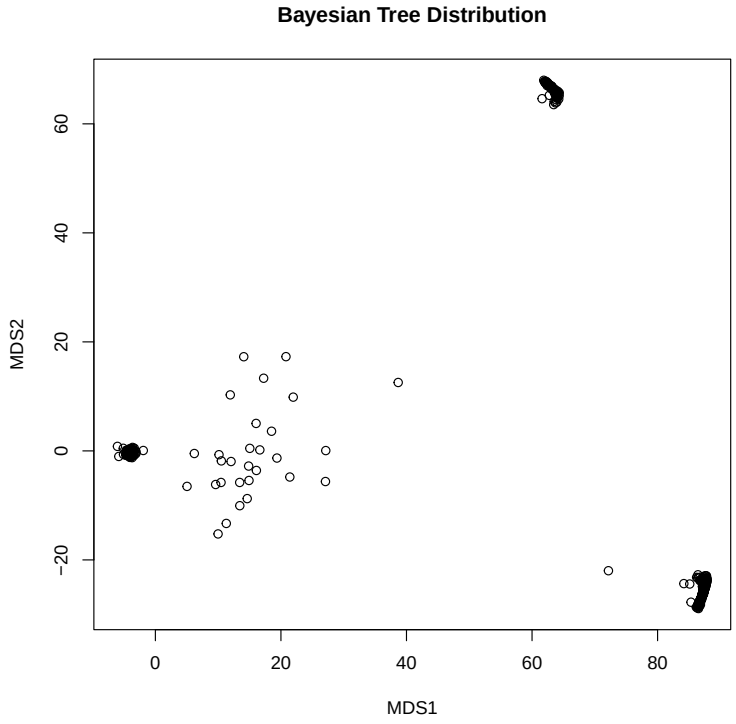


FIGURE 3.4. Multidimensional scaling plot of the Bayesian posterior distribution of trees obtained from an explicitly partitioned Pardo et al.'s (2017) matrix.

(0, '?' or '-'). These assumptions can lead to MrBayes's implicit partitioning scheme differing from an explicit partitioning scheme based on the character descriptions. In fact, for the Pardo et al. (2017) matrix the implicit partitioning analysis yielded a set of 1502 trees, while an analysis with the same settings (see subsection 2.4.2) but with an explicit partitioning scheme, based on the character descriptions, yielded 14871 trees, with an RF=3 between the two analyses' MRCs. However, unlike the analysis with MrBayes's implicit partition scheme, the set of trees inferred with the explicitly partitioned matrix displays a clear underlying island structure (Fig.3.4). Unfortunately, x-RF island extraction analyses had to be aborted due to memory restrictions resulting from the size of the treefile. The less memory intensive clump analyses (see Chapter 4), under a break-only approach, identified six clumps, with two recovering a restricted Lissamphibia topology and four the extended Lissamphibia hypothesis. Given these results, it may be worthwhile to run morphological datasets through MrBayes under implicit and explicit partitioning schemes.

Returning to character coding, most characters in the Pardo et al. (2017) matrix have at least one taxon scored as inapplicable and/or missing (186/345 characters with missing data and 146/345 with missing and inapplicable data), with $\approx 2\%$ and $\approx 25\%$ of the matrix consisting, respectively, of inapplicable and missing data. While these character scores are often treated as

TABLE 3.5. List of characters rescored due to logical inconsistencies. For characters 1–345, any character and/or taxon followed by a *P* was only rescored in the Pardo et al. (2017) matrix, and any followed by an *S* was only rescored in the Schoch et. al (2020) matrix. Characters 346–360 are only present in the Schoch et al. (2020) matrix.

Independent character	Dependent character(s)	Recoded character(s)	Recoded taxa	Recoded state	Reason
10 Alary process	13, 350	13	*Acroplous* (P)	?	Alary process scored as absent (0) in 10, but in 13 scored for premaxillary foramen separating alary processes (1)
		350	*Batropetes*, *Brachydectes*, *Cochleosaurus*, *Edops*, *Greererpeton*, *Proterogyrinus*, *Rhynchonkos*	?	Alary process scored as absent (0) in 10, but as alary process blade-like (0) in 350.
21 Lacrimal bone	1, 19, 20, 22, 23, 24, 25, 35, 36, 38, 43, 44, 45, 219, 220, 221	22	*Celtedens* (S), *Leptodactylus*, *Xenopus*	?	Lacrimal scored as absent (1), but lacrimal suture scored for position.
		24, 35	*Celtedens* (S), *Batrachosuchus*, *Laidleria* (P), *Siderops*	-	Lacrimal scored as absent (1), but lacrimal posterior extension (24) scored for, and infraorbital sulcus (35) defined based on the lacrimal.
		23, 44, 45, 219, 220, 221	*Celtedens* (S)	?	Lacrimal scored as absent (1), but all other characters are defined based on it.
		25, 38	*Ambystoma*, *Celtedens* (S), *Cryptobranchus*, *Leptodactylus*, *Xenopus*	?	Lacrimal scored as absent (1), but the rescored characters are defined based on it.

Continued on the next page

TABLE 3.5. *(cont.)* List of characters rescored due to logical inconsistencies. For characters 1–345, any character and/or taxon followed by a *P* was only rescored in the Pardo et al. (2017) matrix, and any followed by an *S* was only rescored in the Schoch et. al (2020) matrix. Characters 346–360 are only present in the Schoch et al. (2020) matrix.

Independent character	Dependent character(s)	Recoded character(s)	Recoded taxa	Recoded state	Reason
21 (*cont.*) Lacrimal bone		43	*Batrachosuchus*, *Celtedens* (S), *Ichthyophis*, *Laidleria* (P), *Siderops*	?	Lacrimal scored as absent (1), but character 43 is defined based on it.
		59	*Celtedens* (S), *Chinlestegophis*, *Rileymillerus*	?	Lacrimal scored as absent (1), but character 59 corresponds to jugal-lacrimal contact.
106 Basicranium (contact)	107	107	*Acroplous* (P), *Amphibamus*, *Apateon*, *Archegosaurus*, *Australerpeton* (P), *Balanerpeton*, *Batropetes* (S), *Brachydectes* (S), *Capetus* (P), *Dendrerpeton*, *Doleserpeton*, *Gerobatrachus*, *Glanochthon*, *Greererpeton*, *Iberospondylus* (P), *Isodectes* (P), *Limnogyrinus* (P), *Micromelerpeton*, *Micropholis*, *Neldasaurus* (P), *Platyoposaurus* (P), *Platyrhinops*, *Proterogyrinus*, *Rhynchonkos* (S), *Sclerocephalus*, *Trimerorhachis*	-	Based on the scoring for other taxa it seems that 106 is being treated as: Basicranium suture - present/absent. 107 rescored in accordance to this, but character construction/definition should be revised
209 Osteoderms	210, 211, 345	210, 211	*Acanthostomatops*, *Amphibamus*, *Apateon*, *Archegosaurus*, *Australerpeton* (P), *Balanerpeton*, *Batrachosuchus*, *Batropetes* (S), (*cont.*)	-	Osteoderms scored as absent (0), but scored for articulation and width in 210 an 211.

Continued on the next page

TABLE 3.5. *(cont.)* List of characters rescored due to logical inconsistencies. For characters 1–345, any character and/or taxon followed by a *P* was only rescored in the Pardo et al. (2017) matrix, and any followed by an *S* was only rescored in the Schoch et. al (2020) matrix. Characters 346–360 are only present in the Schoch et al. (2020) matrix.

Independent character	**Dependent character(s)**	**Recoded character(s)**	**Recoded taxa**	**Recoded state**	**Reason**
209 (*cont.*) Osteoderms	210, 211, 345	210, 211 (*cont.*)	(*cont.*) *Brachydectes* (S), *Broomistega* (P), *Callistomordax*, *Celtedens* (S), *Capetus* (P), *Chomatobatrachus* (P), *Cochleosaurus*, *Dendrerpeton*, *Doleserpeton*, *Edingerella*, *Eryops*, *Gerobatrachus*, *Glanochthon*, *Greererpeton*, *Iberospondylus* (P), *Isodectes* (P), *Lapillopsis* (P), *Limnogyrinus* (P), *Lydekkerina*, *Mastodonsaurus*, *Metoposaurus*, *Micromelerpeton*, *Micropholis*, *Onchiodon*, *Paracyclotosaurus*, *Platyoposaurus* (P), *Platyrhinops*, *Proterogyrinus*, *Rhineceps* (P), *Rhynchonkos* (S), *Sangaia*, *Sclerocephalus*, *Siderops*, *Trematolestes*, *Trimerorhachis*, *Uranocentrodon*	-	Osteoderms scored as absent (0), but scored for articulation and width in 210 an 211.
		345	*Acanthostomatops*, *Ambystoma*, *Amphibamus*, *Apateon*, *Archegosaurus*, *Australerpeton* (P), *Balanerpeton*, *Batropetes* (S), (*cont.*)	-	Osteoderms scored as absent (0), but scored for fusion to ribs in 345.

Continued on the next page

TABLE 3.5. *(cont.)* List of characters rescored due to logical inconsistencies. For characters 1–345, any character and/or taxon followed by a *P* was only rescored in the Pardo et al. (2017) matrix, and any followed by an *S* was only rescored in the Schoch et. al (2020) matrix. Characters 346–360 are only present in the Schoch et al. (2020) matrix.

Independent character	**Dependent character(s)**	**Recoded character(s)**	**Recoded taxa**	**Recoded state**	**Reason**
209 (*cont.*) Osteoderms		345 (*cont.*)	(*cont.*) *Brachydectes* (S), *Broomistega* (P), *Callistomordax*, *Celtedens* (S), *Cochleosaurus*, *Cryptobranchus*, *Dendrerpeton*, *Doleserpeton*, *Edingerella*, *Eocaecilia*, *Epicrionops*, *Eryops*, *Gerobatrachus*, *Glanochthon*, *Greererpeton*, *Hynobius*, *Iberospondylus* (P), *Ichthyophis*, *Isodectes* (P), *Karaurus*, *Lapillopsis* (P), *Leptodactylus*, *Limnogyrinus* (P), *Lydekkerina*, *Mastodonsaurus*, *Metoposaurus*, *Micromelerpeton*, *Micropholis*, *Onchiodon*, *Paracyclotosaurus*, *Platyoposaurus* (P), *Platyrhinops*, *Proterogyrinus*, *Rhineceps* (P), *Rhynchonkos* (S), *Sclerocephalus*, *Siderops*, *Trematolestes*, *Triadobatrachus*, *Triassurus* (S), *Trimerorhachis*, *Xenopus*	-	Osteoderms scored as absent (0), but scored for fusion to ribs in 345.
216 Supratemporal	47, 50, 53, 54, 56, 62, 217	47	*Batropetes* (S), *Rhynchonkos* (S)	?	Supratemporal scored as absent (1), but 47 defined based on it.

Continued on the next page

TABLE 3.5. *(cont.)* List of characters rescored due to logical inconsistencies. For characters 1–345, any character and/or taxon followed by a *P* was only rescored in the Pardo et al. (2017) matrix, and any followed by an *S* was only rescored in the Schoch et. al (2020) matrix. Characters 346–360 are only present in the Schoch et al. (2020) matrix.

Independent character	Dependent character(s)	Recoded character(s)	Recoded taxa	Recoded state	Reason
216 (*cont.*) Supratemporal	47, 50, 53, 54, 56, 62, 217	50	*Batropetes* (S), *Epicrionops*, *Gerobatrachus*, *Ichthyophis*, *Rhynchonkos* (S)	?	Supratemporal scored as absent (1), but 50 defined based on it.
		53	*Karaurus* (P)	-	Supratemporal scored as absent (1) in 216, but scored for its proportions in character 53.
		54	*Gerobatrachus*, *Karaurus* (P)	-	Supratemporal scored as absent (1), but its width was scored for in 54.
		56	*Batropetes* (S), *Leptodactylus*, *Rhynchonkos* (S), *Xenopus*	-	Supratemporal scored as absent (1), but its contribution to otic notch scored for in 56.
		62	*Batropetes*, (S), *Brachydectes* (S), *Gerobatrachus*, *Rhynchonkos* (S)	-	Supratemporal scored as absent (1), but 62 defined based on its presence.
		217	*Batropetes* (S), *Brachydectes* (S), *Celtedens* (S), *Rhynchonkos* (S)	?	Supratemporal scored as absent (1), but its exposure on the occiput scored for in 217.
230 Postorbital	48, 49, 231, 232	48, 49	*Plagiosuchus*	-	Postorbital scored as absent (1), but scored for shape in 48 and 49.
235 Postparietals	236, 237	236	*Batropetes* (S)	?	Postparitals scored as absent (2), but scored for contact with squamosal.
239 Tabular	63, 65, 66, 67, 71, 73, 234	63	*Ambystoma* (S), *Cryptobranchus* (S), *Hynobius* (S), *Karaurus*, *Kokartus* (S)	-	Tabular scored as absent (1), but scored for contact with squamosal in 63.

Continued on the next page

TABLE 3.5. *(cont.)* List of characters rescored due to logical inconsistencies. For characters 1–345, any character and/or taxon followed by a *P* was only rescored in the Pardo et al. (2017) matrix, and any followed by an *S* was only rescored in the Schoch et. al (2020) matrix. Characters 346–360 are only present in the Schoch et al. (2020) matrix.

Independent character	Dependent character(s)	Recoded character(s)	Recoded taxa	Recoded state	Reason
239 (*cont.*) Tabular		71	*Ambystoma* (S), *Celtedens* (S), *Cryptobranchus* (S), *Eocaecilia* (S), *Hynobius* (S), *Karaurus*, *Kokartus* (S), *Leptodactylus*, *Triadobatrachus*, *Xenopus* (S)	?	Tabular scored as absent (1), but 71 defined based on it.
264 Splenials	265, 310, 322	265	*Eocaecilia*, *Platyrhinops*	-	Splenials scored as absent (2), but scored for exposure in 265.
		310	*Eocaecilia* (P), *Platyrhinops* (P)	-	Splenials scored as absent (2), but scored for contact with first coronoid.
271 Coronoids	147, 148, 272, 273	148	*Celtedens* (S), *Cryptobranchus*, *Hynobius*	-	Coronoids scored as absent (3), but 148 scored for number of coronoid teeth.
		147, 272, 273	*Celtedens* (S)	?	Coronoids scored as absent (3), but characters related to coronoid teeth scored for.
314 Prefrontals	39, 40, 41, 42, 43, 218, 315	314, 315	*Plagiosuchus*	0	All dependent characters scored for, assumed indepedent character was a typographical mistake.
328 Dorsomedial process of sphenethmoid	329	329	*Batropetes* (S), *Brachydectes* (S), *Rhynchonkos* (S)	?	Dorsomedial process scored as absent (0) in 328, but its shape scored for in 329.

Continued on the next page

TABLE 3.5. *(cont.)* List of characters rescored due to logical inconsistencies. For characters 1–345, any character and/or taxon followed by a *P* was only rescored in the Pardo et al. (2017) matrix, and any followed by an *S* was only rescored in the Schoch et. al (2020) matrix. Characters 346–360 are only present in the Schoch et al. (2020) matrix.

Independent character	Dependent character(s)	Recoded character(s)	Recoded taxa	Recoded state	Reason
346 Ectopterygoid	76, 88, 131, 132, 133	88	*Celtedens*, *Doleserpeton*, *Eocaecilia*, *Triadobatrachus*	?	Dependent characters recoded because new presence/absence character not accounted for.
		131, 132, 133	*Doleserpeton*, *Triadobatrachus*	?	Dependent characters recoded because new presence/absence character not accounted for.
348 Cleithrum	175, 176, 177	175, 176, 177	*Triadobatrachus*	?	Dependent characters recoded because new presence/absence character not accounted for.
248, 356† Palatine	80, 102, 117, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 252, 355	123	*Ambystoma*, *Hynobius*	?	Dependent characters recoded because new presence/absence character not accounted for.
		127	*Ambystoma*	-	Dependent characters recoded because new presence/absence character not accounted for.
		252	*Ambystoma*, *Cryptobranchus*, *Hynobius*, *Karaurus*	-	Dependent characters recoded because new presence/absence character not accounted for.

† Characters 248 and 356 both allow the character state 'Palatine absent'. Character 248 recoded by Schoch et al. (2020).

equivalent (e.g., Schoch et al. 2020, see below), they represent two different types of 'unknowns'. Scoring characters as missing, commonly coded as '?', represents lack of knowledge/information. An example of this would be that, due to taphonomic processes, only the cranium of an individual is fossilised. Logic dictates that the organism had, at least, a vertebral column, but we cannot score any postcranial characters because we are missing the necessary biological materials to do so. Scoring characters as inapplicable ('-'), on the other hand, denotes a logical impossibility. Using Maddison's (1993) classic example of tail presence and colour, if we apply conventional coding (Hawkins et al. 1997), sometimes referred as contingent coding (Forey & Kitching 2000) or hierarchical characters (Hopkins & St. John 2021), we would need two characters to explain the variation in tail colour: character A - tail present/absent, and character B - if present, tail blue/red. Thus, character B would be inapplicable for any taxa without a tail, as a tail cannot be both absent AND have colour. Such reconstructions can, however, occur during character optimisation, affecting tree scores (e.g., Vignette 2 in Brazeau et al. 2019, Maddison 1993).

While there is still debate on how best to code variation that, just like the tail colour example, is logically and/or biologically dependent (e.g., Hawkins et al. 1997, Hopkins & St. John 2021, Maddison 1993, Wilkinson 1995*a*), it is sometimes suggested that, once a presence/absence character has been defined, any new character that subdivides the original character's 'presence' state should be hierarchically/conditionally defined (e.g., Hawkins et al. 1997). In other words, if character A is tail presence/absence, we should not define a new three-state character (C) as tail absent, spotted tail, stripy tail, rather it should follow the same character definition as character B, i.e. character C - if present, tail spotted/stripy. In Pardo et al.'s (2017) matrix, this suggestion was not always followed resulting in multiple instances of logically inconsistent/impossible character definitions and/or scores (table 3.5). These coding inconsistencies/impossibilities were rescored based exclusively on the independent character(s) definition/scoring (except for character 314, see table 3.5), no specimens were re-examined and, if a revision of character construction/definition was warranted, dependent characters were rescored conservatively as either inapplicable ('-') or missing ('?'). When the inconsistent/impossible codings are corrected and the matrix re-analysed under default settings, PAUP* yields 108 MPTs in two 1-TBR islands, all displaying the restricted Lissamphibia hypothesis (table 3.4). These islands correspond to the 1-TBR islands-18 and -90 of the original parsimony analysis (Fig.2.3ac).

Lastly, the presence of inapplicable data also raises the question of how best to analyse these characters. PAUP* allows users to choose between treating inapplicable data as missing or as a new character state (pages 40–41 in the manual), with the former as default. When Pardo et al.'s (2017) matrix is analysed with 'GapMode=NewState', PAUP* finds 351 MPTs in two 1-TBR islands (table 3.4). Interestingly, the new set of MPTs displays the restricted Lissamphibia hypothesis, but with the clade nested in Stereospondyli not Dissorophoidea, which can also be seen in island-72 of the default search (Fig.2.3b). This placement is found for both the original and rescored matrices. Thus, as has been shown for other datasets (e.g., Brazeau et al. 2019,

Maddison 1993), choice of how to treat inapplicable data clearly affects how many MPTs are found, their length, and, even, the topologies found. However, for as long as systematists have been debating how to code logically dependent characters, they have also been debating how to analyse these characters once conventional coding has been applied (e.g., Hawkins et al. 1997, Maddison 1993).

Recently, Brazeau et al. (2019) have developed and implemented an algorithm that is aware of inapplicable data during character optimisation, and does not require the relationship between hierarchical characters to be specified (unlike for example De Laet 2015, Goloboff et al. 2021).They found that treating inapplicable data as 'missing', 'new state' or 'inapplicable' often yields different sets of MPTs, for the same dataset, but not a clear pattern in the results' variation, i.e. no approach always found the shortest trees, the most resolved SC, the broadest sampling of treespace, etc. Analysing the original Pardo et al. (2017) matrix with Morphy v.0.2 (Brazeau & Desjardins 2020), with 1000 replicates and random addition sequence, yielded 585 MPTs (351 of which were also found in the original analysis) that display the extended Lissamphibia hypothesis. Using the *xRFislands* function from the *islandNeighbours* R package (Serra Silva & Wilkinson 2021), identifies three island structure patterns for the Morphy MPTs. With an RF=2 threshold the function identifies four islands (72, 108, 162 and 243 trees), at $4 \leq RF \leq 10$ two islands (180 and 405 trees), and at RF=12 all MPTs are part of a single island. The primary topological differences between the 10-RF islands are the placement of (*Rileymillerus*+*Chinlestegophis*+Gymnophiona) within Stereospondyli and the internal relationships of Batrachia, patterns that can also be seen in the 882 MPT set. Thus, the phylogenetic trees inferred from Pardo et al.'s (2017) matrix vary greatly, not only between analyses under different optimality criteria, but also between analyses where the treatment of inapplicable data and polymorphic taxa varies (table 3.4).

3.4.3 Differences between the Pardo et al. (2017) and Schoch et al. (2020) datamatrices

When investigating the phylogenetic placement of the Triassic stem-salamander *Triassurus sixtelae*, Schoch et al. (2020) used a matrix modified from the Pardo et al. (2017) matrix. They bypassed the issues detailed in the previous subsection by rescoring all inapplicable and polymorphic states as missing ('?'), and further modified Pardo et al.'s (2017) matrix by adding 15 new characters, revising the scores for ten others, and adjusting taxonomic sampling to include a chimaeric albanerpetontid, three lepospondyls (*Brachydectes*, *Batropetes* and *Rhynchonkos*) and exclude 19 taxa across Temnospondyli (see supplemetary for Schoch et al. 2020). This 62 taxa and 360 character matrix yields two MPTs that display a restricted Lissamphibia in Dissorophoidea, with albanerpetontids sister to Batrachia, and *Chinlestegophis* as sister to *Rileymillerus*, within Stereospondyli (Fig.3.5). Resampling analyses find low to moderate support for the restricted Lissamphibia hypothesis (bootstrap=77%, delete-half jackknife=59% and Farris jackknife=67%). Bayesian inference analyses of this matrix were attempted with MrBayes, but they did not

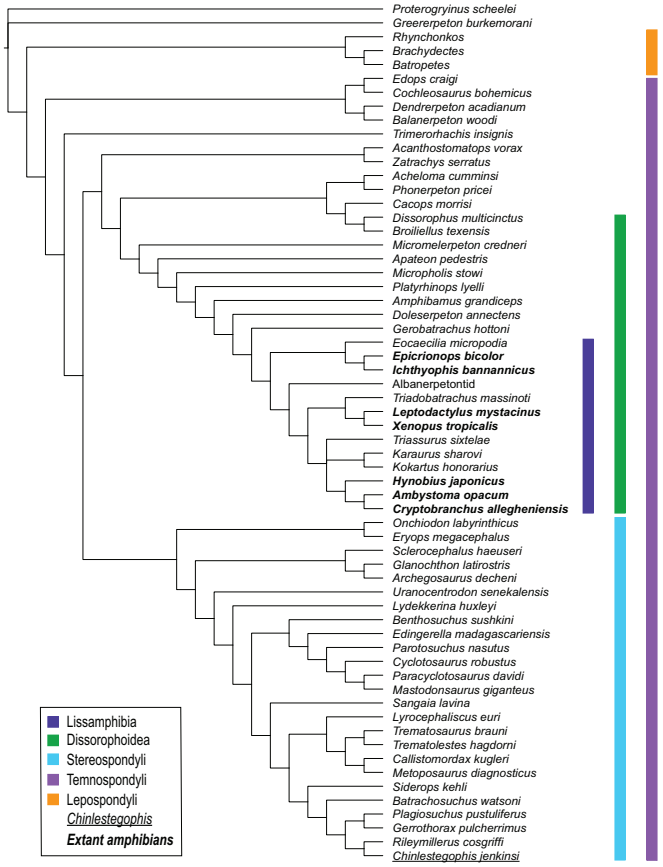


FIGURE 3.5. Stric consensus tree of the two most parsimonious trees yielded by analysis of the Schoch et al. (2020) matrix, with higher order amphibian taxonomy highlighted.

converge (the longest attempt ran for 200 million generations).

Further parsimony analyses reveal that it is the 15 additional characters in Schoch et al.'s (2020) matrix that are responsible for the recovery of the restricted Lissamphibia, since when these characters are removed the extended Lissamphibia hypothesis is inferred. This occurs when analysing either all 62 taxa in Schoch et al.'s (2020) original matrix, or just the 57 taxa shared between Schoch et al.'s (2020) and Pardo et al.'s (2017) matrices. Additionally, while inapplicable data is no longer coded as such (all '-' recoded as '?'), character descriptions were not adjusted and hierarchical relationships between characters are still present in the matrix (table 3.5). This is confirmed by the analysis of the rescored matrix yielding fewer and shorter MPTs than the original matrix (table 3.4), the same pattern found for Pardo et al.'s (2017) dataset. And, just as with the latter dataset, the rescored 345 character matrix no longer recovers an extended Lissamphibia, rather the inferred trees display a restricted Lissamphibia within Dissorophoidea. In summary, despite stronger support for the restricted Lissamphibia hypothesis, minor changes to Schoch et al.'s (2020) data matrix can still recover the extended Lissamphibia hypothesis introduced by Pardo et al. (2017).

3.5 Discussion

To date the presence and effects of tree islands have, primarily, been explored in the contexts of tree search (e.g., Höhna & Drummond 2011, Lakner et al. 2008, Olmstead et al. 1993) and consensus building (e.g., Maddison 1991, Serra Silva & Wilkinson 2021, Sharkey & Leathers 2001, Sumrall et al. 2001). In the latter, the phenomenon of large island bias has lead to the recognition that the MRC is often a poor summary of (multi)sets of trees, as the trees in the largest island(s) drown the signal from trees in other islands. This insight urged workers to advocate either for the use of the SC (Sumrall et al. 2001) or weighted consensuses (e.g., Serra Silva & Wilkinson 2021, Sumrall et al. 2001) if a single summary is desired, or for island-partitioned consensuses (e.g., Maddison 1991, Serra Silva & Wilkinson 2021). This idea of summarising islands, or any such subset, of trees individually rests on the assumption that there is more variation between islands than within, and that "a single consensus may ignore information in the data" (Hendy et al. 1988, p.358). And, as shown in Chapter 2, the "ignore[d] information" can greatly affect the conclusions made about a group's relationships and evolutionary history, with most MPTs inferred (under PAUP*'s default settings) from the Pardo et al. (2017) dataset supporting a greatly extended Lissamphibia while most islands supported the traditional restricted Lissamphibia. It is, thus, not insensible to question whether the presence of islands, more specifically large island bias, might also affect post-inference topology-based taxonomic instability tests.

The split frequency-based LS analyses above suggest that island bias is not an issue for the identification of unstable taxa, but that analyses of the complete (multi)set and each of its islands can be complementary. The results of the analyses on the Pardo et al. (2017) MPTs found that the unpartitioned analyses identify the set of taxa whose placement changes between islands as the least stable taxa (Lissamphibia+*Chinlestegophis*), whereas the partitioned-by-island analyses pick up on locally unstable taxa (table 3.2, Figs.2.2b and 2.3). Thus, the unpartitioned analyses find the taxa responsible for island structure, while partitioned analyses find the taxa responsible for island size(s), i.e. locally unstable. While any taxon's identification as unstable in either the partitioned or unpartitioned analyses warrants further investigation, whether the aim is to understand the causes of or decrease topological variation, the identification of a taxon as unstable at the global AND local levels might be a sensible indicator for its removal.

The consensus optimisation-based RBIC analyses, on the other hand, did not find any complementarity between partitioned and unpartitioned analyses, but did yield a potential indicator for the presence of islands, if these are not known *a priori*. RBIC's default dropset is of one taxon, however it was shown by Wilkinson & Crotti (2017) that this highly conservative parametrisation can lead to the failure to identify all unstable taxa, particularly if they belong to an (internally-stable) unstable clade. For the Pardo et al.'s (2017) MPTs, when the dropset is allowed to go up to 15 taxa, the RBIC identifies Lissamphibia+*Chinlestegophis* as the (group of) taxa whose removal shows the greatest improvement to the SC resolution. This result is unsurprising given they are the taxa driving island structure, but is interesting in that it suggests that the recovery of groups

of unstable taxa, rather than single unstable taxa, might be indicative of island presence. As such, while large island bias does not appear to affect topology-based instability analyses, the latter can be used either to explore taxonomic instability within and between islands (if known), or as an indicator for the possible presence of island structure. Using an RBIC-like method as a test for the potential presence of islands also has the benefit of not requiring the extensive tree-to-tree distance calculations needed to visualise tree space occupation, and thus island structure, with multidimensional scaling (MDS, see figure 2.5).

Given that phylogenetic trees are special cases of phylogenetic networks (Huson et al. 2010), and that with a 50% threshold a consensus network is also the MRC (Holland & Moulton 2003), it is unsurprising that non-tree networks computed from split frequencies are also affected by large island bias. With the two largest, and most topologically similar (Figs.2.2b and 2.3d), islands making up ≈ 80% of all Pardo et al.'s (2017) MPTs, minor increments in split frequency threshold result in major changes to the consensus networks. With the 20% threshold network displaying a tree-like topology (Fig.3.1b), while the 10% threshold consensus consists almost entirely of cycles/3-cubes, resembling an actual net (Fig.3.1c). Thus, in the presence of large island bias consensus networks, and any reticulation network that uses them as input, can be uninformative and, if the goal is to explore the different evolutionary hypotheses supported by the islands, networks are not an effective alternative to partitioned consensus trees approaches. While the networks computed in this chapter might not be representative of the data/island structure underpinning the (multi)sets of trees commonly summarised with consensus networks, it may be beneficial to, prior to any splits-based network analysis, use RBIC-like methods or MDS to ascertain whether island structure is likely to be present.

Tree islands and its effects aside, the Pardo et al. (2017) dataset is particularly interesting due to the myriad ways it can be analysed, even when restricted to heuristic parsimony tree searches (table 3.4). Because the data matrix includes inapplicable data and polymorphic taxa it raises the still open questions of how best to analyse these types of character scores (e.g., Brazeau et al. 2019, Nixon & Davis 1991). Even without addressing the inconsistent use of conventional coding in character construction/descriptions (table 3.5), the very different sets of MPTs recovered, when all (implemented) possible ways of parametrising inapplicable data and polymorphism are compared, show that the existing matrix does not yield robust tree inferences and it should not be used to make definitive statements about lissamphibian relationships. This lack of robust inferences from morphological matrices is not restricted to Pardo et al.'s (2017) data matrix, with Marjanović & Laurin (2019) finding that minor analytical changes to searches on a carefully curated matrix (all characters and their scores revised) for early limbed vertebrates does not yield robust topologies.

The Bayesian analyses of Pardo et al.'s (2017) matrix at first appeared to display less topological variation than the parsimony analyses, with multiple MRCs displaying the extended Lissamphibia hypothesis. However, a closer look at the results of the analysis on the explicitly

partitioned data matrix revealed that, much like in the parsimony analyses, Bayesian analyses can recover more than one 'major' topology from the Pardo et al. (2017) data. Additionally, morphological analyses in MrBayes can only be run with the Mk or Mkv models (Lewis 2001) and always treat polymorphism as uncertainty (pages 11–12 and 85–87 in the manual), which restricts the analyses that can be performed to explore how polymorphic taxa and inapplicable data affect the inferred trees, much as was done in the parsimony context. Also, the performance of Bayesian inference *vs.* parsimony debate (e.g., Goloboff et al. 2017, Wright & Hillis 2014, O'Reilly et al. 2016, Puttick et al. 2017) aside, recent works have contended that the Mk(v) model is not adequate for analyses of morphological datasets (e.g., Goloboff et al. 2019, Goloboff & Arias 2019), which raises the question of whether Bayesian inference on morphological datasets amounts to an exercise in model misspecification. Yet, there are currently no other identifiable and easily implementable models for morphological evolution available. Recently, Tarasov (2019) proposed combining knowledge of ontology and developmental processes with structured (SMM, Nodelman et al. 2002) and hidden Markov models (HMM, Beaulieu & O'Meara 2014) to address the problems of inapplicable data and character construction subjectivity during tree inference. However, its is unclear how scalable this SMM-based approach is, how differently do explicit Mk models and SMMs that condense into a multistate Mk model behave, and how the models behave with empirical data. Thus, there are still a lot of questions surrounding model-based inference of morphological data.

As for character coding, while I looked exclusively at logic violations in the scoring of dependent characters (table 3.5), a more thorough review of character construction and coding (akin to the one seen in Marjanović & Laurin 2019) may yield yet another set of MPTs and supported lissamphian relationships. Along with the ever present possibility of typographical errors (see characters 216 and 314 in table 3.5) and the sometimes uncertain homology resulting from bone loss/fusion (e.g., Maddin et al. 2016, Schultze et al. 2008) and nomenclatural variation (e.g., Abel & Werneburg 2021, Schultze et al. 2008) across taxa, Schoch et al.'s (2020) recoding of some character scores, and their reasons, suggests that there are at least some characters whose scoring is not agreed upon by all workers. Also, it has been shown that palaeontologists and neontologists approach character scoring differently (e.g., Harris 2005, and M Wilkinson 2018, personnal communication, 16 October), which may again influence character scoring and tree inference, given that extinct and extant taxa are present in the Pardo et al. (2017) and Schoch et al. (2020) matrices. As such, a thorough revision of character construction and scoring of both matrices, based on biological and logical criteria, may be warranted.

Thus, the effect of tree islands, particularly large island bias, on tree summaries and instability analyses (and possibly branch support, although further work is needed to confirm this) is an important consideration when exploring (multi)sets of trees and the evolutionary histories they display. And, the drastic change in island structure/size seen between the various inferences on Pardo et al.'s (2017) matrix are a reminder that phylogenetic analyses are only as good as their

underlying data and that analytical settings are not one size fits all.

3.6 Data availability

The data, scripts and results for this chapter are available from the following repository:

- Figshare, `https://doi.org/10.6084/m9.figshare.c.6050033`.

CHAPTER 4

ON PARTITIONING SETS OF PHYLOGENIES WITH NON-IDENTICAL LEAF SETS

4.1 Introduction

Confronted with (multi)sets of inferred phylogenetic trees, common practice in systematic biology has been to represent them with a single consensus or supertree (Felsenstein 2004). In contrast, several researchers have suggested that when sets of trees are not homogenous it is better to partition the trees and summarise the subsets separately (e.g., Hendy et al. 1988, Maddison 1991, Serra Silva & Wilkinson 2021, Sumrall et al. 2001). In practice, most methods for discovering heterogeneity in and partitioning sets of trees have been developed for the special case where the trees have identical leaf sets, the so called 'consensus case' (e.g., Bonnard et al. 2006, Serra Silva & Wilkinson 2021, Stockham et al. 2002). However, sets of trees on non-identical leaf sets are increasingly common, e.g. from genomic scale data. Thus, there is a need for methods for interrogating (multi)sets of trees in this, more general, 'supertreee case' and partitioning them when they are found to be heterogeneous.

Sets of inferred trees may be heterogeneous due to real differences in evolutionary histories, such as those produced by gene duplication and loss, incomplete lineage sorting, etc. (e.g., Chan et al. 2020, Hahn 2007), or because of incorrect inferences, such as those produced by systematic biases and/or stochastic and other errors (e.g., Léveillé-Bourret et al. 2017, Simmons et al. 2022). Though I will not explore these, many important approaches to investigating heterogeneity focus upon the underlying data from which trees are inferred (e.g., statistical binning, Bayzid et al. (2015); ortholog enrichment, Siu-Ting et al. (2019)). Other approaches that focus on the trees themselves without recourse to the underlying data have also been developed, almost exclusively within the consensus context (e.g., islands of trees, Maddison (1991) and Serra Silva & Wilkinson

(2021); families of trees, Hendy et al. (1988); k-medoids clustering, Tahiri et al. (2018); multipolar consensus, Bonnard et al. (2006); consensus networks, Holland & Moulton (2003); tree alignment graphs, Smith et al. (2013)). These can be broadly divided into those based on split compatibility (e.g., multipolar consensus and tree alignment graphs) or on tree-to-tree distances, which seek to produce either consensus networks or to partition (multi)sets of trees into more homogeneous subsets (e.g., families and islands of trees, k-medoids clustering). Here, I present a novel, tree-to-(super)tree distance based approach to partitioning any heterogeneous (multi)set of trees that does not make use of underlying data and which can be applied in cases where the trees have non-identical leaf sets.

Islands of trees are the disconnected components of a graph where vertices correspond to trees and edges connect all trees within a fixed distance threshold (Maddison 1991, Serra Silva & Wilkinson 2021). Trees within islands may be quite dissimilar provided they are 'connected' by a series of sufficiently similar intermediates. Until recently, the fixed distance threshold was restricted to NP-hard branch-rearrangement metrics, mostly limiting their discovery to heuristic searches in parsimony analyses. By generalising the definition of islands to any tree-to-tree distance metric Serra Silva & Wilkinson (2021) freed them from the shackles of branch-rearrangement metrics, making possible the *a posteriori* partitioning of any (multi)set of trees into islands in the consensus case. Given the increasing prevalence of sets of trees with non-identical tips, further generalisation of tree set partitioning methods, such as islands, to the supertree case is desirable.

A first step in generalising any tree-to-tree distance-based method for partitioning sets of trees from the consensus to the supertree case is the generalisation of tree-to-tree distances, originally defined in the consensus case, to pairs of trees with non-identical leaf sets. This generalisation revolves, primarily, around how to make two distinct leaf sets identical, which can be achieved by pruning or grafting leaves to one or both trees, so as to render their leaf sets identical. In pruning, we use the intersection of the leaf sets to identify the subtrees induced by the shared taxa, whereas with grafting, we seek the pairs of (most similar) trees that display the original trees and the union of the leaf set (Cotton & Wilkinson 2007). However, both approaches have difficulty dealing with trees with no or minimal overlap, which is particularly problematic for partitioning methods that require exhaustive pairwise distance matrices, such as islands. Two imperfect solutions are to either set the distance between non-overlapping trees to zero, which can represent absence of conflict (e.g., Robinson-Foulds (RF, Robinson & Foulds 1981) and quartet (QD, Estabrook et al. 1985) distances), or to treat these instances as non-applicable (NA), which are generally dealt with by removal or replacement with imputed values (reviewed in Wagstaff 2004). Neither solution is well suited for methods like islands. Also, even when all distances are defined, the search for islands can be derailed by insufficient overlap between trees because small trees can be sufficiently similar to larger, highly disparate trees, thus placing the larger trees in the same island and masking tree heterogeneity (Fig.4.1). Additionally, distances between

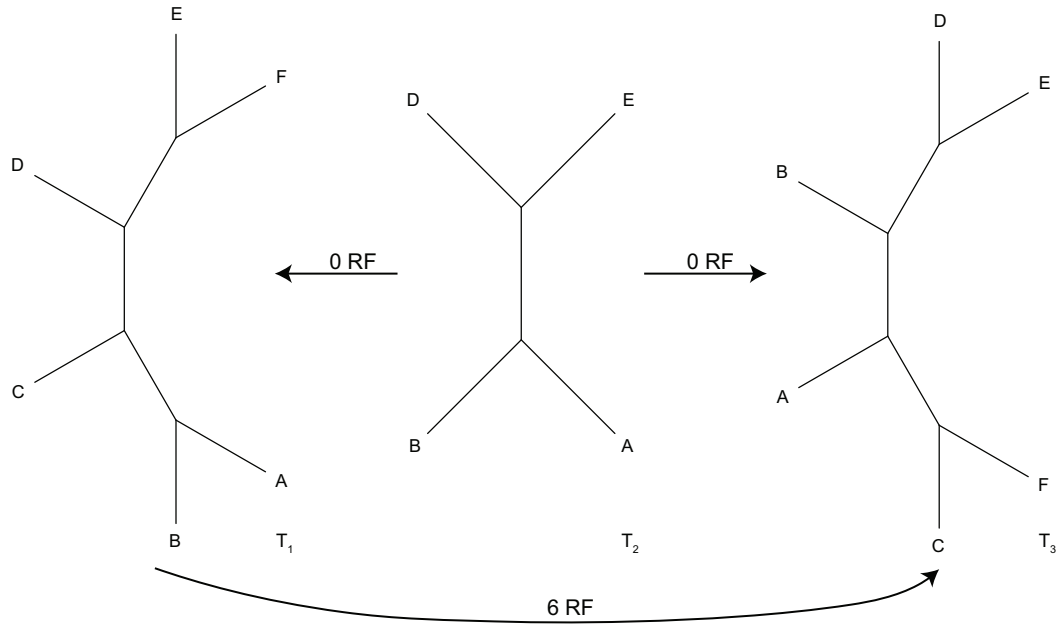


FIGURE 4.1. Small tree problem. Three partially overlapping trees where the two larger trees (T_1 and T_3) are maximally dissimilar under the RF distance, yet equidistant from T_2. In this scenario, despite having no splits in common T_1 and T_3 would be placed in the same RF-island.

different sized trees (can) break the triangle inequality, meaning that the tree-to-tree distances will no longer be metrics. Here, I define a new partitioning strategy, more similar to families of trees, that resolves the non-overlap problem by using a tree-to-supertree distance, eschewing exhaustive pairwise comparisons and the within subset connectivity that characterise islands.

4.2 Defining clumps of trees

When dealing with (multi)sets of trees with the same leaf set it may be possible to identify mutually exclusive and exhaustive subsets that correspond to alternative evolutionary histories and/or analytical artefacts, which cause topological incongruence (e.g., Hendy et al. 1988, Maddison 1991, Serra Silva & Wilkinson 2021). However, identifying those subsets in (multi)sets of trees with non-identical leaf sets requires overcoming a number of methodological obstacles, both universal (e.g., how to define distances between trees with no taxa in common) and method-specific (e.g., small trees can place larger, very dissimilar, trees in the same island, Fig. 4.1). And, while method-specific obstacles might be easily, if imperfectly, tackled/mitigated, they do not hold the answer to the key question of how to effectively compare trees with no, or very little, overlap. For example, akin to how removing trees that do not support a specified topology, have paralogous sequences, or do not contain select taxa is used in various post-processing analyses (e.g., Hime et al. 2021, Siu-Ting et al. 2019), filtering trees under a minimum leaf set size might prevent highly disparate trees from being clustered into the same island of trees. However, filtering trees

by size does not deal with pairs of non-overlapping trees, meaning that the question of how to define tree-to-tree distance(s) between non-overlapping trees remains.

The issue of how to compare trees with no, or minimal, taxonomic overlap can be addressed by using a clustering approach that relies on the distance from a pre-determined tree to all trees in the (multi)set. One such strategy is the use of families of trees (Hendy et al. 1988), where members of a subset are identified based on their distance to a tree T. If a (multi)set's supertree were set as T, it would ensure that, because all trees in the set partially overlap with the supertree, there would be no undefined tree-to-(super)tree distances. Unfortunately, families of trees lack an unambiguous formal definition, making it unclear not only how tree T is selected, but also how clustering proceeds past the identification of the first family. To prevent potential ambiguities while attempting to extend families of trees to sets of trees with non-identical leaf sets, I define a new clustering approach called clumps of trees (a play on clump *n.*: 1.a. a compact mass, 2.a. a cluster of trees and clump *v.*: 2.b. to make into a clump, Oxford English Dictionary (OED)). I define clumps of trees as the sets that are sequentially extracted from a tree distribution, with each subset containing all trees in a distribution within a selected distance of the input set's supertree, and any trees placed in a clump being removed from the partitionable distribution. In set theory terminology, the partitionable tree distribution is the complement of the subset trees ($\mathscr{S}^C$ in the formal definition below), i.e. the trees not yet assigned to a clump.

Formally, given i) a set $\mathscr{T}$ of trees and its supertree $\mathscr{T}_{ST}$, ii) a pairwise distance function $d : \mathscr{T} \times \mathscr{T}_{ST} \rightarrow \mathbb{R}_0^+$, and iii) a threshold $x \in \mathbb{R}_0^+$, I define a sequence of sets $\mathscr{S}$, such that $\mathscr{S}_n = \{d(\mathscr{T}, \mathscr{T}_{ST}) \leq x,\ x \in [d_{min}, +\infty)$ and $\mathscr{T} = \mathscr{S}^C\}$. The *tree clumps* of ($\mathscr{T}$, d, x) correspond to the exhaustive subsets of the sequence of sets $\mathscr{S}$. This definition allows for the threshold to be changed at each clump identification iteration, meaning that individual clumps can be defined on a different threshold distance to the supertree of the previous clump(s)'s $\mathscr{S}^C$. This flexibility in setting the threshold means that no trees are 'unclumpable', i.e. all trees will be placed in a cluster. I propose that the threshold distance be set at the troughs, points at which the sign of the first derivative of the density curve goes from negative to positive. Using histogram bin heights as proxy for the density curve, this allows for two 'strict' scenarios, one where every trough defines a clump (strict trough), the other where a clump is defined only when $y = 0$ (breaks), and a 'loose' scenario where the difference between the nearest peaks and troughs (Δamplitude) is used as a guide of whether to partition the distance distribution (Fig.4.2). As in Serra Silva and Wilkinson's (2021) generalised tree island definition, I do not specify a tree-to-tree distance metric on which to define clumps. As such, any tree-to-tree distance that can be applied to distributions of trees with non-identical leaf sets can be used, and the choice of whether to work with (multi)sets and/or branch lengths is left to the user.

Three special cases follow easily from the definition of clumps of trees: i) at least one supertree corresponds to an input tree, ii) all supertrees correspond to input trees, and iii) if all input trees have the same leaf set, the supertree will also be a consensus tree. While it is not immediately

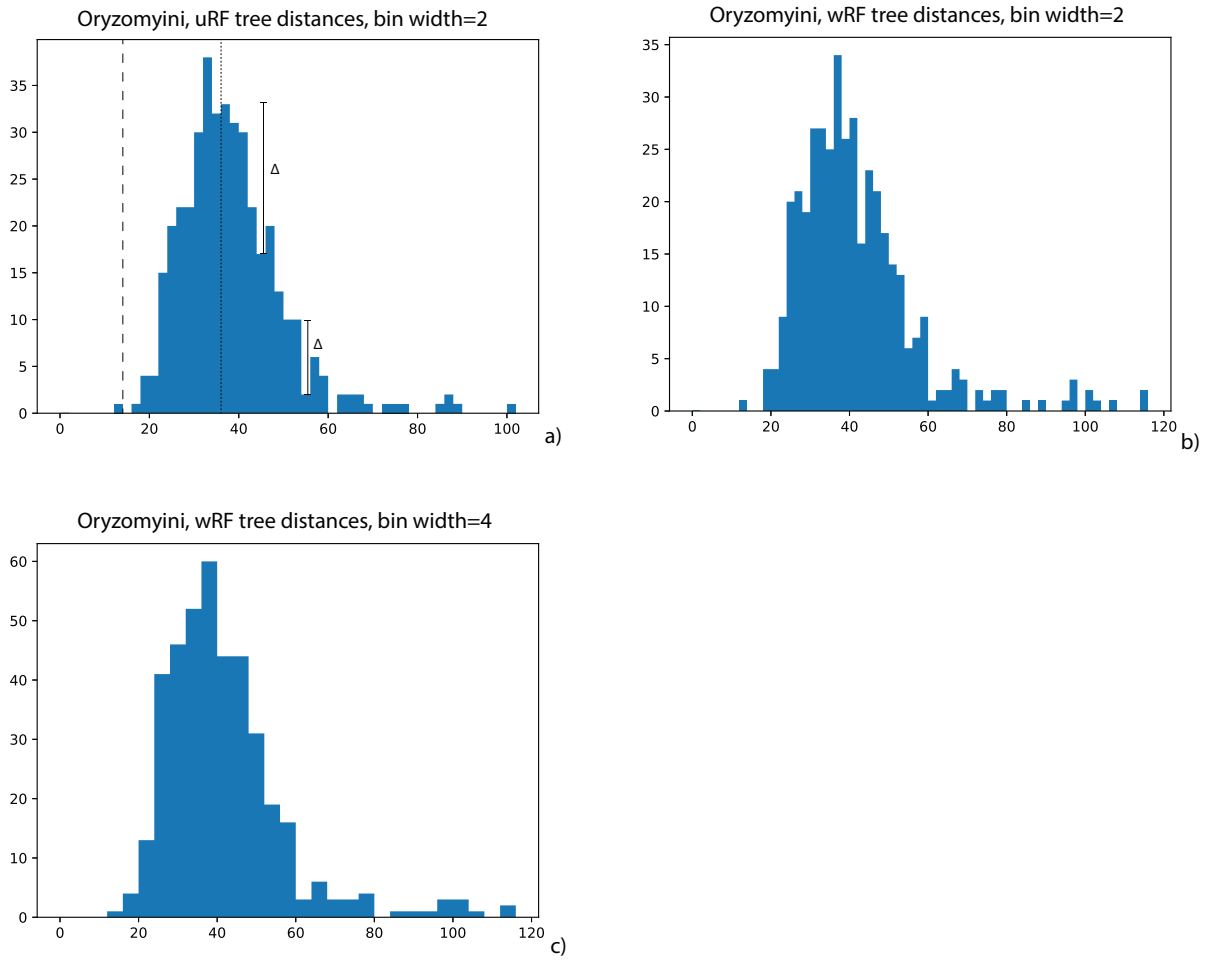


FIGURE 4.2. Example of the histograms produced during the clump extraction pipeline (see subsection 4.4.1), using Percequillo et al.'s (2021) Oryzomini dataset, under a) uncorrected Robinson-Foulds distance (uRF, Robinson & Foulds 1981) and bin width of two, b) weighted RF distance (wRF, see subsubsection 4.4.1.1) and bin width of two, and c) wRF and bin width of four. In a) the dashed line illustrates the break (y=0) partitioning approach, the dotted line illustrates the strict trough approach (after removal of the clump at RF=14), and the line segments illustrate the user-defined Δamplitude approach.

apparent how these special cases might affect clump identification, the choice of supertree (or consensus) method will indubitably affect the identification of clumps, since the use of distinct supertrees to identify clumps from the same tree distribution may yield different tree clump numbers and/or structure.

Additionally, because clumps are defined on the distance to a single tree, like families of trees, the clumps may not be mutually exclusive. While under identical analytical settings the same trees will be placed into the same clumps, changing bin sizes may lead to slightly different clump structures/make-ups being identified, and not necessarily by the process of clumps merging (as

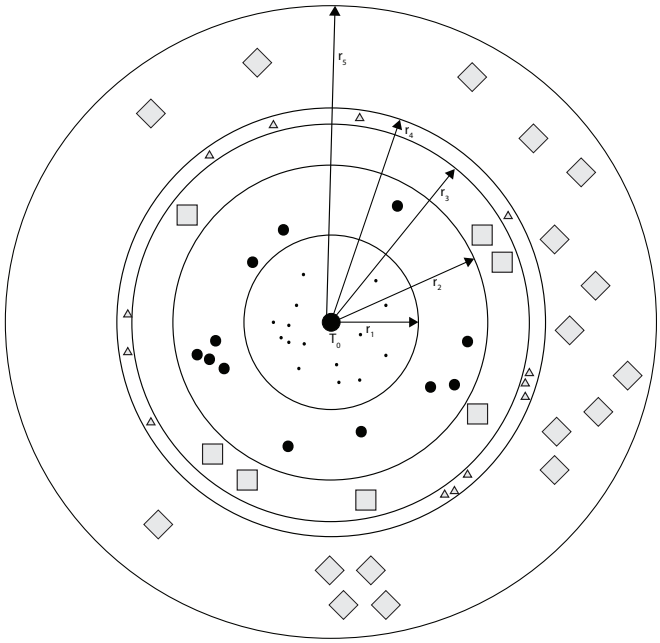


FIGURE 4.3. Stands of trees. Each concentric sphere (stand) contains trees that are within a bounded distance of the focal tree, T_0, and up to twice the stands' radius of each other. For example, the rhombuses are all within r_5 and $r_5 - r_4$ of T_0, but up to $2 \times r_5$ of each other.

is seen when island extraction thresholds are increased, see Chapter 2 section 2.4.3), see table 4.2. This will be of particular importance when dealing with sets of trees with non-identical leaf sets, since, if using uncorrected distances, we expect the smaller subtrees to fit with multiple clusters despite their expected placement within the first clump(s) to be extracted. However, as stated above, trees with fewer than a selected number of leaves can be filtered from the initial tree distribution, thus reducing the number of trees that might fit into multiple clumps.

An alternative to sequential clustering, which will not be tested here, might be to use the density distribution, or histogram, generated by the first clump search to identify concentric areas of trees in the (multi)set, with the bounds of each concentric sphere corresponding to a break or trough in the tree-to-supertree distance distribution (Fig.4.3). To denote the static nature of the identification of these subsets I refer to these as stands, from stand *n.1.*: 5.a. a state of checked or arrested movement, 29. a standing growth spec. one of trees and stand *v.*: 40. of a process: to remain stationary or unchanged (OED). However, it should be noted that such a snapshot of tree space would yield subsets of increasingly different trees with increasing distance from the focal tree (the supertree), as the maximum theoretical distance between trees within any stand will be twice the largest tree(in the stand)-to-supertree distance.

4.3 Examples

To test multiple aspects of clumping I will be using six distinct tree sets from studies that represent, arguably, some of the more commonly used types of phylogenetic data and types of studies these data are used in. The types of phylogenetic data used as the basis for the selected tree distributions include discrete morphological characters, and molecular sequences obtained using Sanger sequencing, target-capture, anchored hybrid enrichment (AHE), ultraconserved element (UCE) enrichment and transcriptomics (table 4.1). These data were employed in studies interested in the phylogenetic placement of particular taxa, answering macroevolutionary questions, and/or exploring how methodological choices affect inference analyses.

First, I will use Percequillo et al.'s (2021) AHE-based dataset to illustrate that clumping analyses can and do identify differing evolutionary histories present in the treeset. Given the supertree/supermatrix conflict they reported for oryzomyine rodents' clade D (Fig. 1 in Percequillo et al. 2021), I expect the clumping analysis to identify at least two major topologies in two clumps. Since all trees in the set have at least 50% of the sampled taxa present, the effects of tree size and complete non-overlap between trees on clumping cannot be addressed with this dataset.

I will be using a second AHE dataset (Hime et al. 2021), primarily, to test whether changing histogram bin width affects clump identification, for details on the analytical pipeline see subsection 4.4.1. While all datasets were clumped under multiple bin widths, the Hime et al. (2021) dataset was chosen specifically to test this analytical parameter for two reasons. First, the dataset was originally used to explore deep-time relationships between Lissamphibia (the least inclusive clade containing Gymnophiona, Caudata and Anura), and was shown to contain at least three major topological variants, corresponding to distinct lissamphibian evolutionary histories: i) Batrachia (Caudata+Anura), ii) Procera (Caudata+Gymnophiona), and iii) Acauda (Anura+Gymnophiona), see figure 1.5. Secondly, of the selected datasets, Hime et al.'s (2021) has the largest trees (table 4.1) and, thus, the widest range of possible RF distances between trees (0–576). As such, by modifying the bin width parameter we can not only compare how the number of clumps changes between analyses, but also which main topological variants are identified, and how the proportion of clumps supporting each hypothesis changes between analyses.

Another type of anchored markers commonly used in phylogenetic analyses are UCEs, which differ from AHEs by having shorter, less divergent sequences (Singhal et al. 2017). Because of UCEs' popularity I analysed two datasets that were used in studies attempting to resolve recalcitrant branches in Microhylidae (Streicher et al. 2020) and Squamata (Streicher & Wiens 2017). The Squamata dataset will be used for the same purposes as the Percequillo et al. (2021) rodent dataset, but looking at higher taxonomic levels and with every species represented by a single tip. Because the largest tree(s) in the microhylid dataset do not exceed 77% taxon sampling (the lowest observed maximum taxon sampling of all datasets), this dataset allows us to explore how much non-effective overlap affects clumping analyses.

The final dataset based on a single type of molecular data I will test is Siu-Ting et al.'s (2019)

TABLE 4.1. Molecular datasets used for clumping analyses. Datasets starting with *KST* refer to Siu-Ting et al. (2019). The datasets starting with *Chan* are from Chan et al. (2020), and the abbreviations stand for the following: *unf* is unfiltered, *75tax* is 75% taxon completeness and *50pars* is 50% parsimony informative. In the datatype column, AHE corresponds to anchored hybrid enrichment, and UCE to ultraconserved elements.

Dataset	Ingroup taxa	Datatype	Trees	Taxon sampling	Supertree size
Percequillo et al. (2021)	Oryzomyini	AHE	402	38–75	75
Hime et al. (2021)	Lissamphibia	AHE	220	201–290	291
Streicher et al. (2020)	Microhylidae	UCE	1796	24–49	64
Streicher & Wiens (2017)	Squamata	UCE	4178	18–34	37
KST2656	Lissamphibia	Transcriptome	2656	4–32	33
KST2019	Lissamphibia	Transcriptome	2019	4–32	33
KST768	Lissamphibia	Transcriptome	768	6–32	33
KST348	Lissamphibia	Transcriptome	348	6–32	33
KST2656-KST2019	Lissamphibia	Transcriptome	637	4–32	33
KST2656-KST768	Lissamphibia	Transcriptome	1888	4–32	33
KST2656-KST348	Lissamphibia	Transcriptome	2308	4–32	33
KST2019-KST348	Lissamphibia	Transcriptome	1671	4–32	33
Chan all	Rhacophoridae	Target-capture, UCE, Sanger	23186	4–50	50
Chan legacy	Rhacophoridae	Sanger	30	4–50	50
Chan exon unf	Rhacophoridae	Target-capture	10291	4–50	50
Chan intron unf	Rhacophoridae	Target-capture	12247	4–50	50
Chan UCE unf	Rhacophoridae	UCE	648	4–50	50
Chan exon 75tax	Rhacophoridae	Target-capture	7487	38–50	50
Chan intron 75tax	Rhacophoridae	Target-capture	7555	38–50	50
Chan UCE 75tax	Rhacophoridae	UCE	532	38–50	50
Chan exon 50pars	Rhacophoridae	Target-capture	6184	6–50	50
Chan intron 50pars	Rhacophoridae	Target-capture	6127	11–50	50
Chan UCE 50pars	Rhacophoridae	UCE	325	23–50	50

transcriptome-based tree (sub)sets, which were originally used to explore the effects of paralog inclusion on topological incongruence, branch support and divergence estimates on lissamphibian relationships. This dataset, KST hereafter, is particularly interesting because it is made up of multiple subsets of a larger tree distribution, which allows for the exploration of whether the number of trees in a set affects clump identification, and might shed light on whether the topologies most similar to a dataset's summary tree are found in the largest clump(s), the most clumps, or a combination of both. And, given that two of the subsets include quartet trees, we can test the assumption that under uncorrected distances the smallest trees are indeed swept into the first clump(s).

While each of the datasets above uses a single type of molecular data, there are also studies that make use of multiple types of nucleotide data. One such study was Chan et al.'s (2020) exploration of the phylogenetic relationships within Rhacophoridae, a family of Old World tree frogs, which used a combination of target-captured exon-only and intron-only, UCE and Sanger sequenced datasets obtained with the FrogCap probe design and pipeline (Hutter et al. 2021). This dataset allows us to explore how common filtering parameters (taxon coverage and proportion of parsimony informative sites) change the tree-to-supertree distance distributions for each type of molecular marker and how this may inform post-processing of tree sets based on multiple types of nucleotide data.

Lastly, the set of most parsimonious trees (MPTs) recovered from the morphological data matrix used by Pardo et al. (2017) to place *Chinlestegophis jenkinsi* in a phylogeny of fossil amphibians will be used to explore some of the special cases detailed in the previous section. Particularly, how the change in supertree method might influence the clumps that are identified. As one of the tree distributions used to illustrate Serra Silva and Wilkinson's (2021, also Chapter 2) generalised definition of islands of trees, I will also compare the output of the clumping analyses, under the strict trough and break scenarios, to the known 10- and 12-RF island sets.

4.4 Extracting clumps from distributions of trees with non-identical leaf sets

4.4.1 *clumpy* Python pipeline

To facilitate the extraction of clumps, I wrote four Python 3.8 (Van Rossum & Drake 2009) pipelines, all using the algorithm below, to identify and extract clumps from any distribution of trees. The consensus pipelines, see section 4.5, can only be applied to trees on the same leaf set, but the supertree pipelines can be applied to any set of trees with complete or partial leaf set overlap. The latter currently uses Astral-III v.5.7.7 (Zhang et al. 2018) to compute the supertree, but the pipeline can be modified to allow any Newick-compatible supertree building software to be called instead of Astral. Also, to allow the pipeline to be used from text-only systems and systems with graphic capabilities two versions of the supertree and consensus pipelines were

written: one prints the tree-to-supertree histograms as text directly on the terminal, using the *plotille* module (Ippen 2013); the other uses the *Matplotlib* module (Hunter 2007) to graph the histograms. At this stage, only the latter version saves the histograms generated during the clumping analyses and all pipelines unroot the input trees prior to analysis. The pipelines are available from the GitHub repository: `https://github.com/anaserrasilva/clumpy`.

4.4.1.1 Weighted Robinson-Foulds distance

Clumps, much like islands and families of trees, can be defined on any tree-to-tree distance. However, because we are dealing with trees with non-identical leaf sets care is required when interpreting and/or using uncorrected tree-to-supertree distances, since they actually correspond to the distance between each input tree and the subtree they induce on the supertree (Cotton & Wilkinson 2007). This leads to situations where the same RF distance between two trees, of different sizes, and the supertree may convey different levels of topological divergence (Fig.4.1). For example, assuming binary trees, a maximally divergent tree-supertree pair with only four shared tips will have an RF of 2, that same RF in any tree-supertree pair with over four tips in common conveys the minimal non-zero topological divergence between trees. This means that the differences in size between input trees and the supertree are not taken into account in uncorrected tree-to-supertree distances.

Using the RF distance, due to its ease of implementation and interpretation, I propose a modification of the uncorrected RF distance (uRF) that accounts for the theoretical maximum RF for the supertree, I am calling this modification the weighted RF distance (wRF). The wRF is computed using

$$wRF = uRF \times \frac{splits\ in\ ST}{splits\ in\ ST_p \cap T},$$

where "splits in $ST_p \cap T$" is the intersection of splits in the pruned supertree ST and in tree T. This calculation is implemented in the supetree-based clump extraction pipelines (see next subsubsection and Fig.4.3) with ETE's v.3.1.1 (Huerta-Cepas et al. 2010) RF calculator, where the uRF is calculated directly and the wRF calculation uses the edge inclusion/exclusion information outputted by the RF calculator. However, other modifications to the RF distance, specifically for trees on non-identical leaf sets, are available (e.g., Llabrés et al. 2021, Tahiri et al. 2022).

4.4.1.2 Clump extraction algorithm

Given a tree distribution and bin width
While the distribution is non-empty:
Step 1:

Build supertree of the tree distribution

Step 2:

For each tree in distribution:

Unroot tree

Compute distance to supertree (trees pruned to shared leaves)

Plot histogram of distances to supertree and choose distance threshold

Step 3:

Create a vector with the tree distribution's length and all values set to property a

(Indices in the tree distribution and property vector have 1:1 correspondence)

Set the first instance of a to b

Find the trees within x-D of first tree and change them to property b in vector

For all but the first tree:

Find the trees within x-D of trees with property b:

Set corresponding vector indices to b

Remove all trees with property b from tree distribution and write to clump file

Step 3 of this pipeline (see Fig.4.3 for a graphical depiction of the pipeline) is a modification of the algorithm used to extract tree islands available through the *islandNeighbours* R package (Serra Silva & Wilkinson 2021, see also section 2.4 in this thesis). As noted above, the pipeline accepts any Newick-formatted tree file but unroots trees prior to analysis. In addition to the RF distance parameter, users can also set their desired histogram bin width (although any integer is accepted, even numbers are recommended when using RF distances).

4.4.2 Clump extractions

Before delving into the biological patterns identified by the clumping of each dataset, it is worth noting that, for the majority of the analyses, as the number of extracted clumps increases so does the dissimilarity between the trees in the clumps and the supertree of all input trees (ST_{all}), with many of the clump supertrees displaying complete breakdown of ingroup monophyly (from species to class). Also, while results for the uRF analyses will not be shown, they confirmed my conjecture that, under an uncorrected tree-to-supertree distance, all quartet trees are placed into the first clump (not the case with the wRF). I did not, however, explore up to what size tree this remains true. A breakdown of the number of clumps extracted from each tree distribution, under wRF and multiple bin widths, can be found in table 4.2 and all the files resulting from the analyses are available in the Figshare Repository `https://doi.org/10.6084/m9.figshare.c.6050033`. Lastly, while the pipelines were designed to allow the users to choose the threshold, and thus which of the suggested approaches to clump identification to use (Fig.4.2a), unless stated otherwise, all results below correspond to the strict trough approach with bin width of two.

Starting with the rodent dataset, before clump supertrees start displaying extensive non-monophyly (*c.* clump 14) the differences in topology between the clumps are mostly restricted to the taxa in clade D (see figure 1 in Percequillo et al. (2021)), particularly *Cerradomys*, *Lundomys* and *Sooretamys*, which are also the taxa driving the supertree/supermatrix incongruence reported by Percequillo et al. (2021). While clades A and B display some instability, it is the taxa in clade C

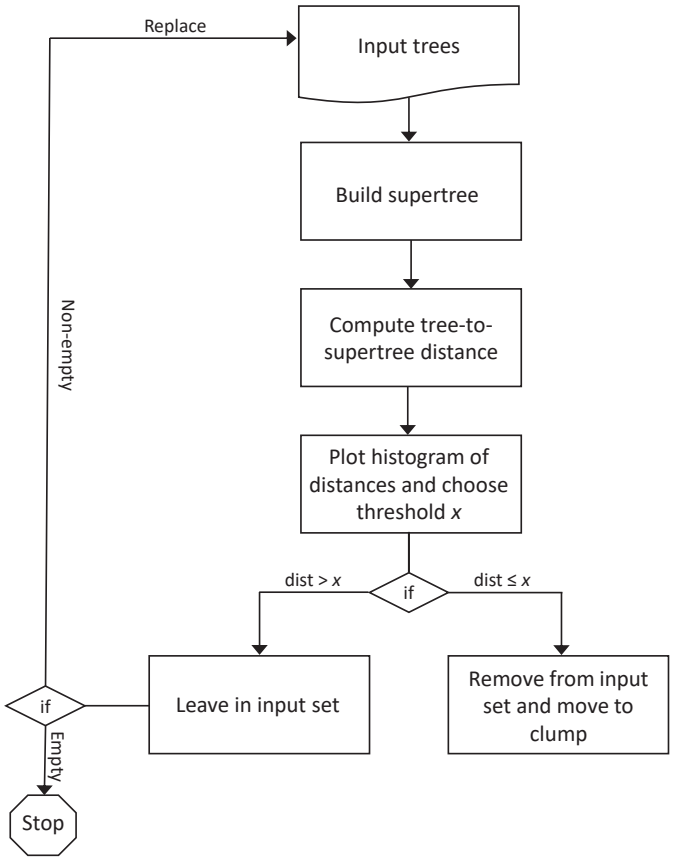


FIGURE 4.4. Clumping pipeline. Graphical depiction of the algorithm in subsubsection 4.4.1.2.

that are the most unstable after those in clade D, with at least three clump supertrees displaying C paraphyletic with D. From a clump identification standpoint, this is an interesting dataset since the first clump is composed of a single tree that is not topologically identical to ST_{all}, yet the supertree of clump-2 is identical to ST_{all}. Moreover, this pattern was found both with the uRF and wRF analyses. Thus, while the first clump contains those trees that are the most similar to ST_{all}, the clump summary that best matches ST_{all} may not correspond to that of the first clump. When bin width is increased, very little information can be gleaned from the outputted clumps. With a bin width of four, two clumps are still recovered that show clade D's instability in comparison to the other oryzomyine clades. But, at higher bin width values, apart from a single clump whose supertree matches ST_{all}, only clumps exhibiting a combination of rooting and non-monophyly of all/most Oryzomyini clades are identified. As such, for this dataset the most informative bin width setting is bin=2.

Unlike the previous dataset, none of the analyses on the Hime et al. (2021) data yielded a clump with its supertree identical to ST_{all}. For the default analysis (bin width=2), the summary trees most similar to ST_{all} correspond to those of clumps five and eight, yet their distance to

TABLE 4.2. Number of clumps identified for each dataset under the strict trough approach using the wRF distance. The *bin#* column names refer to the histogram bin width used for each analysis.

Dataset	bin2	bin4	bin6	bin8	bin10	bin12	bin14
Percequillo et al. (2021)	26	12	9	7	7	5	1
Hime et al. (2021)	41	21	15	13	6	6	11
Streicher et al. (2020)	20	3	1	1	1	1	1
Streicher & Wiens (2017)	10	2	1	1	1	1	1
KST2656	38	5	2	2	1	1	1
KST2019	52	15	4	2	1	1	1
KST768	24	2	2	1	1	1	1
KST348	22	7	1	1	1	1	1
KST2656-KST2019	33	3	1	1	1	1	1
KST2656-KST768	69	12	4	4	1	1	1
KST2656-KST348	20	6	2	2	1	1	1
KST2019-KST348	59	15	17	6	1	1	1
Chan all	18	5	2	1	1	1	1
Chan legacy	13	9	3	7	3	2	2
Chan exon unf	10	4	2	1	1	1	1
Chan intron unf	11	2	2	2	1	1	1
Chan UCE unf	15	13	3	1	1	1	1
Chan exon 75tax	14	2	1	1	1	1	1
Chan intron 75tax	13	1	1	1	1	1	1
Chan UCE 75tax	12	3	2	1	1	1	1
Chan exon 50pars	14	3	1	1	1	1	1
Chan intron 50pars	7	1	1	1	1	1	1
Chan UCE 50pars	14	2	1	1	1	1	1

the supertree is of 40 RF. In the analysis with bin width of 4, the minimum distance between the supertrees of the clumps and ST_{all} is 12 RF and corresponds to clump-2. The analyses with the lowest distance between a clump supertree and ST_{all} are those under bin widths 12 and 14, where the supertree of the first clump is 4 RF away from ST_{all}, with the difference between the supertrees revolving around the relationship between Nyctibatrachidae and Ceratobatrachidae. Another major difference between the analyses with bin widths two and four is the nearly halved number of identified clumps when the histogram bin width is doubled, see table 4.2. Interestingly, when the clumps with outgroup rooting, extensive ingroup non-monophyly and taxon sampling issues are ignored the default analysis yields a breakdown of clumps supporting Batrachia,

Procera or Acauda similar to that reported by Hime et al. (2021), though the proportions are not an exact match to Hime et al.'s 2021 analysis. They identified a proportion of trees supporting Batrachia:Procera:Acauda of approximately 2:1:1, and the default clumping analysis yielded a proportion of 3:1:1. I should note that this discrepancy in the proportitions might be partly due to the fact that I kept all gene trees for the clumping analyses, while Hime et al. (2021) filtered out all trees that could not be used to test the relationship between lissamphibian orders. In the bin width=4 analysis, ignoring clumps without major topological issues, no clumps support Acauda, a single clump supports Procera and nine Batrachia. In both the bin=2 and bin=4 analyses, Gymnophiona is the most internally stable amphibian order, the majority of trees that include Sirenidae and Cryptobranchoidea have the latter as sister to all other caudates, and most of the intraorder taxonomic instability is found within Anura, specifically Neobatrachia. At larger bin widths, the initial clumps all recover Batrachia, but any clump supporting Procera or Acauda has Amniota nested within Lissamphibia and *Latimeria* as sister to it. Unlike the rodent dataset, the non-monophyly of the ingroup (Lissamphibia) is mostly due to paraphyly with the outgroups, not to para- and polyphyly of the ingroup's lower taxa.

The two UCE-only datasets behaved quite differently under the default clump analyses, likely due to their distinct taxon sampling ranges. The Squamata dataset, which had a sampling range of c.48–92%, yielded 10 clumps, with the supertree of the fourth and fifth clumps being identical to ST_{all}. That two adjacent clumps have the same supertree suggests that they should be treated as a single clump and that the strict trough approach might overpartition tree distributions. The supertrees of clumps 1–3 do not have the full complement of taxa, and all three display different topologies within Lacertoidea. In fact, all clump supertrees (except four and five) display alternative topologies for Gekkota and/or Lacertoidea, consistent with the low branch support values reported by Streicher & Wiens (2017). Interestingly, unlike the other analyses reported here, only one clump supertree (clump-8) shows extensive non-monophyly of Squamata named groups. For the analyses with increased bin width, only the bin width=4 yielded more than a single clump, with the first clump having a supertree whose only difference from ST_{all} is the non-monophyly of Dibamidae and clump-2 showing alternate topologies for Gekkota, Lacertoidea and Anguimorpha, and having Dibamidae as sister to Toxicofera.

The microhylid dataset (sampling of 37.5%–76.5%) on the other hand, yielded 20 clumps, under the bin width=2 analysis, with RF=28 being the minimum distance between a clump supertree, with all 64 taxa, and ST_{all}. Most clump supertrees exhibit extensive non-monophyly at the subfamily level, which is likely the result of poor effective overlap within clumps, this is congruent with the large number of low support branches in the ST_{all} (figure 2.B in Streicher et al. 2020) and the supertree/supermatrix incongruence reported in the original paper. Given that no tree in this dataset exceeded a taxon sampling of 77%, the lowest maximum observed sampling of all analysed datasets, it is not unreasonable to think that this is the parameter causing the clumping pipeline's inability to identify 'major' topologies. The uRF analysis of this

dataset yielded similar results, but found only 9 clumps, a minimum distance to ST_{all} of 14, and, more importantly, a considerably lower proportion of unresolved clump summary trees (wRF=0.8, uRF≈0.4). The latter suggests that poor effective overlap within clumps is indeed the culprit of the patterns identified in the wRF analysis, since the numbers of unresolved clump supertrees greatly decreased when clump size increased (uRF). This is further cemented by the recovery of three large clumps (> 300 trees) whose summaries display primarily monophyletic microhylid subfamilies when the bin width is increased to four and a minimal distance to ST_{all} of 24 RF, although any further increase in bin width leads to the recovery of a single clump.

The KST datasets broadly follow the patterns reported by Siu-Ting et al. (2019) for their supertree analyses, with the majority of clump supertrees (that do not have extensive outgroup rooting, non-monophyly and/or taxon sampling issues) supporting the same hypothesis of amphibian relationships as the ST_{all} of each dataset. In other words, most clump supertrees from KST348 and KST2019 support Batrachia, and those for KST768 and KST2656 support Procera. Two things of note that can be found in all analyses of the KST datasets are that Anura is the most internally unstable amphibian clade, and that the supertree of clump-2 is either identical or the most similar to ST_{all}. For KST2019 and KST2656, the latter observation can be explained by 50% of the first clump consisting of quartet trees, which leads to increased topological instability. For KST348 and KST768, the reason is not as clear, though that the proportion of small trees (6-tip) is higher for clump-1 than for clump-2 may contribute to the lower resolution and topological changes compared to ST_{all}. The analyses of the dataset complements showed similar results to those of the original datasets, but it is worth noting that for KST2019-348 the number of clumps supporting Procera just edges out Batrachia, with both KST2019 and KST348 supporting Batrachia. As for the non-default bin width analyses, they revealed a pattern similar to Hime et al.'s (2021), but from a bin width of 10 all datasets recovered a single clump (table 4.2).

The Chan et al. (2020) datasets display multiple trends detailed above, starting with the first clump's supertree not corresponding to the ST_{all}, due to the clump consisting of small trees with poor effective overlap or to it including very few trees (or both, unfiltered UCE dataset). The latter is particularly well illustrated by the 75% taxon sampling datasets, where for the non-UCE datasets the supertree of the first clump is also that of all trees, but for the UCE dataset, whose first clump contains only three trees, it is the second clump's supertree that matches ST_{all}. Most analyses recover the majority of clump supertrees exhibiting extensive non-monophyly issues concentrated toward the tail end of the identified clumps, except for the small tree-rich first clumps. Across all analyses, and prior to breakdown of generic monophyly, instability is concentrated on the three nodes identified by Chan et al. (2020) as the foci of topological instability. However, the topologies of the clump supertrees are not restricted to the five alternatives depicted in their figure 1, rather most are intermediate topologies (for example, one tree with N1 from T4 and N3 from T1). But, apart from the Sanger dataset, at least one clump supertree matching one of Chan et al.'s (2020) observed topologies is identified before

intermediate topologies. Lastly, while all non-Sanger datasets behave similarly, with the initial wRF histograms showing nearly continuous uni- or bimodal distributions with small peripheral peaks, the histogram for the Sanger dataset shows multiple small discontinuous peaks. This is due to the very small number of trees (30) in the Sanger dataset, the wide range of possible RF values (0–94) and the topological incongruence between trees, it is, thus, improbable that it could yield a continuous and/or tight distribution of wRF values. As such, we might expect this pattern to be repeated in other instances where clumps are used to cluster small tree sets. Similar patterns are found in the analyses with bin widths greater than four that do not immediately drop to a single clump.

Overall, the results show that clumping can extract some of the major topologies found in (multi)sets of trees, but the effects of a set's maximum tree size on the analyses need to be explored further. Additionally, (multi)set size is of greater import to the clustering analyses than the methodological sources of the data underpinning the tree sets. But perhaps more importantly, clump informativeness can be manipulated by varying the histogram bin width, with some datasets benefiting from wider bins, like the microhylid trees, whereas others benefit from narrow bins, much like the Oryzomyini tree set.

4.5 Special cases

In the section describing clumps of trees, I introduced a number of special cases that are expected to occur in clumping analyses. The Pardo et al. (2017) MPTs allow us to explore two of those, the case where most supertrees correspond to input trees and the use of consensus trees. The latter also helps to illustrate how changing the supertree method may change the clumps identified. Additionally, because the tree-to-supertree distance histogram for all MPTs shows a distribution made of two connected peaks and a third disconnected peak, this dataset is ideal to illustrate the differences between using the strict trough and break approaches to identify clumps. Under the break approach the *clumpy* pipeline that computes the supertrees using Astral identifies four clumps that correspond to the four 12-RF islands reported in Serra Silva & Wilkinson (2021). The trough analysis, however, identified the three smallest 10-RF islands as clumps, but the two largest clumps do not map onto the two largest 10-RF islands. Thus, while providing information similar to that of islands, regarding major hypotheses for evolutionary histories, clumps of trees on the same leaf set are not always identical to tree islands. However, as seen in Chapter 3, with large tree distributions clump identification is less memory intensive than island extraction analyses.

4.5.1 Using a majority-rule consensus as the supertree

Using the majority-rule consensus (MRC, Margush & McMorris 1981) as the supertree, computed with DendroPy v.4.5.2 (Sukumaran & Holder 2010), one immediate change between

the MRC and Astral results is the right-displacement of the distance distributions. This is due to the MRC not being one of the input trees, unlike most of the summary trees computed by Astral, thus increasing the minimum distance between the 'supertree' and all input trees. As for what clumps are identified, under the break approach the four 12-RF islands are recovered once more, though the distance thresholds differ between the Astral and MRC analyses. The trough analysis yields five clumps, with the three smallest corresponding to the smallest 10-RF islands, but the two largest clumps do not match the two largest 10-RF islands, much like in the supertree case above. Additionally, the make-up of the two largest clumps is distinct from those obtained in the Astral+trough analysis. Thus, by changing the supertree method we can change the number and size of the clumps that are identified from the same dataset.

4.6 Discussion

There is a long history of post-processing (multi)sets of phylogenetic trees to find the most informative summary for these sets by clustering and/or building consensus trees (e.g., Bonnard et al. 2006, Hendy et al. 1988, Maddison 1991, Serra Silva & Wilkinson 2021, Stockham et al. 2002). Many post-processing approaches focus on sets of trees with completely overlapping leaf sets, but with the increased use of phylogenomic datasets comes the need for tools that can be applied to sets of trees with non-identical leaf sets. First, since supertrees are an extension of consensus methods (Gordon 1986), in the context of trees with non-identical leaf sets we may consider them to fulfil the same role as consensus methods do when dealing with sets of trees on the same leaf set. A non-supertree 'consensus' method that may at first appear enticing and would allow for the identification of a set of relationships/taxa present in all input trees is computing the maximum agreement subtree (MAST, Amir & Keselman 1997, Finden & Gordon 1985). However, as the number of input trees and/or polytomies increases, the complexity of computing the MAST for a set of trees becomes NP-hard (reviewed in Deepak et al. 2014), making this approach suboptimal when dealing with large tree distributions, which many phylogenomic datasets are. This makes clustering approaches the better option to post-process sets of trees with non-identical leaf sets.

However, clustering methods are not without challenges. While some of the main considerations for methods to cluster (multi)sets of trees on the same leaf set might entail rootedness and the presence of duplicate trees, for trees with non-identical leaf sets it is also necessary to take the presence of small trees (triplets/quartets), and of trees with little to no taxonomic overlap into account.

The presence of small trees is an important consideration seeing as two trees with highly disparate topologies, supporting distinct evolutionary histories (e.g., T_1 supports Batrachia and T_2 supports Acauda), might still display shared quartet/triplet trees. In island-like clustering methods, these quartets/triplets have the capacity to cluster highly disparate trees into the same

subset, since the small distances between the quartet/triplet and the larger trees artificially lower the 'connectedness' threshold between the large trees (Fig.4.1). By implementing a sequential clustering approach, we need not worry about quartet trees affecting the clustering process, since, when using uncorrected distances, they were always placed in the first cluster(s) to be identified. However, further work is required to check up to what size tree we can expect to find the same pattern exhibited by the quartet trees, and whether size filtering prior to clustering changes the number of clumps identified. If using a corrected distance, like the wRF, quartet trees will be more widely distributed across clumps, provided more than one clump is identified, but each clump will only contain those quartet trees that perfectly match their input supertrees.

An interesting phenomenon that may occur with clumping is the identification of first clumps whose summary tree differs from the ST_{all}, followed by a second clump whose summary tree is identical to the ST_{all}. From the trough-based analyses I identified two distinct scenarios for the occurrence of this phenomenon. One scenario consisted of a small clump followed by a break and a large second clump, and the other scenario of a large first clump followed by a trough and a second large clump. The former was identified from the rodents dataset, where the first clump consisted of a single, large tree that was not identical to the ST_{all}, but the summary tree of the second, larger clump (90 trees) did match the ST_{all}. The latter scenario was found in many of the KST datasets, where the summary tree of a large first clump was highly unresolved, yet the summary of the second clump matched the ST_{all}. In this case, the first clump consisted of many (100s–1000s) trees, which had considerably fewer tips than the ST_{all} with poor effective overlap between them resulting in a highly unresolved summary tree. For both scenarios, it may be sensible to collapse the first two clumps into a single clump, and, more broadly, it may be sensible to collapse any clumps whose input supertrees are identical.

With the exception of the microhylid dataset, under a bin width of two all analyses yielded a collection of clumps from which biological information can be gleaned. For example, from the KST and Hime et al. (2021) datasets it was possible to identify clumps that support the Batrachia, Procera or Acauda hypotheses of amphibian relationships. However, in all analyses a large number of clumps exhibits outgroup non-monophyly, which hinders the identification/interpretation of the summary trees' and the biological information they display, whether it be at the order or genus level, and renders the clumps uninformative. This pitfall might be avoided either by filtering out those unrooted trees without monophyletic outgroups, similar to the filtering used by Hime et al. (2021), or by using ingroup-only trees as input and rooting the clump summary trees with an *a posteriori* method (e.g., MAD, Tria et al. 2017).

The microhylid dataset presents its own set of pitfalls. While not unexpected, given the incongruence between the topologies recovered by the supermatrix and supertree analyses reported by Streicher et al. (2020), the level of dissimilarity between the summary tree of each clump and both the supertree of all trees and the supertree of each complement ($\mathscr{S}^C$) cannot be explained entirely by the presence of small trees, nor by outgroup non-monophyly. Rather, that

no tree in the set has a taxon sampling of over 77% suggests that there is extensive non-effective overlap between the trees in each clump. Especially, given that when bin width is increased, fewer and larger clumps are identified, with resolved and informative clump supertrees. Thus, while more testing will be required, it appears that, though clumping does not require datasets where at least some trees have 100% taxon sampling, it might have an optimal lower bound for the maximum observed taxon sampling, and that choice of bin width is an important analytical parameter. Additionally, the choice of supertree method may also be a consideration, since, as shown by the analyses on the Pardo et al. (2017) MPTs, different supertrees can yield different clump assortments. It is thus possible that if a different supertree method had been used, clump summaries might have imparted more information on alternate microhylid topologies.

While I did not present results, I also tested a fixed threshold clumping method (exclusively under a bin width of two), but the clumps were not informative. While the fixed threshold approach identifies the minimum threshold required to find a clump and the maximum threshold before all trees are placed in a single clump, the clumps themselves are not informative. This is due to the method relying solely on the chosen threshold value, thus yielding clumps that either do not capture a whole peak, capture multiple discontinuous peaks, or a combination of these. Also, the information this method does provide can also be extracted from the histograms plotted by the variable threshold approach(es). However, if a fixed threshold approach is desired the sequence of sets $\mathscr{S}$, from the clump definition given in section 4.2, can be redefined such that $\mathscr{S}_n = \{d(T, \mathscr{T}_{ST}) \leq x \text{ and } \mathscr{T} = \mathscr{S}^C\}$. Beyond the drawbacks above, the use of a single threshold to identify all clumps will also lead to instances where the set of remaining trees are unpartitionable, and will thus be unclumpable under x-D. Although, an intermediate approach between fixed threshold and histogram-based threshold selection might eschew the pitfalls of both threshold setting approaches (the possibility of trees not being placed into clumps and the need for constant user input, respectively). Setting a minimum threshold that can be dynamically increased to the minimum tree-to-supertree distance found for each clump extraction iteration would ensure that all trees can be assigned to a clump. However, an extra step would be necessary to merge sufficiently similar clumps, just as was suggested above for successive clumps whose input set supertrees are identical.

Despite the myriad pitfalls described above, variable threshold clumping can and does yield relevant and useful information of the distinct evolutionary histories encoded in (multi)sets of trees. While the choices of supertree method, partitioning strategy and histogram bin width can influence the clumps identified, this new partitioning strategy has the benefit of not requiring data alignments nor phylograms (e.g., Smith et al. 2020), meaning that it can be applied to any (multi)set of trees. Additionally, by relying on a tree-to-supertree distance the tree order in the input file will not affect the analyses, unlike some of the currently available methods for post-processing trees with non-identical leaf sets (e.g., Smith et al. 2013). However, further work is required to identify the optimal minimum observed taxon sampling and outgroup monophyly

parameters for the method.

While I used a modified RF distance for the analyses, there is a push to develop tree-to-tree distances specifically designed to deal with trees with non-identical leaf sets (e.g., leaf removal, Chauve et al. (2017); completion based RF, Bansal (2020)), which may either lead to other partitioning strategies or to a refinement of the definition of clumps based on the novel tree distance measures, which may in turn make the choice of partitioning scheme and/or histogram bin widths unnecessary. Recently, Tahiri et al. (2022) extended the single *vs*. multiple consensus debate to the supertree context with their proposed extension of k-means clustering to a pruning-based normalised RF distance. However, despite highly promising results, their method is unable to address the issue of distances between trees with two or fewer tips in common. While k-means are less robust than other clustering approaches like k-medoids (Arora et al. 2016), Tahiri et al.'s (2022) work is a huge step in the post-processing of sets of trees with non-identical leaf sets, and may soon lead to a proliferation of generalised clustering approaches and other methods to rival the breadth of tools available for post-processing of trees on the same leaf set.

Lastly, with small alterations to the pipeline using the appropriate supertree and/or consensus software and a corresponding distance measure, the existing clump extracting pipelines can be extended to deal with (multi)sets of rooted trees and of internally labelled trees. The latter might consist of phylogenetic trees with branch labels corresponding to higher taxonomy, thus allowing for trees at distinct taxonomic levels to be compared (e.g., MultiLevelSupertree, Berry et al. 2012), or they might consist of the node-labelled mutation trees commonly used in cancer phylogenetics (e.g., Aguse et al. 2019).

4.7 Data and software availability

The data, scripts and results for this chapter are available on the following repository:

- Figshare, `https://doi.org/10.6084/m9.figshare.c.6050033`.

The *clumpy* Python pipeline is available from:

- GitHub, `https://github.com/anaserrasilva/clumpy`.

CHAPTER 5

DEALING WITH NON-EFFECTIVE OVERLAP IN LARGE-SCALE PHYLOGENETICS

An earlier version of the alignments used in this chapter, restricted to nucleotide data available in GenBank (Benson et al. 2008) on June 27^{th}, 2020, was used as the basis for the Gymnophiona phylogenetic trees in a masters thesis co-supervised by ASSilva:

Battye, S. (2020), Biogeography, life history and conservation of caecilians. MSc Thesis, School of Sciences, University of Wolverhampton, UK.

5.1 Introduction

As systematists continue their efforts to infer the Tree(s) of Life (e.g., Hime et al. 2021, Frost et al. 2006, Shen et al. 2013, Hellemans et al. 2022, Simões & Pyron 2021), they have been beset by the continued identification of recalcitrant nodes (e.g., Cai et al. 2021, Kulkarni et al. 2021, Wurdack & Davis 2009). This taxonomic instability can be caused by data incongruence (e.g., Cannatella et al. 1998, Chan et al. 2020, Hahn 2007), missing data (e.g., Simmons & Goloboff 2014, Roure et al. 2012), and/or analytical artefacts (e.g., Léveillé-Bourret et al. 2017, Simmons et al. 2022). Approaches to summarise the tree (multi)set heterogeneity that often accompanies these recalcitrant nodes have been explored in previous chapters (e.g., islands of trees (Chapter 2, Maddison 1991, Serra Silva & Wilkinson 2021), clumps of trees (Chapter 4)), but methods targeting the causes behind taxonomic instability are also widely available. Multiple tools have been developed to identify and mitigate the effects of data incongruence due to reticulate events—such as the ABBA-BABA test (Patterson et al. 2012), multispecies coalescent co-estimation and summary methods (Bouckaert et al. 2014, Zhang et al. 2018), and phylogenetic networks (Huson et al. 2010, see Chapter 3 for additional references)—and, as briefly alluded to in Chapter 3,

so too have approaches to deal with the effects of missing data (e.g., safe taxonomic reduction (STR, Wilkinson 1995*b*); gene sampling sufficiency (Steel 2016)), which I will be focusing on here. The amount and (non-random) distribution of missing data have both been implicated in taxonomic instability (e.g., Roure et al. 2012, Simmons 2012, Wiens & Morrill 2011), and it has often been suggested that mitigating the negative effects of missing data is as simple as adding more data to the matrix, particularly loci or characters (e.g., Wiens 2003, this thesis: Chapter 3, subsection 3.4.3). However, untargeted increase of taxa and/or loci sampling may not be feasible (e.g., financial concerns, access to rare and/or extinct taxa), and methods that favour targeted sampling might prove more efficient solutions.

A mathematical concept that has garnered renewed attention in the last decade, particularly as it relates to missing data in large-scale phylogenetic analyses, is that of phylogenetically decisive taxon coverage (Sanderson et al. 2010, Steel 2016, Steel & Sanderson 2010), also referred to as effective taxonomic overlap (Wilkinson & Cotton 2006). Where, given a tree T and a taxon coverage pattern S, S is decisive for T if T is the unique supertree of the subtrees induced by S (Dobrin et al. 2018, Sanderson et al. 2010, Steel & Sanderson 2010), figure 5.1. In other words, if the subtrees induced by S are ((A,B),C) and ((A,D),B), S is decisive because T=((A,D),B),C), but not if the subtrees are ((A,B),C) and ((A,B),D) since T={(((A,B),C),D); ((A,B),(C,D)); (((A,B),D),C)}. Thus, decisive taxonomic coverage reflects the pattern, not amount of overlap. This concept informs methods that seek to minimise the negative effects of missing data on tree inference, both on the taxon (e.g., STR and its heuristic extension: Concatabominations, Siu-Ting et al. 2015) and locus sampling fronts (e.g., gene sampling sufficiency).

For partitioned matrices, the (non-)effective overlap across a matrix can also be identified by testing for the presence of tree terraces. These are areas of tree space where trees are all topologically different, yet have exactly the same optimality score and display the same set of subtrees (Sanderson et al. 2011, 2015). These terraces are (sub)sets of 1-NNI islands of trees (Maddison 1991, Serra Silva & Wilkinson 2021), and like them can negatively influence tree search and lead to topological uncertainty (Sanderson et al. 2011, 2015). Tree terraces arise when data partitions (individual characters in parsimony analyses) induce subtrees that do not have a unique parent tree (supertree), meaning that the matrix's taxon coverage is not decisive. Additionally, for terraces to form in maximum likelihood analyses, branch length optimisation must be independent between partitions (edge-unlinked models), and they never form during Bayesian inference analyses (Sanderson et al. 2015). However, because branch length optimisation models are seldom reported, recent explorations of the presence/prevalence of terraces in published datasets tested for the terraces that would have, theoretically, arisen under maximum parsimony or edge-unlinked maximum likelihood analyses (discussed in Dobrin et al. 2018). Thus, recovering the worst case scenarios of taxonomic instability driven by non-effective overlap.

Here, I will be comparing two methods that rely on effective overlap to inform decisions on

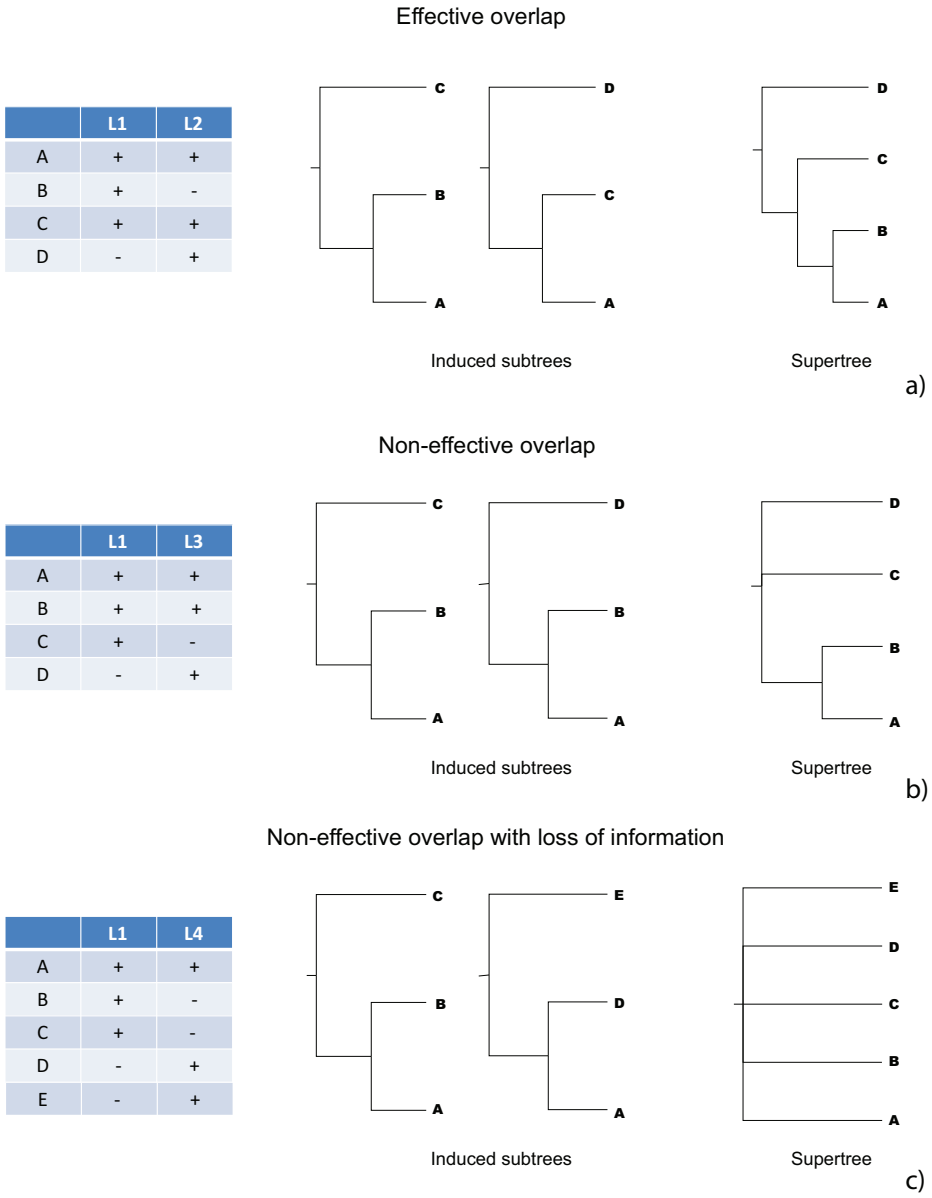


FIGURE 5.1. Patterns of overlap for three sets of subtrees (or data partitions in the supermatrix case). In a) the pattern of overlap is effective, i.e. the taxon coverage is decisive, because the two subtrees induced encode a unique supertree, while in b) the induced subtrees encode three unique supertrees, illustrating non-effective overlap (non-decisive taxon coverage). In c) extreme non-effective overlap with complete loss of topological information is illustrated.

how data should be added to a matrix to minimise taxonomic instability, one focused on taxa and another on loci. I will use a caecilian phylogeny to test whether, with a jackknife approach, Concatabominations—originally developed for morphological datasets but successfully applied to matrix representations (MR) of supertrees (e.g., Akanni et al. 2015, Cardillo et al. 2004)—can be used to identify target loci for increased taxonomic sampling, not just which taxa are unstable. I

will then compare this jackknife approach to the gene sampling sufficiency method, using both the caecilian phylogeny and a selection of the datasets used by Dobrin et al. (2018) to explore the prevalence of terraces in multilocus datamatrices.

5.2 Materials and methods

5.2.1 Gymnophiona

Gymnophiona is the least speciose of the three amphibian orders, with 214 currently valid species in 10 families (AmphibiaWeb 2019, Frost 2022). Caecilians are easily distinguished from frogs and salamanders due to their elongated, annulated and limbless morphology (San Mauro et al. 2014). Caecilians are found primarily in tropical and subtropical environments, with most species being fossorial, except for the secondarily aquatic Typhlonectidae (Kamei et al. 2012, San Mauro et al. 2014). Because of their geographic distribution and habitat(s), caecilians are not easily found or collected, leading to many species being known from a single specimen (e.g., *Amozops amozops* Wilkinson et al. 2021). This means that caecilians are a prime taxon to test targeted taxon/locus sampling. In fact, they have been previously used to explore an information-based approach (Goldman 1998, Goldman et al. 2000) that identifies both the most informative loci and the areas of the tree that would benefit the most from increased taxon sampling (San Mauro et al. 2009, 2012).

5.2.1.1 Data collection and processing

All caecilian nucleotide data available in GenBank (Benson et al. 2008) on March 14^{th}, 2022 were downloaded. To filter synonyms from the downloaded data, the taxonomic information provided by Genbank was checked using the Amphibian Species of the World v.6.0 database (Frost 2022). Unless included in previous comprehensive phylogenetic analyses (e.g., San Mauro et al. 2014), sequences identified to the species level that included *aff.*, *cf.*, or *sp.* in their updated taxonomy were discarded, as were loci with data for fewer than three caecilian species (e.g., tyrosinase). This resulted in a dataset made up of 96 caecilian taxa, 88 currently valid, and 24 loci. These consisted of 15 mitochondrial (12S rRNA, 16S rRNA, ATP6, ATP8, COX1, COX2, COX3, CYTB, ND1, ND2, ND3, ND4, ND4L, ND5, ND6) and 9 nuclear loci (18S rRNA, 28S rRNA, BDNF, CXCR4, H3A, NCX1, RAG1, SIA, SLC8A3).

From the filtered sequence data, only one sequence was selected per species per locus, except for *Gymnopis multiplicata*, as the most recent comprehensive Gymnophiona-focused phylogenetic study found this species to be paraphyletic with *Dermophis mexicanus* (San Mauro et al. 2014). The criteria for within species sequence selection were, in order: i) one voucher specimen - multiple loci, ii) sequence length, iii) most recent sequence, and iv) valid taxon (not synonym) in sequence description. Data was also downloaded for the following outgroups: the sarcopterigians *Latimeria chalumnae* and *Protopterus annectens*; the amniotes *Mus musculus*, *Anolis carolinensis* and

Gallus gallus; the caudates *Lyciasalamandra atifi*, *Andrias davidianus* and *Ranodon sibiricus*; and the anurans *Leiopelma archeyi*, *Bombina orientalis* and *Xenopus laevis*. For GenBank accession numbers, see table 5.1 for nuclear loci (nucDNA), and table 5.2 for mitogenomes and mitochondrial loci (mtDNA).

All loci were aligned using the MAFFT v.7.450 (Katoh & Standley 2013) plugin for Geneious Prime v.2022.0.2 (`https://www.geneious.com`, Kearse et al. 2012), under automatic algorithm selection. To remove hyper-variable regions and trim ragged ends, all multiple sequence alignments were all run through Gblocks v.0.91b (Castresana 2002) under the settings: "minimum number of sequences for a flank position" = $\frac{n}{2}+1$, where n is the total number of sequences in the alignment, "maximum number of contiguous non-conserved positions" = 10, "minimum length of a block" = 5, and "allowed gap positions" = "with half". This resulted in the following alignment sizes: 757 base pairs (bp) for 12S rRNA, 1229bp for 16S rRNA, 684bp for ATP6, 157bp for ATP8, 1540bp for COX1, 688bp for COX2, 784bp for COX3, 1141bp for CYTB, 963bp for ND1, 1035bp for ND2, 343bp for ND3, 1368bp for ND4, 297bp for ND4L, 1781bp for ND5, 492bp for ND6, 1775bp for 18S rRNA, 697bp for 28S rRNA, 713bp for BDNF, 644bp for CXCR4, 328bp for H3A, 1304bp for NCX1, 1509bp for RAG1, 432bp for SIA and 1121bp for SLC8A3. The taxon coverage density (i.e. the average of the per locus taxon coverage, see equation in 5.2.2) for this dataset was 0.47, with 16S rRNA having the highest (0.94) and BDNF and SIA the lowest (0.12) proportions of taxonomic coverage.

5.2.1.2 Phylogenetic analyses

Nucleotide substitution models were fitted using jModelTest v.2.1.7 (Darriba et al. 2012), with number of substitution models set to 3. The best-fit models were HKY+I+G for CXCR4, GTR+I for SIA, GTR+G for 18S rRNA and H3A, and GTR+I+G for all other loci. Gene trees were inferred under MrBayes v.3.2.6 (Ronquist et al. 2012), the analyses were set to two independent runs with four chains and 10 million generations, sampled every 1000 generations with a relative burn-in of 25%, and summarised with the majority-rule consensus (MRC, Margush & McMorris 1981). Run convergence was checked using an average standard deviation of split frequencies (ASDSF) ≤ 0.01, and a potential scale reduction factor (PSRF) of 1.00, if either measure was not met the analyses were run for an additional 10 million generations. A supermatrix analysis was also run under the same settings as the gene tree analyses, with the genes set as partitions. Codon positions were not taken into account due to the use of Gblocks.

The supertree was obtained by using the MRCs of the inferred gene trees as input to the quartet-based summary method Astral-III v.5.6.3 (Zhang et al. 2018). The MRC trees were chosen over the highest posterior probability tree in each gene tree analysis to ensure that the supertree recovered a conservative estimate of caecilian relationships and, because the primary aim of this chapter is to compare two methods whose aim is to identify and minimise non-effective overlap between loci, increase the chances of recovering a partially resolved supertree. If the aim of the

TABLE 5.1. List of taxa used in the caecilian phylogenetic analyses, and GenBank accessions for nuclear loci. See table 5.2 for mitogenomes and mitochondrial loci (mt).

Taxon	**mt**	**18S**	**28S**	**BDNF**	**CXCR4**	**H3A**	**NCX1**	**RAG1**	**SIA**	**SLC8A3**
Atretochoana eiselti	Y				KX757097	KX757110		KX757122	KX757125	KX757134
Boulengerula boulengeri	Y				EF107484		EF107263	EF107322		EF107425
Boulengerula cf. *boulengeri*	Y									
Boulengerula changamwensis	Y							FR691680		
Boulengerula fischeri	Y							FR691676		
Boulengerula niedeni	Y							FR691679		
Boulengerula taitana	Y						HQ444140	DQ320062		
Boulengerula uluguruensis	Y		DQ283488			DQ284138		FR691681	DQ282670	
Brasilotyphlus dubium	Y									
Brasilotyphlus guarantanus	Y									
Caecilia gracilis	Y				KX757102	KX757108		KX757117	KX757126	KX757142
Caecilia isthmica	Y									
Caecilia pulchraserrana	Y									
Caecilia tentaculata	Y		DQ283717		KX757096	DQ284387		KX757118	KX757130	KX757140
Caecilia thompsoni	Y									
Caecilia volcani	Y						HQ444141	HQ444128		
Chikila fulleri	Y						HQ456773	HQ456772		
Chthonerpeton indistinctum	Y				KX757104	KX757113	EF107266	KX757115	KX757127	KX757138
Chthonerpeton viviparum	Y									

Continued on the next page

TABLE 5.1. *(cont.)* List of taxa used in the caecilian phylogenetic analyses, and GenBank accessions for nuclear loci. See table 5.2 for mitogenomes and mitochondrial loci (mt).

Taxon	**mt**	**18S**	**28S**	**BDNF**	**CXCR4**	**H3A**	**NCX1**	**RAG1**	**SIA**	**SLC8A3**
Crotaphatrema lamottei	Y							JN089397		
Crotaphatrema tchabalmbaboensis	Y		DQ283676			DQ284342				
Dermophis mexicanus	Y			JQ073153			EF107261	JQ073273		EF107423
Dermophis oaxacae	Y					DQ284428			DQ282897	
Dermophis parviceps	Y									
Epicrionops marmoratus	Y									
Epicrionops parkeri	Y									
Gegeneophis carnosus	Y							HQ444115		
Gegeneophis danieli	Y							HQ444121		
Gegeneophis goaensis	Y							HQ444119		
Gegeneophis krishni	Y							HQ444122		
Gegeneophis madhavai	Y							HQ444123		
Gegeneophis mhadeiensis	Y							HQ444118		
Gegeneophis orientalis	Y							KP400613		
Gegeneophis pareshi	Y							KP400612		
Gegeneophis ramaswamii	Y							AY456255		
Gegeneophis seshachari	Y							HQ444124		
Geotrypetes seraphini	Y		DQ283662		AY523683	DQ284328	AY523726	DQ320063		EF107361
Grandisonia alternans	Y	GRNRR18S					HQ444143	KP400617		
Grandisonia larvata	Y							KP400618		

Continued on the next page

TABLE 5.1. *(cont.)* List of taxa used in the caecilian phylogenetic analyses, and GenBank accessions for nuclear loci. See table 5.2 for mitogenomes and mitochondrial loci (mt).

Taxon	**mt**	**18S**	**28S**	**BDNF**	**CXCR4**	**H3A**	**NCX1**	**RAG1**	**SIA**	**SLC8A3**
Grandisonia sechellensis	Y							KP400619		
Gymnopis multiplicata 1	Y									
Gymnopis multiplicata 2	Y						HQ444144	HQ444132		
Herpele squalostoma	Y		DQ283679		EF107485	DQ284346	EF107264	FR691675		EF107426
Hypogeophis brevis	Y			MH114002				KP400616		
Hypogeophis montanus	Y			MH113997						
Hypogeophis pti	Y			MH114007						
Hypogeophis rostratus	Y	HYORR18S	DQ283524	MT569251	AY948787	DQ284172	AY948829	AY948930	DQ282687	AY948885
Ichthyophis asplenius	Y				EF107480		EF107257	EF107316		EF107419
Ichthyophis biangularis	Y									
Ichthyophis cardamomensis	Y									
Ichthyophis catlocensis	Y									
Ichthyophis chaloensis	Y									
Ichthyophis glutinosus	Y				AY948794		AY948839	AY456256		AY948901
Ichthyophis kohtaoensis	Y	ITYRR18S			EF107451		EF107225	HQ902530		EF107358
Ichthyophis larutensis	Y									
Ichthyophis longicephalus	Y									
Ichthyophis multicolor	Y									
Ichthyophis nguyenorum	Y									

Continued on the next page

TABLE 5.1. *(cont.)* List of taxa used in the caecilian phylogenetic analyses, and GenBank accessions for nuclear loci. See table 5.2 for mitogenomes and mitochondrial loci (mt).

Taxon	**mt**	**18S**	**28S**	**BDNF**	**CXCR4**	**H3A**	**NCX1**	**RAG1**	**SIA**	**SLC8A3**
Ichthyophis orthoplicatus	Y				EF107481		EF107258	EF107317		EF107420
Ichthyophis supachaii	Y									
Ichthyophis tricolor	Y									
Idiocranium cf. *russelli*	Y									
Indotyphlus battersbyi	Y							HQ444125		
Indotyphlus maharashtraensis	Y							HQ444126		
Luetkenotyphlus brasiliensis	Y				EF107483		EF107262	EF107321		EF107424
Luetkenotyphlus fredi	Y									
Luetkenotyphlus insulanus	Y									
Microcaecilia dermatophaga	Y									
Microcaecilia nicefori	Y									
Microcaecilia sp.	Y									
Microcaecilia unicolor	Y									
Oscaecilia ochrocephala	Y						HQ444146	HQ444137		
Potomotyphlus kaupii	Y				KX757098	KX757109		KX757121	KX757128	KX757135
Praslinia cooperi	Y				EF107489		EF107269	HQ444138		EF107431
Rhinatrema bivittatum	Y		DQ283700		EF107478	DQ284370	EF107255	AY456257	KX757132	EF107417
Rhinatrema gilbertogili	Y				MH177835	MH177838		MH177842		MH177844

Continued on the next page

TABLE 5.1. *(cont.)* List of taxa used in the caecilian phylogenetic analyses, and GenBank accessions for nuclear loci. See table 5.2 for mitogenomes and mitochondrial loci (mt).

Taxon	**mt**	**18S**	**28S**	**BDNF**	**CXCR4**	**H3A**	**NCX1**	**RAG1**	**SIA**	**SLC8A3**
Rhinatrema nigrum	Y						HQ444142	HQ444130		
Rhinatrema shiv	Y									
Rhinatrema uaiuai	Y				MH177834	MH177836		MH177841		MH177840
Schistometopum cf. *gregorii*	Y									
Schistometopum gregorii	Y		DQ283490			DQ284140				
Schistometopum thomense	Y				EF107488		EF107268	HQ444139		EF107430
Scolecomorphus kirkii	Y									
Scolecomorphus uluguruensis	Y				EF107486		EF107265	EF107324		EF107427
Scolecomorphus vittatus	Y		DQ283663		EF107457	DQ284329	EF107232	AY456258	DQ282816	EF107380
Siphonops annulatus	Y						HQ444147	DQ320064		
Siphonops hardyi	Y		DQ283489			DQ284139				
Siphonops paulensis	Y				EF107487	KX757114	EF107267	KX757119	KX757131	EF107429
Typhlonectes compressicauda	Y				KX757099	KX757111		KX757123	KX757129	KX757136
Typhlonectes natans	Y	TYPRR18S	DQ283486	JQ073154	KX757100	DQ284136	EF107229	EF551566		EF107365
Uraeotyphlus bombayensis	Y									
Uraeotyphlus cf. *oxyurus*	Y							AY456259		
Uraeotyphlus gansi	Y									

Continued on the next page

TABLE 5.1. *(cont.)* List of taxa used in the caecilian phylogenetic analyses, and GenBank accessions for nuclear loci. SIA omitted due to space, none of the species below had sequences available for it. See table 5.2 for mitogenomes and mitochondrial loci (mt).

Taxon	**mt**	**18S**	**28S**	**BDNF**	**CXCR4**	**H3A**	**NCX1**	**RAG1**	**SLC8A3**
Uraeotyphlus narayani	Y		DQ283491			DQ284141		DQ282671	
Outgroups									
Gallus gallus	Y	KT445934	KT445934	NM_001398061	NM_204617		DQ987923	NM_001031188	NM_001293097
Anolis carolinensis	Y	AY859624	AY859623	EU402616	JN702390		GU456076	FJ356739	JF804215
Mus musculus	Y	7CPU_S2	NR_003279		BC031665	BC115816	AF004666	BC138342	AF453257
Xenopus laevis	Y	X04025	X59734	EF035623	Y17895		X90839	XELRAG1X	
Bombina orientalis	Y		DQ283741	EF453367	AY364177	HM998943	AY523715	AY583335	AY948867
Leiopelma archeyi	Y		DQ283588		AY523700	HM998942	HM998951	HM998973	EF107408
Lyciasalamandra atifi	Y			KF645905	KF645581		KF645550	KY342294	KF645521
Ranodon sibiricus	Y	AJ279506		HM037760				HM037735	
Andrias davidianus	Y			EU275889	AY948801		AY948847	AY650142	AY948911
Latimeria chalumnae	Y	LTIRREX	LCU34336			DQ284319			
Protopterus annectens	Y				MH329958			KF911897	

TABLE 5.2. List of taxa used in the caecilian phylogenetic analyses, and GenBank accessions for mitogenomes and mitochondrial loci (mtDNA). See table 5.1 for nuclear loci (nucDNA).

Taxon	**nucDNA**	**mitogenome**	**12S**	**16S**	**COX1**	**CYTB**	**ND1**	**ND2**
Atretochoana eiselti	Y		KX757071	KX757082	KX757090			
Boulengerula boulengeri	Y		EF107199					
Boulengerula cf. *boulengeri*	N	GQ244464						
Boulengerula changamwensis	Y		FN652690	FN652722	FR691671	FN652754		
Boulengerula fischeri	Y		FR691654	FR691658	FR691668	FR691663		
Boulengerula niedeni	Y		FN652691	FN652723	FR691670	FN652755		
Boulengerula taitana	Y	AY954504						
Boulengerula uluguruensis	Y	DQ283087	DQ283087	FR691672	FN652748			
Brasilotyphlus dubium	N			MG162601		MG162607		
Brasilotyphlus guarantanus	N			MG162598		MG162606		
Caecilia gracilis	Y	NC_023508						
Caecilia isthmica	N			MN555719	MN555727			
Caecilia pulchraserrana	N			MN555715	MN555723			
Caecilia tentaculata	Y	NC_023507						
Caecilia thompsoni	N			MN555717	MN555725			
Caecilia volcani	Y	GQ244466						
Chikila fulleri	Y	NC_021369						
Chthonerpeton indistinctum	Y	NC_023509						
Chthonerpeton viviparum	N			KU495177	KU494384			
Crotaphatrema lamottei	Y	NC_019596						
Crotaphatrema tchabalmbaboensis	Y		EF219345	EF219346				

Continued on the next page

TABLE 5.2. *(cont.)* List of taxa used in the caecilian phylogenetic analyses, and GenBank accessions for mitogenomes and mitochondrial loci (mtDNA). See table 5.1 for nuclear loci (nucDNA).

Taxon	**nucDNA**	**mitogenome**	**12S**	**16S**	**COX1**	**CYTB**	**ND1**	**ND2**
Dermophis mexicanus	Y	GQ244467						
Dermophis oaxacae	Y		DQ283455	DQ283455				
Dermophis parviceps	N			EU753994		EU754006		
Epicrionops marmoratus	N	KF540151						
Epicrionops parkeri	N			MN555716	MN555724			
Gegeneophis carnosus	Y		HQ443936	HQ443979	HQ444022	HQ444070		
Gegeneophis danieli	Y		HQ443948	HQ443991	HQ444033	HQ444080		
Gegeneophis goaensis	Y		HQ443943	HQ443986	HQ444029	HQ444075		
Gegeneophis krishni	Y		HQ443954	HQ443997	HQ444039	HQ444085		
Gegeneophis madhavai	Y		HQ443956	HQ443999	HQ444041	HQ444087		
Gegeneophis mhadeiensis	Y		HQ443940	HQ443983	HQ444026	HQ444074		
Gegeneophis orientalis	Y		KP400601	KP400605	KP400611	KP400609		
Gegeneophis pareshi	Y		KP400600	KP400604	KP400610	KP400608		
Gegeneophis ramaswamii	Y	AY456250						
Gegeneophis seshachari	Y		HQ443970	HQ444013	HQ444055	HQ444101		
Geotrypetes seraphini	Y	AY954505						
Grandisonia alternans	Y	KU974367						
Grandisonia larvata	Y	GQ244470						
Grandisonia sechellensis	Y	NC_023510						
Gymnopis multiplicata 1	N	GQ244471						
Gymnopis multiplicata 2	Y	KF540153						

Continued on the next page

TABLE 5.2. *(cont.)* List of taxa used in the caecilian phylogenetic analyses, and GenBank accessions for mitogenomes and mitochondrial loci (mtDNA). See table 5.1 for nuclear loci (nucDNA).

Taxon	**nucDNA**	**mitogenome**	**12S**	**16S**	**COX1**	**CYTB**	**ND1**	**ND2**
Herpele squalostoma	Y	NC_019586						
Hypogeophis brevis	Y	KU753817						
Hypogeophis montanus	Y			MH055428				
Hypogeophis pti	Y			MH055418				
Hypogeophis rostratus	Y	GQ244472						
Ichthyophis asplenius	Y	KF540148						
Ichthyophis biangularis	N		AB686145	AB686145		AB686080		
Ichthyophis cardamomensis	N				KP264594	KP264619		
Ichthyophis catlocensis	N				KP264602	KP264626		
Ichthyophis chaloensis	N					KP264625		
Ichthyophis glutinosus	Y	AY456251						
Ichthyophis kohtaoensis	Y	AY458594						
Ichthyophis larutensis	N		AB686156	AB686156		AB686091		
Ichthyophis longicephalus	N			JQ040046				
Ichthyophis multicolor	N		FR715999	FR716007	MG935541	FR716015		
Ichthyophis nguyenorum	N				KP264596	KP264621		
Ichthyophis orthoplicatus	Y		AY101213	AY101233		AY101253		
Ichthyophis supachaii	N				KP264600	KP264624		
Ichthyophis tricolor	N		AF461138	AF461139				
Idiocranium cf. *russelli*	N	KF540156						
Indotyphlus battersbyi	Y		HQ443973	HQ444016	HQ444058	HQ444104		

Continued on the next page

TABLE 5.2. *(cont.)* List of taxa used in the caecilian phylogenetic analyses, and GenBank accessions for mitogenomes and mitochondrial loci (mtDNA). See table 5.1 for nuclear loci (nucDNA).

Taxon	**nucDNA**	**mitogenome**	**12S**	**16S**	**COX1**	**CYTB**	**ND1**	**ND2**
Indotyphlus maharashtraensis	Y	NC_023512						
Luetkenotyphlus brasiliensis	Y	NC_023513						
Luetkenotyphlus fredi	N		MK660790	MK660785	MK660780			
Luetkenotyphlus insulanus	N			KU495583	KU494790			
Microcaecilia dermatophaga	N	NC_023514						
Microcaecilia nicefori	N			MN555722	MN555729			
Microcaecilia sp.	N	GQ244473						
Microcaecilia unicolor	N	NC_023515						
Oscaecilia ochrocephala	Y	GQ244474						
Potomotyphlus kaupii	Y	NC_023516						
Praslinia cooperi	Y	NC_023517						
Rhinatrema bivittatum	Y	AY456252						
Rhinatrema gilbertogili	Y		MH177827	MH177830	MH177833			
Rhinatrema nigrum	Y	GQ244468						
Rhinatrema shiv	N		GU566188	GU566189		GU566190		
Rhinatrema uaiuai	Y		MH177826	MH177828	MH177832			
Schistometopum cf. *gregorii*	N	NC_023518						
Schistometopum gregorii	Y		DQ283089	DQ283089				
Schistometopum thomense	Y	GQ244476						
Scolecomorphus kirkii	N			MT433517				
Scolecomorphus uluguruensis	Y		AY450618	AY450625				

Continued on the next page

TABLE 5.2. *(cont.)* List of taxa used in the caecilian phylogenetic analyses, and GenBank accessions for mitogenomes and mitochondrial loci (mtDNA). See table 5.1 for nuclear loci (nucDNA).

Taxon	**nucDNA**	**mitogenome**	**12S**	**16S**	**COX1**	**CYTB**	**ND1**	**ND2**
Scolecomorphus vittatus	Y	AY456253						
Siphonops annulatus	Y	AY954506						
Siphonops hardyi	Y			DQ283088	KU494789			
Siphonops paulensis	Y		KX757080	EF107203	KX757094	AY954507		
Typhlonectes compressicauda	Y		KX757073	KX757084	KX757092		AY916014	AY916014
Typhlonectes natans	Y	AF154051						
Uraeotyphlus bombayensis	N	NC_023511						
Uraeotyphlus cf. *oxyurus*	Y	AY456254						
Uraeotyphlus gansi	N	NC_023519						
Uraeotyphlus narayani	Y		DQ283090	DQ283090		AY101242		
Outgroups								
Gallus gallus	Y	NC_040970						
Anolis carolinensis	Y	EU747728						
Mus musculus	Y	NC_010339						
Xenopus laevis	Y	MH991335						
Bombina orientalis	Y	NC_006689						
Leiopelma archeyi	Y	NC_014691						
Lyciasalamandra atifi	Y	NC_002756						
Ranodon sibiricus	Y	AJ419960						
Andrias davidianus	Y	KU131042						
Latimeria chalumnae	Y	AB257297						
Protopterus annectens	Y	PRRMTCG						

analyses were to infer the most accurate summary tree possible, an MRC tree with minority components (i.e. greedy consensus) should be used instead (Mirarab 2019). Particularly, since collapsing poorly supported branches ($< 5 - 20\%$) increases the accuracy of the summary trees inferred by Astral, but collapsing moderately supported branches ($< 50 - 75\%$) increases topological errors (Mirarab 2019, Zhang et al. 2018).

5.2.1.3 Testing for taxonomic instability with tree jackknifing

To test the supertree for taxonomic instability caused by non-effective overlap the inferred gene trees were run though the Concatabominations pipeline (Siu-Ting et al. 2015), and visualised with Cytoscape v.3.7.1 (Demchak et al. 2014). Because the pipeline was developed to look for instability in morphological datamatrices, the gene trees used as input for the supertree analysis were converted to their MR using p4 v.1.2.0 (Foster 2004) prior to running the pipeline.

To test whether the Concatabominations pipeline might be used to inform decisions of which gene trees are the best candidates for increased taxonomic sampling in order to solve issues of non-effective overlap, the MR was subjected to a "tree jackknife". In other words, one gene tree was removed from the MR, and the ensuing matrix run through Concatabominations, with the procedure being repeated for each gene tree. If the removal of a gene tree from the matrix increases the amount of instability recovered by the Concatabominations pipeline, then that tree is considered a "stabilising" tree, and the locus is a candidate for a targeted increase in taxonomic sampling.

5.2.1.4 How does Concatabominations work

The Concatabominations pipeline is a heuristic extension of the STR approach to identifying unstable taxa, as mentioned above, and uses compatibility methods (e.g., Meacham & Estabrook 1985) to test whether chimaeric taxa (concatabominations) that combine pairs of "potential taxonomic equivalents that are asymmetric both ways" might increase homoplasy in a dataset, and thus taxonomic instability (Siu-Ting et al. 2015). These pairs of taxa, originally defined in Wilkinson (1995*b*) and referred to as D pairs in Siu-Ting et al. (2015), consist of taxa that share some identical character states, but both have characters that are scored in one taxon and missing from the other. Borrowing Siu-Ting et al.'s (2015) example, if taxon x = ??0111 and taxon y = 0001??, then they are "potential taxonomic equivalents that are asymmetric both ways", and form the concatabomination x+y = 000111. For an explanation on other types of taxonomic equivalence please see Siu-Ting et al. (2015) and Wilkinson (1995*b*) Any taxon that contributes to multiple concatabominations without increasing homoplasy then becomes a candidate for safe removal, i.e. its removal does not affect the relationships of the remaining taxa. In the context of MR, the D pairs correspond to taxa sampled for a limited number of loci, and any taxon identified as a candidate for safe removal, based on its inclusion in chimaeric taxa, to an instance of instability due to non-effective overlap.

On a more practical level, the networks outputted by the Concatabominations pipeline can be visualised and manipulated in Cytoscape. This provides an easily interpretable graphical-depiction of the (unstable) relationships between taxa, and allows users to explore what happens to their dataset as unstable taxa are removed, without the need to infer a new tree each time. However, because Concatabominations is a heuristic method it should be noted that there may be instances where it identifies unstable taxa whose removal will affect the relationships between the remaining taxa (Siu-Ting et al. 2015)—i.e. taxa that cannot be 'safely' removed. Thus, despite the exploratory capabilities of the network visualisation, if the goal is to decrease taxonomic instability by taxon removal, the latter should be removed from the matrix individually and the new matrix (or matrices) rerun through the pipeline to check that only taxa that can be safely removed have been removed.

5.2.2 Comparison of tree jackknifing and gene sampling sufficiency

To illustrate that tree jackknife is a method that can inform targeted data acquisition, I reanalysed a selection of the datasets used by Dobrin et al. (2018) to explore the theoretical terrace space occupied by a variety of phylogenomic datasets, see table 5.3. Datasets that passed gene sampling sufficiency or were terminated in Dobrin et al.'s (2018) analyses were not re-analysed here. Re-analyses of the bat (Shi & Rabosky 2015) and primate (Springer et al. 2012) datasets were abandoned because each tree jackknife run was taking over two weeks to complete, due to the size of the matrices in both datasets (all loci matrix sizes: 815 taxa by 3415 characters for bats; 372 taxa by 10692 characters for primates). To allow the selected datasets to be run through the tree jackknife protocol, individual loci were extracted from the concatenated matrices and all loci with fewer than four taxa (minimum number of taxa required to infer an unrooted tree) were removed from the selected datasets. Gene trees were obtained with RAxML v.8.2.12 (Stamatakis 2014), with the multiple tree search set to 1000 runs without bootstrap and with GTR+I+G as the substitution model for all gene trees, and the best trees were used as input for the Astral analysis.

Gene sampling sufficiency (ζ) was calculated for all reanalysed datasets and for the caecilian matrices, following Dobrin et al. (2018) and Steel (2016). In order to compute sampling sufficiency, it is first necessary to calculate a theoretical minimum number of loci, k_{min}, that must be sampled to ensure that taxon coverage, S, would be decisive for tree T. The k_{min} calculation assumes random taxon coverage, and need not be computed if any locus has been sampled for all taxa, since that is the sufficient condition for S to be decisive (Sanderson et al. 2010, Steel 2016). To calculate k_{min}, I need to first compute the taxon coverage density for the dataset (d), or the average proportion of taxon sampling, with:

$$d = \frac{\sum \frac{n_l}{n}}{k},$$

where n_l is the number of taxa sampled in a locus, n is the total number of sampled taxa, and k the

number of sampled loci. Following Dobrin et al. (2018), I calculated k_{min} using the approximation:

$$k_{min} \approx \frac{log\frac{n^3}{6p}}{-log(1-d^4)},$$

where the new term p is the desired confidence level on the decisiveness of S (i.e. the 'p-value'). The latter was set to 0.05 for all analyses, to allow for direct comparison between the results below and Dobrin et al.'s (2018). With d and k_{min} computed, we can finally calculate gene sampling sufficiency with:

$$\zeta = -ln\frac{k}{k_{min}}.$$

A value of $\zeta \geq 0$ means enough loci have been sampled to achieve taxon coverage decisiveness, and $\zeta < 0$ means that more loci must be sampled.

I also explored the theoretical terrace space associated with the supertrees of the reanalysed datasets and the caecilian trees with *terraphy* (Zwickl 2014). While neither of the phylogenetic analyses in this chapter meet the conditions to form terraces (Bayesian inference is not susceptible to terraces (Sanderson et al. 2015), RAxML's default branch length parameters are linked across partitions (Dobrin et al. 2018, Stamatakis 2014), and concatenated alignments were not used for tree inference), I wanted to compare the theoretical terrace sizes for the filtered and unfiltered matrices of the datasets analysed by Dobrin et al. (2018). This also allowed for the comparison of the resolution of the supertree (branches with a quartet score, QS, bellow 50%, including indeterminate scores ('?') were collapsed) and the strict consensus (SC, Sokal & Rohlf 1981) of the theoretical terraces.

5.3 Results

5.3.1 Phylogenetic analyses

The pre-jackknife supertree analysis yielded a mostly well-resolved phylogeny (Fig.5.2), with the family-level relationships broadly congruent with recent molecular studies (Jetz & Pyron 2018, Pyron 2014, Pyron & Wiens 2011, Roelants et al. 2007, San Mauro et al. 2014). However, the split between Scolecomorphidae and the other Terasomata, recovered by Jetz & Pyron (2018), Kamei et al. (2012), Pyron (2014), and San Mauro et al. (2012, 2014), collapses resulting in the following polytomy ((Chikilidae, Herpelidae), (Scolecomorphidae, (Typhlonectidae, Caeciliidae), other Terasomata)), consistent with the low support found for this split in Pyron & Wiens (2011) and San Mauro et al. (2009). As in Acosta-Galvis et al. (2019), the supertree recovered Siphonopidae as paraphyletic due to the placement of *Microcaecilia nicefori*, and *Brasilotyphlus* nested within *Microcaecilia*. The latter was also recovered in Maciel et al. (2009) and Maciel et al. (2019). As in San Mauro et al. (2014), *Gymnopis multiplicata* is paraphyletic with the genus *Dermophis*. Additionally, unlike Maciel et al. (2019), the supertree does not support the (*Siphonops*, *Luetkenotyphlus*) split. At the intrafamily-level, within Grandisoniidae (formerly Indotyphlidae, Dubois

et al. 2021, Frost 2022), *Hypogeophis* and *Grandisonia* are recovered as paraphyletic in relation to each other (polytomy prevents unambiguous identification of which is the paraphyletic genus), which is consistent with previous work on the Seychelles caecilian radiation (Maddock et al. 2018). Additionally, the phylogenetic placement of *Hypogeophis montanus* is recovered as indeterminate (quartet score = ?) in relation to all other grandisoniids. And in the Typhlonectidae family, *Potomotyphlus* was nested within *Typhlonectes*. Lastly, *Rhinatrema*, *Ichthyophis*, *Boulengerula*, *Caecilia*, *Gegeneophis* and *Luetkenotyphlus* display intrageneric polytomies.

5.3.2 Tree jackknife analyses

The Concatabominations analysis, including all Gymnophiona gene trees, found most taxa to be stable (64 singletons and 11 dyads), except for a cluster of instability with 17 taxa anchored by the Seychelles caecilian *H. montanus*, which was identified as candidate for safe *a priori* deletion (Fig.5.3). Taxa are safe to remove if their omission does not affect the relationships between the remaining species (Wilkinson 1995*b*). Along with *H. montanus*, *Boulengerula changamwensis*, *Brasilotyphlus guarantanus*, *Caecilia thompsoni*, *Chthonerpeton viviparum*, *Dermophis parviceps*, *Epicrionops parkeri*, *Ichthyophis biangularis*, *Ichthyophis longicephalus*, *Ichthyophis nguyenorum*, *Ichthyophis orthoplicatus*, *Indotyphlus battersbyi* and *Scolecomorphus kirkii* were also identified as candidates for safe deletion. These taxa are recovered by Concatabominations as part of dyads, which correspond to sister leaves in the supertree, except for *I. battersbyi*, which is part of the cluster of unstable taxa anchored by *H. montanus*. If the goal of running the taxonomic instability analysis is simply to identify taxa whose deletion increases tree resolution, then deleting *H. montanus* from the Concatabominations network, as well as from the MR, results in a decrease of taxonomic instability, with only singletons and dyads left in the instability network. Inspection of the taxonomic complement of the gene trees confirms that *H. montanus*' instability is due to non-effective overlap, as the taxon is present in only one of 24 gene trees (BDNF), whith the four *Hypogeophis* species included forming a polytomy (tables 5.1 and 5.3).

The tree jackknife procedure identified six gene trees which, when removed, lead to increased taxonomic instability, making them candidate trees for a targeted increase of taxon sampling (table 5.3). These candidate loci are 12S and 16S rRNA, COX1, CYTB, H3A and ND1, with removal of the 16S gene tree showing the largest increase in taxonomic instability (i.e. the best candidate locus). While it might be expected that all candidate loci correspond to the most taxon rich loci (true for four of the six candidate loci), RAG1 and ND2 both have higher taxon sampling than ND1, and H3A is one of the least taxon-rich loci (25 of 105 taxa). Thus, taking the existing taxon sampling into account, only four of the candidate loci are realistic targets for increased sampling: 12S and 16S rRNA, COX1 and CYTB. With 16S rRNA being the most promising candidate for increased taxon sampling.

Preliminary sequence alignments revealed a series of *Hypogeophis* spp. DNA sequences (MH055413–MH055428, Maddock et al. 2018) labelled as 12S rRNA that aligned poorly with

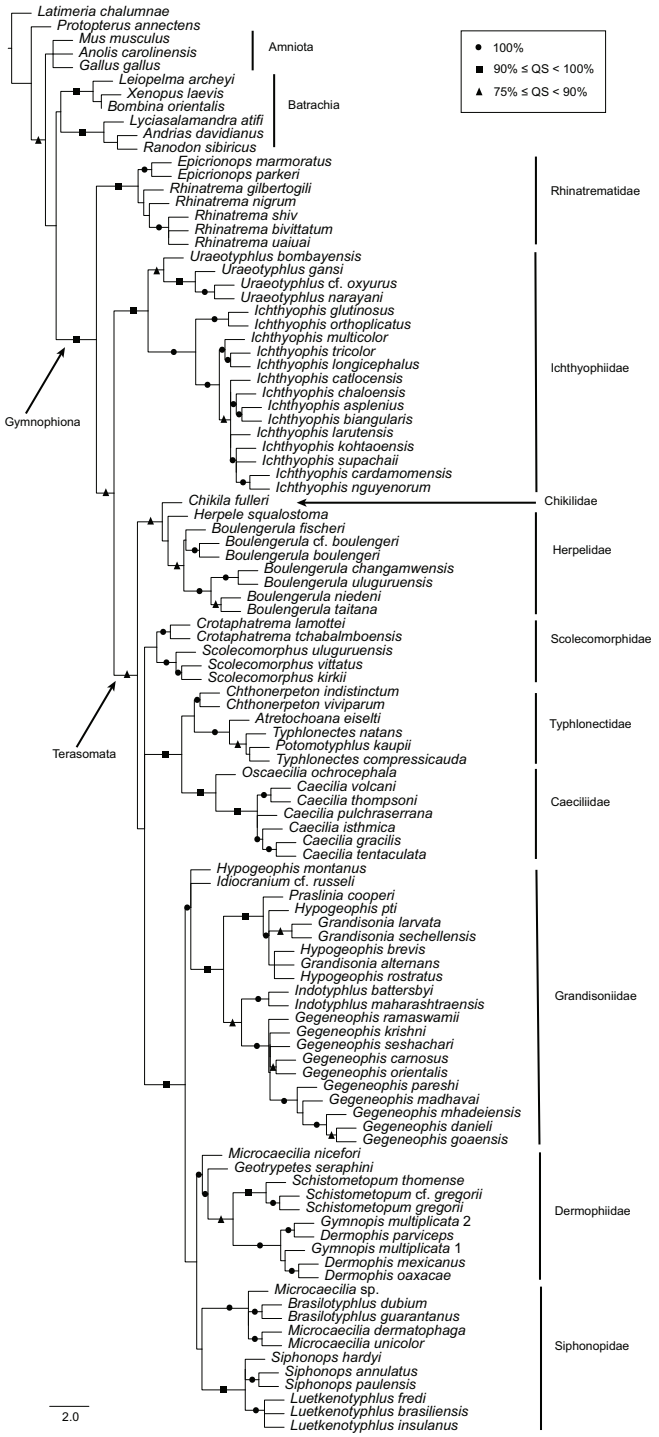


FIGURE 5.2. Supertree of the caecilian gene trees inferred with MrBayes prior to taxonomic instability analyses. All branches with quartet support (QS) below 50% collapsed. Circles denote branches with 100% QS, squares denote 90% ≥ QS < 100% and triangles 75% ≥ QS < 90%. Scale bar represents coalescent units.

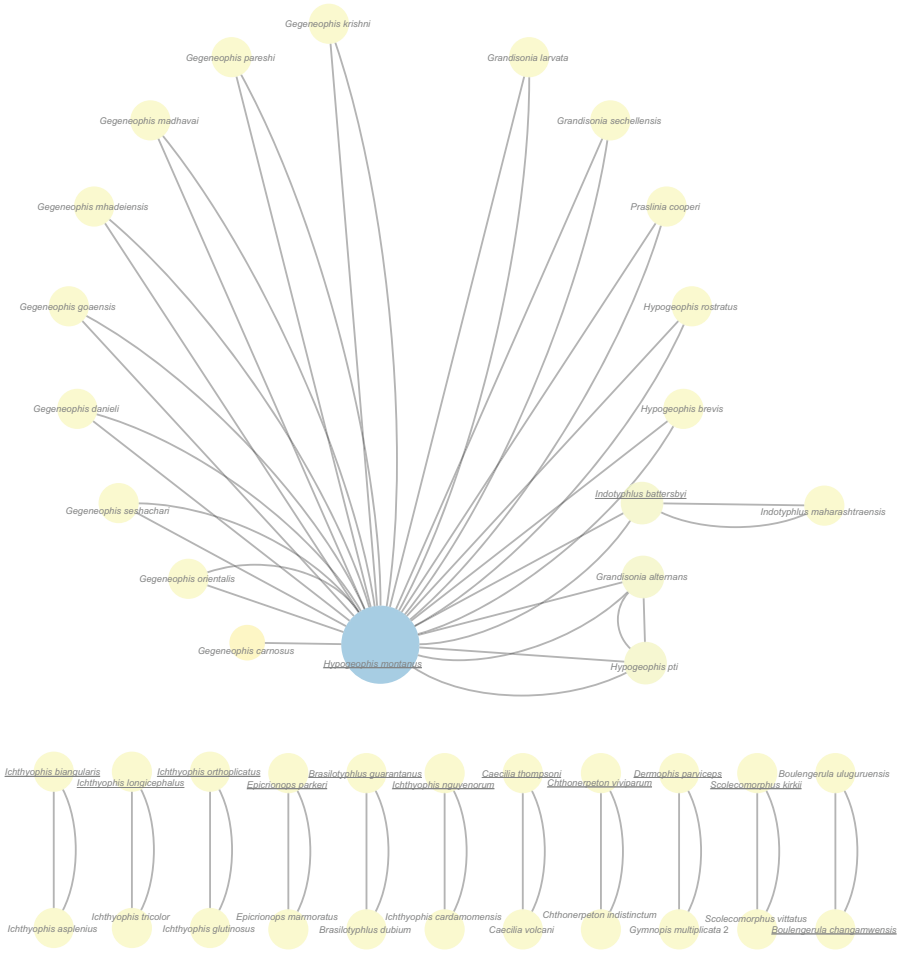


FIGURE 5.3. Concatabominations network of the Gymnophiona supertree with 64 stable, unconnected taxa omitted. Candidates for safe *a priori* deletion are underlined.

the other 12S rRNA sequences, which led me to remove these sequences from the 12S rRNA alignment. However, upon further inspection it was revealed that the sequences were in fact mislabelled 16S rRNA sequences (confirmed with BLAST (Altschul et al. 1990) and by S Maddock 2019, personal communication, 11 December), with MH055425–MH055428 corresponding to *Hypogeophis montanus* sequences. Thus, through serendipitous means, there was available sequence data for the target taxon/locus pair identified by the tree jackknife analyses.

After the addition of data for *H. montanus*, the 16S rRNA alignment had a length of 1142bp and the Concatabominations pipeline yielded a network made-up entirely of singletons and dyads (again consisting of cherries), meaning that all, or most, taxa are taxonomically stable. Additionally, tree resolution (ρ) increased (no *H. montanus* 16S rRNA, $\rho = 0.95$; with *H. montanus* 16S rRNA, $\rho = 0.96$), showing that the effects of non-effective overlap were successfully mitigated (Fig.5.4). Following Dobrin et al. (2018), ρ is defined as the ratio of the splits in a tree and the splits in a binary tree of equal size. While the absolute increase in internal branches is one, there

TABLE 5.3. Concatabominations with tree jackknife results, number of edges excludes self-loops. Candidate loci for targeted taxon sampling highlighted in bold text.

MR	Taxa in network	Edges in network	Taxa in jackknifed locus
All loci	105	59	NA
All loci, no *Hypogeophis montanus*	104	46	NA
no 12S rRNA	105	61	84
no 16S rRNA	103	135	99
no 18S rRNA	105	59	10
no 28S rRNA	105	59	19
no ATP6	105	59	53
no ATP8	105	59	53
no BDNF	104	25	13
no COX1	105	71	86
no COX2	105	59	53
no COX3	105	59	53
no CXCR4	105	59	33
no CYTB	104	63	81
no H3A	105	61	25
no NCX1	105	59	34
no ND1	105	63	54
no ND2	105	59	55
no ND3	105	59	53
no ND4	105	59	53
no ND4L	105	59	53
no ND5	105	59	53
no ND6	105	59	53
no RAG1	105	59	65
no SIA	105	59	13
no SLC8A3	105	59	32
with 16S rRNA for *H. montanus*	105	23	NA

are some notable differences between the Gymnophiona supertree inferred pre- and post-tree jackknife analyses.

In the post-jackknife supertree, the split between Scolecomorphidae and the other Terasomata

is now recovered (with the same topology as Kamei et al. 2012, San Mauro et al. 2014), albeit with very low support (QS=50.2%). Within the Seychelles clade, *H. montanus* is recovered as sister to *Hypogeophis brevis* (also recovered by Maddock et al. 2018), and, while *Hypogeophis* and *Grandisonia* remain paraphyletic, two distinct clades can now be seen—one comprised of *Hypogeophis pti*, *Grandisonia larvata* and *Grandisonia sechellensis* and another made up of *Grandisonia alternans*, *Hypogeophis rostratus*, *H. brevis* and *H. montanus*. However, in Siphonopidae+Dermophidae the branch leading to *M. nicefori* collapsed. Small intragenus changes in extremely short and poorly supported branches were also found between the two supertrees, but these are not easily discernible. While these intragenus areas of instability (e.g., *Ichthyophis*) were not explored, they may result from "incomplete" non-effective overlap, i.e. they cannot be unambiguously placed in relation to their putative closest relatives, but their presence in at least one of the "stabilising" gene trees is masking their instability in the Concatabominations analyses. Also, given that several taxa in these localised areas of instability are taxa with partial gene sequences, they may benefit from more complete locus sequences, rather than from increased taxon/locus sampling. However, the increased resolution within Grandisoniidae after the addition of *H. montanus* data to the 16S rRNA alignment shows that tree jackknifing is a useful tool for identifying taxa with poor overlap between loci, as well as which loci contribute the most to the inferred (supertree) topology.

5.3.3 Gene sampling sufficiency and terraces

In this subsection and any subsequent discussions, I will be ignoring the results for the *Ficus* and scincids datasets (which were not altered from the versions available at `https://github.com/BDobrin/data.sets`), because there are discrepancies between the taxon coverage density (d), k_{min} and gene sampling sufficiency (ζ) reported by Dobrin et al. (2018) and those recalculated by me (table 5.4). Since these values rely only on the number of taxa and loci in each dataset, and the same calculations (d, k_{min}, ζ) were consistent between the two studies for other unedited datasets (Chameleons and Rosaceae), all results discussed below assume that the correct formulae were programmed into Microsoft Excel (Office 365 v.2204) and that the matrices provided in the data repository may have been edited prior to Dobrin et al.'s (2018) analyses. Also, how significant figures were treated, and any resulting rounding errors, cannot be discarded as a contributing factor for at least some of the calculations, see the Gymnophiona entries in table 5.4.

5.3.3.1 Taxon coverage and gene sampling sufficiency

The removal of loci with fewer than four taxa (minimum required for inference of unrooted trees) from the matrices used in Dobrin et al. (2018), resulted in increased taxon coverage density for all datasets (Fig.5.5a). Except for the Chameleons, *Ficus* and Rosaceae datasets, where there was no change, and in the Scincids dataset, where $\Delta d = -0.07$, but as noted above the *Ficus* and Scincid, while reported, will not be discussed. This increase in taxon coverage density in turn

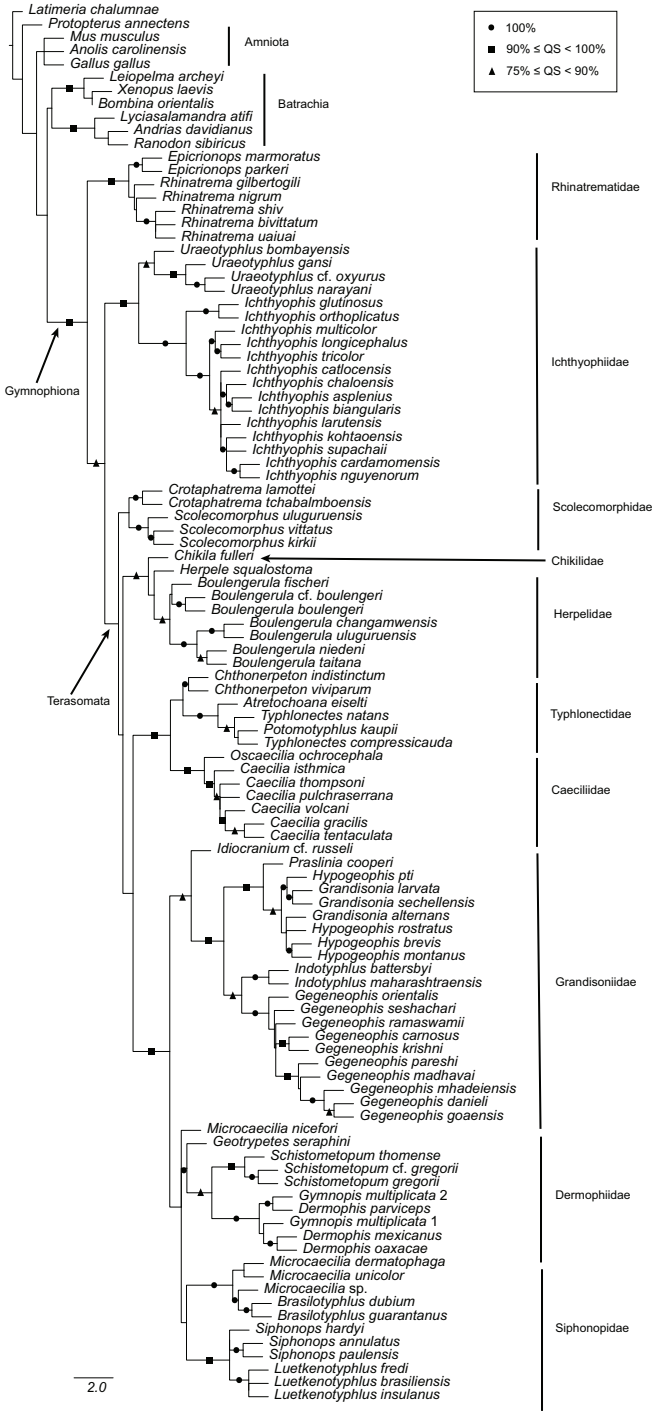
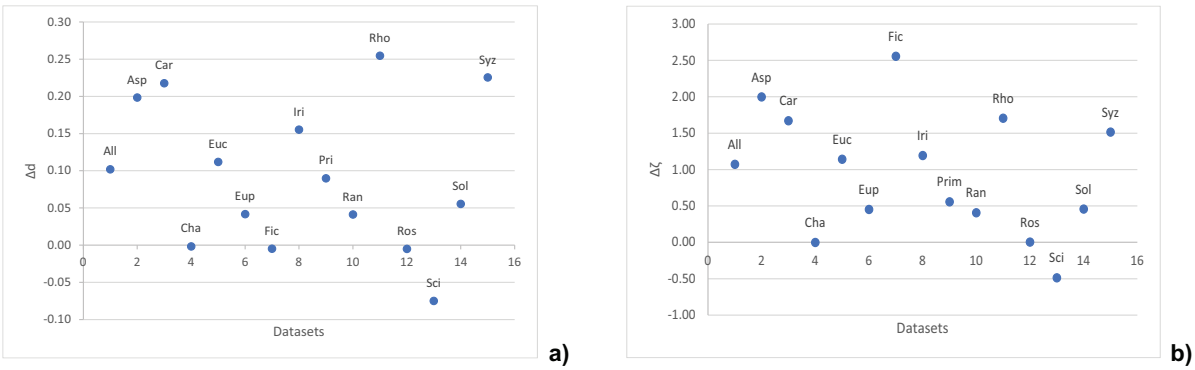


FIGURE 5.4. Supertree of the caecilian gene trees inferred with MrBayes after taxonomic instability analyses. All branches with quartet support (QS) below 50% collapsed. Circles denote branches with 100% QS, squares 90% ≥ QS < 100% and triangles 75% ≥ QS < 90%. Scale bar represents coalescent units.



a) b)

FIGURE 5.5. Change in a) taxon coverage density and b) gene sampling sufficiency between Dobrin et al. (2018) and this study.

lead to an improvement of the gene sampling sufficiency scores (Fig.5.5b), although no dataset was found to have met the $\zeta \geq 0$ requirement for taxon coverage decisiveness (assuming a 5% confidence level, table 5.4). This is in broad agreement with Dobrin et al.'s (2018) results. And like them, the recalculated minimum loci needed for decisiveness (k_{min}) were in the hundreds to thousands of loci, despite many datasets having terraces with fewer than ten trees (Chameleons, *Iris*, *Ranunculus*, *Rhododendron* and *Syzygium*).This discrepancy was also found by Dobrin et al. (2018), again confirming that Sanderson et al.'s (2010) and Steel's (2016) assumption of random uniform taxon sampling is not necessarily reflected by empirical datasets.

5.3.3.2 Tree jackknife *vs.* gene sampling sufficiency

Along with terrace size, the sampling sufficiency values are also unexpectedly high compared to the supertree resolution for the majority of datasets (table 5.5); with two datasets having fully resolved supertrees, yet having a $k_{min} > 1000$ loci. For the Gymnophiona datasets, where k_{min} was 308 pre-jackknife and 307 post-jackknife, with a sampling sufficiency of -2.55 (for both), despite having one locus with $\approx 94\%$ taxon coverage and both inferred supertrees being highly resolved ($\rho \geq 0.95$). This contrasts with the tree jackknife analyses, which identified six "stabilising" loci, with four strong candidates for increased taxon sampling, and one unstable taxon (*H. montanus*). While the addition of a single *H. montanus* sequence to the 16S rRNA alignment lead to an increase in supertree resolution and to Concatabominations no longer finding unstable taxa, in the sampling sufficiency analyses this only lead to k_{min} decreasing by one locus.

For the datasets used in Dobrin et al. (2018), tree jackknife analyses yield results similar to those reported for caecilians, although the effects of targeted increased taxon sampling were not tested for these. However, unlike the hundreds to thousands of loci identified by gene sampling sufficiency, tree jackknifing identified a maximum of four candidate loci for increased taxon

TABLE 5.4. Results of the gene sampling sufficiency (ζ) and terrace analyses on the selection of Dobrin et al.'s (2018) datasets and pre- and post-jackknife Gymnophiona datasets. The length of the concatenated matrices may differ from those reported by Dobrin et al. (2018) because they include columns with no data. The calculations reported in this table include the taxon coverage density (d), tree resolution (ρ) of the terraces' strict consensus (SC), minimum loci required for decisiveness (k_{min}) and gene sampling sufficiency (ζ).

Dataset	**Taxa**	**Loci**	**Length of matrix**	**d**	**Terrace size**	**ρ of SC of terrace**	**k_{min} for p=0.05**	**ζ**	**Reference**
Allium	57	4	4512	0.34	1.2×10^7	0.29	967	-5.49	Zanne et al. (2014)
Asplenium	133	3	3862	0.41	39	0.85	562	-5.23	Zanne et al. (2014)
Caryophyllaceae	224	4	5723	0.51	1.2×10^{14}	0.83	253	-4.15	Zanne et al. (2014)
Chameleons	202	6	5054	0.92	1	1.00	14	-0.83	Tolley et al. (2013)
Eucalyptus	136	4	5639	0.34	53	0.80	1,159	-5.67	Zanne et al. (2014)
Euphorbia	131	6	8805	0.32	10,143	0.85	1,467	-5.50	Zanne et al. (2014)
Ficus	112	5	6245	0.36	112,995	0.86	955†	-5.25†	Zanne et al. (2014)
Iris	137	4	5526	0.49	1	0.99	280	-4.25	Zanne et al. (2014)
Primula	185	5	6629	0.52	1,701	0.90	222	-3.79	Zanne et al. (2014)
Ranunculus	170	6	10264	0.35	3	0.99	1,082	-5.20	Zanne et al. (2014)
Rhododendron	117	4	5691	0.60	9	0.96	108	-3.30	Zanne et al. (2014)
Rosaceae	529	7	11728	0.30	2.2×10^{22}	0.75	2,627	-5.93	Zanne et al. (2014)
Scincids	213	6	5283	0.71†	3	0.99	61†	-2.32†	Rabosky et al. (2014)
Solanum	187	6	9076	0.37	71,775	0.88	939	-5.05	Zanne et al. (2014)
Syzygium	106	3	4252	0.58	3	0.98	131	-3.78	Zanne et al. (2014)
Gymnophiona pre-jackknife	105	24	21782	0.47	2,457	0.81	308	-2.55	this thesis
Gymnophiona post-jackknife	105	24	21695	0.47	81	0.96	307	-2.55	this thesis

†*These values conflict with those reported by Dobrin et al. (2018).*

sampling (table 5.5). Additionally, no more than three unstable taxa were identified per dataset, except for Rosaceae, which had twice as many taxa as the next most speciose dataset and the lowest taxon coverage density of all analysed datasets. While unstable taxa were identified for some datasets with binary supertrees, they corresponded to taxa that were present in a single gene tree, usually with low taxon coverage, similar to the *H. montanus* case. An example of this is the *Eucalyptus* dataset, where *Eucalyptus lehmannii*, the most unstable taxon, is sampled only in the rbcL gene tree, which consists of a quartet tree. Much like in the caecilian alignments, none of the reanalysed datasets included a locus with full taxon coverage and the stabilising loci are not a one-to-one match to the most taxon-rich loci. Thus, tree jackknifing again identifies reasonable numbers of loci and taxa that can be targeted for increased sampling, which are one to three orders of magnitude smaller than the calculated minimum number of loci required for decisiveness.

5.3.3.3 Terraces

As mentioned in section 5.2.2, the analytical parameters of the tree searches in this chapter are not conducive to the formation of terraces, thus the results detailed below are a theoretical exercise rather than a reflection of the 'true' terrace space associated with the concatenated alignments. Also trees with resolution $\geq 99\%$ are treated as binary for the following discussion. As in Dobrin et al.'s (2018) original analyses, terrace size ranged over multiple orders of magnitude, from 1×10^0 to over 2×10^{22} trees. However, unlike taxon coverage density and gene sampling sufficiency, there was not a uniform trend in the direction of change for terrace size. While three of the datasets whose terrace size increased after reanalysis yielded polytomous supertrees (*Allium*, Caryophyllaceae and *Euphorbia*), what unites them with the fourth dataset, *Eucalyptus*, is that all four have at least one taxon that is sampled for only one locus (*Allium triquetum* is sampled in two loci, but one has only six taxa). The presence of polytomies is also not synonymous with larger terrace sizes. The supertrees for *Primula*, Rosaceae and *Solanum* are all partially resolved, yet terrace size decreased for all three datasets, even though their trees were fully resolved for Dobrin et al.'s (2018) analyses. Interestingly, the addition of a single DNA sequence to the caecilian dataset (*H. montanus*' 16S rRNA) led to a two order of magnitude decrease in terrace size (although less drastic than the difference in size between the sets of all possible resolved trees for each Gymnophiona supertree, pre-jackknife = $20^3 = 8000$ and post-jackknife = $4^3 = 64$).

5.4 Discussion

The increased focus on large-scale phylogenetics, whether in numbers of taxa (e.g., Bininda-Emonds et al. 2007, Pyron & Wiens 2011, Wächter & Melzer 2020) or loci (e.g., Chan et al. 2020, Siu-Ting et al. 2019, Streicher et al. 2020), has re-ignited the debates of how best to infer trees

TABLE 5.5. Tree jackknife results and supertree resolution (ρST) for each reanalysed Dobrin et al. (2018) dataset.

Dataset	ρ **ST**	**Candidate loci**	**Unstable taxa**
Allium	0.80	rbcL, matK	*Allium siculum*, *Allium textile*, *Allium triquetum*
Asplenium	0.98	rbcL	*Dicksonia sellowiana*
Caryophyllaceae	0.91	ITS, matK, t-RNAs	*Spergularia marina*
Chameleons	1.00	All	None
Eucalyptus	0.99	None	*Eucalyptus lehmannii*, *Eucalyptus rodwayi*
Euphorbia	0.96	ITS, matK, t-RNAs	*Euphorbia silvifolia*, *Euphorbia polyacantha*
Ficus	0.95	ITS, matK, rbcL, t-RNAs	*Ficus trigonata*, *Ficus bullenei*, *Ficus asperula*
Iris	0.99	All	None†
Primula	0.97	matK, rbcL, t-RNAs	*Primula veitchiana*
Ranunculus	1.00	ITS, matK	None†
Rhododendron	0.99	matK, t-RNAs	*Rhododendron micranthum*
Rosaceae	0.95	18S, matK, t-RNAs, ITS	*Aphanes microcarpa*, *Crataegus columbiana*, *Amelanchier alnifolia*, *Rubus arizonensis*, *Rosa orientalis*, *Rosa corymbifera*, *Geum macrophyllum*, *Geum aleppicum*, *Rubus discolor*
Scincids	0.99	All	None
Solanum	0.96	ITS, matK, rbcL, t-RNAs	*Solanum erianthum*
Syzygium	0.99	All	None†

†*Some taxa in triads or rhombi, but their removal does not affect tree resolution.*

(e.g., de Queiroz & Gatesy 2007, Gatesy et al. 2002, Von Haeseler 2012)[1] and also of how to deal with missing data in these large matrices (e.g., Molloy & Warnow 2017, Roure et al. 2012, Streicher et al. 2016). As seen in Chapter 3, many approaches to dealing with missing data focus on identifying and removing unstable taxa in order to increase tree resolution (e.g., Aberer et al. 2012, Thorley & Wilkinson 1999, Wilkinson 1995*b*) and/or remove long branches (Mai & Mirarab 2018). This focus on taxa and sequence removal, however, may preclude the inclusion of poorly sampled, rare, or recently extinct taxa from otherwise comprehensive phylogenetic studies.

[1]While I did not explore the supertree *vs.* supermatrix debate, I will note that the decision to use supertrees in this chapter was driven exclusively by the need to compute gene trees for the tree jackknife approach.

In contrast, gene sampling sufficiency is an approach that aims to test whether, given a taxon coverage pattern, there are enough sampled loci in a partitioned matrix to ensure taxon coverage decisiveness, i.e. all subtrees induced by the data partitions have the same parent tree (Sanderson et al. 2010, Steel 2016, Steel & Sanderson 2010), and if not, it yields a minimum number of loci that needs to be sampled to ensure decisiveness (Dobrin et al. 2018, Parvini et al. 2021, Steel 2016). This approach was developed to counteract the formation of terraces during tree inference. These are areas of tree space, more specifically of islands of trees *sensu* Maddison (1991), that are made up of trees with exactly the same optimality score, and which arise (partly) due to non-effective taxonomic overlap between data partitions (Sanderson et al. 2011, 2015). Unfortunately, gene sampling sufficiency assumes uniform random taxon sampling of infinite and equally informative loci (Dobrin et al. 2018, Sanderson et al. 2010), which are either a biological impossibility (infinite loci) or unlikely to be met in empirical datasets (random taxon sampling). One more, indirect, assumption is that there is one unique binary phylogenetic tree for any set of taxa, and thus penalises gene tree-species tree discordance and hard polytomies. Additionally, because gene sampling sufficiency looks at the amount, not the pattern, of overlap, it can lead to instances where sampling is inferred to be insufficient, even though a fully resolved, and reasonably well supported, tree is inferred. This can be seen for the *Ranunculus* dataset, where even though the inferred supertree is binary, gene sampling is insufficient ($\zeta = -5.20$), and 1802 loci need to be sampled for decisiveness (Tables 5.4 and 5.5).

On a more practical level, the high numbers of loci needed for taxon coverage decisiveness (as calculated by gene sampling sufficiency) also runs into the problem of how to increase sampling for hard to collect and rare taxa, even if advances in museomics is making it possible to extract molecular data from previously inaccessible specimens (e.g, Colella et al. 2020, Raxworthy & Smith 2021). The large number of 'sufficient' loci also does not take into account the financial costs of undirected loci-rich sequencing approaches (Zaharias et al. 2020), which might deter work in non-model taxa. A potentially less costly approach to mitigating the effects of non-effective overlap due to missing data is to target specific taxa and/or loci for sampling. Goldman (1998) and Massingham & Goldman (2000) proposed an information-based approach to identify the most informative loci in a dataset, as well as areas of a phylogeny that might benefit from increased taxon sampling (i.e. targeted sampling). However, this method did not gain traction and was seldom used with empirical datasets. An exception being San Mauro et al.'s (2009, 2012) application of Goldman's (1998) method to caecilian phylogenetics; where in the 2009 paper they found the Terasomata split to be a candidate for targeted taxon sampling and that the tRNAs and rRNAs were the most informative loci, which they subsequently tested, and confirmed, in their 2012 study. And, while the likelihood and Bayesian supermatrix analyses of San Mauro et al. (2012, 2014) and Kamei et al. (2012) fully supported the split between Scolecomorphidae and the other terasomatans, my supertree analyses only supported this split after the tree jackknifing analyses, and then only very weakly (QS=50.2%). This suggests that the deep-time terasomatan

divergences are not robust to changes in inference method, although the omission of tRNAs might also have contributed to the decreased support for this split. However, despite San Mauro et al.'s (2009, 2012) encouraging results, Goldman's (1998, 2000) approach need for hypothetical taxa to test which areas of a tree would benefit from increased taxon sampling, while easily manageable for a 9-tip tree (San Mauro et al. 2009), can prove daunting for more taxon-rich datasets. The Concatabominations with tree jackknifing approach, proposed here, aims to be a middle ground between the data-only gene sampling sufficiency approach (Steel 2016) and the experiment-based approach from Goldman (1998), by relying only on the set of inferred gene trees, but having the capacity to identify which of those trees approach decisiveness and which contribute noise to the topology.

Since their publication STR and Concatabominations have been more or less successfully applied to identify unstable taxa in morphological datasets (e.g., Moon 2019, Wilson 2002), and also in supertree studies that employed MR-based methods to infer phylogeny (e.g., Cardillo et al. 2004, Davis & Page 2008), or where MR was used solely for the purpose of rogue taxa identification (Akanni et al. 2015). In these studies, once the unstable taxa were identified, all those taxa that could be safely deleted, i.e. their removal would not affect the inferred relationships between the remaining taxa, were removed from the datasets in order to decrease or even eliminate taxonomic instability. However, as noted above my aim was to decrease taxonomic instability caused by missing data without removing taxa, but also without having to increase loci sampling by multiple orders of magnitude, as suggested by the gene sampling sufficiency results (table 5.4). To achieve this, I adapted the idea behind taxon-jackknifing (Lanyon 1985, see Chapter 3 for example of application) to the MR of trees and Concatabominations. In other words, if we can identify which taxa are adding instability to the matrix by sequential removal of taxa, might we be able to identify which gene trees are adding/decreasing instability by sequential removal of trees from the MR?

As seen from the caecilian analyses (Figs. 5.3 and 5.4, table 5.), Concatabominations with tree jackknifing does indeed identify not only unstable taxa (candidates for increased locus sampling), but also the gene trees that would contribute the most to reduce taxonomic instability by increasing their taxon sampling. In other words, which taxa and gene trees would lead to the biggest increase in effective overlap by adding the least possible data. While taxon coverage might be considered a proxy for how good a gene tree would be at reducing non-effective overlap (underpinning of decisiveness, Sanderson et al. 2010, Steel & Sanderson 2010) were data added to it, the analyses show that that alone is insufficient. The tree jackknifing analyses identified 12S and 16S rRNA, COX1, CYTB, H3A and ND1 as the gene trees where increased taxon sampling would be most effective at reducing non-effective overlap, yet the ND2 and RAG1 trees have greater taxon coverage than the latter two. Furthermore, that the addition of a single nucleotide sequence (16S rRNA for *H. montanus*) resulted in increased topological resolution raises the possibility that, by targetting those taxa missing one or more of the "stabilising" loci, resolution

FIGURE 5.6. Cluster of within genus instability between *Ichthyophis* species identified after jackknifing of the tree for 16S rRNA.

can be further increased at a fraction of the cost required to sequence all taxa for all loci selected for analysis, or for the large k_{min} yielded by the gene sampling sufficiency calculations. This would be especially cost effective when dealing with datasets containing hundreds or thousands of loci (Zaharias et al. 2020). However, this strategy is not guaranteed to result in a fully resolved topology, as the jackknifing is targeting non-effective overlap, not data incongruence resulting from evolutionary history. In fact, resolved trees may be biologically unwarranted if events such as explosive radiations or hybridisation occurred anywhere along the tree (e.g., Olave et al. 2015, Stolzer et al. 2012). Evolutionary histories aside, targeted sampling guided by Concatabominations with tree jackknifing can and does minimise the number of polytomies resulting from poor taxon/sequence sampling.

While data generation was beyond the scope of this study, it is worth noting that additional areas of instability recovered by Concatabominations during tree jackknifing (Fig.5.6), may be indicative of "incomplete" non-effective overlap (i.e. taxa whose overlap is effective for the full complement of sampled loci, but not for certain subsets of loci) and might benefit from ensuring they are sampled for all "stabilising" loci. It is also possible that, because Concatabominations is a heuristic method, not all unstable taxa are identified (Siu-Ting et al. 2015), and that by removing trees from the MR, the compatibility scores change drastically between analyses, and thus the set of identified unstable taxa. As such, studies that aim to increase taxon and/or sequence sampling may benefit from using the output of the jackknifed runs not just to test for changes in overall taxonomic stability, but to find taxa that change from stable to unstable between analyses, and might therefore benefit from the addition of sequence data. And, while the exact relationship between terrace size and taxon coverage in empirical datasets remains elusive (Dobrin et al. 2018, this chapter), directed increases in taxon coverage may also result in datasets with decisive

taxon coverage, thus preventing the occurrence of terraces.

Thus, Concatabominations with tree jackknifing is a promising approach to tackling taxonomic instability caused by non-effective overlap by identifying those taxa that would benefit from increased locus sampling, as well as the best loci to sample. Thus providing a cost-effective, and biologically realistic, alternative to gene sampling sufficiency. However, further work is needed to investigate up to what matrix size (taxa and loci) this approach can be efficiently and/or effectively used.

5.5 Data availability

The data, scripts and results for this chapter are available on the following repository:

- Figshare, `https://doi.org/10.6084/m9.figshare.c.6050033`.

CHAPTER 6

COMPARING BRANCH SUPPORT ACROSS PHYLOGENIES

Some of the treefiles used in this chapter were generated for:

Wilkinson, M. and Serra Silva, A. 'Large island bias in Bayesian phylogenetic inference and bootstrapping', *Systematic Biology*. Invited to resubmit.

6.1 Introduction

Phylogenetic trees inferred from empirical data are seldom reported without branch support values, which often consist of frequentist (e.g., bootstrap (Felsenstein 1985); jackknife (Lanyon 1985)) or posterior predictive (e.g., Bayesian analyses) probability methods. Although rarely used outside parsimony, there are non-probabilistic branch support measures, like Bremer's (1988) decay index, and the strict consensus (SC, Sokal & Rohlf 1981) might also be considered a support measure, albeit one that is only interested in plenary support (Nixon & Carpenter 1996). More important than the method used to calculate branch support is what those values can tell us. Poorly supported branches are often indicators of topological incongruence, which can be due to missing data (e.g., Wiens 2003, Wilkinson 1996), analytical artefacts (e.g., Léveillé-Bourret et al. 2017, Simmons et al. 2022) and/or reticulation events (e.g., Chan et al. 2020, Hahn 2007). However, some of the phenomena that lead to poor branch support can also lead to artefactually high, or inflated, support (e.g., Simmons 2012, Simmons & Freudenstein 2011, Simmons & Norton 2013). Thus, level of support *per se* may not be informative, but, much like the (non-) effective overlap discussed in Chapter 5, the non-random distribution of very poorly (or very highly) supported branches might be indicative of issues in the underlying datamatrices. To my knowledge, branch support has not been approached under this rationale, except for work dealing with long branch attraction (e.g., Huelsenbeck 1997, Lartillot et al. 2007), and even then the

focus was on branch lengths and their causes.

While most parsimony and maximum likelihood inference software include bootstrap analyses, there are multiple algorithms implemented for the standard bootstrap (reviewed in Simmons & Norton 2013), with many software packages offering their own versions of 'fast' bootstrap searches (e.g., rapid bootstrap, Stamatakis et al. (2008); ultrafast bootstrap, Minh et al. (2013)). Because all these implementations have their own analytical artefacts, this can lead to different software packages finding different levels of support for trees inferred from the same matrix. This, combined with the fact that bootstrap results can be projected onto a best likelihood tree (e.g., tree search with multiparametric bootstrap in RAxML (Stamatakis 2014)), a majority rule consensus (MRC, Margush & McMorris 1981) of the best likelihood trees (e.g., IQTree, Nguyen et al. 2015) or of the bootstrap trees (e.g., PAUP* Swofford 2003), means that often branch support has to be compared across trees with different topologies. However, what is often done is to project multiple support values onto a single tree topology (e.g., San Mauro et al. 2009, 2012, Taboada et al. 2020), and then compare topologies based on tree-to-tree distances (e.g., Robinson-Foulds distance (RF), Robinson & Foulds 1981) or visual inspections. Unfortunately, the former can tell us only how different trees are not where they differ, and the latter is increasingly harder to do the larger the trees being compared. For example, comparing the topologies of the caecilian trees inferred in Chapter 5 by eye is feasible, but the same would no longer hold for a Tree of Life of Caudata, the next largest order of amphibians with over 770 valid species (Frost 2022). Thus, a tool that allows for the direct comparison of support values across tree distributions would be highly desirable.

In this chapter, I introduce a permutation-based analysis to test for non-random distribution of probabilistic support values across a tree, and a Python (Van Rossum & Drake 2009) script to extract splits and their associated branch support values from collections of Newick-formatted trees on the same leaf set, and discuss possible extensions to trees on non-identical leaf sets.

6.2 Testing for non-random distribution of branch support values

As mentioned throughout the previous chapters, there are multiple causes for topological uncertainty and the resulting poorly resolved branches, and thus for poor split support. However, as seen in Simmons and Freudenstein (2011) and Simmons and Norton (2013), the phenomena causing poor branch support can also artificially inflate support values. Thus, if we are interested in searching for patterns of non-random distribution of branch support, we must search for both very high and very low support. To do this we must know both the location and magnitude of changes in support. In other words, we want to know if and where there are large differences in support between adjacent branches. We can achieve this by calculating the sum of absolute differences in support between adjacent branches (for all branches, except the root), see figure 6.1 for example. These sums provide the magnitudes of support changes, and we can logically expect

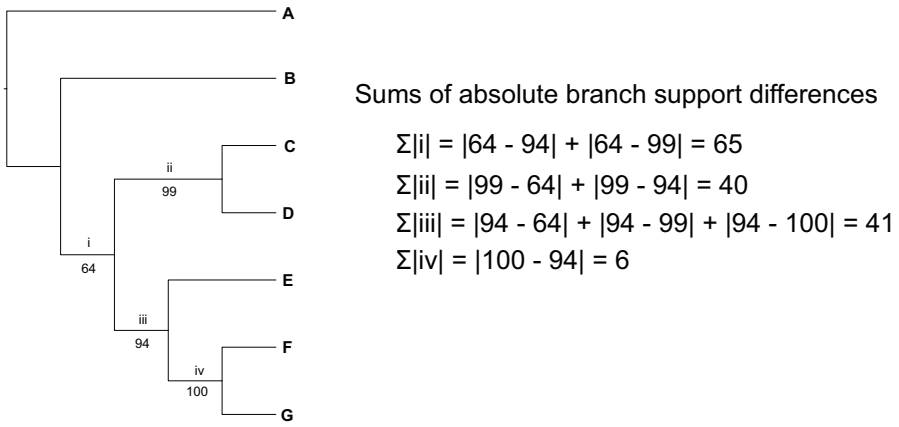


FIGURE 6.1. Example of how the sum of absolute branch support differences for adjacent branches.

that the more similar the support values are across the tree, the closer the sum of magnitudes will get to zero (independent of tree size). Using the magnitude of changes means that we test for both poor and inflated branch support. Permuting the support values and recalculating the sum of magnitudes give us the 'location' of the support changes, i.e. are they clustered, overdispersed or randomly distributed. This magnitude-based (permutation) test of branch support distribution is implemented in the *supportDistribution* script described below.

6.2.1 *supportDistribution* script

Given a binary tree with probabilistic support values and number of permutations (n)
Root trees and set the root's support to 'NA'
Standardise probabilistic support values to $S \in [0, 100]$
Compute patristic distance between all branches and export tab delimited distance matrix
Permute support values (except root's) n times and export tab delimited permutation table
Calculate average branch support across tree
For each branch:
    Identify branch adjacencies from distance matrix
    Calculate the sum of absolute support value differences between adjacent branches
For dispersion test:
    Calculate sum of sums of absolute differences, plot its histogram and save histogram to file
    Calculate z-score
    Calculate one- and two-sided p-values
Calculate number of sum of sums in the permuted data that exactly matches input tree
Write z-score, p-values, average branch support and number of exact matches to text file

The script is implemented in Python v.3.8 (Van Rossum & Drake 2009), and currently requires fully resolved Newick-formatted trees. The dispersion test was based on Martin Hughes' (2015) *Parsimony_conv* R v.3.2.0 (R Core Team 2015) script to test for the non-random distribution of convergent events in a phylogenetic tree (used by the present author in de Almeida Serra Jorge da Silva 2016). Following the original script, the z-scores were multiplied by -1, so that positive scores correspond to overdispersion of large changes in support values, and negative scores to underdispersion (Cooper et al. 2008). The newly added two-sided p-value calculation uses Kulinskaya's (2008) and Kulinskaya and Lewin's (2009) conditional p-value, which works on symmetric and non-symmetric distributions. For symmetric distributions the two-sided p-values will be twice the one-sided p-value. The script has a default number of permutations (999), but users have the ability to select a different number of permutations.

6.2.2 Example

Running two published taxon-rich amphibian trees (Jetz & Pyron 2018, Pyron 2014) shows the possible usefulness of the *supportDistribution* script. The Pyron (2014) tree was inferred, under maximum likelihood, from a matrix with 3310 taxa, a maximum length of 12,809 base pairs (bp), and an average 20% completeness. Given the large amount of missing data, and the presence of branches with bootstrap proportions (BP) < 50%, this is an ideal tree to test for non-random distribution of branch support, and because it was inferred under likelihood it fulfils the script's binary tree requirement. The script found the support in Pyron's (2014) tree to be significantly overdispersed (z-score = 2.478; p-value: one-sided = 0.005 and two-sided = 0.010), a pattern that can be seen in the histogram of the sums of absolute differences for the permuted data (Fig. 6.2a).

The second tree tested, Jetz and Pyron's (2018) tree, was inferred in two stages. First, a tree was inferred under maximum likelihood from a concatenated matrix with 4062 taxa and 15 loci (length of matrix and amount of missing data not provided). Then, 3177 amphibian taxa were added to the inferred backbone based on a taxonomic classification, generating a megaphylogeny *sensu* Smith et al. (2009). Running only the inferred backbone tree through *supportDistribution* again finds the sum of sums of differences to be significantly overdispersed (z-score = 2.690; one-sided = 0.005 and two-sided = 0.010), see figure 6.2b. This result was not unexpected, since the Jetz and Pyron (2018) concatenated matrix built on Pyron's (2014) matrix.

6.3 Comparing probabilistic support across collections of trees

Comparison of branch support across summary trees (consensus or supertrees) often consists of projecting support values onto a single topology (e.g., San Mauro et al. 2009, 2012, 2014), or of side-by-side visual comparison of trees (e.g., caecilian supertrees in Chapter 5). However, for the former strategy different inference software packages and/or optimality criteria can recover

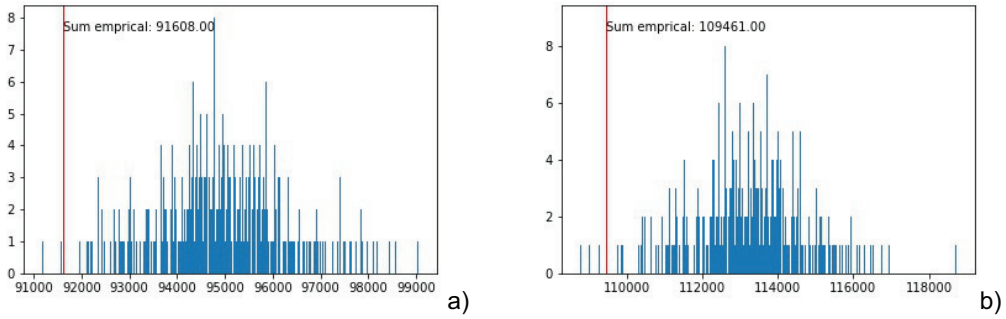


FIGURE 6.2. Histograms resulting from the *supportDistribution* script permutation-based dispersion for a) Pyron's (2014) and b) Jetz and Pyron's (2018) amphibian phylogenies. The red vertical line shows the empirical sum of sums of absolute branch support differences.

different (multi)sets of trees, and thus different summaries. This means that branches that are strongly supported in one summary tree might be absent from another (e.g., Mahony et al. 2022, Roberto et al. 2022). As for the second approach, the more taxa are added to trees the harder they are not only to visualise, but also the harder it is to identify the splits present in the trees and their associated support values. And even though there are a few dedicated tools for side-by-side comparison of large phylogenetic trees (e.g., Phylo.io, Robinson et al. 2016), these are often restricted to pairwise comparisons. It may, thus, be useful to have a tool that allows for the comparison of (multi)sets of summary trees, both the splits that are present and their associated support values.

6.3.1 *splitSupport* script

Given a treefile with a pointer taxon
Create a 'Translation table' (akin to those in Nexus files) with the pointer taxon as 1
(There is a 1:1 correspondence between taxa in a split and taxa in the translation table)
Standardise pointer taxon across all trees and unroot them
Extract slits and attached support from trees:
    If first split:
        Add split to array as the first row and its attached support value to column two
    If split present:
        Add split support to column corresponding to the tree's index
    If new split:
        Add split as new row, fill columns before tree index with zeros
Standardise probabilistic support values to $S \in [0, 100]$
Calculate number and frequency of trees each split is present in
Calculate number (and percentage) of splits present in all trees

Calculate mean, standard deviation, median and mode support for each tree
Save translation, splits and summary statistic tables to tab delimited text file

This small Python 3.8 script extracts the splits present in a collection of trees as a dot-star matrix, their associated support values, and, for probabilistic support measures, calculates a number of summary statistics (see Fig.6.4 for example output). Earlier Python 3 versions can be used, but the script has to be edited to use the *statistics.mode* function rather than *statistics.multimode*, since the latter is only supported from Python v.3.8. Thus, if using earlier Python 3 versions, trees whose support distribution is multimodal will cause the script to abort with an error. The current script accepts only Newick-formatted tree files (rooted or unrooted) and the desired outgroup, or pointer taxon, must be the first taxon on the treefile's opening tree. If a set of trees with no support values is provided the script will generate only the translation table and a table of splits similar to those outputted by PAUP* (Swofford 1991) for bootstrap analyses or the MRC. At present, the script can only work with (multi)sets of trees on the same leaf set and non-probabilistic support values, like decay indices, are treated as if they were proportions/frequencies and any comparisons to probabilistic support measures (and the summary statistics using them) should be interpreted very cautiously.

6.3.2 Example

To illustrate the usefulness of the *splitSupport* script, I will use a set of summary trees inferred from multiple analyses on a data matrix modified from the one used by Wilkinson (1996) to illustrate the effect of unstable taxa (taxon *X*) on branch support measures, table 6.1. The new matrix was modified to encode for two islands of 1-TBR trees (Maddison 1991, Serra Silva & Wilkinson 2021, and Fig.6.3), through the addition of three taxa to one end of the (caterpillar) tree and two characters supporting their placement, but scored as missing for the unstable taxon *X*. In order to generate multiple trees to use as input for the *splitSupport* script, eleven inference and/or support analyses were run on the matrix (order of analysis descriptions is the same as the trees listed in Fig.6.4).

Starting with PAUP* (Swofford 1991), three bootstrap searches were run under default heuristic parsimony settings, with 100 replicates each, and with one analysis set to save a single tree per replicate and the others to save multiple trees ('MulTrees' option). A MrBayes v.3.2.6 (Ronquist et al. 2012) analysis was run under the Mkv+G model, set to two independent runs with four chains for 10 million generations, sampled every 1000 generations with a relative burn-in of 25%. Run convergence was checked using an average standard deviation of split frequencies (ASDSF) ≤ 0.01, and a potential scale reduction factor (PSRF) of 1.00, and the analyses were summarised with the MRC. Three IQ-Tree v.1.6.10 (Nguyen et al. 2015) analyses were run: two nonparametric bootstrap searches, with 100 and 1000 replicates, and an ultrafast bootstrap (Minh et al. 2013) search with 1000 replicates. All analyses were run with the JC2+G+ASC, which is equivalent to the Mkv+G model, and the bootstrap proportions projected onto the MRC. Lastly,

TABLE 6.1. Binary character patterns in the hypothetical data. The data matrix used for analyses contained five copies of each character pattern for a total of 65 characters. All incongruence and uncertainty in these data are concentrated in the single unstable taxon X.

Taxon	Site patterns
O	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
A	1 1 1 1 1 1 0 0 0 0 0 0 0 0 0
B	1 1 1 1 1 0 0 0 0 0 0 0 0 0 0
C	1 1 1 1 0 0 0 0 0 0 0 0 0 0 0
D	1 1 1 0 0 0 0 0 0 0 0 0 0 0 0
E	1 1 0 0 0 0 0 0 0 0 0 0 0 0 0
F	1 0 0 0 0 0 1 0 0 0 0 0 0 0 0
G	1 0 0 0 0 0 1 1 0 0 0 0 0 0 0
H	1 0 0 0 0 0 1 1 1 0 0 0 0 0 0
I	1 0 0 0 0 0 1 1 1 1 0 0 0 0 0
J	1 0 0 0 0 0 1 1 1 1 1 0 0 0 0
K	1 0 0 0 0 0 1 1 1 1 1 1 0 0 0
L	1 0 0 0 0 0 1 1 1 1 1 1 1 0 0
M	1 0 0 0 0 0 1 1 1 1 1 1 1 0 0
X	1 1 1 1 1 1 1 1 1 1 1 1 1 ? ?

four analyses were run on RAxML v.8.2.10 (Stamatakis 2014), two rapid bootstrap searches (Stamatakis et al. 2008) with 100 and 1000 replicates, and two multiparametric bootstrap searches also with 100 and 1000 replicates. All analyses were run under the BIN+G model with the Lewis ascertainment bias correction, and the bootstrap proportions were projected onto the best likelihood tree.

Running the collection of trees resulting from the analyses above through the *splitSupport* script shows they recover trees from the two 1-TBR islands (Figs.6.3 and 6.4), trees 1–3 correspond to the small singleton island and trees 4–11 to trees in the large island. However, from the table of splits and support (Fig.6.4) we can see that for tree 8 (RAxML rapid bootstrap, 100 replicates), the best supported topology is not that onto which the bootstrap proportions have been projected. The *splitSupport* script also shows that tree 4, the tree inferred from the Bayesian analysis, has the highest average support. This shows that unlike most of the other analyses, there is likely to have been extreme island bias during tree search. However, the purpose of this chapter is not to explore the existence of large island bias in bootstrap and Bayesian analyses, rather to introduce tools that allow for comparison of support values across multiple trees, and to test for their non-random distribution in one tree.

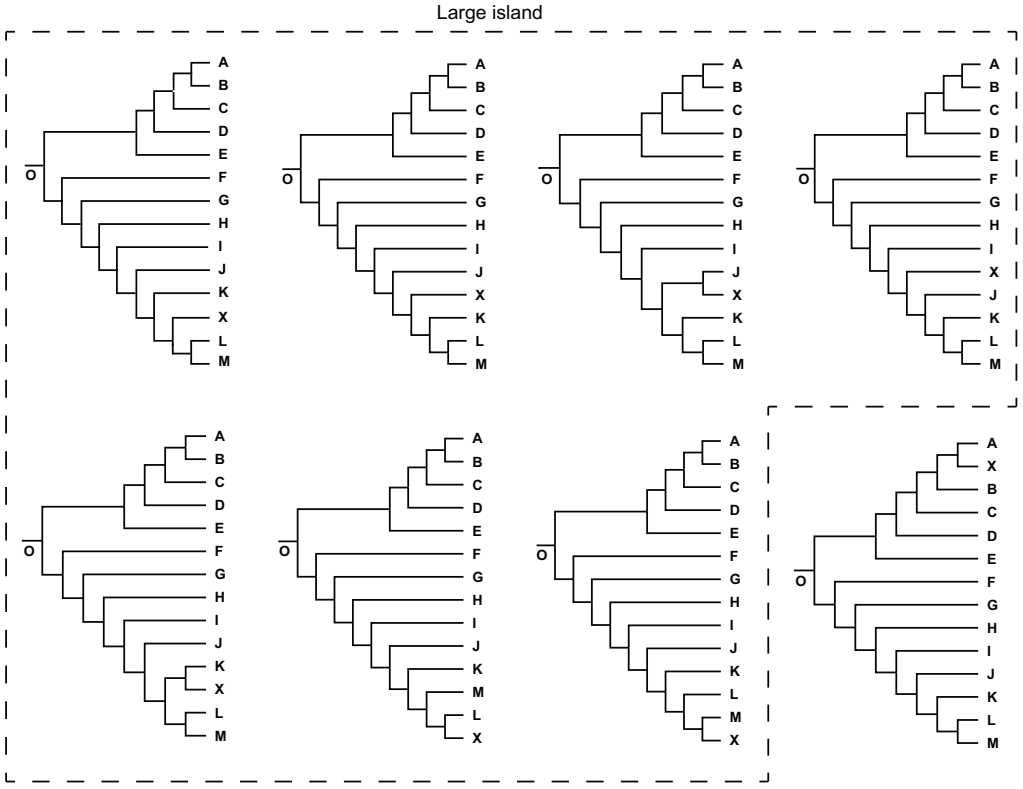
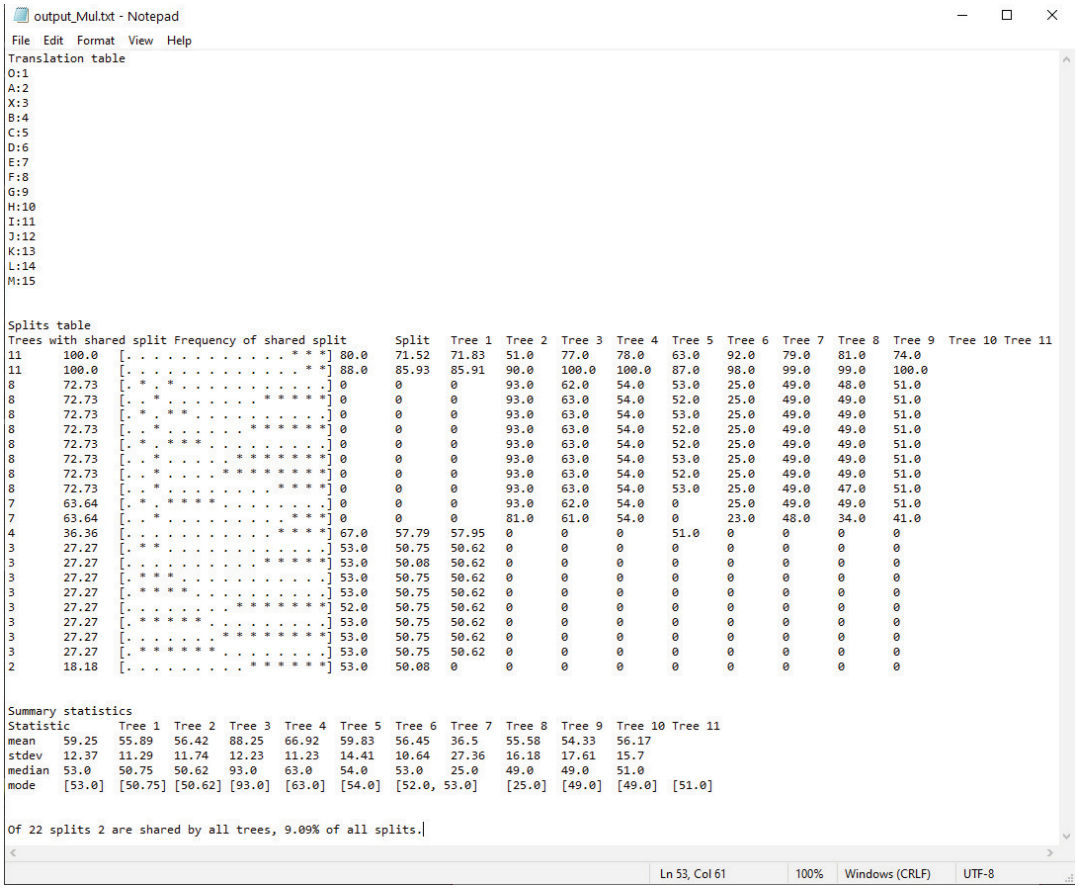


FIGURE 6.3. The eight trees encoded by the hypothetical data matrix and their 1-TBR island structure.

6.4 Discussion

Phylogenetic analyses are often not considered to be complete without the calculation of branch support values, whether they are an integral part of the inference analysis (e.g., posterior predictive probabilities in Bayesian inference) or whether they require additional analyses (e.g., parsimony or likelihood bootstrap searches). These support values, particularly those from resampling analyses, are measures of how robust to perturbation the inferred tree's underlying data are, and are used as guides for how robust any inferred relationship is (Felsenstein 2004). There are, however, many possible causes for the topological instability that leads to low support values, from missing data (e.g., Wiens 2003, Wilkinson 1996) to reticulate events (e.g., Chan et al. 2020, Hahn 2007), often compounded by inference software packages not allowing for hard polytomies (e.g., RAxML). Interestingly, many analytical artefacts, beyond long branch attraction, can lead to inflated support values (e.g., Simmons & Freudenstein 2011, Simmons & Norton 2013). Thus, because both low and high support values can be problematic we may want to focus not on the values alone, but on areas of the tree where drastic shifts in support occur and on how these are distributed across a tree. This is similar to testing the patterns of missing data, i.e. (non-)effective overlap, not its amount.



output_Mul.txt - Notepad

File Edit Format View Help

```
Translation table
O:1
A:2
X:3
B:4
C:5
D:6
E:7
F:8
G:9
H:10
I:11
J:12
K:13
L:14
M:15


Splits table
Trees with shared split Frequency of shared split      Split   Tree 1  Tree 2  Tree 3  Tree 4  Tree 5  Tree 6  Tree 7  Tree 8  Tree 9  Tree 10 Tree 11
11      100.0   [. . . . . * * *] 80.0    71.52   71.83   51.0    77.0    78.0    63.0    92.0    79.0    81.0    74.0
11      100.0   [. . . . . * * *] 88.0    85.93   85.91   90.0    100.0   100.0   87.0    98.0    99.0    99.0    100.0
8       72.73   [. * . * . . . .] 0       0       0       93.0    62.0    54.0    53.0    25.0    49.0    48.0    51.0
8       72.73   [. . * . . . . *] 0       0       0       93.0    63.0    54.0    52.0    25.0    49.0    49.0    51.0
8       72.73   [. * . * * . . .] 0       0       0       93.0    63.0    54.0    53.0    25.0    49.0    49.0    51.0
8       72.73   [. . * . . . . *] 0       0       0       93.0    63.0    54.0    52.0    25.0    49.0    49.0    51.0
8       72.73   [. * . * * * . .] 0       0       0       93.0    63.0    54.0    52.0    25.0    49.0    49.0    51.0
8       72.73   [. . * . . . . *] 0       0       0       93.0    63.0    54.0    53.0    25.0    49.0    49.0    51.0
8       72.73   [. . * . . . . *] 0       0       0       93.0    63.0    54.0    52.0    25.0    49.0    49.0    51.0
8       72.73   [. . * . . . . *] 0       0       0       93.0    63.0    54.0    53.0    25.0    49.0    47.0    51.0
7       63.64   [. * . * * * . .] 0       0       0       93.0    62.0    54.0    0       25.0    49.0    49.0    51.0
7       63.64   [. . * . . . . *] 0       0       0       81.0    61.0    54.0    0       23.0    48.0    34.0    41.0
4       36.36   [. . . . . * * *] 67.0    57.79   57.95   0       0       0       51.0    0       0       0       0
3       27.27   [. * * . . . . .] 53.0    50.75   50.62   0       0       0       0       0       0       0       0
3       27.27   [. . . . . * * *] 53.0    50.08   50.62   0       0       0       0       0       0       0       0
3       27.27   [. * * * . . . .] 53.0    50.75   50.62   0       0       0       0       0       0       0       0
3       27.27   [. * * * * . . .] 53.0    50.75   50.62   0       0       0       0       0       0       0       0
3       27.27   [. . . . . * * *] 52.0    50.75   50.62   0       0       0       0       0       0       0       0
3       27.27   [. * * * * * . .] 53.0    50.75   50.62   0       0       0       0       0       0       0       0
3       27.27   [. . . . . * * *] 53.0    50.75   50.62   0       0       0       0       0       0       0       0
3       27.27   [. * * * * * . .] 53.0    50.75   50.62   0       0       0       0       0       0       0       0
2       18.18   [. . . . . * * *] 53.0    50.08   0       0       0       0       0       0       0       0       0


Summary statistics
Statistic       Tree 1  Tree 2  Tree 3  Tree 4  Tree 5  Tree 6  Tree 7  Tree 8  Tree 9  Tree 10 Tree 11
mean    59.25   55.89   56.42   88.25   66.92   59.83   56.45   36.5    55.58   54.33   56.17
stdev   12.37   11.29   11.74   12.23   11.23   14.41   10.64   27.36   16.18   17.61   15.7
median  53.0    50.75   50.62   93.0    63.0    54.0    53.0    25.0    49.0    49.0    51.0
mode    [53.0]  [50.75] [50.62] [93.0]  [63.0]  [54.0]  [52.0, 53.0]    [25.0]  [49.0]  [49.0]  [51.0]


Of 22 splits 2 are shared by all trees, 9.09% of all splits.
```

Ln 53, Col 61 100% Windows (CRLF) UTF-8

FIGURE 6.4. Example output for the *splitSupport* script using the summary trees of the (multi)sets inferred from the table 6.1 data matrix.

The distribution of shifts in branch support, though not their direction, can be tested with the *supportDistribution* script. The results for the two tested phylogenies, Pyron's (2014) and Jetz and Pyron's (2018) amphibian trees, found their shifts in support values to be significantly overdispersed. This means that there are multiple areas of the tree where poorly and highly supported relationships are in close proximity, but these are not clustered, as might be expected if shifts to low support were due to, for example, incomplete lineage sorting (ILS). The shifts are, however, more broadly distributed than might be expected from chance alone. While it is clear that significant clustering of shifts in support value reduces confidence on the inferred trees, since the clustering is likely linked to either extensive missing data or ILS/reticulate events, the causes for large absolute differences in support values are not clear, and might be linked to the number of shifts not their underlying causes. Thus, while the identification of significantly clustered support value shifts warrants a thorough revision of the data matrix, further work is needed to understand what causes overdispersion of large changes in support values.

A major methodological gap when dealing with support measures is how to compare them

across multiple trees on the same leaf set. While there are tree visualisers like Phylo.io (Robinson et al. 2016) that allow for the direct comparison of phylogenetic trees, they are often restricted to two input trees. With trees of increasingly larger leaf sets being inferred, visual comparison between trees, particularly branch topology and associated support values, becomes increasingly more challenging. To mitigate this challenge, the *splitSupport* script prints a list of the splits present in all input trees (on the same leaf set) and the support for each split in all trees. From figure 6.4 it is clear that two distinct topologies were inferred from the hypothetical data matrix in table 6.1. While this was expected *a priori* for this dataset, if other sets of input trees display similar trends of inferring more than one topology, it may be indicative of the presence of islands and the user would be advised to check the inferred (multi)sets for island structure. Additionally, the script also identifies two trees with very different support values from the other trees in the set. First, the Bayesian analysis has an average branch support of 88.25%, which is 20–30% higher than most other trees, which is not unprecedented given that multiple studies have found that posterior probabilities are often higher than bootstrap proportions for trees inferred from the same matrix (e.g., Simmons et al. 2004, Suzuki et al. 2002). It does, however, call into question the results of the Bayesian analysis, particularly because the dataset encodes for two 1-TBR islands.

The other tree with very different support levels is tree 8, which was inferred with RAxML and the support values were obtained with a rapid bootstrap search. Contrary to the Bayesian analysis, this tree displays lower average support ($\overline{BP}$ = 36.5%) than the other trees, which is probably due to the low number of bootstrap replicates (100) and stochasticity. Because in RAxML, bootstrap proportions are projected onto the best likelihood tree, the most likely explanation for the support values on tree 8 is that the bootstrap favours trees from the singleton island, whereas the tree search identified a tree from the largest island as the optimal tree. This discrepancy between tree and bootstrap searches can be explained by how the rapid bootstrap algorithm works. First, each bootstrap replicate uses the tree found by the previous replicate as a starting tree, and every tenth bootstrap uses a starting tree that is encoded by the original alignment (Stamatakis et al. 2008). Thus, if the tree from the singleton island is drawn from the original alignment more often than a tree from the largest island, the bootstrap distribution can be skewed toward the smallest island simply due to stochasticity. As such, by printing the splits and support values for multiple trees side-by-side the *splitSupport* script allows for the identification of major topologies found by distinct inference analyses, and if trees with very different support values are present, it can direct users to the trees that should be reanalysed or whose analytical parameters/settings need to be revised/rethought. And, while the script outputs only summary statistics, the split support table is exported in a format where it can be extracted and manipulated for further statistical tests.

Despite its current utility, the *splitSupport* script might be further modified to deal with partially resolved trees and with trees on non-identical leaf sets, as it is sometimes necessary

to compare topological and support changes between trees inferred before and after increased taxon sampling. For the former, it will first be necessary to choose whether to treat polytomies as hard (real multifurcations) or soft (topological uncertainty). If dealing with hard polytomies, we might want to test for the non-random distribution of support values and polytomy occurrence, with the latter being achieved by permuting the number of children any branch/node has. Soft polytomies, on the other hand would need to be expanded and a support of 33.33% assigned to each possible resolution, in order to compare them to any other trees where one of the resolutions is present. As for comparing trees on non-identical leaf sets, while there is a preliminary script to extract partial splits, these cannot yet be compared except to search for exact matches, and the extracted support values are thus not necessarily informative. In order to compare partial splits, it will be necessary to implement a pruning, or grafting, step to convert all pairs of trees into trees on the same leaf set (e.g., Bansal 2020, Wilkinson et al. 2005), and from there compare branch support values. However, if grafting is chosen over pruning, branch support values will either need to be imputed, or a classification system will be required, similar to those proposed by Bininda-Emonds (2003) and Wilkinson et al. (2005), to allow for the comparison of branches present in some, but not all, trees.

In summary, while the scripts introduced in this chapter address a methodological gap (testing for non-random distribution of support values, and comparing support values across multiple trees with the same leaf set), they can be further expanded to polytomous trees and sets of trees with non-identical leaf sets, but careful consideration is required to achieve these modifications.

6.5 Data and software availability

The data, scripts and results for this chapter are available on the following repository:

- Figshare, `https://doi.org/10.6084/m9.figshare.c.6050033`.

CHAPTER 7

CONCLUSION AND FUTURE WORK

7.1 General conclusion

With this thesis I attempted to tackle two problems: how to identify and summarise tree (multi)set heterogeneity, and how to identify and mitigate topological incongruence caused by non-effective overlap. While I explored, generalised, and proposed methods to deal with these topics, in the age-old tradition of scientific enquiry, new questions arose with each step forward (or sideways). What I hope was made clear in this work is that, although there is no perfect data and evolutionary phenomena can and do lead to topological incongruence, by trying to understand the patterns of incongruence and their causes, we might eventually be able to efficiently summarise these patterns. We might even one day be capable of restricting all inferred topological incongruence to that caused by processes of evolution, without having to worry about incongruence caused by missing data and analytical artefacts. To be able to do so, however, further work on the topics covered here is needed. Below I leave some of the new questions, and possible research avenues, that arose from the work detailed in the previous chapters.

7.2 Future work

7.2.1 Islands and clumps of trees

In Chapters 2 and 4, I briefly mentioned that the development of generalised tree-to-tree distances, which account for minimal or no overlap between leaf sets, might allow us to extend islands of trees (Maddison 1991, Serra Silva & Wilkinson 2021) to the supertree context, and the use of tree-to-tree distances for internally labelled trees means that islands and clumps might be applicable in cancer phylogenetics. There are, however, many more questions/applications to be

explored. Starting with islands, the work in this thesis focused only on tree sets inferred from morphological data (and one hypothetical dataset, Chapter 6), but how prevalent are islands throughout empirical datasets? Are they more common in parsimony, likelihood or Bayesian searches? Do they, in fact strongly affect resampling analyses, and thus support measures? Is any commonly used bootstrap implementation particularly prone to island bias? Beyond analytical aspects, we may want to explore going one step further than multidimensional scaling (MDS) plots to depict island structure, and instead project islands in three-dimensions (3-D). These 3-D plots might identify gaps in the sampled topologies, particularly when dealing exclusively with (multi)sets of optimal trees. Also, while the concepts of island size, mass and density were defined in Chapter 2, other measures of between and within island dissimilarity might be devised. For example, island diameter, or the greatest within-island tree-to-tree distance, might be a proxy for how many areas of local instability occur in any given island.

As for clumps, beyond the work necessary to establish what the best analytical parameters are, simulation studies, where the causes of topological incongruence are known, might give a better understanding of the biological information being relayed by the set of identified clumps. Might a visualising tool with the capacity to plot within and between clump distances, which does not require imputation of the distance between trees with minimal/no overlap, be developed to project the tree space occupied by (multi)sets of trees with non-identical leaf sets in two- or three-dimensions? How might we measure within and between clump dissimilarity, other than relying on comparisons of clump supertree topologies? Much like islands, there is a breadth of possibilities to explore what information clumps relay, their prevalence in phylogenomic data, etc.

7.2.2 The *Chinlestegophis* conundrum

As mentioned in Chapter 3, a thorough revision of character construction and scoring, of both the Pardo et al. (2017) and Schoch et al. (2020) matrices, is warranted. Particularly in regards to the logically and biologically dependent characters present in both matrices. However, given the uncertainty surrounding the relationships between tetrapod lineages (reviewed in Marjanović & Laurin 2019), we may expect that this will be done incrementally as more fossils are found, which can either reduce the amount of missing data in the existing matrices or increase taxon sampling.

7.2.3 Mitigating non-effective overlap in molecular datasets

Further work on the Concatabominations with tree jackknifing approach should focus, first, on testing it with other empirical datasets, and also on testing the hypothesis that sampling taxa for as many 'stabilising' loci as possible will decrease the overall incongruence caused by non-effective overlap. Another avenue might be to test whether matrix representations (MR) other than MR of splits (reviewed in Wilkinson et al. 2004) might be used to yield smaller Concatabominations-compatible input matrices, while retaining the capacity to identify instances of non-effective

overlap. If such MRs can be encoded, the smaller matrices would lower the computational requirements on the Concatabominations pipeline and the tree jackknifing approach might be efficiently run on matrices larger than those tested in Chapter 5.

7.2.4 Comparison of branch support between trees with non-identical leaf sets

Lastly, beyond the generalisation to trees on non-identical leafs sets discussed in Chapter 6, further work on branch support comparison across trees might focus on exploring how well the existing scripts perform with trees inferred from empirical datasets and on optimising the runtimes of the existing scripts, thus increasing their usefulness.

APPENDIX A

List of Abbreviations

Abbreviation	Meaning	Page
c.	*circa*: approximately	
e.g.	*exempli gratia*: for example	
i.e.	*id est*: that is	
vs.	*versus*	
ILS	Incomplete lineage sorting	2
MPT	Most parsimonious tree	9
RF	Robinson-Foulds tree-to-tree distance	11
NNI	Nearest-neighbour interchange	11
SPR	Subtree prune-regrafting	11
TBR	Tree bisection-reconnection	11
QD	Quartet distance	22
MDS	Multidimensional scaling	24
BIC	Bayesian Information Criterion	25
NJ	Neighbour-joining	26
UPGMA	unweighted pair group method with arithmetic mean	26
HGT	Horizontal gene transfer	28
LS	Leaf stability	31
RBIC	Relative bipartition information content	32
MRC	Majority-rule consensus	32
STR	Safe taxonomic reduction	32
SC	Strict consensus	33
RNR	RogueNaRok	33
LS_{max}	Maximal leaf stability	33
LS_{diff}	Difference between highest split frequencies leaf stability	33
LS_{ent}	Entropic leaf stability	33
RC	Reduced consensus	36

Abbreviation	Meaning	Page
SMM	Structured Markov model	60
HMM	Hidden Markov model	60
NA	Non-applicable	64
AHE	Anchored hybrid enrichment	69
UCE	Ultraconserved element	69
KST	Siu-Ting et al. (2019)	71
uRF	Uncorrected Robinson-Foulds distance	72
wRF	Weighted Robinson-Foulds distance	72
ST_{all}	Supertree of all input trees	73
MAST	Maximum agreement subtree	79
MAD	Minimal ancestor deviation	80
MR	Matrix representation	85
nucDNA	Nuclear DNA	87
mtDNA	Mitochondrial DNA	87
bp	Base pairs	87
ASDSF	Average standard deviation of split frequencies	87
PSRF	Potential scale reduction factor	87
QS	Quartet support	101
BP	Bootstrap proportion	120

APPENDIX B

LIST OF SYMBOLS

Symbol	Usage	Meaning	Page
$=$	$x = y$	x equals y	
$\sim$	$\sim x$	Approximately x	
$\approx$	$x \approx y$	x is approximately equal to y	
$>$	$x > y$	x is greater than y	
$<$	$x < y$	x is less than y	
$\geq$	$x \geq y$	x is greater than or equal to y	
$\leq$	$x \leq y$	x is less than or equal to y	
$\lvert x \rvert$		The absolute value of x	
$\sum$	$\sum x$	Sum of all values of x	
$\overline{x}$		Average of x	
$\in$	$x \in S$	x is an element of set S	
^{C}	S^C	The complement of set S	
$\cap$	$X \cap Y$	The intersection of sets X and Y	
$\rightarrow$	$f : A \rightarrow B$	Function with domain A and codomain B, read as: f from A to B	
$\mathbb{R}_0^+$		The set of non-negative real numbers	
ρ		Tree resolution (ratio of observed to theoretical branches in a tree)	4
Δ	Δx	Change in x	44
ζ		Gene sampling sufficiency	100
k_{min}		Theoretical minimum number of loci to sample	100
d		Taxon coverage density	100

BIBLIOGRAPHY

Abel, P. & Werneburg, I. (2021), 'Morphology of the temporal skull region in tetrapods: research history, functional explanations, and a new comprehensive classification scheme', *Biological Reviews* **96**(5), 2229–2257.

Aberer, A. J., Krompass, D. & Stamatakis, A. (2012), 'Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice', *Systematic Biology* **62**(1), 162–166.

Acosta-Galvis, A. R., Torres, M. & Pulido-Santacruz, P. (2019), 'A new species of *Caecilia* (Gymnophiona, Caeciliidae) from the Magdalena valley region of Colombia', *ZooKeys* **884**, 135–157.

Adams, E. N. (1986), 'N-trees as nestings: complexity, similarity, and consensus', *Journal of Classification* **3**(2), 299–317.

Agnolin, F. L., Motta, M. J., Brissón Egli, F., Lo Coco, G. & Novas, F. E. (2019), 'Paravian phylogeny and the dinosaur-bird transition: an overview', *Frontiers in Earth Science* **6**, 252.

Aguse, N., Qi, Y. & El-Kebir, M. (2019), 'Summarizing the solution space in tumor phylogeny inference by multiple consensus trees', *Bioinformatics* **35**(14), i408–i416.

Akanni, W. A., Siu-Ting, K., Creevey, C. J., McInerney, J. O., Wilkinson, M., Foster, P. G. & Pisani, D. (2015), 'Horizontal gene flow from Eubacteria to Archaebacteria and what it means for our understanding of eukaryogenesis', *Philosophical Transactions of the Royal Society B: Biological Sciences* **370**(1678), 20140337.

Alexander, A. M., Su, Y.-C., Oliveros, C. H., Olson, K. V., Travers, S. L. & Brown, R. M. (2017), 'Genomic data reveals potential for hybridization, introgression, and incomplete lineage sorting to confound phylogenetic relationships in an adaptive radiation of narrow-mouth frogs', *Evolution* **71**(2), 475–488.

Allen, B. L. & Steel, M. (2001), 'Subtree transfer operations and their induced metrics on evolutionary trees', *Annals of Combinatorics* **5**(1), 1–15.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990), 'Basic local alignment search tool', *Journal of Molecular Biology* **215**(3), 403–410.

Amir, A. & Keselman, D. (1997), 'Maximum agreement subtree in a set of evolutionary trees: metrics and efficient algorithms', *SIAM Journal on Computing* **26**(6), 1656–1669.

AmphibiaWeb (2019), *AmphibiaWeb: Information on amphibian biology and conservation*, AmphibiaWeb. Available from: `https://amphibiaweb.org/index.html` (Accessed 15 March 2022)

Anderson, J. S., Reisz, R. R., Scott, D., Fröbisch, N. B. & Sumida, S. S. (2008), 'A stem batrachian from the Early Permian of Texas and the origin of frogs and salamanders', *Nature* **453**(7194), 515–518.

Arora, P., Deepali & Varshney, S. (2016), 'Analysis of k-means and k-medoids algorithm for big data', *Procedia Computer Science* **78**, 507–512.

Assenov, Y., Ramírez, F., Schelhorn, S.-E., Lengauer, T. & Albrecht, M. (2007), 'Computing topological parameters of biological networks', *Bioinformatics* **24**(2), 282–284.

Balakrishnan, V. (1997), *Schaum's Outline of Graph Theory: Including Hundreds of Solved Problems*, McGraw Hill, London.

Bansal, M. S. (2020), 'Linear-time algorithms for phylogenetic tree completion under Robinson-Foulds distance', *Algorithms for Molecular Biology* **15**(1), 1–15.

Baum, B. R. & Ragan, M. A. (2004), 'The MRP method'. In Bininda-Emonds,O. R. P. (ed.), *Phylogenetic supertrees: combining information to reveal the tree of life*', Kluwver Academic Publisher, Dordrecht, pp. 17–34.

Bayzid, M. S., Mirarab, S., Boussau, B. & Warnow, T. (2015), 'Weighted statistical binning: enabling statistically consistent genome-scale phylogenetic analyses', *PLoS One* **10**(6), e0129183.

Beaulieu, J. M. & O'Meara, B. C. (2014), 'Hidden Markov models for studying the evolution of binary morphological characters'. In Garamszegi, L. Z. (ed.) *Modern phylogenetic comparative methods and their application in evolutionary biology*, Springer, Berlin, Heidelberg, pp. 395–408.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. (2008), 'Genbank', *Nucleic Acids Research* **37**(suppl_1), D26–D31.

Berry, V., Bininda-Emonds, O. R. P. & Semple, C. (2012), 'Amalgamating source trees with different taxonomic levels', *Systematic Biology* **62**(2), 231–249.

Bininda-Emonds, O. R. (2003), 'Novel versus unsupported clades: assessing the qualitative support for clades in MRP supertrees', *Systematic Biology* **52**(6), 839–848.

Bininda-Emonds, O. R., Cardillo, M., Jones, K. E., MacPhee, R. D., Beck, R., Grenyer, R., Price, S. A., Vos, R. A., Gittleman, J. L. & Purvis, A. (2007), 'The delayed rise of present-day mammals', *Nature* **446**(7135), 507–512.

Bonnard, C., Berry, V. & Lartillot, N. (2006), 'Multipolar consensus for phylogenetic trees', *Systematic Biology* **55**(5), 837–843.

Bordewich, M. & Semple, C. (2005), 'On the computational complexity of the rooted subtree prune and regraft distance', *Annals of Combinatorics* **8**(4), 409–423.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., 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).

Brazeau, M. D., Guillerme, T. & Smith, M. R. (2019), 'An algorithm for morphological phylogenetic analysis with inapplicable data', *Systematic Biology* **68**(4), 619–631.

Brazeau, M. & Desjardins, C. (2020), 'Morphy'.

Bremer, K. (1988), 'The limits of amino acid sequence data in angiosperm phylogenetic reconstruction', *Evolution* **42**(4), 795–803.

Brown, E. K. & Day, W. H. E. (1984), 'A computationally efficient approximation to the nearest neighbor interchange metric', *Journal of Classification* **1**(1), 93–124.

Bryant, D. (2003), 'A classification of consensus methods for phylogenetics', In Janowitz, M. F., Lapointe, F.-J., McMorris, F. R., Mirkin, B. & Roberts, F.S. (eds.), *Bioconsensus: DIMACS series in discrete mathematics and theoretical computer science*, Vol. 61, American Mathematical Society, Providence, Rhode Island, **61**, 163–184.

Bryant, D. (2004), 'The splits in the neighborhood of a tree', *Annals of Combinatorics* **8**(1), 1–11.

Cai, L., Xi, Z., Lemmon, E. M., Lemmon, A. R., Mast, A., Buddenhagen, C. E., Liu, L. & Davis, C. C. (2021), 'The perfect storm: gene tree estimation error, incomplete lineage sorting, and ancient gene flow explain the most recalcitrant ancient angiosperm clade, Malpighiales', *Systematic Biology* **70**(3), 491–507.

Cannatella, D. C., Hillis, D. M., Chippindale, P. T., Weigt, L., Rand, A. S. & Ryan, M. J. (1998), 'Phylogeny of frogs of the *Physalaemus pustulosus* species group, with an examination of data incongruence', *Systematic Biology* **47**(2), 311–335.

Caparros, M. & Prat, S. (2021), 'A phylogenetic networks perspective on reticulate human evolution', *iScience* **24**(4), 102359.

Cardillo, M., Bininda-Emonds, O. R., Boakes, E. & Purvis, A. (2004), 'A species-level phylogenetic supertree of marsupials', *Journal of Zoology* **264**(1), 11–31.

Castresana, J. (2002), 'Gblocks: selection of conserved blocks from multiple alignments for their use in phylogenetic analysis', *Version 0.91 b. Copyrighted by J. Castresana, EMBL*.

Chan, K. O., Hutter, C. R., Wood Jr, P. L., Grismer, L. L. & Brown, R. M. (2020), 'Target-capture phylogenomics provide insights on gene and species tree discordances in Old World treefrogs (Anura: Rhacophoridae)', *Proceedings of the Royal Society B: Biological Sciences* **287**(1940), 20202102.

Chauve, C., Jones, M., Lafond, M., Scornavacca, C. & Weller, M. (2017), 'Constructing a consensus phylogeny from a leaf-removal distance' (Extended abstract). In Fici, G., Sciortino, M. & Venturini, R. (eds.) *International symposium on string processing and information retrieval*, Springer, Cham, pp. 129–143.

Chernomor, O., Minh, B. Q. & von Haeseler, A. (2015), 'Consequences of common topological rearrangements for partition trees in phylogenomic inference', *Journal of Computational Biology* **22**(12), 1129–1142.

Coiffard, C., Mohr, B. A. & Bernardes-de Oliveira, M. E. (2013), '*Jaguariba wiersemana* gen. nov. et sp. nov., an Early Cretaceous member of crown group Nymphaeales (Nymphaeaceae) from northern Gondwana', *Taxon* **62**(1), 141–151.

Colella, J. P., Tigano, A. & MacManes, M. D. (2020), 'A linked-read approach to museomics: higher quality de novo genome assemblies from degraded tissues', *Molecular Ecology Resources* **20**(4), 856–870.

Cooper, N., Rodríguez, J. & Purvis, A. (2008), 'A common tendency for phylogenetic overdispersion in mammalian assemblages', *Proceedings of the Royal Society B: Biological Sciences* **275**(1646), 2031–2037.

Cotton, J. A. & Wilkinson, M. (2007), 'Majority-rule supertrees', *Systematic Biology* **56**(3), 445–52.

Darlu, P. & Guénoche, A. (2011), 'TreeOfTrees method to evaluate the congruence between gene trees', *Journal of Classification* **28**(3), 390–403.

Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. (2012), 'jModelTest 2: more models, new heuristics and parallel computing', *Nature Methods* **9**(8), 772–772.

Darwin, C. (1859), *On the origin of species by means of natural selection, or, The preservation of favoured races in the struggle for life*, first edn, London : John Murray, London.

DasGupta, B., He, X., Jiang, T., Li, M., Tromp, J. & Zhang, L. (2000), On computing the nearest neighbor interchange distance. In Du, D.-Z., Pardalos, P. M. & Wang, J. (eds.) *Discrete mathematical probelms with medical applications: DIMACS series in discrete mathematics and theoretical computer science*, Vol. 55, American Mathematical Society, Providence, Rhode Island, pp. 125–142.

Davis, K. E. & Page, R. D. (2008), 'Reweaving the tapestry: a supertree of birds', *PLoS Currents* **6**.

de Almeida Serra Jorge da Silva, A. L. (2016), 'Convergent evolution in Anguilliform elongation', Master's thesis, Department of Life Sciences, Imperial College London, London.

De Laet, J. (2015), 'Parsimony analysis of unaligned sequence data: maximization of homology and minimization of homoplasy, not minimization of operationally defined total cost or minimization of equally weighted transformations', *Cladistics* **31**(5), 550–567.

de Lamarck, J.-B. d. M. (1809), *Philosophie zoologique, ou Exposition des considérations relatives à l'histoire naturelle des animaux...*, Vol. 1, Paris: Dentu.

de Queiroz, A. & Gatesy, J. (2007), 'The supermatrix approach to systematics', *Trends in Ecology & Evolution* **22**(1), 34–41.

Deepak, A., Fernández-Baca, D., Tirthapura, S., Sanderson, M. J. & McMahon, M. M. (2014), 'EvoMiner: frequent subtree mining in phylogenetic databases', *Knowledge and Information Systems* **41**(3), 559–590.

Degnan, J. H. & Rosenberg, N. A. (2009), 'Gene tree discordance, phylogenetic inference and the multispecies coalescent', *Trends in Ecology & Evolution* **24**(6), 332–340.

Demchak, B., Hull, T., Reich, M., Liefeld, T., Smoot, M., Ideker, T. & Mesirov, J. P. (2014), 'Cytoscape: the network visualization tool for genomespace workflows', *F1000Research* **3**.

Dobrin, B. H., Zwickl, D. J. & Sanderson, M. J. (2018), 'The prevalence of terraced treescapes in analyses of phylogenetic data sets', *BMC Evolutionary Biology* **18**(1), 46.

dos Santos, C. M. & Falaschi, R. L. (2007), 'Missing data in phylogenetic analysis: comments on support measures', *Darwiniana* **45**(Sup), 25–26.

Dubois, A., Ohler, A. & Pyron, R. A. (2021), 'New concepts and methods for phylogenetic taxonomy and nomenclature in zoology, exemplified by a new ranked cladonomy of recent amphibians (Lissamphibia)', *Megataxa* **5**(1), 1–738.

Elworth, R. A. L., Ogilvie, H. A., Zhu, J. & Nakhleh, L. (2019), 'Advances in computational methods for phylogenetic networks in the presence of hybridization'. In Warnow, T. (ed.) *Bioinformatics and phylogenetics* , Springer, Cham, pp. 317–360.

Estabrook, G. F., McMorris, F. & Meacham, C. A. (1985), 'Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units', *Systematic Zoology* **34**(2), 193–200.

Farris, J. S., Albert, V. A., Källersjö, M., Lipscomb, D. & Kluge, A. G. (1996), 'Parsimony jackknifing outperforms neighbor-joining', *Cladistics* **12**(2), 99–124.

Felsenstein, J. (1985), 'Confidence limits on phylogenies: an approach using the bootstrap', *Evolution* **39**(4), 783–791.

Felsenstein, J. (2004), *Inferring phylogenies*, Sinauer Associates, Sunderland, Massachusetts.

Finden, C. & Gordon, A. (1985), 'Obtaining common pruned trees', *Journal of Classification* **2**(1), 255–276.

Forey, P. L. & Kitching, I. J. (2000), 'Experiments in coding multistate characters'. In Scotland, R. & Toby Pennington, R. (eds.) *Homology and systematics: coding characters for phylogenetic analysis*, CRC Press, London, pp. 54–80.

Foster, P. G. (2004), 'Modeling compositional heterogeneity', *Systematic Biology* **53**(3), 485–495.

Frost, D. R. (2022), 'Amphibian species of the world: an online reference, version 6.0'. Available from: `http://research.amnh.org/vz/herpetology/amphibia` (Accessed 15 March 2022)

Frost, D. R., Grant, T., Faivovich, J., Bain, R. H., Haas, A., Haddad, C. F. B., De SÁ, R. O., Channing, A., Wilkinson, M., Donnellan, S. C., Raxworthy, C. J., Campbell, J. A., Blotto, B. L., Moler, P., Drewes, R. C., Nussbaum, R. A., Lynch, J. D., Green, D. M. & Wheeler, W. C. (2006), 'The amphibian tree of life', *Bulletin of the American Museum of Natural History* **297**, 1–291.

Fürbringer, M. (1888), *Untersuchungen zur Morphologie und Systematik der Vögel: zugleich ein Beitrag zur Anatomie der Stütz-und Bewegungsorgane*, Vol. 15, T. van Holkema, Amsterdam.

Gardner, J. D. (2001), 'Monophyly and affinities of albanerpetontid amphibians (Temnospondyli; Lissamphibia)', *Zoological Journal of the Linnean Society* **131**(3), 309–352.

Gatesy, J., Matthee, C., DeSalle, R. & Hayashi, C. (2002), 'Resolution of a supertree/supermatrix paradox', *Systematic Biology* **51**(4), 652–64.

Goldman, N. (1998), 'Phylogenetic information and experimental design in molecular systematics', *Proceedings of the Royal Society of London. Series B: Biological Sciences* **265**(1407), 1779–1786.

Goldman, N., Anderson, J. P. & Rodrigo, A. G. (2000), 'Likelihood-based tests of topologies in phylogenetics', *Systematic Biology* **49**(4), 652–670.

Goloboff, P. A. (2008), 'Calculating SPR distances between trees', *Cladistics* **24**(4), 591–597.

Goloboff, P. A. & Arias, J. S. (2019), 'Likelihood approximations of implied weights parsimony can be selected over the Mk model by the Akaike information criterion', *Cladistics* **35**(6), 695–716.

Goloboff, P. A., De Laet, J., Ríos-Tamayo, D. & Szumik, C. A. (2021), 'A reconsideration of inapplicable characters, and an approximation with step-matrix recoding', *Cladistics* **37**(5), 596–629.

Goloboff, P. A., Pittman, M., Pol, D. & Xu, X. (2019), 'Morphological data sets fit a common mechanism much more poorly than DNA sequences and call into question the Mkv model', *Systematic Biology* **68**(3), 494–504.

Goloboff, P. A., Torres, A. & Arias, J. S. (2017), 'Weighted parsimony outperforms other methods of phylogenetic inference under models appropriate for morphology', *Cladistics* .

Gordon, A. D. (1986), 'Consensus supertrees: the synthesis of rooted trees containing overlapping sets of labeled leaves', *Journal of Classification* **3**(2), 335–348.

Govek, K., Sikes, C. & Oesper, L. (2018), 'A consensus approach to infer tumor evolutionary histories'. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, Association for Computing Machinery, New York, New York, pp. 63–72.

Graham, S. W., Kohn, J. R., Morton, B. R., Eckenwalder, J. E. & Barrett, S. C. H. (1998), 'Phylogenetic congruence and discordance among one morphological and three molecular data sets from Pontederiaceae', *Systematic Biology* **47**(4), 545–567.

Guénoche, A. (2013), 'Multiple consensus trees: a method to separate divergent genes', *BMC Bioinformatics* **14**(1), 46.

Gunnell, G. F., Smith, R. & Smith, T. (2017), '33 million year old *Myotis* (Chiroptera, Vespertilionidae) and the rapid global radiation of modern bats', *PloS One* **12**(3), e0172621.

Haeckel, E. (1874), *Anthropogenie oder Entwickelungsgeschichte des Menschen: gemeinverständliche wissenschaftliche Vorträge über die Grundzüge der menschlichen Keimes-und Stammes-Geschichte*, Wilheml Engelmann, Leipzig.

Hahn, M. W. (2007), 'Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution', *Genome Biology* **8**(7), 1–9.

Harris, S. R. (2005), 'Character construction in morphological phylogenetics and the affinities of turtles', PhD thesis, School of Earth Sciences, University of Bristol, Bristol.

Hawkins, J. A., Hughes, C. E. & Scotland, R. W. (1997), 'Primary homology assessment, characters and character states', *Cladistics* **13**(3), 275–283.

Hellemans, S., Wang, M., Hasegawa, N., Šobotník, J., Scheffrahn, R. H. & Bourguignon, T. (2022), 'Using ultraconserved elements to reconstruct the termite tree of life', *Molecular Phylogenetics and Evolution* p. 107520.

Hendy, M. D., Little, C. H. C. & Penny, D. (1984), 'Comparing trees with pendant vertices labelled', *SIAM Journal on Applied Mathematics* **44**(5), 1054–1065.

Hendy, M. D., Steel, M. A., Penny, D. & Henderson, I. M. (1988), 'Families of trees and consensus'. In Block, H. H. (ed.) *Classification and related methods of data analysis*, Elsevier, New York, New York, pp.355–362.

Hennig, W. (1966), *Phylogenetic systematics*, University of Illinois Press, Urbana and Chicago, Illinois.

Hibbett, D. S. & Donoghue, M. J. (2001), 'Analysis of character correlations among wood decay mechanisms, mating systems, and substrate ranges in homobasidiomycetes', *Systematic Biology* **50**(2), 215–242.

Hime, P. M., Lemmon, A. R., Lemmon, E. C. M., Prendini, E., Brown, J. M., Thomson, R. C., Kratovil, J. D., Noonan, B. P., Pyron, R. A., Peloso, P. L. et al. (2021), 'Phylogenomics reveals ancient gene tree discordance in the amphibian tree of life', *Systematic Biology* **70**(1), 49–66.

Hitchcock, E. (1840), *Elementary geology*, J. S. & C. Adams, Amherst.

Höhna, S. & Drummond, A. J. (2011), 'Guided tree topology proposals for Bayesian phylogenetic inference', *Systematic Biology* **61**(1), 1–11.

Holland, B. & Moulton, V. (2003), 'Consensus networks: A method for visualising incompatibilities in collections of trees'. In Benson, G. & Page, R.D.M. (eds.) *International workshop on algorithms in bioinformatics*, Springer, Berlin, Heidelberg, pp. 165–176.

Hopkins, M. J. & St. John, K. (2021), 'Incorporating hierarchical characters into phylogenetic analysis', *Systematic Biology* **70**(6), 1163–1180.

Huber, K. T., Oxelman, B., Lott, M. & Moulton, V. (2006), 'Reconstructing the evolutionary history of polyploids from multilabeled trees', *Molecular Biology and Evolution* **23**(9), 1784–1791.

Huelsenbeck, J. P. (1991), 'When are fossils better than extant taxa in phylogenetic analysis?', *Systematic Biology* **40**(4), 458–469.

Huelsenbeck, J. P. (1997), 'Is the Felsenstein zone a fly trap?', *Systematic Biology* **46**(1), 69–74.

Huerta-Cepas, J., Dopazo, J. & Gabaldón, T. (2010), 'ETE: a Python environment for tree exploration', *BMC Bioinformatics* **11**(1), 1–7.

Hunter, J. D. (2007), 'Matplotlib: A 2D graphics environment', *Computing in Science & Engineering* **9**(3), 90–95.

Huson, D. H. & Bryant, D. (2006), 'Application of phylogenetic networks in evolutionary studies', *Molecular Biology and Evolution* **23**(2), 254–267.

Huson, D. H. & Kloepper, T. H. (2007), 'Beyond galled trees-decomposition and computation of galled networks'. In Speed, T., Huang, H. (ed.) *Research in computational molecular biology RECOMB 2007*, Springer, Berlin, Heidelberg, pp. 211–225.

Huson, D. H., Klöpper, T., Lockhart, P. J. & Steel, M. A. (2005), 'Reconstruction of reticulate networks from gene trees'. In Miyano, S., Mesirov, J., Kasif, S., Istrail, S., Pevzner, P.A. & Waterman, M. (eds.) *Research in computational molecular biology RECOMB 2005*, Springer, Berlin, Heidelberg, pp. 233–249.

Huson, D. H., Rupp, R. & Scornavacca, C. (2010), *Phylogenetic networks: concepts, algorithms and applications*, Cambridge University Press, Cambridge.

Hutter, C. R., Cobb, K. A., Portik, D. M., Travers, S. L., Wood Jr., P. L. & Brown, R. M. (2021), 'FrogCap: A modular sequence capture probe-set for phylogenomics and population genetics for all frogs, assessed across multiple phylogenetic scales', *Molecular Ecology Resources* **22**(3), 1100-1119.

Ippen, T. (2013), 'plotille', `https://github.com/tammoippen/plotille`.

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), 850–858.

Jombart, T. (2008), '*adegenet*: a R package for the multivariate analysis of genetic markers', *Bioinformatics* **24**(11), 1403–1405.

Kamei, R. G., Mauro, D. S., Gower, D. J., Van Bocxlaer, I., Sherratt, E., Thomas, A., Babu, S., Bossuyt, F., Wilkinson, M. & Biju, S. D. (2012), 'Discovery of a new family of amphibians from northeast India with ancient links to Africa', *Proceedings of the Royal Society B: Biological Sciences* **279**(1737), 2396–2401.

Katoh, K. & Standley, D. M. (2013), 'MAFFT multiple sequence alignment software version 7: improvements in performance and usability', *Molecular Biology and Evolution* **30**(4), 772–780.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C. et al. (2012), 'Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data', *Bioinformatics* **28**(12), 1647–1649.

Keynes, J. (1921), *A Treatise on Probability*, Macmillan and Co. Ltd., London.

Kulinskaya, E. (2008), 'On two-sided p-values for non-symmetric distributions', *arXiv preprint arXiv:0810.2124* .

Kulinskaya, E. & Lewin, A. (2009), 'Testing for linkage and Hardy-Weinberg disequilibrium', *Annals of Human Genetics* **73**(2), 253–262.

Kulkarni, S., Kallal, R. J., Wood, H., Dimitrov, D., Giribet, G. & Hormiga, G. (2021), 'Interrogating genomic-scale data to resolve recalcitrant nodes in the spider tree of life', *Molecular Biology and Evolution* **38**(3), 891–903.

Lakner, C., van der Mark, P., Huelsenbeck, J. P., Larget, B. & Ronquist, F. (2008), 'Efficiency of Markov chain Monte Carlo tree proposals in Bayesian phylogenetics', *Systematic Biology* **57**(1), 86–103.

Lanyon, S. M. (1985), 'Detecting internal inconsistencies in distance data', *Systematic Zoology* **34**(4), 397–403.

Lartillot, N., Brinkmann, H. & Philippe, H. (2007), 'Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model', *BMC Evolutionary Biology* **7**(1), 1–14.

Lemmon, A. R., Brown, J. M., Stanger-Hall, K. & Lemmon, E. M. (2009), 'The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference', *Systematic Biology* **58**(1), 130–145.

Lewis, P. O. (2001), 'A likelihood approach to estimating phylogeny from discrete morphological character data', *Systematic Biology* **50**(6), 913–925.

Lipschutz, S. (1998), 'Schaum's outline of set theory and related topics', McGraw Hill, London.

Llabrés, M., Rosselló, F. & Valiente, G. (2021), 'The generalized Robinson-Foulds distance for phylogenetic trees', *Journal of Computational Biology* **28**(12), 1181–1195.

Léveillé-Bourret, t., Starr, J. R., Ford, B. A., Moriarty Lemmon, E. & Lemmon, A. R. (2017), 'Resolving rapid radiations within angiosperm families using anchored phylogenomics', *Systematic Biology* **67**(1), 94–112.

Maciel, A. O., de Castro, T. M., Sturaro, M. J., Silva, I. E. C., Ferreira, J. G., dos Santos, R., Risse-Quaioto, B., Barboza, B. A., Oliveira, J. C., Sampaio, I. et al. (2019), 'Phylogenetic systematics of the neotropical caecilian amphibian *Luetkenotyphlus* (Gymnophiona: Siphonopidae) including the description of a new species from the vulnerable Brazilian Atlantic Forest', *Zoologischer Anzeiger* **281**, 76–83.

Maciel, A. O., Mott, T. & Hoogmoed, M. S. (2009), 'A second species of *Brasilotyphlus* (Amphibia: Gymnophiona: Caeciliidae) from Brazilian Amazonia', *Zootaxa* **2226**(1), 19–27.

Maddin, H. C., Jenkins Jr, F. A. & Anderson, J. S. (2012), 'The braincase of *Eocaecilia micropodia* (Lissamphibia, Gymnophiona) and the origin of caecilians', *PLoS One* **7**(12).

Maddin, H. C., Piekarski, N., Sefton, E. M. & Hanken, J. (2016), 'Homology of the cranial vault in birds: new insights based on embryonic fate-mapping and character analysis', *Royal Society Open Science* **3**(8), 160356.

Maddison, D. R. (1991), 'The discovery and importance of multiple islands of most-parsimonious trees', *Systematic Zoology* **40**(3), 315–328.

Maddison, W. (1989), 'Reconstructing character evolution on polytomous cladograms', *Cladistics* **5**(4), 365–377.

Maddison, W. P. (1993), 'Missing data versus missing characters in phylogenetic analysis', *Systematic Biology* **42**(4), 576–581.

Maddock, S. T., Wilkinson, M. & Gower, D. J. (2018), 'A new species of small, long-snouted *Hypogeophis* Peters, 1880 (Amphibia: Gymnophiona: Indotyphlidae) from the highest elevations of the Seychelles island of Mahé', *Zootaxa* **4450**(3), 359–375.

Mahony, M. J., Hines, H. B., Bertozzi, T., Mahony, S. V., Newell, D. A., Clarke, J. M. & Donnellan, S. C. (2022), 'A new species of *Philoria* (Anura: Limnodynastidae) from the uplands of the Gondwana Rainforests World Heritage area of eastern Australia.', *Zootaxa* **5104**(2), 209–241.

Mai, U. & Mirarab, S. (2018), 'TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees', *BMC Genomics* **19**(5), 23–40.

Margush, T. & McMorris, F. (1981), 'Consensus n-trees', *Bulletin of Mathematical Biology* **43**(2), 239–244.

Marjanović, D. & Laurin, M. (2019), 'Phylogeny of Paleozoic limbed vertebrates reassessed through revision and expansion of the largest published relevant data matrix', *PeerJ* **6**, e5565.

Martin, A. P. & Burg, T. M. (2002), 'Perils of paralogy: using HSP70 genes for inferring organismal phylogenies', *Systematic Biology* **51**(4), 570–587.

Massingham, T. & Goldman, N. (2000), 'EDIBLE: experimental design and information calculations in phylogenetics', *Bioinformatics* **16**(3), 294–295.

McGowan, G. J. (2002), 'Albanerpetontid amphibians from the Lower Cretaceous of Spain and Italy: a description and reconsideration of their systematics', *Zoological Journal of the Linnean Society* **135**(1), 1–32.

Meacham, C. A. & Estabrook, G. F. (1985), 'Compatibility methods in systematics', *Annual Review of Ecology and Systematics* pp. 431–446.

Minh, B. Q., Nguyen, M. A. T. & von Haeseler, A. (2013), 'Ultrafast approximation for phylogenetic bootstrap', *Molecular Biology and Evolution* **30**(5), 1188–1195.

Mirarab, S. (2019), 'Species tree estimation using Astral: practical considerations'. `https://arxiv.org/abs/1904.03826`

Molloy, E. K. & Warnow, T. (2017), 'To include or not to include: the impact of gene filtering on species tree estimation methods', *Systematic Biology* **67**(2), 285–303.

Moon, B. C. (2019), 'A new phylogeny of ichthyosaurs (Reptilia: Diapsida)', *Journal of Systematic Palaeontology* **17**(2), 129–155.

Morrison, D. A. (2007), 'Increasing the efficiency of searches for the maximum likelihood tree in a phylogenetic analysis of up to 150 nucleotide sequences', *Systematic Biology* **56**(6), 988–1010.

Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. (2015), 'IQ-Tree: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies', *Molecular Biology and Evolution* **32**(1), 268–274.

Nixon, K. C. & Carpenter, J. M. (1996), 'On consensus, collapsibility, and clade concordance', *Cladistics* **12**(4), 305–321.

Nixon, K. C. & Davis, J. I. (1991), 'Polymorphic taxa, missing values and cladistic analysis', *Cladistics* **7**(3), 233–241.

Nodelman, U., Shelton, C. R. & Koller, D. (2002), 'Continuous time bayesian networks'. In Darwiche, A. & Friedman, N. *UAI02: Proceedings of the eighteenth conference on uncertainty in artificial intelligence*, Morgan Kaufmann Publishers Inc., San Francisco, California, p. 378–387.

Nye, T. M. (2008), 'Trees of trees: an approach to comparing multiple alternative phylogenies', *Systematic Biology* **57**(5), 785–794.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E. & Wagner, H. (2020), *vegan: Community Ecology Package*. R package version 2.5-7. `https://CRAN.R-project.org/package=vegan`

Olave, M., Avila, L. J., Sites, Jack W., J. & Morando, M. (2015), 'Model-based approach to test hard polytomies in the *Eulaemus* clade of the most diverse South American lizard genus *Liolaemus* (Liolaemini, Squamata)', *Zoological Journal of the Linnean Society* **174**(1), 169–184.

Olmstead, R. G., Bremer, B., Scott, K. M. & Palmer, J. D. (1993), 'A parsimony analysis of the *Asteridae sensu lato* based on rbcL sequences', *Annals of the Missouri Botanical Garden* pp. 700–722.

Olmstead, R. G. & Palmer, J. D. (1994), 'Chloroplast DNA systematics: a review of methods and data analysis', *American Journal of Botany* **81**(9), 1205–1224.

O'Reilly, J. E., Puttick, M. N., Parry, L., Tanner, A. R., Tarver, J. E., Fleming, J., Pisani, D. & Donoghue, P. C. (2016), 'Bayesian methods outperform parsimony but at the expense of precision in the estimation of phylogeny from discrete morphological data', *Biology Letters* **12**(4), 20160081.

Oxford University, P. (2000), 'Oxford English Dictionary', p. 1 online resource. `http://www.oed.com`

Pardo, J. D., Small, B. J. & Huttenlocker, A. K. (2017), 'Stem caecilian from the Triassic of Colorado sheds light on the origins of Lissamphibia', *Proceedings of the National Academy of Sciences of the United States of America* **114**(27), E5389–E5395.

Parvini, G., Braught, K. & Fernández-Baca, D. (2021), 'Checking phylogenetics decisiveness in theory and in practice', *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. *In press*. `https://doi.org/10.1109/TCBB.2021.3128381`

Pattengale, N., Aberer, A., Swenson, K., Stamatakis, A. & Moret, B. (2011), 'Uncovering hidden phylogenetic consensus in large data sets', *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **8**(4), 902–911.

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T. & Reich, D. (2012), 'Ancient admixture in human history', *Genetics* **192**(3), 1065–1093.

Percequillo, A. R., do Prado, J. R., Abreu, E. F., Dalapicolla, J., Pavan, A. C., de Almeida Chiquito, E., Brennand, P., Steppan, S. J., Lemmon, A. R., Lemmon, E. M. et al. (2021), 'Tempo and mode of evolution of oryzomyine rodents (Rodentia, Cricetidae, Sigmodontinae): A phylogenomic approach', *Molecular Phylogenetics and Evolution* **159**, 107120.

Pietsch, T. W. (2012), *Trees of life: a visual history of evolution*, Johns Hopkins University Press, Baltimore, Maryland.

Platnick, N. I., Griswold, C. E. & Coddington, J. A. (1991), 'On missing entries in cladistic analysis', *Cladistics* **7**(4), 337–343.

Pollard, D. A., Iyer, V. N., Moses, A. M. & Eisen, M. B. (2006), 'Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting', *PLoS Genetics* **2**(10), e173.

Puttick, M. N., O'Reilly, J. E., Tanner, A. R., Fleming, J. F., Clark, J., Holloway, L., Lozano-Fernandez, J., Parry, L. A., Tarver, J. E., Pisani, D. et al. (2017), 'Uncertain-tree: discriminating among competing approaches to the phylogenetic analysis of phenotype data', *Proceedings of the Royal Society B: Biological Sciences* **284**(1846), 20162290.

Pyron, R. A. (2014), 'Biogeographic analysis reveals ancient continental vicariance and recent oceanic dispersal in amphibians', *Systematic Biology* **63**(5), 779–797.

Pyron, R. A. & Wiens, J. J. (2011), 'A large-scale phylogeny of Amphibia including over 2800 species, and a revised classification of extant frogs, salamanders, and caecilians', *Molecular Phylogenetics and Evolution* **61**(2), 543–83.

R Core Team (2015), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. `https://www.R-project.org/`

R Core Team (2019), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. `https://www.R-project.org/`

Rabosky, D. L., Donnellan, S. C., Grundler, M. & Lovette, I. J. (2014), 'Analysis and visualization of complex macroevolutionary dynamics: an example from australian scincid lizards', *Systematic Biology* **63**(4), 610–627.

Raxworthy, C. J. & Smith, B. T. (2021), 'Mining museums for historical DNA: advances and challenges in museomics', *Trends in Ecology & Evolution* **36**(11), 1049–1060.

Roberto, I. J., Loebmann, D., Lyra, M. L., Haddad, C. B. & Vila, R. W. (2022), 'A new species of *Pristimantis* Jiménez de la Espada, 1870 (Anura: Strabomantidae) from the Brejos de Altitude in northeast Brazil', *Zootaxa* **5100**(4), 521–540.

Robinson, D. F. & Foulds, L. R. (1981), 'Comparison of phylogenetic trees', *Mathematical Biosciences* **53**(1-2), 131–147.

Robinson, O., Dylus, D. & Dessimoz, C. (2016), 'Phylo.io : interactive viewing and comparison of large phylogenetic trees on the web', *Molecular Biology and Evolution* **33**(8), 2163–2166.

Roelants, K., Gower, D. J., Wilkinson, M., Loader, S. P., Biju, S., Guillaume, K., Moriau, L. & Bossuyt, F. (2007), 'Global patterns of diversification in the history of modern amphibians', *Proceedings of the National Academy of Sciences of the United States of America* **104**(3), 887–892.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A. & Huelsenbeck, J. P. (2012), 'MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space', *Systematic Biology* **61**(3), 539–542.

Roure, B., Philippe, H. & Baurain, D. (2012), 'Impact of missing data on phylogenies inferred from empirical phylogenomic data sets', *Molecular Biology and Evolution* **30**(1), 197–214.

Ruta, M. & Coates, M. I. (2007), 'Dates, nodes and character conflict: addressing the lissamphibian origin problem', *Journal of Systematic Palaeontology* **5**(1), 69–122.

San Mauro, D., Gower, D. J., Cotton, J. A., Zardoya, R., Wilkinson, M. & Massingham, T. (2012), 'Experimental design in phylogenetics: testing predictions from expected information', *Systematic Biology* **61**(4), 661–674.

San Mauro, D., Gower, D. J., Massingham, T., Wilkinson, M., Zardoya, R. & Cotton, J. A. (2009), 'Experimental design in caecilian systematics: phylogenetic information of mitochondrial genomes and nuclear RAG1', *Systematic Biology* **58**(4), 425–438.

San Mauro, D., Gower, D. J., Müller, H., Loader, S. P., Zardoya, R., Nussbaum, R. A. & Wilkinson, M. (2014), 'Life-history evolution and mitogenomic phylogeny of caecilian amphibians', *Molecular Phylogenetics and Evolution* **73**, 177–189.

Sanderson, M. J., McMahon, M. M., Stamatakis, A., Zwickl, D. J. & Steel, M. (2015), 'Impacts of terraces on phylogenetic inference', *Systematic Biology* **64**(5), 709–726.

Sanderson, M. J., McMahon, M. M. & Steel, M. (2010), 'Phylogenomics with incomplete taxon coverage: the limits to inference', *BMC Evolutionary Biology* **10**(1), 1–13.

Sanderson, M. J., McMahon, M. M. & Steel, M. (2011), 'Terraces in phylogenetic tree space', *Science* **333**(6041), 448–450.

Schliep, K. P. (2011), 'phangorn: phylogenetic analysis in R', *Bioinformatics* **27**(4), 592–593.

Schoch, R. R., Werneburg, R. & Voigt, S. (2020), 'A Triassic stem-salamander from Kyrgyzstan and the origin of salamanders', *Proceedings of the National Academy of Sciences of the United States of America* **117**(21), 11584–11588.

Schultze, H.-P., Arratia, G. & Wilson, M. V. H. (2008), 'Nomenclature and homologization of cranial bones in actinopterygians'. In Arratia, G., Schultze, H.-P. & Wilson, V. H. (eds.), *Mesozoic fishes 4: homology and phylogeny*, Verlag Dr. Friedrich Pfeil, München, Bavaria , pp. 23–48.

Semple, C. & Steel, M. (2003), *Phylogenetics*, Oxford University Press, Oxford.

Serra Silva, A. & Wilkinson, M. (2021), 'On defining and finding islands of trees and mitigating large island bias', *Systematic Biology* **70**(6), 1282–1294.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. & Ideker, T. (2003), 'Cytoscape: a software environment for integrated models of biomolecular interaction networks', *Genome Research* **13**(11), 2498–2504.

Sharkey, M. J. & Leathers, J. W. (2001), 'Majority does not rule: the trouble with majority-rule consensus trees', *Cladistics* **17**(3), 282–284.

Sharkey, M. J., Stoelb, S., Miranda-Esquivel, D. R. & Sharanowski, B. J. (2013), 'Weighted compromise trees: a method to summarize competing phylogenetic hypotheses', *Cladistics* **29**(3), 309–314.

Shen, X. X., Liang, D., Feng, Y. J., Chen, M. Y. & Zhang, P. (2013), 'A versatile and highly efficient toolkit including 102 nuclear markers for vertebrate phylogenomics, tested by resolving the higher level relationships of the Caudata', *Molecular Biology and Evolution* **30**(10), 2235–2248.

Shi, J. J. & Rabosky, D. L. (2015), 'Speciation dynamics during the global radiation of extant bats', *Evolution* **69**(6), 1528–1545.

Simmons, M. P. (2012), 'Radical instability and spurious branch support by likelihood when applied to matrices with non-random distributions of missing data', *Molecular Phylogenetics and Evolution* **62**(1), 472 – 484.

Simmons, M. P. & Freudenstein, J. V. (2011), 'Spurious 99% bootstrap and jackknife support for unsupported clades', *Molecular Phylogenetics and Evolution* **61**(1), 177–191.

Simmons, M. P. & Goloboff, P. A. (2014), 'Dubious resolution and support from published sparse supermatrices: the importance of thorough tree searches', *Molecular Phylogenetics and Evolution* **78**, 334–348.

Simmons, M. P. & Norton, A. P. (2013), 'Quantification and relative severity of inflated branch-support values generated by alternative methods: an empirical example', *Molecular phylogenetics and Evolution* **67**(1), 277–296.

Simmons, M. P., Pickett, K. M. & Miya, M. (2004), 'How meaningful are Bayesian support values?', *Molecular Biology and Evolution* **21**(1), 188–199.

Simmons, M. P., Springer, M. S. & Gatesy, J. (2022), 'Gene-tree misrooting drives conflicts in phylogenomic coalescent analyses of palaeognath birds', *Molecular Phylogenetics and Evolution* **167**, 107344.

Simões, T. R. & Pyron, R. A. (2021), 'The squamate tree of life', *Bulletin of the Museum of Comparative Zoology* **163**(2), 47–95.

Singhal, S., Grundler, M., Colli, G. & Rabosky, D. L. (2017), 'Squamate conserved loci (SqCL): a unified set of conserved loci for phylogenomics and population genetics of squamate reptiles', *Molecular Ecology Resources* **17**(6), e12–e24.

Siu-Ting, K., Pisani, D., Creevey, C. J. & Wilkinson, M. (2015), 'Concatabominations: identifying unstable taxa in morphological phylogenetics using a heuristic extension to safe taxonomic reduction', *Systematic Biology* **64**(1), 137–43.

Siu-Ting, K., Torres-Sánchez, M., San Mauro, D., Wilcockson, D., Wilkinson, M., Pisani, D., O'Connell, M. J. & Creevey, C. J. (2019), 'Inadvertent paralog inclusion drives artifactual topologies and timetree estimates in phylogenomics', *Molecular Biology and Evolution* **36**(6), 1344–1356.

Smith, S. A., Beaulieu, J. M. & Donoghue, M. J. (2009), 'Mega-phylogeny approach for comparative biology: an alternative to supertree and supermatrix approaches', *BMC Evolutionary Biology* **9**(1), 37.

Smith, S. A., Brown, J. W. & Hinchliff, C. E. (2013), 'Analyzing and synthesizing phylogenies using tree alignment graphs', *PLOS Computational Biology* **9**(9), 1–15.

Smith, S. A., Walker-Hale, N., Walker, J. F. & Brown, J. W. (2020), 'Phylogenetic conflicts, combinability, and deep phylogenomics in plants', *Systematic Biology* **69**(3), 579–592.

Sokal, R. R. & Rohlf, F. J. (1981), 'Taxonomic congruence in the Leptopodomorpha re-examined', *Systematic Zoology* **30**(3), 309–325.

Soltis, D. E. & Kuzoff, R. K. (1995), 'Discordance between nuclear and chloroplast phylogenies in the *Heuchera* group (Saxifragaceae)', *Evolution* **49**(4), 727–742.

Springer, M. S., Meredith, R. W., Gatesy, J., Emerling, C. A., Park, J., Rabosky, D. L., Stadler, T., Steiner, C., Ryder, O. A., Janečka, J. E. et al. (2012), 'Macroevolutionary dynamics and historical biogeography of primate diversification inferred from a species supermatrix', *PloS One* **7**(11), e49521.

Stamatakis, A. (2014), 'RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies', *Bioinformatics* **30**(9), 1312–1313.

Stamatakis, A., Hoover, P. & Rougemont, J. (2008), 'A rapid bootstrap algorithm for the RAxML web servers', *Systematic Biology* **57**(5), 758–771.

Steel, M. (2016), *Phylogeny: discrete and random processes in evolution*, Society for Industrial and Appied Mathematics, Philadelphia, Pennsylvania.

Steel, M. & Sanderson, M. J. (2010), 'Characterizing phylogenetically decisive taxon coverage', *Applied Mathematics Letters* **23**(1), 82–86.

Stockham, C., Wang, L.-S. & Warnow, T. (2002), 'Statistically based postprocessing of phylogenetic analysis by clustering', *Bioinformatics* **18**(suppl_1), S285–S293.

Stolzer, M., Lai, H., Xu, M., Sathaye, D., Vernot, B. & Durand, D. (2012), 'Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees', *Bioinformatics* **28**(18), i409–i415.

Streicher, J. W., Loader, S. P., Varela-Jaramillo, A., Montoya, P. & de Sá, R. O. (2020), 'Analysis of ultraconserved elements supports african origins of narrow-mouthed frogs', *Molecular Phylogenetics and Evolution* **146**, 106771.

Streicher, J. W., Schulte, J. A. & Wiens, J. J. (2016), 'How should genes and taxa be sampled for phylogenomic analyses with missing data? An empirical study in iguanian lizards', *Systematic Biology* **65**(1), 128–145.

Streicher, J. W. & Wiens, J. J. (2017), 'Phylogenomic analyses of more than 4000 nuclear loci resolve the origin of snakes among lizard families', *Biology Letters* **13**(9), 20170393.

Sukumaran, J. & Holder, M. T. (2010), 'Dendropy: a python library for phylogenetic computing', *Bioinformatics* **26**(12), 1569–1571.

Sumrall, C. D., Brochu, C. A. & Merck, J. W. (2001), 'Global lability, regional resolution, and majority-rule consensus bias', *Paleobiology* **27**(2), 254–261.

Suvorov, A., Kim, B. Y., Wang, J., Armstrong, E. E., Peede, D., D'agostino, E. R., Price, D. K., Waddell, P. J., Lang, M., Courtier-Orgogozo, V. et al. (2022), 'Widespread introgression across a phylogeny of 155 *Drosophila* genomes', *Current Biology* **32**(1), 111–123.

Suzuki, Y., Glazko, G. V. & Nei, M. (2002), 'Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics', *Proceedings of the National Academy of Sciences of the United States of America* **99**(25), 16138–16143.

Swofford, D. L. (1991), 'When are phylogeny estimates from molecular and morphological data incongruent?'. In Myamoto, M. M. & Cracraft, J. *Phylogenetic analyses of DNA sequences*, Oxford University Press, New York, New York, pp. 295–333.

Swofford, D. L. (2003), 'PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4.0 a165', Sinauer Associates, Sunderland, Massachusetts.

Taboada, S., Serra Silva, A., Neal, L., Cristobo, J., Ríos, P., Álvarez-Campos, P., Hestetun, J. T., Koutsouveli, V., Sherlock, E. & Riesgo, A. (2020), 'Insights into the symbiotic relationship between scale worms and carnivorous sponges (Cladorhizidae, *Chondrocladia*)', *Deep Sea Research Part I: Oceanographic Research Papers* **156**, 103191.

Tahiri, N., Fichet, B. & Makarenkov, V. (2022), 'Building alternative consensus trees and supertrees using k-means and Robinson and Foulds distance', *Bioinformatics*. *In press*. https://doi.org/10.1093/bioinformatics/btac326

Tahiri, N., Willems, M. & Makarenkov, V. (2018), 'A new fast method for inferring multiple consensus trees using k-medoids', *BMC Evolutionary Biology* **18**(1), 48.

Tarasov, S. (2019), 'Integration of anatomy ontologies and evo-devo using structured Markov models suggests a new framework for modeling discrete phenotypic traits', *Systematic Biology* **68**(5), 698–716.

Taylor, D. W., Brenner, G. J. & Basha, S. H. (2008), '*Scutifolium jordanicum* gen. et sp. nov. (Cabombaceae), an aquatic fossil plant from the Lower Cretaceous of Jordan, and the relationships of related leaf fossils to living genera', *American Journal of Botany* **95**(3), 340–352.

Thorley, J. L. (2000), Cladistic information, leaf stability and supertree construction., PhD thesis, Faculty of Science, University of Bristol, Bristol.

Thorley, J. L. & Wilkinson, M. (1999), 'Testing the phylogenetic stability of early tetrapods', *Journal of Theoretical Biology* **200**(3), 343.

Tolley, K. A., Townsend, T. M. & Vences, M. (2013), 'Large-scale phylogeny of chameleons suggests African origins and Eocene diversification', *Proceedings of the Royal Society B: Biological Sciences* **280**(1759), 20130184.

Tria, F. D. K., Landan, G. & Dagan, T. (2017), 'Phylogenetic rooting using minimal ancestor deviation', *Nature Ecology & Evolution* **1**(1), 1–7.

Trueb, L. & Cloutier, R. (1991), 'A phylogenetic investigation of the inter-and intrarelationships of the Lissamphibia (Amphibia: Temnospondyli)', In Schultze, P.-H & Trueb, L. (eds.), *Origins of the higher groups of tetrapods: controversy and consensus*, Cornell University Press, Ithaca, New York pp. 223–313.

Vallin, G. & Laurin, M. (2004), 'Cranial morphology and affinities of *Microbrachis*, and a reappraisal of the phylogeny and lifestyle of the first amphibians', *Journal of Vertebrate Paleontology* **24**(1), 56–72.

Van Rossum, G. & Drake, F. L. (2009), *Python 3 reference manual*, CreateSpace, Scotts Valley, California.

Von Haeseler, A. (2012), 'Do we still need supertrees?', *BMC Biology* **10**(1), 1–4.

Wächter, D. & Melzer, A. (2020), 'Proposal for a subdivision of the family Psathyrellaceae based on a taxon-rich phylogenetic analysis with iterative multigene guide tree', *Mycological Progress* **19**(11), 1151–1265.

Wagstaff, K. (2004), 'Clustering with missing values: No imputation required'. In Banks, D., McMorris, F. R., Arabie, P. & Gaul, W. (eds.), *Classification, clustering, and data mining applications*, Springer, Berlin, Heidelberg, pp. 649–658.

Whidden, C. & Matsen IV, F. A. (2018), 'Efficiently inferring pairwise subtree prune-and-regraft adjacencies between phylogenetic trees'. In Nebel, M. & Wagner, S. (eds.), *2018 Proceedings of the meeting on analytic algorithmics and combinatorics (ANALCO)*, Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, pp. 77–91.

Wiens, J. J. (2003), 'Missing data, incomplete taxa, and phylogenetic accuracy', *Systematic Biology* **52**(4), 528–538.

Wiens, J. J. & Morrill, M. C. (2011), 'Missing data in phylogenetic analysis: reconciling results from simulations and empirical data', *Systematic Biology* **60**(5), 719–731.

Wilkinson, M. (1994), 'Common cladistic information and its consensus representation: reduced Adams and reduced cladistic consensus trees and profiles', *Systematic Biology* **43**(3), 343–368.

Wilkinson, M. (1995*a*), 'A comparison of two methods of character construction', *Cladistics* **11**(3), 297–308.

Wilkinson, M. (1995*b*), 'Coping with abundant missing entries in phylogenetic inference using parsimony', *Systematic Biology* **44**(4), 501–514.

Wilkinson, M. (1995*c*), 'More on reduced consensus methods', *Systematic Biology* **44**(3), 435–439.

Wilkinson, M. (1996), 'Majority-rule reduced consensus trees and their use in bootstrapping.', *Molecular Biology and Evolution* **13**(3), 437–444.

Wilkinson, M. (2003), 'Missing entries and multiple trees: instability, relationships, and support in parsimony analysis', *Journal of Vertebrate Paleontology* **23**(2), 311–323.

Wilkinson, M. (2006), 'Identifying stable reference taxa for phylogenetic nomenclature', *Zoologica Scripta* **35**(1), 109–112.

Wilkinson, M. & Benton, M. J. (1996), 'Sphenodontid phylogeny and the problems of multiple trees', *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **351**(1335), 1–16.

Wilkinson, M. & Cotton, J. (2006), 'Supertree methods for building the tree of life: divide-and-conquer approaches to large phylogenetic problems'. In Hodkinson, T. R. & Parnell, J. A. N. (eds.), *Reconstructing the tree of life: taxonomy and systematics of species rich taxa*, CRC Press, Boca Raton, Florida, pp. 61–75.

Wilkinson, M., Cotton, J. A., Creevey, C., Eulenstein, O., Harris, S. R., Lapointe, F. J., Levasseur, C., McInerney, J. O., Pisani, D. & Thorley, J. L. (2005), 'The shape of supertrees to come: tree shape related properties of fourteen supertree methods', *Systematic Biology* **54**(3), 419–31.

Wilkinson, M., Cotton, J. A. & Thorley, J. L. (2004), 'The information content of trees and their matrix representations', *Systematic Biology* **53**(6), 989–1001.

Wilkinson, M. & Crotti, M. (2017), 'Comments on detecting rogue taxa using roguenarok', *Systematics and Biodiversity* **15**(4), 291–295.

Wilkinson, M., Pisani, D., Cotton, J. A. & Corfe, I. (2005), 'Measuring support and finding unsupported relationships in supertrees', *Systematic Biology* **54**(5), 823–831.

Wilkinson, M., Reynolds, R. P. & Jacobs, J. F. (2021), 'A new genus and species of rhinatrematid caecilian (Amphibia: Gymnophiona: Rhinatrematidae) from Ecuador.', *Herpetological Journal* **31**(1).

Wilson, J. (2002), 'Sauropod dinosaur phylogeny: critique and cladistic analysis', *Zoological Journal of the Linnean Society* **136**(2), 215–275.

Wright, A. M. & Hillis, D. M. (2014), 'Bayesian analysis using a simple likelihood model outperforms parsimony for estimation of phylogeny from discrete morphological data', *PLoS One* **9**(10), e109210.

Wurdack, K. J. & Davis, C. C. (2009), 'Malpighiales phylogenetics: gaining ground on one of the most recalcitrant clades in the angiosperm tree of life', *American Journal of Botany* **96**(8), 1551–1570.

Zaharias, P., Pante, E., Gey, D., Fedosov, A. E. & Puillandre, N. (2020), 'Data, time and money: evaluating the best compromise for inferring molecular phylogenies of non-model animal taxa', *Molecular Phylogenetics and Evolution* **142**, 106660.

Zanne, A. E., Tank, D. C., Cornwell, W. K., Eastman, J. M., Smith, S. A., FitzJohn, R. G., McGlinn, D. J., O'Meara, B. C., Moles, A. T., Reich, P. B. et al. (2014), 'Three keys to the radiation of angiosperms into freezing environments', *Nature* **506**(7486), 89–92.

Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S. (2018), 'Astral-III: polynomial time species tree reconstruction from partially resolved gene trees', *BMC Bioinformatics* **19**(6), 15–30.

Zwickl, D. (2014), 'Terraphy: various analyses and utilities related to phylogenetic terraces'. `https://github.com/zwickl/terraphy/`