Systematic Bias in Phylogenetic Methods: Investigating the Adequacy of the Treelikeness Assumption

Caitlin Ann Cherryh

December 2024

A thesis submitted for the degree of Doctor of Philosophy of
The Australian National University

This page intentionally left blank

Declaration

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university. To the best of the author's knowledge, it contains no material previously published or written by another person, except where due reference is made in the text.

Caitlin Ann Cherryh

December 2024

I acknowledge the traditional owners of the land on which I lived and studied, the Ngunnawal and Ngambri people, and the traditional owners of lands throughout Australia where the published studies I used within my thesis were conducted. I extend my respect to elders past, present and emerging.

This thesis is dedicated to Zachary Ryan Dowton, the light of my life and the sun in my sky.

Acknowledgements

My PhD was funded by an Australian Government Research Training Program scholarship and by the Australian National University. I am grateful to both for allowing me to pursue this opportunity.

I have enormous gratitude for my panel chair and primary supervisor, Rob Lanfear. I am extremely grateful to Rob for believing in my potential and guiding me towards being the more confident, more capable researcher he saw I could be. Rob has modelled how to be a good supervisor, a good researcher, a good colleague and a good supervisor. Every day, Rob lives by his values and acts with integrity. Through working with Rob over the past few years I have learnt now only to be a good biologist, but how to be a good person. I am deeply lucky to have such a fantastic supervisor.

Thank you to each of my panel members: Minh Bui, Maja Adamska and Barbara Holland. I am grateful for the expertise and support of my panel members as I explored the world of phylogenetics and systematic bias methods. Minh's expertise in software, algorithms, and methods helped me better understand current methods and the gaps in them, as well as showing me how to have computational rigor in the methods I developed. Maja encouraged my interest in sponges and supported me with her expertise as I gradually shifted my research focus towards the animal tree of life. Barbara taught me to think critically about the underpinnings of models and simulations, and always had time for a zoom chat or to email the perfect paper for the concept I was struggling with. Each of my panel members has helped me to shape this thesis and to grow as a biologist.

I would like to thank all members of the Lanfear and Bui labs, including past, present, and honorary members. Throughout my PhD I have been privileged to work with and learn from my colleagues: Yilin Bai, James Barbetti, Eleanor Beavan, Ashutosh Das, Piyumal Demotte, Yanghe Dong, Polly Hannaford, Jeremias Ivan, Frederick Jaya, Hashara Kumarasinghe, Suha Naser-Khdour, Matthew McCauley, Jinghua Mu, Van Nguyen-Hoang, Nhan Trong Ly, Huaiyen Ren, Weiwen Wang, Thomas Wong, and Zora Zhuang. I greatly appreciate the kindness of Eleanor Beavan, Suha Naser-Khdour and Weiwen Wang, who helped me find my feet within RSB and navigate the HDR system. Particular thanks to Thomas Wong for his dedicated work to developing the MAST implementation in IQ-Tree, especially for his quick bug fixes and collaborative approach while we worked on my third chapter.

The Division of Ecology and Evolution at the Research School of Biology has been an incredible place to study. I am grateful to every student and member of staff who contributed to the positive and engaging environment where I have been able to learn and grow. Particular thanks to those responsible for managing the school, the division, and all the HDR convenors: Michael Jennions, Scott Keogh, Craig Moritz, Celeste Linde, Rod Peakall, and Spencer Whitney. Thanks also to the administrative, support and IT staff who kept everything running smoothly even during a global pandemic.

Many people outside my department have been generous with their time and energy. My entry to biology and scientific research was made possible by Dr Clare Holleley, who sparked my

love for the natural world and helped me create a pathway to where I am today. My lovely mentor, Dr Tory Clarke, has helped me navigate the PhD process and the complexities of a career in science. Dr Erin Hahn has been a voice of wisdom when I am struggling and my role model for conducting a science career with integrity, empathy, and passion. I am grateful to both Clare and Erin for the opportunities they have created for me and the belief they have in me. I am also grateful to the staff of the Australian National Wildlife Collection (ANWC) at CSIRO. My first biology research project was at ANWC, and I could not have asked for a more supportive and engaging environment to enter the world of evolutionary biology.

I have so much appreciation for all of my lovely friends who have supported me throughout the PhD process. Special thanks to Jack Berry; Dash Bruce and Breeze Mojel; Andrew Bugg and Christie Ng; Jacob Caluzzi and Katelyn Kummer; Hannah Carle and Jacob Ross; Jaki Coppen and Marlon Leicester; Ivo de los Santos Vekemans and Laura Mouat; Ashley Drury; Roshan Fernandez; Evie Forward and Mason Rose-Campbell; Margaret Harrison; Sophie Holland; Tim McInerney and Ada Kapetanović; Russell Nash and Genevieve Ridgeway; Sylvius Leonard, Samantha Fogarty and their daughters Dorothy and Margaret; Ryan Pemberton; Bryce Pemberton and Nancy Li; Jess Petrie; Sam and Caitlin Launt; Cassandra Taylor; Oscar Tigwell; Jess Youngberry; Ada Quinn and Lissa Fraser; Oliver Stuart; and Melissa von Moger. In particular I am grateful to Kyara and Kevin Chong, and their daughters Olivia and Lily. Kyara has been my cheerleader for over a decade, and her friendship has always brought lightness and love to my life. I am so grateful for the Chong's love and support throughout this journey. Thank you for being part of my family.

Thank you to my parents, Kathy Henderson and Jamie Cherryh, and my siblings Sophie and Michael Cherryh, for their support and love. My parents always encouraged my interest in science and learning, which started my path towards evolutionary biology and scientific research. I am grateful to the support my parents offered my husband and I over the past few chaotic years. I have been privileged to grow alongside my siblings and to navigate the world together as adults.

My wonderful parents-in-law Pete and Liz Dowton graciously welcomed me into their home during the pandemic. I am so grateful for their love, support, and practical advice while I navigated a global pandemic and a PhD program. Thanks also to my brother-in-law Chris Dowton, who has been a voice of reason to me in stressful times and is always keen to discuss the intricacies of baking. The Dowtons welcomed me into their family with open arms and I am honoured to be included.

Finally, all my love and all my gratitude to my husband, Zachary Ryan Dowton. Zac has always encouraged me to dream and supported me to thrive, and I am so grateful for his love and care. This thesis would not exist without Zac – each page, each paragraph was made possible through his unyielding support. Thank you, Zac. I love you.

Abstract

Phylogenetics is the science of estimating relationships between different individuals or groups using genetic sequence data. As phylogenetic analyses are attempting to reconstruct evolutionary history using only a snapshot of data, models and assumptions about the processes and patterns of evolution are applied. One assumption incorporated into many phylogenetic methods is the treelikeness assumption, which states that each site in an alignment shares an identical evolutionary history, which fits a single bifurcating tree. However, treelikeness is violated by biological processes such as introgression, hybridisation, and incomplete lineage sorting. This has the potential to reduce the accuracy of phylogenetic inference, potentially resulting in inaccurate trees and impacting downstream analyses.

In Chapter One, I perform a comprehensive benchmarking of tests for treelikeness. Various test statistics to quantify the treelikeness of an alignment have been proposed, but to my knowledge there has been no systematic comparison of the behaviour of these tests under controlled conditions with gradated treelikeness. I developed 2 simulation schemes: one to assess the absolute adequacy of each test to detect changes in treelikeness, and one to assess test performance under biologically feasible conditions. I also introduce a new test statistic for quantifying treelikeness, which I call the tree proportion, along with a parametric bootstrap to assess the statistical significance of test statistic results for empirical datasets. I found three test statistics performed well when considering behaviour under both simulation schemes, underlying methodology, and existing implementations: the δ plot (Holland 2002), site concordance factors (Minh 2020, Mo 2023) and my new test tree proportion. When applying these three tests to empirical datasets I found that the null hypothesis of treelikeness was rarely rejected, due in part to underlying construction of metrics and due to the increased complexity moving from simulated to empirical multiple sequence alignments.

In Chapter Two, I explored the impacts of intra-locus recombination on phylogenetic tree accuracy. Intra-locus recombination violates the treelikeness assumption, but is known to occur frequently within empirical sequence alignments. I applied 3 existing tests for recombination to each locus from 4 empirical sequence alignments, and constructed “clean” subsets of loci for each alignment by excluding all putatively-recombinant loci. I then compared trees estimated from the unfiltered dataset to trees estimated from the clean subsets. In general, trees estimated from clean datasets using concatenation methods were similar or identical to trees estimated from the unfiltered dataset, providing that there were sufficient loci. However, under summary methods (also known as two step methods), I identified several statistically significant and biologically meaningful differences between trees estimated from the clean and unfiltered datasets.

In Chapter Three, I extend the treelikeness assumption by allowing a single multiple sequence alignment to have multiple distinct evolutionary histories. The metazoan tree is an unsolved

and contentious problem in phylogenetics. This chapter assesses whether a single tree is adequate to represent the evolutionary history of the Metazoa. I used the Mixtures Across Sites and Trees (MAST) model (Wong et al. 2024), a multitree mixture model which uses mixtures of bifurcating trees with independent models of evolution to represent the evolutionary histories within a dataset. I applied the MAST model with 4 classes of substitution models to 14 empirical phylogenetic datasets previously used to estimate the relationships between metazoan clades. I found that multi-tree models were overwhelmingly preferred (46/54 analyses). These results suggest that using current phylogenetic models, a single bifurcating tree is insufficient to describe the complex evolutionary history of Metazoa.

In Chapter Four, I extended my analyses from Chapter Three to investigate the causes of conflicting signal within metazoan datasets using concordance factors. I took 12 datasets previously used to estimate the metazoan phylogeny, and estimated species trees and gene trees with a concatenated model, partitioned model, and C60 model. I then examined the variation in phylogenetic signal using gene and quartet concordance factors. I found substantial conflicting phylogenetic signal within the published empirical datasets. My results suggest widespread incomplete lineage sorting contributes to the difficulty of resolving deep nodes within the metazoan phylogeny.

Together, these chapters show the diverse impacts of the treelikeness assumption on phylogenetic inference and suggest that treelikeness should be considered during phylogenetic tree inference. This thesis presents methods to investigate and interpret the treelikeness of any multiple sequence alignment, and suggests approaches for managing and analysing non-treelike data.

List of Abbreviations

Abbreviation	Definition
AA	Amino acid
AIC	Akaike information criterion
AICc	Corrected Akaike information criterion
ASTRAL	Accurate Species Tree ALgorithm
AU	Approximately unbiased
BIC	Bayesian information criterion
BILAT	Bilateria
BL	Branch length
bp	Base pairs
C-gene	Coalescence gene
C20	A 20-profile protein mixture model
C60	A 60-profile protein mixture model
CAT	A protein profile mixture model
CF	Concordance factor
CF4	A five-profile protein mixture model
CNID	Cnidaria
CONCAT	Concatenated
CTA	Constrained topology analyses
CTEN	Ctenophora
CTEN+PORI	A monophyletic consisting of Ctenophora and Porifera
DNA	Deoxyribonucleic acid
EHO	A three-matrix protein mixture model
EX_EHO	A six-category protein mixture model
EX2	A two-matrix protein mixture model
EX3	A three-matrix protein mixture model
F81	A nucleotide substitution model
gCF	Gene concordance factor
gDF1	Gene discordance factor 1
gDF2	Gene discordance factor 2
gDFP	Gene discordance factor (paraphyletic)
GENECONV	A software program for detecting gene conversion within genetic sequences
GTEE	Gene tree estimation error
GTR	General time reversible model of nucleotide substitution
GTR20	General time reversible model of amino acid substitution
HGT	Horizontal gene transfer
ILS	Incomplete lineage sorting
IQ-Tree2	Important Quartets Tree 2
JC	Jukes-Cantor model of nucleotide substitution
JC69	Alternate abbreviation for the Jukes-Cantor model of nucleotide substitution
JTT	An amino acid substitution model
JTTDCMut	A modified version of the JTT amino acid substitution model
LBA	Long branch attraction
LG	An amino acid substitution model
LG4M	Four matrix protein mixture model with gamma rate heterogeneity
LPP	Local posterior probability – a measure of branch support
LW	Likelihood weight
Ma	Millions of years ago

MAST	Mixtures Across Sites and Trees
MaxChi	Maximum Chi-squared – a method of recombination detection
MFP	Command line option to perform model selection with ModelFinder in IQ-Tree
ML	Maximum likelihood
ms	Software to generate samples from a population under a Wright-Fisher neutral model
MSC	Multi-species coalescent
MSNC	Multi-species network coalescent
mtZOA	An amino acid substitution model for mitochondrial regions of metazoan species
Myr	Million years
NNI	Nearest-neighbour interchange
PHI	Pairwise Homoplasy Index – a method of recombination detection
PM	Profile mixture
PMB	An amino acid substitution model
PMSF	Posterior mean site frequency
PLAC	Placozoa
PORI	Porifera
qCF	Quartet concordance factor
qDF1	Quartet discordance factor 1
qDF2	Quartet discordance factor 2
QuIBL	Quantifying Introgression via Branch Lengths
RF	Robinson-Foulds
RHAS	Rate Heterogeneity Across Sites
RSS	Residual sum of squares
rtREV	An amino acid substitution model designed for retroviruses
SOM	Sister to all other metazoan clades
SRH	Stationary, Reversible and Homogeneous
TIGER	Tree Independent Generation of Evolutionary Rates
TSS	Total sum of squares
UFB	Ultra-fast bootstrap – a measure of branch support
UL2	Unsupervised learning variant of the EX2 model
UL3	Unsupervised learning variant of the EX3 model
WAG	An amino acid substitution model
wRF	Weighted Robinson-Foulds
δ plot	Delta plots

Table of Contents

List of Tables

List of Supplementary Tables

List of Figures

List of Supplementary Figures

This page intentionally left blank

Introduction

Phylogenetics is the study of evolutionary history, with the aim of inferring evolutionary relationships between groups of individuals or species (Delsuc et al. 2005; Jermiin et al. 2020). Prior to the 20th century, organisms were categorised using morphology. The use of DNA sequences to infer evolutionary relationships between taxa began in the late 20th century, but these early molecular phylogenies were limited to a single locus and simple evolutionary models. More recently, with the advent of next generation sequencing and increased computational power, phylogenetic analyses are conducted using genome-scale data (Philippe et al. 2005; Kapli et al. 2020).

i Introduction to Phylogenetics

Historically, relationships between different species were determined using morphology. Species with homologous traits were placed together within the hierarchy of organisms (Rieppel 2020). In the 19th century, evolutionary trees were used to represent these relationships (Ragan 2009). Darwin's notebook contains perhaps the most famous example (Figure 1a), a sketch representing the relationships as a branching tree (Darwin 1987; Ragan 2009). Another well-known example of an early tree was produced by Haeckel (1866) (Figure 1b), who used morphology to construct a genealogical tree of living organisms (Dayrat 2003). In the 1950s and 1960s, biologists beginning to infer evolutionary relationships using comparisons of genetic sequences (Crick 1958; Zuckerkandl and Pauling 1965; Fitch and Margoliash 1967; Cobb 2017). The development of Sanger sequencing in the 1970s (Sanger and Coulson 1975; Sanger et al. 1977) enabled accurate large-scale DNA sequencing and dramatically increased the availability of genetic sequences (Heather and Chain 2016). In the latter half of the twentieth century, a variety of methods for estimating molecular phylogenies were developed including maximum likelihood, parsimony, pairwise sequence similarity, or sequence compatibility methods (Edwards and Cavalli-Sforza 1963, 1964; Fitch and Margoliash 1967; Le Quesne 1969; Felsenstein 1973, 1981, 1983, 1984; Chakraborty 1977). These phylogenies were generally inferred from one locus (or very few loci), using simple models of evolution (Fitch and Margoliash 1967; Woese and Fox 1977; Brown et al. 1982; Doolittle and Feng 1987). Phylogenetic methods and sequencing technology continued to develop throughout the late 20th century, and phylogenetic trees were established as an important tool for studying evolution (Nei and Kumar 2001). Facilitated by the availability of next generation sequencing and high performance computing (Philippe et al. 2005; Kapli et al. 2020), modern phylogenies are inferred from massive datasets containing hundreds or thousands of taxa and thousands of loci (Leebens-Mack et al. 2019a; Kawahara et al. 2023;

Stiller et al. 2024). The increasing amount of sequence data has resulted in the development of probabilistic models of sequence evolution, and more robust statistical methods for phylogenetic inference.

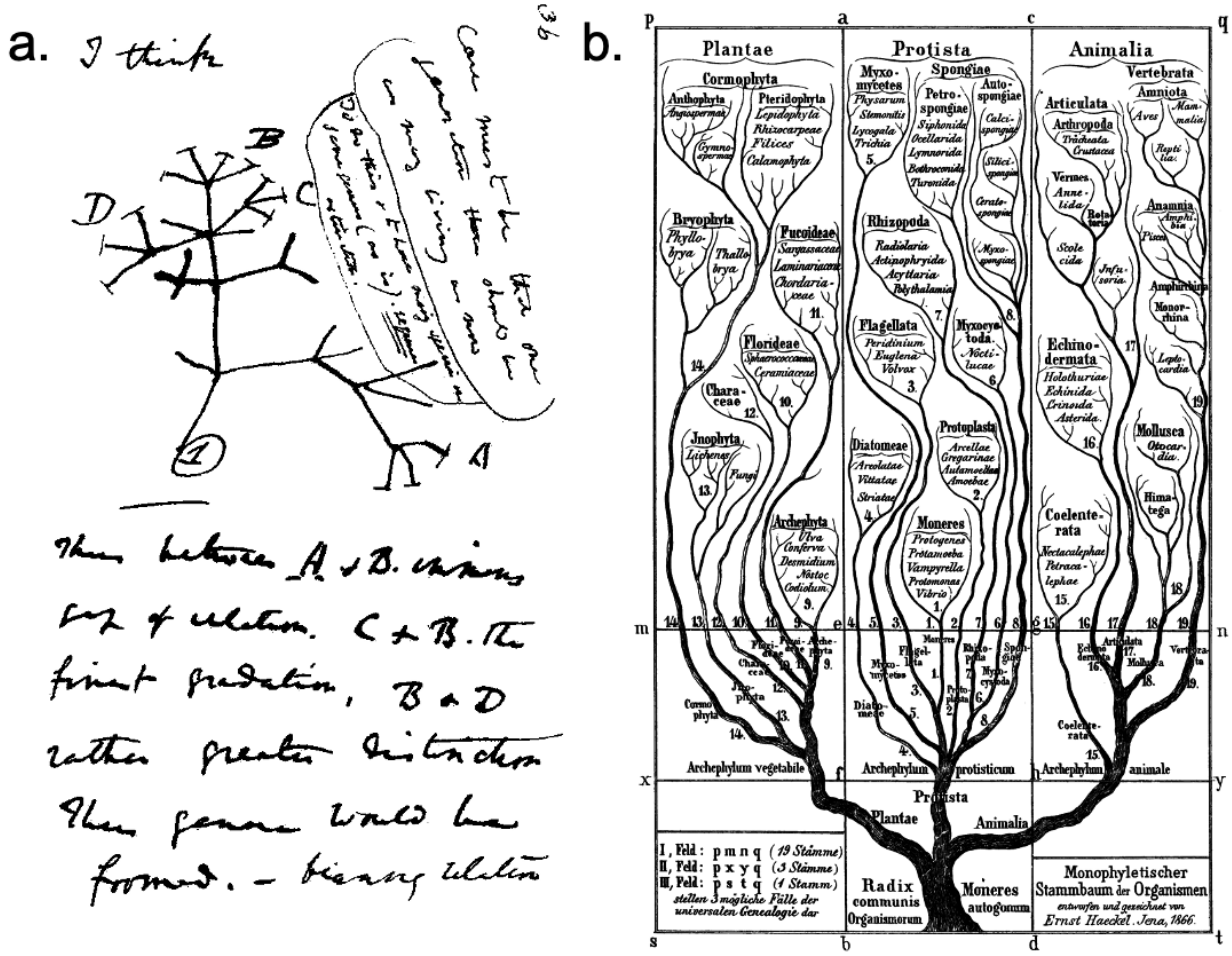


Figure 1: Famous examples of evolutionary trees

a. "I think" from Darwin's notebook (Darwin 1987, Notebook B, p. 36). Text reads: "*I think [the] case must be that one generation then should be as many living as now. To do this & to have many species in same genus (as is) requires extinction. Thus between A & B immense gap of relation. C & B the finest gradation, B & D rather greater distinction. Thus genera would be formed. - bearing relation to ancient types with several extinct forms...*" (Darwin 1987, Notebook B, p. 36 - 37). Figure in public domain (Racconish / Public Domain / Wikimedia Commons).

b. *Monophyletischer Stammbaum der Organismen* (monophyletic genealogical tree of organisms) (Haeckel 1866, Vol. II, Plate II), depicting three clades of living organisms: Plants, Animals, and Protista. Figure in public domain (Cmdrjameson / Public Domain / Wikimedia Commons).

A range of different phylogenetic methods have been developed. Broadly, these methods fall into two classes: distance-based and character-based. Distance-based methods calculate the genetic distance between every pair of taxa, then construct a tree from the resulting distance matrix (Kapli et al. 2020). The most common distance method is Neighbor-Joining (Saitou and Nei 1987) (>72,000 citations), which is available in multiple software programs including BioNJ (Gascuel 1997), SplitsTree (Huson 1998; Huson and Bryant 2022), amd PAUP* (Swofford and Sullivan 2003; Swofford 2019). Character-based methods infer phylogenies by comparing the character states for all sequences within the alignment at individual sites (nucleotides or amino acid residues) (Yang and Rannala 2012). Commonly used character-based methods include maximum parsimony, maximum likelihood, and Bayesian inference. Maximum parsimony is a mathematically simple and computationally efficient method of tree estimation, which infers a tree that minimises the total number of state changes required to explain the alignment (Felsenstein 1983). Multiple software programs are available to infer trees with maximum parsimony, including MPBoot (Hoang et al. 2018b), PAUP* (Swofford and Sullivan 2003; Swofford 2019), and UShER (Turakhia et al. 2021). Maximum likelihood and Bayesian inference differ from maximum parsimony, as the former methods apply both an explicit model of evolution and a likelihood function.

Maximum likelihood tree reconstruction methods were developed in the early 1980s (Felsenstein 1981). Today, maximum likelihood is commonly used to infer trees in empirical phylogenetic studies (Cunha et al. 2022; Kawahara et al. 2023; Sanderson et al. 2023; Wolfe et al. 2023; Stiller et al. 2024). Maximum likelihood methods use a model of sequence evolution and a likelihood function to infer the tree that maximises the likelihood (Yang and Rannala 2012; Kapli et al. 2020). This has two steps: performing tree search to find the maximum likelihood tree, and optimising the branch lengths (Yang and Rannala 2012). The most commonly used software for maximum likelihood tree inference are IQ-TREE (Nguyen et al. 2015; Minh et al. 2020b) and RAxML (Stamatakis 2014; Kozlov et al. 2019) with >25K and >30K citations respectively. Other programs include PAML (Yang 2007), PAUP* (Swofford 2019), PhyML (Guindon et al. 2010) and FastTree (Price et al. 2010) (the latter of which infers approximately-maximum-likelihood trees).

Bayesian inference methods were developed in the late 1990s (Rannala and Yang 1996). Rather than estimating a single tree, Bayesian methods use information from a prior and the likelihood function to create a posterior distribution of parameters that are plausible under the model for the data (Rannala and Yang 2017). Use of a prior allows researchers to include known information for the tree and model (Yang and Rannala 2012). The posterior distribution contains a set of trees, each with an associated posterior probability. This distribution of trees

is a benefit of Bayesian analyses, compared to the point estimate inferred by maximum likelihood analyses. A distribution of trees is hard to analyse so generally the posterior distribution of trees is summarised in some way for publication, such as reporting the tree with the highest probability or the set of trees whose probability sums to a threshold (e.g., 90% or 95%) (Cranston and Rannala 2007). Bayesian inference is commonly used to infer trees from empirical phylogenetic datasets (Simion et al. 2017a; Laumer et al. 2018a, 2019a; Cunha et al. 2022; Stiller et al. 2024). Software for Bayesian inference include MrBayes (Huelsenbeck and Ronquist 2001; Ronquist et al. 2012), RevBayes (Höhna et al. 2016), BEAST (Bouckaert et al. 2019), and PhyloBayes (Lartillot et al. 2009; Lartillot 2020a). Bayesian inference programs are frequently used and highly cited, with >50K citations for MrBayes and >20K citations for BEAST. An advantage of Bayesian methods is that they allow complex site heterogeneous models of evolution (Yang and Rannala 2012; Kapli et al. 2020). The main weaknesses of Bayesian inference are the computational and time requirements. Bayesian methods are extremely computationally intensive, and are often computationally intractable for genome-scale datasets (Yang and Rannala 2012).

Both maximum likelihood and Bayesian method use substitution models to estimate the evolutionary history of a given alignment. A range of different models of DNA and protein sequence evolution have been developed for tree estimation. Most commonly used models of sequence evolution contain a number of components: a substitution rate matrix *Q*, a frequency vector, and the distribution of rates heterogeneity across sites (Kalyaanamoorthy et al. 2017). The rate matrix described is an *n* by *n* matrix describing the probability of changing from one nucleotide or amino acid to another in a particular time period. The frequency vector describes the frequencies of each character state in the alignment. Finally, the distribution of rate heterogeneity across sites is a model describing the relative rate of evolution for different sites in the alignment (Sullivan and Joyce 2005). In general, the model for a given alignment is selected by applying a program such as ModelFinder (Kalyaanamoorthy et al. 2017) or ModelTest-NG (Darriba et al. 2020) to estimate the best option for each of these model components.

The simplest DNA rate matrix is the Jukes-Cantor model (Jukes and Cantor 1969), which has equal substitution rates between all nucleotides and equal frequencies of each nucleotide. Other models were developed to accommodate the biological processes of sequence evolution. For example, transition mutations are more likely to occur than transversions, and the K80 model (Kimura 1980) models this with equal base frequencies but different rates for transitions and transversions. Similarly, Felsenstein (1981) kept all substitution rates equal but accommodated the biochemical processes of nucleotides by allowing for uneven base

frequencies. Each DNA substitution matrix contains 6 substitution rates and 4 nucleotide frequencies, and a variety of DNA substitution models with different *Q* matrices and frequency vectors exist such as HKY85 (Hasegawa et al. 1985), TN93 (Tamura and Nei 1993) or SYM (Zharkikh 1994). The most general DNA substitution model is the general time reversible (GTR) model (Tavaré 1986), which permits different rates of substitution for each pair of nucleotides and different frequencies for each nucleotide. More complicated DNA models have been developed to accommodate heterogeneity in evolutionary processes. Mixture models accommodate heterogeneity by including multiple classes, and the likelihood at each site is calculated as a weighted sum across the classes (Pagel and Meade 2005). For example, the GHOST model (Crotty et al. 2020) accommodates rate variation by including multiple classes which each have a set of model parameters and branch lengths.

Protein *Q* matrices are similar to DNA *Q* matrices, but include a substitution rate for each pair of the 20 amino acids (Arenas 2015). The simplest protein *Q* matrix is the Bishop-Friday model, which assumes all substitution rates are equal and all amino acid residue frequencies are equal (Bishop et al. 1987, 1997). Many protein models are determined using large empirical protein datasets, including the Dayhoff (Dayhoff et al. 1978; Kosiol and Goldman 2005) and WAG (Whelan and Goldman 2001) models. Models are available for specific parts of the genome such as the mitochondrial genome (Le et al. 2017) or for specific clades such as plants (Minh et al. 2021), insects (Minh et al. 2021) or viruses (Dimmic et al. 2002; Nickle et al. 2007; Dang et al. 2010). Custom empirical amino acid substitutions can be estimated from protein alignments using QMaker (Minh et al. 2021). Other protein models account for protein structural constraints and the physicochemical properties of amino acids (Braun 2018; Chi et al. 2018). Protein mixture models contain multiple *Q* matrices to better mimic protein evolution. For example, the EX2 model (Le et al. 2008b) contains two *Q* matrices: one for exposed amino acid residues, and one for buried residues. Other protein mixture models which model amino acid residues in different environments include the EX3 and EHO models (Le et al. 2008b). Other mixture models accommodate differences in substitution rate across sites, such as the four-matrix mixture models LG4M and LG4X which have different *Q* matrices for sites that evolve at different rates (Le et al. 2012). Profile mixture models assume that there are distinct classes of sites that differ in state frequency (Baños et al. 2024). The most well-known profile mixture model is the CAT model (Lartillot and Philippe 2004), which uses mixtures of stationary probability classes to account for different evolutionary pressures in different biochemical environments. The CAT model is applied during Bayesian inference with a Markov Chain Monte Carlo framework, which simultaneously infers the number of frequency vectors, the frequency vectors, the site classes, the tree, and the branch lengths (Lartillot and Philippe

2004). The CAT model requires an alignment large enough to characterise patterns of homogeneity, but small enough that convergence is reached (Le et al. 2008a; Baños et al. 2024). The C10-C60 models are variants of the CAT model developed for maximum likelihood inference, where each model is a mixture of precomputed frequency profiles (Le et al. 2008a).

Some models of substitution explicitly account for rate heterogeneity across sites (RHAS), such as the GHOST model mentioned above (Crotty et al. 2020). Rate heterogeneity can also be added to a *Q* matrix model such as the nucleotide GTR model or the amino acid GTR20 model. The most commonly used models of RHAS are the invariant sites model (Hasegawa et al. 1985), the discrete gamma model (Yang 1994), and the free rate model (Yang 1995; Soubrier et al. 2012). The invariant site model assumes that a portion of the sites in the alignment do not change (Hasegawa et al. 1985). The discrete gamma model uses a set number of categories to approximate a gamma distribution of rates, allowing sites to evolve at different rates (Yang 1994). The invariant site model and the discrete gamma model can be combined to model both a portion of invariant sites and different rates for the variable sites (Gu et al. 1995). The free rate model also estimates a number of categories with different substitution rates, but relaxes the assumption that the rates must fit a gamma distribution (Yang 1995; Soubrier et al. 2012).

Molecular phylogenetics studies take genetic samples at the end point of a process, and attempt to infer the process that occurred. Box (1979) famously states “All models are wrong, but some models are useful”. It is impossible to create a model that fully captures the evolutionary history of a group using genetic sequences. As a result, there are a number of errors that can impact the accuracy of tree inference.

ii Systematic Bias in Phylogenetic Methods

Phylogenetic models cannot capture the full complexity and stochasticity of biological processes. Instead, phylogenetic models aim to capture the most important aspects of the evolutionary process (Kelchner and Thomas 2007; Brown and Thomson 2018). While stochastic error can be mitigated by increasing the size of datasets (Rokas and Carroll 2005), systematic error occurs due to the limitations of current phylogenetic methods to resolve ancestral relationships (Philippe et al. 2005; Steenwyk et al. 2023). Systematic bias can be introduced by analytical factors such as taxon sampling, site/gene filtering, ortholog misidentification, alignment error, and model misspecification (Rokas and Carroll 2005; Philippe et al. 2011b; Tan et al. 2015; Molloy and Warnow 2018; Redmond and McLysaght 2021a; McCarthy et al. 2023).

In a maximum likelihood framework, increasing the complexity of phylogenetic models by incorporating more parameters introduces a cost – the variance associated with those parameters will increase, and consequently a more complex model is not always better (Burnham and Anderson 2002; Kelchner and Thomas 2007). A wide range of substitution models have been developed for both DNA and amino acid datasets, with each model designed to capture aspects of the evolutionary process that impact tree inference. Generally, the model of evolution for a particular dataset is determined by testing the fit of hundreds or even thousands of models, before selecting the model that has the best relative fit according to a certain statistical test (Kelchner and Thomas 2007; Jermiin et al. 2020). These algorithms are unable to pick the best model if it is not included in the list of models to test, in which case the model assumptions will be violated by the evolutionary history of the dataset (Jermiin et al. 2020). Many software programs have been developed to assess model fit such as MODELTEST (Posada and Crandall 1998), jModelTest (Darriba et al. 2012), ModelFinder (Kalyaanamoorthy et al. 2017), PartitionFinder (Lanfear et al. 2017), and ModelTeller (Abadi et al. 2020). These programs generally apply statistical tests to assess relative model fit, commonly the Akaike information criterion (AIC) (Akaike 1974), the Akaike information criterion with correction for small sample size (AICc) (Sugiura 1978; Hurvich and Tsai 1989), or the Bayesian information criterion (BIC) (Schwarz 1978). Most model selection does not include a goodness-of-fit test, and consequently the best-performing model is not guaranteed to have adequate fit (Gatesy 2007).

Allowing the data to reject the model is a vital step of data analysis and phylogenetics (Goldman 1993). An ideal phylogenetic method will include a goodness-of-fit test to allow the data to reject the model (Penny 1982; Goldman 1993). A number of methods have been developed to test the adequacy of substitution models and the adequacy of inferred trees. However, these approaches are rarely included when estimating phylogenetics from empirical datasets (Brown and Thomson 2018; Jermiin et al. 2020).

All sequence models contain assumptions about the process of evolution. The first general test for model adequacy in phylogenetics was developed by Goldman (1993), and simultaneously tests the model of sequence evolution and all other assumptions of the phylogenetic method to determine whether that model is an adequate fit for the data. Since then, other tests for model adequacy have been developed to test different aspects of phylogenetic models. Model adequacy tests are available for Bayesian inference, such as the tests developed by Brown and ElDabaje (2009) and Brown (2014) allow researchers to test the plausibility of Bayesian models. Many model adequacy tests are available to test the assumptions underlying maximum likelihood tree inference. The SRH assumptions (stationary,

reversible and homogeneous assumptions) are included in many substitution models. The SRH assumptions state that the frequencies of the character states are constant throughout the tree (stationary), that the substitution rates are identical throughout a branch (homogeneous), and that substitution rates are identical for forward and backward mutations (reversible). The adequacy of the SRH assumptions can be assessed using the matched-pairs tests of homogeneity (Ababneh et al. 2006; Naser-Khdour et al. 2019) or SeqVis (Ho et al. 2006). Other tests assess individual components of the SRH assumptions individually (Kumar and Gadagkar 2001; Duchêne et al. 2017). Applying testing model adequacy and phylogenetic assumptions can inform choice of model for phylogenetic inference. For example, the GHOST model models sequences that evolved heterogeneously by incorporating a mixture of trees and a mixture of models (Crotty et al. 2020). Sequences that evolved non-reversibly would benefit from non-reversible models such as the Lie-Markov models for nucleotide sequences (Woodhams et al. 2015) or the models developed in nQMaker for amino acid sequences (Dang et al. 2022).

Goodness-of-fit tests have also been developed to assess the adequacy of inferred phylogenetic trees. Different topology tests such as the KH, SH, and approximately unbiased tests have been developed to compare multiple trees (Hasegawa and Kishino 1989; Kishino and Hasegawa 1989; Shimodaira and Hasegawa 1999; Shimodaira 2002). The most commonly applied is the approximately unbiased (AU) test (Shimodaira 2002), which compares a tree to a dataset to test the hypothesis that the data could come from that tree. The AU test assumes that the evolutionary history of the alignment is a single tree. This approach does not incorporate horizontal inheritance or allow gene histories to vary. The Quartet Network goodness-of-fit test was developed to test the absolute fit of a species network to the multispecies network coalescent (MSNC), which determines whether a candidate network adequately explains observed quartet frequencies (Stenz et al. 2015; Cai and Ané 2021). As trees are networks without reticulate branches, this test can also be applied to test the adequacy of any tree under the multispecies coalescent (MSC). This allows the comparison of multiple candidate networks to a single dataset, facilitating investigation into the placement and timing of reticulation events.

One assumption included in many sequence and tree models is the treelikeness assumption. Although purely treelike evolution is likely to be extremely rare, the treelikeness assumption is incorporated into many phylogenetic methods. In this thesis, I explore the treelikeness assumption from four directions and determine how violation of the treelikeness assumption impacts tree inference.

iii The Treelikeness Assumption

Many phylogenetic methods assume that every site in an alignment shares an identical evolutionary history that fits a single bifurcating tree. This is called the treelikeness assumption. The treelikeness assumption is violated by biological processes such as introgression, hybridisation, horizontal gene transfer and incomplete lineage sorting. Treelikeness may also be disrupted by analytical processes, such as alignment error, site filtering, or model misspecification. When these processes are not considered in phylogenetic methods, the accuracy of the estimated phylogenetic tree may be decreased.

The treelikeness assumption is incorporated into many phylogenetic methods. The treelikeness assumption is built into concatenation methods, where each gene in an alignment is concatenated into a single supermatrix. This ignores any underlying heterogeneity between and within genes that arises due to biological evolutionary processes and could decrease tree accuracy (Bryant and Hahn 2020). Concatenated maximum likelihood methods are inconsistent when genes have conflicting evolutionary histories, and when species trees have closely spaced divergences (Kubatko and Degnan 2007; Roch and Steel 2015; Mendes and Hahn 2018).

Coalescent methods, which include Bayesian methods and some summary methods (also known as two-step or gene tree/species tree methods), allow the evolutionary history of each gene to vary independently according to a specified model of population genetics. However, these methods generally assume that the evolutionary history of each locus is treelike. Empirical studies have shown that evolutionary history can vary within a gene, with different exons from the same gene having different evolutionary histories (Mendes and Hahn 2016; Scornavacca and Galtier 2017; Mendes et al. 2019). Smith et al. (2020) analysed genes from 13 empirical datasets covering a broad range of clades and found that 0.6 – 100% of genes contained intragenic conflict. Including genes with multiple underling evolutionary histories violates the assumptions of the coalescent and has previously been termed “concatalescence”, a term which describes the hybrid method of using concatenated genes with a coalescent tree estimation method (Gatesy and Springer 2013; Springer and Gatesy 2016). As previous studies indicate that many loci used for coalescent tree estimation violate the treelikeness assumption (Smith et al. 2020), the accuracy of coalescent tree inference may be negatively impacted.

Some modern phylogenetic methods directly allow a single sequence alignment to have multiple underlying evolutionary histories. Bayesian methods that include the multispecies

coalescent allow genes to have different evolutionary histories. The multispecies coalescent (MSC) is a stochastic model that describes the genealogical history of sequences within a population being traced backwards through time (Rannala et al. 2020). Methods such as StarBEAST (Ogilvie et al. 2017; Douglas et al. 2022) and BPP (Flouri et al. 2018) include the MSC, which allows the evolutionary history of different genes to vary, and explicitly accommodates coalescent processes such as incomplete lineage sorting. One critique of the MSC is that it assumes an individual locus is non-recombining, so analysing genes (which can span vast regions of the genome) violates the model assumption (Gatesy and Springer 2013; Springer and Gatesy 2016). Simulation studies suggest that recombination impacts tree inference when the level of incomplete lineage sorting is high (Lanier and Knowles 2012). The levels of recombination and incomplete lineage sorting within a dataset should be considered before applying these methods.

Phylogenetic networks extend trees by allowing for reticulate evolutionary events including recombination, hybridisation, gene duplication or horizontal gene transfer (HGT). There are multiple types of phylogenetic networks, all designed to identify and represent conflicting phylogenetic signals. Split networks such as Neighbor-Net (Bryant and Moulton 2004) or split decomposition (Bandelt and Dress 1992; Dopazo et al. 1993) don't explicitly represent the evolutionary history of a set of taxa, but instead represent incompatible or ambiguous signals (Huson 1998; Huson and Bryant 2006). Alternatively, reticulate networks such as hybridisation or recombination networks explicitly represent the evolutionary history of a group, including reticulate events such as hybridisation, HGT, or introgression (Huson and Bryant 2006). Multiple programs have been developed to infer reticulate networks, including the maximum likelihood method NetRAX (Lutteropp et al. 2022), and the MSNC methods SNaQ (Solís-Lemus and Ané 2016; Solís-Lemus et al. 2017) and NANUQ (Allman et al. 2019). Phylogenetic networks are a common way to estimate, interpret or communicate conflicting signals within an alignment. However, many network methods are limited to small numbers of taxa due to the computational demands of network inference (Blair and Ané 2020).

Another recently-developed method is the Mixtures Across Sites and Trees (MAST) model, which specifically relaxes the treelikeness assumption by modelling the underlying evolutionary history of a group as a mixture of 2 or more bifurcating trees (Wong et al. 2024). Each tree in the mixture has an independent topology. Other parameters such as branch lengths, substitution model, model of rate heterogeneity, or state frequency may be unlinked and therefore separate for each tree, or linked between trees (Wong et al. 2024). MAST differs from explicit phylogenetic networks as the mixture of trees approach is agnostic to the cause or pattern of conflicting phylogenetic signal, and thus is able to represent almost any biological

process or analytical error that reduces treelikeness. MAST therefore facilitates concatenated analysis of sequence alignments with reticulate evolutionary histories.

Most phylogenetic analyses are performed to identify the primary tree, which represents the evolutionary history of the majority of the sites within a multiple sequence alignment (Baum, 2007). These phylogenetic datasets often have complicated evolutionary histories. However, current phylogenetic methods incorporate the treelikeness assumption, so applying these models to phylogenetic datasets may result in incorrect or misleading estimates of phylogenetic trees. Therefore, identifying the primary tree of a multiple sequence alignment is more complicated than estimating the phylogenetic tree using existing methods. We need to understand the extent to which violation of the treelikeness assumption impacts tree estimation, and we need methods that can handle non-treelike data.

iv The Animal Tree of Life

The group of all animals, the Metazoa, consists of five clades: Porifera, Ctenophora, Placozoa, Cnidaria, and Bilateria. Sponges are grouped into the clade Porifera, which has traditionally been considered the oldest and structurally simplest Animal clade (Borchiellini et al. 2000, 2008). Ctenophora contains the comb jellies, which are free-living marine organisms with diverse morphology (Jékely et al. 2015). Placozoa is a small clade of tiny multicellular marine animals (Schierwater et al. 2021). Cnidaria comprises aquatic animals such as jellyfish, sea anemones and corals (Kayal et al. 2013). Finally, Bilateria include all animals that have bilateral symmetry during embryonic development, such as vertebrates, arthropods and molluscs (Dunn et al. 2014). The Metazoa is an ancient clade estimated to be 650 – 800 million years old (dos Reis et al. 2015; Erwin 2015). During the Ediacaran or Cambrian era, the metazoan clade underwent a rapid radiation, resulting in a dramatic expansion in the number of species and body plans seen in the fossil record (Budd 2008; Erwin et al. 2011; dos Reis et al. 2015).

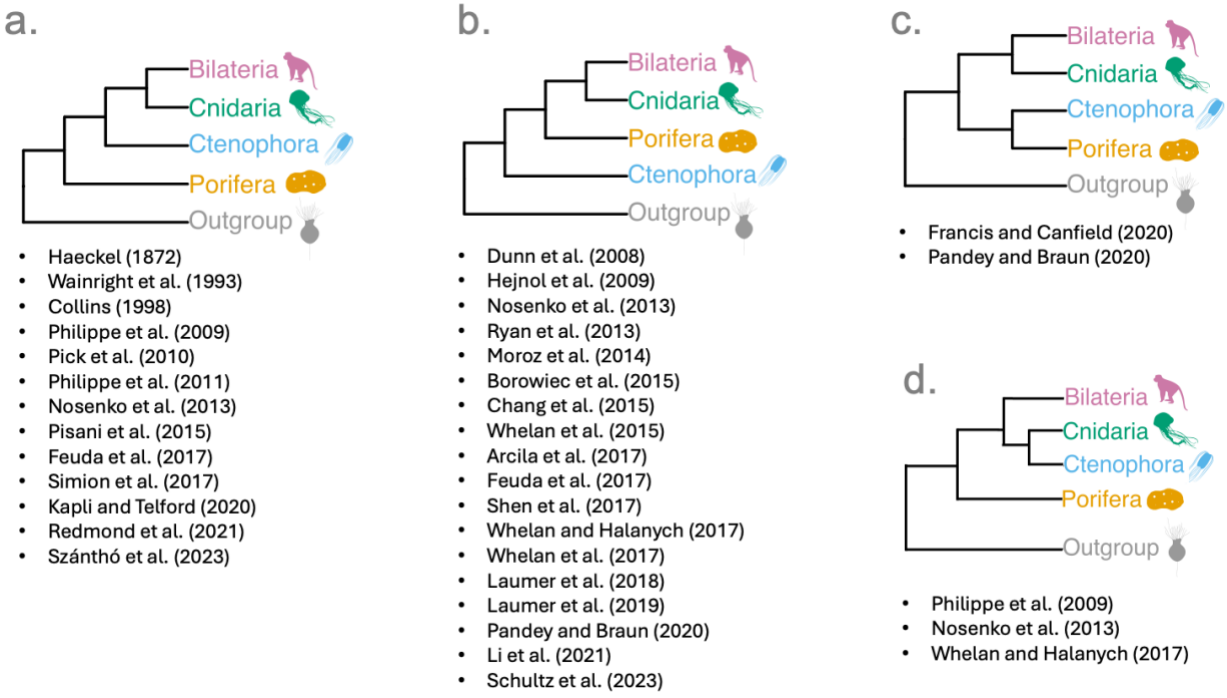


Figure 2: Different hypotheses for the relationships among the major groups of Metazoa

Each tree represents a single hypotheses of animal evolution. Each hypothesis is presented with a selection of published papers (not an exhaustive list) which include a phylogenetic tree supporting that hypothesis. When a paper included phylogenetic trees supporting different hypotheses of evolution, that paper was listed under each topology.

a. Porifera diverges first (Haeckel 1872; Wainright et al. 1993; Collins 1998; Philippe et al. 2009, 2011b; Pick et al. 2010; Nosenko et al. 2013a; Pisani et al. 2015; Feuda et al. 2017a; Simion et al. 2017a; Kapli and Telford 2020; Redmond and McLysaght 2021a; Szánthó et al. 2023).

b. Ctenophora diverges first (Dunn et al. 2008; Hejnol et al. 2009; Nosenko et al. 2013a; Ryan et al. 2013; Moroz et al. 2014; Borowiec et al. 2015; Chang et al. 2015; Whelan et al. 2015b, 2017a; Arcila et al. 2017; Feuda et al. 2017a; Shen et al. 2017; Whelan and Halanych 2017; Laumer et al. 2018a, 2019a; Pandey and Braun 2020; Li et al. 2021; Schultz et al. 2023).

c. A monophyletic clade consisting of Porifera and Ctenophora diverges first (Francis and Canfield 2020; Pandey and Braun 2020).

d. Porifera diverges before all other animals, and Cnidaria and Ctenophora form a monophyletic clade named *Coelenterata* (Philippe et al. 2009; Nosenko et al. 2013a; Whelan and Halanych 2017).

The question of which metazoan clade diverged first, before all other animals, is of particular interest (Figure 2). Historically, the sponge clade Porifera was considered the sister group to all other metazoans based on morphological analyses and early molecular studies (Haeckel 1866; Wainright et al. 1993; Collins 1998; Reynolds 2019). However, later phylogenetic analyses identified the comb jelly clade, Ctenophora, as the sister to all other metazoans (Dunn et al. 2008; Hejnol et al. 2009). Studies have since found support for a range of groups as the sister to all other metazoan clades, including Porifera (Pisani et al. 2015; Arcila et al. 2017; Simion et al. 2017a; Redmond and McLysaght 2021a); Ctenophora (Ryan et al. 2013; Moroz

et al. 2014; Borowiec et al. 2015; Shen et al. 2017); and a monophyletic clade of Porifera and Ctenophora (Shen et al. 2017; Francis and Canfield 2020). The uncertainty around the metazoan phylogeny is ongoing and impacts other fields such as medical research, as it limits understanding of the evolution of complex traits like nervous or digestive systems (Philippe et al. 2009; King and Rokas 2017).

Multiple factors in the evolutionary history of metazoans contribute to the difficulty of resolving their phylogenetic history. First, the metazoan clade underwent a rapid radiation, resulting in short branches at the base of the tree. These short branches result in discordant evolutionary histories between genes, due to evolutionary processes such as incomplete lineage sorting, hybridisation, and introgression (King and Rokas 2017; Pandey and Braun 2021). Second, the metazoan clades diverged over 500 million years ago. Therefore, metazoan taxa are distantly related, reducing the available number of shared orthologous loci between taxa (Pett et al. 2019; McCarthy et al. 2023). Third, the deep timescales involved when examining relationships between metazoan clades complicates tree inference.

The deep divergence time of the metazoan root results in differences between taxa in substitution rate and GC content (Gouy et al. 2015; King and Rokas 2017). Biases in GC content have been shown to impact tree accuracy (Romiguier et al. 2013; Shen et al. 2016; Romiguier and Roux 2017; Rousselle et al. 2018). If models that don't allow for heterogeneous base frequencies are used on datasets with GC bias, unrelated taxa with similar base frequencies will be incorrectly placed together (Phillips et al. 2004). Heterogeneous substitution rates can result in long branch attraction due to mutational saturation (Lartillot et al. 2007; Kapli et al. 2020). Two lineages with naturally high substitution rates will appear to be closely related as they have similar characters at each site, but this shared similarity is not due to a shared common ancestor. Most phylogenetic methods will identify the convergent signal as homologous, resulting in a tree with the fast-evolving species incorrectly placed closely together (Lartillot et al. 2007; Simion et al. 2020). As the branch leading to the outgroup is generally long, species with high substitution rates can be pulled towards the root of the tree (Philippe and Laurent 1998). The branch leading to the Ctenophore clade is long compared to other branches within the metazoan tree due to the high substitution rate of the Ctenophora genome (Kohn et al. 2012; Wang and Cheng 2019), which could lead to long branch attraction and erroneously pull the Ctenophore clade towards the root of the tree (Pisani et al. 2015; Kapli and Telford 2020).

There is substantial conflicting signal present within metazoan phylogenetic datasets. Shen et al. (2017) investigated 8 phylogenetic datasets used to estimate the metazoan tree, and found

that that 42.5 – 69.7% of genes and 39.8 – 59.6% of sites supported Ctenophora as sister to all other metazoan clades, with the rest supporting either Porifera or a monophyletic clade of Ctenophora and Porifera as the sister group to all other metazoan clades. This could explain why model choice has been shown to impact metazoan tree topology, with partitioned site-homogeneous models supporting Ctenophora-sister (Ryan et al. 2013; Moroz et al. 2014; Whelan et al. 2015b, 2017a), and site-heterogeneous models (e.g. the CAT model) supporting Porifera-sister (Pisani et al. 2015; Feuda et al. 2017a; Simion et al. 2017a). Previous studies have found that filtering sites and genes also impacts the topology of the metazoan tree. Francis and Canfield (2020) showed that removing the 1.7% of sites strongly supporting either the Ctenophora-sister or the Porifera-sister hypothesis resulted in a highly-supported tree with a monophyletic clade of Ctenophora and Porifera as the sister to all other metazoans. Nosenko et al. (2013a) found that varying gene sampling and choice of outgroup resulted in differing but well-supported topologies. Finally, McCarthy et al. (2023) examined 5 matrices that had previously been used to estimate the metazoan tree, and found that in 2/5 cases the sister to all metazoans changed after removing loci with insufficient evidence for orthology. These results demonstrate that minor analytical changes can bias the resulting metazoan tree topology, suggesting the presence of non-treelike data either due to a complex evolutionary history (e.g., reticulate evolution, homoplasy, or incomplete lineage sorting) or systematic bias (e.g., model misspecification).

In this thesis, I use the Metazoa as a case study. I selected the Metazoa because this clade has an interesting evolutionary history and unresolved phylogeny. Metazoan phylogenetic studies are good examples of typical approaches to tree inference. First, the metazoan tree has been investigated many times, leading to a wealth of published datasets available for reanalysis (see Figure 2). The breadth of phylogenetic studies that have inferred the metazoan tree allows me to compare my results to previous studies assessing systematic bias on the same clade, and to contextualise the results of my analyses. Second, metazoan datasets are a similar size to many empirical phylogenetic datasets, with comparable numbers of taxa, sites, and loci. This will ensure that any methods I develop are sufficient for datasets being used for phylogenetic inference. Third, multiple studies have identified substantial heterogeneous signal within metazoan datasets (Arcila et al. 2017; Shen et al. 2017; Redmond and McLysaght 2021a; Szánthó et al. 2023). This heterogeneous signal clearly violates the treelikeness assumption, making the Metazoa a good candidate to explore methods to quantify, manage, and mitigate treelikeness.

v Motivation and aims

In this thesis, I aim to reduce the systematic bias associated with estimating phylogenetic trees from non-treelike data by developing new approaches to quantify and manage non-treelike data. In particular, I focus on systematic bias within the metazoan tree of life and determine the impact of non-treelike data on tree topology within this clade. In addition, I propose recommendations for estimating accurate phylogenetic trees from empirical datasets, which are likely to contain some level of non-treelike evolutionary history due to the complex and stochastic processes of evolution.

In Chapter One, I present a new test designed specifically to quantify the treelikeness of any sequence alignment. The new test, called the tree proportion, represents the extent to which the evolutionary history of a given sequence alignment can be explained by a single bifurcating tree. The tree proportion is calculated by determining the extent to which a single maximum spanning tree can capture the information present in a single split network. A split is a bipartition of a set of taxa, and a split network is an implicit representation of evolution used to represent conflicting phylogenetic signal within a dataset (where parallel edges are used to represent splits computed from the data) (Huson and Bryant 2006). Given any split network, a maximum spanning tree is the tree with the largest (maximum) sum of branch weights. A tree proportion value of 1 indicates that an alignment is perfectly treelike (i.e., every split in the split network is present in the maximum spanning tree), and the tree proportion will decrease as the treelikeness of an alignment decreases. I also present a parametric bootstrap for use which can be used to statistically test the null hypothesis of treelikeness for any empirical alignment. In addition to presenting a new test for treelikeness, I perform a comprehensive benchmarking of 7 existing tests for treelikeness plus the newly proposed tree proportion. There are a variety of published tests for treelikeness with different underlying approaches, but some have inconsistent performance or have low power, and none have become widely adopted. In addition, there is a lack of benchmarking to compare the performance of each test. To systematically vary the treelikeness of simulated alignments, I present two simulation schemes: one with dramatically reduced treelikeness caused by increasing the number of randomly generated evolutionary histories present within a single alignment, and one designed to mimic biological processes by incorporating a single introgression event into each alignment. I show that three metrics performed adequately as metrics for treelikeness under a range of scenarios: the tree proportion, δ plots, and site concordance factors. This chapter includes examples quantifying the treelikeness of simulated and empirical alignments, and demonstrates how to apply and interpret tests for treelikeness.

In Chapter Two, I investigate the impact of filtering recombinant loci on species tree accuracy for 4 empirical phylogenetic datasets. Recombination violates the treelikeness assumption and therefore has the potential to reduce species tree accuracy. My study was designed to check whether species tree topology does change when recombinant genes are included or removed from empirical phylogenetic datasets, and assess the frequency and extent of topological changes. To do this, I selected 4 empirical phylogenetic datasets and 3 tests for recombination. The four datasets all contain loci consisting of concatenated exons, and were selected to include both shallow and deep datasets from both animals and plants: tomatoes (Pease et al. 2016a), primates (Vanderpool et al. 2020b), green plants (Leebens-Mack et al. 2019a), and metazoans (Whelan et al. 2017a). The three tests for recombination have been highly cited, previously validated, and have different theoretical approaches: the maximum chi-squared method (MaxChi) (Maynard Smith 1992), GENECONV (Sawyer 1989, 2000), and the Pairwise Homoplasy Index (PHI) (Bruen et al. 2006). For each combination of dataset and test for recombination, I grouped the loci into two subsets: loci identified as putatively recombinant by that test, and loci not identified as putatively recombinant by that test. I then estimated both a maximum likelihood tree in IQ-TREE (Minh et al. 2020b) and a two-step tree in ASTRAL (Mirarab et al. 2014; Zhang et al. 2018b). In general, the impact of excluding loci with evidence of recombination on species tree topology is small. However, in some specific cases exclusion of putatively recombinant loci results in biologically and statistically significant differences in tree topology.

In Chapter Three, I relax the treelikeness assumption and allow a single phylogenetic dataset to have multiple underlying evolutionary histories by applying the Mixtures Across Sites and Trees (MAST) model. This study is designed to assess the adequacy of a single-tree model in the complex and often-contested phylogeny of all animals. First, I assess the consistency of tree topology under a single-tree model by estimating 364 trees from each combination of 26 models of sequence evolution and 14 empirical metazoan phylogenetic datasets. I then apply the MAST method to the same 14 datasets to relax the assumption of treelikeness. The MAST method (Wong et al. 2024) uses a mixture of bifurcating trees to represent multiple evolutionary histories for a single concatenated alignment. I defined two MAST analyses: the 2-tree and the 5-tree model. The 2-tree model contains two hypotheses of metazoan evolution: Ctenophora-sister (i.e., Ctenophora as the first clade to diverge) (Figure 3a) or Porifera-sister (Figure 3b). The 5-tree model contains 5 hypotheses of evolution: Ctenophora-sister (Figure 3a); Porifera-sister (Figure 3b); a monophyletic clade consisting of Ctenophora and Porifera as sister to all other animals (Figure 3c); Ctenophora-sister with paraphyletic Porifera (Figure 3d); and Porifera-sister with paraphyletic Porifera (Figure 3e). I find that the multi-tree model is

overwhelmingly preferred over the single-tree model, and propose that a single tree is inadequate to represent the evolutionary history of the Metazoa.

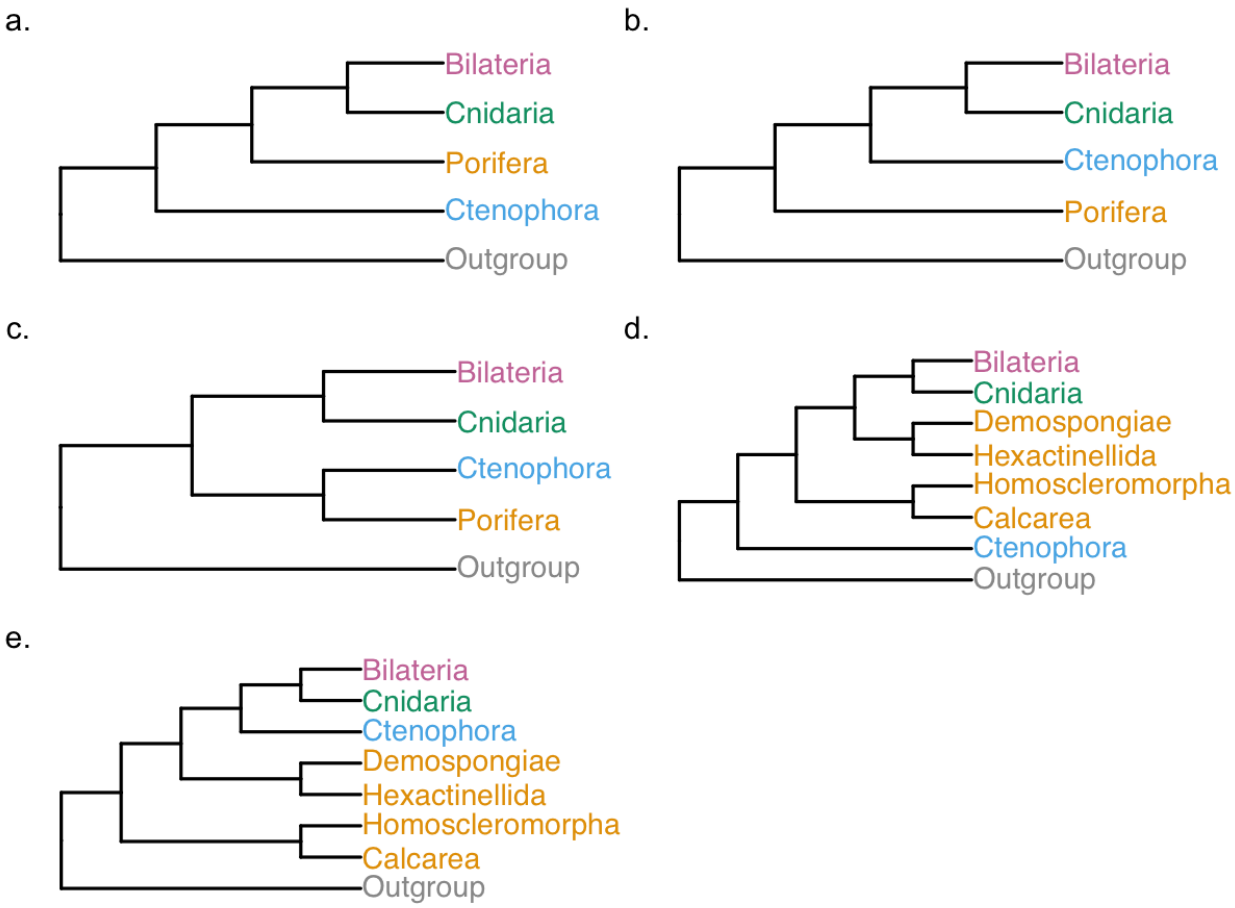


Figure 3: The five different hypotheses for the relationships among the major groups of Metazoa analysed within this thesis.

a. Ctenophora diverges before all other animals.

b. Porifera diverges before all other animals.

c. A monophyletic clade consisting of Ctenophora and Porifera diverges before all other animals.

d. Ctenophora diverges before all other animals. Porifera is paraphyletic. The Porifera clades Homoscleromorpha and Calcarea form a monophyletic clade, which diverges before the monophyletic clade consisting of the Porifera clades Demospongiae and Hexactinellida.

e. Porifera is paraphyletic, with one of two Porifera clades diverging before all other animals. The Porifera clades Homoscleromorpha and Calcarea form a monophyletic clade, which diverges first. The Porifera clades Demospongiae and Hexactinellida form a monophyletic clade.

In Chapter Four, I investigate the causes of conflicting signal within the metazoan tree of life. In this chapter, I quantify the phylogenetic signal within different genes from the same dataset, with the aim of determining whether incomplete lineage sorting contributes to the low topological consistency of metazoan species trees. The short divergence time between Ctenophora and Porifera contributes to the difficulty of resolving which clade diverged from all other metazoans first. Previous studies have suggested incomplete lineage sorting contributes to the difficulty of resolving these relationships, due to the rapid radiation of metazoan species resulting in short branches at the base of the metazoan tree (King and Rokas 2017; Francis and Canfield 2020). I apply concordance factors to 12 previously published empirical phylogenetic datasets to determine whether incomplete lineage sorting contributes to the inference of conflicting metazoan topologies. I estimate gene and quartet concordance factors at key branches of the metazoan tree. Model misspecification has previously been suggested as a cause of the Ctenophora-sister topology, with site-heterogeneous models fitting metazoan datasets better than Partitioned models (Kapli and Telford 2020; Redmond and McLysaght 2021a). To account for model misspecification, I performed my concordance factors analyses with both a Partitioned model and the C60 model. I identify substantial conflicting signal within each of the 12 metazoan datasets, which is consistent with incomplete lineage sorting at the base of the metazoan tree. My results suggest that traditional phylogenomic approaches are unlikely to resolve the evolutionary history of the Metazoa, and alternate approaches designed to detect deep phylogenetic signal may be required.

Chapter One:
A Comparison of Methods for Quantifying Treelikeness

Caitlin Cherryh[1]*, Bui Quang Minh[2], Robert Lanfear[1]

[1] Division of Ecology and Evolution, Research School of Biology, Australian National University, Canberra, Australia

2 School of Computing, Australian National University, Canberra, Australia

* Corresponding author: caitlin.cherryh@anu.edu.au

Contributions:

Caitlin Cherryh designed the simulations, wrote the R scripts, performed the simulations and empirical analysis, interpreted the results, and drafted the manuscript. Minh Bui and Robert Lanfear assisted with conceptual development, experimental design, and editorial comments.

1.1 Abstract

Many phylogenetic methods assume that all sites in an alignment share an evolutionary history that conforms to a single bifurcating tree, i.e., the treelikeness assumption. However, the treelikeness assumption is often violated in empirical data by evolutionary processes such as incomplete lineage sorting, introgression, recombination, and horizontal gene transfer. Many approaches have been used to quantify, visualize, and test for treelikeness, but there is a lack of benchmarking to compare these tests and quantify their behaviour in different situations, and few tests provide metrics that are intuitive to interpret. In this chapter, I address these issues by introducing a new measure of treelikeness I call the tree proportion, which represents the extent to which a single bifurcating tree explains the evolutionary history of a single alignment. I then benchmark the behaviour of this measure and 7 existing measures of treelikeness using simulated data with increasingly non-treelikeness by concatenating alignments generated under random trees or under a multispecies coalescent model with incomplete lineage sorting and introgression. My benchmarking shows that three measures are particularly useful for estimating the treelikeness of an alignment: the tree proportion, δ plots, and the site concordance factor. These tests showed a correlation with treelikeness under a wide range of scenarios. The remaining metrics were not consistently correlated with the treelikeness of an alignment, and I discuss the underlying reasons for and implications of these results.

Keywords: Phylogenetic methods, Phylogenetic inference, Model adequacy, Gene flow, Parametric bootstrap

1.2 Introduction

A phylogenetic tree is an estimate of the evolutionary relationships between a group of species, populations, or individuals (Simion et al. 2020). As all aspects of biology have been shaped by evolutionary processes, many areas of biology (for example conservation, behavioural ecology, and epidemiology) require accurate phylogenetic trees (Jermiin et al. 2020). Due to the complexity of the evolutionary process, all phylogenetic methods make some simplifying assumptions. One common assumption is that all sites in a sequence alignment share the same evolutionary history, and that this history conforms to a single bifurcating tree. This is called the treelikeness assumption (Jermiin et al. 2020). Perfectly treelike empirical alignments may be rare due to biological processes such as incomplete lineage sorting (ILS) (Rokas and Carroll 2006), hybridization (Sanderson et al. 2023), introgression (Steenwyk et al. 2023), or horizontal gene transfer (HGT) (Gogarten and Townsend 2005). As a result, even short

alignments such as those representing a single gene may have an evolutionary history represented by more than one bifurcating tree (Mallet et al. 2016; Scornavacca and Galtier 2017; Mendes et al. 2019).

The treelikeness assumption is incorporated into many methods for estimating phylogenetic trees. For example, concatenation methods assume that all sites from all loci in a concatenated alignment share a single evolutionary history. This approach has been criticized because concatenating alignments from loci with different evolutionary histories clearly violates the treelikeness assumption, and the histories of individual loci may vary dramatically, potentially resulting in incorrect phylogenetic inferences (Weisrock et al. 2012; Zhao et al. 2016; Shi and Yang 2018; Wu et al. 2018). Coalescent methods incorporate non-treelikeness due to ILS by allowing the topology of each gene tree to vary under a specific model of evolution. In other words, they to some extent relax the treelikeness assumption between loci. However, these methods still make the treelikeness assumption within loci – i.e., they assume that all the sites from a single locus share a single evolutionary history. In empirical datasets, different exons within the same gene have been shown to have different evolutionary histories (Scornavacca and Galtier 2017; Mendes et al. 2019), i.e., the treelikeness assumption may be violated within individual loci. This problem has been termed “concatalescence” and can undermine the accuracy of coalescent analyses by reducing the accuracy of gene trees (Gatesy and Springer 2013, 2014). Performing a coalescent analysis becomes a trade-off between using loci large enough to contain sufficient phylogenetic signal and small enough to be a single non-recombining unit (Springer and Gatesy 2018). As a result, both concatenation and coalescent methods can be vulnerable to errors introduced by violation of the treelikeness assumption. Quantifying the level of conflicting signal within an alignment prior to tree estimation facilitates appropriate selection of phylogenetic method and phylogenetic models.

In this chapter I introduce a new measure for quantifying treelikeness called the “tree proportion”, which provides an intuitive measure of the extent to which a single alignment is treelike. The tree proportion describes the extent to which the evolutionary history of an alignment can be captured by a single bifurcating tree. A tree proportion of 1 indicates that an alignment is perfectly treelike, and a tree proportion of 0 indicates that there is no treelike signal in the dataset at all. The calculation of tree proportion works by determining the extent to which a single maximum spanning tree captures the information present in a split network (see Methods for details).

Despite the potential utility of tests for treelikeness, a key limitation in choosing a test for treelikeness is the current lack of benchmarking. To the best of my knowledge, there has been

no study comparing multiple metrics for treelikeness on a set of well-characterized datasets in which treelikeness is systematically varied. In this chapter, I address this by benchmarking eight tests for treelikeness against a wide range of simulated datasets in which I systematically vary treelikeness. Although simulations cannot fully capture the complexity of biological evolutionary processes, I argue that any treelikeness metric that cannot identify large changes in treelikeness in simulated data is likely to perform poorly on empirical datasets. In addition to my benchmark, I present a parametric bootstrap approach to determine whether the null hypothesis of treelikeness holds for any given sequence alignment. I then use this parametric bootstrap along with the three most useful tests of treelikeness identified by my benchmarking to assess the treelikeness of two empirical metazoan datasets. The first dataset was developed by Whelan et al. (2017a) to estimate the evolutionary history of the Metazoa, and the second was a filtered version of that same dataset by McCarthy et al. (2023), who removed genes with insufficient evidence of orthology.

Together, my results provide information that phylogeneticists can use to test the treelikeness of any phylogenetic dataset. Allowing the data to reject model assumptions such as treelikeness is an important step in the phylogenetic protocol that is often skipped (Brown and Thomson 2018; Jermiin et al. 2020). If datasets are incorrectly assumed to be treelike, the conclusions drawn from the resulting tree may be compromised (Brown and Thomson 2018; Jermiin et al. 2020). Therefore, testing the treelikeness of empirical datasets prior to tree estimation will help inform choice of phylogenetic methods, and hopefully result in more accurate inferences.

1.3 Existing Metrics for Quantifying Treelikeness

In this chapter, I differentiate between a “tree” (i.e., a phylogenetic tree with branch lengths and topology) and the “tree topology” or simply “topology” (i.e., a phylogenetic tree without branch lengths but with defined relationships between labelled tips).

A range of tests for treelikeness with different methodologies have been proposed (Table 1). These methods can be broadly grouped by the approach that they take. Early tests for treelikeness used multivariate or regression analysis. Cavalli-Sforza and Piazza (1975) published a test for treelikeness, which they define as the validity of a given tree for an observed distance matrix. This test uses multivariate analysis to ask to what extent using a tree to represent a distance matrix results in a loss of information. Similarly, Cunningham (1978) suggested a method for testing the validity of a tree as a representation of an alignment by comparing a distance matrix estimated from the sequences to one calculated from the tree. Both tests are designed to compare the validity of a tree as the representation for a dataset. While these models will detect violation of the treelikeness assumption, they will also detect violations of other assumptions which would decrease the fit of a single tree for a given dataset, such as model misspecification.

Table 1: Summary of existing metrics for treelikeness discussed in this manuscript.

Test statistic	Citation	Method
Cavalli-Sforza-Piazza metric	(Cavalli-Sforza and Piazza 1975)	Regression analysis
Cunningham metric	(Cunningham 1978)	Regression analysis
Eigen quartet metric	(Eigen et al. 1988)	Quartet
δ plot	(Holland et al. 2002)	Quartet
Q-residual	(Gray et al. 2010)	Quartet
Likelihood mapping	(Strimmer and von Haeseler 1997)	Quartet
Treeness triangles	(White et al. 2007)	Ternary plot
Split networks	(Huson and Bryant 2006)	Split network
Network Treelikeness Test	(Huson and Bryant 2006)	Split network
TIGER	(Cummins and McInerney 2011)	Character-state distribution
Site concordance factors	(Minh et al. 2020a; Mo et al. 2023)	Concordance factor
Reticulation index	(Cai et al. 2021)	Triplet frequency

Other tests use quartets to test for treelikeness. These methods work by calculating the treelikeness of individual quartets of four taxa from the dataset, and then summarizing this information across quartets. Eigen et. al. (1988) proposed one such test in which an alignment is deemed treelike if each quartet of four taxa in the alignment is itself treelike, where treelikeness is calculated by comparing the distances between each of the four sequences. For example, given a quartet of four taxa in an alignment i, j, k and l one can calculate four

distances: $d_{ij} + d_{kl}$, $d_{ik} + d_{jl}$, and $d_{il} + d_{ik}$. If these distances are not identical then the quartet is classified as non-treelike (Eigen et al. 1988). The δ plot method (Holland et al. 2002) similarly quantifies the treelikeness of an alignment using a mathematical approach that assesses the treelikeness of all possible quartets of taxa within an alignment, by quantifying the extent to which each quartet fails to pass the four-point condition (which is satisfied for four taxa i, j, k and l when there is no unique maximum value for the three sums of distances between taxa $d_{ij} + d_{kl}, d_{ik} + d_{jl}, d_{il} + d_{jk}$) (Buneman 1971; Lapointe and Kirsch 1995). However, there is no consensus on a threshold above which alignments are classified as non-treelike for the δ plot method, and as a result different studies have made chosen very different cut-off points (Grimm and Renner 2013; Short et al. 2014; Kozak et al. 2015).

Another quartet method is the Q-residual, which has been applied to cultural and linguistic evolution to quantify treelikeness (Gray et al. 2010; Syrjänen et al. 2021). The Q-residual and δ plot values are calculated similarly. Given a quartet of four taxa in an alignment i, j, k and l, one can calculate four distances: $d_{ij} + d_{kl}$, $d_{ik} + d_{jl}$, and $d_{il} + d_{jk}$. If one assumes that $d_{ij} + d_{kl} \leq d_{ik} + d_{jl} \leq d_{il} + d_{jk}$, then the δ_q score (δ plot value for this quartet) is $\delta_q = \frac{(d_{il}+d_{jk})-(d_{ik}+d_{jl})}{(d_{il}+d_{jk})-(d_{ij}+d_{kl})}$, whereas for the same quartet the formula to calculate the Q-residual is $Q - residual = \left((d_{il} + d_{jk}) + (d_{ik} + d_{jl})\right)^2$.

Finally, likelihood mapping (Strimmer and von Haeseler 1997) was designed to visualize the phylogenetic information content of an alignment (Salzburger et al. 2002; Baric et al. 2003; Steiner and Dreyer 2003) by visualising the treelikeness of individual quartets. Likelihood mapping provides a general overview of the level of conflict within quartets for any given alignment. Given any quartet of four taxa, first the maximum likelihood for each of the three unrooted tree topologies is calculated, and then the posterior probability of each tree is calculated using Bayes' theorem. Each quartet is then plotted onto a ternary graph with each vertex representing one of the three possible tree topologies. Quartets in the corners favour one tree and are deemed treelike, whereas quartets in the centre of the plot have a star-like topology and are deemed non-treelike. Likelihood mapping has been applied as a proxy for the overall treelikeness of an alignment or the phylogenetic signal of an alignment, with the proportion of treelike quartets used as a test statistic to determine the treelikeness or phylogenetic content of an alignment (Nadan et al. 2003; Nikolaev et al. 2007; Skaloud and Peksa 2010; Kim et al. 2013; Vanhove et al. 2015; Vďačný 2017; Prasanna et al. 2020; Cunha et al. 2022).

Another class of treelikeness tests are designed to help visualize conflicting signals within a dataset, without using quartets. Treeness Triangles plots (White et al. 2007) visualize the phylogenetic signal of a dataset. Given an alignment and a tree, the three vertices on the Treeness Triangle plot represent the proportion of signal supporting the exterior branches, the proportion of signal supporting interior branches, and the sum of the residual signals. Each point in the Treeness Triangle represents a single tree, allowing for comparison between different alignments or different models of evolution.

Phylogenetic networks also used to visualize conflicting signal within an alignment. A phylogenetic network is any network where taxa are represented by nodes and the relationships between taxa are represented by edges. One type of phylogenetic networks is the split network, which use splits to represent incompatibilities in a dataset. A split is a bipartition of a set of taxa – each edge within a phylogenetic tree or network defines a single split. Split networks naturally display conflicting signals and relationships within an alignment and can be used to visualize the treelikeness of a dataset by looking at the number, size, and position of conflicting splits in the network (Bryant and Moulton 2004; Huson and Bryant 2006). A split network is an implicit representation of evolution, as individual nodes within the network do not represent ancestral species. Alternatively, reticulate networks explicitly represent evolutionary events such as recombination or hybridisation by adding edges to a phylogenetic tree (Huson and Bryant 2006), and therefore nodes within a reticulate network explicitly represent ancestral species.

Split networks can serve as the basis for quantitative tests for treelikeness. The Network Treelikeness Test (Huson and Bryant 2006) determines whether it is likely that data originated from a given tree, starting by estimating a Neighbor-Net network. In the Network Treelikeness Test, a Neighbor-Net network is inferred from the dataset, and then used to construct a consensus network. A confidence network is constructed by determining the confidence interval for the weight of each split in a split network (Huson and Bryant 2006). Next, the confidence network is tested to see if it contains a given tree: i.e., that each split in the tree is in the confidence network, that the branch length for each split is contained within the confidence intervals for the weight of that split in the confidence network, and that every split in the confidence network that is not in the tree has a confidence interval containing 0 (Huson and Bryant 2006). If the above three conditions are not met, the Network Treelikeness Test rejects the null hypothesis that the data originated on a tree. Unfortunately, previous studies have shown that this test is conservative and has low power to reject non-treelike datasets (Huson and Bryant 2006).

In addition to explicit tests of the treelikeness assumption several other phylogenetic tools have been proposed to detect signals that are likely to change as the treelikeness of an alignment varies. TIGER (Cummins and McInerney 2011) is a tree-independent method to determine similarity between characters and was designed to detect rapidly evolving sites. TIGER has been applied to sort sites into categories based on evolutionary rates, and remove sites for phylogenetic analysis (He et al. 2014; Xi et al. 2014; Heikkilä et al. 2015; Burki et al. 2016; Foster et al. 2017). As treelike datasets are expected to have higher internal consistency, TIGER values have previously been applied as a proxy for treelikeness in linguistic datasets (Syrjänen et al. 2021; List 2022). Site concordance factors (sCF) (Minh et al. 2020a; Mo et al. 2023) may also be expected to vary consistently with treelikeness. For each branch in a tree, the site concordance factor represents the proportion of decisive sites which agree with that branch, where decisive sites are defined as those with sufficient information to have potentially supported that branch. Site concordance factors will tend to be high when treelikeness is high (as all sites in the alignment will share an evolutionary history) and become lower as treelikeness is reduced. Finally, the Reticulation Index (Cai et al. 2021) was designed to quantify introgression at each node of a species tree using triplet frequency. The Reticulation Index estimates the proportion of gene flow (i.e., the proportion of asymmetric triplets) at each node, resulting in an estimation of the levels of reticulate evolution across a species tree. As this method works by comparing triplets (Cai et al. 2021), the species tree must contain at least one descendant of both the donor and recipient lineage to identify the signal of gene flow.

1.4 Materials and Methods

1.4.1 Tree Proportion: A New Metric for Quantifying Treelikeness

I introduce a new metric designed to quantify treelikeness in a sequence alignment, called tree proportion (Figure 4). First, an input alignment sequence is used to estimate a phylogenetic network (such as Neighbor-Net implemented in the SplitsTree software). Second, a maximum spanning tree is constructed from the split network using a modified version of Kruskal's algorithm (Kruskal 1956) to select the set of compatible splits with the highest sum of split weights. The tree proportion metric quantifies the proportion of split weights from the split network within the maximum spanning tree.

Tree proportion is defined as follows. For a split network denoted by (S, λ) where S is the set of non-trivial splits in the network and λ is a split weight function, and a maximum spanning tree (T) the tree proportion is calculated as:

Equation 1

$$\boldsymbol{Tree\ proportion} = \frac{\sum_{S \cap T} \boldsymbol{\lambda(\sigma)}}{\sum_{S} \boldsymbol{\lambda(\sigma)}}$$

In Equation 1, the set of splits in the tree is denoted by $S \cap T$. For any split σ, the split weight (i.e., the edge length) is given by $\lambda(\sigma)$.

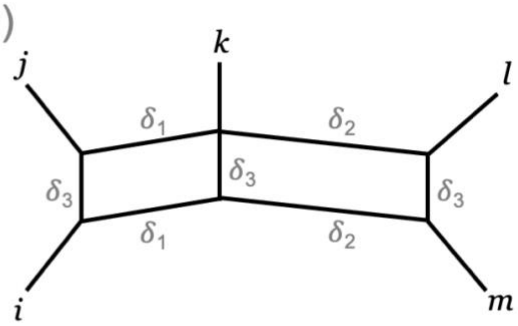
In other words, tree proportion is the proportion of the total weight of non-trivial splits in the network that are represented by the maximum spanning tree (Figure 4, Equation 1).

The R function to calculate the tree proportion of any alignment is available at the "Code" folder of the GitHub repository https://github.com/caitlinch/treelikeness-metrics/, in the scripts func_metrics.R and func_tree_proportion.R. The tree proportion function was written in R v4.3.1 (R Core Team 2018), and utilizes the R packages ape v5.7.1 (Paradis and Schliep 2019) and phangorn v2.11.1 (Schliep 2011; Schliep et al. 2017).

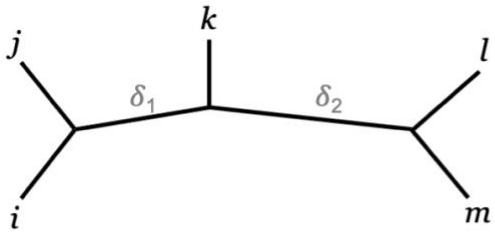
a)

i	A T T G C C
j	A T T G C A
k	T G T G C A
l	T G C A G A
m	T G C A G C

b)



c)



d)

Split	Bipartition	Weight (λ)
δ_1	$ij \mid klm$	0.4
δ_2	$ijk \mid lm$	0.6
δ_3	$im \mid jkl$	0.2

e)

$$\textit{Tree proportion} = \frac{\lambda(\delta_1) + \lambda(\delta_2)}{\lambda(\delta_1) + \lambda(\delta_2) + \lambda(\delta_3)}$$

$$\textit{Tree proportion} = \frac{0.4 + 0.6}{0.4 + 0.6 + 0.2}$$

$$\textit{Tree proportion} = 0.83$$

Figure 4: Illustration of computing tree proportion from a sequence alignment of five taxa.

a) An example of a DNA sequence alignment for 5 taxa.

b) The Neighbor-Net network with three splits ($\delta_1, \delta_2, \delta_3$) estimated from the alignment in Figure 4a.

c) A maximum spanning tree with two splits (δ_1, δ_2) constructed greedily from the Neighbor-Net network in Figure 4b by taking the set of compatible splits with the maximum possible sum of split weights.

d) The weights and bipartitions (i.e., partitions of the set of taxa into two parts) for the splits present in the Neighbor-Net network and maximum spanning tree, estimated from the alignment in Figure 4a.

e) Example calculation for the tree proportion, using the three splits in the Neighbor-Net network (Figure 4b) and the two splits in the maximum spanning tree (Figure 4c). The tree proportion is calculated by dividing the total sum of split weights from the maximum spanning tree by the sum of split weights in the Neighbor-Net network.

In principle, the tree proportion can be calculated from any combination of a network and a tree estimated from the same alignment. Here, I calculate the tree proportion by first estimating a Neighbor-Net network (Bryant and Moulton 2004) from an alignment using SplitsTree4 v4.18.2 (Huson 1998; Huson and Bryant 2006). I then estimate a maximum spanning phylogenetic tree from the Neighbor-Net network using a modified version of Kruskal's

algorithm (Kruskal 1956). Tree estimation from a set of splits has previously been proposed (Bandelt and Dress 1992; Bastkowski et al. 2014), and estimating a maximum spanning tree via greedy selection of splits has been shown to perform well for split decomposition networks (Bandelt and Dress 1992). To estimate a maximum spanning tree from the network, I use a modified version of Kruskal's algorithm (Kruskal 1956). I first extract the set of splits from the Neighbor-Net network and order them from strongest to weakest by split weight. Next, I add one split at a time, beginning with the strongest split. If the added split is compatible with the existing set of splits, I add it to the tree. If it is not compatible, the split is discarded. This continues until all splits have been processed and the maximum spanning tree has been generated. This algorithm is not guaranteed to maximise split support, but will be a good starting point to evaluate the level of conflicting signal within an alignment.

The tree proportion is simply the sum of the weights of all non-trivial splits (i.e., I exclude all splits leading to leaf nodes) in the maximum spanning tree divided by the sum of the weights of all non-trivial splits in the Neighbor-Net network (Figure 4). A tree proportion of 1 indicates that all non-trivial splits present in the network are also present in the tree, hence the alignment is perfectly treelike. As the amount of conflicting signal within an alignment increases, the tree proportion decreases towards zero.

By starting with a split network, the tree proportion method assumes that all conflicting phylogenetic signal is included in the distance matrix and hence the network. However, some loss of information in converting from an alignment to a distance matrix to a split network is inevitable (Felsenstein 2004). Therefore, the tree proportion method is only able to detect changes in treelikeness that are captured by transformation from alignment into a split network.

1.4.2 Simulations

I performed two sets of simulations designed to create alignments which varied predictably in their treelikeness. The first set of simulations concatenates alignments generated from random trees. In this case, for an alignment of a given length the minimum number of trees is 1 (in which case the alignment is treelike), and the maximum number is equal to the length of the alignment (i.e., every site is simulated from a different random tree – in this case the alignment is maximally non-treelike). The second set of simulations concatenates alignments generated from gene trees simulated under the multi-species coalescent model with introgression. The principle here is the same as with the first set of simulations, but the trees are now constrained by a biologically inspired model in which the generated trees are much more similar to each other than randomly generated trees. Thus, although both approaches will generate monotonic

changes in treelikeness, treelikeness should decrease more quickly in the former approach (using random trees) than in the latter approach (using trees from the coalescent with introgression).

1.4.2.1 Random tree simulations

The first set of simulations decreased treelikeness by increasing the number of random trees within an alignment. I simulated alignments with all combinations of the following parameters: 5, 10, 20, 50, and 100 taxa; three tree depths in substitutions per site (0.01, 0.1, 1); an alignment length of 10kbp; and 1, 2, 4, 5, 8, 10, 16, 20, 25, 40, 50, 80, 100, 125, 200, 250, 400, 500, 625, 1000, 1250, 2000, 2500, 5000, 10000 random trees per simulation (i.e., all whole-number divisors of the total alignment length). I note that for a rooted tree with 5 taxa there are only 105 tree topologies, so 5-taxon alignments included replicate tree topologies. I performed 10 replicates for each combination of number of taxa, tree depth, and number of trees. This resulted in a total of 3750 simulated alignments.

To simulate a single dataset, I generated random rooted trees using the "rmtree" function from the R package ape v5.6.2 (Paradis and Schliep 2019) and scaled all trees to the relevant tree depth. After generating the random trees, I constructed one alignment per tree by simulating along each tree such that the alignment length was l/x (where l is the length of the total alignment and x is the number of trees in the simulation). DNA sequence alignments with the Jukes-Cantor model of sequence evolution (Jukes and Cantor 1969) were simulated in Alisim (Ly-Trong et al. 2022), using IQ-Tree version 2.2.0 (Minh et al. 2020b). I then concatenated each alignment in a simulation to create single supermatrix for each simulation. This procedure generates datasets with gradated treelikeness. At one extreme is a 10,000 base pair long alignment where each site shares an identical underlying tree (perfectly treelike). At the other extreme is a 10,000 base pair alignment where each site is simulated from a different randomly generated tree.

Custom scripts were written to perform these analyses in R v4.3.1 (R Core Team 2018) using the R packages ape v5.7.1 (Paradis and Schliep 2019), phangorn v2.11.1 (Schliep 2011; Schliep et al. 2017) and phytools v1.9.16 (Revell 2012). All code to replicate these simulations is available from the "Code" folder of the GitHub repository https://github.com/caitlinch/treelikeness-metrics, in the files 01_simulations.R and func_simulating_alignments.R. Simulation results are available from the Figshare repository https://doi.org/10.6084/m9.figshare.26054467.

1.4.2.2 Introgression simulations

The second set of simulations is similar to the first, except that instead of using random trees, the trees are generated from a biologically inspired model. As above, I set the possible number of taxa *n* at five values (5, 10, 20, 50, 100). To better mimic empirical phylogenetic datasets, I fixed the total alignment length at 40,000 bp long. Each alignment consisted of 200 simulated genes, with a fixed gene length of 200 bp. For each simulated alignment, I used the function "sim.bd.taxa.age" from the R package TreeSim v2.4 (Stadler 2011, 2017) to simulate species trees with *n* species under a birth-death process, with a fixed time since the origin of the process. I used a Yule process (i.e., an extinction rate of zero) and two speciation rates, 0.01 and 1. These speciation rate values were selected to cover a broad range of speciation rates from empirical studies (Etienne et al. 2011; Stadler and Bokma 2013; Silvestro et al. 2018; Condamine et al. 2019).

Table 2: Divergence times and empirical tree depths for three clades from the animal tree of life.

Clades were selected to have a range of tree ages included in simulations. 'Myr is millions of years. 'Subs/site' is substitutions per site.

Simulated tree age	Clade	Divergence time		Empirical tree depth		Simulation tree depth
		Time (Myr)	Citations	Depth (subs/site)	Citations	
500	Bilateria	573 – 656	(Peterson et al. 2004)	0.313 – 0.422	(Simion et al. 2017a, 2017b)	0.5
		400 – 700	(Chernikova et al. 2011)	0.735 – 0.832	(Laumer et al. 2018a, 2018b)	
		688 – 596	(dos Reis et al. 2015)	0.331 – 0.338	(Laumer et al. 2019a, 2019b, 2019c)	
50	Primates	68.2 – 81.2	(Pozzi et al. 2014)	0.083	(Vanderpool et al. 2020a, 2020b)	0.05
		~60	(Vanderpool et al. 2020b, 2020a)			
		61.20 – 66.17	(Wisniewski et al. 2022)			
5	Human+ Chimp+ Gorilla	5 – 7	(Glazko and Nei 2003)	0.004	(Vanderpool et al. 2020a, 2020b)	0.005
		7 – 13	(Langergraber et al. 2012)			
		~6.6	(Amster and Sella 2016)			

To ensure my simulations covered a range of biologically plausible conditions, I used three sets of simulation conditions, each based on a different clade from the animal tree of Life (Table 2). I selected three tree ages: 5, 50, and 500 million years (Myr). The 5 Myr trees were designed to mimic the age of the Human-Chimp-Gorilla clade, which has been estimated to diverge 5-13 millions of years ago (Ma) (Glazko and Nei 2003; Langergraber et al. 2012; Amster and Sella 2016; Vanderpool et al. 2020b). The 50 Myr trees were designed to mimic the age of the Primates clade, which has been estimated to diverge 60 – 81.2 Ma (Pozzi et al. 2014; Vanderpool et al. 2020b; Wisniewski et al. 2022). Finally, the 500 Myr trees were designed to mimic the age of the Bilateria clade, which has been estimated to diverge 400-700 Ma (Peterson et al. 2004; Chernikova et al. 2011; dos Reis et al. 2015; Simion et al. 2017a; Laumer et al. 2018a, 2019a).

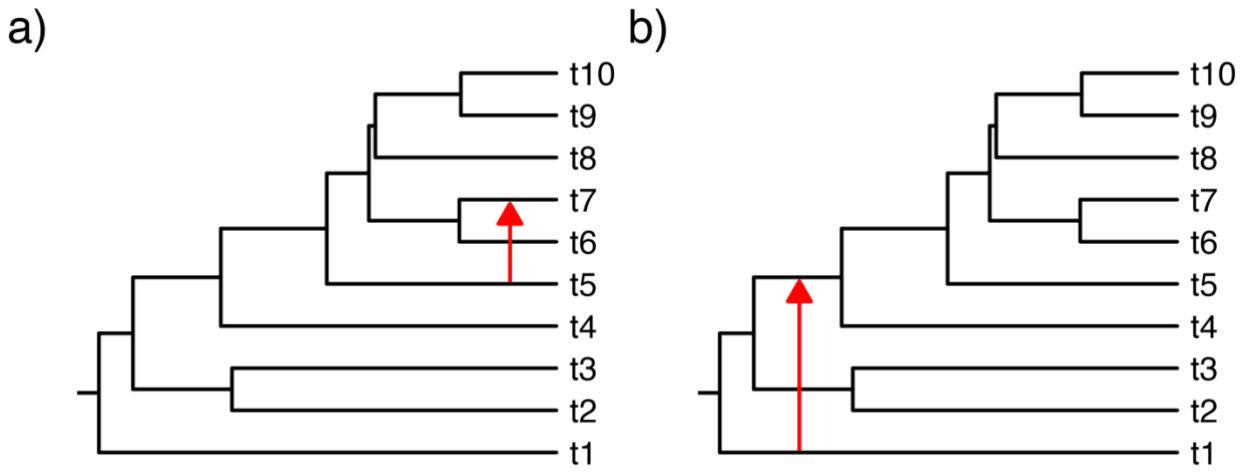


Figure 5: Illustration of simulated introgression events.

a) A randomly generated 10-taxa tree undergoing a recent introgression event, where genetic material from taxa t5 is introgressed into taxa t7. A recent introgression event occurs between any set of two randomly selected tips, provided that the tips are not sister species.

b) A randomly generated 10-taxa tree undergoing an ancient introgression event, where genetic material from the species t1 is introgressed into the ancestor to taxa t4 t5, t6, t7, t8, t9 and t10. An ancient introgression event occurs towards the root of the tree where there are three distinct lineages, and introgression is between two of the three lineages (provided that the two lineages are not sister species).

Next, I added a single introgression event to the tree. Introgression events were classified as either recent or ancient (Figure 5). For a recent event, two random tips were selected such that the pair of tips were not sister taxa (Figure 5a). I calculated the timing of each branching event, then set the timing for recent introgression events halfway between the tips and the most tip-wise node involving either of the taxa. Ancient events were set such that introgression occurred as close to the root as possible between two non-sister taxa (Figure 5b). To select

two lineages for the ancient introgression event, I identified lineages involved in the second and third branching events (where the first branching event is defined as the event that creates the root of the tree). If four taxa were present in those two events, I selected one taxon from each event at random. If three taxa were involved in the event, I selected two taxa such that the event took place between two taxa that were not sister lineages. The timing of an ancient event was set halfway during the period that the two taxa involved in the introgression event were present on the tree. For both recent and ancient introgression events, I varied the proportion of recombinant sequence r moving from one species to the other from 0 to 50%, with r = (0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5). When the proportion of recombinant sequence r = 0, no introgression event was added to the alignment.

From each species tree, I generated 200 gene trees in ms (Hudson 2002). By simulating gene trees under the coalescent I introduced conflicting phylogenetic signal due to ILS in all alignments, regardless of whether an introgression event was added. To simulate gene trees for each simulation replicate, I generated a command line for ms based on the scaled random coalescent tree using a custom-written R function ("ms.generate.trees" function from code/func_simulating_alignments.R, https://github.com/caitlinch/treelikeness-metrics). In ms, I set the population size for each of the populations (i.e., each of the tips) equal to 1 with no migration.

To convert trees from coalescent units to substitutions per site, I scaled each of the 200 gene trees for an alignment by the simulation tree depth for the specified simulated tree age. To obtain a reasonable simulation tree depth, I determined empirical tree depths for each of the three clades underlying my simulations (Table 2). The simulated tree depths selected (in substitutions per site) were 0.5 for the 500 Myr simulations, 0.05 for the 50 Myr simulations, and 0.005 for the 5 Myr simulations (Table 2). I estimated these tree depths from either estimating trees from published empirical datasets or directly from existing published phylogenetic trees. To identify the tree depth in substitutions per site, I took existing phylogenetic studies that investigated each of three clades from Table 2.

For the 50 and 5 Myr simulations, I used primates dataset of Vanderpool et al. (2020a, 2020b) and estimated a partitioned tree in IQ-Tree with ModelFinder and 1000 ultrafast bootstrap replicates (Chernomor et al. 2016; Kalyaanamoorthy et al. 2017; Biczok et al. 2018; Hoang et al. 2018a; Minh et al. 2020b) using the command "`iqtree2 -s alignment.fa -p partition.nex -m MFP+MERGE`". For the 500 Myr simulations, I used the following published trees downloaded from the relevant data repositories: tree_90sp_CAT, tree_90sp_CAT_heterop60, tree_90sp_CAT_heterop70 and tree_90sp_LGF-PARTITION

(Simion et al. 2017a, 2017b); Tplx_phylo_d1_withbnni_Tadhonly.phylip.treefile, Tplx_phylo_d1.phylip (Laumer et al. 2018a, 2018b); and nonbilateria_MARE_BMGE.IQTree, nonbilateria_MARE_cho_BMGE.IQTree (Laumer et al. 2019a, 2019b, 2019c). To calculate the empirical tree depth, I opened each tree in R, used the ape function "extract.clade" to extract only the clade of interest, and then calculated the maximum branching time for that clade using the ape function "branching.times". I calculated the empirical tree depth of the Human-Chimp-Gorilla clade as 0.004 substitutions per site (Vanderpool et al. 2020a, 2020b). I calculated the empirical tree depth of the Primates clade as 0.083 (Vanderpool et al. 2020a, 2020b). Finally, I calculated the empirical tree depth of the Bilateria clade as 0.313 – 0.422 (Simion et al. 2017a, 2017b), 0.735 – 0.832 (Laumer et al. 2018a, 2018b), or 0.331 – 0.338 (Laumer et al. 2019a, 2019b, 2019c). These three cases were used as the biological grounding for this set of simulations.

After scaling gene tree depth, I simulated DNA along each gene tree using Alisim with the Jukes-Cantor model as above, except that every simulated gene was 200 base pairs long. Finally, I concatenated all genes for that simulated alignment into a single supermatrix. This resulted in a single concatenated alignment of 40,000 bp. I performed 10 replicates for each combination of number of taxa, tree age, speciation rate, proportion of recombinant sequence, and timing of recombination event (either recent or ancient) resulting in a total of $10 \times 5 \times 3 \times 2 \times 11 \times 2 = 6600$ simulated alignments. However, I excluded all simulations with 5 taxa and an ancient introgression event (a total of $10 \times 1 \times 3 \times 2 \times 11 \times 1 = 660$ alignments), as for a 5-taxon tree there is little difference between recent and ancient events. This resulted in a total of 5940 simulated alignments.

To generate the alignments, I wrote custom R v4.3.1 (R Core Team 2018) script using the R packages ape v5.6.2 (Paradis and Schliep 2019), phangorn v2.7.1 (Schliep 2011), phytools v1.9.16 (Revell 2012) and TreeSim v2.4 (Stadler 2011, 2017). All code to replicate these simulations is available from the GitHub repository https://github.com/caitlinch/treelikeness-metrics, in the files code/01_simulations.R and code/func_simulating_alignments.R. Simulation results are available from the Figshare repository https://doi.org/10.6084/m9.figshare.26054467.

1.4.3 Tests for Treelikeness

I reviewed the literature and identified 9 existing approaches which had been used to measure or test for treelikeness (although not all were explicitly designed for this purpose): Treeness Triangles (White et al. 2007), the Reticulation Index (Cai et al. 2021), Cunningham metric

(Cunningham 1978), likelihood mapping (Strimmer and von Haeseler 1997), δ plots (Holland et al. 2002), the Network Treelikeness Test (Huson and Bryant 2006), Q-residual (Gray et al. 2010), TIGER (Cummins and McInerney 2011), and site concordance factors (sCF) (Minh et al. 2020a). From these 9, I excluded Treeness Triangles (White et al. 2007) as I could not access the original software implementation, and I excluded the Reticulation Index (Cai et al. 2021), because calculating this metric required multiple steps that have the potential to violate the treelikeness assumption (e.g., by estimating gene trees, this method assumes that each gene is treelike). I then focused on comparing the remaining 7 approaches and my new approach described above.

Cunningham (1978) defines a method to compare the goodness-of-fit of different representations of the same dataset, by calculating the extent to which each representation accounts for the variance within the data. In this chapter, I represented this by calculating the R^2 value for an alignment using the observed (i.e., alignment) and predicted (i.e., tree) pairwise distances.

To calculate the Cunningham metric, I wrote a custom R function cunningham.test, available in the file code/func_metrics.R from the project GitHub repository https://github.com/caitlinch/treelikeness-metrics. To calculate the Cunningham metric for an alignment, I first calculated the observed distances from the alignment (d_{ij}) by calculating a pairwise distance matrix from the alignment using the dist.ml function from the R package phangorn v2.7.1 (Schliep 2011). I then estimated a maximum likelihood tree from the alignment using IQ-Tree v2.2.0 with ModelFinder (Kalyaanamoorthy et al. 2017; Minh et al. 2020b), although any tree estimation method would be sufficient. From that maximum likelihood tree, I calculate the predicted distances (p_{ij}) by calculating a pairwise distance matrix from the tree. I then calculate the total sum of squares (TSS) (Equation 2) and residual sum of squares (RSS) (Equation 3) and use these to calculate the R^2 (Equation 4). I report the R^2 value for each alignment as the result of the Cunningham metric. When the choice of model is appropriate for the data, the Cunningham metric is between 0 and 1. An increasing Cunningham metric value indicates that the tree is a good fit for the alignment.

Equation 2 $$\boldsymbol{TSS = \sum d_{ij}^2}$$

Equation 3 $$\boldsymbol{RSS = \sum \left(p_{ij} - d_{ij}\right)^2}$$

Equation 4 $$\boldsymbol{R^2 = \frac{TSS - RSS}{RSS}}$$

I performed likelihood mapping using the implementation in IQ-Tree v2.2.0 (Strimmer and von Haeseler 1997; Minh et al. 2020b). I report the proportion of fully resolved quartets (the number of fully resolved quartets divided by the total number of quartets) as the test statistic for the likelihood mapping method. This approach has been used in previous studies to quantify the treelikeness of the phylogenetic signal within a given dataset (Vďačný 2017; Bourke et al. 2021). The proportion of resolved quartets will be between 0 and 1, with higher values indicating a higher proportion of treelike quartets within the alignment.

To calculate the mean δ_q value (Holland et al. 2002) for each alignment, I first used the dist.ml function from R package phangorn (Schliep 2011) with substitution model set to "JC69" to calculate a distance matrix from the alignment. Then I applied the "delta.plot" function from R package ape v5.6.2 (Paradis et al. 2004) to calculate the δ_q values for the alignment (i.e., the treelikeness of each possible quartet). The test statistic reported is the mean δ_q value from an alignment. The mean δ_q value ranges from 0 to 1, with a value of 0 indicating perfect treelikeness. As treelikeness decreases, the mean δ_q value increases.

To calculate the Network Treelikeness Test, I used SplitsTree4 v4.18.2 (Huson 1998; Huson and Bryant 2006, 2022) to infer a split network and then construct a confidence network using 100 bootstraps. I then identified the set of splits with confidence intervals excluding 0. If this set is compatible, I accept the null hypothesis that this dataset is treelike. As a test statistic, I report the proportion of treelike alignments for each set of experimental conditions (i.e., the proportion of the 10 replicate alignments for each set of simulation conditions that are classified as treelike by the network treelikeness). The proportion of treelike alignments is between 0 and 1, with a value of 1 indicating every alignment for that set of experimental conditions is classified as treelike by the Network Treelikeness Test.

I used the Phylogemetric python library (Greenhill 2016, 2021) to calculate the Q-residual value for each alignment. I report the mean Q-residual for each alignment as the test statistic. The mean Q-residual value is bound between 0 and 1, and as treelikeness of quartets decreases the mean Q-residual value increases.

To calculate the TIGER value for each alignment I used the fast_tiger software (Frandsen 2015). Phylogenetically uninformative sites (i.e., sites that do not provide information about the evolutionary relationships between taxa in an alignment) skew TIGER values, especially when the proportion of uninformative sites is high (List 2022). Therefore, I removed phylogenetically uninformative sites from each alignment using the "pis" function from the R package ips v0.0.11 (Heibl 2008) before applying fast_tiger to calculate the TIGER value. As a test statistic, I report

the mean TIGER value for each alignment. TIGER values range from 0 to 1, with TIGER values increasing as the treelikeness increases.

I calculated site concordance factors (Minh et al. 2020a; Mo et al. 2023) using IQ-Tree2 v2.2.2 (Minh et al. 2020b). As calculating the sCFs requires a reference tree, I used IQ-Tree2 to estimate a maximum likelihood tree for each alignment, using ModelFinder to identify the best model (Kalyaanamoorthy et al. 2017). I then calculated the maximum likelihood based sCF using IQ-Tree2 and the flag "`--scfl`". For any given replicate alignment, the test statistic is the mean sCF calculated from the reference tree. The mean sCF ranges from ~0.3 to 1, and as treelikeness increases the mean sCF should increase.

To apply all tests for treelikeness and compile the results, I wrote a pipeline in R v4.3.1 (R Core Team 2018) using the R packages ape v5.7.1 (Paradis and Schliep 2019), ips v0.0.11 (Heibl 2008), phangorn v2.11.1 (Schliep 2011; Schliep et al. 2017), and phytools v1.9.16 (Revell 2012). Code to apply test statistics and compile results is available in the "Code" folder of the GitHub repository (https://github.com/caitlinch/treelikeness_metrics) in the files 02_apply_treelikeness_metrics.R, func_metrics.R, and func_tree_proportion.R. Data analysis and plotting were performed using the R packages ggplot2 v3.4.2 (Wickham 2016), ggtree v3.8.2 (Yu 2020, 2022; Xu et al. 2022), patchwork v1.1.2 (Pedersen 2022), reshape2 v1.4.4 (Wickham 2007), and scales v1.2.1 (Wickham and Seidel 2022). The scripts to replicate plotting and data analysis (03_data_analysis.R, 05_figures.R and func_data_analysis.R) are available from the GitHub repository as above. Results are available from the Figshare repository https://doi.org/10.6084/m9.figshare.26054467.

1.4.4 Empirical Analysis with Parametric Bootstrap

I selected two datasets for empirical phylogenetic analysis that had previously applied to estimate the metazoan tree of life, a contentious tree with unresolved topology and conflicting phylogenetic signal (Shen et al. 2017; Kapli and Telford 2020; Redmond and McLysaght 2021a). Phylogenetic studies have found support for placing a range of different groups as the sister to all other metazoan clades including Porifera (Philippe et al. 2009, 2011b; Pisani et al. 2015; Simion et al. 2017a), Ctenophora (Dunn et al. 2008; Hejnol et al. 2009; Ryan et al. 2013; Borowiec et al. 2015), or a monophyletic group consisting of Porifera and Ctenophora (Francis and Canfield 2020). In addition, metazoan tree topology is affected by choices made during phylogenomic analysis including the selection of loci (Pandey and Braun 2020; McCarthy et al. 2023) or the filtering of sites (Francis and Canfield 2020).

The first dataset "Metazoa_Choano_RCFV_strict.phy" (Whelan et al. 2017a, 2017b), includes 76 taxa and 117 genes (for a total of 49388 sites). The second dataset is "Whelan2017MCRS_filtered.fasta" (McCarthy 2022; McCarthy et al. 2023), which takes the Whelan 2017 original dataset and applies a test of orthologous signal, excluding genes with insufficient signal (i.e., genes unable to recover 3 or more of the 6 distinct clades within the metazoan tree of life). The filtered dataset contains 76 taxa but only 42 genes (for a total of 22820 sites). I refer to these datasets as the "Original" and "Orthology-enriched" datasets respectively. I separated each concatenated alignment into individual genes using the partition files from the original publications, resulting in 117 alignments for the Whelan et al. (2017a) original dataset and 42 for the filtered dataset of McCarthy et al. (2023).

I identified three test statistics to apply to empirical data based on simulation results (see 1.5.1 below): tree proportion, sCFs, and δ plots. No test statistic performed as an ideal test statistic for treelikeness (i.e., a monotonic change as treelikeness increased) for the introgression simulations, so only random tree simulation performance was considered to identify the three best-performing test statistics. To test the statistical significance of each test statistic value, I performed a parametric bootstrap with 100 bootstrap replicates for each alignment (i.e., each gene from each dataset). A bootstrap replicate is a simulated sequence alignment with the same underlying parameters as the empirical alignment (tree, number of sites, model of substitution, rate parameters, placement of gaps and unknown characters). Thus, each bootstrap replicate is treelike because it is simulated along a single bifurcating tree. In this way, the bootstrap replicates give the range of each test statistic which is plausible under treelike simulations. The observed test statistics can then be compared to these null distributions to ask whether each observed alignment is treelike, or whether treelikeness can be rejected in the statistical sense for that alignment. I generated 100 parametric bootstrap replicate alignments in Alisim (Ly-Trong et al. 2022) using the command "`iqtree2 -s alignment.fa --alisim param_bs -m MFP --num-alignments 100 --out-format fasta`". In this command, "`alignment.fa`" is the path to the alignment file for the gene of interest; "`param_bs`" is the prefix for simulated alignments; "`-m MFP`" indicates IQ-Tree2 should run ModelFinder and apply the best model to generate simulated sequences; "`--num-alignments 100`" sets the number of simulated alignments at 100; and "`--out-format fasta`" species that simulated alignments will be in FASTA format. Alisim generates 100 simulated alignments with the same length, number of taxa, underlying tree and model, and gap location as the input alignment. I then calculate the tree proportion, mean sCF value, and mean δ plot value for the input alignment and the 100 simulated parametric bootstrap alignments for each gene in each dataset.

I calculated the lower-tail p-value for each test statistic for each alignment (i.e., each gene in each dataset) to determine whether the null hypothesis of treelikeness was rejected. To determine the statistical significance of a given test statistic, I took the 101 test statistic values for any given gene (1 value from the empirical alignment and 100 values from the parametric bootstrap replicates) and computed the p-value by calculating the proportion of bootstrap replicate test statistic values with values less than or equal to the observed test statistic value.

The pipeline was written in R v4.3.1 (R Core Team 2018) with the packages ape v5.7.1 (Paradis and Schliep 2019), ggplot2 v3.3.4 (Wickham 2016), ggpubr v0.6.0.999 (Kassambara 2023), patchwork v1.1.3 (Pedersen 2022), phangorn v2.11.1 (Schliep 2011), seqinr v4.2.30 (Charif and Lobry 2007), and reshape2 v1.4.4 (Wickham 2007). The custom R code to replicate the empirical analysis is available from the project GitHub repository (https://github.com/caitlinch/treelikeness_metrics), in the files code/04_empirical_data_test.R, code/func_parametric_bootstrap.R, and code/func_empirical.R. Gene alignments and results of the parametric bootstrap are available from the project Figshare repository (https://doi.org/10.6084/m9.figshare.26054467).

1.5 Results

The ideal performance for a test statistic of treelikeness would be for the test statistic to show a monotonic change as either the number of random trees is increased (random tree simulations) or the proportion of introgression increases from 0% to 50% (introgression simulations). The direction of the monotonic change will depend on the test statistic: e.g., for tree proportion a treelike alignment has a test statistic value of 1 and the tree proportion decreases as the treelikeness decreases, but for the δ plots method an alignment with perfectly treelike quartets will have a mean δ_q value of 0 and the mean δ_q value will increase as the treelikeness of the quartets decreases.

1.5.1 Random Tree Simulations

For the random tree simulations, the most successful metrics were the tree proportion, mean δ_q value and mean sCF values (Figure 6). These test statistics clearly and consistently changed in the predicted direction up to some limit as the number of trees increased for all tree depths and all numbers of taxa (although all statistics also decreased as the tree depth increased). The tree proportion test statistic performed similarly under different numbers of taxa and under different tree depths. The range in mean δ values was lower at the highest tree depth of 1 substitution per site, with a higher initial mean δ value of 0.25 – 0.37 compared to the initial values of approximately 0.12 at lower tree depths (Figure 6). Similarly, the mean sCF value was lower at the highest tree depth of 1 substitution per site (Figure 6, initial sCF value of around 0.5) , compared to the initial values 0.01 substitutions per site (sCF of 0.87 – 1) and 0.1 substitutions per site (sCF of 0.75 – 0.87).

The remaining five tests performed poorly (Figure 6). The Cunningham metric test statistic increased in value as the number of trees increased only for the highest tree depth of 1 substitution per site. The proportion of treelike alignments (calculated from the network treelikeness test) had no clear trend when the number of taxa was low (5 taxa) or the tree depth was low (0.01 substitutions per site). The proportion of resolved quartets (calculated from likelihood mapping) and the mean Q-residual value had minimal response to decreasing treelikeness at low and moderate tree depths (0.01 and 0.1 substitutions per site). At the highest tree depth of 1 substitution per site the proportion of resolved quartets initially increased as the number of underlying trees in an alignment increased, whereas the mean Q-residual change showed no response. Finally, the mean TIGER value decreased as the number of trees increased at low and moderate tree depths (0.01 and 0.1 substitutions per site), but the magnitude of the difference was correlated to the number of taxa in an alignment.

As the number of taxa in the alignment increased, the total range of mean TIGER values decreased. Finally, for the highest tree depth of 1 substitution per site, the mean TIGER value varied very little as the treelikeness of the alignments decreased, and instead was dependent on the number of taxa in the alignment. In general, the ability of all test statistics to detect changes in treelikeness decreased when the number of trees in an alignment increased above 10 – 20 trees (Figure 6). Some test statistics performed successfully up to 50 trees, such as the mean site concordance factor and the proportion of treelike alignments.

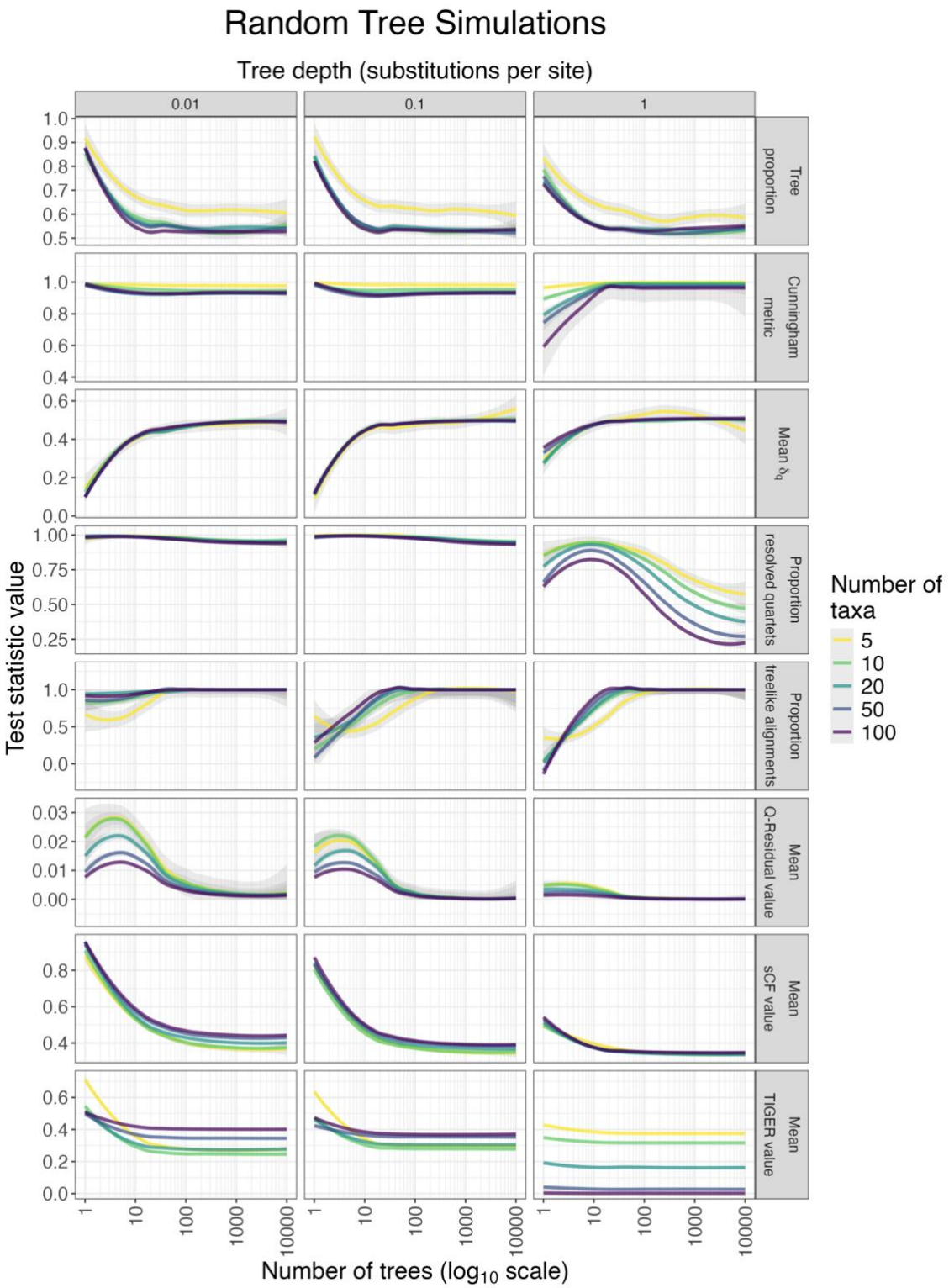


Figure 6: Treelikeness test statistic values for alignments with decreasing treelikeness due to an increasing number of random concatenated trees.

The y axis differs for each test statistic. Each line is a smoothed conditional mean calculated using locally estimated scatterplot smoothing and the formula y~x. The light gray bands represent the 95% confidence intervals for each line. The x-axis is shown on a log10 scale to better depict the large range in the number of trees variable. Each column is a different tree depth (in substitutions per site), and each row is a different test statistic. 'sCF' is the site concordance factor.

1.5.2 Introgression Simulations

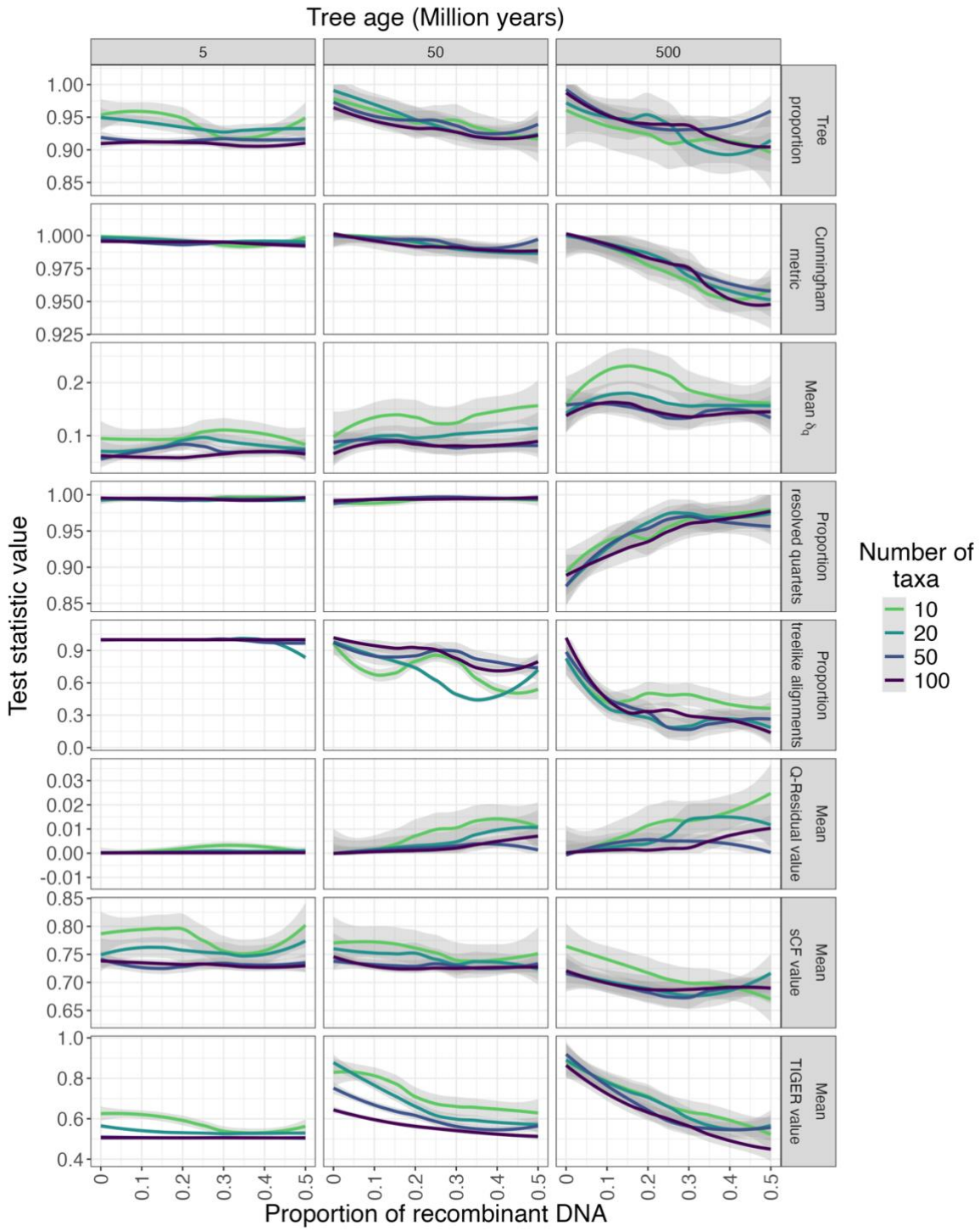


Figure 7: Treelikeness test statistic results for alignments with one ancient introgression event and a speciation rate of 1.

The y axis differs for each test statistic. The proportion of recombinant DNA ranged from 0 to 0.5 (half the alignment), and as the proportion of recombinant DNA increases the treelikeness of the alignment decreases. Each line is a smoothed conditional mean calculated using locally estimated scatterplot smoothing and the formula y~x. The light grey bands represent the 95% confidence intervals for each line. Each column is a different tree age (in millions of years), and each row is a different test statistic.

In general, most test statistics were unable to detect decreased treelikeness under a single ancient introgression event (Figure 7, Supplementary Figure 1). Most test statistics performed best when the tree age was high (500 Myr) and when the speciation rate was 1. At large tree ages (500 Myr), four of the test statistics varied in value as the proportion of introgressed DNA reached 50%: Cunningham metric, mean δ value, proportion of resolved quartets (calculated from likelihood mapping), and mean site concordance factor (Figure 7, Supplementary Figure 1). When the speciation rate was 1 and the tree age was moderate or high (50 or 500 Myr), the tree proportion decreased slightly as the proportion of introgressed DNA increased. The mean Q-residual value did not respond consistently to decreasing treelikeness.

The ability to detect decreased treelikeness due to a single recent introgression event varied (Figure 8, Supplementary Figure 2). The test statistic with the largest difference in value as the proportion of recombinant DNA increased was the proportion of treelike alignments (calculated from the network treelikeness test). The proportion of treelike alignments performed best when the tree age was low (5 Myr), or the speciation rate was low (0.1). Both the tree proportion and the mean δ value varied as the proportion of recombinant DNA increased. The magnitude of change in test statistic value was highest for alignments with 5 or 10 taxa and there was very little change in test statistic value for alignments with 50 or 100 taxa. Neither the Cunningham metric, the mean Q-residual, or the proportion of resolved quartets (calculated from the likelihood mapping) varied substantially as the proportion of recombinant DNA increase. At low to moderate tree ages (5 or 50 Myr) and low numbers of taxa, the mean site concordance factor monotonically decreased as the proportion of recombinant DNA in the alignments increased. Otherwise, both the mean TIGER value and the mean site concordance factor showed a non-monotonic response, but the overall value for these two test statistics was more dependent on tree age or the number of taxa then the proportion of recombinant DNA (Figure 8, Supplementary Figure 2).

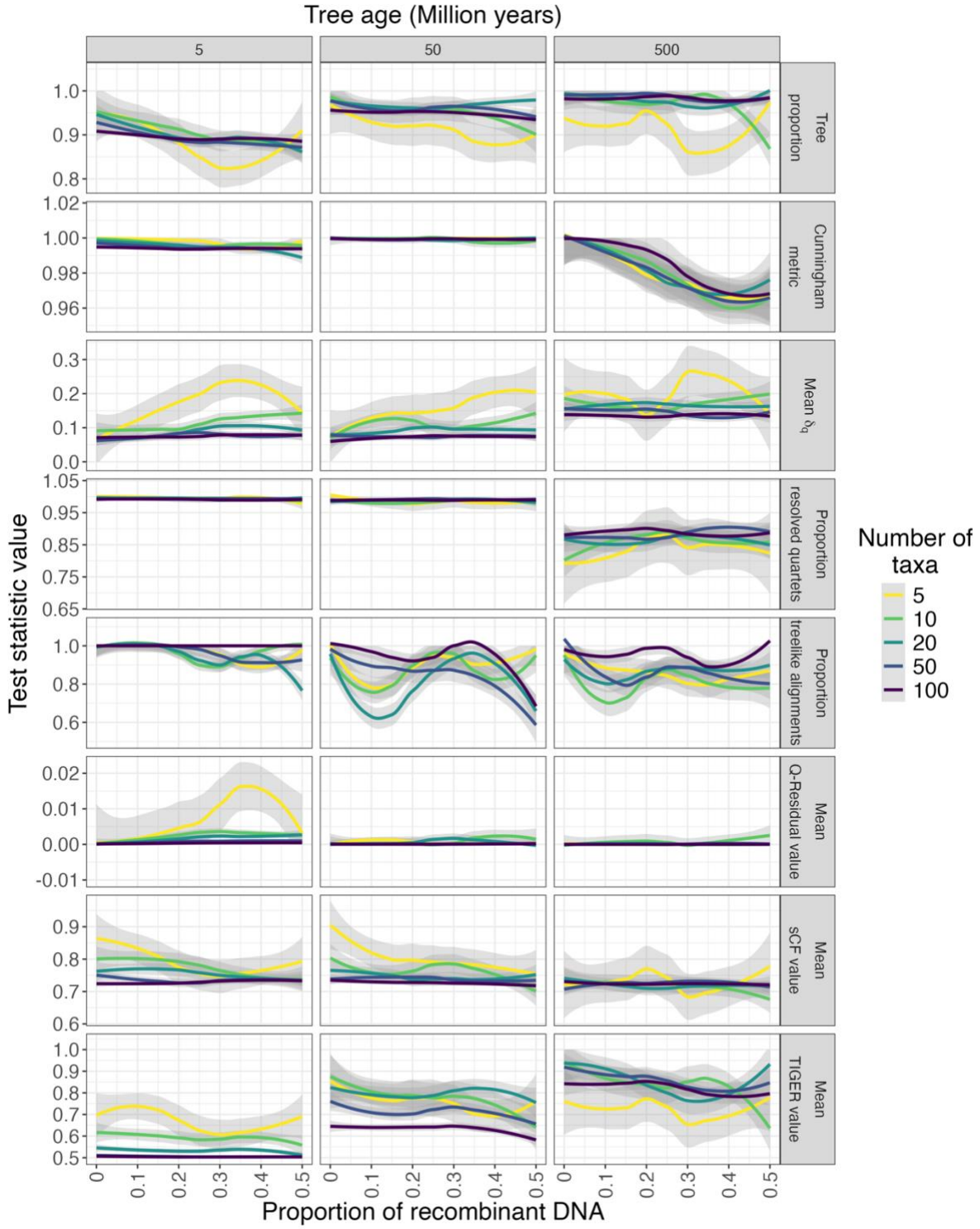


Figure 8: Treelikeness test statistic results for alignments with one recent introgression event and a speciation rate of 1.

The y axis differs for each test statistic. The proportion of recombinant DNA ranged from 0 to 0.5 (half the alignment), and as the proportion of recombinant DNA increases the treelikeness of the alignment decreases. Each line is a smoothed conditional mean calculated using locally estimated scatterplot smoothing and the formula y~x. The light grey bands represent the 95% confidence intervals for each line. Each column is a different tree age (in millions of years), and each row is a different test statistic.

1.5.3 Empirical Analysis

I identified three tests that showed consistent monotonic changes in test statistic value across the simulations: tree proportion, mean sCF, and mean δ_q value. Each of these test statistics performed better under the random tree simulations, but each performed well under a subset of parameters for the introgression simulations. I applied each of these three test statistics along with a parametric bootstrap with 100 replicates to every gene in two empirical datasets. I selected datasets previously used to estimate the metazoan tree of life: the first from Whelan et al. (2017a), and a version of this first dataset edited by McCarthy et al. (2023) to remove genes with insufficient evidence of orthology. I call these the “original” dataset and the “orthology-enriched” dataset respectively. The mean tree proportion was significantly higher (i.e., more treelike) in the orthology-enriched dataset (0.7485) then in the original dataset (0.7166) (Figure 9, p=7E-04). Similarly, mean δ_q values were lower (i.e., more treelike) in the orthology-enriched dataset (mean test statistic value of 0.3647) than in the original dataset (mean test statistic value of 0.3844) (Figure 9a, p=2E-03). The mean SCF values did not differ significantly in the two datasets (sCF of 52.34 in the original dataset, and 52.03 in the orthology-enriched dataset; p=0.6). The parametric bootstrap shows that proportion of genes for which treelikeness could be rejected ($p < 0.05$) was very similar between both datasets (Figure 9bc), regardless of the test statistic used. However, the tree proportion rejected the hypothesis of treelikeness far less often than the other two test statistics.

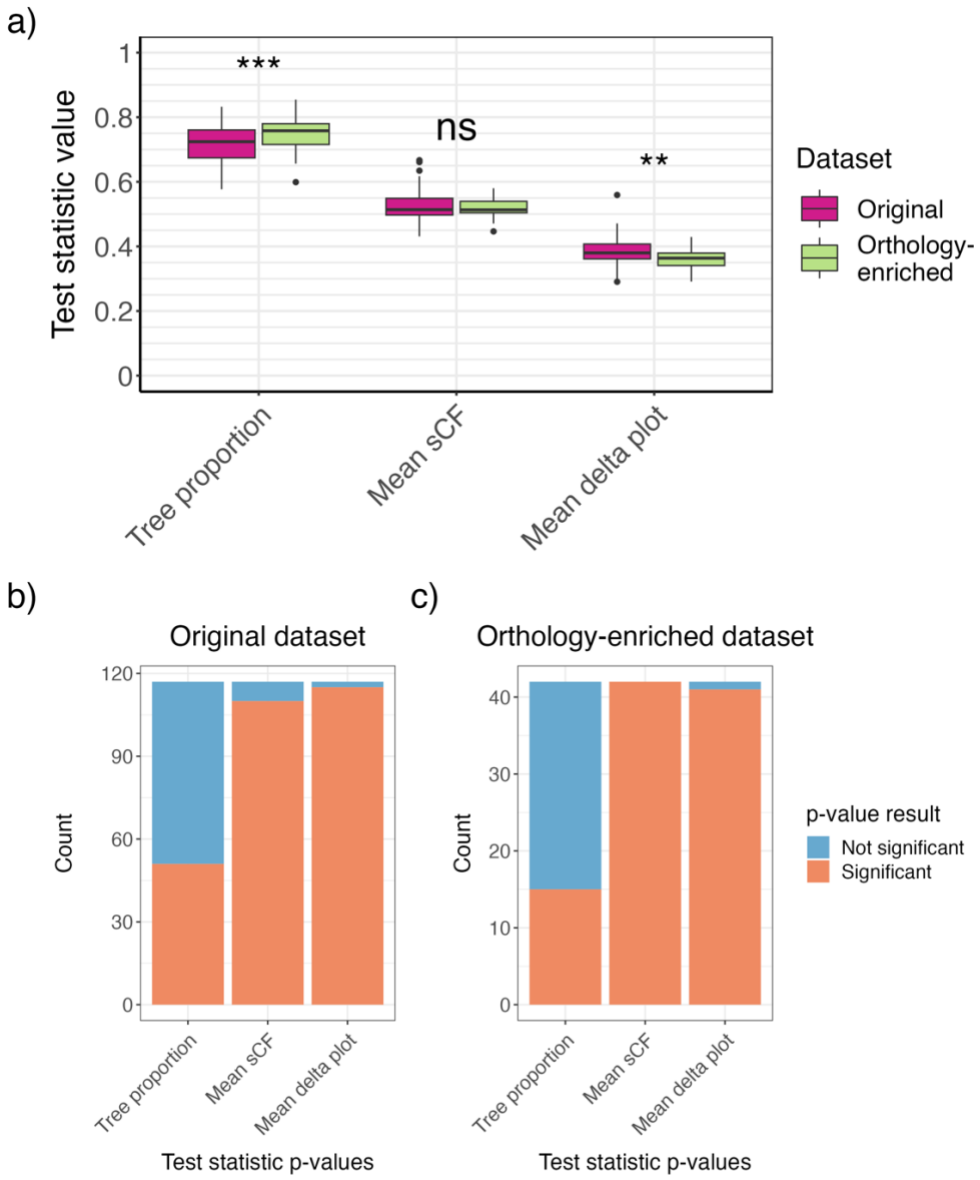


Figure 9: Applying tree proportion, mean sCF and the mean δ plot value to 2 empirical alignments.

"Original dataset", refers to the Metazao_Choano_RCFV_strict.phy alignment (Whelan et al. 2017a). The "Orthology-enriched dataset" is an edited version of the original dataset where genes with insufficient orthologous signal were removed (McCarthy et al. 2023).

a) Distribution of the test statistic values for three best-performing test statistic values identified from the simulation study (tree proportion, mean sCF and mean δ plot value). Each test statistic was applied to each gene from the original and orthology-enriched datasets. The original dataset has 117 genes, whereas the orthology-enriched dataset has 42 genes. To make the range of all test statistics 0 – 1, for this plot site concordance values were divided by 100. ns: p> 0.05, *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001.

b) Stacked bar chart showing p-values (calculated from parametric bootstrap) for each gene and each test statistic in the Original dataset. Genes with a statistically significant p-value (p < 0.05) reject the null hypothesis of treelikeness. Genes with a not significant p-value (p > 0.05) accept the null hypothesis of treelikeness

c) Stacked bar chart showing p-values (calculated from parametric bootstrap) for each gene and each test statistic in the Orthology-enriched dataset.

1.6 Discussion

This chapter aims to benchmark test statistics for quantifying treelikeness, to facilitate selection of appropriate methods for phylogenetic tree estimation. In this chapter, I used two kinds of simulations to assess the performance of a range of methods that have been applied to quantify treelikeness, as well as a new approach which I call the tree proportion. The best performing test statistics in my simulations were tree proportion, sCF and mean δ_q values. However, the performance of all test statistics differed substantially under the different simulation conditions. I first discuss the methods that performed poorly in my simulations, before discussing the methods that performed well and my empirical analysis.

As the Network Treelikeness Test of Huson and Bryant (2006) returns a binary result (i.e., either treelike or non-treelike), the treelikeness of individual alignments cannot be quantified using this test. In addition, the power to accurately determine the treelikeness of any particular alignment was inconsistent. This limited power to classify any individual alignment as treelike was previously noted by Huson and Bryant (2006), who rejected this test as a test statistic for treelikeness.

The Cunningham metric performed particularly poorly as a test of treelikeness. The Cunningham metric considers pairwise distances between taxa, which includes terminal branches. If the terminal branches are a large proportion of the distance, the Cunningham metric will have low power to detect decreased treelikeness. This limits the usefulness of the Cunningham metric as a general measure for treelikeness, as test statistic values cannot be directly compared for clades with different terminal branch lengths. In addition, calculating the Cunningham metric requires estimating a phylogenetic tree. The Cunningham metric value will be dependent on the method and model used for tree estimation, which may bias the pairwise distance matrix calculated from the tree. Estimating a tree will also increase the time and computational resources required to estimate the Cunningham metric, particularly for large datasets.

Likelihood mapping was designed to visualize phylogenetic signal within an alignment (Strimmer and von Haeseler 1997), and previous studies have applied this method to assess the phylogenetic signal of an alignment or support of individual branches within a tree (Nikolaev et al. 2007; Skaloud and Peksa 2010; Kim et al. 2013; Vanhove et al. 2015; Vďačný 2017; Prasanna et al. 2020; Cunha et al. 2022). In my simulations, I used the proportion of treelike quartets as a proxy for treelikeness and found no correlation between the proportion of treelike quartets and simulated treelikeness. One reason for this is that an alignment may have a high

proportion of resolved quartets but a low treelikeness, if the resolved quartets support conflicting relationships. In addition, the proportion of quartets affected by a single recent introgression event is low – only the quartets that contain both taxa impacted by the introgression event will be impacted, which is a relatively small proportion of quartets especially for larger trees. This occurs as individual quartets may be perfectly treelike, despite the presence of a introgression event (i.e., a recombination breakpoint) within the alignment. Given a tree with n tips and a single recent introgression event between two taxa, the probability of selecting a quartet that is not impacted by the introgression event (i.e., a quartet that does not contain either taxon involved in the introgression event) is:

$$p = \frac{\binom{n-2}{4}}{\binom{n}{4}} = \frac{(n-2)!}{4!\,\big(n-(2+4)\big)!} \times \frac{4!\,(n-4)!}{n!}$$

As the number of taxa in an alignment n increases, the proportion of probability of selecting a quartet not impacted by the introgression event increases – at 10 taxa p = 0.333, and at 100 taxa p = 0.921. The results of likelihood mapping are therefore dependent on both the number of taxa, and the number of taxa involved in events that cause decreased treelikeness. The proportion of treelike quartets therefore cannot accurately quantify the treelikeness of the alignments simulated in this study. Adding additional introgression events complicates this calculation, but given the number of taxa included in modern phylogenomic studies it's reasonable to assume the overall proportion of treelike quartets remains low. However, likelihood mapping is a useful visualisation tool for understanding the treelikeness of quartets within an alignment, especially when used in conjunction with other methods.

TIGER was originally designed to detect rapidly evolving sites by binning sites based on the site-wise TIGER values, reducing noise and homoplasy by leaving only slowly evolving homologous sites (Cummins and McInerney 2011). I found that TIGER did not correlate with treelikeness in either of my simulation scenarios. However, I found increasing the number of taxa in an alignment or the tree age lowered the power of the TIGER test in my simulation studies, as the number of sites with identical site patterns consequently decreased. Previous studies applying the TIGER test to remove fast-evolving sites observed loss of phylogenetic signal in the filtered alignments (Sharma et al. 2015; Simmons and Gatesy 2015; Klimov et al. 2018), consistent with the idea that filtering sites using TIGER values did not increase treelikeness. Considering these previous analyses and my simulations, I do not recommend TIGER as a test statistic for treelikeness.

Q-residual values represent the treelikeness of each quartet, and have been used as a proxy for treelikeness in linguistic (Gray et al. 2010; Lee and Hasegawa 2013; Syrjänen et al. 2021) and biological datasets (Crouch 2014). I found that mean Q-residual performed poorly for the introgression simulations. For the random tree simulations, the Q-residual increased in value as the number of trees increased, but after ~10 random evolutionary histories were present in the alignment, the Q-residual value decreased. This is a limitation of the test statistic to detect the constant monotonic increase in the number of evolutionary histories present within a single alignment. This limitation may arise as the Q-residual includes only the largest two out of the three sums of the path lengths within a quartet, compared to the δ plot method which considers all three distances and performed well in my simulations. My simulations and previous comparisons of the two metrics on linguistic datasets (Gray et al. 2010; Holman et al. 2011; Syrjänen et al. 2021) find that δ_q values are a better measure of treelikeness than Q-residuals. Additionally, Holman et. al. (2011) find the Q-residual values are sensitive to the length of the terminal branches of a quartet, whereas δ_q values are not. These results and my chapter together suggest that δ plots are more useful as measures of treelikeness than Q-residuals.

The three best performing test statistics in my simulations were δ plots, sCFs, and tree proportion. The δ plot method has previously been applied to estimate the treelikeness of empirical phylogenetic alignments (Göker and Grimm 2008; Hernández-López et al. 2013; Folk et al. 2017; Charr et al. 2020). I found that mean δ_q performed well as a test statistic for treelikeness for the random tree simulations. The δ plot method calculates the treelikeness of each quartet in an alignment, which results in a single introgression being included in multiple quartets, making detection of a single reticulation event possible when calculating the mean δ_q value. My simulations found the absolute δ_q value for treelike alignments depends on the number of taxa, tree age and speciation rate. Several factors contribute to this result. First, although a single introgression event will be included in many quartets, the proportion of quartets affected will decrease as the total number of taxa increases, resulting in less change to the mean δ_q. Second, as the tree age increases, the number of substitutions on each branch increases, resulting in a higher initial δ_q value as the increased substitutions result in increased quartet treelikeness. Finally, the timing of speciation events within the tree will impact the initial δ_q value, with trees where speciation events are placed closer to the root have more time to diverge, resulting in higher mean δ_q values.

The mean sCF performed well as a test for treelikeness for the random tree simulations, although the initial mean sCF (when the number of trees is 1) decreases as the tree depth increases. This is likely due to homoplasy as recurrent mutations decrease phylogenetic signal

(Mossel and Steel 2005), reducing the proportion of decisive sites concordant with each branch in the tree and thus the sCF value. I note that this suggests that despite a recent update to the way that the sCF is calculated (Mo et al. 2023) this change does not render the sCF completely immune to the effects of homoplasy. In this chapter I investigated the use of mean sCF for quantifying treelikeness, but concordance factors (CFs) can also be used to quantify topological variation and investigate evolutionary processes (Baum 2007; Lanfear and Hahn 2024). Previously, gene and site CFs have been used to assess patterns of concordance and discordance at different branches within the phylogenies of the Primates (Vanderpool et al. 2020b), the carnivorous plant clade Nepenthaceae (Murphy et al. 2020), the wasp clade Chalcidoidea (Cruaud et al. 2024) and the Golden-backed frogs (Chan et al. 2020). Finally, as sCFs represent topological variation rather than sampling variation, increases in dataset size do not impact sCF values as much as bootstrap support or posterior probabilities. Larger datasets have less sampling variance, and therefore consistently high branch support values (Thomson and Brown 2022; Lanfear and Hahn 2024). In summary, sCFs are a useful tool for assessing phylogenetic signal more generally, and treelikeness more specifically.

Tree proportion performed well for the random tree simulations under all conditions. Tree proportion was also able to detect decreased treelikeness in the introgression simulations for certain combinations of tree age and speciation rate. For the recent introgression simulations, tree proportion performed best when the number of taxa or the tree age was low. This is expected, because a single introgression event will have a higher proportional impact on the tree proportion statistic when there are fewer taxa in the tree, and when the tree is shorter. In this chapter, tree proportion was calculated using a Neighbor-Net network, which could only capture the two conflicting splits with the highest weights, even if those weights represented a small proportion of either the total phylogenetic signal or the total number of splits. Consequently, the maximum spanning tree included around half of the splits from the Neighbor-Net network, resulting in a minimum tree proportion value of 0.5 as seen in my random tree simulations. Implementing tree proportion using a triplet or quartet phylogenetic network method could increase the proportion of phylogenetic signal from the alignment present in the network, as networks estimated using the quartet methods QNet or FlatNJ have more splits than those estimated from the same data using Neighbor-Net (Grünewald et al. 2007; Balvočūtė et al. 2014).

Our simulations included alignments with varying treelikeness, number of taxa, tree depths and introgression events. Future work could investigate further modelling of biologically realistic evolutionary histories, particularly in the space between the two extremes investigated in this chapter or by determining the adequacy of treelikeness tests for clades with complex

evolutionary histories. My random tree simulations represent the most extreme case of decreased treelikeness: each tree is randomly generated, and non-recombining regions of the alignment are as short as 1bp. While single genes in empirical datasets do contain multiple evolutionary histories (Mendes and Hahn 2016; Scornavacca and Galtier 2017; Mendes et al. 2019; Smith et al. 2020), those evolutionary histories are unlikely to be completely random. However, testing the adequacy of treelikeness metrics under the most extreme case is useful, as tests that cannot detect dramatic changes in treelikeness are unlikely to detect more subtle variations. On the other end of the spectrum, my introgression simulations were limited to a single introgression event. Empirical phylogenetic datasets can contain complex patterns of reticulation such as the multiple introgression events present within the tomato (Pease et al. 2016a) or butterfly clades (Edelman et al. 2019), or the combination of ancient and recent introgression events present within the primates (Vanderpool et al. 2020b).

I applied the three best-performing tests for treelikeness (tree proportion, sCFs, and δ plots) with a parametric bootstrap to two empirical datasets: the amino acid dataset of Whelan et al. (2017a, 2017b), and a filtered version of that dataset with misidentified orthologs removed (McCarthy 2022; McCarthy et al. 2023). For both datasets, most genes (> 94%) were classified as non-treelike by the sCF and mean δ_q test statistics, while far fewer were classified as non-treelike by tree proportion: 43.6% for the original dataset and 35.7% for the orthology-enriched dataset. Metazoan datasets contain substantial conflicting phylogenetic signal, with 42.5 – 69.7% of genes and 39.8 – 56.9% of sites from 8 datasets supporting Ctenophores as the sister to all other animals, with other genes and sites supporting alternate topologies (Shen et al. 2017). Other studies have shown that single genes contain multiple evolutionary histories (Scornavacca and Galtier 2017; Mendes et al. 2019), including Smith et al. (2020) who found between 0.6 and 100% of genes from 13 empirical datasets contained intragenic conflict. These results strongly suggest that the treelikeness assumption is violated in empirical data, raising the question of whether current tree inference methods could be improved in the face of such common non-treelikeness.

I performed two classes of introgression simulations in this chapter. To compare the simulations with results from existing studies, both tree depth and the number of species need to be considered. The three simulation depths were selected to be around 5 Myr (based on the Human-Chimp-Gorilla clade, or more generally a clade of closely related species), 50 Myr (the Primate clade, or more generally at the taxonomic level of Order), and 500 Myr (the animals, or more generally at the taxonomic level of Kingdom). This approach allows me to capture a broad range of phylogenetic analyses. For example, one set of introgression simulations included 100 taxa and a tree depth of 5 Myr. Large, shallow phylogenies are

common in epidemiology and public health (Hadfield et al. 2018; Li et al. 2020a). For example, during the recent coronavirus pandemic, millions of viral sequences were collected over a 5 year period (Oude Munnink et al. 2021). Epidemiological datasets contain a number of samples several magnitudes higher and timescales several magnitudes lower than the empirical metazoan datasets analysed in this chapter. The other extreme case for introgression simulations in this chapter is a deep phylogeny (500 Myr) with 5 or less taxa. Due to the massive amounts of genetic sequence data available for modern phylogenetic studies (Kapli et al. 2020), most modern phylogenetic studies would include more than 5 taxa. Even early phylogenetic analyses of the animal tree of life included >20 taxa, although these analyses were limited to a handful of loci.

In my recent introgression simulations, each introgression event was restricted temporally but the relationship between tips involved in the introgression event were not fixed. As such, any pair of tips in the tree could have a recent introgression event added, as long as the pair of tips involved did not form a cherry (i.e., a pair of sister tips). Introgression has been detected in a range of organisms: the dabbling duck genus *Anas* (Lavretsky et al. 2014); the macaque genus *Macaca* (Vanderpool et al. 2020b); the cat genus *Felidae* (Li et al. 2019); Chinese horseshoe bats (Mao et al. 2017); the North American whiptail lizards (Barley et al. 2022); the butterfly genus *Heliconius* (Thawornwattana et al. 2022); the fruit fly genus *Drosophila* (Suvorov et al. 2022); the tomato clade (Hibbins and Hahn 2021); the carnivorous pitcher plant genus *Nepenthes* (Scharmann et al. 2021); and the baobabs (Karimi et al. 2019). Within these examples, there's variation in the number of species involved in introgression events, relationship between species involved in an introgression event, and proportion of introgressed DNA. Introgression between closely-related species is well documented, such as the introgression between three macaque species identified by Vanderpool et al. (2020b) or the *erato-sara* clade of *Heliconius* butterflies (Thawornwattana et al. 2022). This case occurs in my introgression simulations when the randomly selected pair of tips are closely related. Introgression events can also happen between more distantly related taxa. For example, Lavretsky et al. (2014) studied the mallard complex consisting of around 20 species and subspecies of the dabbling duck genes *Anas*, and found the mallard had undergone extensive hybridisation with duck species from across the world. In the 20-taxon case for my recent introgression simulations, this would be equivalent to allowing an introgression event between any pair of taxa. For the larger cases of 50 and 100 taxa, my simulations are more similar to introgression observed in Eucalypts. Eucalypts are a group of more than 700 Australian trees and shrubs, currently separated into 7 genera (Crisp et al. 2024). Eucalypts are known to hybridise, and over 1300 agricultural hybrid clones of *Eucalyptus* species have been

developed (Dale and Dieters 2007). In an agricultural breeding program aiming to develop salt and drought tolerant hybrids, the species *Eucalyptus camaldulensis* was crossed with the species *Eucalyptus grandis* and *Eucalyptus globulus* (Dale and Dieters 2007). Each of these species is in different sects of the *Eucalyptus* genus (Crisp et al. 2024). By not restricting the relationship between taxa undergoing an introgression event in my simulations, I have been able to reflect the diversity of introgression events from the natural world.

In this chapter, I introduced a new metric for quantifying treelikeness called the tree proportion. I found that the tree proportion performed well in simulations with large decreases in treelikeness but had limited ability to detect the small decrease in treelikeness caused by a single introgression event. The tree proportion has several strengths. First, the method is tree independent. By estimating the maximum spanning tree using a modified version of Kruskal's algorithm (Kruskal 1956), the tree proportion is a measure of the maximum proportion of relationships in an alignment that can be captured by a tree, given a particular network representation of that alignment. Second, the tree proportion is easy to interpret: a tree proportion value of 1 indicates that the alignment is perfectly treelike, and the tree proportion decreases as the level of conflicting signal within an alignment increases. Finally, by removing the trivial splits present in both the phylogenetic network and the maximum spanning tree, the tree proportion is limited to conflicting signals within the phylogenetic network. I recommend applying a parametric bootstrap with at least 100 replicate alignments, to provide context to the tree proportion by accounting for factors including tree topology, model of evolution, number of sites, and position of gaps or unknown sites.

Each of the test statistics in this chapter are calculated from an alignment. These test statistics are therefore susceptible to alignment error, which can be substantial. Liu et al. (2011) applied a range of sequence alignment software to 6 datasets (ranging from 117–1028 taxa and 4722–10,738 sites) and measured the error rate, which they defined as the proportion of truly homologous pairs of nucleotides (defined by the reference alignment) that are missing in the estimated alignment. They found alignment error rates of 23–40%, depending on the choice of software and dataset (Liu et al. 2011). Given these results, it's reasonable to assume that the majority of empirical phylogenetic datasets contain some level of alignment error. The extent to which each test statistic is both impacted by alignment error and able to detect alignment error depends on the underlying approach of each method. Tree proportion is agnostic to the source of decreased treelikeness, but as the method relies upon a split network the tree proportion is limited to detecting only events that introduce splits into the split network. Small errors in alignment are therefore unlikely to be detected by the tree proportion. However, error on the scale of 20–40% as in Liu et al. (2011) may be identified. Simulations would be

necessary to benchmark performance of the tree proportion under different levels of alignment error.

This chapter provides a framework for empiricists to make informed choices about the appropriate phylogenetic pipeline for different datasets. Many studies have applied the metrics for treelikeness in this chapter (Nikolaev et al. 2007; Skaloud and Peksa 2010; Kim et al. 2013; Vanhove et al. 2015; Vďačný 2017; Prasanna et al. 2020; Cunha et al. 2022), demonstrating the general desire to investigate phylogenetic datasets and select appropriate phylogenetic methods. However, until now there has been no empirical approach to facilitate selection of one of the many existing metrics for treelikeness. This chapter is a starting point to establishing a comprehensive testing suite to assess systematic bias in phylogenetic datasets. In this chapter I focused on benchmarking test statistics for treelikeness in a particular context, i.e., loci that have multiple underlying tree topologies. However, there are several causes of non-treelikeness including hybridisation and incomplete lineage sorting. Ideally, this chapter will be used in conjunction with other methods designed specifically for biological processes such as recombination (Etherington et al. 2005; Bruen et al. 2006; Kosakovsky Pond et al. 2006; Martin et al. 2015; Blischak et al. 2018; Lam et al. 2018); introgression (Patterson et al. 2012; Pfeifer and Kapan 2019; Hibbins and Hahn 2022); and incomplete lineage sorting (DeBiasse et al. 2014; Kuritzin et al. 2016; Rosenzweig et al. 2022). The results of the testing suite will then inform the tree inference process. For example, these tests may inform selection of substitution models such as GHOST (Crotty et al. 2020) which accommodates heterotachous evolutionary processes, or use of Qmaker (Minh et al. 2021) or nQmaker (Dang et al. 2022) to infer a substitution model specifically for that dataset. Alternatively, alignments with hybridisation or recombination events would benefit from explicit network inference methods such as SNaQ (Solís-Lemus et al. 2017) or NetRAX (Lutteropp et al. 2022).

An important stage in the phylogenetic pipeline is testing the underlying assumptions of the phylogenetic model prior to tree estimation (Misof et al. 2014; Jermiin et al. 2020). Testing the treelikeness of an alignment allows the data to reject the underlying model assumptions and helps inform selection of tree estimation methods. For example, alignments with low treelikeness could benefit from analysis with models that relax the treelikeness assumption, such as the MAST (Wong et al. 2024) or GHOST (Crotty et al. 2020) models, or the use of coalescent and species network models if individual loci are shown to be treelike (Solís-Lemus and Ané 2016; Solís-Lemus et al. 2017; Flouri et al. 2018; Wen et al. 2018; Zhang et al. 2018a; Rannala et al. 2020; Douglas et al. 2022). I hope this chapter facilitates assessment of treelikeness prior to tree estimation and serves as an aid for selecting tests for treelikeness based on the characteristics of individual datasets.

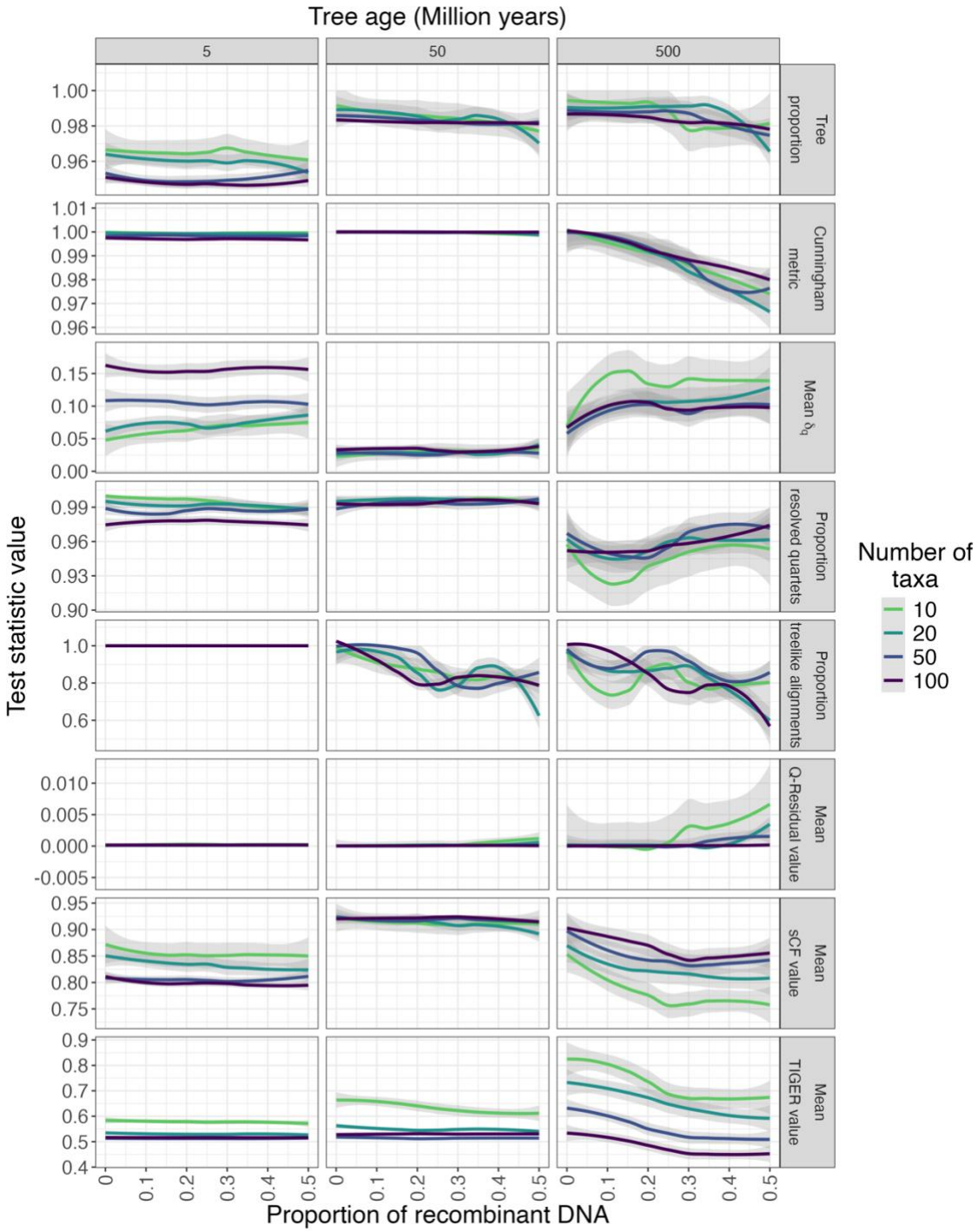
1.7 Data Availability Statement

All code to replicate these analyses are available from the GitHub repository for this project (https://github.com/caitlinch/treelikeness-metrics). Results are available at the Figshare repository (https://doi.org/10.6084/m9.figshare.26054467) which includes simulation parameters, simulation results, gene alignments, test statistic values for individual genes, and parametric bootstrap results. The empirical datasets used in this study were obtained from the data repositories of the original publications: from Figshare for the Whelan et al. (2017b) dataset (https://doi.org/10.6084/m9.figshare.4484138.v1); and from Github for the McCarthy et al. (2023) enriched orthology dataset (https://github.com/chmccarthy/ATOLRootStudy). Implementations of treelikeness metrics applied in this study were obtained from the following locations: likelihood mapping is implemented within IQ-Tree2 (Strimmer and von Haeseler 1997; Minh et al. 2020b), available at http://www.iqtree.org/; Q-residuals were implemented using Phylogemetric v1.0.0 (Greenhill 2016, 2021), available at https://github.com/SimonGreenhill/phylogemetric; δ plots were implemented using the "delta.plot" function from R package ape v5.6.2 (Paradis et al. 2004; Paradis and Schliep 2019), available at https://cran.r-project.org/web/packages/ape/index.html; the Network Treelikeness Test was calculated using tools in SplitsTree4 v4.18.2 (Huson 1998; Huson and Bryant 2006, 2022), available at https://software-ab.cs.uni-tuebingen.de/download/splitstree4/welcome.html; site concordance factors were implemented using IQ-Tree (Minh et al. 2020a; Mo et al. 2023), available at http://www.iqtree.org/; and TIGER was implemented in fast-TIGER (Frandsen 2015), available at https://github.com/pbfrandsen/fast_TIGER.

1.8 Acknowledgments

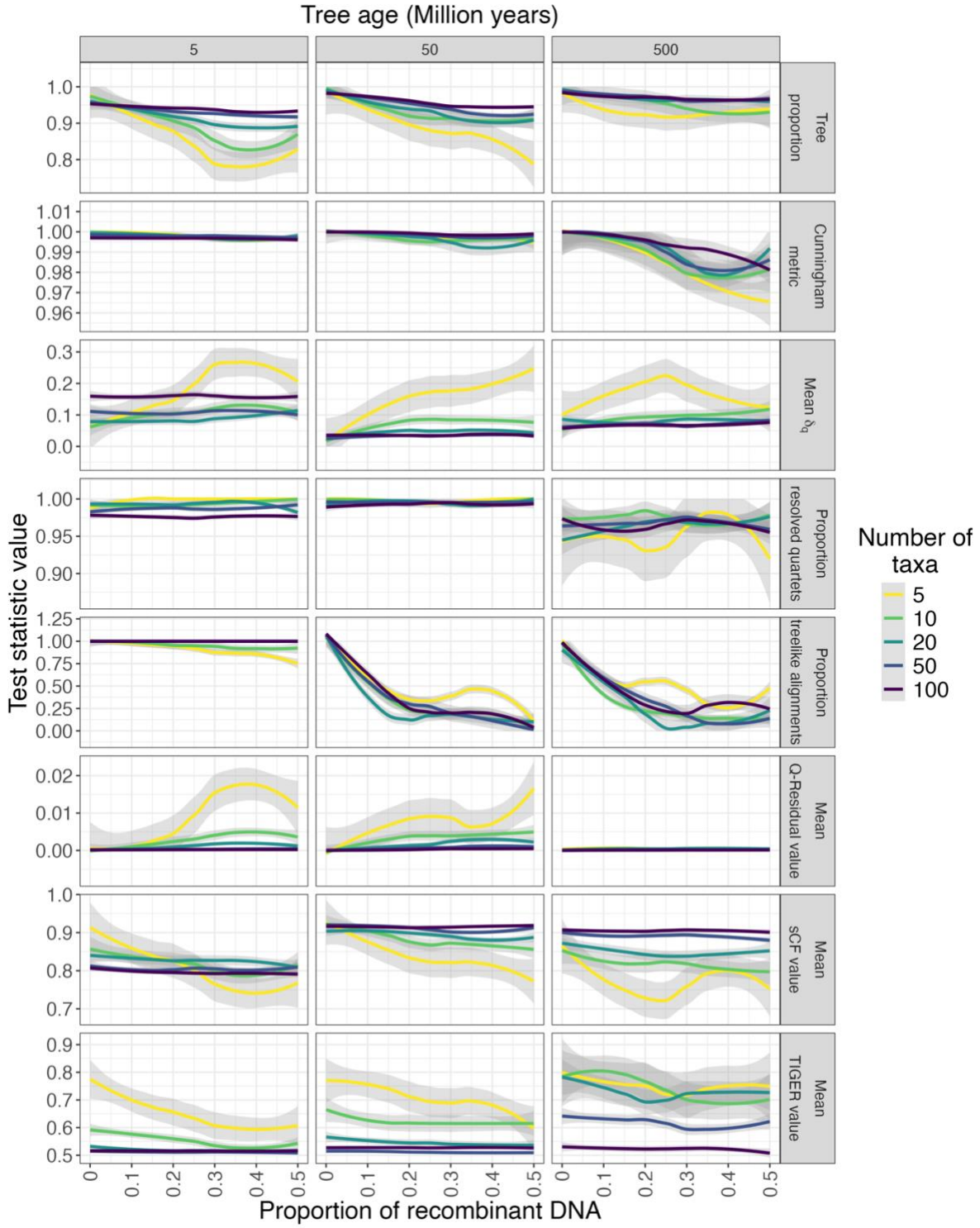
I would like to thank Barbara Holland, Maja Adamska, James Barbetti, Fred Jaya, Nhan Trong Ly and Thomas Wong for their comments on early versions of this manuscript; Teresa Neeman for statistical advice; Zachary Ryan Dowton for assistance with probability and combinatorics; and Daniel Huson for implementing a command line for the consensus network method in SplitsTree4. This work was funded by an Australian Government Research Training Program scholarship (to C.C.).

1.9 Supplementary Figures



Supplementary Figure 1: Treelikeness test statistic results for alignments with one ancient introgression event and a speciation rate of 0.1.

The y axis differs for each test statistic. The proportion of recombinant DNA ranged from 0 to 0.5 (half the alignment), and as the proportion of recombinant DNA increases the treelikeness of the alignment decreases. Each line is a smoothed conditional mean calculated using locally estimated scatterplot smoothing and the formula y~x. The light grey bands represent the 95% confidence intervals for each line. Each column is a different tree age (in millions of years), and each row is a different test statistic.



Supplementary Figure 2: Treelikeness test statistic results for alignments with one recent introgression event and a speciation rate of 0.1.

The y axis differs for each test statistic. The proportion of recombinant DNA ranged from 0 to 0.5 (half the alignment), and as the proportion of recombinant DNA increases the treelikeness of the alignment decreases. Each line is a smoothed conditional mean calculated using locally estimated scatterplot smoothing and the formula y~x. The light grey bands represent the 95% confidence intervals for each line. Each column is a different tree age (in millions of years), and each row is a different test statistic.

Chapter Two:
Removing Recombinant Loci has Minimal Impact on Species Tree Topologies Estimated from Empirical Data

Caitlin Cherryh[1]*, Bui Quang Minh[2], Robert Lanfear[1]

[1] Division of Ecology and Evolution, Research School of Biology, Australian National University, Canberra, Australia

[2] Research School of Computer Science, Australian National University, Canberra, Australia

* Corresponding author: caitlin.cherryh@anu.edu.au

Contributions:

Caitlin Cherryh designed the simulations, wrote the R scripts, performed the simulations and empirical analysis, interpreted the results, and drafted the manuscript. Minh Bui and Robert Lanfear assisted with conceptual development, experimental design, and editorial comments.

2.1 Abstract

Many methods for inferring phylogenetic trees assume that there is no within-locus recombination, but this assumption is rarely tested in empirical analyses. Including loci that violate this assumption in species tree estimation may impact the accuracy of the resulting species trees by affecting the gene trees from which they are estimated. In theory, species trees estimated from gene trees using summary methods will be more impacted than those estimated using concatenation methods. In this chapter, I ask whether excluding loci that show evidence of recombination changes the conclusions of phylogenetic analyses from empirical datasets. To do this, I apply three tests for recombination (PHI, MaxChi and GENECONV) to four empirical phylogenetic datasets and estimate trees using both concatenation and summary (also known as two-step or gene tree/species tree) methods.

Our results suggest that filtering loci for evidence of recombination sometimes results in highly supported differences between trees. In some cases, excluding putatively recombinant loci resulted in several biologically meaningful differences in tree topology. However, for almost all combinations of dataset and test for recombination, filtering resulted in minimal differences in branch length, branch support values, or quartet concordance factors.

Continued development and use of Multi-species Coalescent Network methods, and/or testing for recombinant loci prior to applying other species-tree estimation methods, is likely to improve the accuracy of phylogenetic inferences for summary methods in cases where introgression and concatalescence have been prevalent.

Keywords: Systematic bias, Treelikeness, Phylogenomics, Reticulate evolution, Phylogenetic methods, Data filtering

2.2 Introduction

Phylogenomic datasets are getting larger due to advances in sequencing and computing technology. As datasets increase in size they also increase in complexity, because larger datasets capture more biological processes that result in heterogeneous signals across the genome (Bravo et al. 2019). Phylogenetic models capture key details of biological processes but are nevertheless simplifications of the true complexity of evolution. This simplification can sometimes lead to systematic bias when model assumptions are violated by empirical data (Brown and Thomson 2018). To detect and reduce systematic bias, the data should ideally be allowed to reject the model when the model assumptions are violated (Penny 1982; Goldman

1993). However, this step is sometimes missing from phylogenetic protocols (Jermiin et al. 2020).

Most phylogenetic methods make assumptions about the absence of recombination in alignments. In concatenation approaches, alignments are combined across all loci in the dataset, and it is assumed that all sites in all loci share a common evolutionary history. This assumption is violated by empirical data when evolutionary histories vary between or within loci due to biological processes such as horizontal gene transfer, introgression, and incomplete lineage sorting (ILS) (Dasmahapatra et al. 2012; Mallet et al. 2016; Edelman et al. 2019). Coalescent approaches use stochastic methods to describe the how the genealogical relationships between species or individuals may have arisen from a common ancestor (Liu et al. 2015b). These methods incorporate gene tree heterogeneity due to ILS by allowing each gene to have a different tree topology and estimating the species tree under a model that explicitly accounts for ILS among loci. These methods can be single step, where gene trees and the species tree are simultaneously estimated, or summary methods (also known as two-step or gene/tree species tree methods), where gene trees are estimated and then used as input data to estimate a species tree (Liu et al. 2015b; Bryant and Hahn 2020). This allows coalescent methods to account for one of the processes (ILS) that can cause evolutionary histories to vary between loci. However, these methods still assume that the evolutionary history of each locus can be described by a single tree.

Recombination within loci can cause different regions within the same locus to have different evolutionary histories (Scornavacca and Galtier 2017; Mendes et al. 2019; Smith et al. 2020), which could impact tree topology as well as measures of support and concordance. However, removing gene trees estimated from recombinant genes may impact the measures of branch support (e.g., posterior probability or bootstrap values), measures of variation (e.g., concordance factors (Baum 2007; Lanfear and Hahn 2024)) and branch lengths (such as those in ASTRAL estimated from quartet concordance factors (Sayyari and Mirarab 2016)). Conversely, I expect removing genes with evidence of recombination to have little impact on concatenated tree topology (i.e., the topology of the tree estimated using a concatenation method). Concatenation methods assume from the outset that a single tree is sufficient to represent the evolutionary history of all loci, which will often result in them identifying the tree supported by the majority of phylogenetically informative sites (Mendes and Hahn 2018; Bryant and Hahn 2020). These methods do not incorporate within- or between-locus recombination, so removing recombinant genes will reduce the sample size, but providing that there is sufficient phylogenetic signal within the remaining genes, there should be little impact on tree topology estimated from concatenation methods.

Using single-locus alignments which contain multiple evolutionary histories in coalescent analyses has been called “concatalescence”, and violates the underlying assumptions of the multispecies coalescent (MSC) model (Gatesy and Springer 2013, 2014; Springer and Gatesy 2016). Concatalescence may be a relatively minor problem for phylogenetic datasets in which recombination is rare, such as datasets that use short loci, cover short timescales, and for which recombination is rare in the sampled loci. However, other datasets may be more prone to concatalescence. For example, datasets which use protein-coding genes often form loci by concatenating exons that can be hundreds of thousands of base pairs apart in the genome, and the samples may cover huge timescales (e.g. hundreds of millions of years). Both of these factors increase the probability of recombination occurring within a single-locus alignment (Gatesy and Springer 2013, 2014; Springer and Gatesy 2016; Scornavacca and Galtier 2017). Including concatalescent genes in coalescent phylogenetic analyses violates the underlying assumptions of the method and could result in inaccurate species trees. One option is therefore to attempt to filter out loci with evidence of recombination prior to species tree inference.

Previous work investigating the effect of including recombinant loci in Bayesian coalescent species tree analyses found that including recombinant loci resulted in a weak negative effect on the accuracy of species tree estimation, but the increase in accuracy due to adding more loci always outweighed any decrease from adding recombinant loci (Lanier and Knowles 2012). However, these simulations have been criticized for underestimating recombination rates and missing simulations for deep phylogenetic relationships (Gatesy and Springer 2014; Springer and Gatesy 2018). An empirical investigation using into the phylogeny of the mallard complex compared Bayesian trees (estimated in *BEAST) from both an unfiltered dataset and a filtered dataset made of putatively non-recombinant loci, and found that while some clades were recovered identically in both analyses, the placement of a number of taxa varied between the two trees (Lavretsky et al. 2014). However, the dataset for this study was small (64 individuals from 16 operational taxonomic units and 20 nuclear loci), so differences in topology between the two trees could also be due to differences in the sampling between the filtered and unfiltered datasets. Another study on the phylogeny of the Ecuadorian plant *Lachemilla* found that all phylogenetic analyses recovered the same four major clades, but differences in tree estimation method (concatenated tree estimated in RAxML or summary trees estimated using ASTRAL, MP-EST or SVDquartets) and data filtering (whole dataset; hybrid species removed; or hybrid species and putatively recombinant loci removed) resulted in different relationships between clades (Morales-Briones et al. 2018). Thus, previous work certainly demonstrates that recombination has the potential to affect species tree accuracy, but more

work is needed to clarify the generality of these conclusions, and the extent to which they depend on how putatively recombinant loci are detected.

In this chapter, I evaluate the impact of detecting and removing putatively recombinant loci on species tree estimation for three recombination detection methods and four empirical phylogenetic datasets. The datasets were carefully selected to represent different phylogenetic depths and recombination rates. For each dataset I estimate species trees using concatenation (IQ-TREE) and summary (ASTRAL) methods, using either the whole dataset or the putatively recombinant/non-recombinant sets of loci as identified by each of the three recombination detection methods I consider. I compare the resulting trees and calculate the adequacy of each tree using either the approximate unbiased (AU) test for concatenated trees (Shimodaira 2002) or the goodness of fit test for species trees (Stenz et al. 2015; Cai and Ané 2021). My results suggest that filtering loci for evidence of recombination sometimes results in highly supported differences between trees. Filtering putatively recombinant loci had a greater impact on summary trees estimated in ASTRAL than on concatenated trees for 3 datasets, but the opposite pattern was observed for the largest dataset of green plants. In all but one analyses, removing putatively recombinant loci did not result in statistically significant differences in branch support value, branch length, or quartet concordance factor.

2.3 Materials and Methods

I selected four empirical datasets where within-locus recombination was likely to be an issue. I applied three different tests for recombination to each locus. For each test, I split the loci into two subsets depending on whether each locus passed or failed that test. I then estimated species trees from each subset of loci using both concatenation and summary tree methods. Finally, for all the trees estimated from each dataset I compared the tree topology, branch lengths, and support values to determine the effects of removing putatively recombinant loci on tree estimation. For the summary trees, I also estimated and compared the quartet concordance factors.

2.3.1 Dataset and Locus Selection

I selected four empirical transcriptome datasets where each locus was made up of long loci (e.g., large genomic regions or concatenated exons) such that within-locus recombination was likely to be an issue. The four datasets comprise both two evolutionarily deep and two shallow clades, one each from both plants and animals. (Pease et al. 2016a; Whelan et al. 2017a; Leebens-Mack et al. 2019a; Vanderpool et al. 2020b). I defined shallow datasets as encompassing a single order or genus with a common ancestor within the last 100 million years. I defined deep datasets as those encompassing a large clade or kingdom with a common ancestor around 500-1000 million years ago. I chose two shallow datasets with evidence of recent introgression: a Primate dataset (Vanderpool et al. 2020b) and a Tomato dataset (Pease et al. 2016a). In addition, I selected two deep datasets which had previously been analysed to investigate the relationships between well-established clades: Plants (Leebens-Mack et al. 2019a) and Metazoans (Whelan et al. 2017a). This allowed me to compare the differences of the impacts of gene-filtering on topology estimation at different time scales and in different clades.

I assessed each locus in each dataset for inclusion in downstream analyses. I sought to exclude loci which may have strong effects on species tree estimation that may be independent of the inclusion or exclusion of putatively recombinant loci. To do this, I estimated a gene tree from each locus in each dataset using IQ-TREE (Nguyen et al. 2015; Minh et al. 2020b). I then excluded loci that were flagged by IQ-TREE during gene tree estimation with warnings that could have follow-on effects on tree estimation such as: if estimated model parameters were at a boundary that could cause numerical instability; if saturated sites led to a very long branch (more than 9.8 estimated substitutions per site); if pairwise maximum likelihood (ML) distances

were very high indicating saturation; if 1 or more states were rarely present which could present numerical issues; or if the locus was unable to run successfully in IQ-TREE.

I used a custom-written R v4.0.3 script (R Core Team 2018) with the packages ape v5.5 (Paradis and Schliep 2019), phangorn v2.7.0 (Schliep 2011), phytools 0.7.70 (Revell 2012), and seqinr v4.2.5 (Charif and Lobry 2007) to estimate the gene tree for each locus using IQ-TREE, and to remove loci based on the IQ-TREE .iqtree and .log file warnings. The script "1_Recombination_Detection.R" is available from the GitHub repository https://github.com/caitlinch/gene_filtering.

2.3.1.1 Primates dataset

I selected the Vanderpool et al. (2020b) DNA dataset, which I label the "Primate" dataset, consisting of the protein-coding sequences for 1730 genes, obtained from the genomes of 26 primates and 3 non-primates (Vanderpool et al. 2020a). I estimated maximum likelihood gene trees using the same method as Vanderpool et al. (2020b), using IQ-TREE v2.0 (Minh et al., 2020b; Nguyen et al., 2015) with models of sequence evolution selected using ModelFinder (Kalyaanamoorthy et al. 2017). I checked all warnings in the IQ-TREE .log and .iqtree files and excluded 8 loci: 7 with estimated model parameters at a boundary that could cause numerical instability, and one which had saturated sites for one sequence resulting in a very long branch (more than 9.8 estimated substitutions per site). This left me with a final dataset of 1722 loci, where each loci is the protein-coding sequence for a gene. The mean locus length was 1020.22 bp.

2.3.1.2 Tomato dataset

I selected the Pease et al. (2016a, 2016b) dataset, which I label the "Tomato" dataset. This dataset consisted of whole transcriptomes from 30 samples covering the 13 species of the wild tomato clade Solanum sect. Lycopersicon. I excluded one sample, LA1360, that was found to be a hybrid by Pease et al. (2016a) and consequently excluded from their consensus phylogeny. Neither my concatenated nor my summary method approaches accounted for hybrid species origin, so including known hybrids was likely to mislead both approaches. The consensus phylogeny of Pease et al. (2016a) was estimated from 2,745 non-overlapping genomic windows of the chromosomes with sequences for all 29 species. While each genomic window spanned 100 kb of a given chromosome, the alignment for each window included only the sites from the transcriptome (smallest window = 681, largest window = 57,569 bp, mean window size = 21,226.1 bp).

To ensure I used the same loci in my experiments, I replicated the process of generating genomic windows described by Pease et al (2016a). First, I used MVFTools (https://www.github.com/jbpease/mvftools) (Pease and Rosenzweig 2018; Pease 2021) to separate the mvf-formatted alignment Pease_etal_Tomato29acc_HQ.mvf.gz from Pease et al. (2016b) into non-overlapping 100kb genomic windows using the following python3 command:

```
python3 mvftools/mvftools.py InferTree
--mvf Pease_etal_Tomato29acc_HQ.mvf.gz --windowsize 100000
--out trees100k.txt --contig-ids 1,2,3,4,5,6,7,8,9,10,11,12
--sample-indices
0,1,2,3,4,5,6,7,8,9,10,11,12,14,15,16,17,18,19,20,21,22,23,24,25,26,
27,28,29
--raxml-path raxmlHPC-AVX2
```

This resulted in a separate alignment for each of the 8000 possible non-overlapping 100kb genomic windows. To identify the 2745 windows used in the original paper I wrote a custom R script 0_Pease2016_data_formatting.R (available from this project's GitHub repository https://github.com/caitlinch/gene_filtering) with the R packages ape v5.5 (Paradis and Schliep 2019) and phangorn v2.7.0 (Schliep 2011). The InferTree function in MVFTools generates trees for each 100kb window with RAxML (Stamatakis 2014), but for consistency with the other datasets I estimated the tree for each 100kb window in IQ-TREE as above. I checked all warnings in the IQ-TREE .log and .iqtree files as above and excluded 1 window with very high pairwise ML distances (indicating saturation). This left me with a filtered dataset of 2744 loci, where each locus is a genomic window extracted from the transcriptome. The mean locus length was 21,030.27 bp.

To visualise the distribution of gene trees from the Tomato dataset, I generated a cloudogram (Bouckaert 2010) of the gene tree topologies (Supplementary Figure 19) using the densiTree function Phangorn v2.6.3 (Schliep 2011).

2.3.1.3 Metazoan dataset

From Whelan et. al. (2017a) I selected the alignment used to estimate the main phylogeny ("Metazoa_Choano_RCFV_strict"), which I label the "Metazaon" dataset. The dataset consists of 117 genes for 76 taxa, and was constructed from transcripts from new and publicly available species to establish the placement of Ctenophora (comb jellies) and determine the sister group to all animals (Whelan et al. 2017a). I downloaded the alignment and the associated partitioning scheme (Metazoa_Choano_RCFV_strict_Models.txt) from the online repository (Whelan et al. 2017b). Using the partitioning scheme, I separated the alignment into individual genes using a custom-written R v4.0.3 script 0_Whelan2017_data_formatting.R available from

the GitHub repository https://github.com/caitlinch/gene_filtering, and the R packages phylotools v0.2.2 (Zhang 2017) and phangorn v2.7.0 (Schliep 2011). I estimated a maximum likelihood gene tree for each gene in IQ-TREE as above. I checked all warnings in the IQ-TREE .log and .iqtree files as above and excluded 6 loci: 4 with estimated model parameters that could lead to numerical instability, and 2 with a very long branch (more than 9.8 estimated substitutions per site). This left me with a final dataset of 111 loci (where each locus is a protein-coding sequence). The mean locus length was 430.59 amino acid residues.

2.3.1.4 Plants dataset

I used the "alignments-FAA-masked" version of the One Thousand Plants dataset (Leebens-Mack et al. 2019a), available from (Leebens-Mack et al. 2019b), and referred to here as the "Plants" dataset. This is the dataset used for the main results in that paper and consists of 410 genes from 1178 species, from transcriptomes and 31 published genomes. I estimated maximum likelihood gene trees for each gene in IQ-TREE as above. I checked all warnings in the IQ-TREE .log and .iqtree files as above and excluded 19 loci. 15 of these had one or more overly long branches (more than 9.8 estimated substitutions per site), 2 had states that were rarely present which may cause numerical problems, and IQ-TREE was unable to run on 2 loci. This left me with a filtered dataset of 391 loci (where each locus is a protein-coding sequence). The mean locus length was 380.99 amino acid residues.

2.3.2 Tests for Recombination

2.3.2.1 Identifying putatively recombinant loci

There are many tests for detecting recombination or introgression in multiple sequence alignments, such as 3SEQ (Lam et al. 2018), GARD (Kosakovsky Pond et al. 2006), gmos (Domazet-Lošo and Domazet-Lošo 2016), Hyde (Blischak et al. 2018), Patterson's D statistic (Patterson et al. 2012; Pfeifer and Kapan 2019), RAT (Etherington et al. 2005), and the compilation of methods in RDP5 (Martin et al. 2020). My criteria for tests for recombination detection were that the test was widely used, the test had been validated for multiple datasets or experimental conditions, and that the tests selected had different underlying methods. I selected three tests from the literature: the maximum chi-squared method (MaxChi) (Maynard Smith 1992), GENECONV (Sawyer 1989) and the Pairwise Homoplasy Index (PHI) (Bruen et al. 2006). Each of these tests has been highly cited: over 1700 citations for the maximum chi-squared method, over 1000 citations for GENECONV and over 1300 citations for PHI. Each of these tests had also been validated in previous studies. GENECONV and the MaxChi test have previously been found to reliably detect recombination in both simulations and empirical

datasets (Posada and Crandall 2001; Posada 2002). The PHI test is more powerful when sequences are closely related (Bruen et al. 2006), whereas MaxChi and GENECONV perform better when sequences are more divergent (Posada and Crandall 2001). I applied the MaxChi and PHI tests using PHIPack (Bruen 2005), and GENECONV using the GENECONV v1.8 implementation (Sawyer 2000).

I used a custom-written R v4.0.3 script (R Core Team 2018) with the packages ape v5.5 (Paradis and Schliep 2019), phangorn v2.7.9 (Schliep 2011), phytools 0.7.70 (Revell 2012), and seqinr v4.2.5 (Charif and Lobry 2007) to apply each test for recombination to each locus. The script "1_Recombination_Detection.R" is available from the GitHub repository https://github.com/caitlinch/gene_filtering.

2.3.2.2 Applying three tests of recombination to each dataset

Applying all three tests to each dataset led me to create 9 subsets of loci for each dataset. Each locus could be included in multiple subsets, depending on the results from the recombination detection tests. The first subset, Unfiltered, contained all loci in that dataset. The next six subsets were associated with the loci that passed or failed each of the three tests. For each test I created one subset which contained loci that passed the test (labelled as "P_test_name", e.g. "P_PHI"); and one which contained loci that failed the test (i.e., were identified as putatively recombinant; labelled "F_test_name", e.g. "F_PHI"). Specifically, P_PHI contained all loci that had a non-significant p-value for the PHI test, and F_PHI contained all loci that had a significant p-value for the PHI test and thus reject the null hypothesis of the PHI test that distant sites in the alignment have the same genealogical correlation as adjacent sites (Bruen et al. 2006). P_MaxChi contained all loci that had a non-significant p-value for the MaxChi test, and F_MaxChi contained all loci that had a significant p-value for the MaxChi and thus reject the null hypothesis that there is no mosaic structure present within the alignment (Maynard Smith 1992). P_GENECONV contained all loci that had a non-significant p-value for both the inner fragments and outer fragments, meaning that GENECONV found no evidence for gene conversion events between ancestors of two species within the alignment (inner) or evidence of gene conversion events that may have originated outside of the alignment (outer). F_GENECONV contained all loci that had a significant p-value for either of both GENECONV test (either inner or outer), and thus reject the null hypothesis that no signs of gene conversion are present within the alignment. The remaining two subsets of loci, were based on combinations of all three tests of recombination: P_All contained the loci that passed all three tests (i.e., showed no significant evidence of recombination on any test) and F_All contained all loci that failed one or more test (i.e., returned a significant p-value for at least one

test). This resulted in a total of nine subsets of loci for each dataset: Unfiltered, P_PHI, F_PHI, P_MaxChi, F_MaxChi, P_GENECONV, F_GENECONV, P_All, and F_All. I refer to the group of subsets with putatively recombinant loci removed as the P_test subsets.

The tests for introgression did not successfully run on some loci. In these cases, I set the result for that test statistic for that locus to NA (as the putative recombination status of the loci could not be identified). For the GENECONV test only, I assigned loci that were unable to obtain both an inner and outer p-value (had either an NA inner p-value or NA outer p-value or both) to the NA subset. For the All category where all three tests were applied concurrently, I assigned any loci with NA for one or more test to the NA subset.

2.3.3 Species Tree Inference and Comparison

I estimated a tree for each of the nine subsets of loci for each of the four datasets using both a concatenated analysis in IQ-TREE (Minh et al. 2020b) and a gene-tree/species-tree analysis in ASTRAL-III (Zhang et al. 2018b).

2.3.3.1 Species tree estimation

I used both concatenation and summary methods to estimate species trees for each of the 9 subsets of loci across the 4 datasets (36 analyses per tree estimation method). To ensure my trees were not unduly affected by sampling error, I restricted those analyses to only those subsets which contained at least 50 loci, resulting in 11 subsets of loci being removed (i.e., 25 analyses per tree estimation method remaining; Table 3). In previous studies, 50 loci has been chosen as the minimum number for estimating species trees with summary methods (Nute et al. 2018; Zhang et al. 2018b). To ensure each comparison across subset trees was sensible, if the P subset for any test contained fewer than 50 loci, I did not estimate a tree from the corresponding F subset regardless of the number of loci in that subset. For clarity, I denote each of the trees with the inference method (ASTRAL or CONCAT) and a subscript that corresponds to the subset of the data from which they were generated. For example: the trees estimated from the complete set of loci are $ASTRAL_{Unfiltered}$ and $CONCAT_{Unfiltered}$; the trees estimated from the subset of loci that pass the PHI test are labelled $ASTRAL_{P\_PHI}$ and $CONCAT_{P\_PHI}$; and the trees estimated from the subset of loci that fail the PHI test are labelled $ASTRAL_{F\_PHI}$ and $CONCAT_{F\_PHI}$.

I used ASTRAL v5.7.5 (Mirarab et al. 2014; Sayyari and Mirarab 2016; Zhang et al. 2018b; Mirarab 2023) to estimate trees using the command "`java -jar astral.5.7.5.jar -i subset_trees.text -o subset_tree.tre 2> subset_tree.log`", where

`subset_trees.txt` is the file containing the gene trees for that subset of loci and `subset_tree.tre` is the output tree file. I extracted the local posterior probability (Lpp) for each analysis from the ASTRAL output tree. ASTRAL does not calculate the length of terminal branches, so for tree comparison I arbitrarily assigned each terminal branch a branch length of 0.1 coalescent units.

For the Primate, Tomato and Metazoan datasets I estimated a concatenated tree for each subset in IQ-TREE v2.0-rc1 (Nguyen et al. 2015; Minh et al. 2020b) with 1000 ultrafast bootstraps (UFB) (Hoang et al. 2018a) using a partitioned analysis (Chernomor et al. 2016; Biczok et al. 2018). The IQ-TREE command was "`iqtree -p partitions.nex -m MERGE -bb 1000 -nt AUTO`". The partition file specified the model of substitution for each locus as the best model from gene tree estimation as identified by ModelFinder (Kalyaanamoorthy et al. 2017).

The size of the Plants dataset and the complexity of the models selected for each locus meant estimating a concatenated tree in IQ-TREE was intractable. Thus for this dataset I estimated a maximum likelihood tree for each subset using RAxML-NG (Kozlov et al. 2019). The best fitting models of substitution identified for each locus of the Plants dataset using ModelFinder contained free rate models, and the additional numbers of rate categories needed for these models made estimating trees for this dataset computationally intractable. Therefore, for this dataset I selected the best-fitting model (by BIC score) from ModelFinder that did not incorporate free rate models for each locus. Due to the computational requirements of this dataset, I estimated tree topology without bootstraps using one parsimony starting tree with the command "`raxml-ng --search --msa supermat.phy --model partition.txt --brlen scaled --tree pars{1} --threads 50 --lh-epsilon 1`". The R script "2.5_ExtractingModels_DeepDatasets.R" to extract the best model without a free rate parameter for each locus in the Plants dataset is available from the GitHub repository https://github.com/caitlinch/gene_filtering.

I used a custom-written R script (R Core Team 2018) with the packages ape v5.5 (Paradis and Schliep 2019), phangorn v2.7.0 (Schliep 2011), phytools 0.7.70 (Revell 2012), phylotools 0.2.2 (Zhang 2017) and seqinr v4.2.5 (Charif and Lobry 2007) to estimate the species trees (by running ASTRAL and IQ-TREE) for each analysis and compare the resulting trees. The script "2_Species_Tree_Estimation.R" is available from the GitHub repository https://github.com/caitlinch/gene_filtering. All trees are available from the Figshare repository https://doi.org/10.6084/m9.figshare.26087437.

2.3.3.2 Species tree comparison

To determine whether removing putatively recombinant loci from species tree analysis resulted in changes to tree topology, I compared the topologies of the pass tree for each test to the topology of the unfiltered tree (the tree estimated from all loci). For example, this means that for the trees estimated in ASTRAL for one dataset, I compared each of the trees estimated from putatively non-recombinant loci ($ASTRAL_{P\_PHI}$, $ASTRAL_{P\_MAXCHI}$, $ASTRAL_{P\_GENECONV}$) to the tree estimated from all loci ($ASTRAL_{Unfiltered}$).

Where computationally feasible, I used the goodness of fit test (Ané 2021) to compare species trees estimated in ASTRAL-III. The goodness of fit test quantifies the fit between multi-locus data and patterns expected under multispecies network coalescent for a given phylogenetic tree or network (Stenz et al. 2015; Cai and Ané 2021) and is available in the Julia package QuartetNetworkGoodnessFit (Ané 2021). I ran this test in Julia v1.6.1 (Bezanson et al. 2017) with the packages PhyloNetworks.jl v0.14.0 (Solís-Lemus et al. 2017) and DataFrames v 1.1.1 (Bouchet-Valat and Kamiński 2023). The number of taxa in the Plants and Metazoan datasets were too large to reasonably apply this test, so I tested only the Primates and Tomatoes datasets. The trees for each dataset were rooted at the outgroup from the original study: Mus musculus for Vanderpool et al. (2020b) and LA4116, LA2951 and LA4126 for Pease et al. (2016a). The goodness of fit test is designed for ultrametric trees or networks, and when validating the test with simulations Cai and Ané (2021) modified branch lengths in their trees to ensure all root-to-tip paths had equal length. I used the phytools v0.7-70 function `force.ultrametric(tree, method = "extend")` to convert trees to ultrametric trees by extending all external edges of the tree to match the external edge with the greatest height (Revell 2012). For the four tests (PHI, MaxChi, GENECONV and All tests), I compare each of the three trees ($ASTRAL_{P\_test}$, $ASTRAL_{F\_test}$, and $ASTRAL_{Unfiltered}$) individually to the quartet concordance factors calculated from $ASTRAL_{P\_test}$. I applied the goodness of fit test with 100 simulated replicate datasets.

The goodness of fit test outputs a p-value, a Z statistic to describe the deviation of the proportion of outliers from the expected proportion of outliers ($p < 0.05$), and a $\hat{\sigma}$ value used to correct for the non-independence of quartets. A non-significant p-value for this test indicates that the null hypothesis cannot be rejected. In this case, the null hypothesis is that the data evolved along the provided tree or network. I then calculated the statistic $Z/\hat{\sigma}$ for each tree to correct the Z statistic for dependence (Cai and Ané 2021). I extracted the p-value and the statistic $Z/\hat{\sigma}$ from the results of each tree (i.e., the three trees $ASTRAL_{P\_test}$, $ASTRAL_{F\_test}$, and $ASTRAL_{Unfiltered}$).

To statistically compare the concatenated trees, I used the IQ-TREE implementation of the approximately unbiased (AU) test (Shimodaira 2002). For each test (PHI, MaxChi, GENECONV and All tests), I used the alignments of the subset of loci that passed that test ($CONCAT_{P\_test}$) to ask whether the tree estimated from that subset of loci could reject the tree estimated from the unfiltered set of loci ($CONCAT_{Unfiltered}$) and the tree estimated from loci that failed the test ($CONCAT_{F\_test}$). I used the command "`iqtree_path -p partition.nex -z trees.text -n 0 -zb 10000 -zw -au`". The AU test implementation calculates p-values, such that any tree with a statistically significant p-value (<0.05) is rejected in favour of the maximum-likelihood tree ($CONCAT_{P\_test}$ in this case).

Finally, I analysed the differences in the topologies of trees estimated from different subsets of data. As I am interested in whether removing genes with evidence of recombination impacts species tree topology, I focus on comparing the Unfiltered tree with each of the trees estimated from a subset of loci which passed one or more of the tests for recombination. Depending on tree estimation method, I calculated the distance between P_test tree, and either $ASTRAL_{Unfiltered}$ or $CONCAT_{Unfiltered}$. I calculated the Robinson-Foulds (RF) distance (Robinson and Foulds 1981) and the branch weighted Robinson-Foulds (wRF) distance using the RF.dist and wRF.dist functions in the R package phangorn (Schliep 2011). I also calculated the normalised Robinson-Foulds (nRF) distance (Steel and Penny 1993), which scales the RF distance by the total number of splits in both trees (Equation 5).

Equation 5

$$nRF\ distance = \frac{RF\ distance}{2(2n-3)}$$

where n is the number of taxa in either tree

In addition, I identified the branches that differed between each subset tree and either $ASTRAL_{Unfiltered}$ or $CONCAT_{Unfiltered}$ using the `distinct.edges()` function in the R package distory (Chakerian and Holmes 2020). For each pair of trees, I identified and extracted both the congruent branches (i.e., branches present in both trees) and conflicting branches (i.e., branches present in one tree only). I also extracted the length of each branch and the branch support value (either the UFB value for CONCAT trees or the Lpp value for ASTRAL trees). I manually compared all branches that had been identified as distinct (i.e., branches that were present in only one of two trees being compared). For example, I manually compared each branch that was present in $ASTRAL_{Unfiltered}$ but not $ASTRAL_{P\_MaxChi}$ and vice versa. For each distinct branch, I compared the branch length and support value and determined the extent of the impact on tree topology, specifically whether the change resulted in taxonomic

reclassification. The taxonomic rank of each dataset differed: species were classified at the section level in the Tomatoes dataset; the family level in the Primates dataset; the family, class or order level for the Plants dataset; and the infrakingdom or phyla level for the Metazoan dataset.

For the Plants dataset, I followed the procedure for the main figure from Leebens-Mack et. al. (2019a) and assigned each taxon to the taxonomic group specified by the Very Brief Classifications column in the supplementary material file annotations.csv (Leebens-Mack et al. 2019b). For the three remaining datasets (Tomatoes, Primates, and Metazoan), I assigned each taxon to taxonomic groups using the main phylogeny in the results from the original manuscript.

To identify important differences in the tree, I extracted all conflicting branches (i.e., branches that were present in only one of either the Unfiltered tree or the P_test subset tree) with high support (Lpp of > 0.9 or UFB > 90). I call these branches “highly supported differences”. The Lpp threshold of 0.9 was previously validated (Sayyari and Mirarab 2016), and I applied an identical threshold for UFB, although I note that UFB values ≥ 95 are considered high support (Minh et al. 2013; Hoang et al. 2018a). I identified 40 highly supported differences in total (Supplementary Table 3). Finally, I compared the branch length and branch support values for congruent and conflicting branches.

I applied my analyses using custom-written R version 4.0.3 scripts (R Core Team 2018) with the packages ape v5.5 (Paradis and Schliep 2019), distory v1.4.4 (Chakerian and Holmes 2020), dplyr v1.1.4 (Wickham et al. 2021), ggplot2 v3.3.3 (Wickham 2016), phangorn v2.7.0 (Schliep 2011), phytools v0.7.70 (Revell 2012), phylotools v0.2.2 (Zhang 2017), treespace v1.1.4.3 (Jombart et al. 2017), TreeTools v1.10.0 (Smith and Paradis 2023), and seqinr v4.2.5 (Charif and Lobry 2007) to compare species trees and generate plots. Plots were constructed in R using ggplot v3.3.6 (Wickham 2016), ggtern v3.3.5 (Hamilton and Ferry 2018), and patchwork v1.1.1 (Pedersen 2022). Phylogenetic trees were plotted in R using the packages colorBlindness v0.1.9 (Ou 2021), dplyr v1.0.9 (Wickham et al. 2021) , ggplot2 v3.3.6 (Wickham 2016), ggtext v0.1.1 (Wilke 2020), ggtree v.2.4.1 (Yu et al. 2017, 2018; Yu 2020), glue v1.6.2 (Bryan 2022), grid v4.0.3 (R Core Team 2018) and patchwork v1.1.1 (Pedersen 2022). The scripts “3_Species_Tree_Comparison.R”, “4_DataAnalysis.R”, “5_Densitrees.R”, “5_Tanglegrams.R” and “5_Plots.R” are available from the GitHub repository https://github.com/caitlinch/gene_filtering.

2.3.3.3 Quartet concordance factors

I quantified the change in quartet concordance factors (qCF) for between each subset tree and the $ASTRAL_{Unfiltered}$ tree. in ASTRAL v5.7.5 using the command "`java -jar astral.5.7.5.jar -i subset_gene_trees.txt -t 2 -o subset_qCF.tre 2> subset_qCF.log`".

For each pair of trees, I identified and extracted both the congruent and conflicting branches, along with the qCF for each branch. I compared the qCF values of congruent and conflicting branches for all datasets and recombination tests as above. I performed my analyses with a custom written R v4.0.3 script and the packages mentioned above. The script "6_Quartet_Concordance.R" is available from the GitHub repository https://github.com/caitlinch/gene_filtering. All qCFs are available from the Figshare repository https://doi.org/10.6084/m9.figshare.26087437.

2.4 Results

2.4.1 All datasets contained putatively recombinant loci

Table 3: The percent of loci identified as recombinant for four different empirical phylogenetic datasets by three recombination detection tests.

Each Pass row presents the percentage of loci that passed that test and were identified as non-recombinant. Each Fail row presents the percentage of loci that failed that test and were identified as non-recombinant. The NA row presents the percentage of loci that failed to run for each test. Loci that failed to run were not included in tree estimation. The GENECONV test reports two p-values (one for the recombination present within the dataset, and one for signs of recombination between one or more taxa in the dataset and taxa outside the dataset). For this test, the Pass row represents the number of loci that passed both tests (no signs of recombination within or outside the dataset). All loci identified as putatively recombinant (either inside or outside the dataset) were assigned to the GENECONV Fail category. Bold values indicate subsets with less than fifty loci, which were excluded from tree estimation. All percentages are rounded to 2 decimal places.

Percent of loci (%)		Dataset			
		Primate	Tomato	Metazoan	Plants
PHI	Pass	95.01	50.22	83.78	89.00
	Fail	4.94	49.71	**0.90**	**0.51**
	NA	0.05	0.07	15.32	10.49
MaxChi	Pass	68.82	56.38	82.88	98.98
	Fail	31.18	43.55	**16.22**	**0.26**
	NA	0	0.07	0.09	0.77
GENECONV	Pass	58.94	35.86	51.35	**6.65**
	Fail	40.48	64.12	**33.33**	**0.26**
	NA	0.58	0	15.32	93.09
All Tests	Pass	43.37	18.77	**36.04**	**1.28**
	Fail	50.99	80.16	**33.33**	**0**
	NA	0.06	0.07	30.63	92.72
Total number of loci		1722	2744	111	391

Detected levels of recombination varied between recombination detection tests and datasets (Table 3). The Tomato dataset had the highest mean percentage of putatively recombinant loci, followed by the Primates, the Metazoans, and the Plants. Of the three tests, GENECONV identified the most loci as potentially recombinant, followed by MaxChi and PHI. GENECONV also failed to run on the highest proportion of loci compared to the PHI test and the MaxChi test. The combination of dataset and test with the highest proportion of loci unable to run was GENECONV and the Plants dataset. Both the PHI and GENECONV tests failed to run on around 15% of the Metazoan dataset loci.

The Tomato dataset had the highest proportion of putatively recombinant genes for all three tests for recombination. To visualise the distribution of topologies for this dataset I plotted a

cloudogram of all gene trees within the Tomato dataset (Supplementary Figure 19), which revealed substantial gene tree heterogeneity.

2.4.2 Filtering putatively recombinant loci impacts the topology of ASTRAL trees more than concatenated trees

After subsets with fewer than fifty loci were excluded from further analyses (see Methods), I was left with 9/9 subsets for the Primate dataset, 9/9 subsets for the Tomato dataset, 4/9 subsets for the Metazoan dataset and 3/9 subsets for the Plants dataset. One summary (ASTRAL) tree and one maximum likelihood (CONCAT) tree were estimated from each subset, resulting in 18 trees for the Primate and Tomato datasets, 8 trees for the Metazoan dataset and 6 trees for the Plants dataset.

For the Primates and Tomatoes datasets, the difference in tree topologies between filtered (i.e., using loci with no evidence of recombination) and unfiltered datasets measured using ASTRAL trees were equal to or larger than those estimated from the CONCAT trees (compare values in Supplementary Table 1 and Supplementary Table 2). This is the case for all three tests of recombination and for every dataset, although I note that for Primates all trees estimated from P_test subsets were identical to trees estimated from the unfiltered datasets (Figure 10, Supplementary Figure 3, Supplementary Figure 7, Supplementary Figure 8). The Plants dataset had greater RF distances for the CONCAT trees, and the Metazoan dataset had no consistent trend (Supplementary Table 1, Supplementary Table 2).

There were few topological differences between trees estimated from different subsets of the Tomatoes dataset (Figure 11, Supplementary Figure 4, Supplementary Figure 9, Supplementary Figure 10). Of the $CONCAT_{P\_test}$ trees, 3/4 (P_PHI, P_MaxChi, P_GENECONV) were identical to the $CONCAT_{Unfiltered}$ tree (Supplementary Table 2). The remaining tree $CONCAT_{P\_All}$ had one branch different to $CONCAT_{Unfiltered}$. $ASTRAL_{P\_test}$ trees were different to $ASTRAL_{Unfiltered}$ by only 1-3 conflicting branches (Supplementary Table 1). Differences in tree topology resulted in different resolutions of the Peruvianum clade, and particularly the placement of the two taxa *Solanum peruvianum 2744* and *Solanum huaylasense 1364*. I identified 15 highly supported differences (Lpp > 0.9) within ASTRAL trees estimated from the Tomatoes dataset (Supplementary Table 3). Each of these highly supported differences resulted in different topology of the Peruvianum clade in the $ASTRAL_{P\_test}$ and $ASTRAL_{Unfiltered}$ trees.

For the Metazoan dataset, the distance between the Unfiltered and P_test trees depended on the choice of recombination test (Figure 12, Supplementary Figure 5, Supplementary Figure

11, Supplementary Figure 12). Trees estimated from P_GENECONV had higher RF distances (ASTRAL: 30, CONCAT: 24) than those estimated from the P_PHI (ASTRAL: 4, CONCAT: 2) or P_MaxChi (ASTRAL: 6, CONCAT: 8) subsets. The Metazoan dataset had the highest number of highly supported differences (Supplementary Table 3), with 8 highly supported differences from ASTRAL trees (Lpp > 0.9) and 16 from CONCAT trees (UFB >90). All highly supported differences from the ASTRAL trees and 15/16 highly supported differences from the CONCAT trees resulted in different topologies of the Ctenophora clade. Only 1 highly supported differences from the $CONCAT_{P\_GENECONV}$ tree disrupted the relationships between clades, corresponding to an alternate placement of *Trichoplax adhaerens*, the single-taxon Placozoa clade (Supplementary Figure 14). In the Metazoan tree $CONCAT_{P\_GENECONV}$, the relationships between clades are defined by the tree **(Outgroup, (Ctenophora, (Placozoa, (Porifera, (Cnidaria, Bilateria)))))** whereas in the tree $CONCAT_{Unfiltered}$ I observe the tree **(Outgroup, (Ctenophora, (Porifera, (Placozoa, (Cnidaria, Bilateria)))))**.

The largest RF distances were observed when comparing trees estimated from different subsets of the Plants dataset (Figure 4, Supplementary Figure 6, Supplementary Figure 13). In the Plants dataset, RF distances between the P_test tree and the Unfiltered tree were higher for CONCAT trees (P_PHI: 62, P_MaxCHI: 58) than for ASTRAL trees (P_PHI: 42, P_MaxCHI: 18). The differences in branch length varied, with higher wRF distances for ASTRAL (P_PHI: 85.18, P_MaxCHI: 11.48) than for CONCAT trees (P_PHI: 2.89, P_MaxCHI: 1.47). While there were the highest number of differences between trees estimated from different subsets of the Plants dataset, most of these branches had low support. I identified a single highly supported difference in the Plants ASTRAL trees (Lpp > 0.9). This branch occurred within the Malpighiales clade (Classification from Leebens-Mack et al. (2019a): CoreEudicots/Rosids Malpighiales) and places the species *Phyllanthus sp.* and *Bischofia javanica* as sisters within a cherry for the tree $ASTRAL_{P\_PHI}$ (but not for $ASTRAL_{Unfiltered}$) (Supplementary Figure 15).

I compared the normalised Robinson-Foulds (nRF) distances from each dataset to determine whether tree depth impacted ASTRAL tree topology (Supplementary Table 1). The dataset with the highest mean nRF distance was the Metazoans (mean nRF = 0.0447), followed by the Tomatoes dataset (mean nRF = 0.0363) and finally the Plants dataset (mean nRF = 0.0065). I also calculated the nRF distance for CONCAT trees (Supplementary Table 2). The Metazoan dataset had the highest mean nRF distance (mean nRF = 0.0383), followed by the Plants dataset (mean nRF = 0.0125) and finally the Tomatoes dataset (mean nRF = 0.0045). Although the Plants had the highest RF distances for both ASTRAL and CONCAT, the differences between trees estimated from different subsets were low after adjusting for the comparatively large number of taxa in this dataset. The Primates dataset had the lowest

normalised Robinson-Foulds distance for both ASTRAL and CONCAT as all trees estimated from P_test subsets were identical to the unfiltered tree (ASTRAL: mean nRF = 0, CONCAT: mean nRF = 0).

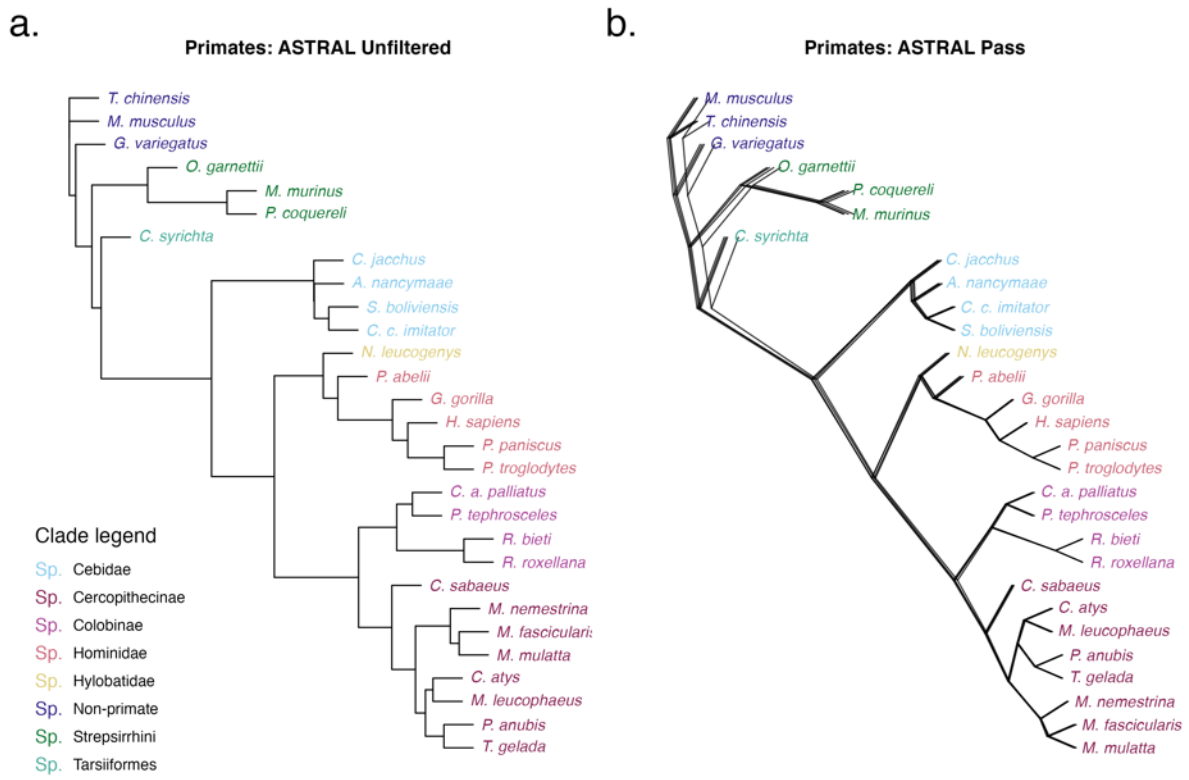


Figure 10: Comparing the ASTRAL tree estimated from the Unfiltered Primates dataset ($ASTRAL_{Unfiltered}$) with the four trees estimated from subsets of putatively non-recombinant loci. Each of the four trees estimated from the P_test subsets have identical topology to the $ASTRAL_{Unfiltered}$ tree.

ASTRAL does not estimate terminal branch lengths, so terminal branch lengths were arbitrarily assigned a length and added for plotting purposes only.

a. ASTRAL tree estimated from the Unfiltered Primates dataset. Tip labels are coloured by clade according to the Clade Legend.

b. Densitree for the ASTRAL Primate trees estimated from putatively non-recombinant loci. Each tree was estimated in ASTRAL from subsets of loci that passed each test (P_PHI, P_MaxChi, P_GENECONV, P_All). Tip labels are coloured by clade as above. For the Primates dataset, all ASTRAL P_test trees have identical topology to the $ASTRAL_{Unfiltered}$ tree.

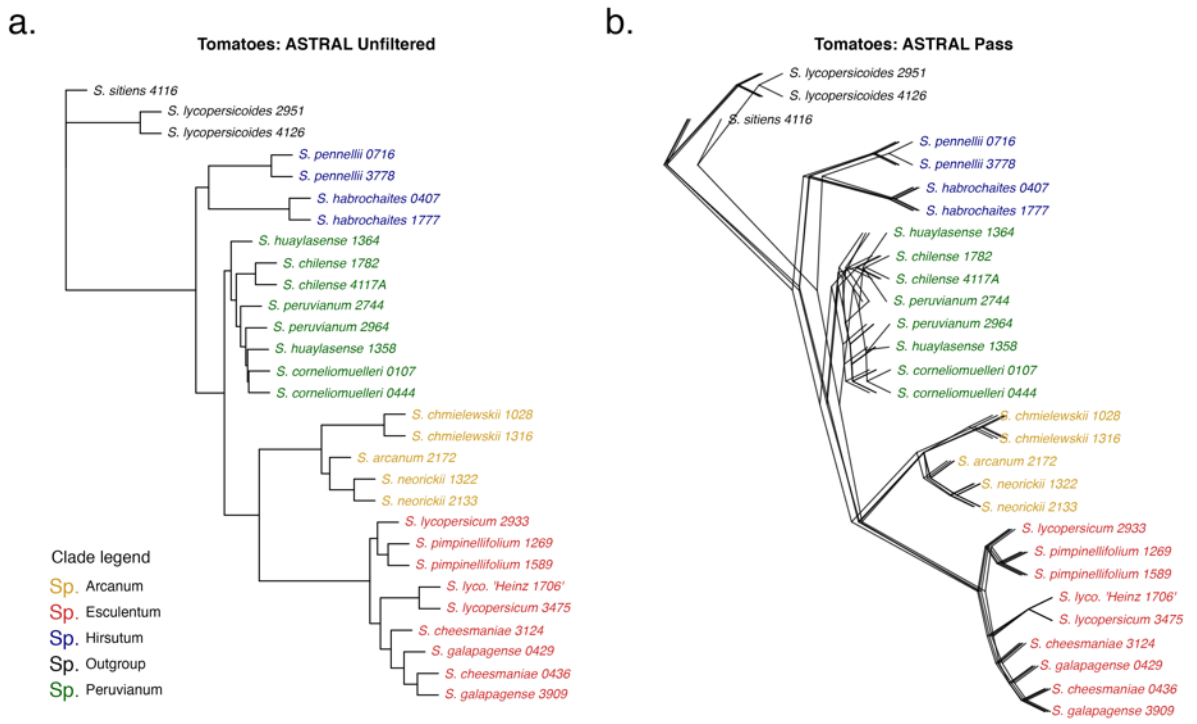


Figure 11: Comparing the ASTRAL tree estimated from the Unfiltered Tomatoes dataset ($ASTRAL_{Unfiltered}$) with the four trees estimated from subsets of putatively non-recombinant loci. The four trees estimated from the P_test subsets have identical topology to the $ASTRAL_{Unfiltered}$ tree for the Arcanum, Esculentum, Hirsutum and Outgroup clades. Compared to the tree $ASTRAL_{Unfiltered}$, the four trees estimated from the P_test subsets have different topology of the Peruvianum clade.

ASTRAL does not estimate terminal branch lengths, so terminal branch lengths were arbitrarily assigned a length and added for plotting purposes only.

a. ASTRAL tree estimated from the Unfiltered Tomatoes dataset. Tip labels are coloured by clade according to the Clade Legend.

b. Densitree for the ASTRAL Tomatoes trees estimated from putatively non-recombinant loci. Each tree was estimated in ASTRAL from subsets of loci that passed each test (P_PHI, P_MaxChi, P_GENECONV, P_All). Tip labels are coloured by clade as above. For the Tomatoes dataset, all ASTRAL P_test trees differ in topology to the $ASTRAL_{Unfiltered}$ tree only in the Peruvianum clade.

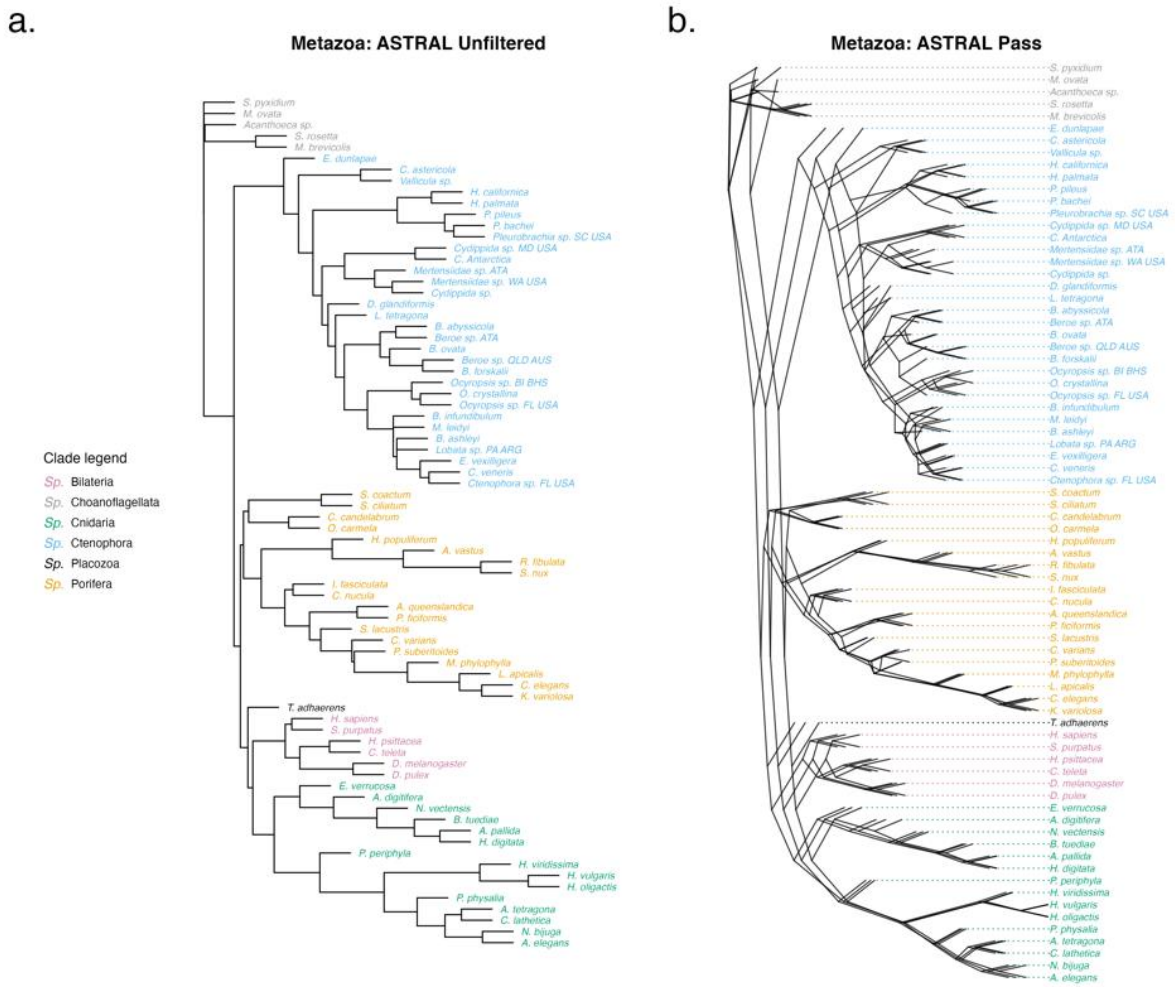


Figure 12: Comparing the ASTRAL tree estimated from the Unfiltered Metazoan dataset ($ASTRAL_{Unfiltered}$) with the four trees estimated from subsets of putatively non-recombinant loci. Each of the four trees estimated from the P_test subsets have identical relationships between established metazoan clades as the $ASTRAL_{Unfiltered}$ tree. The vast majority of topological differences between the P_test trees and the $ASTRAL_{Unfiltered}$ tree occur within the Ctenophora clade.

ASTRAL does not estimate terminal branch lengths, so terminal branch lengths were arbitrarily assigned a length and added for plotting purposes only.

a. ASTRAL tree estimated from the Unfiltered Metazoan dataset. Tip labels are coloured by clade according to the Clade Legend.

b. Densitree for the ASTRAL Metazoan trees estimated from putatively non-recombinant loci. Each tree was estimated in ASTRAL from subsets of loci that passed each test (P_PHI, P_MaxChi, P_GENECONV). Tip labels are coloured by clade as above. The majority of topological differences between the $ASTRAL_{P\_test}$ trees and the $ASTRAL_{Unfiltered}$ tree occur within the Ctenophora clade. There are also differences in the topology of the outgroup Choanoflagellata, and within Porifera in the placement of *C. varians* and *P. suberitoides.*

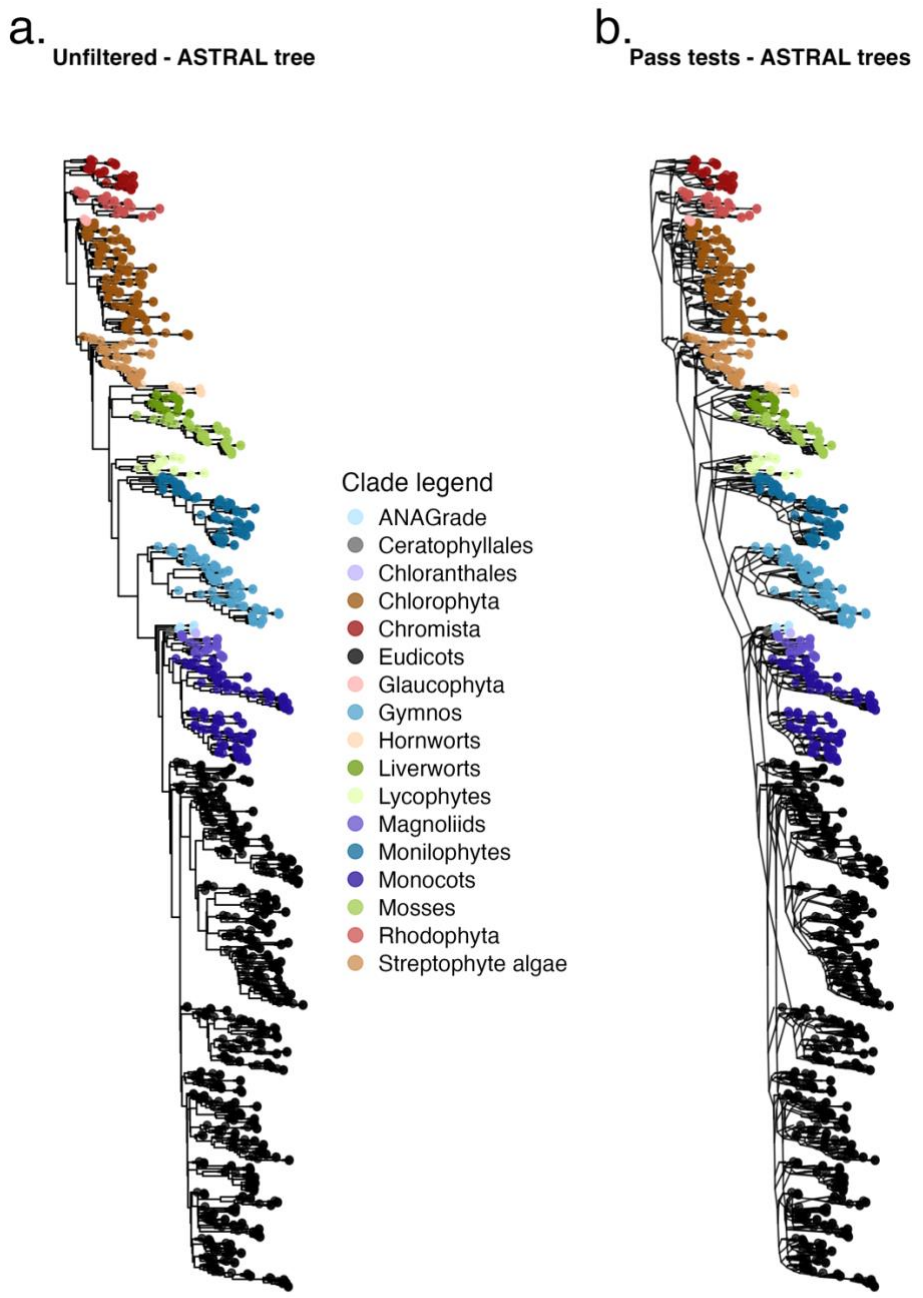


Figure 13: Comparing the ASTRAL tree estimated from the Unfiltered Plants dataset ($ASTRAL_{Unfiltered}$) with the four trees estimated from subsets of putatively non-recombinant loci. The relationships between well-established clades of the Plants dataset are identical when comparing the $ASTRAL_{Unfiltered}$ tree with the trees estimated from the P_test subsets.

ASTRAL does not estimate terminal branch lengths, so terminal branch lengths were arbitrarily assigned a length and added for plotting purposes only.

a. ASTRAL tree estimated from the Unfiltered Plants dataset. Tip labels are coloured by clade according to the Clade Legend.

b. Densitree for the ASTRAL Plants trees estimated from putatively non-recombinant loci. Each tree was estimated in ASTRAL from subsets of loci that passed each test (P_PHI, P_MaxChi). Tip labels are coloured by clade as above. The relationships between clades of the Plants dataset are consistent between $ASTRAL_{P\_test}$ trees.

2.4.3 Gene filtering very rarely impacted statistical tests of tree adequacy

I applied goodness of fit tests to determine whether removing putatively-recombinant loci resulted in statistically significant changes in tree topology. I applied the Quartet Network goodness of fit test to ASTRAL trees and the AU test to CONCAT trees. There were no differences in test results for filtered and unfiltered trees (Supplementary Table 1). The Quartet Network goodness of fit test had a statistically significant p-value ($p < 0.05$) for every ASTRAL tree in both the Primates and Tomatoes datasets (Supplementary Table 1), meaning that all ASTRAL trees were inadequate to explain observed patterns of evolution. I was unable to apply the goodness of fit test to the Metazoan and Plants datasets.

In total, 3 trees were rejected by the AU test (Supplementary Table 2). Of all CONCAT trees estimated from P_test subsets, only 1 tree was rejected by the AU test: the Metazoan $CONCAT_{Unfiltered}$ was rejected ($p = 0.006$) when compared to the tree $CONCAT_{P\_GENECONV}$ under the loci from subset P_GENECONV. Two Tomatoes dataset trees estimated from F_test subsets were also rejected by the AU test (Supplementary Table 2).

2.4.4 Filtering had little impact on branch properties

I investigated the impact of filtering putatively recombinant genes on branch properties for congruent branches (i.e., by taking branches present in both the Unfiltered and P_test trees, and comparing the distribution of branch properties) and for conflicting branches (i.e., by comparing the properties of branches that appeared only in the P_test tree with those of branches that appeared only in the Unfiltered tree). Branch lengths for all four datasets were generally not impacted by choice of recombination test, and conflicting branches were shorter than congruent branches (Figure 14, Supplementary Figure 16).

Similarly, filtering did not generally impact distributions of branch support (Lpp) for either conflicting or congruent branches in ASTRAL trees (Figure 15). Lpp values were high for congruent branches across all datasets and tests for recombination (Figure 15), although some outlier branches had Lpp values as low as 0.25 (Metazoan dataset) or 0.03 (Plants dataset). For the Metazoan subsets P_MaxChi and P_PHI, conflicting branches in the $ASTRAL_{P\_test}$ trees had higher Lpp than conflicting branches in the $ASTRAL_{Unfiltered}$ tree. Examining distributions of branch support (UFB) in CONCAT trees (Supplementary Figure 17), I found that congruent branches had high UFB values and that there was minimal difference in the distributions of congruent branches from $CONCAT_{P\_test}$ and $CONCAT_{Unfiltered}$. Test of recombination impacted the distribution of UFB values for the Metazoan dataset (Supplementary Figure 17). Conflicting branches in the tree $CONCAT_{P\_MaxChi}$ were lower than

those in $CONCAT_{Unfiltered}$, whereas conversely conflicting branches in the tree $CONCAT_{P\_PHI}$ were higher than those in $CONCAT_{Unfiltered}$.

I calculated qCFs for ASTRAL trees (Figure 16) and found that qCF values were similar for filtered and unfiltered trees. In general, qCF values were not impacted by choice of recombination test and qCF values for conflicting branches were lower than those for congruent branches. Test for recombination did impact qCF values for conflicting branches in the Metazoan dataset (Figure 16). The recombination tests MaxChi and PHI had slightly higher qCFs in the tree estimated from the subset of loci that passed the test ($ASTRAL_{P\_test}$) compared to the Unfiltered tree ($ASTRAL_{Unfiltered}$), whereas the results for GENECONV had larger range in qCF values for conflicting branches.

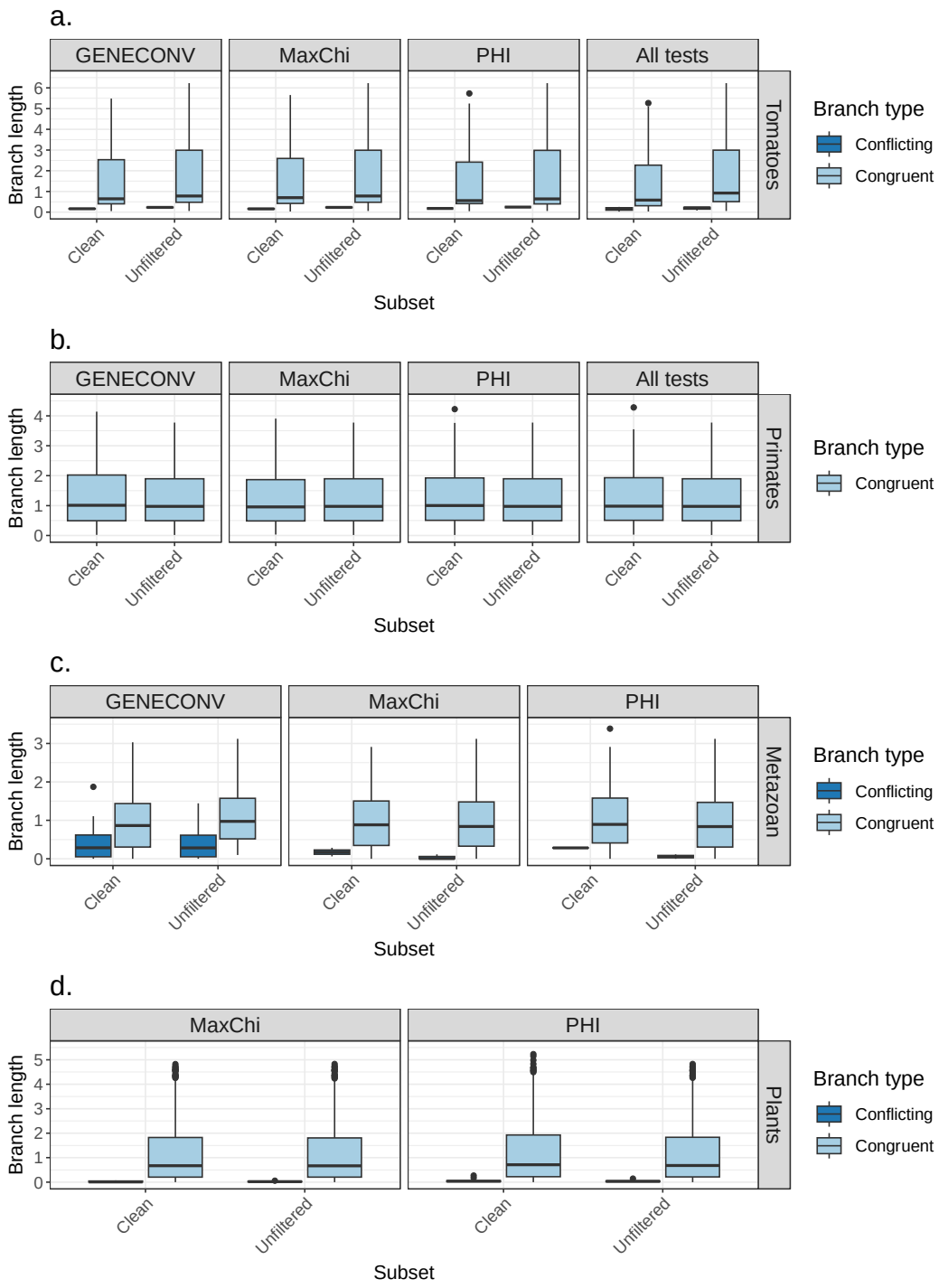


Figure 14: Branch lengths for congruent and conflicting branches in ASTRAL trees for all four datasets.

Dataset names are listed on the right of each panel. Each P_test subset tree (denoted "Clean") was compared to the tree estimated from the unfiltered dataset ("Unfiltered"). Congruent branches were branches present in both the unfiltered tree and the subset tree. Conflicting branches were present in only the unfiltered tree or the subset tree. Where trees were identical to the tree estimated from the unfiltered dataset, no conflicting branches exist and no boxplot for conflicting branches is present.

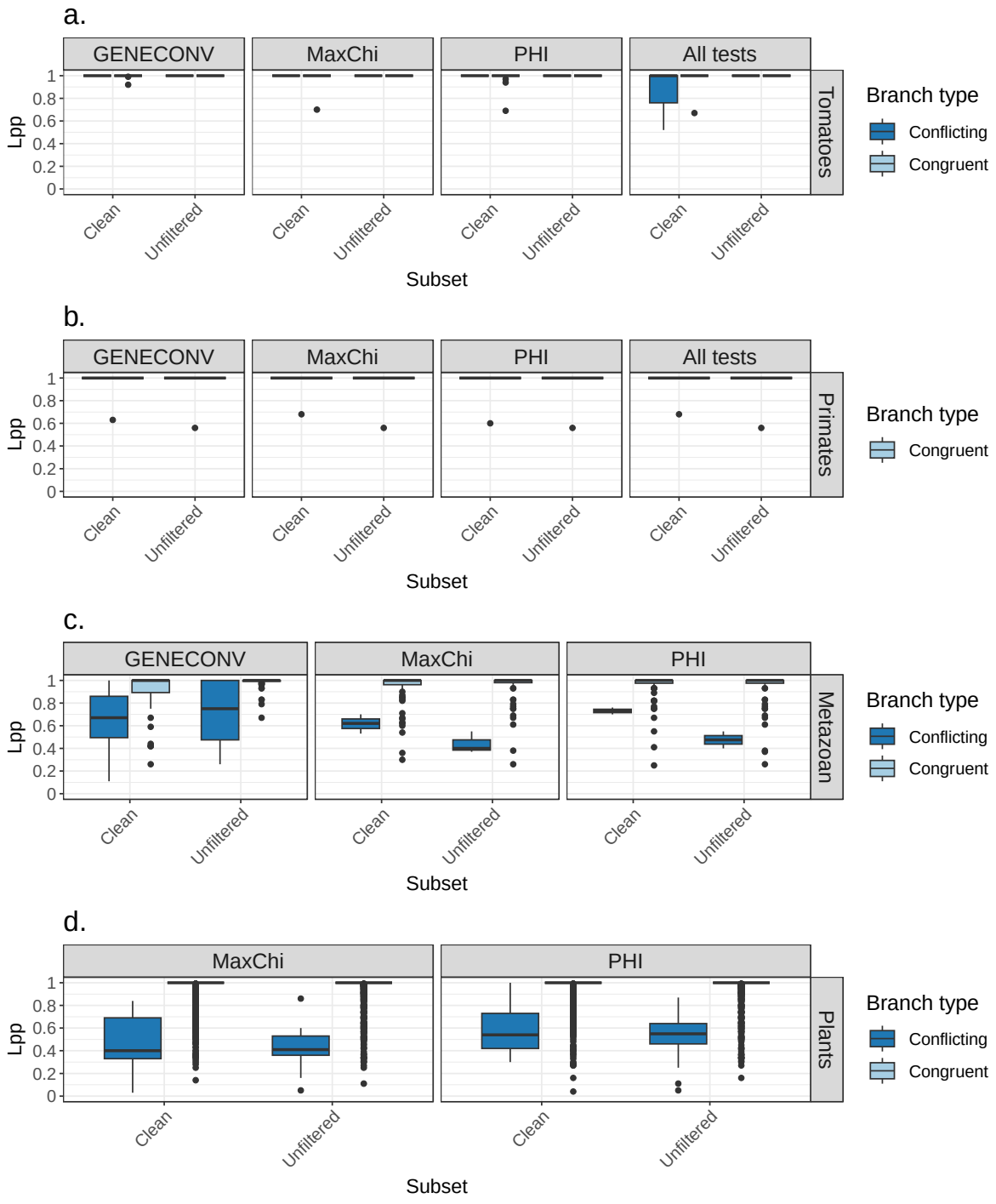


Figure 15: The local posterior probabilities (Lpp) for congruent and conflicting branches in ASTRAL trees estimated from all four datasets.

Dataset names are listed on the right of each panel. Each P_test subset tree (denoted "Clean") was compared to the tree estimated from the unfiltered dataset ("Unfiltered"). Congruent branches were branches present in both the unfiltered tree and the subset tree. Conflicting branches were present in only the unfiltered tree or the subset tree. Where trees were identical to the tree estimated from the unfiltered dataset, no conflicting branches exist and no boxplot for conflicting branches is present.

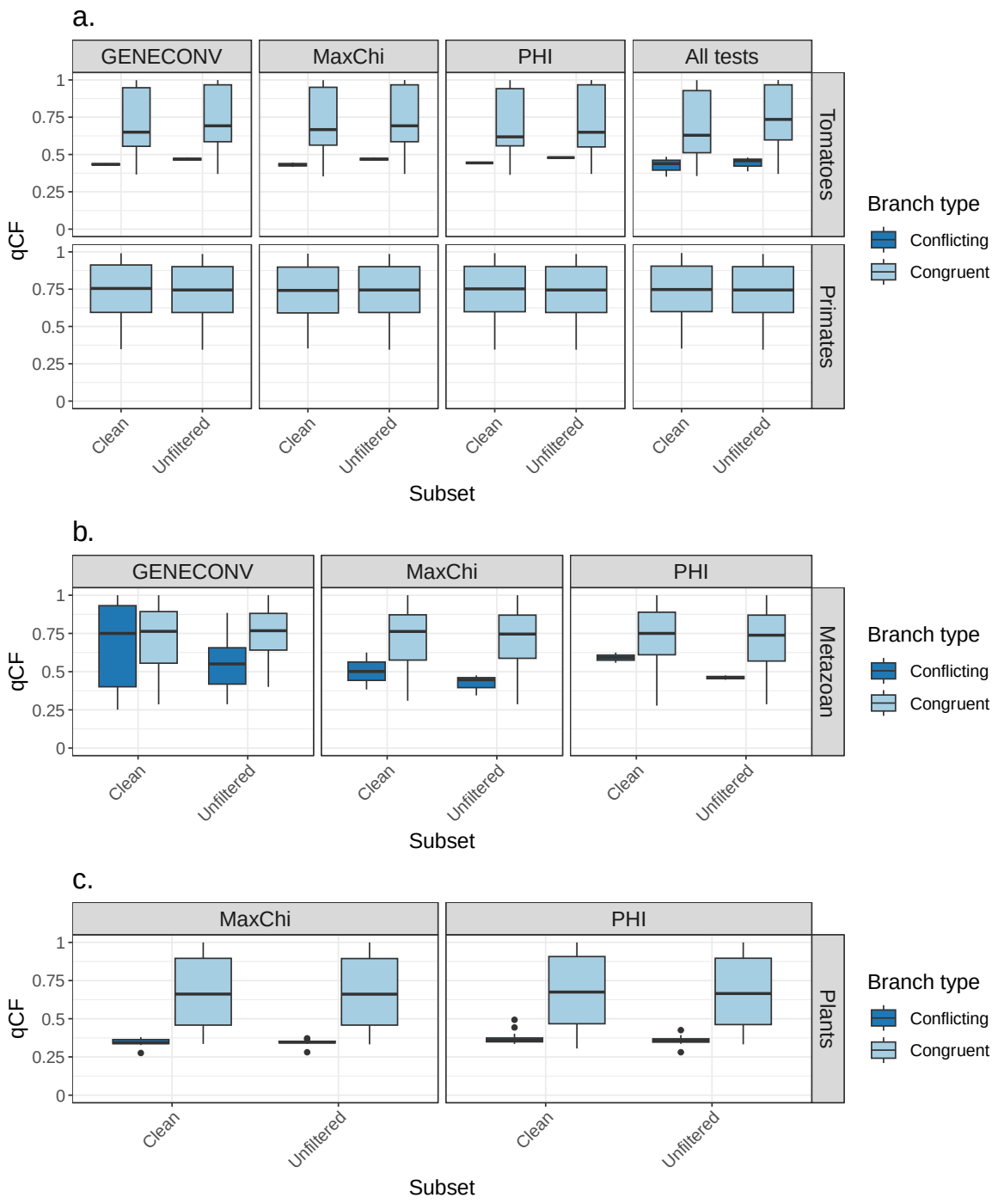


Figure 16: Quartet concordance factors for congruent and conflicting branches in ASTRAL trees for all four datasets.

Dataset names are listed on the right of each panel. Each P_test subset tree (denoted "Clean") was compared to the tree estimated from the unfiltered dataset ("Unfiltered"). Congruent branches were branches present in both the unfiltered tree and the subset tree. Conflicting branches were present in only the unfiltered tree or the subset tree. Where trees were identical to the tree estimated from the unfiltered dataset, no conflicting branches exist and so no boxplot for conflicting branches is present.

2.5 Discussion

In this chapter, I investigated the impact of removing putatively recombinant loci on species tree estimation for four empirical phylogenetic datasets. Across three different recombination detection methods, three out of four datasets showed larger differences in tree topology for summary trees inferred with ASTRAL than for concatenated trees estimated in IQ-TREE. Although filtering putatively recombinant loci tended to result in a relatively small number of topological differences, there was little difference in the branch length, branch support, or qCF value. I accounted for the number of taxa using the normalised RF distance and found the Plants dataset had the smallest proportion of branches affected by filtering (Supplementary Table 1, Supplementary Table 2). The Tomatoes dataset was most impacted by tree inference method, with higher nRF for ASTRAL trees (mean nRF = 0.0363) than for CONCAT trees (mean nRF = 0.0045) (Supplementary Table 1, Supplementary Table 2).

Loci in each dataset were excluded from analysis when applying a test for recombination was unsuccessful. After applying a test for recombination to each locus, I categorised the result as Pass (i.e., the locus was not identified as putatively recombinant, Fail (i.e., the locus was identified as putatively recombinant), or NA (i.e., the test was unable to successfully run and therefore could not categorise the locus). The MaxChi test had a low proportion of unsuccessful runs (≤1% of loci in each dataset). The shallow datasets analysed in this chapter, the Primate and Tomato datasets, had a low proportion of unsuccessful loci of <0.1% for all three tests. However, applying the PHI and GENECONV tests to the deep datasets resulted in much higher proportions of loci unable to be categorised. For the PHI test, 15.3% of Metazoan loci and 10.5% of Plants loci failed to run successfully. For GENECONV, 15.3% of Metazoan loci and 93.1% of Plants loci failed to run successfully. I suspect this is due to the low ratio of sites to taxa for these two datasets. When testing PHI, Bruen et al. (2006) considered both simulated (5 – 50 taxa, 1000 sites per simulation replicate) and empirical datasets (8 datasets with 8 – 45 taxa and alignments 444 – 2553 sites long). In his paper outlining methods for detecting gene conversion, Sawyer (1989) analysed three alignments of a single locus with between 7 and 13 taxa. The datasets analysed by Sawyer (1989) and Bruen et al. (2006) are much more comparable to the shallow datasets included in this chapter (both Tomatoes and Primates datasets included 29 taxa) than the deep datasets. The Metazoan dataset has 76 taxa and 111 loci, with an average loci length of 430.59 amino acid residues. The Plants dataset has 1178 species and 391 loci, with an average locus length of 380.99 amino acid residues. These loci seem too short to estimate a gene tree for those numbers of taxa, meaning there may simply be insufficient phylogenetic signal to infer patterns of recombination. All loci unable to be classified as putatively recombinant or not putatively

recombinant were excluded from tree estimation. As the Metazoan and Plants datasets had low numbers of loci compared to the Primates and Tomatoes datasets, the failure of these tests had a proportionally larger impact on tree estimation and comparison.

Empirical phylogenetic analyses often pay more attention to branches with high support values (Simon 2022; Thomson and Brown 2022). My analyses revealed that filtering putatively recombinant loci frequently resulted in highly supported differences among tree topologies. For each method of tree estimation (ASTRAL and IQ-TREE), I performed 13 comparisons between an Unfiltered tree and a tree estimated from a P_test subset (4 for both Tomatoes and Primates, 3 for Metazoans, 2 for Plants). For concatenated trees, 2/13 comparisons had a highly supported difference: the Metazoan subsets P_GENECONV and P_PHI. There were more highly supported differences for ASTRAL trees (6/13): Tomato subsets P_GENECONV, P_MaxCHI, P_PHI, and P_All; Metazoan subset P_GENECONV; and Plant subset P_PHI. In all but 1 case (comparing Metazoan CONCATP_GENECONV to CONCATUnfiltered, which I discuss below), filtering to remove putatively recombinant loci did not impact relationships between major clades (i.e., did not result in taxonomic reclassification of any taxa or change relationships between well-established clades). The consequences of my chapter therefore depend on the purpose of phylogenetic inference. For all datasets except the Metazoans, the consistency in relationship between clades under both concatenation and summary methods is reassuring for studies conducting downstream analyses. However, the majority of highly supported differences occurred within two particular clades (Ctenophora in the Metazoan dataset: 24 highly supported differences, Peruvianum in the Tomato dataset: 15 highly supported differences) (Supplementary Table 3). For studies interested in the resolution of these clades, this heterogeneity in phylogenetic signal is concerning. In these cases, I recommend thoroughly exploring phylogenetic signal within the data (for example, applying tests for recombination as I did here) and examining previous studies of the same system to look for evidence of biological processes and topological variation. The results of this data exploration can be used to determine appropriate phylogenetic models and tree inference methods (Brown and Thomson 2018; Jermiin et al. 2020). Finally, if filtering loci is performed, I recommend presenting both trees estimated from the filtered and unfiltered data. These recommendations are consistent with the new phylogenetic protocol proposed by Jermiin et al. (2020), which focuses on performing reproducible and transparent phylogenetics.

Filtering putatively recombinant loci had no impact on the topology of the Primate tree. Tree topology was dependant on tree inference method. All trees estimated using the summary method had identical topology (Figure 10, Supplementary Figure 3), and all trees estimated using the concatenated method had identical topology (Supplementary Figure 7,

Supplementary Figure 8). I found only one branch difference between the concatenated and summary trees estimated from the Primates dataset. This difference occurs within the Cebidae clade of the New World Monkeys, represented in the Primates dataset by 4 taxa (Supplementary Figure 18). Vanderpool et al. (2020b) also observed the same difference in topology, finding the concatenated ML trees had a balanced 4-taxa tree for this clade, compared to the ASTRAL trees which resolved the clade as a fully unbalanced tree (also known as a caterpillar tree). There is substantial conflicting signal within this clade (Schrago and Seuánez 2019; Vanderpool et al. 2020b), with similar proportions of gene trees and parsimony informative sites supporting each of the three possible topologies (Vanderpool et al. 2020b), resulting in bias during tree reconstruction (Kubatko and Degnan 2007; Bryant and Hahn 2020).

Aside from this clade, all primate trees were identical regardless of the test of recombination or the tree inference method, and the topology of the Primates was not impacted by removing putatively recombinant loci. The Primates dataset I analysed was constructed by identifying single-copy orthologs from primate and non-primate genomes, with an average ortholog length of 1018 base pairs (bp), and an average of 178 parsimony informative sites (Vanderpool et al. 2020a, 2020b). These sequences are similar in length to sequences used in previous studies to identify substantial intra-genic conflict (Scornavacca and Galtier 2017; Mendes et al. 2019; Smith et al. 2020). Given that for primates the mean length of an average non-recombining segment without incomplete lineage sorting is less than 100 bp (Hobolth et al. 2011), loci within the Primates dataset should be long enough to include conflicting phylogenetic signal. However, both my results and Smith et al. (2022) found that almost identical trees were inferred regardless of data filtering or tree estimation method, although the latter study focused on inclusion/exclusion of paralogs and gene families rather than intra-loci recombination as I did here. In both my chapter and the original Primates study (Vanderpool et al. 2020b), branch support values were high for both concatenated and summary trees (almost all local posterior probability = 1, same for UFB = 100), indicating low sampling variation (Thomson and Brown 2022). Despite the known presence of introgression and ILS within the Primates dataset (Tung and Barreiro 2017; Schrago and Seuánez 2019; Vanderpool et al. 2020b; Smith et al. 2022), filtering putatively recombinant loci does not impact tree inference under either concatenated or summary tree methods. I hypothesise that there is sufficient phylogenetic signal in this dataset to infer a backbone that is robust to the stochastic error introduced by different sampling schemes.

Comparing subset trees to the Unfiltered Tomatoes dataset tree, I found filtering impacted trees inferred using summary methods more than trees inferred using concatenation methods.

Concatenated trees estimated from subsets of the Tomatoes dataset were either identical (P_GENECONV, P_MaxChi, P_PHI) or one branch away (P_All) to the tree estimated from the Unfiltered dataset. However, inferring summary trees from the Tomatoes dataset subsets resulted in at least 1 conflicting branch and at least 1 highly supported difference for each tree estimated from a Clean subset (P_GENECONV, P_MaxChi, P_PHI, P_All). Choice of test of recombination impacted tree topology: the trees $ASTRAL_{P\_GENECONV}$ and $ASTRAL_{P\_MaxChi}$ were identical, whereas $ASTRAL_{P\_PHI}$ and $ASTRAL_{P\_All}$ had distinct unique topologies. Of the 16 conflicting branches I observed when comparing Clean and Unfiltered ASTRAL trees, 15 were highly supported difference with local posterior probability of 1. All conflicting branches for the Tomato dataset occurred in the Peruvianum clade, which has previously been difficult to resolve due to incomplete lineage sorting and a history of gene flow and introgression (Peralta et al. 2005; Rodriguez et al. 2009; Labate et al. 2014; Pease et al. 2016a). In particular, I note that choice of recombination test impacted the placement of taxa *Solanum peruvianum 2744* and *Solanum huaylasense 1364*, which have undergone recent hybridisation (Labate et al. 2014; Pease et al. 2016a) of the 10 unique conflicting branches, 9 related to placement of the taxa *S. peruvianum 2744* and *S. huaylasense 1364*. I investigated this further by plotting all gene trees as a cloudogram and observed substantial conflict within gene tree topologies across all clades of the Tomatoes (Supplementary Figure 19). The Tomatoes dataset was generated by mapping RNA-Seq reads to the *Solanum lycopersicum 'Heinz' 1706* reference genome and processing mapped reads into 2,745 non-overlapping genomic windows of 100 kbp (Pease et al. 2016a). The rapid radiation of Tomato clades resulted in ILS shown by the high discordance within short branches of the tree (Pease et al. 2016a). Given that ASTRAL is explicitly designed to accommodate ILS (Mirarab et al. 2014; Zhang et al. 2018b), the difference in ASTRAL tree topology should stem from the differences in the recombination tests in the putatively recombinant loci that they identify. The consistency of impacted taxa (*S. peruvianum 2744* and *S. huaylasense 1364*) across the filtered subsets, and the identical topology of trees $ASTRAL_{P\_GENECONV}$ and $ASTRAL_{P\_MaxChi}$, suggests that applying the tests for recombination to the Tomatoes dataset has revealed genuine biological signal that was masked by inclusion of recombinant loci in the $ASTRAL_{Unfiltered}$ tree.

The topology of the metazoan tree is contentious, with support for a number of different phylogenetic hypotheses including alternate placements of Placozoa (Schierwater et al. 2009, 2021; Nosenko et al. 2013a; Whelan et al. 2015a). Previous studies have found substantial conflicting signal within metazoan datasets (Nosenko et al. 2013a; Pandey and Braun 2020). In particular, Shen et al. (2017) found support for three alternate, conflicting hypotheses for the metazoan tree (Ctenophora as sister to all other animals i.e., Ctenophora sister; Porifera-

sister; and a monophyletic clade of Ctenophora+Porifera as sister). I estimated both summary and concatenated trees from different Clean subsets of the metazoan datasets and found that the relationships between well-established clades were stable for 5/6 trees, with all differences in these trees impacting relationships between Ctenophora species. Only one tree, $CONCAT_{P\_GENECONV}$, had an alternate relationship between clades due to the movement of *Trichoplax adhaerens*, the single representative of the Placozoa (Supplementary Figure 14). Previous studies have found that data filtering can result in different relationships between metazoan clades. Francis and Canfield (2020) were able to produce a highly-supported tree by removing 1.7% of sites that strongly favoured alternate topologies. Alternatively, McCarthy et al. (2023) took 5 alignments that had previously been used to estimate the metazoan phylogeny, rejected the 25–52% of loci that successfully recovered 3 or more monophyletic clades, and observed a different tree topology was inferred for 2/5 datasets (McCarthy et al. 2023). Despite finding similar proportions of loci rejected by the tests for recombination to those of McCarthy et al. (2023), I found that the topological differences were limited to the Ctenophora clade and did not impact relationships between the major clades.

Regardless of tree inference method all subset trees estimated from the Plants dataset were different to the corresponding Unfiltered tree, and concatenated (CONCAT) trees had a higher number of conflicting branches than summary (ASTRAL) trees. The Plants dataset consists of transcriptomes (Leebens-Mack et al. 2019a), meaning that loci are likely to include multiple conflicting evolutionary histories. The Plants dataset had the lowest ratio of loci to taxa (391 loci to 1178 taxa), and the Clean subsets included 89.0% (P_PHI) and 98.98% (P_MaxChi) of those loci. Examining the ASTRAL trees, I identified a single highly supported difference (Lpp > 0.9) from the 2353 branches ($2 \times 1178 - 3$) of the Plants dataset. This highly supported difference occurred in the Malpighiales clade of the $ASTRAL_{P\_PHI}$ tree (Supplementary Figure 15). Evolutionary relationships within the Malpighiales have been difficult to resolve due to a rapid radiation during the mid-Cretaceous period (Davis et al. 2005), leading to high levels of ILS and gene tree estimation error (Cai et al. 2021). However, even under the tree that has the highest RF distance to the Unfiltered tree ($CONCAT_{P\_PHI}$), the proportion of branches impacted is only 1.3% (31 conflicting branches/2353 total branches) with only one highly supported difference (Lpp > 0.9). My chapter corroborates Sanderson et al. (2010), who found that tree reconstruction was possible for the majority of branches in a tree given a sparse matrix, providing there was sufficient overlap among taxa. The occupancy of taxa in the final Plants matrix varied (minimum number of taxa per locus = 38, maximum = 1131, median = 937.5), and therefore the decrease in phylogenetic signal will also depend on which loci are removed. Despite the well-known challenges of sparse matrices, it is perhaps encouraging

that filtering sometimes more than 10% of putatively recombinant loci, the impacts on the tree topology of the plants dataset were relatively minor.

Each test for recombination uses a different underlying method. The PHI test measures the mean refined incompatibility score, which can be interpreted as the minimum number of homoplasies present to describe the history of any two sites (Bruen et al. 2006). The MaxChi method identifies recombination breakpoints by comparing the number of segregating sites on either side of a potential breakpoint from a pair of sequences with the number of segregating sites in all other sequences (Maynard Smith 1992). Finally, GENECONV (Sawyer 1989) looks for gene conversion events by identifying long aligned segments shared between two sequences. The proportion of loci rejected by each test and the topological impact of each test varied across datasets. The Plants dataset had the smallest proportion of loci rejected (0.26–0.51%) for the three tests for recombination, which is surprising given that this dataset is comprised of transcriptomes and covered a huge evolutionary timescale, presenting substantial opportunity for within-locus recombination. This surprising result suggests that the power of all three tests to detect recombination may be very limited in certain situations.

The Tomatoes dataset had the highest proportion (43.55–64.12%) of loci rejected. This may be due to the long mean locus length of 21k bp (over 21 times longer than the second-largest mean locus length of ~1k bp for the Primates dataset). Pease et al. (2016a) applied the D-statistic and D_{FOIL} (Pease and Hahn 2015) and identified introgression in 44% of Tomato dataset loci, similar to the proportion of Tomato loci rejected by the MaxChi test (43.55%). The Tomatoes have undergone extensive hybridisation and a rapid radiation (Labate et al. 2014; Pease et al. 2016a). The Plants dataset was assembled to estimate the phylogeny of green plants (Viridiplantae), which is estimated to include 450,000–500,000 species (Corlett 2016; Lughadha et al. 2016). Leebens-Mack et al. (2019a) did note some gene tree discordance within their dataset which they attribute to causes including rapid diversification, reticulate evolution, and gene duplication and loss. However, they found generally congruent results across multiple analyses (Leebens-Mack et al. 2019a). These suggested causes of gene tree discordance would not be detected by the tests for recombination applied here, which each identify specific signals of recombination.

Previous studies have investigated the impact of recombination on tree accuracy using simulations. Lanier and Knowles (2012) investigated the effect of including recombinant loci in Bayesian coalescent species tree analyses. Including recombinant loci resulted in a weak negative effect on the accuracy of species tree estimation, but the increase in accuracy due to adding more loci always outweighed any decrease from adding recombinant loci (Lanier and

Knowles 2012). These simulations were criticised for underestimating recombination rates and missing simulations for deep phylogenetic relationships (Gatesy and Springer 2014; Springer and Gatesy 2018). Using empirical datasets from different clades of different depths allowed me to ensure that recombination rates and other properties of the alignment were reflective of real biological processes. By applying widely-used tests for recombination to published empirical datasets, my chapter allows me to determine the impact of recombination on tree topology in the context of common phylogenetic pipelines. One limitation of empirical studies is that the true tree is unknown, meaning the extent of variation from the true tree cannot be quantified. This is true of all empirical studies comparing trees estimated using different models or different subsets of data. To manage this uncertainty, trees can be applied using approaches such comparing the likelihood values of the different trees (de Oliveira Martins et al. 2015) or applying tests for tree adequacy such as the AU test (Shimodaira 2002). In this chapter, I applied the AU test (Shimodaira 2002) and the goodness-of-fit test (Ané 2021; Cai and Ané 2021) to compare the set of trees inferred from different subsets of each alignment. In addition, I compared tree topologies using the RF distance, weighted RF distance and normalised RF distance (Robinson and Foulds 1981; Steel and Penny 1993). This allowed me to quantify the variation in tree topologies among trees inferred from different subsets of each dataset. While the true tree is unknown, for the datasets analysed in this chapter topological differences in the tree inferred after removing putatively-recombinant loci were generally localised to a particular clade associated with historical introgression or hybridisation events. The impact of these topology differences depends on the purpose for tree inference. For example, for the four datasets analysed in my chapter, the relationships between clades were relatively stable between the different trees. However, if the particular relationship between certain species is important for downstream analyses, these differences in topology will have a larger impact and will necessitate further investigation.

The datasets in this chapter were selected specifically due to the potential for within-locus recombination, as each locus was made up of large genomic regions or concatenated exons. Each test for recombination applied in this chapter aimed to determine whether an individual loci was a coalescence gene (c-gene), that is, a segment of the genome where each site has an identical evolutionary history (which means there has been no recombination). When analysing protein-coding genes, the number of base pairs in the gene is much smaller than the genomic region over which the exons span. Springer and Gatesy (2016) analysed a mammal dataset created by Song et al. (2012) and found the mean gene length in the genome (start codon to stop codon) was 139.6 kb, dramatically larger than the reported length of 3.1 kb. In addition, Springer and Gatesy (2016) note that exons in some genes were separated by

more than 100,0000 base pairs, and as such it is unreasonable to assume that exons within a gene share an identical evolutionary history. Analysing such loci under a coalescent model has been termed “concatalescence” (Gatesy and Springer 2013), and as each locus cannot be a single c-gene (due to the large region of the genome each locus spans) this violates the assumptions underlying the coalescent model. In general, concatalescence arises due to the compromise between maximising phylogenetic signal (which is generally increased when there are more sites in an alignment) and minimising noise and error (by attempting to capture c-genes) (Bryant and Hahn 2020). Exons are less likely to contain conflicting evolutionary histories, but may not contain sufficient phylogenetic signal to resolve gene trees with many taxa (Mendes et al. 2019; Bryant and Hahn 2020). Unfortunately, the complexity and stochasticity of biological evolutionary processes means that no phylogenetic method can truly account for the levels of heterogeneous signal within empirical datasets.

Over the timescales of the four datasets involved in this chapter, especially for the Metazoan and Plants datasets, the proportion of gene identified as putatively recombinant is surprisingly low. Previously, Springer and Gatesy (2016) investigated the impact of recombination on the mammal dataset of Song et al. (2012), which contains 447 protein-coding genes for 37 species. Using the recombination rate of primates, Springer and Gatesy showed that the expected length of a c-gene within the mammals dataset was 12 base pairs – i.e., that on average across the alignment, there is a recombination breakpoint every 12 base pairs. This occurs due to a phenomenon called “recombination ratchet” (Springer and Gatesy 2018). To understand recombination ratchet, suppose that there is a three taxon tree with topology $((A,B),C)$ and a short internal branch between species A and B, such that the branch is susceptible to ILS and not all gene trees agree with the species tree. This creates a number of breakpoints between c-genes in the genome, with different c-genes supporting different topologies. Now, imagine adding another identical clade of three taxa $((D,E),F)$ to the tree. While the recombination events in the two clades are independent, the alignment of taxa A,B,C,D,E,F contains all the recombination breakpoints from both clades. By adding three extra taxa, the number of recombination breakpoints in the alignment is doubled. Springer and Gatesy (2018) simulated this exact example, and showed that adding the clade $((D,E),F)$ halved c-gene length. As a consequence of recombination ratchet, c-gene size shrinks as the number of taxa in an alignment grows (Springer and Gatesy 2018). Comparing this prediction to the results of this chapter suggests that the three tests for recombination used within this chapter have low power. Even when applying the most conservative measure of rejecting any loci that failed one of more tests (F_{All}), there are a substantial proportion of loci classified as putatively non-recombinant. In particular, I would expect the number of genes classified as

putatively recombinant to be higher for the Metazoan and Plants datasets, which have deeper evolutionary timescales than the Tomatoes and Primate datasets. Contrary to expectations, the Plants dataset had the lowest proportion of genes identified as putatively recombinant. This may be due to a weakness of the tests to classify genes in alignments with low phylogenetic signal (1178 taxa), as the ratio of locus length to number of taxa is substantially lower for the Plants dataset than for any other dataset (0.478 bp per taxon).

One further cause for concern is accuracy of the alignment process. In the last two decades, the amount of sequencing data available has exponentially increased, resulting in massive phylogenomic datasets that cannot be manually curated (Ranwez and Chantret 2020). Due to the sheer scale of modern phylogenomic datasets and the number of processes within a standard sequence alignment pipeline, it would be surprising for an alignment to be published without any mistakes. Some studies have re-analysed published phylogenetic datasets and identified alignment errors that directly impact conclusions drawn from the original study. For example, Springer and Gatesy note multiple alignment errors in the mammal dataset of Song et al. (2012), including: 21 loci with switched taxonomic names; 8 duplicated loci; 26 loci with visible misalignments; and multiple loci with >50% missing data for individual taxa. Similarly, Philippe et al. (2011b) reanalysed the previously published metazoan datasets of Dunn et al. (2008) and Schierwater et al. (2009) and found poor orthology in both datasets, which they attributed to contamination and paralogy that was missed during the original alignment process. Other issues present in the two datasets included high proportions of missing data (55.5% of characters in the Dunn et al. (2008) dataset); incorrect amino acids due to frameshift and translation errors (4800 amino acid sites or 0.66% of the Dunn et al. (2008) alignment); and both taxon misidentification and inclusion of paralogous loci in the Schierwater et al. (2009) alignment. In both reanalysis studies (Philippe et al. 2011b; Springer and Gatesy 2016), conclusions drawn from the tree inferred from the updated datasets contradict findings from the original studies. Systematic errors including alignment error and ortholog misidentification have been previously identified to introduce incongruence and increase intragenic conflict in empirical phylogenomic datasets (Smith et al. 2020; McCarthy et al. 2023). This creates a substantial problem when attempting to uncover sources of incongruence, and particularly when attempting to determine historical reticulation events such as those occurring within the Primates and Tomatoes datasets in this chapter. To mitigate these risks, I first suggest that researchers interested in detangling reticulate biological events take extra care during the alignment process. Philippe et al. (2011b) found that the manually-assessed dataset of Schierwater et al (2009) had higher rates of errors, so ideally an automated pipeline with quality checks will be used with manual alignment checks as a supplement. Second, I suggest that

studies investigating recombination validate the alignment quality in putatively-recombinant genes, to distinguish between those with genuine biological signal of reticulate evolution and those with alignment errors.

In this chapter, I investigated the impact of removing putatively recombinant loci on tree topology and branch properties for both concatenated and summary trees, for three tests of recombination and four empirical multiple sequence alignments. Removing loci is known to negatively impact tree accuracy by increasing sampling error, especially when the sample of loci removed is biased (Sanderson et al. 2010; Molloy and Warnow 2018; Chan et al. 2020). I found that generally, when comparing a tree estimated from an unfiltered dataset to a tree estimated from a subset of the dataset with putatively recombinant loci removed, there were more topological differences in summary trees than concatenated trees. The number of differences was small, with only 0–4 conflicting branches when compared the filtered tree to the unfiltered tree. In the worst case, removing putatively recombinant loci for a Plants dataset with 1178 taxa and 391 loci, I found filtering impacted only 1.3% of branches. I recommend trees should be estimated from the unfiltered and filtered dataset when tests for recombination are applied, and that differences in trees should be considered in the context of the evolutionary history of those species. Finally, my results suggest that the individual tests for recombination used within this chapter have low power, even in the most generous case where any loci identified as putatively recombinant by one or more test is rejected.

2.6 Data Availability Statement

All scripts to replicate my analyses is available in the GitHub repository https://github.com/caitlinch/gene_filtering. All alignments, trees, quartet concordance factors and results csv files are available from the Figshare repository https://doi.org/10.6084/m9.figshare.26087437. The datasets analysed within this article are available from the original publications: for Tomatoes, the Dryad repository https://doi.org/10.5061/dryad.182dv; for Primates, the Dryad repository https://doi.org/10.5061/dryad.rfj6q577d; for Metazoans, the Figshare repository https://doi.org/10.6084/m9.figshare.4484138.v1; and for Plants, the Data Commons repository https://doi.org/10.25739/8m7t-4e85.

2.7 Acknowledgments

This work was supported by an Australian Government Research Training Scholarship (to C.C.). The authors thank James Pease, Matt Hahn, and Dan Vanderpool for their assistance with empirical phylogenetic datasets; Sha (Joe) Zhu for his advice on the Hybrid-Lambda

software; and Cécile Ané for her advice on the QuartetNetworkGoodnessFit.jl package. Thanks to Barbara Holland, Maja Adamska, James Barbetti, Frederick Jaya, Huiayan Ren, Nhan Trong Ly, Rahil Vora and Thomas Wong for comments on early versions of this manuscript.

2.8 Supplementary Tables

Supplementary Table 1: Analysis of the goodness of fit for summary (ASTRAL) species trees under different filtering methods for four different datasets.

This test was only applied to the Primates and Tomatoes dataset. $Z/\hat{\sigma}$ and p-value calculated using the goodness of fit test (Cai and Ané 2021). The test statistic $Z/\hat{\sigma}$ corrects the uncorrected Z statistic (used to determine whether there are more outliers than expected) for dependence between quartets. A statistically significant p-value indicates the tree does not adequately fit the data. All trees in the Primates and Tomatoes datasets were inadequate. The Robinson-Foulds (RF), normalised RF (nRF) and weighted RF (wRF) distance for each tree were calculated between that tree and the tree estimated from the unfiltered dataset (Unfiltered), therefore no RF/wRF distances were calculated for trees estimated from the Unfiltered subset. Only categories with 50 or more loci were included in comparison analyses. ASTRAL does not output terminal branch lengths so these were arbitrarily assigned

Dataset	Test	Subset	$Z/\hat{\sigma}$	p-value	RF dist.	nRF dist.	wRF dist.
Primates	PHI	P_PHI	13.59	2.26E-42	-	-	-
		F_PHI	39.1	0	2	0.018	15.673
		Unfiltered	13.68	6.37E-43	0	0	1.550
	MaxChi	P_MaxChi	14.26	2.00E-46	-	-	-
		F_MaxChi	20.24	2.09E-91	2	0.018	2.599
		Unfiltered	15.52	1.34E-54	0	0	0.848
	GENECONV	P_GENECONV	15.27	6.29E-53	-	-	-
		F_GENECONV	17.34	1.09E-67	2	0.018	3.992
		Unfiltered	13.41	2.48E-41	0	0	1.669
	All tests	P_All	15.24	9.94E-53	-	-	-
		F_All	12.81	7.30E-38	2	0.018	2.598
		Unfiltered	9.73	1.09E-22	0	0	1.333
Tomatoes	PHI	P_PHI	19.72	6.95E-87	-	-	-
		F_PHI	25.97	6.05E-149	4	0.036	11.224
		Unfiltered	23.12	1.56E-118	2	0.018	4.956
	MaxChi	P_MaxChi	15.29	4.57E-53	-	-	-
		F_MaxChi	27.68	6.73E-169	6	0.055	11.763
		Unfiltered	26.14	7.15E-151	4	0.036	4.873
	GENECONV	P_GENECONV	12.5	3.68E-36	-	-	-
		F_GENECONV	23.83	7.66E-126	4	0.036	7.357
		Unfiltered	19.98	4.40E-89	4	0.036	4.854
	All tests	P_All	10.37	1.62E-25	-	-	-
		F_All	18.17	4.10E-74	6	0.055	11.927
		Unfiltered	21.26	1.21E-100	6	0.055	9.248
Metazoan	PHI	P_PHI	-	-	-	-	-
		Unfiltered	-	-	4	0.013	9.386
	MaxChi	P_MaxChi	-	-	-	-	-
		Unfiltered	-	-	6	0.020	7.73
	GENECONV	P_GENECONV	-	-	-	-	-
		Unfiltered	-	-	30	0.101	26.10
Plants	PHI	P_PHI	-	-	-	-	-
		Unfiltered	-	-	42	0.009	85.18
	MaxChi	P_MaxChi	-	-	-	-	-
		Unfiltered	-	-	18	0.004	11.48

Supplementary Table 2: Analysis of the goodness of fit of each concatenated (IQ-TREE) species tree, calculated using the AU test.

Only categories with 50 or more loci were included in comparison analyses. Each test contains either two (the unfiltered tree; the P_test tree) or three (the unfiltered tree; the P_test tree; and the F_test tree) trees. Rows in bold represent trees with a statistically significant p-value (<0.05) that are rejected by the AU test. The expected likelihood weight is shown for each subset for each dataset. Δ from max log likelihood is the difference between the log likelihood for that tree and the maximal log likelihood in the set of trees being compared. The Robinson-Foulds (RF), normalised RF (nRF) and weighted RF (wRF) distance were calculated between that tree and the tree estimated from the unfiltered dataset (Unfiltered), therefore no RF/wRF distances were calculated for trees estimated from the Unfiltered subset. Only categories with 50 or more loci were included in comparison analyses.

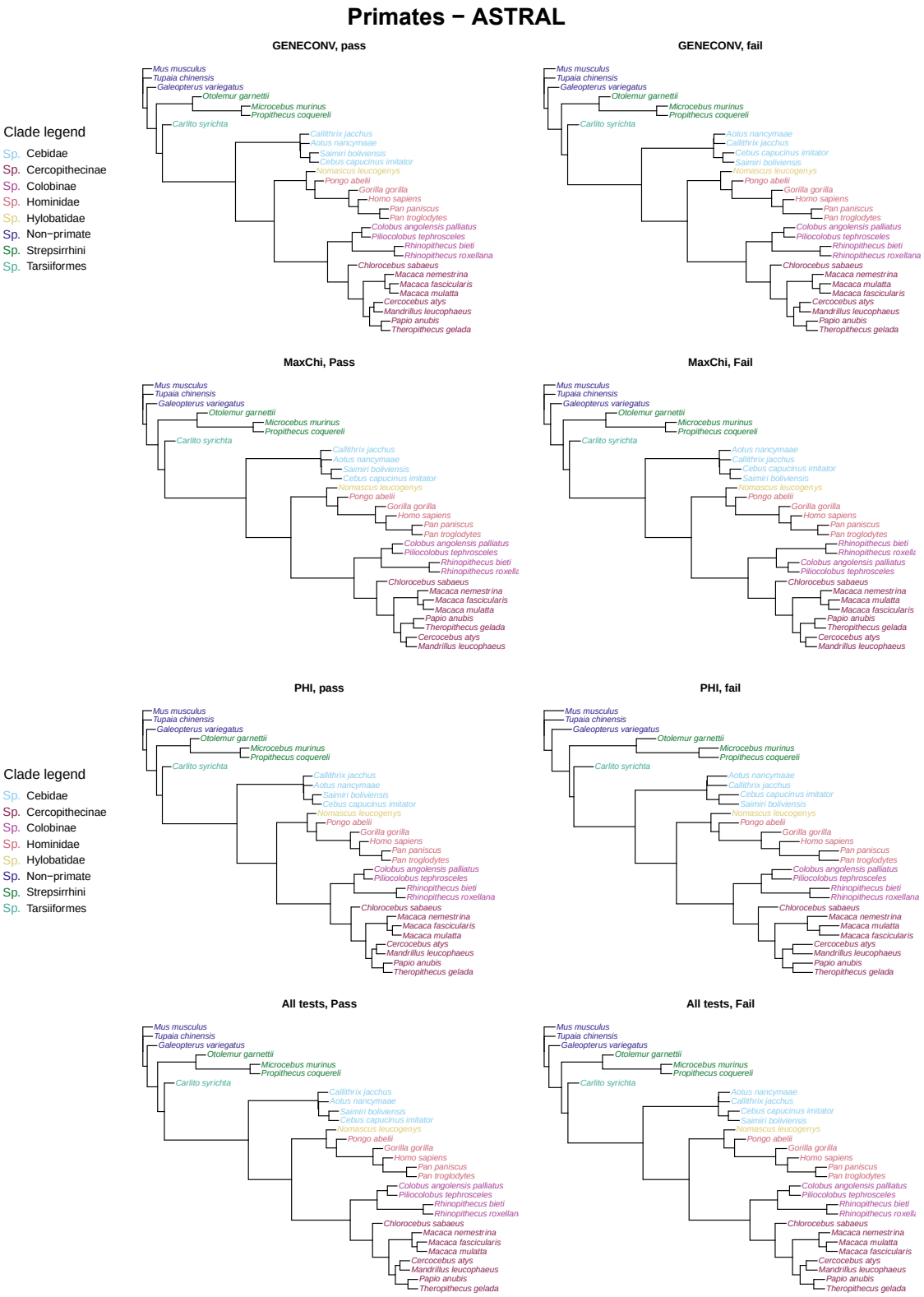
Dataset	Test	Subset	Δ from Max LogL	AU test p-val.	RF dist.	nRF dist.	wRF dist.
Primates	PHI	P_PHI	0	0.504	-	–	-
		F_PHI	**6835.9**	**0**	**4**	**0.036**	**0.354**
		Unfiltered	3.18E-04	0.496	0	0	0.037
	MaxChi	P_MaxChi	0	0.497	-	–	-
		F_MaxChi	2.20E-04	0.498	0	0	0.056
		Unfiltered	2.20E-04	0.463	0	0	0.033
	GENECONV	P_GENECONV	0	0.501	-	–	-
		F_GENECONV	7.42E-04	0.465	0	0	0.033
		Unfiltered	7.28E-04	0.491	0	0	0.028
	All tests	P_All	5.40E-05	0.493	-	–	-
		F_All	0	0.489	0	0	0.051
		Unfiltered	2.83E-04	0.476	0	0	0.041
Tomatoes	PHI	P_PHI	0.041	0.447	-	–	-
		F_PHI	**4026.1**	**0**	**6**	**0.055**	**0.011**
		Unfiltered	0	0.553	0	0	0.005
	MaxChi	P_MaxChi	0	0.518	-	–	-
		F_MaxChi	**3612.4**	**0**	**6**	**0.055**	**0.008**
		Unfiltered	0.013	0.482	0	0	0.003
	GENECONV	P_GENECONV	0	0.624	-	–	-
		F_GENECONV	277.98	0.099	4	0.036	0.005
		Unfiltered	0.033	0.593	0	0	0.003
	All tests	P_All	0	0.725	-	–	-
		F_All	52.11	0.277	2	0.018	0.014
		Unfiltered	52.11	0.277	2	0.018	0.013
Metazoan	PHI	P_PHI	0	0.616	-	–	-
		Unfiltered	1.93	0.384	2	0.007	0.51
	MaxChi	P_MaxChi	0	0.526	-	–	-
		Unfiltered	1.86	0.474	8	0.027	1.11
	GENECONV	P_GENECONV	0	0.994	-	–	-
		Unfiltered	**98.63**	**0.006**	**24**	**0.081**	**2.55**
Plants	PHI	P_PHI	0	0.695	-	–	-
		Unfiltered	169.43	0.305	62	0.013	2.89
	MaxChi	P_MaxChi	98.90	0.380	-	–	-
		Unfiltered	0	0.620	58	0.012	1.47

Supplementary Table 3: Conflicting branches with high support (local posterior probability > 0.9 or ultrafast bootstrap > 90) for all datasets and recombination tests.

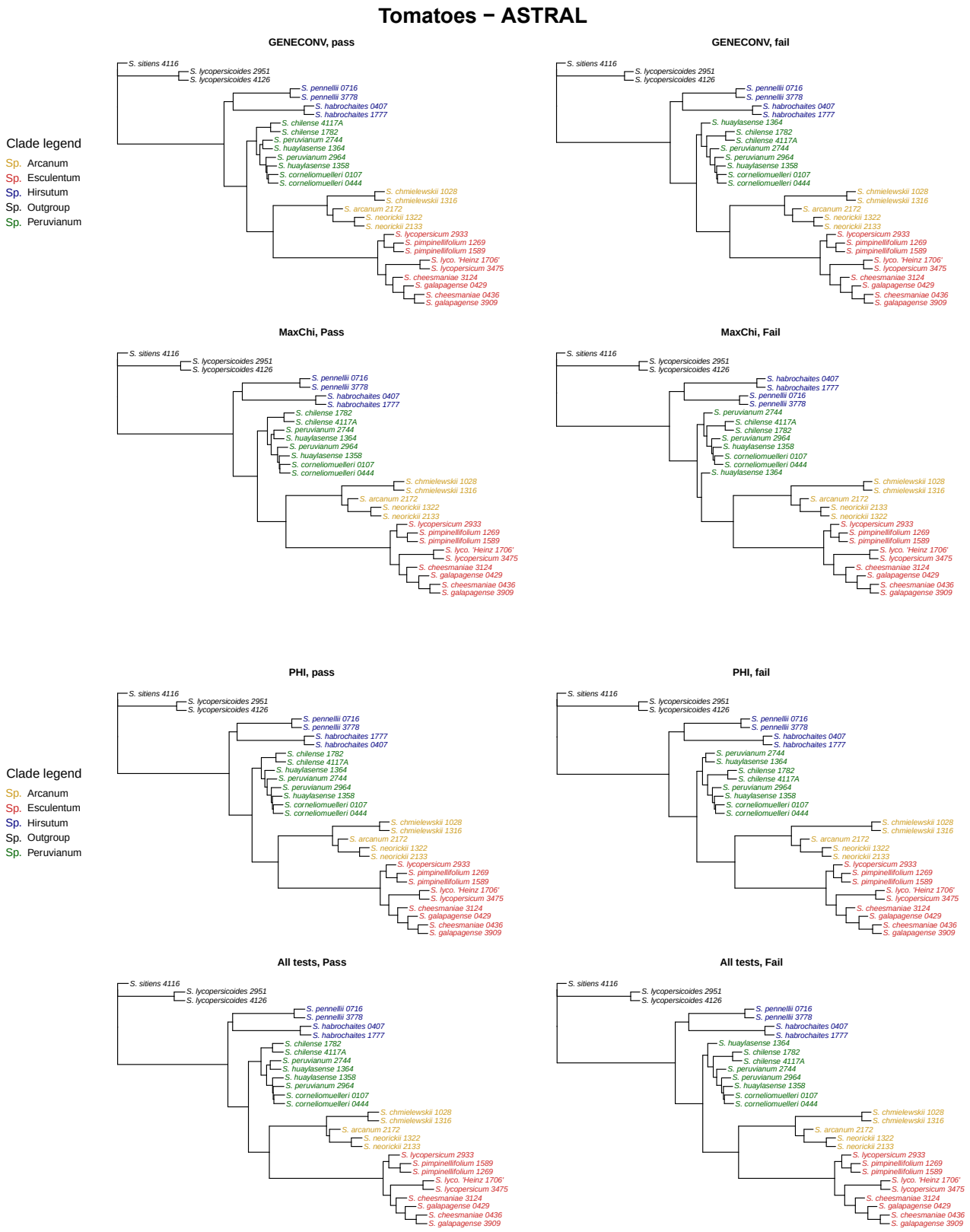
"Tree subset" indicates the tree containing the highly supported difference (i.e., the branch is present in the "Tree" and absent in the "Comparison tree"). "Branch Support" is local posterior probability (LPP) for ASTRAL trees or ultrafast bootstrap (UFB) for CONCAT trees. Branch lengths are in coalescent units for ASTRAL trees and substitutions per site for CONCAT trees.

Dataset	Tree subset	Comparison tree subset	Branch Support	Branch support value	Branch length
Tomatoes	P_PHI	Unfiltered	LPP	1	0.181
Tomatoes	Unfiltered	P_PHI	LPP	1	0.247
Tomatoes	P_MaxChi	Unfiltered	LPP	1	0.186
Tomatoes	P_MaxChi	Unfiltered	LPP	1	0.133
Tomatoes	Unfiltered	P_MaxChi	LPP	1	0.247
Tomatoes	Unfiltered	P_MaxChi	LPP	1	0.208
Tomatoes	P_GENECONV	Unfiltered	LPP	1	0.176
Tomatoes	P_GENECONV	Unfiltered	LPP	1	0.150
Tomatoes	Unfiltered	P_GENECONV	LPP	1	0.247
Tomatoes	Unfiltered	P_GENECONV	LPP	1	0.208
Tomatoes	P_All	Unfiltered	LPP	1	0.256
Tomatoes	P_All	Unfiltered	LPP	1	0.171
Tomatoes	Unfiltered	P_All	LPP	1	0.247
Tomatoes	Unfiltered	P_All	LPP	1	0.208
Tomatoes	Unfiltered	P_All	LPP	1	0.084
Metazoan	P_GENECONV	Unfiltered	LPP	1	1.107
Metazoan	P_GENECONV	Unfiltered	LPP	1	1.871
Metazoan	P_GENECONV	Unfiltered	LPP	0.92	0.666
Metazoan	Unfiltered	P_GENECONV	LPP	1	0.472
Metazoan	Unfiltered	P_GENECONV	LPP	1	1.167
Metazoan	Unfiltered	P_GENECONV	LPP	1	1.077
Metazoan	Unfiltered	P_GENECONV	LPP	1	0.755
Metazoan	Unfiltered	P_GENECONV	LPP	1	1.444
Plants	P_PHI	Unfiltered	LPP	1	0.269
Metazoan	P_PHI	Unfiltered	UFB	94	0.009
Metazoan	P_GENECONV	Unfiltered	UFB	100	0.128
Metazoan	P_GENECONV	Unfiltered	UFB	100	0.050
Metazoan	P_GENECONV	Unfiltered	UFB	99	0.048
Metazoan	P_GENECONV	Unfiltered	UFB	95	0.036
Metazoan	P_GENECONV	Unfiltered	UFB	98	0.022
Metazoan	Unfiltered	P_GENECONV	UFB	92	0.027
Metazoan	Unfiltered	P_GENECONV	UFB	100	0.062
Metazoan	Unfiltered	P_GENECONV	UFB	100	0.047
Metazoan	Unfiltered	P_GENECONV	UFB	100	0.034
Metazoan	Unfiltered	P_GENECONV	UFB	99	0.020
Metazoan	Unfiltered	P_GENECONV	UFB	93	0.015
Metazoan	Unfiltered	P_GENECONV	UFB	100	0.023
Metazoan	Unfiltered	P_GENECONV	UFB	100	0.023
Metazoan	Unfiltered	P_GENECONV	UFB	100	0.031
Metazoan	Unfiltered	P_GENECONV	UFB	93	0.015

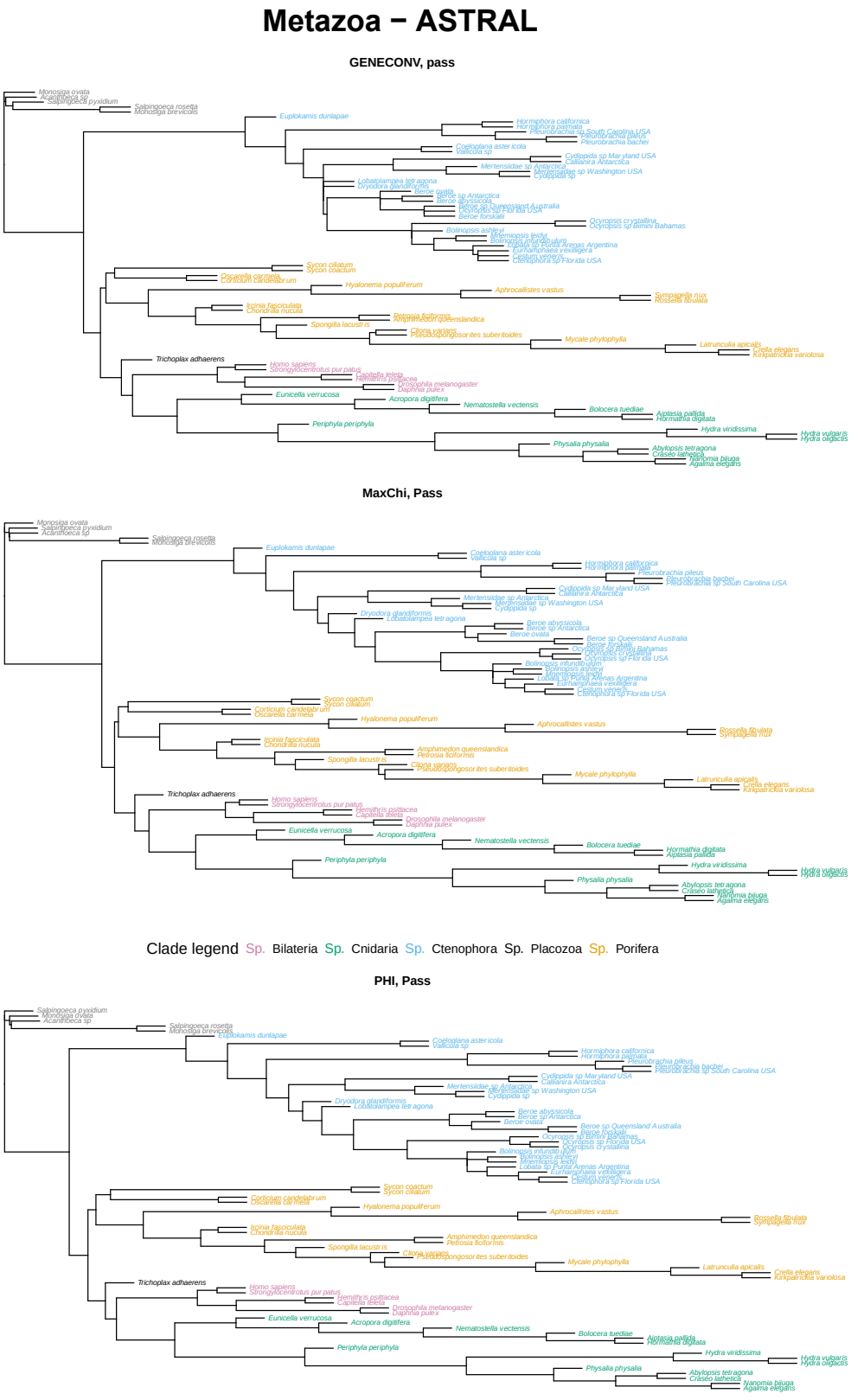
2.9 Supplementary Figures



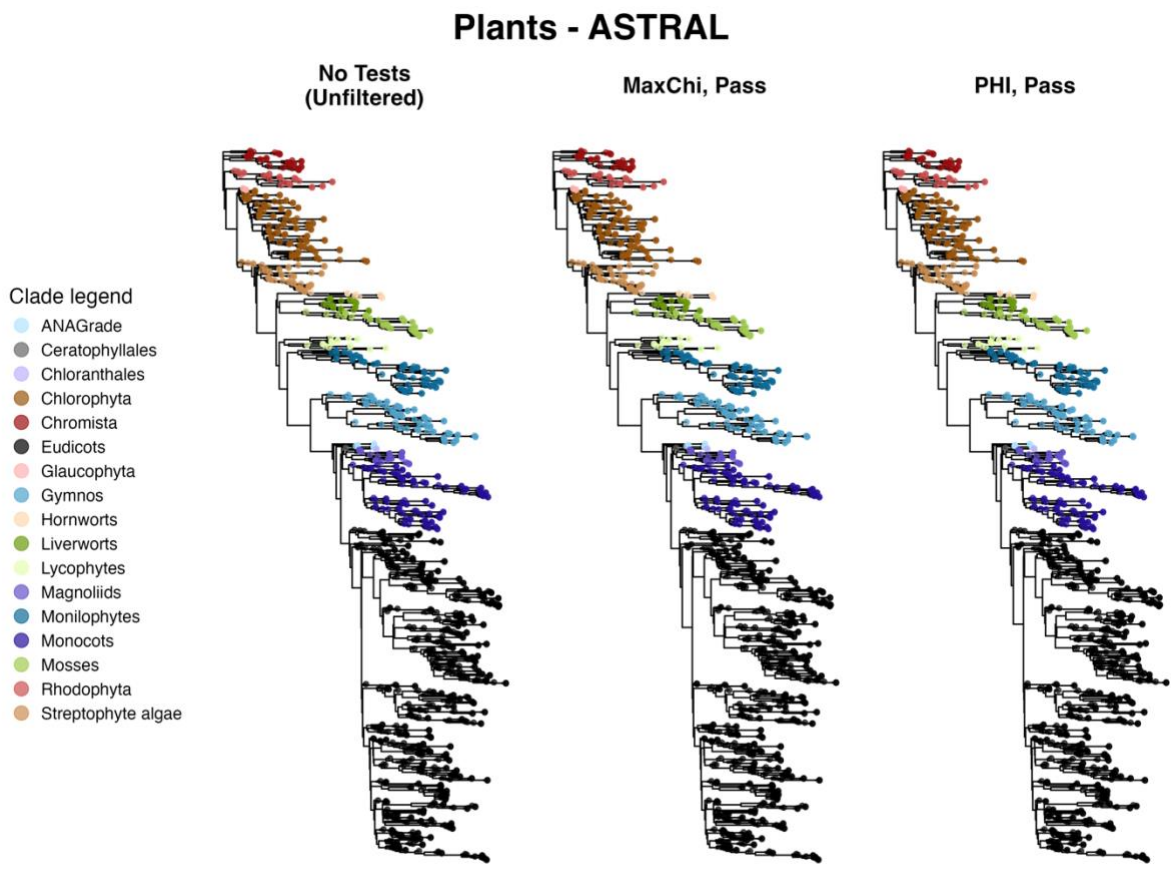
Supplementary Figure 3: ASTRAL trees estimated from the Primates dataset using the subsets P_GENECONV, F_GENECONV, P_MaxChi, F_MaxChi, P_PHI, F_PHI, P_ALL and F_ALL. Only the topology of the Cebidae clade (shown in light blue) differs between trees.



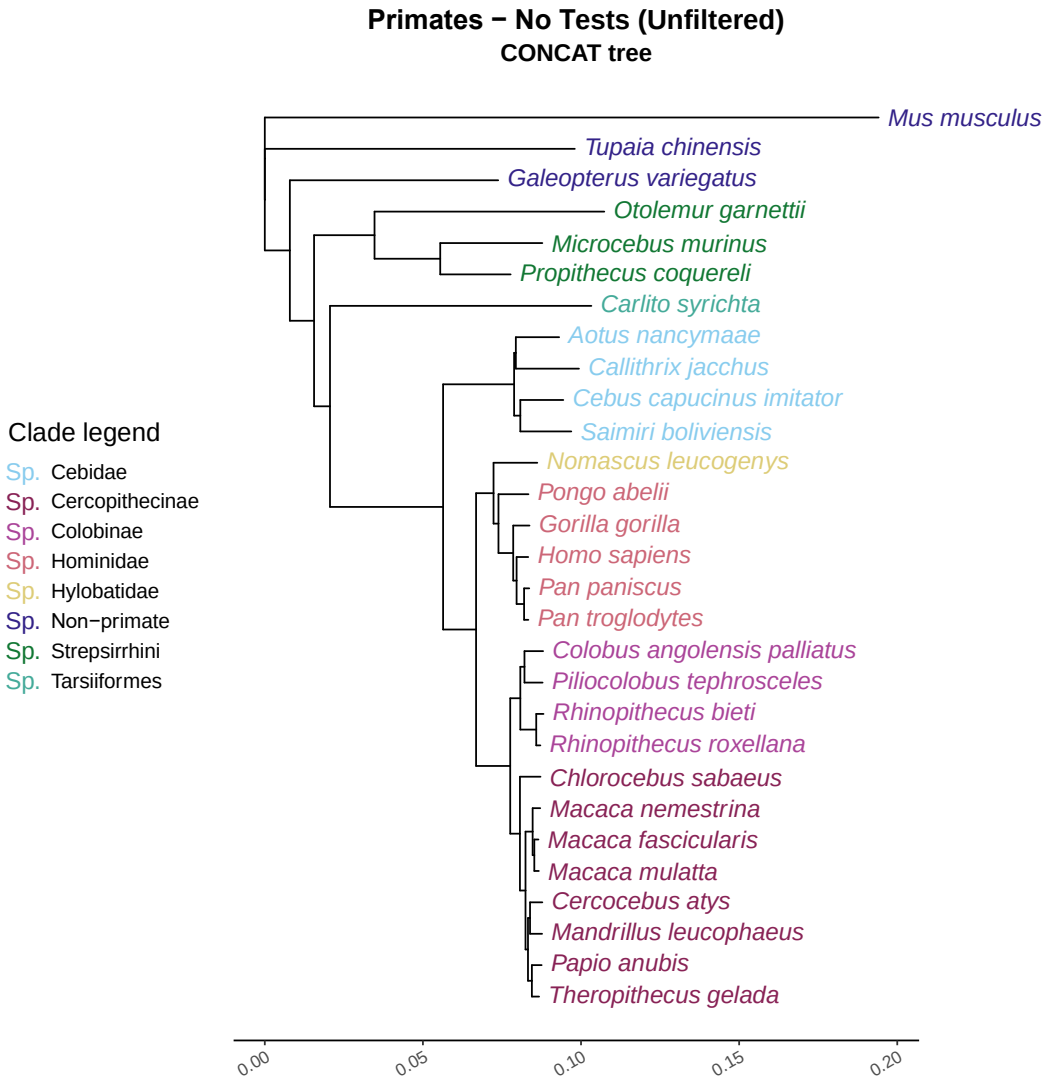
Supplementary Figure 4: ASTRAL trees estimated from the Tomatoes dataset using the subsets using the subsets P_GENECONV, F_GENECONV, P_MaxChi, F_MaxChi, P_PHI, F_PHI, P_ALL and F_ALL.
Only the topology of the Peruvianum clade (shown in green) differs between trees.



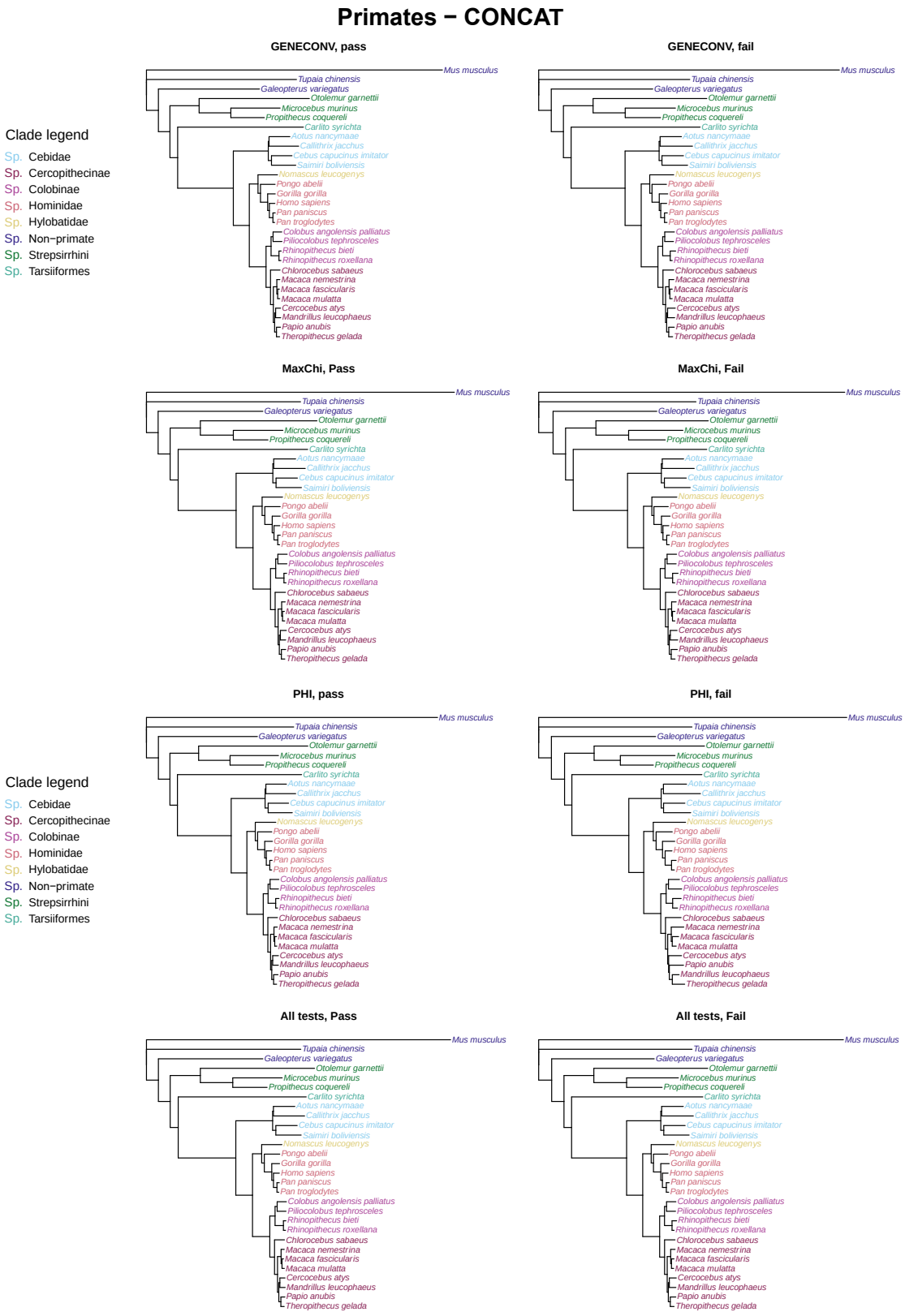
Supplementary Figure 5: ASTRAL trees estimated from the Metazoan dataset from the P_GENECONV dataset (top), P_MaxChi subset (middle) and P_PHI subset (bottom). All three trees have the same relationships between the 5 Metazoan clades (Bilateria, Cnidaria, Ctenophora, Placozoa and Porifera). The majority of differences between trees occur in the Ctenophora clade (shown in blue).



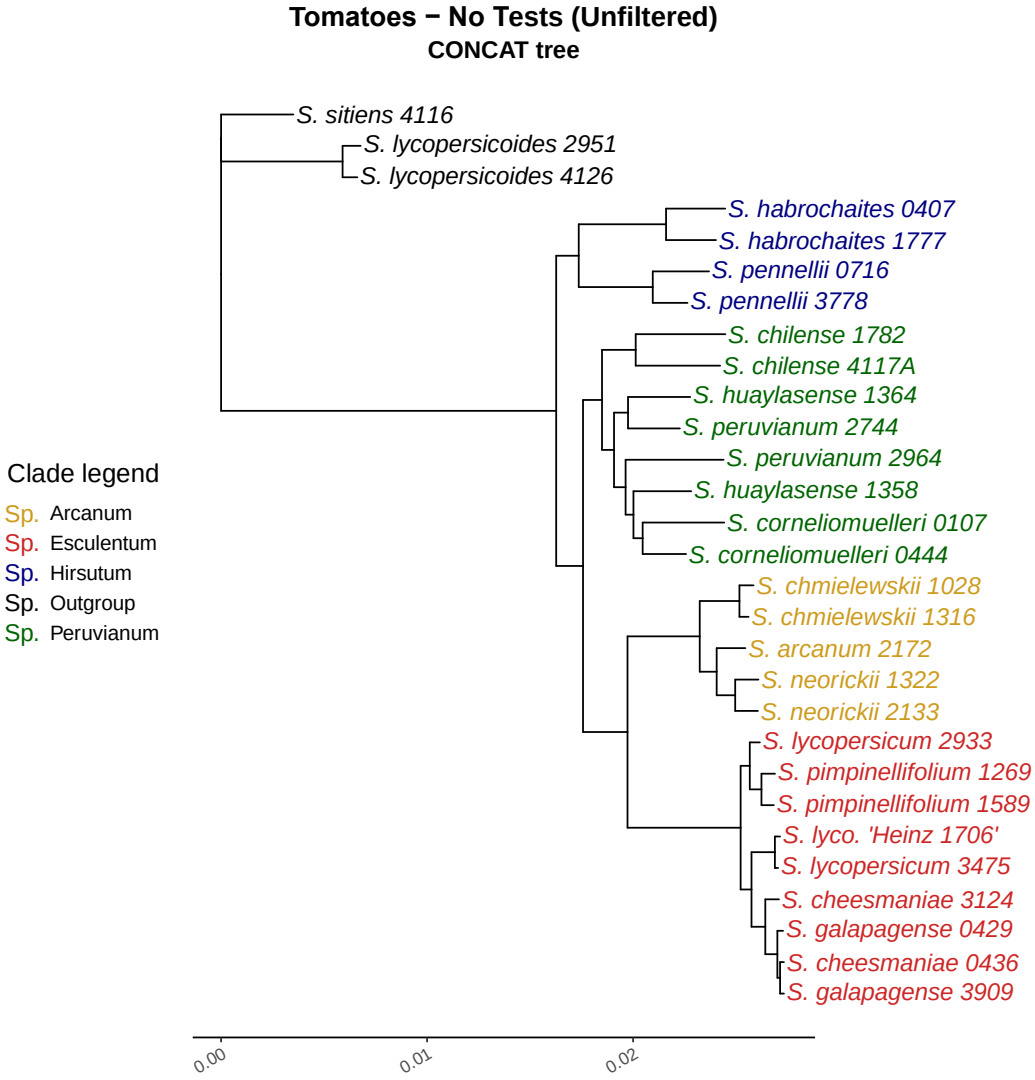
Supplementary Figure 6: ASTRAL trees estimated from the Plants dataset, from the Unfiltered dataset (left), the P_MaxChi subset (centre) and the P_PHI subset (right).
The three trees have the same relationships between clades.



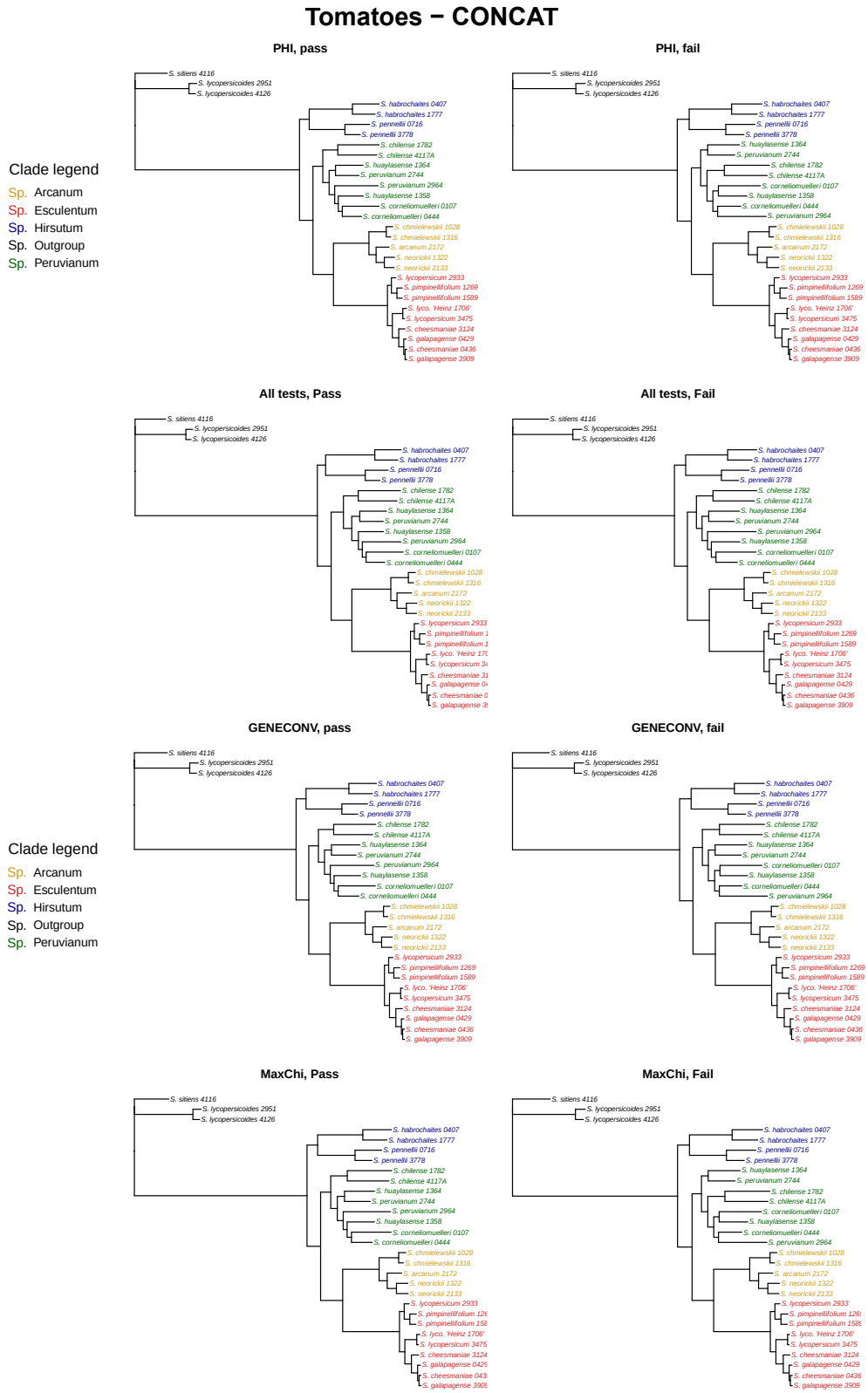
Supplementary Figure 7: Concatenated tree estimated from the unfiltered Primates dataset



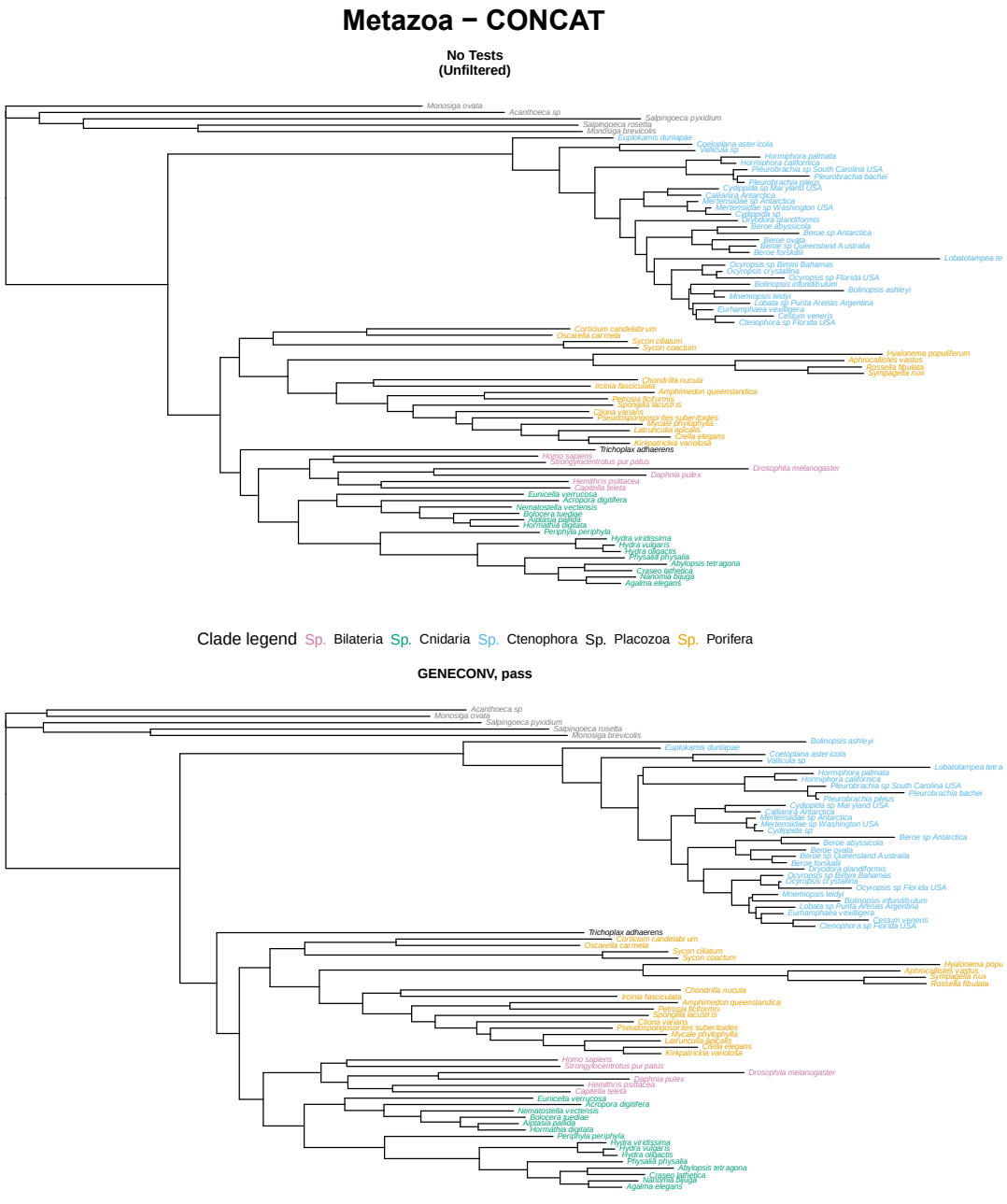
Supplementary Figure 8: Concatenated trees estimated from the Primates dataset using the subsets P_GENECONV, F_GENECONV, P_MaxChi, F_MaxChi, P_PHI, F_PHI, P_ALL and F_ALL.
Only the topology of the Cercopithecinae clade (shown at bottom of tree) differs between trees.



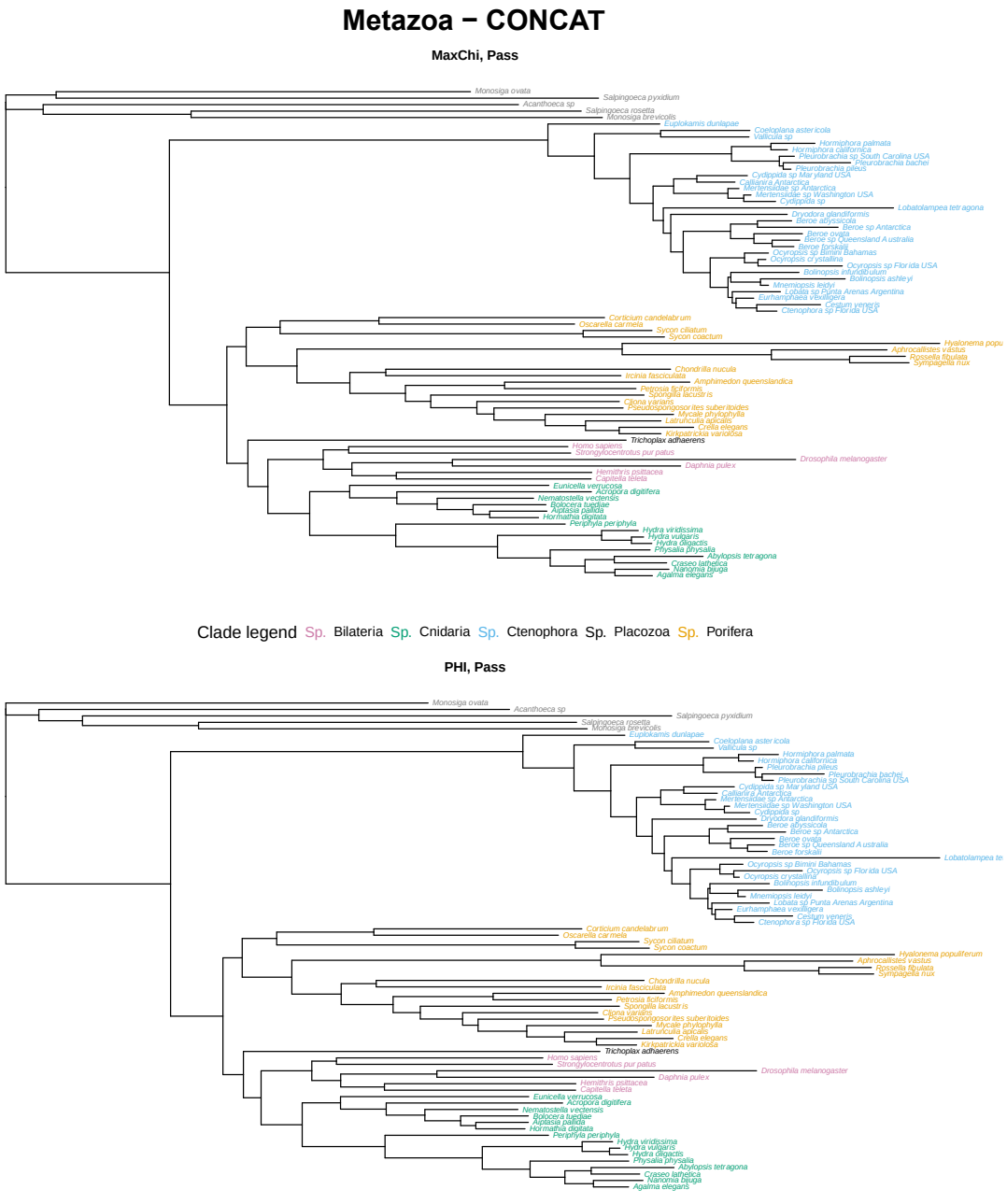
Supplementary Figure 9: Concatenated tree estimated from the unfiltered Tomatoes dataset



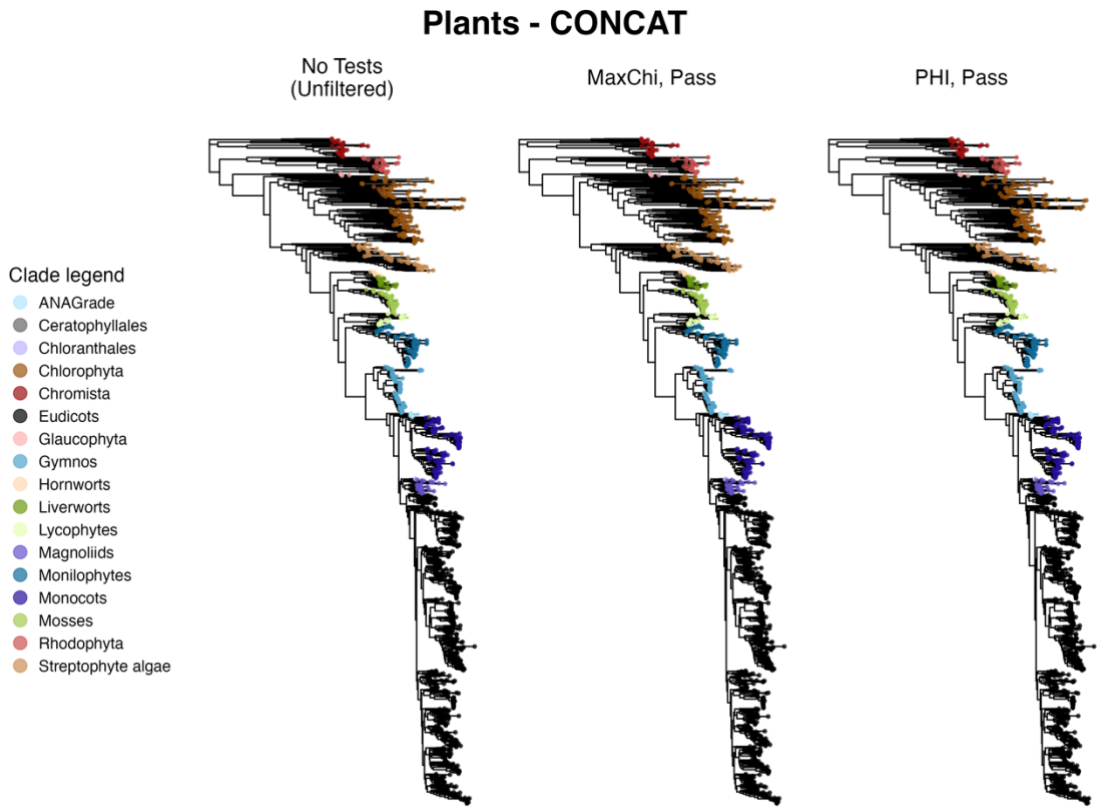
Supplementary Figure 10: Concatenated trees estimated from the Tomatoes dataset using the subsets P_GENECONV, F_GENECONV, P_MaxChi, F_F_MaxChi, P_PHI, F_PHI, P_ALL and F_ALL.
Only the topology of the Peruvianum clade (shown in green) differs between trees.



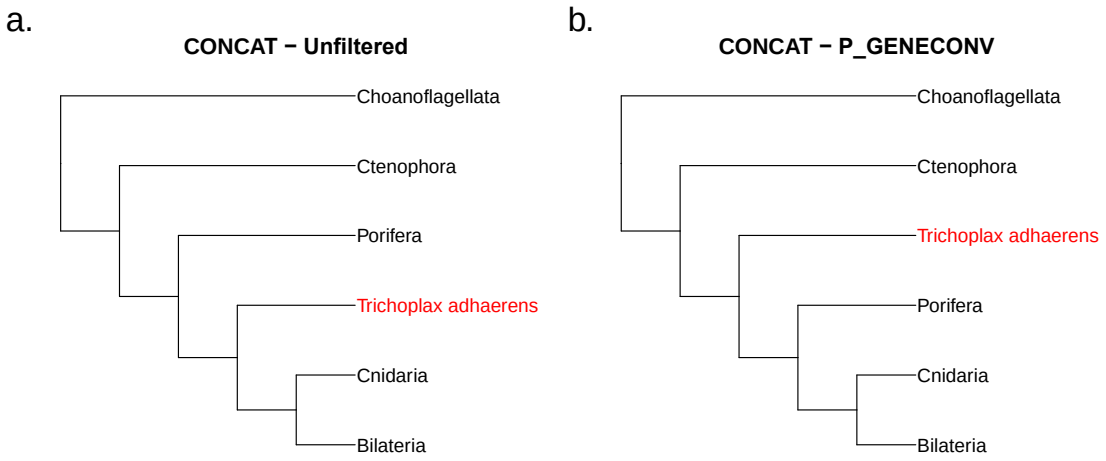
Supplementary Figure 11: Concatenated trees estimated from the Metazoan dataset, estimated from the Unfiltered dataset (top) and P_GENECONV subset (bottom).
The bottom tree CONCAT$_{P_GENECONV}$ has a different placement of Placozoa to the top tree CONCAT$_{Unfiltered}$. The majority of differences between trees occur in the Ctenophora clade (shown in blue).



Supplementary Figure 12: Concatenated trees estimated from the Metazoan dataset, estimated from the P_MaxChi subset (top) and P_PHI subset (bottom).
The two trees $CONCAT_{P\_MaxChi}$ and $CONCAT_{P\_PHI}$ have the same arrangement of Metazoan clades as the tree $CONCAT_{Unfiltered}$ (Supplementary Figure 11). The majority of differences between trees occur in the Ctenophora clade (shown in blue).



Supplementary Figure 13: Concatenated trees estimated from the Plants dataset, from the Unfiltered dataset (left), the P_MaxChi subset (centre) and the P_PHI subset (right). The three trees have the same relationships between clades.

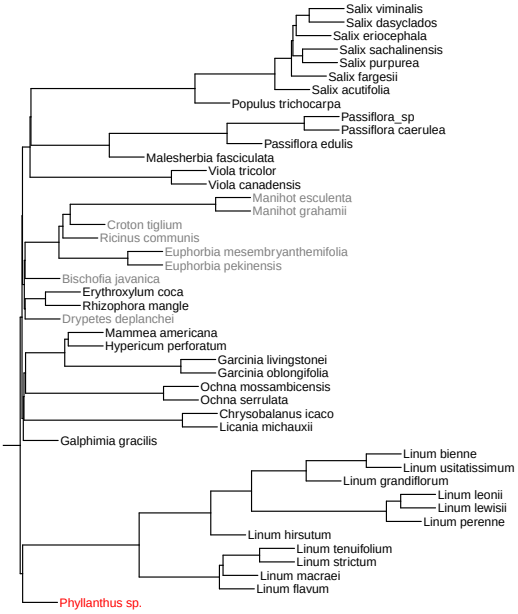


Supplementary Figure 14: One outlier branch (conflicting branch with ultrafast bootstrap > 90) from CONCAT trees estimated from subsets of the Metazoan dataset.

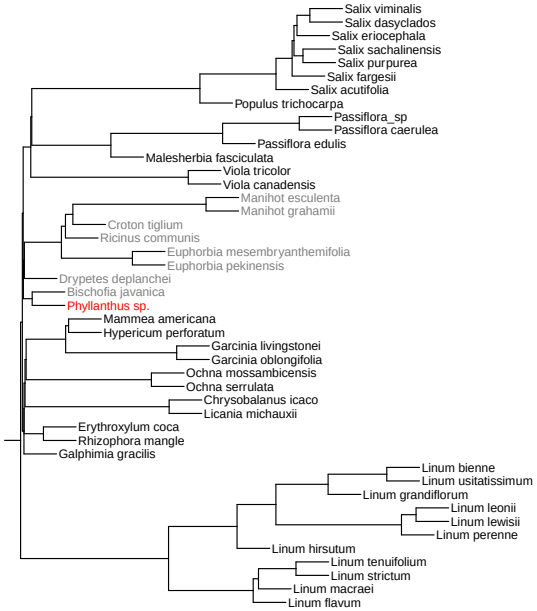
The outlier branch occurs on tree CONCAT$_{P_GENECONV}$ and involves alternate placement of the Placozoa clade (represented by the species *Trichoplax adhaerens)*. Tips within all clades except Placozoa have been collapsed into clades (Choanoflagellata, Ctenophora, Porifera, Cnidaria and Bilateria).

a. The CONCAT tree estimated from the Unfiltered Metazoan dataset
b. The CONCAT tree estimated from the P_GENECONV subset of the Metazoan dataset

a.



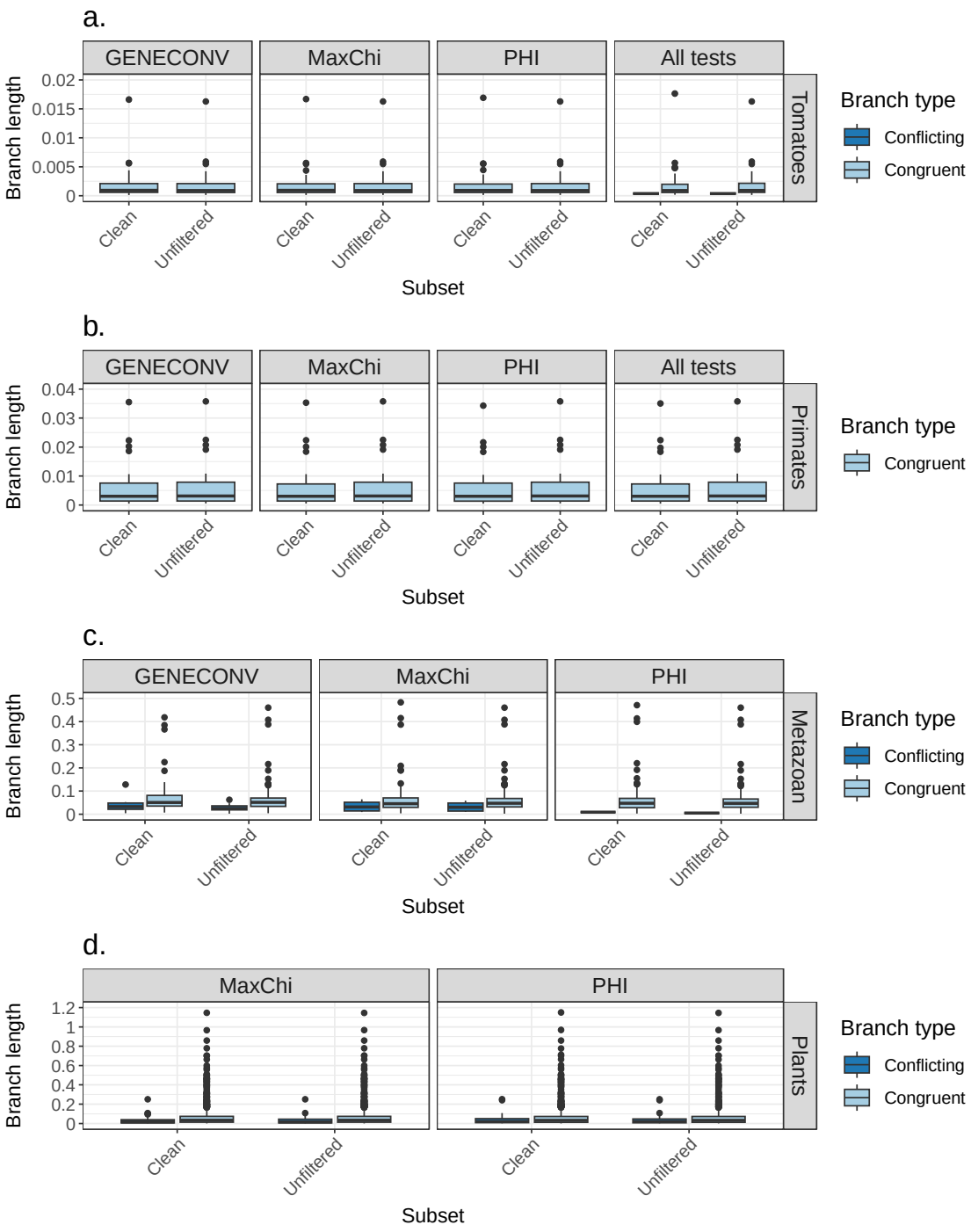
b.



Supplementary Figure 15: The single outlier branch (conflicting branch with posterior probability > 0.9) from ASTRAL trees estimated from subsets of the Plants dataset.

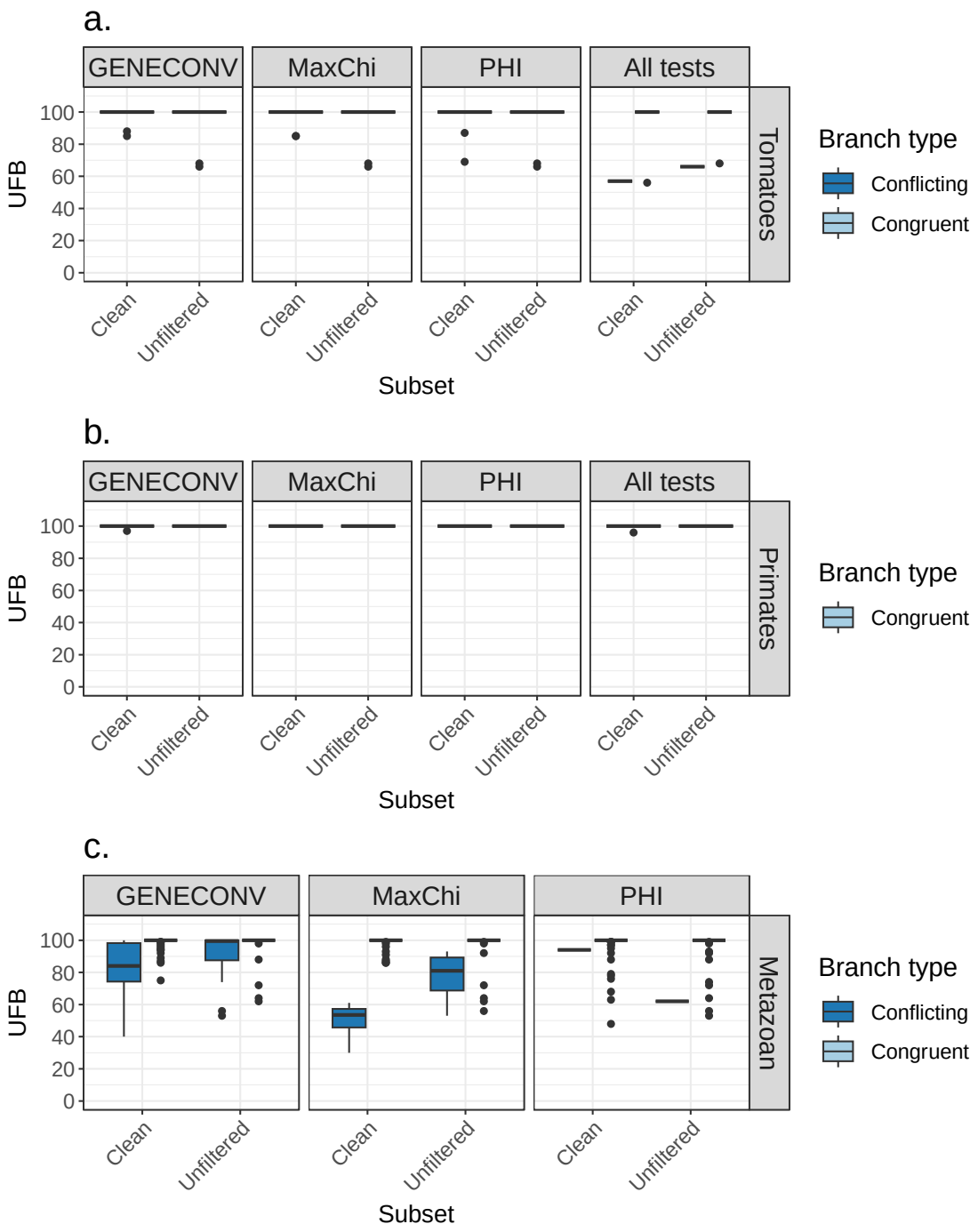
Only a subset of tips in the tree are shown, from the clade Malpighiales (classification from Leebens-Mack et al. (2019a): CoreEudicots/Rosids). The outlier branch occurs on tree ASTRAL$_{P_PHI}$. All tips in the clade descending from the outlier branch are shown in grey, except the tip *Phyllanthus sp.* which has different placement in the two trees.

a. The ASTRAL tree estimated from the Unfiltered dataset, showing different placement of *Phyllanthus sp.*
b. The ASTRAL tree estimated from the P_PHI subset



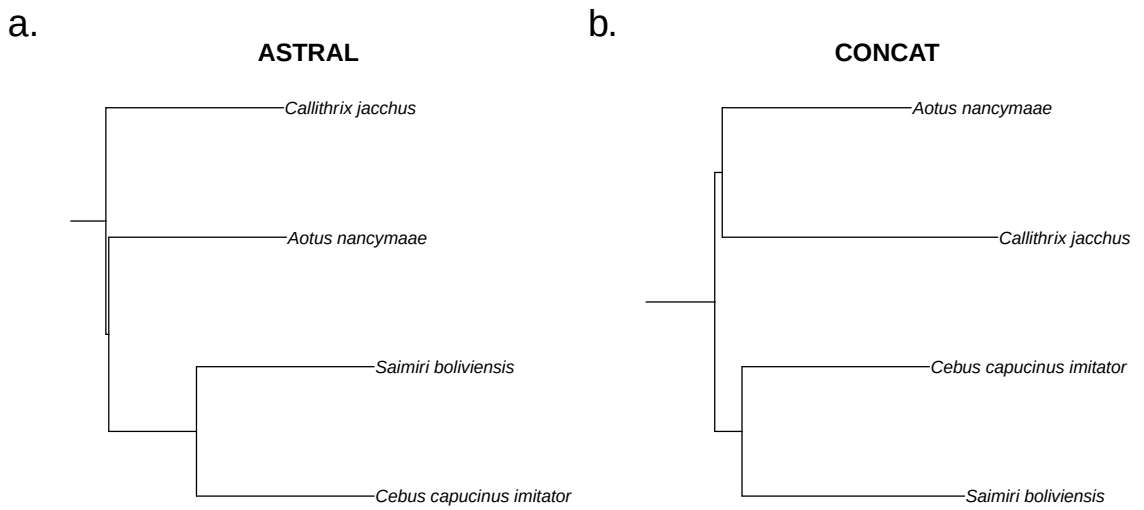
Supplementary Figure 16: Branch lengths for congruent and conflicting branches in CONCAT trees for all four datasets.

Dataset names are listed on the right of each panel. Each subset tree was compared to the tree estimated from the unfiltered dataset. Congruent branches were branches present in both the unfiltered tree and the subset tree. Conflicting branches were present in only the unfiltered tree or the subset tree. Where trees were identical to the tree estimated from the unfiltered dataset, no conflicting branches exist and no boxplot for conflicting branches is present.



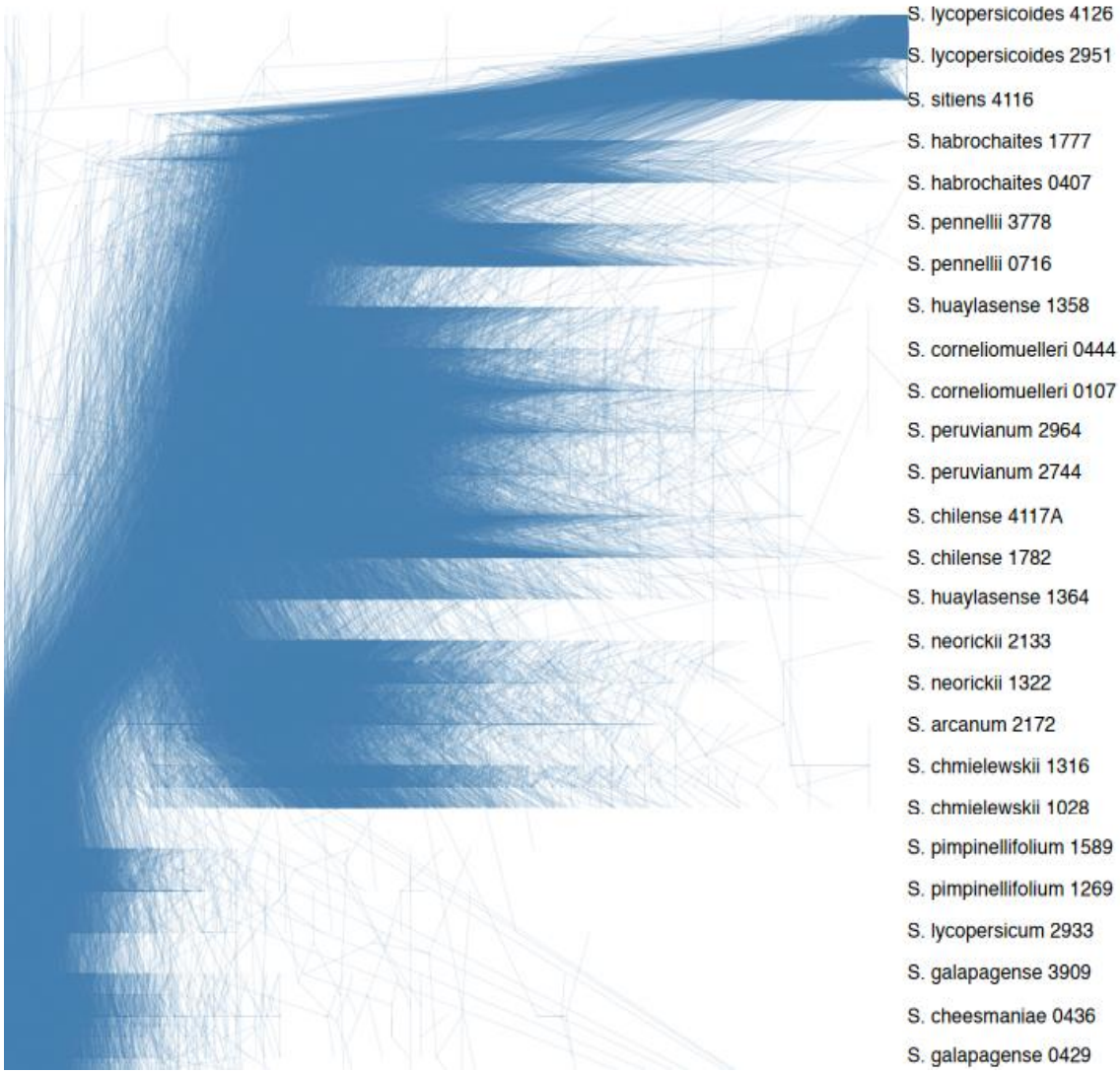
Supplementary Figure 17: The Ultrafast Bootstrap (UFB) support for congruent and conflicting branches in CONCAT trees for all four datasets.

Dataset names are listed on the right of each panel. Each subset tree was compared to the tree estimated from the unfiltered dataset. Congruent branches were branches present in both the unfiltered tree and the subset tree. Conflicting branches were present in only the unfiltered tree or the subset tree. Where trees were identical to the tree estimated from the unfiltered dataset, no conflicting branches exist and so no boxplot for conflicting branches is present.



Supplementary Figure 18: Topology of the Cebidae clade from the Unfiltered Primates dataset under different tree inference methods

a. Clade topology extracted from summary tree estimated in ASTRAL ($ASTRAL_{Unfiltered}$). ASTRAL does not output terminal branch lengths, so these were arbitrarily set to 1 for plotting purposes only
b. Clade topology extracted from concatenated tree ($CONCAT_{Unfiltered}$) estimated in IQ-Tree2



Supplementary Figure 19: Cloudogram showing all gene trees from the Tomatoes dataset simultaneously. Plot generated using the densiTree function in R package Phangorn v2.6.3 (Schliep 2011) with idea from DensiTree (Bouckaert 2010).

Chapter Three:
A Single Tree is Insufficient to Describe Evolutionary Relationships Between Animal Clades

Caitlin Cherryh[1]*, Thomas Wong[1], Davide Pisani[2], Bui Quang Minh[3], Robert Lanfear[1]

[1] Division of Ecology and Evolution, Research School of Biology, Australian National University, Canberra, Australia

[2] Palaeobiology Research Group. School of Biological Sciences and School of Earth Sciences, University of Bristol, Bristol, UK

[3] School of Computing, Australian National University, Canberra, Australia

* Corresponding author: caitlin.cherryh@anu.edu.au

Contributions:

Caitlin Cherryh collated the datasets, wrote the R scripts, performed the analysis, interpreted the results, and drafted the manuscript. Thomas Wong created and updated the MAST implementation, and assisted with MAST analyses. Davide Pisani and Minh Bui assisted with conceptual development and experimental design. Robert Lanfear contributed to conceptual development, experimental design, and provided editorial comments.

3.1 Abstract

Reconstructing the root of the tree of all animals is a central question in evolutionary biology. Early morphological and molecular studies tended to place Porifera (sponges) as the sister group to all other animals. More recently, phylogenomic studies have found support for a range of different clades as sister group to all other animals, including Ctenophora, Porifera, Ctenophora + Porifera or Porifera + Placozoa. In addition, the Porifera are sometimes resolved as a monophyletic clade and sometimes a paraphyletic clade. Phylogenomic studies have tended to recover each tree with very high bootstrap support but have also demonstrated that the topology obtained can depend on relatively minor changes to the underlying dataset and/or the model of sequence evolution used. Due to the complexity of reconstructing short branches at deep evolutionary timescales, there is little consensus on the relationships between these clades. Here, I use multitree mixture models to show that all combinations of previously analysed datasets and models exhibit substantial support for a wide range of hypotheses. To do this, I apply the recently developed multitree mixture model MAST (Mixtures Across Sites and Trees) to all combinations of 14 previously published datasets and 26 models of molecular evolution. The MAST method uses a mixture of bifurcating trees to represent multiple evolutionary histories for a single concatenated alignment. I apply MAST to five common hypotheses for the topology of early animal evolution and determine the relative weights of each hypothesis for 14 datasets and 4 classes of model. I find that multi-tree models are overwhelmingly preferred over single-tree analyses (47/56 analyses). For the 41 combinations of dataset and model class where I estimated both 2- and 5-tree MAST models, the 5-tree models were preferred (28/41) over the 2-tree (5/41) and single-tree (8/41) models. My results suggest that using current methods, a single phylogenetic tree is insufficient to describe the evolutionary history of Metazoa.

Keywords: Metazoa, Mixture model, Model adequacy, Early animal evolution, Heterogeneous phylogenetic signal, Model misspecification

3.2 Introduction

Determining the root of the metazoan tree is an open problem in phylogenetics. Historically, morphological analyses and early molecular phylogenies placed Porifera (sponges) as the sister group to all other metazoan clades (Haeckel 1866; Wainright et al. 1993; Collins 1998). Later phylogenetic analyses identified Ctenophores (comb jellies) as the sister group to all other metazoan clades (Ryan et al. 2013; Moroz et al. 2014). Since then, phylogenetic studies have found support for placing a range of different groups as the sister to all other metazoan

clades (Figure 17), including Porifera (Philippe et al. 2009; Pick et al. 2010; Philippe et al. 2011b; Pisani et al. 2015; Simion et al. 2017a; Arcila et al. 2017; Redmond and McLysaght 2021a), Ctenophora (Dunn et al. 2008; Hejnol et al. 2009; Ryan et al. 2013; Moroz et al. 2014; Borowiec et al. 2015; Whelan et al. 2015b; Shen et al. 2017), or a monophyletic group consisting of Porifera and Ctenophora (Shen et al. 2017; Francis and Canfield 2020). Alternative topological hypotheses have also been suggested (Schierwater et al. 2009), although these are less commonly recovered from phylogenomic analyses. In addition to this uncertainty around the sister group of all other metazoan clades, the monophyly or paraphyly of the sponges has been contested, with some studies finding support for monophyletic sponges (Philippe et al. 2009; Pick et al. 2010) and others for paraphyletic sponges (Sperling et al. 2007, 2009; Borchiellini et al. 2008). Interestingly, Nosenko et. al. (2013a) found different partitioning schemes of the same dataset could lead to support for either the monophyly or paraphyly of sponges, suggesting that model choice may play a central role in determining the outcomes of phylogenomic analyses of these questions. Due to the complexity in reconstructing these relationships, there is no consensus on the evolutionary history of these clades. Perhaps most surprisingly, support for different hypotheses has been shown to depend on choices made during almost every step of the phylogenomic analysis pipeline, including: the choice of loci (Laumer et al. 2018a; Pandey and Braun 2020), the filtering of sites from each locus (Francis and Canfield 2020; McCarthy et al. 2023), taxon sampling (Philippe et al. 2011b), the choice of outgroup (Nosenko et al. 2013a), and the choice of the model of substitution (Whelan and Halanych 2017; Kapli and Telford 2020; Redmond and McLysaght 2021a; Li et al. 2021).

Estimating the true evolutionary history of metazoan clades is complicated due to conflicting signals within phylogenetic datasets. Shen et. al. (2017) examined 8 alignments of metazoan species and found that 42.5 – 69.7% of genes and 39.8 – 59.6% of sites supported Ctenophora as sister to all other metazoan clades, with the rest supporting either Porifera or a monophyletic clade of Ctenophora and Porifera as the sister group to all other metazoan clades. This perhaps explains why previous studies have found that filtering genes and sites can impact the topology of the metazoan tree. One study found removing all sites that strongly supported either Ctenophora-sister or Porifera-sister resulted in a highly-supported tree with the monophyletic clade of Ctenophora and Porifera as the sister to all other animals (Francis and Canfield 2020). Additionally, Nosenko et. al. (2013a) found that varying the gene sampling and choice of outgroup resulted in different conflicting but well-supported topologies.

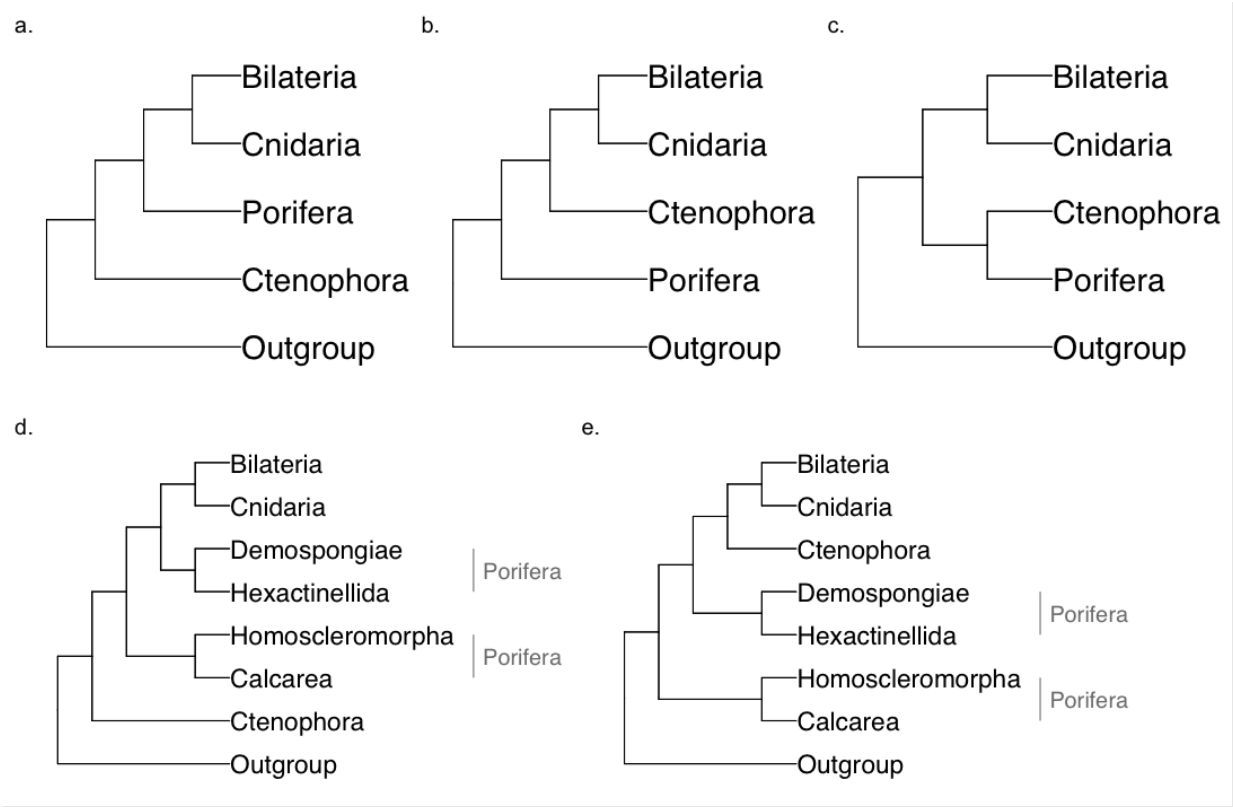


Figure 17: 5 possible alternative hypothesis topologies for early animals.

a. Ctenophora is the sister group to all other metazoans.
b. Porifera is the sister group to all other metazoans.
c. A monophyletic clade consisting of Porifera and Ctenophora is the sister group to all other metazoans.
d. Porifera is a paraphyletic clade (Porifera clades marked in grey). Ctenophora is the sister group to all other metazoans.
e. Porifera is a paraphyletic clade (Porifera clades marked in grey). The monophyletic clade consisting of the Porifera clades Homoscleromorpha and Calcarea is the sister group to all other metazoans (including the remaining Porifera clades Demospongiae and Hexactinellida).

Choice of substitution model also influences the relationships estimated between metazoan clades. Studies using partitioned site-homogeneous models tend to find Ctenophora as the sister to all other metazoan species (Ryan et al. 2013; Moroz et al. 2014; Whelan et al. 2015b, 2017a), while those using site-heterogeneous models such as the CAT model tend to recover Porifera-sister (Pisani et al. 2015; Simion et al. 2017a; Feuda et al. 2017a). A reanalysis of 36 alignments from 15 published studies found that Porifera-sister is recovered under site-heterogeneous CAT models with specific outgroups, while Ctenophora-sister is recovered across the range of substitution models and outgroups (Li et al. 2021). Similarly, Redmond and McLysaght (2021a) found that for three metazoan alignments, partitioned analyses with site-homogeneous models recovered Ctenophora-sister but better-fitting site-heterogeneous CAT models recovered Porifera-sister. In addition, the specific CAT model used has been

shown to impact metazoan tree topology estimates. Whelan and Halanych (2017) reanalysed two empirical phylogenetic datasets with CAT-GTR and CAT-F81 models and found the Philippe et al. (2009) dataset inferred a Ctenophora-sister for both models, but the Nosenko et al. (2013a) dataset inferred Ctenophora-sister for the CAT-GTR model and paraphyletic Porifera as sister for the CAT-F81 model. These studies highlight that the choice of substitution model can have an important impact on the outcome of phylogenomic studies of the relationships between metazoan clades.

A major consideration when selecting models to infer the metazoan tree is whether to use a site-homogeneous or site-heterogeneous model of sequence evolution. Previously, a range of approaches have been applied to infer a maximum likelihood tree for the Metazoa. Studies using site-homogeneous models apply either a single concatenated model (Dunn et al. 2008; Hejnol et al. 2009) or more commonly a partition model (Borowiec et al. 2015; Whelan et al. 2015b, 2017a), and those using site-heterogeneous models apply either profile mixture models (Laumer et al. 2018a) or an approximation of profile mixture models (Laumer et al. 2019a). However, site-homogeneous models often underestimate saturation and convergent evolution (Lartillot et al. 2007; Pisani et al. 2015; Telford et al. 2015). Studies investigating the heterogeneous signal within the metazoan dataset have found that site-homogeneous models have inadequate fit, resulting in long branch attraction and biasing tree topology (Kapli and Telford 2020; Redmond and McLysaght 2021; Szánthó et al. 2023). Any study investigating the evolutionary history of the Metazoa must consider impact of model misspecification on tree topology, particularly the potential for model misspecification from site-homogeneous models.

To my knowledge, all previous phylogenetic analyses of the Metazoa have applied a single-tree framework. The maximum likelihood and Bayesian methods used for tree inference in previous studies have the underlying assumption that the evolutionary history of the provided species fits a single bifurcating tree. However, there is substantial conflicting signal within metazoan phylogenetic datasets (see above). If proportions of the genome supporting different evolutionary hypotheses are similar, that could explain how slight differences in model or site filtering result in different topologies. Despite the conflict present within metazoan datasets, previous phylogenetic studies of this group have high branch support values across the tree. Branch support is often assessed using bootstrap support (Felsenstein 1985) or posterior probability (Larget and Simon 1999), which are designed to determine the stochastic error due to sampling (Thomson and Brown 2022). However, bootstrap support values are not a measure of topological truth, as large genomic datasets have lower stochastic error and therefore can have very high bootstrap values even in the face of substantial conflict within a dataset (Rokas and Carroll 2006; Thomson and Brown 2022). The combined factors of

conflicting signal within metazoan datasets, limits of branch support values, and assumption of a single phylogenetic tree has resulted in multiple highly-supported but conflicting estimates of the metazoan phylogeny. In this chapter, I relax the treelikeness assumption to assess whether a single tree is appropriate to represent the evolutionary history of Metazoa, or whether multiple trees are necessary.

The Mixtures Across Sites and Trees (MAST) method (Wong et al. 2024) is a new generalization of phylogenetic methods which allows multiple evolutionary histories for a single concatenated alignment by explicitly representing each evolutionary history with a separate bifurcating tree. The MAST model works under a concatenation framework but relaxes the assumption that all sites within an alignment must be represented by a single bifurcating tree. Given an input alignment and a set of tree topologies, the MAST model calculates the likelihood of each site under each tree, the maximum-likelihood weight of each tree, and other details including branch lengths and parameters of the substitution model (Wong et al. 2024). The MAST model does not explicitly model biological processes such as incomplete lineage sorting (ILS) or recombination. However, MAST can represent many causes of conflict within a dataset by incorporating multiple input tree topologies (Wong et al. 2024). In other words, the MAST model facilitates investigation into conflicting and discordant phylogenetic signal without the need to explicitly model biological processes. A mixture of trees approach is well-suited to question of early animal evolution, as previous studies have shown that a single maximum likelihood tree is unlikely to represent all of the phylogenetic information present within an alignment of metazoan taxa (Philippe et al. 2011b; Nosenko et al. 2013a; Shen et al. 2017; Francis and Canfield 2020; Redmond and McLysaght 2021a). However, those studies have lacked the tools to evaluate the extent of the support for different hypotheses within a single statistical framework. The MAST model provides that framework.

In this chapter, I use the MAST model to investigate whether a mixture of trees model represents the evolutionary history of the Metazoa better than the single tree models that have been used previously. To do this, I extend the previously published MAST model (Wong et al. 2024), such that identical branches in the input trees must have equal lengths. In this way, the only difference between the input trees are the branches that differ between them, allowing me to evaluate support for the existence of multiple evolutionary histories while retaining most of the other assumptions made by single-tree models. I selected 14 published amino acid alignments and identified 26 amino acid substitution models used to investigate the evolutionary history of the Metazoa in these studies. Across all combinations of dataset and model (two of which failed to complete) I found that a multi-tree model was preferred 47 times out of 56, with a single-tree model preferred just 9 times. This strongly suggests that regardless

of the choice of loci in a dataset, choice of model, and choice of data filtering methods, a single phylogenetic tree is insufficient to represent the evolutionary history of the Metazoa.

3.3 Methods

I identified 5 hypotheses of animal evolution that are frequently recovered from phylogenetic studies (Figure 17). Most metazoan phylogenetic studies find a monophyletic clade of either Ctenophora (Figure 17a) or Porifera (Figure 17b) as the sister to all other metazoans. Another observed topology finds a monophyletic clade combining Ctenophora and Porifera as the sister to all other metazoans (Figure 17c). Alternatively, the Porifera clade may be paraphyletic. In this case there are two possible hypotheses: Ctenophora as sister to other metazoans including paraphyletic sponges (Figure 17d), or a paraphyletic sponges as sister to all other metazoans (specifically, the clades Calcarea and Homoscleromorpha form a monophyletic grouping as the sister to other metazoans) (Figure 17e).

Here I provide a brief overview of my approach, and provide full details of each step below. First, I identified 14 existing phylogenetic datasets and 25 substitution models that were previously used to estimate the metazoan phylogeny across these studies. I divided these models into four groups, where comparisons within each group could be made using the BIC. For each combination of alignment and substitution model group I first estimated a maximum likelihood tree in IQ-Tree2. I selected the best model in each group, then estimated 5 constrained maximum likelihood trees which match the 5 most commonly-recovered topological hypotheses for metazoans shown in Figure 1. For each combination of 14 datasets and four model groups (56 in total) I then used the MAST model to compare the fit of a one-tree model, a two-tree model (containing just the trees shown in Figure 1a and 1b) and a five-tree model (all trees in Figure 1), and to estimate the associated weights of each tree in each case. The AU test works on the strict assumption that a single tree underlies the evolutionary history of the sequence, and is often used to compare trees in a phylogenetic framework (Shimodaira 2002; Planet 2006). Comparing BIC of the 1-tree, 2-tree and 5-tree model is a more rigorous framework as it relaxes the treelikeness assumption, allowing me to consider alternative topologies and alternative mixtures of trees simultaneously.

3.3.1 Datasets and matrix selection

I gathered papers from the literature investigating the relationships between the clades at the root of the metazoan tree. I selected all papers that: included 1 or more taxa in the Ctenophora, Cnidaria, Porifera, Placozoa and Bilateria clades; had amino acid alignments; and had alignments deposited and available in supplementary information or an external repository. This resulted in a list of 15 studies, of which two (Hejnol et al. 2009; Simion et al. 2017a, 2017b)

were computationally intractable for my analyses. The 13 remaining studies and corresponding alignments are detailed in Table 4.

Table 4: Alignments selected for analysis.

Manuscript refers to the original publication of each alignment. Repository is the source of the alignment file, which may be supplementary material in the original manuscript, a data repository for the original manuscript, or a data repository for a different manuscript. Matrix name is the name of each alignment file at the source for that alignment file. The number of taxa and number of amino acid sites for each alignment is also listed.

Manuscript	Repository	Matrix name	Number of taxa	Number of sites
Dunn *et al.* (2008)	Li et. al. (2020b)	Dunn2008	64	21152
Philippe *et al.* (2009)	Philippe et. al. (2009)	Philippe_etal_superalignment	55	30257
Pick *et al.* (2010)	Li et. al. (2020b)	Pick2010	83	19002
Philippe *et al.* (2011b)	Philippe et. al. (2011b)	UPDUNN_MB	77	18463
Nosenko *et al.* (2013a)	Nosenko et. al. (2013b)	nonribosomal_9187_smatrix	71	9189
Nosenko *et al.* (2013a)	Nosenko et. al. (2013b)	ribosomal_14615_smatrix	71	14614
Ryan *et al.* (2013)	Redmond and McLysaght (2021b)	REA_alignment_ includingXenoturbella	61	88384
Moroz *et al.* (2014)	Li et. al. (2020b)	ED3d	46	22772
Borowiec *et al.* (2015)	Borowiec et. al. (2016)	Best108	36	41808
Chang *et al.* (2015)	Feuda et. al. (2017b)	Chang_AA	77	51940
Whelan *et al.* (2015b)	Whelan et. al. (2016)	Dataset10	70	59733
Whelan *et al.* (2017a)	Whelan et. al. (2017b)	Metazoa_Choano_RCFV_strict	76	49388
Laumer *et al.* (2018a)	Laumer et. al. (2018b)	Tplx_phylo_d1	59	73547
Laumer *et al.* (2019a)	Laumer et. al. (2019b, 2019c)	nonbilateria_MARE_BMGE	51	61096

Where more than one alignment was created for a single study (e.g., due to different filtering schemes), I selected the alignment used for the main phylogenetic tree figure in the results section. For some papers, there was no obvious main phylogeny, or the main phylogeny did not include the relevant clades. In this case, I selected the alignment that best met my requirements. For Moroz et. al. (2014), the phylogenetic trees estimated from the alignments were included as Extended Figures 3a and 3d. I selected the alignment used for Extended Figure 3d as it included more taxa (12 Ctenophora species instead of 2). Laumer et. al. (2019a) included multiple alignments including different clades, and I selected the most taxon-rich alignment used to estimate the phylogeny seen in Figure 5, which focuses on the relationships

between metazoan clades. Finally, I selected both the ribosomal and non-ribosomal gene matrices from Nosenko et. al. (2013a), as they had different tree topologies. This resulted in a total of 14 alignments from 13 studies (Table 1). The number of taxa in the alignments varied between 36 and 94 species, and the number of sites between 9 kb and 270 kb. I did not modify or filter alignments in any way, except to update taxon names to be consistent across datasets and allow the alignments to be batch processed using scripts. Results from the original studies are summarised in Supplementary Table 4. My alignments are available from the Figshare repository for this project (https//doi.org/10.6084/m9.figshare.26087386).

3.3.2 Substitution model selection

Previous studies into the question of metazoan relationships have used a wide variety of substitution models. In this chapter I sought to examine the impact of model choice by applying all previously-used amino-acid models to the list of 14 datasets described above, including in papers that have re-analysed those datasets (Pisani et al. 2015; Feuda et al. 2017a; Shen et al. 2017; Whelan and Halanych 2017; Francis and Canfield 2020; Li et al. 2021; Redmond and McLysaght 2021a). To do so, I extracted a list of all substitution models used to estimate metazoan phylogenies from each paper and removed any gamma, site frequency, or rate heterogeneity across site parameters. To create my list of models I wrote custom R v4.3.1 scripts using the packages ape v5.7.1 (Paradis and Schliep 2019), phangorn v2.11.1 (Schliep 2011; Schliep et al. 2017), phylotools v0.2.2 (Zhang 2017), stringr v1.5.1 (Wickham 2023) and TreeTools v1.10.0 (Smith and Paradis 2023). To replicate this process, the R scripts data_dataset_info.R, 00_Li_extracting_all_models.R, 00_Redmond_extracting_all_models.R, and func_data_processing.R are available at the GitHub repository for this project (https://github.com/caitlinch/metazoan-mixtures). This resulted in a list of 25 amino acid substitution models: GTR20, JTT, JTTDCMut, LG, mtZOA, Poisson, PMB, rtREV, WAG, CF4, EHO, EX_EHO, EX2, EX3, LG4M, UL2, UL3, Poisson+C20, Poisson+C60, LG+C20, LG+C60, PMSF LG+C20, PMSF LG+C60, PMSF Poisson+C20, PMSF Poisson+C60.

I split these models into four categories described below, as the fit of different classes of models cannot be compared using information criteria such as BIC (Crotty and Holland 2022): the profile mixture (PM) class; the posterior mean site frequency (PMSF) class; the mixture class; and the Q class. First, the profile mixture (PM) class contains all C20/C60 models (Le et al. 2008a) previously applied to the metazoan phylogeny: Poisson+C20, Poisson+C60, LG+C20, and LG+C60. These empirical profile mixture models are an empirically determined, pre-learnt version of the site-heterogeneous CAT model that can be applied during maximum likelihood tree estimation while explicitly accounting for differences in evolutionary pressure

between sites (Lartillot and Philippe 2004, 2006; Lartillot et al. 2007; Le et al. 2008a). The posterior mean site frequency (PMSF) class (Wang et al. 2018) category contains the following models: PMSF LG+C20, PMSF LG+C60, PMSF Poisson+C20, and PMSF Poisson+C60. The PMSF models are a rapid and computationally-efficient approximation for the empirical profile mixture models, where each site in the alignment is assigned an amino acid profile calculated from an input mixture model and a guide tree (usually calculated with an empirical exchange-rate matrix model such as LG) (Wang et al. 2018). The Mixture class contains the protein mixture models: CF4, EHO, EX_EHO, EX2, EX3, LG4M, UL2, and UL3. Protein mixture models include multiple substitution matrices (between 2 and 6 matrices for the models in this chapter), and the likelihood of each site is calculated as a weighted average over the set of matrices (Le et al. 2008b). Finally, the Q category contains the empirical exchange-rate matrices: GTR20, JTT, JTTDCMut, LG, mtZOA, Poisson, PMB, rtREV, and WAG. These models consist of a single matrix specifying transition rates between each pair of amino acids, and a set of state frequencies (Minh et al. 2021). To check whether I was missing any models I also added ModelFinder (Kalyaanamoorthy et al. 2017) to the list of models, meaning I performed a model search in IQ-Tree with ModelFinder to identify the best model for each dataset. As the MAST model implementation in IQ-Tree (Wong et al. 2024) cannot use partitioned models, I took a concatenated approach and applied a single substitution model to all the sites in an alignment.

3.3.3 Maximum likelihood tree estimation

For each dataset I estimated a single maximum likelihood tree from each of the 25 substitution models listed above, plus a maximum likelihood tree estimatedthe with the model selected by ModelFinder. Estimating maximum likelihood trees using PMSF models required estimating the site frequency profiles, resulting in a three-step tree estimation process. Trees estimated from the other three model classes required only a single IQ-Tree command.

For each model in the PM, Mixture, and Q category (*n*=21 models), I estimated a concatenated maximum likelihood tree in IQ-Tree2 v2.2.0 (Minh et al. 2020b) for each dataset. The command line used was "`iqtree2 -s alignment.fa -mset model --mrate E,I,G,I+G,R,I+R -bb 1000`", where "`model`" indicates the model of interest and "`alignment.fa`" indicates the alignment file. Here, "`-mset model`" restricts model selection to only the model specified (e.g., if the model is LG, the command is "`-mset LG`" and only the LG model is considered). I specified the rate heterogeneity types for model selection with the mrate option using the command "`--mrate E,I,G,I+G,R,I+R`", which resulted in selection of the best model of rate heterogeneity by ModelFinder for each analysis. For each analysis I performed 1000

ultrafast bootstraps (UFB) using the command "`-bb 1000`" (Hoang et al. 2018a). For the ModelFinder tree estimation only, I used the command "`iqtree2 -s alignment.fa -m MFP --mrate E,I,G,I+G,R,I+R -bb 1000`", where the command "`-m MFP`" results in model selection using ModelFinder.

For each model in the PMSF category (*n*=4), I followed a three-step process to estimate the maximum likelihood tree. First, a guide tree was estimated using the alignment and a simpler substitution model using the IQ-Tree2 command "`iqtree2 -s alignment.fa -m 'LG+F+G' -pre guidetree`", where "`alignment.fa`" was the alignment file and "`guidetree`" was the prefix for IQ-Tree2 to include in the output files. Second, the guide tree, the complex model and the alignment are used to estimate the mixture model parameters and infer the site-specific frequency profile. I used the command line "`iqtree2 -s alignment.fa -m model -ft guidetree.treefile -n 0 -pre sitefreqs`", where "`model`" was the relevant PMSF class model, "`-ft guidetree.treefile`" specified the guide tree estimated in the previous step, and "`sitefreqs`" was the output prefix for the site-specific frequency profile. As this step is memory intensive, I specify 0 threads using "`-nt 0`", which stops the analysis after estimating the site frequencies file. Finally, the inferred frequency model is used to estimate a phylogenetic tree with 1000 ultrafast bootstrap replicates using the command "`iqtree2 -s alignment.fa -m model -fs sitefreqs.sitefreq -b 1000`". The site-specific frequency profile was specified using the command "`-fs sitefreqs.sitefreq`", and the PMSF model was specified using "`-m model`". For the PMSF models I specified the model of rate heterogeneity as "`+F+R4`", as previous analyses of metazoan phylogenetic datasets applied the PMSF model with four rate classes (R4) (Laumer et al. 2018a) and included state frequencies from the alignment (Laumer et al. 2018a; Kapli and Telford 2020).

In total, I estimated $14 \times 26 = 364$ maximum likelihood trees. To estimate maximum likelihood trees I wrote custom R v4.3.1 scripts using the packages ape v5.7.1 (Paradis and Schliep 2019), phylotools v0.2.2 (Zhang 2017) and stringr v1.5.1 (Wickham 2023). The R scripts to replicate my analyses 01_estimate_all_trees_parallel.R, 01_estimate_PMSF_trees.R, func_estimate_trees.R, func_data_processing.R, and func_pmsf_trees.R are available at the GitHub repository for this project (https://github.com/caitlinch/metazoan-mixtures).

3.3.4 Hypothesis tree estimation

Conducting MAST and AU test analyses requires a set of hypothesis trees for each dataset. I estimated constrained maximum likelihood trees using a guide tree for each of the 5

hypotheses of metazoan evolution (Figure 17). I refer to these constrained trees as "hypothesis trees". Due to computational limitations, I limited these analyses to one model per class for each dataset. I selected the best model from each of the four classes for each dataset, by comparing the BIC for each tree estimated from a model in this class and selecting the model with the lowest BIC. For each best model, I identified the model of rate heterogeneity and state frequencies determined by ModelFinder and used these parameters to estimate constrained trees (see below). For example, the best Mixture class model for the Ryan 2013 dataset was LG4M, and the model selected by IQ-Tree including rate heterogeneity and state frequency parameters was "LG4M + R6{0.16,0.05,0.22,0.23,0.21,0.61,0.23,1.25,0.15,2.44,0.04,4.54}".

I found that the MAST model was unable to run when the best GTR models included an invariant sites model (about half the GTR models). To address this, when GTR was the best Q class model I estimated a new best GTR model for each alignment without the +I model of rate heterogeneity.

To construct the guide trees, I manually classified the taxa in each alignment into the following clades: Outgroup, Ctenophora, Porifera, Placozoa, Cnidaria, and Bilateria. I then constructed a multifurcating guide tree for each of the five alternative phylogenetic hypotheses for each alignment. Two datasets (Dunn 2008 and Borowiec 2015) contained a single taxon from the Porifera, and therefore only the first three hypotheses could be tested for these datasets. In total, I constructed $12 \times 5 + 2 \times 3 = 66$ guide trees. Placozoa was not included in either the alternative phylogenetic hypotheses or the multifurcating constraint trees, as the placement of Placozoa varies in different phylogenetic analyses of the Metazoa (Philippe et al. 2009; Pick et al. 2010; Nosenko et al. 2013a; Moroz et al. 2014; Simion et al. 2017a). As guide trees do not need to contain all taxa in an alignment, Placozoa is included in tree estimation even when excluded from the guide tree, and its position is estimated by maximum likelihood.

I estimated hypothesis trees in IQ-Tree2 v2.2.0, using the command line "`iqtree2 -s alignment.fa -m best_model -g guide_tree.nex`". Here, "`alignment.fa`" was the alignment, `best_model` was the best model for a given class including state frequency and rate parameters, and "`guide_tree.nex`" was the guide tree constraining maximum likelihood tree estimation. For the PMSF model class, I added the command "`-fs sitefreqs.sitefreq`" to specify the site-specific frequency profile associated with the best model. I had 4 classes of models, 12 datasets with 5 guide trees and 2 datasets with 3 guide trees, resulting in a total of $(4 \times 12 \times 5) + (4 \times 2 \times 3) = 264$ hypothesis trees.

I wrote a custom R v4.3.1 script with the R packages ape v5.7.1 (Paradis and Schliep 2019) and stringr v1.5.1 (Wickham 2023) to estimate hypothesis trees. The scripts available to reproduce these analyses are available in the files 02_estimate_hypothesis_trees.R, func_estimate_trees.R, func_data_processing.R and data_dataset_info.R at the GitHub repository for this project (https://github.com/caitlinch/metazoan-mixtures). I plotted results using custom R v3.4.1 scripts using the packages ape v5.7.1 (Paradis and Schliep 2019), dplyr v1.1.4 (Wickham et al. 2021), ggplot2 v3.4.4 (Wickham 2016), ggtree v3.8.2 (Yu et al. 2017; Xu et al. 2022), patchwork v.1.1.3 (Pedersen 2022), reshape2 v1.4.4 (Wickham 2007), and readxl v1.4.3 (Wickham and Bryan 2023). The scripts to replicate my plotting are available on the GitHub repository for this project in the files 05_plots.R, 05_plots_5trees.R, and func_plotting.R.

3.3.5 Applying the MAST model

I planned to estimate 1-tree, 2-tree MAST models, and 5-tree MAST models for each combination of dataset and model class. I found estimating 5-tree MAST models from PM models was computationally intractable (not completed when run with >200 threads for >4 weeks), limiting me to just 2-tree MAST models for the PM mixture class. For the other three model classes (PMSF, Mixture, and Q) I estimated both 2-tree and 5-tree MAST models for each dataset. In total, I performed $(1 \times 1 \times 14) + (3 \times 2 \times 14) = 98$ MAST analyses.

The Mixtures Across Sites and Trees (MAST) model was introduced by Wong et al. (2024). Given a set of input trees, MAST will optimise the mixture of trees including the relative weights of each input tree, the branch lengths in each topology, and the parameters of the evolutionary model (Wong et al., 2024). The most general MAST model has unlinked parameters, so each class has independent tree topology corresponding to one of the input trees, branch lengths, substitution model, state frequencies, and the rate heterogeneity across sites model. In this chapter, I apply a more restrictive linked MAST model where the substitution model, state frequencies, and rate heterogeneity models are identical across all classes. In addition, I introduce the branch-length-restricted model ("+TR"), where branch lengths are linked across topologies such that a branch b_i of tree T_i is considered equal to the length of branch b_j in tree T_j if both branches b_i and b_j split the trees T_i and T_j into the same sets of taxa (in other words, where the same branch appears in more than one tree in the MAST model, that branch must have the same length in all trees).

I applied the MAST model (Wong et al. 2024) using IQ-Tree2 v2.2.6.hmmster (now available as IQ-Tree v2.3.0 at https://github.com/iqtree/iqtree2/releases/tag/v2.3.0). To run the MAST

model I used the command "`iqtree2_MAST -s alignment.fa -m 'best_model+TR' -te hypothesis_trees.nex -blmin 100/N`". For PMSF models, I added the command "`-fs sitefreqs.sitefreq`" to specify the site-specific frequency profile for that dataset. Here, `iqtree2_MAST` refers to the IQ-Tree2 version containing the MAST implementation. The set of hypothesis trees (either 2 or 5) swas specified with the command "`-te hypothesis_trees.nex`". As detailed above, I fixed the model of evolution for each class as the best model from that class including rate heterogeneity and state frequencies, and for the PM class models only I included weights and rates for each profile. I used the command "`-m 'best_model+TR'`" to specify that best model with the branch-length restricted MAST model, as detailed above. Using the command "`-blmin 100/N`" I set the minimum branch length for each run to 100/N, where N is the number of sites in that alignment. This ensured that each branch contained at least one substitution, thus conservatively ensuring that the use of multiple trees in an analysis results in a non-zero likelihood cost. Lastly, I increased the weight to 0.0001 for any PM profiles with a weight of 0, as weights of exactly 0 caused the software to crash. This impacted the PM model parameter weights for 4 datasets: Dunn et al. (2008), Moroz et al. (2014), Nosenko et al. (2013a) non-ribosomal and Ryan et al. (2013).

Three combinations of PM model class and dataset were unable to run successfully. In two cases, MAST repeatedly crashed (i.e., failed to optimise weights) for the best PM model. For these analyses, I instead applied MAST with the second-best model from that category: for the Nosenko 2013 non-ribosomal matrix, I replaced the best model Poisson+C60 with the second-best model LG+C60. For the Laumer 2018 matrix, I replaced the best model LG+C60 with the second-best model LG+C20. Finally, the Pick 2010 matrix was unable to identify a maximum-likelihood 2-tree mixture model in a tractable time, and I excluded this analysis after waiting >8 weeks. Consequently, the Pick 2010 dataset has no multi-tree models for the PM model class.

I wrote a custom-written R v4.2.1 script with the R packages ape v5.7.1 (Paradis and Schliep 2019) and stringr v1.5.0 (Wickham 2023) to apply the MAST model to each alignment and extract the results. I plotted results as described above. The scripts available to repeat these analyses are available in the files 02_estimate_hypothesis_trees.R, func_estimate_trees.R, and func_data_processing.R at the GitHub repository for this project (https://github.com/caitlinch/metazoan-mixtures). Results and output files are available from the Figshare repository (https://doi.org/10.6084/m9.figshare.26087386).

3.3.6 AU tests

I performed the approximately unbiased (AU) test (Shimodaira 2002) for each set of trees used in either a 2-tree or 5-tree MAST mixture. The AU test is commonly used in phylogenetics to compare trees (Planet 2006). The AU test asks which of the alternative phylogenetic hypotheses can be rejected in favour of the ML tree given the data and model, and assuming that a single tree can represent the dataset and that the single tree is the ML tree under the best model ($p<0.05$). For each combination of dataset and model class, I performed the AU test once with the two hypothesis trees used for the 2-tree MAST model, and once with the 5 hypothesis trees used for the 5-tree MAST model. This resulted in a total of $14 \times 4 \times 2 = 112$ AU tests.

To perform AU tests, I used IQ-Tree2 v2.2.0 with the command "`iqtree2 -s alignment.fa -m best_model -z hypothesis_trees.nex -n 0 -zb 10000 -zw -au -pre prefix`", where "`alignment.fa`" is the alignment, "`best_model`" is the best model for that model class as detailed above, "`hypothesis_trees.nex`" is the nexus file containing the five maximum likelihood hypothesis trees for this combination of alignment and initial model, and prefix is a unique identifier for this set of input parameters. When the best model was a PMSF model, I added the command "`-fs sitefreqs.sitefreq`" to specify the site-specific frequency profile associated with the best model.

I wrote a custom-written R v4.2.1 script with the R packages ape v5.7.1 (Paradis and Schliep 2019) and stringr v1.5.0 (Wickham 2023) to apply the tree topology tests to each alignment and extract the results. The scripts available to repeat these analyses are available in the files 02_estimate_hypothesis_trees.R, func_estimate_trees.R, and func_data_processing.R at the GitHub repository for this project (https://github.com/caitlinch/metazoan-mixtures). I plotted results as described above. Results and output files are available from the Figshare repository (https://doi.org/10.6084/m9.figshare.26087386).

3.4 Results

3.4.1 Multi-tree models are generally preferred over single-tree models

Table 5: BIC scores for single-tree and multi-tree models, across 14 empirical phylogenetic datasets and 4 classes of substitution model.

For each class, the hypothesis tree with the lowest BIC is noted as the single-tree model and the hypothesis is listed in the Tree topology column. Any analyses with >1 tree were estimated using the MAST model. Any MAST analyses that were not possible to estimate are excluded from this table. The analysis with the lowest BIC for each combination and dataset and model class are shown in bold. For each combination of dataset and model class, the Δ BIC column indicates the difference between the best BIC and each BIC.

Dataset	Model class	Best model	Number of trees	Tree topology	BIC	Δ BIC
Dunn 2008	Q	GTR20+F+R7	1	CTEN	1386708.27	16.22
			2	–	1386714.02	21.97
			5	**–**	**1386692.05**	**0**
	Mixture	UL3+R7	**1**	**CTEN**	**1370933.49**	**0**
			2	–	1370943.06	9.58
			5	–	1370937.30	3.81
	PMSF	Poisson+SSF+F+R4	**1**	**CTEN**	**1215369.63**	**0**
			2	–	1215387.67	18.04
			5	–	1215407.06	37.43
	PM	C60+LG+F+R7	1	CTEN	1351879.28	574.63
			2	**–**	**1351304.65**	**0**
Philippe 2009	Q	GTR20+F+R7	1	CTEN	1668411.75	999.1
			2	–	1667815.03	402.38
			5	**–**	**1667412.65**	**0**
	Mixture	UL3+R7	1	CTEN	1649108.25	537.42
			2	–	1648869.60	298.77
			5	**–**	**1648570.83**	**0**
	PMSF	Poisson+SSF+F+R4	1	CTEN	1454859.44	51.43
			2	–	1454819.18	11.17
			5	**–**	**1454808.01**	**0**
	PM	C60+LG+F+R7	1	CTEN	1614346.69	642.14
			2	**–**	**1613704.55**	**0**
Pick 2010	Q	GTR20+F+R8	1	CTEN	1689817.07	677.68
			2	–	1689459.14	319.74
			5	**–**	**1689139.40**	**0**
	Mixture	UL3+R8	1	CTEN	1665577.75	425.25
			2	–	1665330.88	178.37
			5	**–**	**1665152.50**	**0**
	PMSF	Poisson+SSF+F+R4	1	CTEN	1511872.00	62.04
			2	–	1511813.70	3.73
			5	**–**	**1511809.96**	**0**
	PM	C60+LG+F+R9	**1**	**CTEN**	**1633387.59**	**0**
Philippe 2011	Q	GTR20+F+R8	1	CTEN	1680883.42	490.41
			2	–	1680553.14	160.13
			5	**–**	**1680393.012**	**0**
	Mixture	UL3+R8	1	CTEN	1656124.71	723.25
			2	–	1655650.88	249.42
			5	**–**	**1655401.46**	**0**
	PMSF	Poisson+SSF+F+R4	1	CTEN	1502156.97	21.8
			2	**–**	**1502135.16**	**0**
			5	–	1502166.53	31.37

Dataset	Model class	Best model	Number of trees	Tree topology	BIC	Δ BIC
	PM	C60+LG+F+R9	1	PORI	1621446.25	577.44
			2	**–**	**1620868.80**	**0**
Nosenko 2013 non-ribosomal	Q	GTR20+F+R6	1	CTEN	584617.40	188.6
			2	–	584530.74	101.93
			5	**–**	**584428.81**	**0**
	Mixture	UL3+R7	1	CTEN	575544.40	68.19
			2	–	575506.59	30.39
			5	**–**	**575476.21**	**0**
	PMSF	Poisson+SSF+F+R4	**1**	**CTEN**	**508995.61**	**0**
			2	–	508999.62	4.02
			5	–	509039.15	43.55
	PM	C60+LG+F+R7	1	CTEN	565224.87	537.76
			2	**–**	**564687.11**	**0**
Nosenko 2013 ribosomal	Q	GTR20+F+R7	1	CTEN	1034699.36	583.42
			2	–	1034665.78	549.84
			5	**–**	**1034115.93**	**0**
	Mixture	UL3+R7	1	CTEN	1022597.41	96.71
			2	–	1022577.47	76.77
			5	**–**	**1022500.70**	**0**
	PMSF	Poisson+SSF+F+R4	1	CTEN	917539.04	32.3
			2	–	917545.56	38.83
			5	**–**	**917506.74**	**0**
	PM	C60+LG+F+R6	1	CTEN	1005383.01	567.34
			2	**–**	**1004815.68**	**0**
Ryan 2013	Q	GTR20+F+R6	1	CTEN	4613310.04	1114.57
			2	–	4612664.86	469.4
			5	**–**	**4612195.47**	**0**
	Mixture	LG4M+R6	1	CTEN	4578211.43	563.14
			2	–	4577984.89	336.6
			5	**–**	**4577648.29**	**0**
	PMSF	Poisson+SSF+F+R4	1	CTEN	3997079.43	42.81
			2	–	3997042.92	6.31
			5	**–**	**3997036.62**	**0**
	PM	C60+LG+F+R7	1	CTEN	4527897.15	768.5
			2	**–**	**4527128.65**	**0**
Moroz 2014	Q	GTR20+F+R5	1	CTEN	1004366.95	32.96
			2	**–**	**1004333.99**	**0**
			5	–	1004377.30	43.31
	Mixture	LG4M+R5	1	CTEN	997482.91	1.9
			2	**–**	**997481.01**	**0**
			5	–	997513.30	32.29
	PMSF	Poisson+SSF+F+R4	**1**	**PORI**	**855875.99**	**0**
			2	–	855895.85	19.86
			5	–	855962.53	86.54
	PM	C60+LG+F+R5	1	PORI	990832.96	596.02
			2	**–**	**990236.94**	**0**
Borowiec 2015	Q	GTR20+F+R6	1	CTEN	2836519.09	429.17
			2	–	2836374.23	284.3
			5	**–**	**2836089.92**	**0**
	Mixture	LG4M+R6	1	CTEN	2815868.70	249.48
			2	–	2815802.79	183.57
			5	**–**	**2815619.22**	**0**
	PMSF	Poisson+SSF+F+R4	**1**	**CTEN**	**2521422.70**	**0**
			2	–	2521447.77	25.07
			5	–	2521466.21	43.51

Dataset	Model class	Best model	Number of trees	Tree topology	BIC	Δ BIC
	PM	C60+LG+F+R7	1	CTEN	2786191.13	615.54
			2	**–**	**2785575.59**	**0**
Chang 2015	Q	GTR20+F+R8	1	CTEN	4600731.49	1665.43
			2	–	4600190.44	1124.38
			5	**–**	**4599066.06**	**0**
	Mixture	UL3+R9	1	CTEN	4545324.86	1079.9
			2	–	4544998.78	753.82
			5	**–**	**4544244.96**	**0**
	PMSF	LG+SSF+F+R4	1	CTEN	4177836.72	125.52
			2	–	4177813.72	102.53
			5	**–**	**4177711.20**	**0**
	PM	C60+LG+F+R9	1	CTEN	4453142.27	682.23
			2	**–**	**4452460.04**	**0**
Whelan 2015	Q	GTR20+F+R9	1	CTEN	5783687.66	958.22
			2	–	5783429.43	699.99
			5	**–**	**5782729.44**	**0**
	Mixture	LG4M+R8	1	CTEN	5735143.02	588.84
			2	–	5734986.74	432.56
			5	**–**	**5734554.18**	**0**
	PMSF	Poisson+SSF+F+R4	1	CTEN	5294238.12	16.85
			2	–	5294238.82	17.55
			5	**–**	**5294221.27**	**0**
	PM	C60+LG+F+R8	1	CTEN	5678552.48	651.29
			2	**–**	**5677901.19**	**0**
Whelan 2017	Q	GTR20+F+R7	1	CTEN	3536075.89	131.273
			2	**–**	**3535944.62**	**0**
	Mixture	LG4M+R8	1	CTEN	3513036.91	283.96
			2	–	3512956.04	203.09
			5	**–**	**3512752.95**	**0**
	PMSF	Poisson+SSF+F+R4	**1**	**CTEN**	**3189240.35**	**0**
			2	–	3189251.66	11.31
			5	–	3189293.35	53
	PM	C60+LG+F+R7	1	CTEN	3482116.85	654.3
			2	**–**	**3481462.55**	**0**
Laumer 2018	Q	GTR20+F+R7	1	CTEN	6003757.92	1550.13
			2	–	6002739.69	531.9
			5	**–**	**6002207.78**	**0**
	Mixture	LG4M+R7	1	CTEN	5970627.93	1065.69
			2	–	5969880.22	317.97
			5	**–**	**5969562.24**	**0**
	PMSF	Poisson+SSF+F+R4	**1**	**CTEN**	**5470394.61**	**0**
			2	–	5470418.86	24.26
			5	–	5470499.56	104.95
	PM	C20+LG+F+R8	1	CTEN	5931973.09	212.16
			2	**–**	**5931760.93**	**0**
Laumer 2019	Q	GTR20+F+R7	1	CTEN	4241166.02	202.28
			2	**–**	**4240963.74**	**0**
	Mixture	LG4M+R6	1	CTEN	4209812.58	856.93
			2	–	4209699.34	743.69
			5	**–**	**4208955.65**	**0**
	PMSF	Poisson+SSF+F+R4	**1**	**CTEN**	**3789600.80**	**0**
			2	–	3789617.07	16.27
			5	–	3789684.80	84
	PM	C60+LG+F+R7	1	CTEN	4167453.35	666.1
			2	**–**	**4166787.25**	**0**

In general, multi-tree models had better fit than single-tree models across datasets and model classes (Table 5). In three model classes (Q, Mixture, PM) multi-tree models with the most trees were overwhelmingly preferred. For both the Q and PM model classes, the analysis with the lowest BIC was almost always a multi-tree model. For the Q model class, 13/14 times the best BIC came from the 2-tree (*n*=1), or 5-tree (*n*=12) MAST model. For the PM model class, all datasets with a completed 2-tree MAST model (*n*=13) had lowest BIC. The Mixture and PMSF classes had both single- and multi-tree models with the lowest BIC. For the Mixture model class, most datasets (*n*=12) had the lowest BIC for the 5-tree MAST model. However, the Dunn 2008 dataset had lowest BIC for the Ctenophora-sister hypothesis tree and the Moroz 2014 dataset had lowest BIC for the 2-tree MAST model. Finally, the PMSF model class had the highest proportion of single-tree models with lowest BIC (*n*=7), with 6 observations of the Ctenophora-sister and 1 observation of the Porifera-sister hypothesis trees. The remaining datasets within the PMSF model class had lowest BIC for either the 2-tree (*n*=1) or 5-tree (*n*=6) MAST models.

3.4.2 Different combinations of models and datasets support different evolutionary histories for metazoan clades

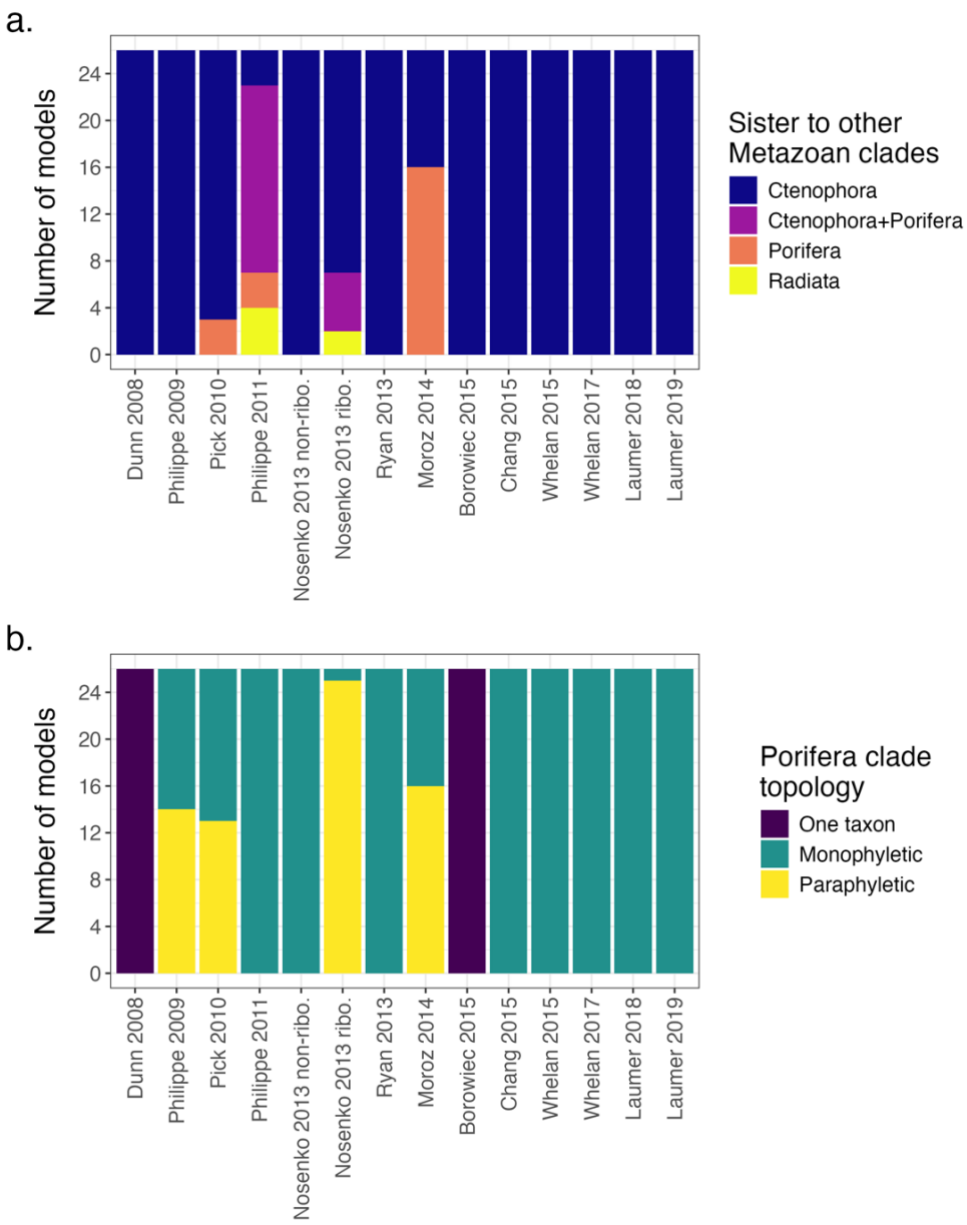


Figure 18: Metazoan phylogeny topology and Porifera clade topology under different substitution models

a. Maximum likelihood tree topology for trees estimated from each combination of 25 substitution models and 14 datasets. Each bar shows the tree topologies obtained for a single dataset. Sister to all other metazoan clades indicates the first clade to diverge. The "Ctenophora+Porifera" clade denotes a single monophyletic clade consisting of both the Porifera and Ctenophora clades. The "Radiata" clade denotes a monophyletic clade consisting of Porifera, Ctenophora, Cnidaria and Placozoa (although Placozoa was not included in all datasets).

b. Topology of the Porifera (sponge) clade for trees estimated from each combination of 25 substitution models and 14 datasets. Each bar shows the Porifera topologies obtained for a single dataset. Two datasets contained only 1 Porifera taxon, which meant the Porifera clade topology could not be obtained.

For the majority of datasets (10/14), the group inferred as sister to all other metazoans was stable regardless of model choice (Figure 18a, Supplementary Figure 20, Supplementary Tables 5-7). In only four datasets did the best single tree recovered depend on the model choice: Pick 2010, Philippe 2011, Nosenko 2013 non-ribosomal and Moroz 2014. Both the Pick 2010 and Nosenko 2013 non-ribosomal datasets inferred the Ctenophora-sister tree for most models of sequence evolution. The Philippe 2011 and Moroz 2014 datasets were the only two datasets that inferred a topology other than Ctenophora-sister most of the time. The preferred topology (as a percentage of the total number of trees) for Philippe 2011 was the monophyletic clade of Ctenophora and Porifera as the SOM, and Porifera-sister for Moroz 2014. For the latter two datasets, the preferred topology was influenced by model choice. For the Philippe 2011 dataset, the most common SOM observed under amino-acid exchange rate matrices and protein mixture models was a monophyletic clade consisting of Ctenophora and Porifera. However, under PM and PMSF models, the preferred topologies were Porifera-sister or Radiata. For the Moroz 2014 dataset, the preferred SOM was Ctenophora for amino-acid exchange rate matrices and Porifera for protein mixture models, PM models, and PMSF models.

Excluding the two datasets with only a single Porifera taxon (Dunn 2008 and Borowiec 2015), the majority of datasets (8/14) inferred a monophyletic clade for Porifera under each model of substitution (Figure 18b, Supplementary Figure 21, Supplementary Tables 5-7). Four datasets recovered paraphyletic Porifera (Philippe 2009, Pick 2010, Nosenko 2013 ribosomal, and Moroz 2014), and in each case paraphyletic Porifera was recovered more often than monophyletic Porifera. The Nosenko 2013 ribosomal dataset preferred paraphyletic Porifera for 25/26 models and inferred monophyletic Porifera only under the LG4M model (a protein mixture model). For the Philippe 2009 and Pick 2010 datasets, paraphyletic Porifera was inferred only under amino-acid exchange rate matrices and protein mixture models, with monophyletic Porifera inferred under PM and PMSF models. There was no clear trend in Porifera clade topology for the Moroz 2014 dataset.

Each of the 364 trees had a high ultrafast bootstrap support value of 100 at the key node in the phylogenetic tree, i.e., the node separating the first metazoan clade to diverge and all other metazoan taxa.

3.4.3 MAST weights show support for multiple trees within a mixture

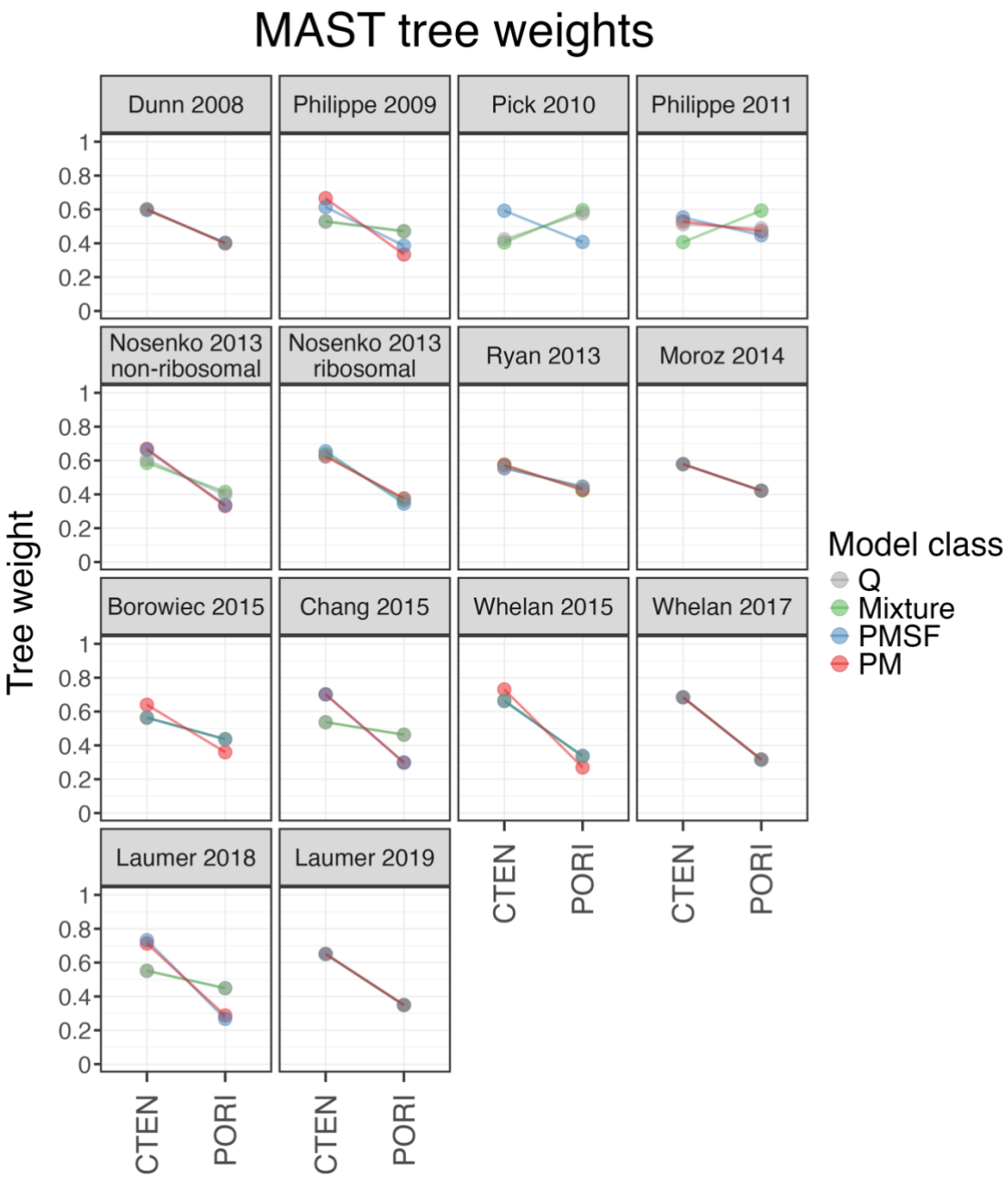


Figure 19: 2-tree MAST model tree weights for the Ctenophora-sister and Porifera-sister hypotheses, for 14 phylogenetic datasets and 4 classes of substitution model.

I applied the MAST model to each dataset with either 2 or 5 hypothesis trees (Figure 19, Supplementary Figure 22, Supplementary Tables 8 and 9). I applied the 2-tree MAST model to all combinations of dataset and model class. When the MAST model was applied with just two trees (Ctenophora-sister and Porifera-sister) I found the weight for each tree is between 0.270 and 0.732 (Figure 19). The average weight for the Ctenophora-sister tree was 0.604 and the average weight for the Porifera-sister tree was 0.396 (Supplementary Table 8). In 12/14

datasets, the tree weight for Ctenophora-sister was higher than the tree weight for Porifera-sister. In two datasets, I found the tree with the highest support was dependent on model of substitution. For Pick 2010, the PMSF model had higher weight for Ctenophora-sister (0.5924), but the Mixture and Q model classes had higher weight for Porifera-sister (0.5951 and 0.5771 respectively). The 2-tree MAST model was computationally intractable for the Pick 2010 dataset. For the Philippe 2011 dataset, the Mixture model class had higher weight for Porifera-sister (0.5933), and all other model classes had higher weight for Ctenophora-sister (Q: 0.5128; PMSF: 0.5525; PM: 0.5287).

The 5-tree MAST model was run for all datasets with the PMSF, Mixture, and Q model classes. Across trees, weights for the 5-tree MAST model ranged from 0.014 to 0.6211 with an average of 0.212 (Supplementary Table 9). The tree with the biggest range in tree weights was Ctenophora-tree, with a range of values from 0.014 – 0.6211. The average weight for the five hypotheses were 0.24 (Ctenophora-sister), 0.176 (Porifera-sister), 0.221 (Ctenophora+Porifera-sister), 0.244 (Ctenophora-sister, paraphyletic Porifera) and 0.182 (Porifera-sister, paraphyletic Porifera). There was more variation in the weight of the Ctenophora-sister tree (Supplementary Figure 22), with a range in Ctenophora-sister tree weights of 0.0568 – 0.6211. Variation in tree weights was not consistently linked to the model of substitution.

3.4.4 Most trees are rejected by the AU test

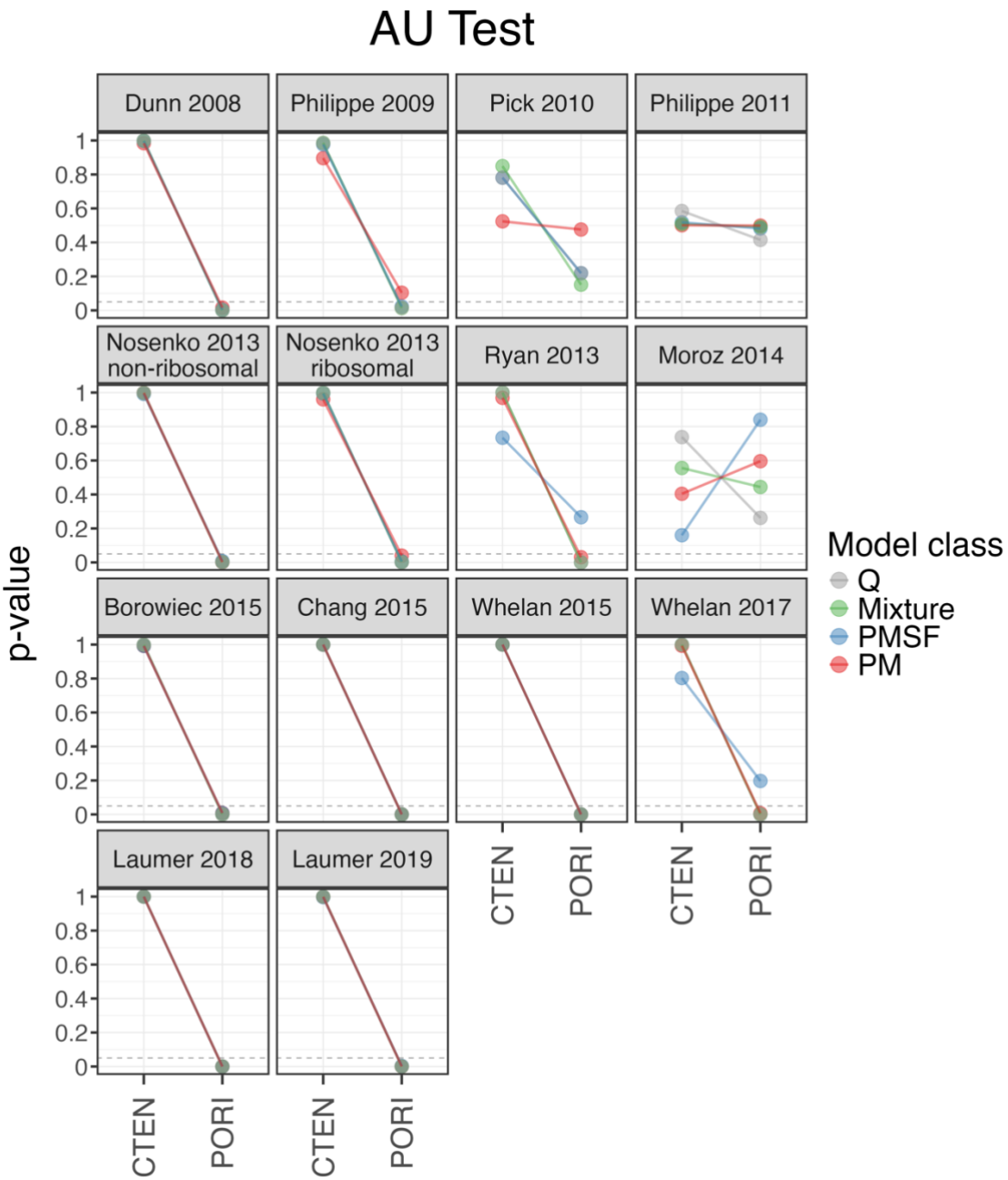


Figure 20: Examining support for the Ctenophora-sister and Porifera-sister hypothesis under a single-tree model, for 14 empirical datasets and 4 model classes.

Each point represents a p-values from the AU test. The grey dashed line indicates the statistical significance threshold of p<0.05. Any point under that line is rejected by the AU test.

I applied the AU test to the trees in the 2-tree and 5-tree models (Figure 20, Supplementary Figure 23, Supplementary Tables 10 and 11). When comparing results from the two tree AU test results trees (Figure 20, Supplementary Table 10) there were two cases: either the Porifera-sister tree was rejected (41/56 analyses), or neither tree was rejected (15/56 analyses). For three datasets (Pick 2010, Philippe 2011, Moroz 2014), neither the Ctenophora-

sister nor the Porifera-sister tree was rejected by the AU test. For the remaining 11 datasets, I observed the Porifera-sister tree rejected by the AU test for all 4 model classes in 8 datasets, and for 3 model classes in 3 datasets.

The AU test results were similar when comparing the full set of 5 evolutionary hypotheses (Supplementary Figure 23, Supplementary Table 11). The Ctenophora-sister tree was not rejected by any combination of dataset and model class. Only 5 combinations of dataset and model class did not have any trees rejected by the AU test: Pick 2010 Q model class; Philippe 2011 Mixture model class; and Moroz 2014 PM, PMSF, and Mixture model classes. These three datasets had the lowest proportion of trees rejected by the AU test: 3/20 for Pick 2010, 6/20 for Philippe 2011, and 2/20 for Moroz 2014. In all other combinations of model class and dataset, 1 or more trees were rejected by the AU test.

3.5 Discussion

In this chapter, I compared the fit of 1-tree, 2-tree, and 5-tree models on a wide range of previously published datasets, and using the full complement of available substitution models. I found that multi-tree models were preferred over single-tree models in 47/56 analyses (Table 5). In addition, when both 2- and 5-tree MAST models were applied, 5-tree models were preferred in the majority (29/41) of cases.

Almost all phylogenetic analyses aim to infer a single species tree that represents the majority of the evolutionary history of the sequences (Baum 2007). However, my results suggest that a single phylogenetic tree is an inadequate fit for metazoan datasets, regardless of the model of substitution. Due to the levels of conflicting signal in metazoan datasets, increasing the size of multiple sequence alignments by adding more genes or taxa will not resolve the metazoan phylogeny. This has previously been noted by Philippe et al. (2011b), who analysed existing metazoan phylogenetic datasets and showed that many genes were saturated or contained little phylogenetic signal. While I also identify substantial heterogeneous signal, my analysis suggests that the difficulty in tree inference is not due just to saturated sequences but reflects genuine phylogenetic signal. If a mixture of trees best explains the evolutionary history of the Metazoa, that would explain results from previous studies that found different model parameters or data filtering resulted in a completely different, highly supported topology (Francis and Canfield 2020; Pandey and Braun 2020; Li et al. 2021; Redmond and McLysaght 2021a; McCarthy et al. 2023). I show that different models of evolution result in different tree weights within the mixture model (Figure 19, Supplementary Figure 22). Inferring single maximum likelihood trees with different models of evolution will result in slightly higher support for one hypothesis, explaining some of the variation in empirical metazoan phylogenies.

Our findings impact the theoretical approach to downstream inferences, such as those investigating the development of complex traits. Given that a single tree does not explain the evolutionary history of Metazoa, these studies must account for heterogeneous evolutionary signal. Rather than using a large genomic dataset, these analyses could carefully select data based on the research question, for example by looking at genes known to be related to the nervous system. Similar approaches have previously been applied to empirical phylogenetic studies. Chen et al. (2015) found that a multiple sequence alignment for jawed vertebrates had substantial conflicting signal, which they managed by selected question-specific genes to resolve individual nodes of the phylogeny. My results could be combined with the approach of Chen et al. (2015) to estimate a mixture of different hypotheses selected for specific evolutionary questions using carefully selected genes.

The impact of model choice and model complexity on metazoan tree topology has been previously investigated (Whelan and Halanych 2017; Kapli and Telford 2020; Li et al. 2021; Redmond and McLysaght 2021a; Szánthó et al. 2023). In my chapter, I found that the majority of maximum likelihood trees inferred the Ctenophora-sister topology for 12/14 datasets. Only two datasets had different topologies consistently under different classes of models: Philippe 2011 and Moroz 2014. In both these datasets, the proportion of trees inferring the Ctenophora-sister topology decreased as model complexity increased. The results of these two datasets align with Redmond and McLysaght (2021a), who applied different models of evolution with increasing complexity to three metazoan datasets and found that branch support for Ctenophora-sister reduced as more complex (i.e., site heterogeneous) models of substitution were applied. I found that model choice does impact the inferred metazoan tree, but this appears to be due to heterogeneous evolutionary processes rather than model choice alone.

Results from the PMSF model class differed from other model classes, with half (7/14) of datasets preferring a single-tree model when the PMSF model was used to estimate trees and mixtures. The PMSF model explicitly accounts for heterogeneity in substitution rate, and was developed as a faster and less computationally-intensive alternative to profile mixture models (Wang et al. 2018). The PMSF model calculates amino acid profiles using a input model and a guide tree estimated from a simple model of evolution (e.g., LG) (Wang et al. 2018). My results suggest that the process of estimating the PMSF model biases the estimated amino acid profiles, resulting in a preference for the single-tree model. Similar biases have been identified in other phylogenetic models. Frandsen et al. (2015) found that some partitioning scheme algorithms included a bias towards the starting tree that was used to infer the partitioning scheme. When maximum likelihood trees were estimated with partitions using these algorithms, the trees inferred were more similar to the starting tree than expected by chance (Frandsen et al. 2015). A similar bias could be occurring during PMSF model estimation, which also relies on a starting tree. Further investigation into the impact and extent of this bias is required.

The MAST model allows me to make valid comparisons of the weights in different joint models, something which is not possible in single-tree models. Under a 5-tree model, 3/5 trees have monophyletic Porifera and 2/5 have paraphyletic Porifera (Figure 17). By summing the weights of the two categories of trees, I can explore support for the monophyly of the Porifera. I found that all 5-tree mixture models for datasets with more than 1 Porifera taxa had support for both monophyletic and paraphyletic Porifera, and that the combined weights of trees with monophyletic Porifera ranged from 0.372 (Laumer 2019 dataset, Mixture model) to 0.817 (Laumer 2019 dataset, PMSF model). My results suggest that the differences in Porifera

topology inferred in different phylogenetic studies reflect heterogeneous evolutionary signal. There is evidence for both paraphyletic (Sperling et al. 2007, 2009, 2010; Borchiellini et al. 2008; Erwin et al. 2011) and monophyletic Porifera (Erpenbeck and Wörheide 2007; Dohrmann et al. 2008; Philippe et al. 2009; Pick et al. 2010; Gazave et al. 2012; Ryan et al. 2013; Simion et al. 2017a; Whelan et al. 2017a; Laumer et al. 2018a) within the literature, although the majority of recent studies support monophyletic Porifera. Topology of the Porifera is dependent on the data included in an alignment, and sites with higher saturation support Ctenophora-sister whereas genes associated with translation support Porifera-sister (Nosenko et al. 2013a). My results suggest the Porifera clade has a conflicting evolutionary history best modelled by a mixture of trees. As a single tree is unable to represent this conflicting signal, different choices during tree inference such as choice of genes and substitution model could result in higher likelihood for either monophyletic or paraphyletic Porifera, resulting in the conflicting Porifera topologies inferred in previous studies.

My analysis found that all 364 maximum likelihood trees (estimated from 14 datasets with 26 models) had 100% ultra-fast bootstrap (UFB) support at the branch separating the first clade to diverge from all other animals. However, in 10/14 datasets, I found that the model of substitution impacted the tree topology, indicating that in these datasets branch support did not correlate with obtaining the true tree topology. My results concur with the simulation study of Kapli and Telford (2020), who simulated alignments with a site-heterogeneous model of evolution under both the Ctenophora-sister tree and the Porifera-sister tree, before estimating trees with both a site-homogeneous model and site-heterogeneous model. They found that when estimating trees with a site-homogeneous model from alignments simulated along the Porifera-sister, the Ctenophora-sister tree with high bootstrap support values was inferred 98% of the time. Given that metazoan species diverged on a deep evolutionary time scale and that the group underwent a rapid radiation, there are few orthologs present across the breadth of the metazoan taxa (Pett et al. 2019). This not only limits phylogenetic resolution but increases the difficulty of obtaining independent and identically distributed sites. Short recalcitrant branches have previously been shown to have high bootstrap support for conflicting topologies in empirical phylogenetic datasets (Chan et al. 2020; Roycroft et al. 2020). This occurs as bootstrap support is a measure of robustness of the data to perturbation, rather than a measure of the true evolutionary history (Holmes 2003; Simon 2022; Thomson and Brown 2022). Combined, the high bootstrap support for conflicting tree topologies within the metazoan tree reinforces that high bootstrap support is not an indication of whether the true evolutionary relationship has been recovered.

My analyses suggest that a larger proportion of the sites in each of the 14 datasets I examined tended to be assigned to the Ctenophora-sister topology than the Porifera-sister topology. For example, the 2-tree MAST models had higher average tree weights on for the Ctenophora-sister tree than for the Porifera-sister tree (0.604 and 0.396 respectively) For the 5-tree MAST model, there was less variation in tree weight for the 5 hypothesis trees (Figure 19, Supplementary Figure 22) with an average tree weight across dataset, model class and tree topology of 0.212. However, when applying the AU test for the trees in the 2-tree and 5-tree MAST model (Figure 20, Supplementary Figure 23), I found that the majority of Ctenophora-sister trees were not rejected by the AU test, and the majority of trees with other hypothesis topologies were. Finally, I found that under a single-tree maximum likelihood model (Figure 18), the most common tree topology estimated was Ctenophora-sister (317/364 trees or 87%). Taken together, my results suggest that most datasets contain substantial signal for both the Ctenophora-sister and Porifera-sister topologies, and that while both are required to adequately explain most datasets under most models of evolution, a larger fraction of most datasets is consistent with the Ctenophora-sister than the Porifera-sister hypothesis.

In this chapter, I applied the AU test to the trees in the 2-tree and 5-tree models. The AU test is often applied to assess the adequacy of phylogenetic trees (James et al. 2006; Shulaev et al. 2011; Espeland et al. 2018; Hughes et al. 2018; Zhang et al. 2018c; Coleman et al. 2021). The AU test assumes that a single tree is the best fit for an alignment and that the single tree is the maximum likelihood tree for that alignment (Shimodaira 2002; Schmidt 2009). I found that the Ctenophora-sister tree with monophyletic Porifera was rarely rejected, but the other four tree topologies were often rejected. As the AU test assumes the evolutionary history of a given alignment is treelike, the AU test cannot determine whether any single tree is sufficient to describe the evolutionary history of an alignment. I applied a statistical approach that simultaneously assessed tree adequacy (by comparing the BIC of multiple single-tree models) and the validity of the treelikeness assumption (by comparing the BIC of single-tree and multi-tree models). My results show even if a tree represents a portion of phylogenetic signal detected by the MAST model, it can still be rejected by the AU test. I recommend results of the AU test are carefully interpreted within the context that the test was designed, and that alternative methods of assessing phylogenetic signal are applied to datasets with substantial heterogeneous phylogenetic signal.

The results of this chapter do not necessarily suggest that the Metazoa have undergone non-treelike evolution. Instead, my results suggest that studies inferring a single phylogeny for the Metazoa are excluding substantial proportions of phylogenetic signal and present an approach that explicitly allows for the incorporation of this conflicting signal. These results are aligned

with previous work showing substantial heterogeneous phylogenetic signal within sequence alignments. Previous studies of the metazoan phylogeny have identified different evolutionary histories within different regions of the genome. Protein structural environment has been shown to impact metazoan tree topology, with residues on the surface of proteins supporting the Ctenophora-sister hypothesis and buried residues supporting the Porifera-sister hypothesis (Pandey and Braun 2020). The support for different hypotheses of evolution also varies between protein environments, with equal support for minority trees in buried sites or sites in sheet and coil environments, but unequal support for minority trees in exposed sites resulting in variation in support for different tree topologies under different models of substitution (Pandey and Braun 2021). Additionally, sites with high saturation correlate with the Ctenophora-sister hypothesis, whereas genes involved in translation support the Porifera-sister hypothesis (Nosenko et al. 2013a). These results are consistent, as exposed sites evolve faster than buried sites and are more likely to be saturated, whereas housekeeping genes are conserved and evolve slowly (Nosenko et al. 2013a; Pandey and Braun 2021). The MAST model provides a new process-agnostic approach to investigate different evolutionary histories within a single alignment, which model the observed complex patterns of evolution seen in empirical data sets. For example, MAST allows researchers to explicitly include different evolutionary histories for different protein structural environments. MAST ties into the broader movement within phylogenetics of assessing whether a single tree can adequately represent complex evolutionary histories, and can be used in conjunction with other tools to explore conflicting or heterogeneous phylogenetic signal.

Comparing the original studies for the datasets analysed in this chapter, there were four approaches to maximum likelihood analysis. First, applying a single concatenated model (such as GTR+G, LG+G+I, or RTRev+G+F) (Dunn et al. 2008; Hejnol et al. 2009; Nosenko et al. 2013a; Ryan et al. 2013; Moroz et al. 2014; Chang et al. 2015). Second, applying a partition model (Borowiec et al. 2015; Whelan et al. 2015b, 2017a). Third, applying a C20 or C60 model (Laumer et al. 2018a). Fourth, applying a PMSF approximation of a C60 model (Laumer et al. 2019a). The MAST results from my study allow comparison with the original papers for three out of four of those approaches. One limitation of my chapter is the application only of concatenated maximum likelihood models, in contrast to previous studies that applied partition models (Redmond and McLysaght 2021a) or CAT models (Whelan and Halanych 2017; Li et al. 2021). This limitation was necessary, as the MAST model does not support partitioned analyses (Wong et al. 2024). However, when interpreting the MAST results in this chapter, the inadequacy of a single concatenated alignment to represent metazoan loci must be considered. By using MAST with a concatenated model, my aim was to expand previous

concatenated analyses of these datasets by estimating a mixture of a set number of pre-defined trees, instead of inferring a single tree from the alignment. The limitation of this approach is the difficulty in untangling the adequacy of a single tree from the adequacy of a single (perhaps incorrect) evolutionary model.

While my chapter does not include a partitioned analysis, it does include the site-heterogeneous models that have previously been found to best fit metazoan multiple sequence alignments. Redmond and McLysaght (2021a) tested 3 metazoan datasets with the same classes of models that I included in my analysis: site-homogeneous models (e.g., Dayhoff, WAG, LG), multi-matrix models (e.g., EHO, UL3), multi-profile models with Poisson matrix (e.g., C20, C60) and multi-profile models with non-Poisson matrix (e.g., LG+C60). Site-heterogeneous models consistently had the best fit, with most genes in all 3 datasets preferring multi-profile models with non-Poisson matrices (Redmond and McLysaght, 2021a). Similarly, Kapli and Telford (2020) compared the fit of site-homogeneous and site-heterogeneous models for 3 metazoan datasets and found site-homogeneous models consistently had better fit. My analyses include both mixture models and profile mixture models, and both these model classes preferred multi-tree models over single-tree models. Given these results, a partitioned model is unlikely to change the inferences drawn from my results. Previous studies have investigated the impact of partitioning schemes with different classes of models for metazoan datasets (Redmond and McLysaght 2021a), and these results found that site-heterogeneous models with non-Poisson matrices (e.g., C60+LG) fit best for metazoan alignments. MAST already includes a computationally intensive optimization process, and adding a partition scheme will increase the number of simultaneous parameters to infer, particularly if different trees in the mixture have different partition models. Future development of the MAST model may enable combining partitioning schemes, site-heterogeneous substitution models, and the mixture of trees approach.

This chapter aligns with the broader movement assessing systematic bias within the metazoan tree, and within phylogenetics more generally. The animal phylogeny is particularly difficult to resolve, due to the combination of distantly related taxa, rapid radiation causing short branches between clades, variation in evolutionary rates, and long timescales (more than 500 million years) (King and Rokas 2017). Previous studies have investigated a range of factors to improve resolution in this tree, including improving orthology of genes (Pett et al. 2019; McCarthy et al. 2023); models of sequence evolution (Whelan and Halanych 2017; Kapli and Telford 2020; Li et al. 2021; Redmond and McLysaght 2021a); compositional heterogeneity (Szánthó et al. 2023); investigating the signal at different protein structural environments (Pandey and Braun 2020, 2021); long-branch attraction (Kapli and Telford 2020); and support

for each topology at a site-by-site basis (Shen et al. 2017). In the majority of these studies, both Ctenophora-sister and Porifera-sister are estimated under different analysis parameters. While slight changes in analysis choice have been shown to change the tree topology obtained for this group, many previous phylogenetic studies into the Metazoa propose support for either Ctenophora-sister or Porifera-sister. The complexity of resolving the relationships appears to stem from underlying phylogenetic signal for both Ctenophora-sister and Porifera-sister. When the metazoan evolutionary history is restricted to a single tree, the treelike model of evolution will be unable to account for a substantial proportion of phylogenetic signal. With current methods unable to decisively resolve this tree, I suggest that downstream inferences and analyses should consider the implications of both evolutionary histories.

3.6 Data availability

The alignments used in this chapter are available from either the original publication of each dataset, or from the repository of a later reanalysis (see Table 4). All scripts to replicate my analyses are available at the GitHub repository https://github.com/caitlinch/metazoan-mixtures. All materials necessary to replicate my analyses are available from the Figshare repository https://doi.org/10.6084/m9.figshare.26087386, including alignments, ML trees, constrained trees, MAST model output files, AU test output files, and input/output csv files.

3.7 Acknowledgements

The authors thank to Maja Adamska, Barbara Holland, and Eleonora Rossi for helpful feedback throughout this project. This study was funded by an Australian Government Research Training Program scholarship (to CC).

3.8 Supplementary Tables

Supplementary Table 4: Summary of results from previous analyses of 14 empirical phylogenetic datasets.

Matrix column lists first author and year of publication for each matrix. Two matrices are included from the Nosenko et al. (2013a), so I also note whether each Nosenko 2013 matrix is ribosomal or non-ribosomal. Citations for each manuscript and repository are listed in Table 4.

Original manuscript conclusions were drawn from the main phylogeny estimated using that matrix. Where possible, I selected the matrix used to estimate the main phylogenetic figure in each study. See Methods for detailed descriptions of matrix selection.

Sister to all other metazoans lists the first monophyletic clade to diverge from all other animals. Porifera topology notes whether the sponge clade was inferred as Monophyletic or Paraphyletic, or whether only a single Porifera taxon was present. Ctenophora and Cnidaria monophyletic column denotes whether the phylogeny inferred a single monophyletic clade including Ctenophora and Cnidaria (True or False).

Matrix	Previously published datasets incorporated	Original manuscript conclusions		
		Sister to all other metazoans	Porifera topology	Ctenophora and Cnidaria monophyletic
Dunn 2008	-	Ctenophora	Monophyletic	False
Philippe 2009	(Baurain et al. 2007; Lartillot and Philippe 2008)	Porifera	Monophyletic	True
Pick 2010	(Dunn et al. 2008)	Porifera	Monophyletic	False
Philippe 2011	(Dunn et al. 2008; Philippe et al. 2009)	Porifera	Monophyletic	False
Nosenko 2013 non-ribosomal	-	Ctenophora	Paraphyletic	False
Nosenko 2013 ribosomal	-	Placozoa+Porifera	Monophyletic	True
Ryan 2013	(Hejnol et al. 2009; Srivastava et al. 2010)	Ctenophora	Monophyletic	False
Moroz 2014	-	Ctenophora	Monophyletic	False
Borowiec 2015	-	Ctenophora	One taxon	False
Chang 2015	(Philippe et al. 2011a)	Ctenophora	Monophyletic	False
Whelan 2015	-	Ctenophora	Monophyletic	False
Whelan 2017	-	Ctenophora	Monophyletic	False
Laumer 2018	-	Ctenophora	Monophyletic	False
Laumer 2019	-	Ctenophora	Monophyletic	False

Supplementary Table 5: Summary of maximum likelihood tree topology and Porifera clade topology for 364 maximum likelihood trees (from 14 empirical phylogenetic datasets with 26 models of sequence evolution).

Topology columns denote which clade diverged first: CTEN (Ctenophora), PORI (Porifera), CTEN+PORI (the monophyletic clade containing both Ctenophora and Porifera) or RADIATA (the combined clades of Placozoa, Porifera, Ctenophora and Cnidaria). UFB value for this divergence was 100 for all trees (all datasets and all substitution models).

Porifera topology denotes the topology of the Porifera clade: either monophyletic (MONO) or paraphyletic (PARA). Two datasets contained a single Porifera taxa, in which case the Porifera topology was marked ONE.

Ctenophora and Cnidaria topology columns denotes the relationship between the Ctenophora and Cnidaria clades. I classified these potential relationships into two classes: either these clades could form a monophyletic clade (MONO), or they could have any other relationship (PARA).

Dataset	Topology				Porifera topology			Ctenophora and Cnidaria topology	
	CTEN	PORI	CTEN+ PORI	RADIATA	MONO	PARA	ONE	MONO	PARA
Dunn 2008	26	0	0	0	0	0	26	0	26
Philippe 2009	26	0	0	0	12	14	0	0	26
Pick 2010	23	3	0	0	13	13	0	0	25
Philippe 2011	3	3	16	4	26	0	0	3	23
Nosenko 2013 non-ribosomal	26	0	0	0	26	0	0	0	26
Nosenko 2013 ribosomal	20	0	4	2	1	25	0	2	24
Ryan 2013	26	0	0	0	26	0	0	0	26
Moroz 2014	11	15	0	0	11	15	0	0	26
Borowiec 2015	26	0	0	0	0	0	26	0	26
Chang 2015	26	0	0	0	26	0	0	0	26
Whelan 2015	26	0	0	0	26	0	0	0	26
Whelan 2017	26	0	0	0	26	0	0	0	26
Laumer 2018	26	0	0	0	26	0	0	0	26
Laumer 2019	26	0	0	0	26	0	0	0	26

Supplementary Table 6: Output topology for each combination of model and sequence alignment.

Each branch defining the split between the first clade to diverge and all other metazoans clades has 100% ultrafast bootstrap (UFB) support. CTEN denotes output topology has Ctenophora as sister to all other metazoans; PORI denotes Porifera-sister; CTEN+PORI denotes a clade of Ctenophora and Porifera as the sister to all other metazoans (note: in this case, it is not necessary that Porifera is a monophyletic clade); and RADIATA denotes the sister clade consisted of the Ctenophora, Cnidaria, Porifera (and Placozoa, if present) clades.

Initial Model	Datasets						
	Dunn 2008	Philippe 2009	Pick 2010	Philippe 2011	Nosenko 2013 non-ribosomal	Nosenko 2013 ribosomal	Ryan 2013
PMSF Poisson+C20	CTEN	CTEN	CTEN	PORI	CTEN	CTEN	CTEN
PMSF Poisson+C60	CTEN	CTEN	CTEN	PORI	CTEN	CTEN	CTEN
PMSF LG+C20	CTEN	CTEN	CTEN	RADIATA	CTEN	CTEN	CTEN
PMSF LG+C60	CTEN	CTEN	CTEN	RADIATA	CTEN	CTEN	CTEN
Poisson+C20	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN
Poisson+C60	CTEN	CTEN	CTEN	PORI	CTEN	CTEN	CTEN
LG+C20	CTEN	CTEN	CTEN	RADIATA	CTEN	RADIATA	CTEN
LG+C60	CTEN	CTEN	PORI	RADIATA	CTEN	RADIATA	CTEN
CF4	CTEN	CTEN	CTEN	CTEN+PORI	CTEN	CTEN+PORI	CTEN
EHO	CTEN	CTEN	CTEN	CTEN+PORI	CTEN	CTEN+PORI	CTEN
EX_EHO	CTEN	CTEN	CTEN	CTEN+PORI	CTEN	CTEN+PORI	CTEN
EX2	CTEN	CTEN	CTEN	CTEN+PORI	CTEN	CTEN	CTEN
EX3	CTEN	CTEN	CTEN	CTEN+PORI	CTEN	CTEN	CTEN
GTR20	CTEN	CTEN	CTEN	CTEN+PORI	CTEN	CTEN	CTEN
JTT	CTEN	CTEN	CTEN	CTEN+PORI	CTEN	CTEN	CTEN
JTTDCMut	CTEN	CTEN	CTEN	CTEN+PORI	CTEN	CTEN	CTEN
LG	CTEN	CTEN	CTEN	CTEN+PORI	CTEN	CTEN	CTEN
LG4M	CTEN	CTEN	CTEN	CTEN+PORI	CTEN	CTEN	CTEN
mtZOA	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN
PMB	CTEN	CTEN	PORI	CTEN+PORI	CTEN	CTEN	CTEN
Poisson	CTEN	CTEN	CTEN	CTEN+PORI	CTEN	CTEN	CTEN
rtREV	CTEN	CTEN	CTEN	CTEN+PORI	CTEN	CTEN	CTEN
UL2	CTEN	CTEN	CTEN	CTEN+PORI	CTEN	CTEN+PORI	CTEN
UL3	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN
WAG	CTEN	CTEN	PORI	CTEN+PORI	CTEN	CTEN	CTEN
ModelFinder	CTEN	CTEN	CTEN	CTEN+PORI	CTEN	CTEN	CTEN

Supplementary Table 6 (continued): Output topology for each combination of model and sequence alignment.

Each branch defining the split between the first clade to diverge and all other metazoans clades has 100% ultrafast bootstrap (UFB) support. CTEN denotes output topology has Ctenophora as sister to all other metazoans; PORI denotes Porifera-sister; CTEN+PORI denotes a clade of Ctenophora and Porifera as the sister to all other metazoans (note: in this case, it is not necessary that Porifera is a monophyletic clade); and RADIATA denotes the sister clade consisted of the Ctenophora, Cnidaria, Porifera (and Placozoa, if present) clades.

Initial Model	Datasets						
	Moroz 2014	Borowiec 2015	Chang 2015	Whelan 2015	Whelan 2017	Laumer 2018	Laumer 2019
PMSF Poisson+C20	PORI	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN
PMSF Poisson+C60	PORI	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN
PMSF LG+C20	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN
PMSF LG+C60	PORI	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN
Poisson+C20	PORI	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN
Poisson+C60	PORI	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN
LG+C20	PORI	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN
LG+C60	PORI	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN
CF4	PORI	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN
EHO	PORI	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN
EX_EHO	PORI	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN
EX2	PORI	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN
EX3	PORI	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN
GTR20	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN
JTT	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN
JTTDCMut	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN
LG	PORI	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN
LG4M	PORI	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN
mtZOA	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN
PMB	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN
Poisson	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN
rtREV	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN
UL2	PORI	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN
UL3	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN
WAG	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN
ModelFinder	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN	CTEN

Supplementary Table 7: Porifera topology for each combination of model and sequence alignment.

MONO denotes that Porifera taxa form a monophyletic clade; PARA denotes that Porifera taxa form a paraphyletic clade; and ONE indicates that the alignment contained only one Porifera taxa.

Initial Model	Datasets						
	Dunn 2008	Philippe 2009	Pick 2010	Philippe 2011	Nosenko 2013 non-ribosomal	Nosenko 2013 ribosomal	Ryan 2013
PMSF Poisson+C20	ONE	MONO	MONO	MONO	MONO	PARA	MONO
PMSF Poisson+C60	ONE	MONO	MONO	MONO	MONO	PARA	MONO
PMSF LG+C20	ONE	MONO	MONO	MONO	MONO	PARA	MONO
PMSF LG+C60	ONE	MONO	MONO	MONO	MONO	PARA	MONO
Poisson+C20	ONE	MONO	MONO	MONO	MONO	PARA	MONO
Poisson+C60	ONE	MONO	MONO	MONO	MONO	PARA	MONO
LG+C20	ONE	MONO	MONO	MONO	MONO	PARA	MONO
LG+C60	ONE	MONO	MONO	MONO	MONO	PARA	MONO
CF4	ONE	PARA	PARA	MONO	MONO	PARA	MONO
EHO	ONE	PARA	PARA	MONO	MONO	PARA	MONO
EX_EHO	ONE	PARA	PARA	MONO	MONO	PARA	MONO
EX2	ONE	PARA	PARA	MONO	MONO	PARA	MONO
EX3	ONE	PARA	MONO	MONO	MONO	PARA	MONO
GTR20	ONE	PARA	PARA	MONO	MONO	PARA	MONO
JTT	ONE	PARA	PARA	MONO	MONO	PARA	MONO
JTTDCMut	ONE	PARA	PARA	MONO	MONO	PARA	MONO
LG	ONE	PARA	PARA	MONO	MONO	PARA	MONO
LG4M	ONE	MONO	MONO	MONO	MONO	MONO	MONO
mtZOA	ONE	MONO	PARA	MONO	MONO	PARA	MONO
PMB	ONE	MONO	PARA	MONO	MONO	PARA	MONO
Poisson	ONE	PARA	MONO	MONO	MONO	PARA	MONO
rtREV	ONE	PARA	PARA	MONO	MONO	PARA	MONO
UL2	ONE	PARA	MONO	MONO	MONO	PARA	MONO
UL3	ONE	MONO	MONO	MONO	MONO	PARA	MONO
WAG	ONE	PARA	PARA	MONO	MONO	PARA	MONO
ModelFinder	ONE	PARA	PARA	MONO	MONO	PARA	MONO

Supplementary Table 7 (continued): Porifera topology for each combination of model and sequence alignment.

MONO denotes that Porifera taxa form a monophyletic clade; PARA denotes that Porifera taxa form a paraphyletic clade; and ONE indicates that the alignment contained only one Porifera taxa.

Initial Model	Datasets						
	Moroz 2014	Borowiec 2015	Chang 2015	Whelan 2015	Whelan 2017	Laumer 2018	Laumer 2019
PMSF Poisson+C20	MONO	ONE	MONO	MONO	MONO	MONO	MONO
PMSF Poisson+C60	MONO	ONE	MONO	MONO	MONO	MONO	MONO
PMSF LG+C20	MONO	ONE	MONO	MONO	MONO	MONO	MONO
PMSF LG+C60	PARA	ONE	MONO	MONO	MONO	MONO	MONO
Poisson+C20	PARA	ONE	MONO	MONO	MONO	MONO	MONO
Poisson+C60	PARA	ONE	MONO	MONO	MONO	MONO	MONO
LG+C20	PARA	ONE	MONO	MONO	MONO	MONO	MONO
LG+C60	PARA	ONE	MONO	MONO	MONO	MONO	MONO
CF4	PARA	ONE	MONO	MONO	MONO	MONO	MONO
EHO	PARA	ONE	MONO	MONO	MONO	MONO	MONO
EX_EHO	PARA	ONE	MONO	MONO	MONO	MONO	MONO
EX2	PARA	ONE	MONO	MONO	MONO	MONO	MONO
EX3	PARA	ONE	MONO	MONO	MONO	MONO	MONO
GTR20	MONO	ONE	MONO	MONO	MONO	MONO	MONO
JTT	MONO	ONE	MONO	MONO	MONO	MONO	MONO
JTTDCMut	PARA	ONE	MONO	MONO	MONO	MONO	MONO
LG	PARA	ONE	MONO	MONO	MONO	MONO	MONO
LG4M	PARA	ONE	MONO	MONO	MONO	MONO	MONO
mtZOA	PARA	ONE	MONO	MONO	MONO	MONO	MONO
PMB	MONO	ONE	MONO	MONO	MONO	MONO	MONO
Poisson	MONO	ONE	MONO	MONO	MONO	MONO	MONO
rtREV	MONO	ONE	MONO	MONO	MONO	MONO	MONO
UL2	PARA	ONE	MONO	MONO	MONO	MONO	MONO
UL3	MONO	ONE	MONO	MONO	MONO	MONO	MONO
WAG	MONO	ONE	MONO	MONO	MONO	MONO	MONO
ModelFinder	MONO	ONE	MONO	MONO	MONO	MONO	MONO

Supplementary Table 8: Tree weights for each dataset and model class under 2-tree MAST model. Best model determined as the model in each category with the lowest BIC score. The tree weights are the proportion of each tree in the tree mixture. For the overall topology of the five hypothesis trees, see Figure 1. Datasets including only one sponge taxon have NA for Tree 4 and Tree 5, as inclusion of the paraphyletic sponge hypotheses required multiple sponge taxa. The MAST model was not run for PM class models due to computational constraints.

Dataset	Model class	Best model	Tree weight		Largest weight
			Tree 1 CTEN	Tree 2 PORI	
Dunn 2008	PM	LG+C60+F+R7	0.6001	0.3999	CTEN
	PMSF	Poisson+C60+F+R4	0.5965	0.4035	CTEN
	Mixture	UL3+R7	0.5999	0.4001	CTEN
	Q	GTR20+F+R7	0.59994	0.4006	CTEN
Philippe 2009	PM	LG+C60+F+R7	0.6660	0.3340	CTEN
	PMSF	Poisson+C60+F+R4	0.6141	0.3859	CTEN
	Mixture	UL3+R7	0.5277	0.4723	CTEN
	Q	GTR20+F+R7	0.5305	0.4695	CTEN
Pick 2010	PMSF	Poisson+C60+F+R4	0.5924	0.4076	CTEN
	Mixture	UL3+R8	0.4049	0.5951	PORI
	Q	GTR20+F+R8	0.4229	0.5771	PORI
Philippe 2011	PM	LG+C60+F+R9	0.5287	0.4713	CTEN
	PMSF	Poisson+C60+F+R4	0.5525	0.4475	CTEN
	Mixture	UL3+R8	0.4067	0.5933	PORI
	Q	GTR20+F+R8	0.5128	0.4872	CTEN
Nosenko 2013 non-ribosomal	PM	LG+C60+F+R7	0.6688	0.3312	CTEN
	PMSF	Poisson+C60+F+R4	0.6625	0.3375	CTEN
	Mixture	UL3+R7	0.5856	0.4144	CTEN
	Q	GTR20+F+R6	0.6037	0.3963	CTEN
Nosenko 2013 ribosomal	PM	LG+C60+F+R6	0.6239	0.3761	CTEN
	PMSF	Poisson+C60+F+R4	0.6545	0.3455	CTEN
	Mixture	UL3+R7	0.6376	0.3624	CTEN
	Q	GTR20+F+R7	0.6268	0.3732	CTEN
Ryan 2013	PM	LG+C60+F+R7	0.5715	0.4285	CTEN
	PMSF	Poisson+C60+F+R4	0.5559	0.4441	CTEN
	Mixture	LG4M+R6	0.5769	0.4231	CTEN
	Q	GTR20+F+R6	0.5533	0.4467	CTEN
Moroz 2014	PM	LG+C60+F+R5	0.5784	0.4216	CTEN
	PMSF	Poisson+C60+F+R4	0.5784	0.4216	CTEN
	Mixture	LG4M+R5	0.5784	0.4216	CTEN
	Q	GTR20+F+R5	0.5785	0.4215	CTEN
Borowiec 2015	PM	LG+C60+F+R7	0.6403	0.3597	CTEN
	PMSF	Poisson+C60+F+R4	0.5639	0.4361	CTEN
	Mixture	LG4M+R6	0.5637	0.4363	CTEN
	Q	GTR20+F+R6	0.5626	0.4374	CTEN
Chang 2015	PM	LG+C60+F+R9	0.7016	0.2984	CTEN
	PMSF	LG+C60+F+R4	0.7016	0.2984	CTEN
	Mixture	UL3+R9	0.5369	0.4631	CTEN
	Q	GTR20+F+R8	0.5370	0.4630	CTEN
Whelan 2015	PM	LG+C60+F+R8	0.7304	0.2696	CTEN
	PMSF	Poisson+C60+F+R4	0.6615	0.3385	CTEN
	Mixture	LG4M+R8	0.6654	0.3346	CTEN
	Q	GTR20+F+R9	0.6650	0.3350	CTEN
Whelan 2017	PM	LG+C60+F+R7	0.6840	0.3160	CTEN
	PMSF	Poisson+C60+F+R4	0.6841	0.3159	CTEN
	Mixture	LG4M+R8	0.6846	0.3154	CTEN
	Q	GTR20+F+R7	0.6832	0.3168	CTEN

Supplementary Table 8 (continued): Tree weights for each dataset and model class under 2-tree MAST model. Best model determined as the model in each category with the lowest BIC score. The tree weights are the proportion of each tree in the tree mixture. For the overall topology of the five hypothesis trees, see Figure 1. Datasets including only one sponge taxon have NA for Tree 4 and Tree 5, as inclusion of the paraphyletic sponge hypotheses required multiple sponge taxa. The MAST model was not run for PM class models due to computational constraints.

Laumer 2018	PM	LG+C20+F+R8	0.7127	0.2873	CTEN
	PMSF	Poisson+C60+F+R4	0.7322	0.2678	CTEN
	Mixture	LG4M+R7	0.5514	0.4486	CTEN
	Q	GTR20+F+R7	0.5515	0.4485	CTEN
Laumer 2019	PM	LG+C60+F+R7	0.6509	0.3491	CTEN
	PMSF	Poisson+C60+F+R4	0.6509	0.3491	CTEN
	Mixture	LG4M+R6	0.6509	0.3491	CTEN
	Q	GTR20+F+R7	0.6509	0.3491	CTEN

Supplementary Table 9: Tree weights for each dataset and model class under 5-tree MAST model.

Best model determined as the model in each category with the lowest BIC score. The tree weights are the proportion of each tree in the tree mixture. For the overall topology of the five hypothesis trees, see Figure 1. Datasets including only one sponge taxon have NA for Tree 4 and Tree 5, as inclusion of the paraphyletic sponge hypotheses required multiple sponge taxa. The MAST model was not run for PM class models due to computational constraints. The tree with the largest weight is noted for each MAST model.

Dataset	Model class	Best model	Tree weight					Topology with largest tree weight
			Tree 1	Tree 2	Tree 3	Tree 4	Tree 5	
Dunn 2008	PMSF	Poisson+C60+F+R4	0.3953	0.2675	0.3372	NA	NA	CTEN
	Mixture	UL3+R7	0.3952	0.2634	0.3414	NA	NA	CTEN
	Q	GTR+F+R7	0.3952	0.2635	0.3413	NA	NA	CTEN
Philippe 2009	PMSF	Poisson+C60+F+R4	0.2011	0.1557	0.198	0.221	0.2241	PORI
	Mixture	UL3+R7	0.1874	0.1942	0.2001	0.2151	0.2032	CTEN
	Q	GTR20+F+R7	0.3064	0.1717	0.2031	0.1569	0.1620	CTEN
Pick 2010	PMSF	Poisson+C60+F+R4	0.14	0.0665	0.39	0.063	0.3404	CTEN+PORI
	Mixture	UL3+R8	0.072	0.5286	0.0865	0.109	0.204	PORI
	Q	GTR20+F+R8	0.1687	0.1747	0.2567	0.1961	0.2038	CTEN+PORI
Philippe 2011	PMSF	Poisson+C60+F+R4	0.2043	0.1504	0.2912	0.1792	0.1749	CTEN+PORI
	Mixture	UL3+R8	0.3032	0.129	0.1695	0.2008	0.1975	CTEN
	Q	GTR20+F+R8	0.2654	0.1260	0.0902	0.4164	0.1020	CTEN
Nosenko 2013 non-ribosomal	PMSF	Poisson+C60+F+R4	0.2058	0.1501	0.1711	0.2514	0.2216	CTEN
	Mixture	UL3+R7	0.2156	0.1499	0.2208	0.2171	0.1967	CTEN+PORI
	Q	GTR20+F+R6	0.2068	0.2003	0.1718	0.2175	0.2036	CTEN
Nosenko 2013 ribosomal	PMSF	Poisson+C60+F+R4	0.1731	0.1393	0.1775	0.2419	0.2682	CTEN
	Mixture	UL3+R7	0.1946	0.1387	0.2006	0.2549	0.2112	CTEN
	Q	GTR20+F+R7	0.1712	0.1393	0.3448	0.1849	0.1597	CTEN+PORI
Ryan 2013	PMSF	Poisson+C60+F+R4	0.1873	0.1641	0.2313	0.2195	0.1978	CTEN+PORI
	Mixture	LG4M+R6	0.191	0.1488	0.25	0.2216	0.1886	CTEN+PORI
	Q	GTR20+F+R6	0.1579	0.2043	0.2220	0.2611	0.1548	CTEN
Moroz 2014	PMSF	Poisson+C60+F+R4	0.21	0.1581	0.2178	0.2253	0.1887	CTEN
	Mixture	LG4M+R5	0.2101	0.158	0.2179	0.2253	0.1886	CTEN
	Q	GTR20+F+R5	0.2093	0.1571	0.2176	0.2266	0.1894	CTEN
Borowiec 2015	PMSF	Poisson+C60+F+R4	0.6211	0.1466	0.2323	NA	NA	CTEN
	Mixture	LG4M+R6	0.3932	0.3051	0.3017	NA	NA	CTEN
	Q	GTR20+F+R6	0.3906	0.3057	0.3037	NA	NA	CTEN
Chang 2015	PMSF	LG+C60+F+R4	0.2088	0.1215	0.1994	0.2827	0.1875	CTEN
	Mixture	UL3+R9	0.0568	0.204	0.2964	0.3305	0.1123	CTEN
	Q	GTR20+F+R8	0.0140	0.2228	0.2992	0.3614	0.1026	CTEN
Whelan 2015	PMSF	Poisson+C60+F+R4	0.1962	0.1321	0.1883	0.2717	0.2118	CTEN
	Mixture	LG4M+R8	0.1972	0.1322	0.1899	0.2694	0.2113	CTEN
	Q	GTR20+F+R9	0.1969	0.1326	0.1901	0.2694	0.2210	CTEN
Whelan 2017	PMSF	Poisson+C60+F+R4	0.2012	0.1324	0.176	0.272	0.2184	CTEN
	Mixture	LG4M+R8	0.201	0.1323	0.176	0.2724	0.2183	CTEN
	Q	GTR20+F+R7	1	1.1 E-04	2.1 E-55	5.1 E-09	1.9 E-60	CTEN
Laumer 2018	PMSF	Poisson+C60+F+R4	0.5559	0.144	0.1048	0.1386	0.0567	CTEN
	Mixture	LG4M+R7	0.212	0.1481	0.1752	0.2688	0.196	CTEN
	Q	GTR20+F+R7	0.2334	0.1373	0.1329	0.3235	0.1729	CTEN
Laumer 2019	PMSF	Poisson+C60+F+R4	0.4797	0.2072	0.1301	0.1495	0.0335	CTEN
	Mixture	LG4M+R6	0.1164	0.0799	0.1757	0.5366	0.0914	CTEN
	Q	GTR20+F+R7	0.1976	0.1237	0.2230	0.2984	0.1574	CTEN

Supplementary Table 10: AU test results for the two hypothesis trees used in the 2-tree MAST model.

For the overall topology of the two hypothesis trees, see Figure 1. The AU test columns denote the p-value for each hypothesis tree generated from the dataset and the best model, with p < 0.05 indicating a certain tree is rejected. Significant p-values are shown in bold.

Dataset	Model class	Best model	AU test p-value	
			Tree 1 Ctenophora-sister	Tree 2 Porifera-sister
Dunn 2008	PM	LG+C60+F+R7	0.984	**0.016**
	PMSF	Poisson+C60+F+R4	0.997	**0.0033**
	Mixture	UL3+R7	0.999	**0.0005**
	Q	GTR20+F+R7	1	**3.42E-04**
Philippe 2009	PM	LG+C60+F+R7	0.896	0.104
	PMSF	Poisson+C60+F+R4	0.977	**0.023**
	Mixture	UL3+R7	0.986	**0.0138**
	Q	GTR20+F+R7	0.984	**0.0163**
Pick 2010	PM	LG+C60+F+R9	0.524	0.476
	PMSF	Poisson+C60+F+R4	0.781	0.219
	Mixture	UL3+R8	0.849	0.151
	Q	GTR20+F+R8	0.781	0.219
Philippe 2011	PM	LG+C60+F+R9	0.501	0.499
	PMSF	Poisson+C60+F+R4	0.518	0.482
	Mixture	UL3+R8	0.509	0.491
	Q	GTR20+F+R8	0.585	0.415.
Nosenko 2013 non-ribosomal	PM	LG+C60+F+R7	0.998	**0.0017**
	PMSF	Poisson+C60+F+R4	0.992	**0.0078**
	Mixture	UL3+R7	0.998	**0.0017**
	Q	GTR20+F+R6	0.999	**6.70E-04**
Nosenko 2013 ribosomal	PM	LG+C60+F+R6	0.96	**0.0401**
	PMSF	Poisson+C60+F+R4	0.995	**0.0054**
	Mixture	UL3+R7	0.998	**0.0019**
	Q	GTR20+F+R7	1	**4.44E-05**
Ryan 2013	PM	LG+C60+F+R7	0.969	**0.0306**
	PMSF	Poisson+C60+F+R4	1	**1.18E-06**
	Mixture	LG4M+R6	0.734	0.266
	Q	GTR20+F+R6	1	**9.03E-06**
Moroz 2014	PM	LG+C60+F+R5	0.404	0.596
	PMSF	Poisson+C60+F+R4	0.556	0.444
	Mixture	LG4M+R5	0.16	0.84
	Q	GTR20+F+R5	0.738	0.262
Borowiec 2015	PM	LG+C60+F+R7	0.992	**0.0082**
	PMSF	Poisson+C60+F+R4	0.999	**0.0008**
	Mixture	LG4M+R6	0.992	**0.0079**
	Q	GTR20+F+R6	1	**4.32E-05**
Chang 2015	PM	LG+C60+F+R9	1	**1.15E-90**
	PMSF	LG+C60+F+R4	1	**2.84E-50**
	Mixture	UL3+R9	1	**2.53E-55**
	Q	GTR20+F+R8	0.998	**0.0020**
Whelan 2015	PM	LG+C60+F+R8	1	**6.69E-62**
	PMSF	Poisson+C60+F+R4	1	**7.18E-39**
	Mixture	LG4M+R8	1	**2.30E-35**
	Q	GTR20+F+R9	1	**7.01E-61**
Whelan 2017	PM	LG+C60+F+R7	0.993	**0.0067**
	PMSF	Poisson+C60+F+R4	0.803	0.197
	Mixture	LG4M+R8	1	**0.0001**
	Q	GTR20+F+R7	1	**1.54E-05**

Supplementary Table 10 (continued): AU test results for the two hypothesis trees used in the 2-tree MAST model.

For the overall topology of the two hypothesis trees, see Figure 1. The AU test columns denote the p-value for each hypothesis tree generated from the dataset and the best model, with p < 0.05 indicating a certain tree is rejected. Significant p-values are shown in bold.

Laumer 2018	PM	LG+C20+F+R8	1	**1.11E-05**
	PMSF	Poisson+C60+F+R4	1	**5.35E-06**
	Mixture	LG4M+R7	1	**0.0003**
	Q	GTR20+F+R7	1	**2.47E-05**
Laumer 2019	PM	LG+C60+F+R7	0.999	**0.0006**
	PMSF	Poisson+C60+F+R4	0.997	**0.0032**
	Mixture	LG4M+R6	1	**3.93E-05**
	Q	GTR20+F+R7	1	**1.67E-94**

Supplementary Table 11: AU test results for the five hypothesis trees used in the 5-tree MAST model.

For the overall topology of the two hypothesis trees, see Figure 1. The AU test columns denote the p-value for each hypothesis tree generated from the dataset and the best model, with p < 0.05 indicating a certain tree is rejected. Significant p-values are shown in bold. Datasets including only one sponge taxon have NA for Tree 4 and Tree 5, as inclusion of the paraphyletic sponge hypotheses required multiple sponge taxa.

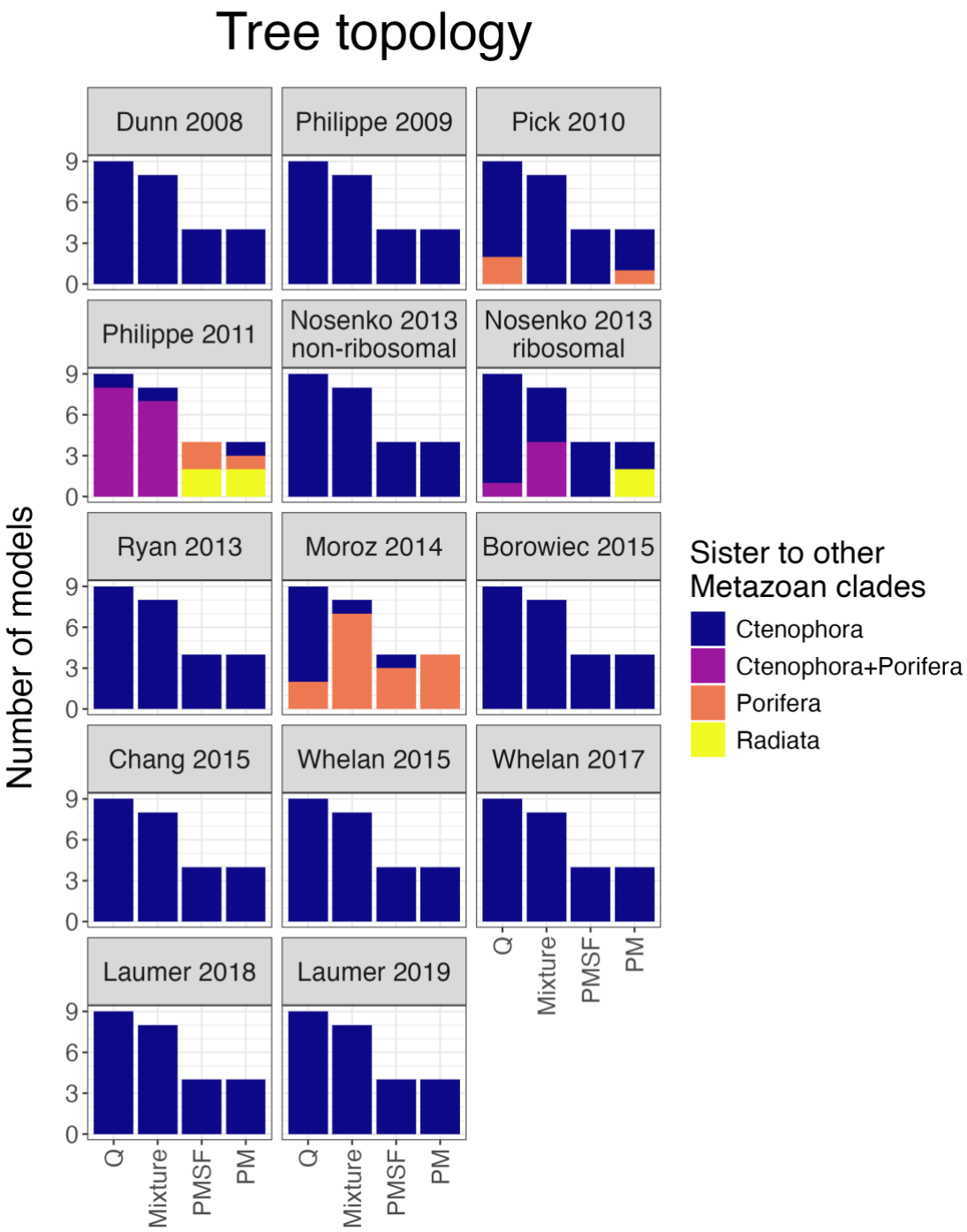
Dataset	Model class	Best model	AU test p-value				
			Tree 1	Tree 2	Tree 3	Tree 4	Tree 5
Dunn 2008	PM	LG+C60+F+R7	0.981	**0.0320**	**0.0297**	NA	NA
	PMSF	Poisson+C60+F+R4	0.996	**0.0049**	**0.0049**	NA	NA
	Mixture	UL3+R7	0.994	**0.0003**	**0.0072**	NA	NA
	Q	GTR20+F+R7	0.994	**5.79E-07**	**6.03E-03**	NA	NA
Philippe 2009	PM	LG+C60+F+R7	0.937	0.154	**0.0045**	0.0878	**0.0130**
	PMSF	Poisson+C60+F+R4	0.986	**0.0386**	**0.0149**	**0.0218**	**0.0075**
	Mixture	UL3+R7	0.960	**0.0145**	**0.0007**	0.0986	**0.0174**
	Q	GTR20+F+R7	0.963	**0.0183**	**0.0032**	0.076	**0.0011**
Pick 2010	PM	LG+C60+F+R9	0.645	0.589	0.202	**0.0113**	0.0808
	PMSF	Poisson+C60+F+R4	0.848	0.354	0.178	**0.0401**	0.0994
	Mixture	UL3+R8	0.925	0.189	**0.0002**	0.0834	0.0988
	Q	GTR20+F+R8	0.885	0.342	0.0605	0.148	0.0982
Philippe 2011	PM	LG+C60+F+R9	0.322	0.319	0.752	**0.0037**	**0.0077**
	PMSF	Poisson+C60+F+R4	0.582	0.567	0.351	**0.0098**	**0.0295**
	Mixture	UL3+R8	0.477	0.496	0.605	0.114	0.114
	Q	GTR20+F+R8	0.482	0.402	0.626	**0.0205**	**0.0126**
Nosenko 2013 non-ribosomal	PM	LG+C60+F+R7	0.983	**0.0050**	**0.0020**	**0.0276**	**0.0032**
	PMSF	Poisson+C60+F+R4	0.994	**0.0112**	**0.0010**	**0.0137**	**0.0016**
	Mixture	UL3+R7	0.927	**0.0015**	**0.0005**	0.0891	**0.0002**
	Q	GTR20+F+R6	0.947	**3.73E-04**	**2.42E-08**	0.0611	**6.50E-07**
Nosenko 2013 ribosomal	PM	LG+C60+F+R6	0.980	0.0949	**0.0433**	**0.0368**	**0.0217**
	PMSF	Poisson+C60+F+R4	0.998	**0.0065**	**0.0003**	**0.0069**	**0.0058**
	Mixture	UL3+R7	0.999	**0.0025**	**0.0021**	**0.0014**	**0.0025**
	Q	GTR20+F+R7	0.930	**9.93E-04**	0.1120	**0.0321**	**4.64E-03**
Ryan 2013	PM	LG+C60+F+R7	0.986	**0.0303**	**0.0007**	**0.0017**	**0.0049**
	PMSF	Poisson+C60+F+R4	0.835	0.3590	0.08120	0.08930	**0.0403**
	Mixture	LG4M+R6	0.997	**7.54E-08**	**7.93E-07**	**0.0026**	**1.68E-06**
	Q	GTR20+F+R6	0.935	**2.52E-05**	**5.80E-147**	0.0659	**6.12E-04**
Moroz 2014	PM	LG+C60+F+R5	0.538	0.677	0.486	0.146	0.255
	PMSF	Poisson+C60+F+R4	0.152	0.725	0.468	0.0776	0.418
	Mixture	LG4M+R5	0.695	0.625	0.121	0.138	0.266
	Q	GTR20+F+R5	0.845	0.376	**0.0295**	**0.0919**	0.172
Borowiec 2015	PM	LG+C60+F+R7	0.930	**0.0084**	0.0941	NA	NA
	PMSF	Poisson+C60+F+R4	0.969	**0.0074**	**0.0398**	NA	NA
	Mixture	LG4M+R6	0.995	**0.0018**	**0.0056**	NA	NA
	Q	GTR20+F+R6	0.998	**8.13E-03**	**0.0027**	NA	NA
Chang 2015	PM	LG+C60+F+R9	1	**2.47E-08**	**5.84E-95**	**6.56E-07**	**1.88E-05**
	PMSF	LG+C60+F+R4	1	**0.0003**	**5.49E-52**	**1.80E-43**	**4.16E-66**
	Mixture	UL3+R9	1	**2.82E-43**	**3.61E-08**	**0.0003**	**2.76E-42**
	Q	GTR20+F+R8	0.999	**3.06E-09**	**1.48E-12**	**6.59E-04**	**5.9E-38**
Whelan 2015	PM	LG+C60+F+R8	1	**3.60E-50**	**1.96E-55**	**3.26E-61**	**2.49E-110**
	PMSF	Poisson+C60+F+R4	1	**9.11E-08**	**8.02E-68**	**9.51E-07**	**0.0002**
	Mixture	LG4M+R8	1	**9.46E-116**	**1.02E-43**	**4.99E-131**	**1.22E-109**
	Q	GTR20+F+R9	1	**2.66E-83**	**5.59E-62**	**2.20E-06**	**1.25E-71**
Whelan 2017	PM	LG+C60+F+R7	0.992	**0.0084**	**1.75E-07**	**1.89E-99**	**0.0009**
	PMSF	Poisson+C60+F+R4	0.782	0.236	**0.0052**	**0.0005**	**0.0006**
	Mixture	LG4M+R8	1	**3.72E-10**	**2.49E-06**	**1.97E-05**	**2.28E-05**
	Q	GTR20+F+R7	1	**1.09E-04**	**2.11E-55**	**5.06E-09**	**1.94E-60**

Supplementary Table 11 (continued): AU test results for the five hypothesis trees used in the 5-tree MAST model.

For the overall topology of the two hypothesis trees, see Figure 1. The AU test columns denote the p-value for each hypothesis tree generated from the dataset and the best model, with p < 0.05 indicating a certain tree is rejected. Significant p-values are shown in bold. Datasets including only one sponge taxon have NA for Tree 4 and Tree 5, as inclusion of the paraphyletic sponge hypotheses required multiple sponge taxa.

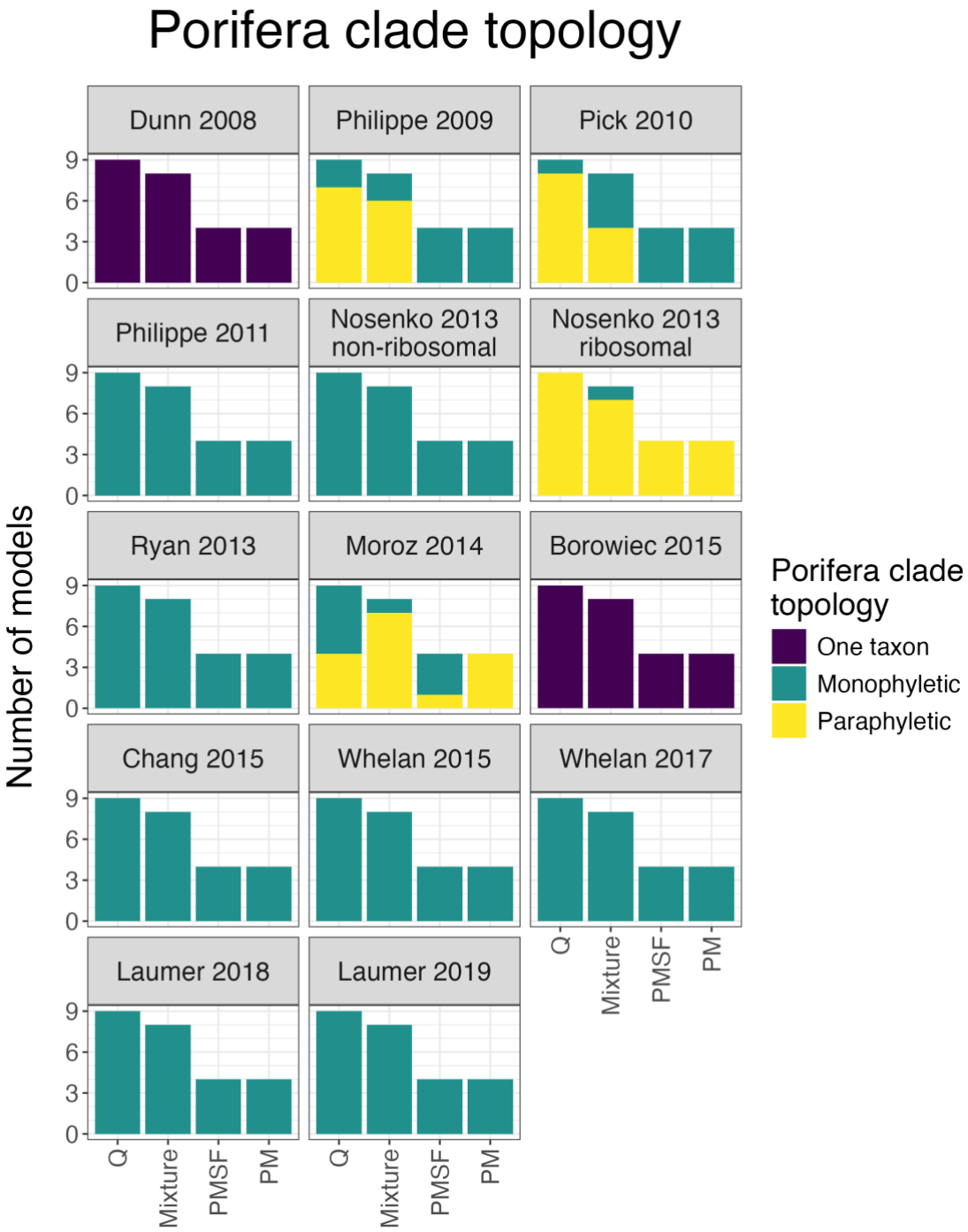
Laumer 2018	PM	LG+C20+F+R8	1	**4.10E-43**	**2.63E-05**	**4.60E-82**	**4.44E-05**
	PMSF	Poisson+C60+F+R4	1	**0.0001**	**9.94E-56**	**4.11E-113**	**2.07E-49**
	Mixture	LG4M+R7	1	**0.0001**	**4.52E-08**	**2.16E-08**	**2.25E-05**
	Q	GTR20+F+R7	1	**4.59E-07**	**3.11E-79**	**4.76E-04**	**2.52E-04**
Laumer 2019	PM	LG+C60+F+R7	1	**0.0005**	**0.0005**	**6.24E-08**	**1.12E-92**
	PMSF	Poisson+C60+F+R4	0.997	**0.0046**	**0.0004**	**4.79E-06**	**0.0011**
	Mixture	LG4M+R6	1	**0.0004**	**1.20E-05**	**2.19E-05**	**1.13E-70**
	Q	GTR20+F+R7	1	**2.40E-06**	**1.90E-73**	**8.04E-06**	**1.44E-05**

3.9 Supplementary Figures



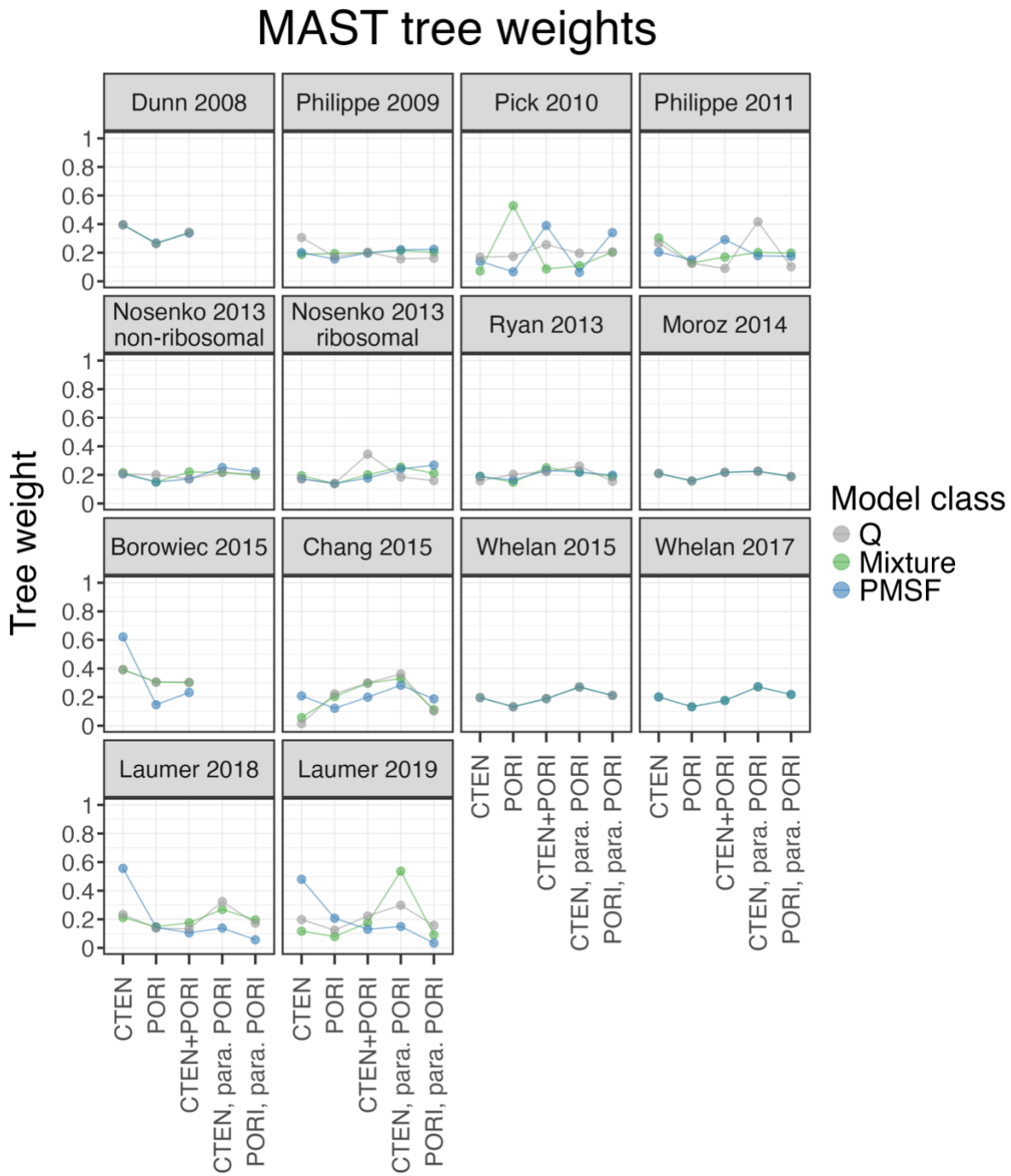
Supplementary Figure 20: Maximum likelihood tree topology for 350 trees estimated from different combinations of model of evolution and dataset.

Each bar shows the tree topologies obtained for a single dataset. For each dataset I calculated 25 different trees, each with a different model of evolution. Sister to all other metazoan clades indicates the first clade to diverge. The "Ctenophora+Porifera" clade denotes a single monophyletic clade consisting of both the Porifera and Ctenophora clades. The "Radiata" clade denotes a monophyletic clade consisting of Porifera, Ctenophora, Cnidaria and Placozoa (although Placozoa was not included in all datasets). Models are grouped on the x-axis by model class.



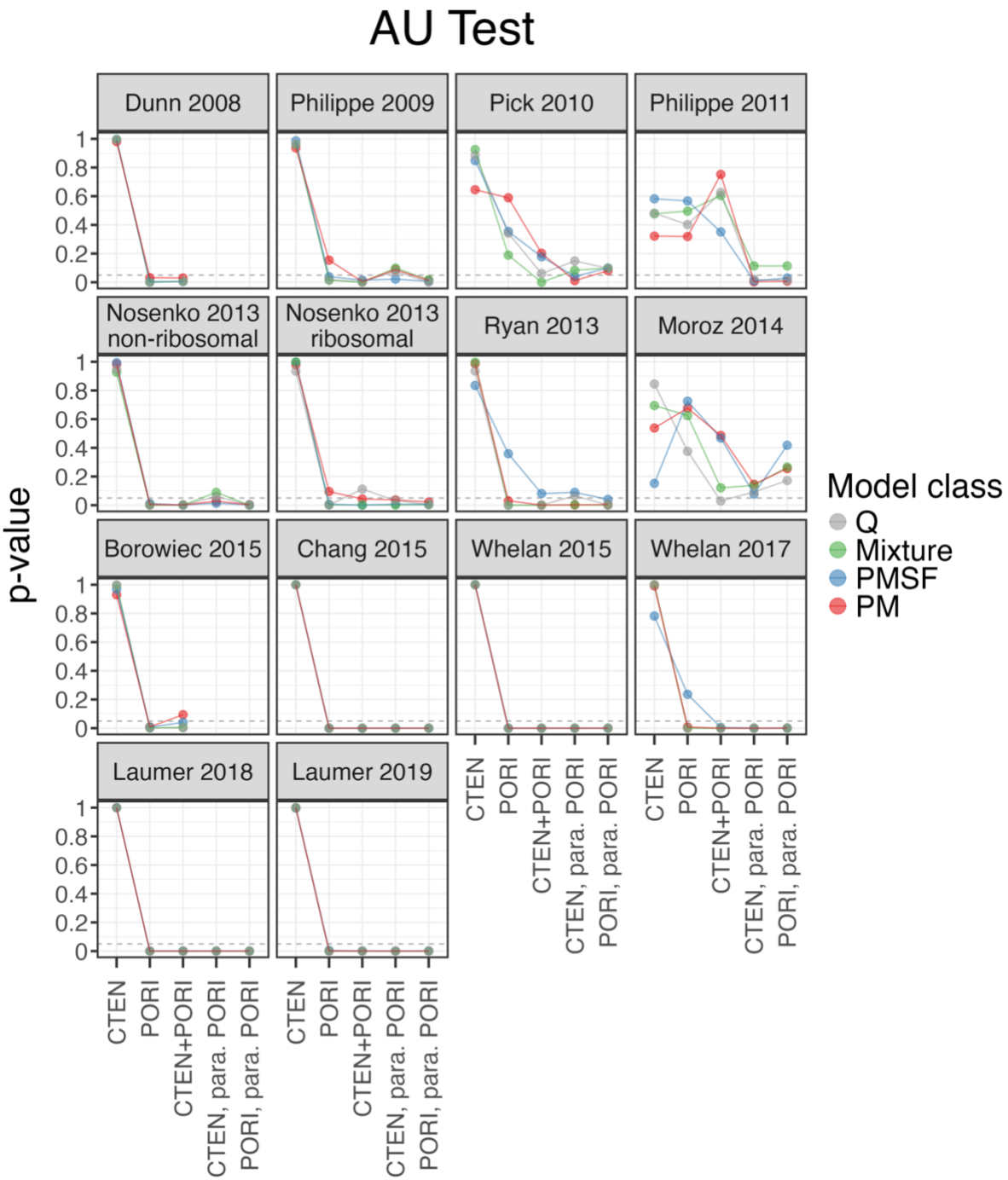
Supplementary Figure 21: Topology of the Porifera (sponge) clade for 350 trees estimated from different combinations of model of evolution and dataset.

Each bar shows the Porifera topologies obtained for a single dataset. For each dataset I calculated 25 different trees, each with a different model of evolution. Two datasets contained only 1 Porifera taxon, which meant the Porifera clade topology could not be obtained. Models are grouped on the x-axis by model class.



Supplementary Figure 22: 5-tree MAST model tree weights for 14 phylogenetic datasets and 3 classes of model.

The five hypotheses are shown on the x-axis: "CTEN" denotes Ctenophora-sister with monophyletic Porifera; "PORI" denotes Porifera-sister with monophyletic Porifera; "CTEN+PORI" means that the sister to all other metazoans is a monophyletic clade consisting of both Ctenophora and Porifera; "CTEN, para. PORI" denotes Ctenophora-sister with a paraphyletic Porifera clade; and "PORI, para. PORI" denotes Porifera-sister with a paraphyletic Porifera clade. Note that two datasets contained a single Porifera taxa (Dunn 2008 and Borowiec 2015), therefore estimating trees with paraphyletic Porifera was not possible for these datasets. For each combination of model class and dataset, the MAST tree weights sum to 1.



Supplementary Figure 23: AU test results for the 5 hypothesis trees from 14 phylogenetic datasets and 4 classes of model.

Each point is the p-value from the AU test for that combination of dataset, model class, and evolutionary hypothesis. The grey dashed line indicates the statistical significance threshold of p<0.05. Any point under that line is rejected by the AU test. The five evolutionary hypotheses are shown on the x-axis: "CTEN" denotes Ctenophora-sister with monophyletic Porifera; "PORI" denotes Porifera-sister with monophyletic Porifera; "CTEN+PORI" means that the sister to all other metazoans is a monophyletic clade consisting of both Ctenophora and Porifera; "CTEN, para. PORI" denotes Ctenophora-sister with a paraphyletic Porifera clade; and "PORI, para. PORI" denotes Porifera-sister with a paraphyletic Porifera clade. Note that two datasets contained a single Porifera taxa (Dunn 2008 and Borowiec 2015), therefore estimating trees with paraphyletic Porifera was not possible for these datasets.

Chapter Four:
Evaluating Hypotheses of Early Animal Evolution using Measures of Topological Variation

Caitlin Cherryh[1]*

1 Division of Ecology and Evolution, Research School of Biology, Australian National University, Canberra, Australia

* Corresponding author: caitlin.cherryh@anu.edu.au

Contributions:

Caitlin Cherryh designed the experiment, curated the datasets, wrote the R scripts, performed the analysis, interpreted the results, created the figures, and wrote the manuscript.

4.1 Abstract

The metazoan phylogeny is particularly difficult to resolve due to rapid radiation at the base of the animal tree and the deep evolutionary timescales involved. In particular, there is controversy over which metazoan clade was first to diverge: the sponge clade Porifera, the comb jellies clade Ctenophora, or a monophyletic clade consisting of both Porifera and Ctenophora. Previous studies have suggested that systematic bias due to incomplete lineage sorting and/or long branch attraction may contribute to the difficulty of resolving these relationships. In this chapter I use concordance factors to infer evolutionary patterns by investigating topological variance at key branches in the animal phylogeny. Using 12 empirical matrices previously used to estimate the metazoan phylogeny, I estimate gene and quartet concordance factors from maximum likelihood trees for three hypotheses of metazoan evolution. I also apply a constrained topology analysis to assess signal within individual genes in each matrix. As model choice has previously been shown to impact metazoan phylogenetic inference, I performed these analyses under both a Partitioned model and a site-specific C60 model. In consensus with previous studies, I identified substantial conflicting signal within empirical metazoan phylogenetic datasets. I found that both gene and quartet concordance factors were smaller under a C60 model than a Partitioned model for the hypothesis where Ctenophora diverges first, suggesting that inference of this topology is biased by model misspecification. This chapter confirms heterogeneous signal is widespread within the Metazoa, contributing to the difficult of resolving relationships between animal clades.

Keywords: Discordance, Gene tree heterogeneity, Animal tree of life, Model misspecification, Systematic bias, Constrained tree analysis

4.2 Introduction

The relationships between major animal clades are an unresolved controversy in evolutionary biology. The metazoan phylogeny consists of five main clades: Ctenophores, Porifera, Placozoa, Cnidaria, and Bilateria. Understanding the relationship between metazoan clades has implications for the origin and evolution of animal traits such as the muscles, a through-gut, and the nervous system (Telford et al. 2015; Schultz et al. 2023). The debate centres on which animal clade was first to diverge, making it the sister to all other metazoans (SOM). Historically, morphological data has placed the sponges (Porifera) as the SOM (Haeckel 1872; Reynolds 2019). Early phylogenetic studies supported this finding (Wainright et al. 1993; Collins 1998). However, later phylogenetic studies found support for Ctenophores as the SOM (Dunn et al. 2008; Ryan et al. 2013). Since then, there have been studies supporting Porifera

as the SOM (Pisani et al. 2015; Simion et al. 2017a; Feuda et al. 2017a; Kapli and Telford 2020; Redmond and McLysaght 2021a) and Ctenophora as the SOM (Whelan et al. 2015b, 2017a; Laumer et al. 2018a, 2019a; Li et al. 2021; Schultz et al. 2023). Another proposed topology proposes the SOM is a monophyletic clade consisting of the Ctenophora and Porifera clades (Shen et al. 2017; Francis and Canfield 2020). Alternative topological hypotheses have also been proposed (Martindale et al. 2002; Schierwater et al. 2009), but lack widespread support. Due to the complexity in reconstructing the metazoan tree of life, there is no consensus on the relationships between these clades.

Choice of substitution model, gene and site sampling, and outgroup selection can all impact the estimated topology of the metazoan tree. For example, trees estimated with partitioned site-homogeneous models tend to find Ctenophores as the SOM (Dunn et al. 2008; Hejnol et al. 2009; Ryan et al. 2013; Chang et al. 2015; Whelan et al. 2015b; Borowiec et al. 2015; Whelan et al. 2017a), whereas trees estimated with site-heterogeneous models tend to find Porifera as the SOM (Pisani et al. 2015; Simion et al. 2017a; Feuda et al. 2017a; Redmond and McLysaght 2021a). Other studies have shown that as model misspecification decreases (i.e., by incorporating site-heterogeneous models), the support for Ctenophora as SOM decreases (Simion et al. 2020; Redmond and McLysaght 2021a). Choice of outgroup also biases tree estimation, with previous studies inferring Porifera as the SOM for Choanoflagellate and holozoan outgroups (Philippe et al. 2009; Nosenko et al. 2013a; Li et al. 2021; McCarthy et al. 2023), and Ctenophora as the SOM for outgroups that include the more-distantly related Fungi clade (Ryan et al. 2013; Pisani et al. 2015; Whelan et al. 2015b).

Empirical metazoan phylogenetic datasets contain conflicting signals supporting both Ctenophora and Porifera as the SOM (Shen et al. 2017). Consequently, which genes and sites are included in an analysis can change the conclusions of phylogenetic inference. For example, Francis and Canfield (2020) found that removing sites that strongly supported either Ctenophora or Porifera as the SOM resulted in a highly supported tree with a monophyletic clade including both Ctenophora *and* Porifera as the sister of all other animals. Another contributing factor is the difficulty of detecting orthologs over such distantly related timescales and such deep divergence times (King and Rokas 2017; Pett et al. 2019). Ortholog misidentification results in datasets with excess internal incongruence, impacting tree inference and downstream analysis. A re-analysis of 5 metazoan phylogenetic datasets found that only 17% to 33% of orthogroups were retained when applying a strict check for orthologous signal, and in 2/5 datasets the SOM changed from Ctenophora to Porifera after filtering (McCarthy et al. 2023). Clearly, dataset construction has an impact on the level of systematic bias within a phylogenetic dataset.

Long branch attraction (LBA) rather than genuine phylogenetic signal has repeatedly been suggested a bias which can lead analyses to erroneously recover the Ctenophora as SOM (Philippe et al. 2009; Pick et al. 2010; Nosenko et al. 2013a; Pisani et al. 2015; Halanych et al. 2016; Kapli and Telford 2020; Szánthó et al. 2023), although some authors suggest the argument for LBA is exaggerated (Whelan et al. 2015b). LBA is a form of systematic error introduced by certain phylogenetic methods (including parsimony, distance methods and maximum likelihood) that arises when homoplasy is mistaken for true phylogenetic signal (Susko and Roger 2021). Under LBA, similarity due to convergent evolution is interpreted as shared ancestry, and fast-evolving taxa are incorrectly grouped together (Kapli et al. 2021). Previous studies have shown high rates of molecular evolution in the Ctenophora clade (Kohn et al. 2012; Arafat et al. 2018; Wang and Cheng 2019), which may cause LBA to 'pull' the Ctenophora clade towards the root of the metazoan tree (Kapli and Telford 2020). Assessing whether LBA results in Ctenophora as the SOM is difficult, as removing the potentially problematic branch that leads to the Ctenophora clade without removing all Ctenophora taxa is impossible (Kapli and Telford 2020). Previous studies have shown that support for Ctenophora as the SOM increases as the outgroup becomes increasingly distantly related (Nosenko et al. 2013a; Pisani et al. 2015), which is consistent with the expectations of topological impact due to LBA at the branch leading to the Ctenophora clade. In addition, Philippe et al. (2009) estimated a phylogenetic tree with Porifera as SOM using two different outgroups, and found that bootstrap values for deep nodes in the metazoan tree were higher for a closely-related outgroup (Choanoflagellates) than a more distantly-related outgroup (which included Fungi and Ichthyosporea taxa).

Incomplete lineage sorting (ILS) has also been suggested as a contributor to the difficulty of reconstructing the metazoan phylogeny. ILS is a source of phylogenetic conflict that occurs during speciation, particularly when speciation events are closely spaced (such as in rapid radiations) or when ancestral populations are large (Degnan and Rosenberg 2009). Under ILS some alleles from the ancestral population are retained in descendant species after consecutive speciation events, which causes gene trees to differ from the species tree (Steenwyk et al. 2023). The short branches at the root of the metazoan tree suggests that metazoans went through a rapid evolutionary radiation (King and Rokas 2017) leading to stronger discordance among loci due to ILS. The base of the metazoan tree contains gene tree–species tree discordance consistent with the effects of ILS (Ewing et al. 2008; Pandey and Braun 2021). The substantial gene tree heterogeneity within metazoan datasets also suggests the presence of ILS (Shen et al. 2017; Redmond and McLysaght 2021a; Li et al. 2021), with genes supporting both Ctenophora and Porifera as the SOM inferred.

The inferred metazoan tree is affected by multiple forms of systematic bias including model misspecification, LBA, and ILS. Kapli and Telford (2020) analysed the impact of both LBA and model misspecification by simulating metazoan datasets with site-heterogeneous models and either Ctenophora or Porifera as SOM. These simulations found that trees inferred with the incorrect site-homogeneous model had shorter internal branch lengths, and used this result to investigate the impact of internal branch length on tree inference (Kapli and Telford 2020). For data simulated with Porifera as SOM, Porifera was recovered as the SOM the majority of the time under a site-heterogeneous model (88%), but not under a site-homogeneous model (7%) (Kapli and Telford 2020). Conversely, when the data was simulated with Ctenophora as the SOM, Ctenophora was always recovered as the SOM under both site-homogeneous and site-heterogeneous models (Kapli and Telford 2020). These results demonstrate the difficulty of disentangling the multiple causes of conflicting signal when attempting to resolve the metazoan tree. The interaction of model misspecification and LBA was also investigated by Redmond and McLysaght (2021a), who examined the adequacy of the performance of different substitution models on three published metazoan datasets. They found that that site-homogeneous mixture models with a non-Poisson model (e.g., LG+C60, WAG+CF4) had the best fit and were best able to withstand LBA (Redmond and McLysaght 2021a). In addition, they found that support for Ctenophora as SOM reduced as model complexity increased, suggesting that this topology is inferred due to systematic error (Redmond and McLysaght 2021a).

Simulation studies have demonstrated the difficulty of resolving phylogenetic trees from data that contains ILS. ILS decreases as the length of branches increases and/or the effective population size decreases. When time between speciation events is short, there is insufficient time for genetic drift to act and fix alleles within the population, and therefore ILS is high (Maddison and Knowles 2006). Common approaches to vary the amount of ILS in simulations are to vary the population size (Liu et al. 2015a), to extend the length of terminal branches (Knowles et al. 2018), or to scale all branches in the tree proportional to a set of tree depths (Maddison and Knowles 2006; McCormack et al. 2009; Huang et al. 2010; Tonini et al. 2015). These methods focus on the impact of ILS across the tree. A study by Liu et al. (2015a) tested the impact of ILS and LBA on tree inference, by generating trees with short internal branches and differing levels of ILS before extending selected terminal branches (Liu et al. 2015a). They found the combination of short internal branches, high ILS, and long terminal branches misled concatenated tree estimation methods (Liu et al. 2015a). Similarly, another study simulating both ILS and LBA found that trees with interspersed long and short branches were difficult to resolve correctly (Kück et al. 2012). In the extreme case of a very short branch preceded and

followed by a long branch, maximum likelihood (ML) methods were unable to recover the true tree even for alignments >100K base pairs (bp) long (Kück et al. 2012). The metazoan tree topology fits this pattern of short internal branches interspersed with longer internal branches, combined with long terminal branches. Determining the correct phylogeny for such deep divergences is extremely difficult due to the conflicting signal present within the dataset (Degnan and Rosenberg 2009). The purpose of this chapter is therefore to examine whether metazoan tree inference is impacted by ILS, and to determine the extent to which ILS could mislead relationships between major metazoan clades.

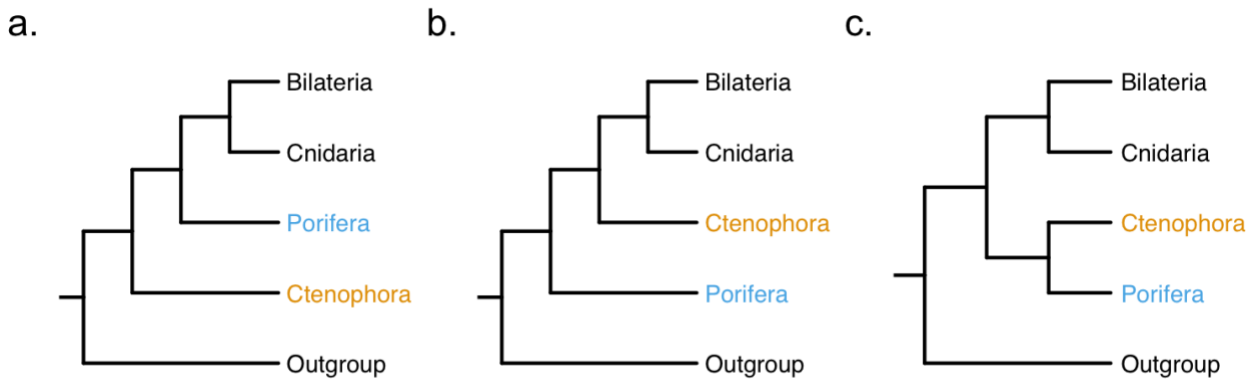


Figure 21: Three hypotheses for the evolutionary history of the metazoan tree of life.

a. Ctenophora is the sister of all other animals
b. Porifera is the sister of all other animals
c. A monophyletic clade consisting of both Ctenophora and Porifera is the sister of all other animals

To investigate whether ILS and LBA contribute to the difficulty of resolving the animal tree of life I reanalysed 12 published empirical matrices with both partitioned and C60 models, then used concordance factors (CF) to examine the discordance around key branches of the metazoan phylogeny. I focused on a key branch within the tree that may impact tree inference, based on previous work that identifies ILS and LBA as causes of systematic bias. This key branch defines the relationships between four clades: Outgroup; Ctenophora; Porifera; and the monophyletic clade consisting of Bilateria and Cnidaria. The three arrangements of these four clades around the key branch form my set of hypotheses of early animal evolution: Ctenophora as SOM (Figure 21a); Porifera as SOM (Figure 21b); and a monophyletic clade consisting of Ctenophora and Porifera as the SOM (Figure 21c). To assess topological variation, I calculated both gene concordance factors (gCF) (Minh et al. 2020a) and quartet concordance factors (qCF) (Mirarab et al. 2014) for each of these three topologies. As biological discordance contains information about evolutionary processes (Lanfear and Hahn

2024), in this chapter I use CFs to investigate patterns of discordance within the metazoan phylogeny. To investigate model misspecification, I calculate CFs under both a Partitioned and C60 model. I find substantial conflicting signal within all 12 empirical phylogenetic datasets. My results identified substantial conflicting signal within the Metazoa, consistent with ILS.

4.3 Methods

4.3.1 Overview

In this chapter, I aimed to identify whether ILS impact metazoan tree inference, by using signals of biological discordance to infer evolutionary processes. To do this, I calculated both gene and quartet concordance factors from 12 empirical multiple sequence alignments that had previously been used to estimate the metazoan phylogeny. Gene and quartet CFs are calculated using a species tree and a set of locus trees (Minh et al. 2020a; Lanfear and Hahn 2024). To examine the proportion of the dataset which agrees with each of three hypotheses of early animal evolution (Figure 21), I calculated concordance factors using constrained maximum likelihood trees (so that I could identify which concordance/discordance factor was associated with which topology) and unconstrained gene trees. To assess the impact of systematic bias due to model misspecification, I performed this analysis under both Partitioned and C60 models of sequence evolution.

4.3.2 Dataset Selection

I curated 12 existing phylogenetic matrices that were previously used to estimate the metazoan phylogeny (Table 6). To select these papers, I reviewed the literature to identify papers investigating relationships between clades at the root of the metazoan tree. I selected all papers that: included 1 or more taxa in the Bilateria, Cnidaria, Ctenophora, and Porifera clades; had amino acid (AA) alignments; and had both the alignment and partition files deposited and available in supplementary information or an external repository. This resulted in a list of 13 studies. Two of these studies contained large matrices which I removed from consideration (Hejnol et al. 2009; Simion et al. 2017a, 2017b). These two matrices had 1487 partitions (Hejnol et al. 2009) and 1719 partitions (Simion et al. 2017a), which were computationally intractable for my chapter (which required estimating 6 ML trees under a C60 model for each dataset).

Table 6: Empirical phylogenetic alignments selected for analysis.

Manuscript refers to the original publication of each alignment. Repository is the source of the alignment file, which may be supplementary material in the original manuscript, a data repository for the original manuscript, or a data repository for a different manuscript. Matrix name is the name of each alignment file at the source for that alignment file.

Manuscript	Repository	Matrix name	Number of taxa	Number of sites	Number of genes
Dunn *et. al.* (2008)	Li *et. al.* (2020b)	Dunn2008	64	21152	150
Philippe *et. al.* (2009)	Philippe *et. al.* (2009)	Philippe_etal_superalignment	55	30257	128
Philippe *et. al.* (2011b)	Philippe *et. al.* (2011b)	UPDUNN_MB	77	18463	150
Nosenko *et. al.* (2013a)	Nosenko *et. al.* (2013b)	nonribosomal_9187_smatrix	71	9189	35
Nosenko *et. al.* (2013a)	Nosenko *et. al.* (2013b)	ribosomal_14615_smatrix	71	14614	87
Ryan *et. al.* (2013)	Redmond and McLysaght (2021b)	REA_alignment_includingXenoturbella	61	88384	406
Moroz *et. al.* (2014)	Li *et. al.* (2020b)	ED3d	46	22772	114
Borowiec *et. al.* (2015)	Borowiec *et. al.* (2016)	Best108	36	41808	108
Chang *et. al.* (2015)	Feuda *et. al.* (2017b)	Chang_AA	77	51940	200
Whelan *et. al.* (2015b)	Whelan *et. al.* (2016)	Dataset10	70	59733	89
Whelan *et. al.* (2017a)	Whelan *et. al.* (2017b)	Metazoa_Choano_RCFV_strict	76	49388	117
Laumer *et. al.* (2018a)	Laumer *et. al.* (2018b)	Tplx_BUSCOeuk	59	94444	303

I selected one matrix from each other study, usually the matrix from which the authors drew their primary conclusions. Where more than one matrix was available for a single study (e.g., due to different filtering or sampling schemes), I selected the alignment used for the main phylogenetic tree figure in the results section. For some papers, there was no obvious main phylogeny, or the main phylogeny did not include the relevant clades. In this case, I selected the alignment that best met my requirements. For Moroz et. al. (2014), the phylogenetic trees estimated from the alignments were included as Extended Figures 3a and 3d. I selected the alignment used for Extended Figure 3d as it included more taxa (12 Ctenophora species instead of 2). For the Laumer 2018 dataset, a partition file was not available for the matrix in the main phylogenetic tree figure, so I selected an alternate matrix. Finally, I selected both the ribosomal and non-ribosomal matrices from the Nosenko 2013 dataset, as trees estimated from these two matrices have different topologies (Nosenko et al. 2013a). I did not modify or filter alignments in any way, except to update taxon names to be consistent across matrices so that the alignments could be batch processed using scripts. My alignments and partition

files are available from the FigShare repository "Ancient ILS" at https://doi.org/10.6084/m9.figshare.25965172.v2.

4.3.3 Tree estimation

I used IQ-Tree2 v2.2.2.6 (Chernomor et al. 2016; Minh et al. 2020b) to estimate maximum likelihood trees from the 12 empirical matrices. I inferred maximum likelihood trees from each alignment in IQ-TREE under both a Partitioned model and the site-specific C60 model (Le et al. 2008a).

To estimate the best partitioning scheme for each alignment in IQ-Tree2, I used the command "`iqtree2 -s alignment.fa -S partition.nex -m MFP`", where `alignment.fa` is an alignment file and `partition.nex` is the partition file from that alignment. I used the partition file from the original study to identify partition start and end points. I used the command option "`-m MFP`" to estimate the best-fit model for each partition with ModelFinder (Kalyaanamoorthy et al. 2017). In IQ-TREE, I estimated a maximum likelihood tree from the same alignment using that partitioning scheme using the command "`iqtree2 -s alignment.fa -p best_partition.nex`", where `alignment.fa` is the alignment file and `best_partition.nex` is the best partitioning scheme for that alignment. I applied an edge proportional Partitioned model using the "`-p`" command, which allowed independent rates of evolution for different partitions.

For each alignment, I estimated a maximum likelihood tree under a C60 model in IQ-TREE using the command "`iqtree2 -s alignment.fa -mset Poisson+C60,LG+C60 -mrate E,I,G,I+G,R,I+R`", where `alignment.fa` is the alignment file. I constrained model selection to either the Poisson+C60 or LG+C60 model as both have previously been applied to estimate the metazoan phylogeny (Laumer et al. 2018a, 2019a; Kapli and Telford 2020). Using the "`-mrate`" command, I tested the +I, +G, +I+G, +R, and +I+R models of rate heterogeneity among sites. As each of the best C60 models contained a "`+F`" parameter, the weights in the mixture model were automatically optimised by IQ-TREE without needing to call "`-mwopt`" explicitly.

I wrote a custom R v4.2.2 (R Core Team 2018) script to automate the process of tree estimation. The script to generate constraint trees and construct IQ-TREE command lines (01_empirical_tree_estimation.R) is available at the GitHub repository for this project https://github.com/caitlinch/ancient_ILS. All trees are available from the Figshare repository for this project (https://doi.org/10.6084/m9.figshare.25965172.v2).

4.3.4 Gene tree estimation

Both gene and quartet concordance factors are calculated using gene trees. I estimated two gene trees in IQ-TREE for each gene in each of the 12 empirical matrices: one using the best-fit model for that gene as determined by ModelFinder, and one using the best C60 model. The 12 empirical matrices included a total of 1187 genes, and I estimated two trees from each gene resulting in a total of 2374 gene trees.

I used partition files from the original analyses to split the 12 empirical matrices into 1187 alignments, each consisting of a single ortholog or orthologous group. To estimate maximum likelihood gene trees with ModelFinder, I used the command "`iqtree2 -s partition_alignment.fa -m MFP`", where `partition_alignment.fa` is the alignment file for a single partition. To estimate gene trees with C60 models, I used the command "`iqtree2 partition_alignment.fa -m best_C60_model`", where, `partition_alignment.fa` is the alignment for a single partition. The term "`best_C60_model`" represents is the best C60 model identified by ModelFinder for the alignment, including rates and weights for each of the 60 C60 parameters, and rates and weights for the rate heterogeneity across sites model.

I wrote custom R v4.2.2 (R Core Team 2018) scripts to automate the process of gene tree estimation. The scripts to generate constraint trees and construct IQ-TREE command lines (02_gene_tree_estimation.R, 02_C60_gene_tree_estimation.R) are available at the GitHub repository for this project https://github.com/caitlinch/ancient_ILS. All gene trees are available from the Figshare repository for this project (https://doi.org/10.6084/m9.figshare.25965172.v2).

4.3.5 Concordance factors

Gene and quartet CFs are an estimation of the proportion of the genome for which a given clade is true (Lanfear and Hahn 2024). Estimating gCFs and qCFs requires a species tree (described below) and a set of gene trees (described above). To assess the topological variation in the 12 empirical matrices, I calculated gene and quartet concordance factors at key branches in the metazoan phylogeny. I was interested in CFs and branch lengths for three branches: the key branch, the branch leading to the Ctenophora clade, and the branch leading to the Porifera clade.

4.3.5.1 Species tree estimation

The key branch is defined by four clades: Outgroup; Porifera; Ctenophora; and the monophyletic clade consisting of Bilateria and Cnidaria. The three alternate arrangements of these clades around the key branch represent the hypotheses of early animal evolution in my chapter (Figure 21). To determine which CF was associated with which evolutionary hypothesis, I calculated CFs for each of the three resolutions of the key branch (Figure 21). I constructed a multifurcating guide tree for each hypothesis to fix relationships between clades without impacting relationships inside each clade, then added these trees as a constraint for the tree estimation process. The guide trees constrained relationships between the Outgroup, Bilateria, Cnidaria, Ctenophora, and Porifera clades such that each guide tree represented one of the three evolutionary hypotheses from Figure 21. Some matrices included one or more taxa in the clade Placozoa. The placement of Placozoa within the metazoan tree is unresolved (Schierwater et al. 2021). Therefore, I did not constrain Placozoa within my guide trees, which allowed placement of Placozoa taxa to be inferred during tree estimation.

Under a Partitioned model, I estimated trees for each hypothesis of early animal evolution with each alignment. I estimated trees in IQ-TREE with the command "`iqtree2 -s alignment.fa -p best_partition.nex -g guide_tree.nex`", where `alignment.fa` is the alignment file, `best_partition.nex` is the best partitioning scheme for that alignment (as identified during tree estimation, see above), and `hypothesis_tree.nex` is the multifurcating guide tree which constrains tree estimation to one of the hypotheses of early animal evolution. I applied an edge proportional Partitioned model using the "-p" command, which allowed independent rates of evolution for different partitions.

For each alignment, I also estimated trees under a C60 model for each hypothesis of early animal evolution. I used the C60 model (including all weights and rates) selected by ModelFinder (as identified during tree estimation, see above). I estimated trees in IQ-TREE with the command command "`iqtree2 -s alignment.fa -m best_C60_model -g hypothesis_tree.nex`", where `alignment.fa` is the alignment file, `best_C60_model` is the best C60 model for that alignment (e.g., LG+C60+R5), and `hypothesis_tree.nex` is the multifurcating guide tree which constrains tree estimation to one of the hypotheses of early animal evolution.

4.3.5.2 Concordance factor estimation

The gene concordance factor (gCF) for each branch is the proportion of decisive gene trees that support a given branch in a reference tree. The gene discordance factors (gDF1 and gDF2) are the proportion of genes that support the alternate topologies obtained by performing an NNI move around the branch of interest. Finally, the paraphyletic gene concordance factor (gDFP) is the proportion of gene trees where one or more clades in the gene tree is not monophyletic compared to the four clades around the branch of interest of the reference topology (Lanfear and Hahn 2024). As all genes are included in these four categories, gCF+gDF1+gDF2+gDFP=1 (Minh et al. 2020a). I calculated 2 sets of gCFs from each of the 12 empirical matrices: one from the constrained trees estimated from Partitioned models, and one from the constrained trees estimated from C60 models. I calculated gCFs in IQ-TREE v2.2.2.6 using the command "`iqtree2 -te hypothesis_tree.nex --gcf gene_trees.nex`", where `hypothesis_tree.nex` is the maximum likelihood tree for one of the three hypotheses of animal evolution (Figure 21), and `gene_trees.nex` is the set of gene trees.

Similar to gCFs, quartet concordance factors represent quartet variance for the main topology (qCF) and two alternate topologies arranged around the same branch (qDF1, qDF2). Quartet concordance factors were estimated in ASTRAL v5.6.9 using the command line "`java -jar astral.5.7.8.jar -q hypothesis_tree.nex -i gene_trees.nex -t 2 -o quartet_concordance.nex 2> quartet_concordance.log`". For a given alignment, `hypothesis_tree.nex` is the maximum likelihood tree for one of the three hypotheses of animal evolution (Figure 21), `gene_trees.nex` is the set of gene trees file, `quartet_concordance.nex` is the output tree and `quartet_concordance.log` is the output log.

Under a Partitioned model, 5 matrices (Chang 2015, Laumer 2018, Nosenko 2013 ribosomal, Philippe 2009, and Whelan 2015) had different placement of Placozoa in each of the constrained maximum likelihood trees such that the key branch did not connect the same 4 clades in each of the 3 constrained trees. For these 5 matrices, before calculating CFs as described above, I removed Placozoa taxa from the constrained ML tree and from each of the gene trees using the function DropTip from the R package TreeTools v1.10.0 (Smith and Paradis 2023). I could remove Placozoa taxa from these analyses because my investigation focuses on the SOM and is not concerned with the placement of Placozoa. The C60 analyses for these matrices did not have this issue and therefore Placozoa was left in the C60 constrained ML trees and gene trees.

In this chapter, I denote which CF belongs to which constrained topology by labelling each concordance factor with the corresponding topology. For example, the gCF and qCF for the constrained tree topology with Ctenophora as the SOM are given by gCF_{CTEN} and qCF_{CTEN} respectively. When Porifera is constrained as the SOM, the CFs are gCF_{PORI} and qCF_{PORI}. Finally, for the constrained tree with monophyletic clade of Ctenophora and Porifera as SOM, the CFs are $gCF_{CTEN+PORI}$ and $qCF_{CTEN+PORI}$.

4.3.5.3 Reproducing concordance factor analyses

To conduct the concordance factor analyses described above, I wrote custom R v4.2.2 (R Core Team 2018) scripts to remove Placozoa taxa from the Partitioned model gene trees and ML trees for the matrices specified above. I also wrote custom R v4.2.2 scripts to automate the process of concordance factor estimation, and to extract the concordance factors and branch lengths for the three branches of interest. The scripts to construct IQ-TREE and ASTRAL command lines (03_empirical_concordance_factors), and to analyse and plot the results (05_plot_figures_cf.R and 05_plot_trees.R) and are available at the GitHub repository for this project https://github.com/caitlinch/ancient_ILS. To perform my analyses I used the R packages ape v5.7.1 (Paradis and Schliep 2019), castor v1.8.0 (Louca 2023), dplyr v1.1.4 (Wickham et al. 2021), phangorn v2.11.1 (Schliep 2011), phytools v2.1.1 (Revell 2012), seqinr v4.2.36 (Charif and Lobry 2007), stringr v1.5.1 (Wickham 2023), and TreeTools v1.10.0 (Smith and Paradis 2023). I performed data analyses and plotting in R, using the packages ggplot2 v3.5.0 (Wickham 2016), ggpubr v0.6.0 (Kassambara 2023), ggtree v3.4.4 (Yu et al. 2017; Xu et al. 2022), patchwork v1.2.0 (Pedersen 2022), and reshape2 v1.4.4 (Wickham 2007). All gCF and qCF results are available from the Figshare repository for this project (https://doi.org/10.6084/m9.figshare.25965172.v2).

4.3.6 Constrained gene tree estimation

To further assess the support for the three hypotheses of early animal evolution (Figure 21), I compared which hypothesis topologies were more likely for each gene. I estimated three constrained gene trees from each gene, one for each hypothesis of animal evolution in Figure 21, and recorded the log-likelihood and BIC values for each constrained gene tree. For each gene, I determined which hypothesis was most likely by identifying which of the constrained gene trees had the lowest BIC. I also compared the BIC from the unconstrained gene tree with the BIC from the constrained gene trees, and noted whether the gene tree with the lowest BIC was constrained or unconstrained. Due to the large computational and time requirements for estimating gene trees with the C60 model, I conducted the constrained tree analysis only for the Partitioned model.

For each gene, I constructed multifurcating guide trees for each of the hypotheses of animal evolution. To estimate constrained maximum likelihood gene trees in IQ-TREE, I used the command line "`iqtree2 -s gene_alignment.fa -m ModelFinder_model -g gene_guide_tree.nex`". Here, `gene_alignment.fa` is the alignment for a single gene, and `gene_guide_tree.nex` is the guide tree for that gene. I set the model of sequence evolution "`ModelFinder_model`" equal to the model for that gene from the best partitioning scheme for that alignment, including rate or gamma parameters.

I removed 5 genes from the Whelan 2015 matrix, as I was unable to estimate constrained gene trees due to errors in the guide trees. I estimated constrained gene trees for the remaining 1882 genes, resulting in a total of $1882 \times 3 = 5646$ constrained gene trees.

I wrote a custom R v4.2.2 (R Core Team 2018) scripts to automate the process of constructing gene trees and calling IQ-TREE to estimate gene trees, and to perform the BIC comparison. The scripts to generate constraint trees and construct IQ-Tree command lines (02_gene_tree_estimation.R, 04_single_gene_processing.R, and 05_plot_figures_cf.R) are available at the GitHub repository for this project https://github.com/caitlinch/ancient_ILS. Plotting was performed as described above. Results from the constrained tree analysis are available from the Figshare repository (https://doi.org/10.6084/m9.figshare.25965172.v2).

4.4 Results

4.4.1 Gene and quartet concordance factor values suggest substantial contributions of ILS

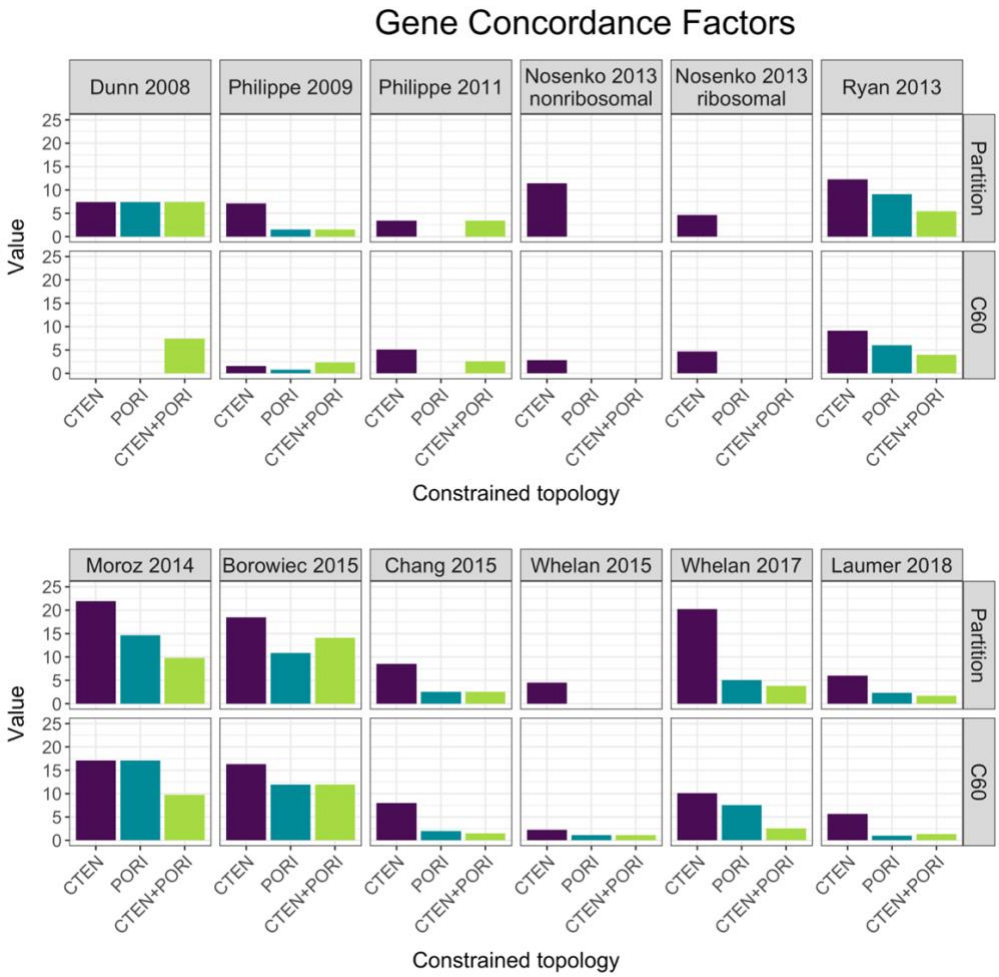


Figure 22: Gene concordance factors (gCFs) around the key branch from 12 empirical phylogenetic matrices.

For each matrix, the results from the Partitioned model are shown above results from the C60 model. In the x axis, CTEN refers to constrained trees estimated with the Ctenophora-as the SOM; PORI refers to the constrained trees estimated with Porifera as the SOM; and CTEN+PORI refers to constrained trees estimated with the monophyletic clade of Ctenophora and Porifera as the SOM.

I found substantial variation in gCFs for the different datasets (Figure 22). The majority of analyses (16/22) had a non-zero gCF value for each of the three resolutions of the key branch. In 6 cases, only one gCF was non-zero: either gCF_{CTEN} (5/6) or $gCF_{CTEN+PORI}$ (1/6). For most datasets and both models of evolution (18/22), gCF_{CTEN} had the largest value (Figure 22). Comparing gCF values under both models, I found gCF values for the same dataset were generally larger under the Partitioned model. For the majority of matrices (10/12), gCF_{CTEN} was

higher for trees estimated under a Partitioned model than for trees estimated under the site-specific C60 model (Figure 22). The gCF_{PORI} values were higher under the C60 model for 4/12 datasets, lower under the C60 model for 5/12 datasets, and identical for 3/12 datasets. Finally, $gCF_{CTEN+PORI}$ values were higher under the C60 model for 2/12 datasets, lower under the C60 model for 6/12 datasets, and identical for 4/12 datasets.

The majority of genes did not include the key branch (Supplementary Figure 24), with mean gDFP values for each matrix ranging from 54.9 – 95.5. The gDFP values were higher for trees estimated under a C60 model than under a Partitioned model for the majority of matrices (9/12).

The highest gCF for each combination of dataset and substitution model ranged from 2.25 to 21.95 (Figure 22). The middle gCF ranged from 0 to 17.07, and the smallest gCF from 0 to 11.96. The difference between the highest and lowest gCF ranged from 0–16.45.

Regardless of substitution model, the majority of matrices had highest qCF values for Ctenophora as the SOM (Partitioned: 11/12 matrices; C60: 8/12 matrices) (Figure 23). For both substitution models, the second-highest qCF was qCF_{PORI} (Partitioned: 9/12 matrices; C60: 7/12 matrices) and the lowest qCF was $qCF_{CTEN-PORI}$ (Partitioned: 9/12 matrices; C60: 8/12 matrices).

Comparing qCF values calculated from constrained trees under the Partitioned and C60 models, I observed the same trend in 7/12 datasets (Figure 23). For these datasets, increasing model complexity resulted in lower qCF_{CTEN} values and higher qCF_{PORI}/$qCF_{CTEN+PORI}$ values. Of the remaining matrices, 4/5 had one qCF value increase and the other two decrease when the model was changed (Philippe 2009, Philippe 2011b, Ryan 2013, Moroz 2014).

While qCF_{CTEN} was generally higher than the other qCFs, I observed that the difference between the highest and lowest qCF for each combination of dataset and model was generally low, ranging from 0.018 – 0.266 (Figure 23). The highest qCF for each combination of dataset and substitution model ranged from 0.342–0.487, the middle qCF from 0.264–0.375, and the lowest qCF from 0.193–0.326.

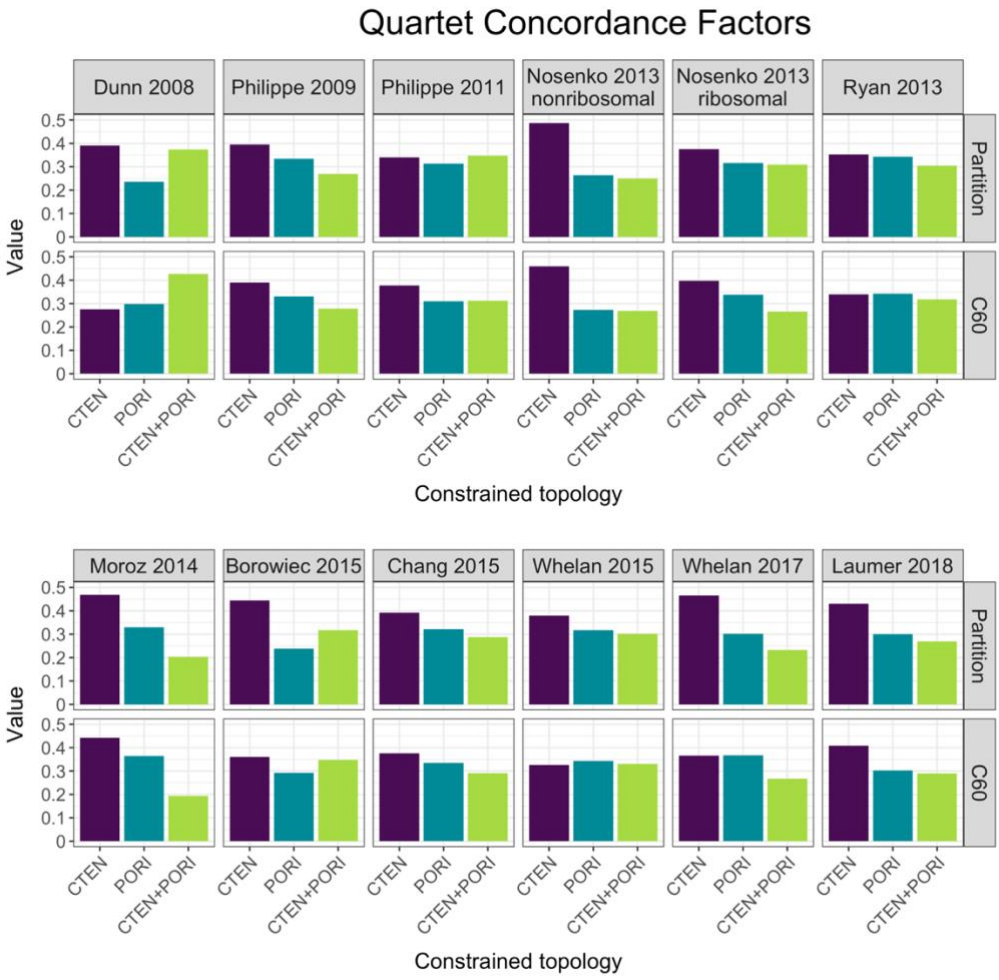


Figure 23: Quartet concordance factors (qCFs) around the key branch from 12 empirical phylogenetic matrices.

For each dataset, the results from the Partitioned model are shown above results from the C60 mode. In the constrained topology axis, CTEN refers to constrained trees estimated with the Ctenophora-as the SOM; PORI refers to the constrained trees estimated with Porifera as the SOM; and CTEN+PORI refers to constrained trees estimated with the monophyletic clade of Ctenophora and Porifera as the SOM.

4.4.2 All hypothesis topologies are included in the set of gene trees for any given dataset

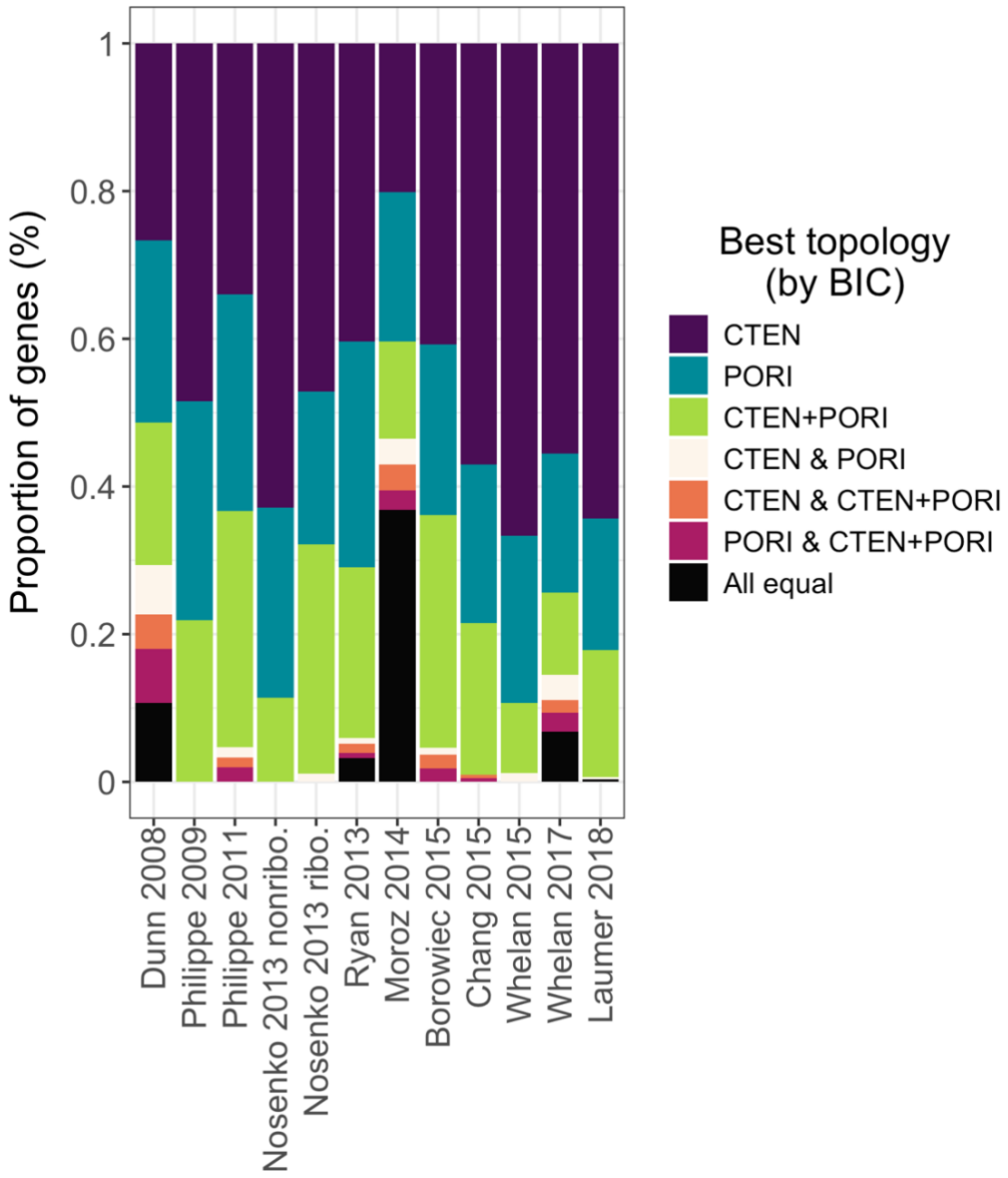


Figure 24: Best topology (i.e., which metazoan clade diverged first) for each gene in 12 empirical matrices. The best topology was defined as the topology with the lowest BIC.

I applied ModelFinder in IQ-Tree2 to estimate the best model for each gene and applied that model to all three constrained tree inferences. CTEN refers to constrained trees estimated with the Ctenophora-as the SOM; PORI refers to the constrained trees estimated with Porifera as the SOM; and CTEN+PORI refers to constrained trees estimated with the monophyletic clade of Ctenophora and Porifera as the SOM. Best topology for each gene determined by comparing the BIC values for each of the three constrained trees (CTEN, PORI and CTEN+PORI) and selecting the tree with the lowest BIC. Where more than one tree tied for equal lowest BIC, I list both topologies (with topologies separated by "&"). Where all BIC values were equal for a given gene, the best topology was marked as "All equal". The terms "nonribo." and "ribo." refer to "non-ribosomal" and "ribosomal" respectively.

All three hypothesis topologies are represented within the gene trees for each dataset (Figure 24). Within each dataset, there were at least 4 genes with the lowest BIC score for each of the

3 hypotheses of evolution (Figure 24). On average, around half of genes had lowest BIC for the hypothesis of Ctenophora as SOM (46.60%), which was higher than the proportion of genes supporting either Porifera (24.2%) or Ctenophora+Porifera (20.9%) as the SOM. The remaining genes (8.3%) had equal BIC for 2 or more constrained tree topologies. Different datasets had different proportions of genes supporting each of the three hypotheses. The hypothesis of Ctenophora as SOM had the highest proportion of genes with the lowest BIC score (20.18 – 66.67%), followed by Porifera as SOM (17.82 – 30.54%), then by the monophyletic clade of Ctenophora and Porifera as SOM (11.11 – 31.38%).

Finally, for each gene I compared the BIC scores from constrained and unconstrained trees (Supplementary Figure 25). For 10/12 matrices, the majority of genes had better BIC scores for unconstrained trees (52 – 88.6%). The exceptions were Moroz 2014 and Borowiec 2015, where the proportion of unconstrained genes with best BIC scores were 43.9% and 50% respectively.

4.5 Discussion

In this chapter, I applied gene and quartet concordance factors to quantify biological variation in 12 empirical matrices previously used to estimate the metazoan phylogeny. I found the majority of datasets had non-zero values for all three concordance factors at the key branch, and that the CF values were highest for the resolution of the key branch with Ctenophora as sister to all other metazoans.

Concordance factors differ from branch support, in that the biological variation should be consistent for alignments with different lengths, whereas branch support is the inverse of sampling variance and as such is high for alignments with large numbers of sites (Thomson and Brown 2022). Biological discordance, such as the discordance contained in gene trees and quartets, can reveal evolutionary processes introduced by biological processes such as incomplete lineage sorting or introgression (Lanfear and Hahn 2024). I compared CFs at the key branch of the metazoan phylogeny to assess topological discordance and evolutionary patterns. In general, CFs estimated under a C60 model had a smaller difference between the largest and smallest CF values for each gene than CFs estimated under a Partitioned model. This pattern held for both gCFs and qCFs. Site-heterogeneous models have been shown to reduce systematic errors in metazoan tree inference (Redmond and McLysaght 2021a). Therefore, CFs calculated from trees inferred under site-homogeneous (e.g., Partitioned models) should be more impacted by systematic bias than those inferred under site-heterogeneous models (e.g., C60), and comparing the CFs between models provides an indication of the extent and type of systematic bias.

Comparing C60 and Partition models does not guarantee that the effects of model misspecification can be determined. One key issue with this approach is that I did not validate the absolute model fit of either model. Allowing the data to reject model assumptions is an important step in the phylogenetic protocol that is often skipped (Brown and Thomson 2018; Jermiin et al. 2020). In this chapter the Partition model for each dataset was selected using ModelFinder in IQ-Tree2 (Chernomor et al. 2016; Kalyaanamoorthy et al. 2017; Minh et al. 2020b) which applies a greedy strategy from PartitionFinder (Lanfear et al. 2012, 2017) to select models for each gene and merge genes into partitions. This approach tests relative model fit, but does not include an absolute goodness-of-fit test to assess whether the final partition model is adequate to explain the data. Consequently, the final model is not guaranteed to have absolute fit for any of the datasets (Gatesy 2007). I also did not test the absolute fit of the C60 models applied in this chapter. Although the absolute fit of the C60 and Partition models is not guaranteed for the datasets used in these studies, previous studies of

the metazoan tree have compared the fit of site-heterogeneous and site-homogeneous models. Kapli and Telford (2020) compared the performance of site-heterogeneous and site-homogeneous models on previously published metazoan datasets and simulated datasets using Bayesian and concatenated tree inference methods. They found that site-homogeneous models consistently under-estimated branch lengths for data that evolved heterogeneously (Kapli and Telford 2020). In addition, site-homogeneous models also resulted in incorrect topologies with >90% bootstrap support values at the nodes of interest (Kapli and Telford 2020). Redmond and McLysaght (2021a) assessed model adequacy for three previously published metazoan datasets using a 4-tiered model system, where each tier introduced increasingly site-heterogeneous models of substitution. Site-heterogeneous models consistently had the best fit for all three datasets (Redmond and McLysaght 2021a). Based on the results of these previous studies, it seems reasonable to assume that even if the C60 model is not an absolute fit for any of the datasets investigated in this chapter, that the fit of the C60 models is better than that of the Partition models. As such the results from the Partition models can be interpreted in comparison to the better-fitting C60 models, as long as the possible limitations of the C60 models are acknowledged when interpreting the results.

I applied concordance factors to determine whether the substantial gene tree heterogeneity present within metazoan datasets is due to ILS. If ILS is the only process responsible, I would expect similar values for the second and third-highest CF (Lanfear and Hahn 2024), such that the CF for one hypothesis was consistently the largest and the other two CFs were smaller and identical in value. Under extreme ILS, all CFs would be identical. My results did not identify the clear pattern in CF values that arises under pure ILS. I observed that the second- and third-highest CFs were non-zero for most datasets. Comparing the difference between the largest and second-largest CF with the difference between the second-largest and third-largest CF, I observed that the former difference was bigger for 19/24 gCF analyses and 14/24 qCF analyses (Figure 22, Figure 23). This suggests that ILS occurred along with additional evolutionary processes, resulting in the patterns of discordance observed in this chapter.

Both gCFs and qCFs are impacted by gene tree estimation error (GTEE) (Lanfear and Hahn 2024). Phylogenetic signal is correlated with alignment length, therefore smaller genes contain less phylogenetic signal (Shen et al. 2016) and are more likely to suffer from GTEE. Gene length in the 12 matrices included in this chapter ranged from 36–1820 sites, with mean gene length of 246.9 sites. My results found that the majority of genes in each dataset (53.7 – 97.1%) were paraphyletic at the key branch, meaning that the majority of genes were unable to resolve well-established metazoan clades. This combination of short alignments and high proportion of paraphyletic gene trees is indicative of widespread GTEE (Lanfear and Hahn 2024). The

consistent heterogeneity of phylogenetic signal observed in this chapter, plus the consistency of highest CF values for the hypothesis where Ctenophora diverges before all other animals, suggests that this chapter identified phylogenetic signal and not just GTEE. Detangling GTEE from the complex evolutionary history of the Metazoa is extremely difficult due to the rapid radiation that occurred at the base of the metazoan tree (King and Rokas 2017). Accommodating this signal may require alternative phylogenetic methods that relax the treelikeness assumption, such as phylogenetic networks (Solís-Lemus et al. 2017; Poormohammadi et al. 2020; Lutteropp et al. 2022) or the MAST model (Wong et al. 2024).

As the metazoan phylogeny is notoriously difficult to resolve, previous studies have applied methods using constrained single genes or sites to dissect phylogenetic signal. These approaches are collectively named “Constrained Topology Analyses” (CTA), and aim to determine the evolutionary relationships of a contentious clade by adding up small contributions from small subsets of data under the assumption that the majority of these subsets will support the evolutionary truth (Simion et al. 2020). CTA approaches are limited for two main reasons: first the lack of guarantee that the majority signal is phylogenetic in nature rather than noise or systematic bias, and second that there is sufficient information to estimate accurate gene trees (Simion et al. 2020). Previous studies applying CTA approaches to the Metazoa tend to find support for Ctenophora as the SOM (Arcila et al. 2017; Shen et al. 2017). Shen et al. (2017) took 8 empirical phylogenetic datasets previously used to estimate relationships between metazoan clades and reanalysed each dataset with a CTA approach. They found that the hypothesis of Ctenophora as SOM was supported by 42.5 – 69.7% of genes and 39.8 – 56.9% of sites, and that genes supporting this hypothesis were more informative i.e., they had a large difference in log likelihood between the best constrained tree and other constrained trees estimated from the same gene (Shen et al. 2017). Similarly, Arcila et al. (2017) applied a CTA approach to the Whelan et al. (2015b) dataset and found that 64.1% of genes support the Ctenophora as SOM hypothesis.

I applied a CTA approach to genes from 12 empirical phylogenetic matrices previously used to estimate the metazoan tree, by estimating three constrained gene trees from each gene and identifying the hypothesis that resulted in the lowest BIC score (Figure 24). By including 12 matrices from 11 manuscripts published over a 20 year period from 2008 – 2018, my analysis expanded on those of Arcila et al. (2017) and Shen et al. (2017), who investigated 1 matrix and 8 matrices (from 3 manuscripts) respectively. I found between 20.18 – 66.67% of genes supported Ctenophora as the SOM, depending on dataset. Two matrices (Dunn 2008 and Philippe 2011) analysed only in my chapter had less than 40% of genes supporting the Ctenophora as SOM topology, although both still had the highest number of trees inferred

supporting this topology. There are two main differences between my analysis and the previous metazoan CTA analyses. First, I identified that for 0 – 46.5% of genes in each dataset, multiple constrained topologies had identical best BIC scores. This varied between datasets, with 0% of genes in the Nosenko 2013 non-ribosomal and Philippe 2009 matrices having multiple constrained trees tie for best BIC score, compared to 46.5% of genes in the Moroz 2014 dataset. Second, I identified that over half of genes for each dataset did not support any of the three hypothesis topologies when trees were unconstrained (Partitioned model: 53.7 – 95.5%; C60 model: 56.1 – 97.14%) (Figure 22).

My chapter reinforces the heterogeneity of phylogenetic signal within metazoan datasets. My chapter corroborates Simion et al. (2020), who reanalysed the datasets included in Shen et al. (2017) to identify sources of error within CTA approaches, and found that the majority of constrained gene trees (>93%) were rejected when compared to the unconstrained topology. My CTA was limited to Partitioned models, but previous research suggests that the proportion of constrained gene trees rejected when compared to the unconstrained topology increases as model complexity increases (Simion et al. 2020). This is consistent with previous findings that site-heterogeneous models have the best fit for metazoan datasets, and that trees estimated from simpler models suffer from systematic bias due to model misspecification (Redmond and McLysaght 2021a). The conflicting signals detected by the CTA in this chapter suggest that inferences of metazoan datasets should explicitly accommodate the conflicting signal identified in this chapter.

A limitation of this chapter's approach is that the concordance factors reveal the extent of discordance within metazoan datasets but cannot definitively identify the historical biological processes that resulted in that discordance. If ILS was the only cause of discordance, both discordance factors would be equal (Lanfear and Hahn 2024). However, my results suggest the presence of another factor impacting concordance factors, either biological or analytical. Many methods to detect and quantify ILS have been developed (Lee et al. 2012; Song et al. 2012; Knowles et al. 2018; Sayyari et al. 2018; Morales-Briones et al. 2021; Stiller et al. 2024). Further analyses of the metazoan datasets analysed in this chapter could include additional methods for detecting or quantifying ILS such as QuIBL ("Quantifying Introgression via Branch Lengths") which distinguishes between introgression and ILS by comparing the distribution of internal branch lengths for three taxon trees (Edelman et al. 2019), or by comparing the genetic distance between sequences from two species to the null distribution obtained from simulations under the multispecies coalescent (Joly et al. 2009).

Unfortunately, the factors that make inferring an accurate metazoan phylogenetic tree difficult also complicate inference of historical biological processes. There have been previous studies attempting to untangle the multiple biological and analytic processes contributing to discordant phylogenetic signal. Cai et al. (2021) applied an approach similar to the CF method in their study investigating the biological and analytic factors contributing to the difficulty of estimating a tree for the flowering plant clade Malpighiales. Cai et al. (2021) applied a triplet frequency based method to test deviation from the multispecies coalescent. Frequencies of each triplet are calculated for each gene tree, with the expectation that under solely ILS the frequency of the minor discordant triplets with be equal. To evaluate the biological causes of discordant triplets, Cai et al. (2021) applied a simulation-based approach that included a parametric bootstrap to allow them to compare their results to a simulation-based null distribution, allowing them to compare the observed triplet frequencies to the expected triplet frequencies under ILS. The methods applied by Cai et al. (2021) allowed them to estimate levels of ILS across the inferred phylogeny, and distinguish between ILS, gene tree estimation error and gene flow. As the Metazoa also has a complex evolutionary history, a similar approach of integrating empirical and simulation data would allow investigation into the contribution of the suggested causes of heterogeneous phylogenetic signal.

Determining orthology of gene in Metazoa datasets is particularly difficult. Few orthologs are shared across the Metazoa due to the deep divergences between clades and the rapid evolution of the Ctenophora clade (Pett et al. 2019). Multiple studies have identified orthology issues within metazoan datasets (Philippe et al. 2011b; Simion et al. 2017a; Redmond and McLysaght 2021a). McCarthy et al. (2023) took 5 previously published metazoan datasets and tested the orthology of each gene by identifying the number of established metazoan clades (Bilateria, Ctenophora, Cnidaria, Porifera, Outgroup) recovered in each gene tree, and retaining only genes that recovered 3 or more monophyletic clades. Between 17 – 33% of orthogroups from the original study were retained, equivalent to 25–52% of the original dataset size (McCarthy 2023). The difficulty of orthology inference for metazoan datasets has been demonstrated through simulation study by Natsidis et al. (2021), who simulated the evolution of sets of orthologous genes along the metazoan tree of life. Natsidis et al. (2021) performed 200 simulations, each of 5000 sets of orthologs, and allowed different genes to have different rates of evolution and different degrees of site rate heterogeneity (Natsidis et al. 2021). Success in recovering the correct orthogroups depended on the gene rate multiplier, with increasing numbers of errors as the gene rate multiplier increased. At the highest values of gene rate multiplier, 250,000 orthogroups were estimated from a single simulation replicate

(Natsidis et al. 2021). Given the empirical and simulation study results, it seems reasonable to assume that some level of orthology errors are present within any metazoan dataset.

The complexity of identifying orthologs within the metazoan tree has consequences on downstream inferences. For example, misidentification of orthologs has been shown to impact species tree estimation (Brown and Thomson 2017; Siu-Ting et al. 2019; Natsidis et al. 2021; McCarthy et al. 2023). In this chapter, my approach assumed that the estimated concordance factors reflected the evolutionary history of orthologous loci. Misidentification of orthologs in the datasets included in this chapter would impact concordance factors, resulting in apparent support for ILS. There are two potential approaches to detangle the impact of ILS and the impact of ortholog misidentification. The first is a simulation study, similar to that of Natsidis et al. (2021). This would allow comparison of the concordance factors between the true simulated orthogroups and the recovered estimated orthogroups, with any difference attributable to errors in the ortholog identification process. The second is a study of published empirical metazoan datasets, similar to McCarthy et al. (2023). After applying a stringent filter to remove inadequate orthogroups, concordance factors of the original and filtered datasets can be compared. One limitation of the empirical approach is dataset size. Of the metazoan datasets analysed by McCarthy et al. (2023), 4/5 original datasets had less than 210 loci. Assuming similar results to McCarthy et al. (2023), around 25-50% of the loci will be removed during the filtering process, resulting in small sample sizes particularly proportional to the number of taxa (usually around 50 – 100) in metazoan datasets. The degree of concordance within in a dataset does not change as the amount of data used to estimate concordance factors increases (Lanfear and Hahn 2024). However, estimating accurate concordance factors relies upon not only sufficient number of genes but also taxon sampling, so it may be difficult to estimate accurate concordance factors from sparse datasets with few loci.

Inferring the evolutionary history of the Metazoa is challenging due to the discordant phylogenetic signal and the large number of steps with potential for introducing systematic errors including as ortholog identification, alignment error, gene tree estimation and model misspecification. In this chapter, I applied CFs to investigate potential sources of systematic bias within this clade. I show that substantial topological discordance is present within each of the 12 matrices I investigated, potentially due to rapid radiation at the base of the metazoan tree. I also find CFs at the key branch decrease when a site-heterogeneous model is applied, consistent with LBA. My results suggest that traditional phylogenomic approaches are unlikely to resolve the evolutionary history of the Metazoa, and alternate approaches designed to detect deep phylogenetic signal may be required.

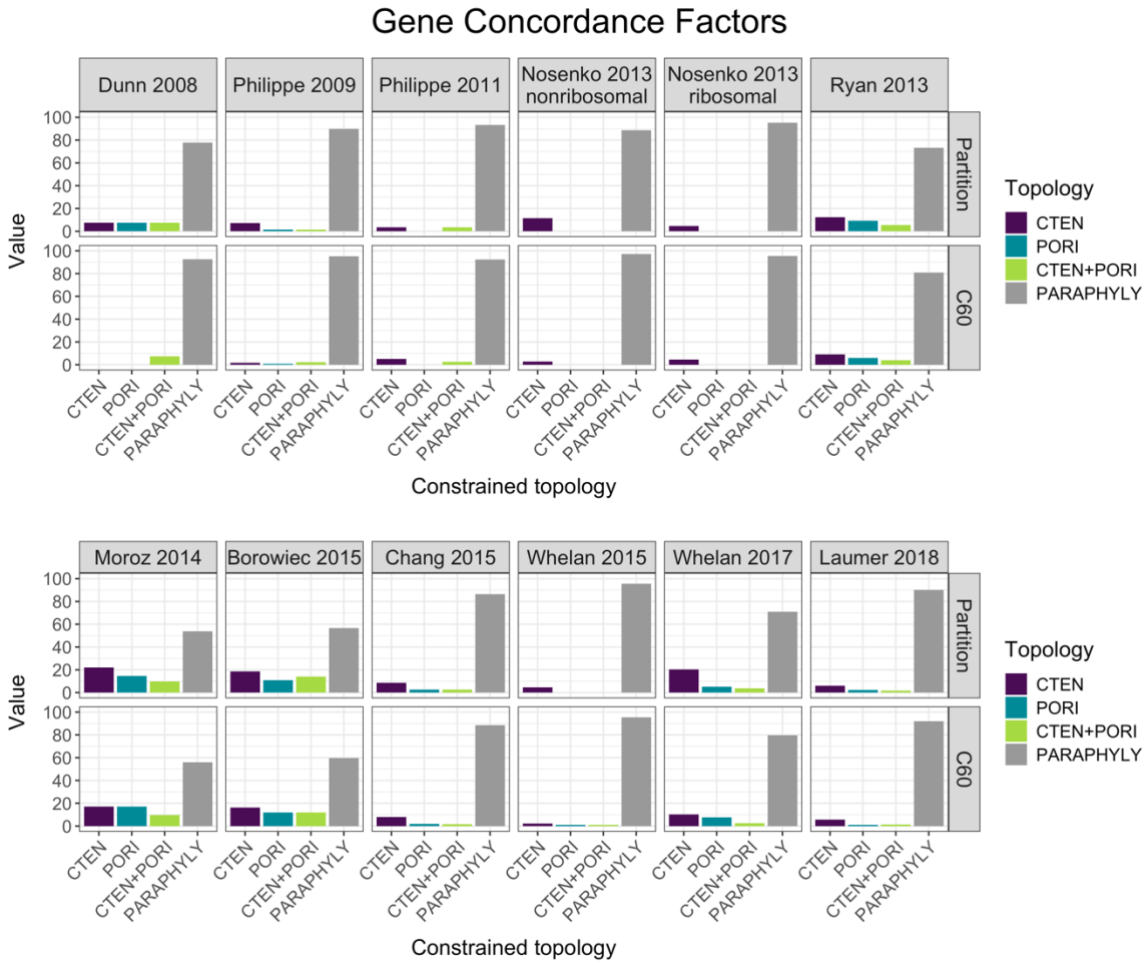
4.6 Data availability

All alignments were downloaded from the sources listed in Table 6. All scripts to replicate these analyses are available at the GitHub repository for this project (https://github.com/caitlinch/ancient_ILS). The alignments I used and other files generated during this analysis are available at the Figshare repository for this project “Ancient ILS” v2 (https://doi.org/10.6084/m9.figshare.25965172.v2), including alignments, partition files, trees, gene trees, and concordance output files.

4.7 Acknowledgments

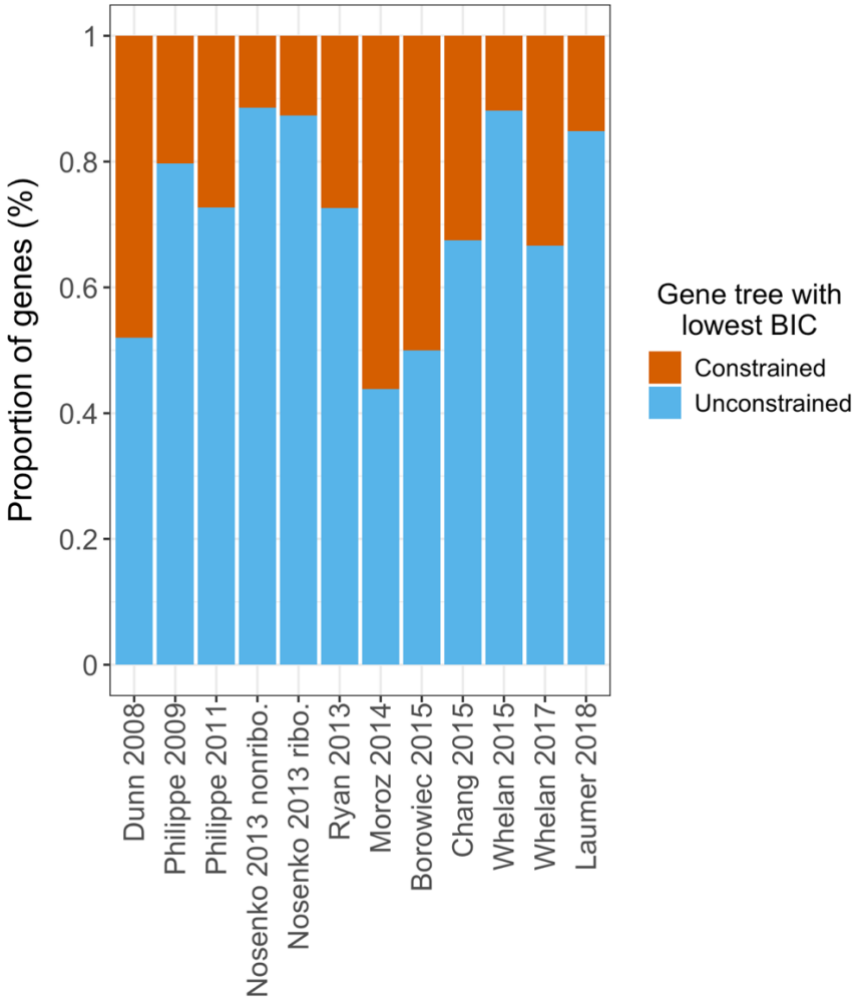
Thanks to Minh Bui, Maja Adamska and Barbara Holland for their comments on early versions of this chapter. Thanks to Rob Lanfear for advice on experimental design and for providing editorial comments. This research was funded by an Australian Government Research Training Program scholarship (to C.C.).

4.8 Supplementary figures



Supplementary Figure 24: Gene concordance factors (gCFs) including the gDFP (gene discordance factor – paraphyletic) around the key branch from 12 empirical phylogenetic matrices.

For each matrix, the results from the Partitioned model are shown above results from the C60 model. In the x axis, CTEN refers to constrained trees estimated with the Ctenophora-as the SOM; PORI refers to the constrained trees estimated with Porifera as the SOM; and CTEN+PORI refers to constrained trees estimated with the monophyletic clade of Ctenophora and Porifera as the SOM.



Supplementary Figure 25: Type of gene tree (either Constrained or Unconstrained) with best BIC value for genes in 12 empirical matrices.

One unconstrained tree and three constrained trees were estimated from each gene, and BIC scores from the constrained and unconstrained trees were compared to identify the best (i.e., lowest) BIC score for each gene. The terms "nonribo." and "ribo." refer to "non-ribosomal" and "ribosomal" respectively.

Discussion

This thesis aims to investigate the prevalence and impact of violating the treelikeness assumption during phylogenetic inference, and to apply new and established methods to assess phylogenetic signal and reduce systematic bias in non-treelike data. I used techniques from mathematics, statistics, computer science and evolutionary biology to investigate sources of reduced treelikeness and develop methods for working with non-treelike data.

The majority of phylogenetic methods apply models to simplify the complex and stochastic processes of evolution. No model is able to fully capture the complexity of evolutionary processes, as noted by Box (1979) who stated “All models are wrong but some are useful.” In the past decade, there has been growing awareness of the importance of assessing model adequacy and testing goodness of fit within the phylogenetic pipeline (Brown and Thomson 2018; Jermiin et al. 2020; Simion et al. 2020). A variety of tests have been developed to aid assessment of different assumptions included in phylogenetic models such as heterotachy (Wang et al. 2007; Crotty et al. 2020) or the stationary, reversible, and homogeneous assumptions (Ababneh et al. 2006; Naser-Khdour et al. 2019, 2021). Ideally, these tests are applied prior to or after tree estimation (depending on the test), with test results informing tree estimation and interpretation. Investigating phylogenetic assumptions prior to tree inference aids in selecting reasonable methods and parameters for phylogenetic parameters (Jermiin et al. 2020). In particular, the massive amounts of genomic data available have resulted in an awareness of the heterogeneity of phylogenetic signal, and development of methods to accommodate that heterogeneity (Delsuc et al. 2005; Jeffroy et al. 2006; Steenwyk et al. 2023).

The treelikeness assumption states that every site in an alignment shares an identical evolutionary history that fits a single bifurcating tree. This assumption is commonly included in phylogenetic methods. Concatenation methods assume that the entire alignment is treelike, and summary methods (also known as gene tree/species tree methods or two step methods) assume that each locus is treelike. The treelikeness of any multiple sequence alignment is violated by processes that disrupts purely vertical genetic transmission including recombination, introgression, hybridisation, or incomplete lineage sorting (ILS). Previous studies have shown that violation of the treelikeness assumption in empirical multiple sequence alignments is widespread. Previously, Smith et al. (2020) analysed 13 datasets from across the tree of life and found that 1 – 92% of datasets contained intragenic conflict. Similarly, exons from the same gene have been found to have different evolutionary histories (Scornavacca and Galtier 2017; Mendes et al. 2019). The level of treelikeness will be different in different datasets, depending on the evolutionary history of the selected taxa and the

evolutionary depth of the clade under consideration. As the number of taxa or sites included in an alignment increases, the frequency of recombination breakpoints increases and the length of any treelike segment of the alignment decreases (Springer and Gatesy 2016, 2018). In theory, any alignment used for tree inference is unlikely to be treelike (Gatesy and Springer 2014; Springer and Gatesy 2016, 2018). In practice, the extent to which treelikeness impacts tree inference remains relatively poorly studied.

In this thesis, I demonstrated multiple ways to assess the treelikeness of phylogenetic datasets and determine the impact on tree inference. In Chapter One, I identified three test statistics for treelikeness that can be applied to any multiple sequence alignment and developed a statistical test to assess the adequacy of the treelikeness assumption using a parametric bootstrap. These three tests are process-agnostic, in that they each aim to quantify treelikeness without detecting or quantifying any particular cause of reduced treelikeness. I assessed the performance of these tests on simulated and empirical data. I also provide an R implementation for the new test I developed, called the “tree proportion”. To my knowledge, this chapter presents the first comprehensive benchmarking of test statistics to quantify treelikeness. While there is potential to improve the performance of the best three tests, this chapter is a helpful reference for biologists to assess phylogenetic assumptions prior to tree estimation. This chapter fits within the broader movement towards reducing bias by assessing model assumptions prior to tree estimation (Jermiin et al. 2020).

Second, I present a number of methods for investigating specific causes of decreased treelikeness. In Chapter Two, I detect decreased treelikeness caused by recombination events and determine whether removing non-treelike loci improved tree inference. In Chapter Four I use topological variation at key branches within empirical phylogenies to identify evolutionary processes that result in decreased treelikeness. Both of these chapters aim to understand historical biological processes, and understand the impact these processes had on tree inference. However, in both chapters I investigated the signals of evolutionary history using existing published multiple sequence alignments. As analyses in these chapters were applied to completed sequence alignments, my results are impacted by choices made during dataset construction. The process of creating a multiple sequence alignment has many steps, each of which may bias phylogenetic tree inference (Ranwez and Chantret 2020). For example, Philippe et al. (2011b) inferred a different topology in 2/3 alignments that had previously been used to estimate the animal tree of life after correcting for sequencing errors, orthology errors, missing data, taxon sampling, incorrectly named taxa, and multiple substitutions. Existing multiple sequence alignments have been constructed with certain analyses in mind, which informs choices made during the alignment process (Morrison 2006; Ranwez and Chantret

2020). Creating a new alignment would result in more control over each of those intermediate choices, removing a potential source of bias and ensuring the final alignment is sufficient to investigate the particular processes of interest.

Finally, in Chapter Three I relax the treelikeness assumption completely and consider whether a single tree is sufficient to describe the evolutionary history of the animal tree of life. The Mixtures Across Sites and Trees (MAST) model (Wong et al. 2024), is a process-agnostic multitree mixture model which represents the evolutionary history of a given alignment as a mixture of bifurcating trees. I applied the MAST model to estimate the metazoan tree of life, a contentious and unresolved phylogeny. My results showed that for most datasets, a mixture of 2 or more trees was a better fit than a single tree. In previous analyses, the inferred metazoan tree topology depends on factors including model choice (Feuda et al. 2017a; Whelan and Halanych 2017; Kapli and Telford 2020; Redmond and McLysaght 2021a); outgroup choice (Nosenko et al. 2013a; Pisani et al. 2015; Li et al. 2021); orthology identification (Pett et al. 2019; Natsidis et al. 2021; McCarthy et al. 2023); matrix completeness (Sanderson et al. 2010; Roure et al. 2013); and alignment error (Pisani et al. 2015). Multiple studies have reanalysed previously-published metazoan datasets with different filtering or model parameters and inferred trees with different evolutionary hypotheses to the original analysis (Philippe et al. 2011b; Li et al. 2021; Redmond and McLysaght 2021a; McCarthy et al. 2023). Additionally, previous studies have established the existence of heterogeneous phylogenetic signal within empirical datasets (Shen et al. 2017; Smith et al. 2020), and particularly the high levels of conflicting signal within metazoan datasets (Arcila et al. 2017; Shen et al. 2017; Kapli and Telford 2020; Szánthó et al. 2023). On average, applying the MAST model resulted in a mixture of trees with a tree weight of around 40% for Porifera as the sister of all other Metazoa (SOM), and 60% for Ctenophora. Given the similarity of the two tree weights, slight changes in the phylogenetic pipeline could plausibly flip the preferred tree topology if it is inferred assuming that a single bifurcating tree topology is adequate to represent the data. In some studies, such as those investigating the evolution of complex traits such as nervous or gut systems, a single metazoan tree is often the basis for drawing evolutionary hypotheses. Given that my work shows overwhelming evidence that there more than one bifurcating tree is required to explain metazoan evolution, I recommend either considering multiple evolutionary pathways using the mixture weights, or focusing on specific genes that are known to be important in the system under consideration.

Throughout this thesis, I have used the metazoan phylogeny as an example of a clade with a complex evolutionary history. The Metazoa contains all animal clades: Porifera (sponges), Ctenophora (comb jellies), Cnidaria (aquatic animals such as jellyfish, sea anemones or

corals), Placozoa (simple blob-like marine organisms), and Bilateria (all other animals e.g., vertebrates, arthropods, molluscs, etc.). The phylogenetic placement of these clades has potential implications for understanding the evolution of complex traits such as central nervous systems or digestive systems (Ryan and Chiodin 2015; Presnell et al. 2016). The relationships between animal clades are currently unresolved and subject to contentious debate, particularly focused on whether Ctenophora or Porifera was the first clade to diverge from all other animals.

Multiple factors contribute to the difficulty of resolving the metazoan tree. Short branches separating clades at the base of the metazoan tree suggest that the Metazoa underwent a rapid radiation (Rokas et al. 2005; Rokas and Carroll 2006). Tree inference is further complicated by the ancient timing of these diversifications and the variation in evolutionary rates across the Metazoa, particularly the rapid rate of evolution for Ctenophora species (Kohn et al. 2012; Wang and Cheng 2019). This results in substitutional saturation and homoplasy within metazoan multiple sequence alignments, which can reduce phylogenetic accuracy (Philippe et al. 2011b; Dunn et al. 2014; King and Rokas 2017). Further adding to the complexity of species tree estimation, phylogenetic signal varies within metazoan genomes. Protein structural environment impacts tree inference, with different trees inferred from exposed sites (i.e., sites located on the protein surface) and the slower-evolving buried sites (Pandey and Braun 2020, 2021). Similarly, choice of genes can bias tree inference. Mitochondrial genomes are fast-evolving in Bilateria, thus selection of mitochondrial or nuclear genes during dataset construction will impact downstream analyses including tree inference (Philippe et al. 2011b). The functional class of genes selected for analysis is also important, with one study showing distinct trees were inferred from two matrices consisting of only ribosomal or only non-ribosomal genes respectively (Nosenko et al. 2013a). Each of these factors further complicates tree inference. Analysing metazoan datasets is therefore a useful case-study which helps to develop a set of tools that can be applied to other complex and contentious datasets. While the approaches applied in this thesis were unable to definitively resolve the animal tree of life, the methods applied throughout enabled an investigation into the causes and impacts of heterogeneous phylogenetic signal. These methods can be applied to other contentious clades to interrogate conflicting signal and investigate incongruence.

Given the levels of discordant signal within metazoan datasets, my work suggests that resolving this tree may be a misleading aim, and that work should instead focus on resolving the number and identity of trees necessary to explain the evolutionary history of metazoan genomes. There exist a number of methods which can be used to aid with this goal, many of which go well beyond what is presented in this thesis. One approach is to use gene content,

where each character represents the binary presence or absence of homologous gene families. One Bayesian analysis of homologous gene content family estimated Porifera as the SOM (Pett et al. 2019), as did an alternative analysis combining gene content and morphological traits (Juravel et al. 2023). However, gene content inferences rely on accurate orthology inference, and therefore systematic biases in orthology inference will also impact trees estimated from gene content data (Natsidis et al. 2021). Determining orthology of genes within the Metazoa is difficult due to the deep evolutionary timescale and rapid radiation at the root of the metazoan tree, with a previous study finding only 17–33% of loci in 5 published metazoan datasets could resolve ≥3 metazoan clades (McCarthy et al. 2023). In addition, morphological data also has the potential to bias inference of metazoan tree topology (Neumann et al. 2021). Another approach is to assess conserved synteny across the metazoan tree. Schultz et al. (2023) applied this approach and found support for Ctenophora as the SOM, identifying 7 sets of genes where Ctenophora shared linkages with single cell eukaryotes, and all other clades (Bilateria, Cnidaria, Placozoa, and Porifera) were united. However, this approach also relies on accurate identification of orthologous genes and is therefore also subject to systematic bias in orthology identification.

An alternate approach for evaluating deep relationships within protein sequences is to incorporate structural protein information. Protein structure is more constrained and therefore more conserved than genetic sequences, so comparing protein structures may reveal deep evolutionary signals (Illergård et al. 2009). Information about protein structure has previously been incorporated into evolutionary analyses (Lake et al. 1984; Thorne et al. 1996; Choi et al. 2007; Le and Gascuel 2010; Lai et al. 2020), but recent methods allow phylogenies to be directly estimated from protein structures (Malik et al. 2020; Puente-Lelievre et al. 2024). Alignments can be generated from protein structures, for example by calculating pairwise structural comparisons resulting in a distance matrix with a set of distances between the central carbon in each amino acid (Malik et al. 2020), or by encoding tertiary structures into character states (Puente-Lelievre et al. 2024; van Kempen et al. 2024). A recent study of the ferritin-like superfamily of proteins compared an AA alignment and a tertiary protein structure alignment, and found that the tree best matching the hypothesised evolutionary history of this group was obtained by combining the AA and structure alignments, plus applying custom substitution models (Puente-Lelievre et al. 2024). The ease of estimating phylogenies from protein structures is increasing due to the development of tools to construct structural alignments and phylogenies (Pei et al. 2008; Cao et al. 2023; Malik et al. 2023; Moi et al. 2023; Puente-Lelievre et al. 2024; van Kempen et al. 2024), and the existence of databases such as the Protein Data Bank (Berman et al. 2003; wwPDB consortium 2019; Baskaran et al. 2024) and the AlphaFold

Protein Structure Database (Jumper et al. 2021; Varadi et al. 2022) consisting of >200,000 and >360,000 protein structures respectively. Protein structure may be particularly useful in the case of deep divergences where protein sequences are poorly conserved. For example, two studies have applied structural phylogenetics to investigate the relationships within the Ferritin-like superfamily (Lundin et al. 2012; Puente-Lelievre et al. 2024). Ferritin-like proteins have low sequence similarity as proteins in this family have evolved diverse functions over a deep evolutionary timescale, such that any phylogeny for this group of proteins must include viruses, bacteria, and eukaryotes (Lundin et al. 2012). The evolutionary timescale of the Metazoa is relatively short compared to the Ferritin-like superfamily but there is still substantial discordance of phylogenetic signal, and structural phylogenetics may identify conserved phylogenetic signal that other phylogenetic analyses (such as phylogenomic tree reconstruction, CFs, or CTA) are unable to reconstruct.

Throughout this thesis, I have focused on the impact of non-treelikeness for maximum likelihood (ML) and summary (also known as two-step or gene-tree/species-tree) methods. Other phylogenetic methods include distance methods, parsimony, and Bayesian inference (Kapli et al. 2020). Bayesian methods use statistical distributions to represent model parameters and account for uncertainty during tree estimation (Yang and Rannala 2012; Rannala et al. 2020). The output is a posterior distribution of trees with estimated probabilities, which helps account for uncertainty around the true evolutionary history (Huelsenbeck et al. 2000). Bayesian methods are commonly used for tree inference in empirical phylogenetic studies (Miya et al. 2005; McGuire et al. 2007; Zuriaga et al. 2009; Lavretsky et al. 2014; Mitchell et al. 2014; Cannon et al. 2016; Laumer et al. 2018a, 2019a; Allio et al. 2020; Roycroft et al. 2020; Hánová et al. 2021; Li et al. 2021; Strassert et al. 2021; Ballesteros et al. 2022; Barley et al. 2022; Wolfe et al. 2023; Höhna and Hsiang 2024) and multiple software programs have been developed to perform Bayesian inference including MrBayes (Ronquist et al. 2012), BEAST (Bouckaert et al. 2019), StarBeast (Douglas et al. 2022), and PhyloBayes (Lartillot 2020a). Unfortunately Bayesian inference are very computationally intensive due to the high number of parameters, which complicates phylogenetic inference of large genomic datasets (Lartillot 2020a; Barido-Sottani et al. 2023). Notably, multi-tree mixture models such as those used in Chapter Three are not yet implemented in a Bayesian framework.

Like most evolutionary models, Bayesian inference methods include underlying assumptions. Previous studies have identified causes of systematic bias in Bayesian analyses including missing data (Lemmon et al. 2009; Roure et al. 2013), model misspecification (Lemmon and Moriarty 2004), and long branch attraction (Kolaczkowski and Thornton 2009). Unlike most ML methods some Bayesian methods incorporate the multispecies coalescent (MSC) (Lartillot

2020b; Rannala et al. 2020). The MSC is a stochastic model that tracks the genealogical history of individuals sampled within a population backwards in time (Bryant and Hahn 2020). Importantly, the MSC allows for different loci to have different evolutionary history and therefore explicitly incorporates non-treelike evolutionary processes. I expect that Bayesian methods incorporating the MSC would be more robust to non-treelike data, but further research is necessary to determine the impact of treelikeness on these methods. I expect that the majority of phylogenetic relationships would be similar regardless of tree inference method (similar to the results in Chapter Two), but the resolution of nodes impacted by biological processes such as ILS or recombination may change. This expectation is supported by a previous study, which investigated the difference in three tree estimation methods (ML, maximum parsimony, and Bayesian inference) by estimating a tree from each method for 157 empirical phylogenetic datasets (Torres et al. 2021). They found that the majority of nodes were the same regardless of tree estimation method (>89%), and nodes that differed tended to be relationships that were already considered contentious.

The aim of phylogenetics is to reconstruct the evolutionary history of species, populations or sites in a genome. Advances in sequencing technologies have resulted in the availability of enormous amounts of genomic data (Kapli et al. 2020). Despite the increase in dataset size, tree inference is still complicated due to the complexity and heterogeneity of evolutionary processes (Steenwyk et al. 2023). Although the treelikeness assumption is incorporated into many phylogenetic methods, purely treelike evolution is rare due to biological processes such as introgression and ILS. Therefore, considering heterogeneous signal is important when conducting phylogenomic analyses. In this thesis I have explored different causes of decreased treelikeness across the tree of life, and investigated different methods to quantify and mitigate decreased treelikeness in phylogenetic datasets. I hope this thesis increases awareness of the importance of the treelikeness assumption in phylogenetic methods, and encourages researchers to assess the adequacy of the treelikeness assumption prior to tree inference.

References

Ababneh F., Jermiin L.S., Ma C., Robinson J. 2006. Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics. 22:1225–1231.

Abadi S., Avram O., Rosset S., Pupko T., Mayrose I. 2020. ModelTeller: model selection for optimal phylogenetic reconstruction using machine learning. Mol. Biol. Evol. 37: 3338–3352.

Akaike H. 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control. 19:716–723.

Allio R., Scornavacca C., Nabholz B., Clamens A.-L., Sperling F.A., Condamine F.L. 2020. Whole genome shotgun phylogenomics resolves the pattern and timing of swallowtail butterfly evolution. Syst. Biol. 69:38–60.

Allman E.S., Baños H., Rhodes J.A. 2019. NANUQ: a method for inferring species networks from gene trees under the coalescent model. Algorithms Mol. Biol. 14:24.

Amster G., Sella G. 2016. Life history effects on the molecular clock of autosomes and sex chromosomes. Proc. Natl. Acad. Sci. U. S. A. 113:1588–1593.

Ané C. 2021. cecileane/QuartetNetworkGoodnessFit.jl. Available from https://github.com/cecileane/QuartetNetworkGoodnessFit.jl.

Arafat H., Alamaru A., Gissi C., Huchon D. 2018. Extensive mitochondrial gene rearrangements in Ctenophora: insights from benthic Platyctenida. BMC Evol. Biol. 18:65.

Arcila D., Ortí G., Vari R., Armbruster J.W., Stiassny M.L.J., Ko K.D., Sabaj M.H., Lundberg J., Revell L.J., Betancur-R R. 2017. Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. Nat. Ecol. Evol. 1:0020.

Arenas M. 2015. Trends in substitution models of molecular evolution. Front. Genet. 6:1:9.

Ballesteros J., Santibanez-Lopez C., Baker C., Benavides L., Cunha T., Gainett G., Ontano A., Setton E., Arango C., Gavish-Regev E., Harvey M., Wheeler W., Hormiga G., Giribet G., Sharma P. 2022. Comprehensive species sampling and sophisticated algorithmic approaches refute the monophyly of Arachnida. Mol. Biol. Evol. 39:msac021.

Balvočūtė M., Spillner A., Moulton V. 2014. FlatNJ: a novel network-based approach to visualize evolutionary and biogeographical relationships. Syst. Biol. 63:383–396.

Bandelt H.-J., Dress A.W.M. 1992. Split decomposition: a new and useful approach to phylogenetic analysis of distance data. Mol. Phylogenet. Evol. 1:242–252.

Baños H., Susko E., Roger A.J. 2024. Is over-parameterization a problem for profile mixture models? Syst. Biol. 73:53–75.

Baric S., Salzburger W., Sturmbauer C. 2003. Phylogeography and evolution of the Tanganyikan Cichlid genus Tropheus based upon mitochondrial DNA Sequences. J. Mol. Evol. 56:54–68.

Barido-Sottani J., Schwery O., Warnock R.C.M., Zhang C., Wright A.M. 2023. Practical guidelines for Bayesian phylogenetic inference using Markov Chain Monte Carlo (MCMC) [version 1; peer review: 3 approved, 1 approved with reservations]. Open Res. Eur. 3.

Barley A.J., Nieto-Montes de Oca A., Manríquez-Morán N.L., Thomson R.C. 2022. The evolutionary network of whiptail lizards reveals predictable outcomes of hybridization. Science. 377:773–777.

Baskaran K., Ploskon E., Tejero R., Yokochi M., Harrus D., Liang Y., Peisach E., Persikova I., Ramelot T.A., Sekharan M., Tolchard J., Westbrook J.D., Bardiaux B., Schwieters C.D., Patwardhan A., Velankar S., Burley S.K., Kurisu G., Hoch J.C., Montelione G.T., Vuister G.W., Young J.Y. 2024. Restraint validation of biomolecular structures determined by NMR in the Protein Data Bank. Structure. 32:824-837.e1.

Bastkowski S., Spillner A., Moulton V. 2014. Fishing for minimum evolution trees with Neighbor-Nets. Inf. Process. Lett. 114:13–18.

Baum D.A. 2007. Concordance trees, concordance factors, and the exploration of reticulate genealogy. TAXON. 56:417–426.

Berman H., Henrick K., Nakamura H. 2003. Announcing the worldwide Protein Data Bank. Nat. Struct. Mol. Biol. 10:980–980.

Bezanson J., Edelman A., Karpinski S., Shah V.B. 2017. Julia: a fresh approach to numerical computing. SIAM Rev. 59:65–98.

Biczok R., Bozsoky P., Eisenmann P., Ernst J., Ribizel T., Scholz F., Trefzer A., Weber F., Hamann M., Stamatakis A. 2018. Two C++ libraries for counting trees on a phylogenetic terrace. Bioinformatics. 34:3399–3401.

Bishop M.J., Friday A.E., Brenner S. 1997. Evolutionary trees from nucleic acid and protein sequences. Proc. R. Soc. B Biol. Sci. 226:271–302.

Bishop M.J., Friday A.E., Patterson C. 1987. Tetropad relationships: the molecular evidence. Molecules and morphology in evolution: conflict or compromise? Cambridge, England.: University Press. p. 123–139.

Blair C., Ané C. 2020. Phylogenetic trees and networks can serve as powerful and complementary approaches for analysis of genomic data. Syst. Biol. 69:593–601.

Blischak P.D., Chifman J., Wolfe A.D., Kubatko L.S. 2018. HyDe: a python package for genome-scale hybridization detection. Syst. Biol. 67:821–829.

Borchiellini C., Chombard C., Lafay B., Boury-esnault N. 2000. Molecular systematics of sponges (Porifera). Hydrobiologia. 420:15–27.

Borchiellini C., Manuel M., Alivon E., Boury-Esnault N., Vacelet J., Le Parco Y. 2008. Sponge paraphyly and the origin of Metazoa. J. Evol. Biol. 14:171–179.

Borowiec M.L., Lee E.K., Chiu J.C., Plachetzki D.C. 2015. Extracting phylogenetic signal and accounting for bias in whole-genome data sets supports the Ctenophora as sister to remaining Metazoa. BMC Genomics. 16:987.

Borowiec M.L., Lee E.K., Chiu J.C., Plachetzki D.C. 2016. Data from: Extracting phylogenetic signal and accounting for bias in whole-genome data sets supports the Ctenophora as sister to remaining Metazoa. Version 1, Dataset. Dryad. Available from https://doi.org/10.5061/dryad.k6tq2. .

Bouchet-Valat M., Kamiński B. 2023. DataFrames.jl: flexible and fast tabular data in Julia. J. Stat. Softw. 107:1–32.

Bouckaert R., Vaughan T.G., Barido-Sottani J., Duchêne S., Fourment M., Gavryushkina A., Heled J., Jones G., Kühnert D., Maio N.D., Matschiner M., Mendes F.K., Müller N.F., Ogilvie H.A., Plessis L. du, Popinga A., Rambaut A., Rasmussen D., Siveroni I., Suchard M.A., Wu C.-H., Xie D., Zhang C., Stadler T., Drummond A.J. 2019. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLOS Comput. Biol. 15:e1006650.

Bourke B.P., Justi S.A., Caicedo-Quiroga L., Pecor D.B., Wilkerson R.C., Linton Y.-M. 2021. Phylogenetic analysis of the Neotropical Albitarsis Complex based on mitogenome data. Parasit. Vectors. 14:589.

Box G.E.P. 1979. Robustness in the strategy of scientific model building. In: Launer R.L., Wilkinson G.N., editors. Robustness in Statistics. Academic Press. p. 201–236.

Braun E.L. 2018. An evolutionary model motivated by physicochemical properties of amino acids reveals variation among proteins. Bioinformatics. 34:i350–i356.

Bravo G.A., Antonelli A., Bacon C.D., Bartoszek K., Blom M.P.K., Huynh S., Jones G., Knowles L.L., Lamichhaney S., Marcussen T., Morlon H., Nakhleh L.K., Oxelman B., Pfeil B., Schliep A., Wahlberg N., Werneck F.P., Wiedenhoeft J., Willows-Munro S., Edwards S.V. 2019. Embracing heterogeneity: coalescing the tree of life and the future of phylogenomics. PeerJ. 7:e6399.

Brown J.M. 2014. Detection of implausible phylogenetic inferences using posterior predictive assessment of model fit. Syst. Biol. 63:334–348.

Brown J.M., ElDabaje R. 2009. PuMA: Bayesian analysis of partitioned (and unpartitioned) model adequacy. Bioinformatics. 25:537–538.

Brown J.M., Thomson R.C. 2017. Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. Syst. Biol. 66:517–530.

Brown J.M., Thomson R.C. 2018. Evaluating model performance in evolutionary biology. Annu. Rev. Ecol. Evol. Syst. 49:95–114.

Brown W.M., Prager E.M., Wang A., Wilson A.C. 1982. Mitochondrial DNA sequences of primates: tempo and mode of evolution. J. Mol. Evol. 18:225–239.

Bruen T. 2005. PhiPack. Online, available at https://www.maths.otago.ac.nz/~dbryant/software.html.

Bruen T.C., Philippe H., Bryant D. 2006. A simple and robust statistical test for detecting the presence of recombination. Genetics. 172:2665.

Bryan H.J. 2022. _glue: interpreted string literals_. Version 1.6.2. Available from https://CRAN.R-project.org/package=glue.

Bryant D., Hahn M.W. 2020. The concatenation question. In: Scornavacca C., Delsuc F., Galtier N., editors. Phylogenetics in the Genomic Era. No commercial publisher | Authors open access book. p. 3.4:1-3.4:23.

Bryant D., Moulton V. 2004. Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. Mol. Biol. Evol. 21:255–265.

Budd G.E. 2008. The earliest fossil record of the animals and its significance. Philos. Trans. R. Soc. B Biol. Sci. 363:1425–1434.

Buneman P. 1971. The recovery of trees from measures of dissimilarity. In: Hodson F.R., Kendall D.G., Tăutu P., editors. Mathematics in the archaeological and historical sciences: proceedings of the Anglo-Romanian Conference, Mamaia. Edinburgh: Edinburgh University Press. p. 387–395.

Burki F., Kaplan M., Tikhonenkov D.V., Zlatogursky V., Minh B.Q., Radaykina L.V., Smirnov A., Mylnikov A.P., Keeling P.J. 2016. Untangling the early diversification of eukaryotes: a phylogenomic study of the evolutionary origins of Centrohelida, Haptophyta and Cryptista. Proc. R. Soc. B Biol. Sci. 283:20152802.

Burnham K.P., Anderson D.R. 2002. Model selection and multimodel inference : a practical information-theoretic approach. New York, NY: Springer.

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. Syst. Biol. 70:491–507.

Cai R., Ané C. 2021. Assessing the fit of the multi-species network coalescent to multi-locus data. Bioinformatics. 37:634–641.

Cannon J.T., Vellutini B.C., Smith J., Ronquist F., Jondelius U., Hejnol A. 2016. Xenacoelomorpha is the sister group to Nephrozoa. Nature. 530:89–93.

Cao W., Wu L.-Y., Xia X.-Y., Chen X., Wang Z.-X., Pan X.-M. 2023. A sequence-based evolutionary distance method for phylogenetic analysis of highly divergent proteins. Sci. Rep. 13:20304.

Cavalli-Sforza L.L., Piazza A. 1975. Analysis of evolution: evolutionary rates, independence and treeness. Theor. Popul. Biol. 8:127–165.

Chakerian J., Holmes S. 2020. distory: distance between phylogenetic histories. R package version 1.4.4. Available from https://CRAN.R-project.org/package=distory.

Chakraborty R. 1977. Estimation of time of divergence from phylogenetic studies. Can. J. Genet. Cytol. 19:217–223.

Chan K.O., Hutter C.R., Wood P.L., Grismer L.L., Brown R.M. 2020. Larger, unfiltered datasets are more effective at resolving phylogenetic conflict: Introns, exons, and UCEs resolve ambiguities in Golden-backed frogs (Anura: Ranidae; genus Hylarana). Mol. Phylogenet. Evol. 151:106899.

Chang E.S., Neuhof M., Rubinstein N.D., Diamant A., Philippe H., Huchon D., Cartwright P. 2015. Genomic insights into the evolutionary origin of Myxozoa within Cnidaria. Proc. Natl. Acad. Sci. U. S. A. 112:14912.

Charif D., Lobry J.R. 2007. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Bastolla U., Porto M., Roman H.E., Vendruscolo M., editors. Structural approaches to sequence evolution: Molecules, networks, populations. New York: Springer Verlag. p. 207–232.

Charr J.-C., Garavito A., Guyeux C., Crouzillat D., Descombes P., Fournier C., Ly S.N., Raharimalala E.N., Rakotomalala J.-J., Stoffelen P., Janssens S., Hamon P., Guyot R. 2020. Complex evolutionary history of coffees revealed by full plastid genomes and 28,800 nuclear SNP analyses, with particular emphasis on Coffea canephora (Robusta coffee). Mol. Phylogenet. Evol. 151:106906.

Chen M.-Y., Liang D., Zhang P. 2015. Selecting question-specific genes to reduce incongruence in phylogenomics: a case study of jawed vertebrate backbone phylogeny. Syst. Biol. 64:1104–1120.

Chernikova D., Motamedi S., Csürös M., Koonin E.V., Rogozin I.B. 2011. A late origin of the extant eukaryotic diversity: divergence time estimates using rare genomic changes. Biol. Direct. 6:26.

Chernomor O., von Haeseler A., Minh B.Q. 2016. Terrace aware data structure for phylogenomic inference from supermatrices. Syst. Biol. 65:997–1008.

Chi P.B., Kim D., Lai J.K., Bykova N., Weber C.C., Kubelka J., Liberles D.A. 2018. A new parameter-rich structure-aware mechanistic model for amino acid substitution during evolution. Proteins Struct. Funct. Bioinforma. 86:218–228.

Choi S.C., Hobolth A., Robinson D.M., Kishino H., Thorne J.L. 2007. Quantifying the impact of protein tertiary structure on molecular evolution. Mol. Biol. Evol. 24:1769–1782.

Cobb M. 2017. 60 years ago, Francis Crick changed the logic of biology. PLOS Biol. 15:e2003243.

Coleman G.A., Davín A.A., Mahendrarajah T.A., Szánthó L.L., Spang A., Hugenholtz P., Szöllősi G.J., Williams T.A. 2021. A rooted phylogeny resolves early bacterial evolution. Science. 372.

Collins A.G. 1998. Evaluating multiple alternative hypotheses for the origin of Bilateria: An analysis of 18S rRNA molecular evidence. Proc. Natl. Acad. Sci. U. S. A. 95:15458–15463.

Condamine F.L., Rolland J., Morlon H. 2019. Assessing the causes of diversification slowdowns: temperature-dependent and diversity-dependent models receive equivalent support. Ecol. Lett. 22:1900–1912.

Corlett R.T. 2016. Plant diversity in a changing world: status, trends, and conservation needs. Plant Divers. 38:10–16.

Cranston K.A., Rannala B. 2007. Summarizing a posterior distribution of trees using agreement subtrees. Syst. Biol. 56:578–590.

Crick F.H. 1958. “On protein synthesis”. The biological replication of macromolecules. Symp. Soc. Exp. Biol. 12:138–163.

Crisp M.D., Minh B.Q., Choi B., Edwards R.D., Hereward J., Kulheim C., Lin Y.P., Meusemann K., Thornhill A.H., Toon A., Cook L.G. 2024. Perianth evolution and implications for generic delimitation in the eucalypts (Myrtaceae), including the description of the new genus, Blakella. J. Syst. Evol. 62:942–962.

Crotty S., Holland B. 2022. Comparing partitioned models to mixture models: do information criteria apply? Syst. Biol. 71:1541–1548.

Crotty S.M., Minh B.Q., Bean N.G., Holland B.R., Tuke J., Jermiin L.S., Haeseler A.V. 2020. GHOST: recovering historical signal from heterotachously evolved sequence alignments. Syst. Biol. 69:249–264.

Crouch J.A. 2014. Colletotrichum caudatum s.l. is a species complex. IMA Fungus. 5:17–30.

Cruaud A., Rasplus J.-Y., Zhang J., Burks R., Delvare G., Fusu L., Gumovsky A., Huber J.T., Janšta P., Mitroiu M.-D., Noyes J.S., van Noort S., Baker A., Böhmová J., Baur H., Blaimer B.B., Brady S.G., Bubeníková K., Chartois M., Copeland R.S., Dale-Skey Papilloud N., Dal Molin A., Dominguez C., Gebiola M., Guerrieri E., Kresslein R.L., Krogmann L., Lemmon E., Murray E.A., Nidelet S., Nieves-Aldrey J.L., Perry R.K., Peters R.S., Polaszek A., Sauné L., Torréns J., Triapitsyn S., Tselikh E.V., Yoder M., Lemmon A.R., Woolley J.B., Heraty J.M. 2024. The Chalcidoidea bush of life: evolutionary history of a massive radiation of minute wasps. Cladistics. 40:34–63.

Cummins C.A., McInerney J.O. 2011. A method for inferring the rate of evolution of homologous characters that can potentially improve phylogenetic inference, resolve deep divergence and correct systematic biases. Syst. Biol. 60:833–844.

Cunha T.J., Reimer J.D., Giribet G. 2022. Investigating sources of conflict in deep phylogenomics of Vetigastropod snails. Syst. Biol. 71:1009–1022.

Cunningham J.P. 1978. Free trees and bidirectional trees as representations of psychological distance. J. Math. Psychol. 17:165–188.

Dale G., Dieters M. 2007. Economic returns from environmental problems: breeding salt- and drought-tolerant eucalypts for salinity abatement and commercial forestry. Ecol. Eng. 31:175–182.

Dang C.C., Le Q.S., Gascuel O., Le V.S. 2010. FLU, an amino acid substitution model for influenza proteins. BMC Evol. Biol. 10:99.

Dang C.C., Minh B.Q., McShea H., Masel J., James J.E., Vinh L.S., Lanfear R. 2022. nQmaker: estimating time nonreversible amino acid substitution models. Syst. Biol. 71:1110–1123.

Darriba D., Posada D., Kozlov A.M., Stamatakis A., Morel B., Flouri T. 2020. ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. Mol. Biol. Evol. 37:291–294.

Darriba D., Taboada G.L., Doallo R., Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat. Methods. 9:772–772.

Darwin C. 1987. Charles Darwin's notebooks, 1836–1844: geology, transmutation of species, metaphysical enquiries. London: British Museum (Natural History).

Dasmahapatra K.K., Walters J.R., Briscoe A.D., Davey J.W., Whibley A., Nadeau N.J., Zimin A.V., Hughes D.S.T., Ferguson L.C., Martin S.H., Salazar C., Lewis J.J., Adler S., Ahn S.-J., Baker D.A., Baxter S.W., Chamberlain N.L., Chauhan R., Counterman B.A., Dalmay T., Gilbert L.E., Gordon K., Heckel D.G., Hines H.M., Hoff K.J., Holland P.W.H., Jacquin-Joly E., Jiggins F.M., Jones R.T., Kapan D.D., Kersey P., Lamas G., Lawson D., Mapleson D., Maroja L.S., Martin A., Moxon S., Palmer W.J., Papa R., Papanicolaou A., Pauchet Y., Ray D.A., Rosser N., Salzberg S.L., Supple M.A., Surridge A., Tenger-Trolander A., Vogel H., Wilkinson P.A., Wilson D., Yorke J.A., Yuan F., Balmuth A.L., Eland C., Gharbi K., Thomson M., Gibbs R.A., Han Y., Jayaseelan J.C., Kovar C., Mathew T., Muzny D.M., Ongeri F., Pu L.-L., Qu J., Thornton R.L., Worley K.C., Wu Y.-Q., Linares M., Blaxter M.L., ffrench-Constant R.H., Joron M., Kronforst M.R., Mullen S.P., Reed R.D., Scherer S.E., Richards S., Mallet J., Owen McMillan W., Jiggins C.D. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature. 487:94.

Davis C.C., Webb C.O., Wurdack K.J., Jaramillo C.A., Donoghue M.J. 2005. Explosive radiation of Malpighiales supports a mid-cretaceous origin of modern tropical rain forests. Am. Nat. 165:E36–E65.

Dayhoff M.O., Schwartz R.M., Orcutt B.C. 1978. A model of evolutionary change in proteins. In: Dayhoff M.O., editor. Atlas of Protein Sequence and Structure. National Biomedical Research Foundation: Washington, DC. p. 345–352.

Dayrat B. 2003. Roots of phylogeny: how did Haeckel build his trees? Syst. Biol. 52:515–527.

DeBiasse M.B., Nelson B.J., Hellberg M.E. 2014. Evaluating summary statistics used to test for incomplete lineage sorting: mito-nuclear discordance in the reef sponge Callyspongia vaginalis. Mol. Ecol. 23:225–238.

Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24:332–340.

Delsuc F., Brinkmann H., Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. 6:361–375.

Dimmic M.W., Rest J.S., Mindell D.P., Goldstein R.A. 2002. rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. J. Mol. Evol. 55:65–73.

Dohrmann M., Janussen D., Reitner J., Collins A.G., Wörheide G. 2008. Phylogeny and evolution of glass sponges (Porifera, Hexactinellida). Syst. Biol. 57:388–405.

Domazet-Lošo M., Domazet-Lošo T. 2016. gmos: rapid detection of genome mosaicism over short evolutionary distances. PLOS ONE. 11:e0166602.

Doolittle R.F., Feng D.F. 1987. Reconstructing the evolution of vertebrate blood coagulation from a consideration of the amino acid sequences of clotting proteins. Cold Spring Harb. Symp. Quant. Biol. 52:869–874.

Dopazo J., Dress A., von Haeseler A. 1993. Split decomposition: a technique to analyze viral evolution. Proc. Natl. Acad. Sci. U. S. A. 90:10320–10324.

Douglas J., Jiménez-Silva C.L., Bouckaert R. 2022. StarBeast3: adaptive parallelized Bayesian inference under the multispecies coalescent. Syst. Biol. 71:901–916.

Duchêne D.A., Duchêne S., Ho S.Y.W. 2017. New statistical criteria detect phylogenetic bias caused by compositional heterogeneity. Mol. Biol. Evol. 34:1529–1534.

Dunn C.W., Giribet G., Edgecombe G.D., Hejnol A. 2014. Animal phylogeny and its evolutionary implications. Annu. Rev. Ecol. Evol. Syst. 45:371–395.

Dunn C.W., Hejnol A., Matus D.Q., Pang K., Browne W.E., Smith S.A., Seaver E., Rouse G.W., Obst M., Edgecombe G.D., Sørensen M.V., Haddock S.H.D., Schmidt-Rhaesa A., Okusu A., Kristensen R.M., Wheeler W.C., Martindale M.Q., Giribet G. 2008. Broad

phylogenomic sampling improves resolution of the animal tree of life. Nature. 452:745–749.

Edelman N.B., Frandsen P.B., Miyagi M., Clavijo B., Davey J., Dikow R.B., García-Accinelli G., Van Belleghem S.M., Patterson N., Neafsey D.E., Challis R., Kumar S., Moreira G.R.P., Salazar C., Chouteau M., Counterman B.A., Papa R., Blaxter M., Reed R.D., Dasmahapatra K.K., Kronforst M., Joron M., Jiggins C.D., McMillan W.O., Di Palma F., Blumberg A.J., Wakeley J., Jaffe D., Mallet J. 2019. Genomic architecture and introgression shape a butterfly radiation. Science. 366:594.

Edwards A.W.F., Cavalli-Sforza L.L. 1963. The reconstruction of evolution. Heredity. 18:553.

Edwards A.W.F., Cavalli-Sforza L.L. 1964. Reconstruction of evolutionary trees. In: Heywood V.H., McNeill J., editors. Phenetic and Phylogenetic Classification. London: Systematics Association. p. 67–76.

Eigen M., Winkler-Oswatitsch R., Dress A. 1988. Statistical geometry in sequence space: a method of quantitative comparative sequence analysis. Proc. Natl. Acad. Sci. U. S. A. 85:5913.

Erpenbeck D., Wörheide G. 2007. On the molecular phylogeny of sponges (Porifera). Zootaxa. 1668:107–126.

Erwin D.H. 2015. Early metazoan life: divergence, environment and ecology. Philos. Trans. R. Soc. B Biol. Sci. 370:20150036.

Erwin D.H., Laflamme M., Tweedt S.M., Sperling E.A., Pisani D., Peterson K.J. 2011. The Cambrian conundrum: early divergence and later ecological success in the early history of animals. Science. 334:1091–1097.

Espeland M., Breinholt J., Willmott K.R., Warren A.D., Vila R., Toussaint E.F.A., Maunsell S.C., Aduse-Poku K., Talavera G., Eastwood R., Jarzyna M.A., Guralnick R., Lohman D.J., Pierce N.E., Kawahara A.Y. 2018. A comprehensive and dated phylogenomic analysis of butterflies. Curr. Biol. 28:770-778.e5.

Etherington G.J., Dicks J., Roberts I.N. 2005. Recombination Analysis Tool (RAT): a program for the high-throughput detection of recombination. Bioinformatics. 21:278–281.

Etienne R.S., Haegeman B., Stadler T., Aze T., Pearson P.N., Purvis A., Phillimore A.B. 2011. Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. Proc. R. Soc. B Biol. Sci. 279:1300–1309.

Ewing G.B., Ebersberger I., Schmidt H.A., von Haeseler A. 2008. Rooted triple consensus and anomalous gene trees. BMC Evol. Biol. 8:118.

Felsenstein J. 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst. Biol. 22:240–249.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368–376.

Felsenstein J. 1983. Parsimony in systematics: biological and statistical issues. Annu. Rev. Ecol. Syst. 14:313–333.

Felsenstein J. 1984. Distance methods for inferring phylogenies: a justification. Evolution. 38:16–24.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution. 39:783–791.

Felsenstein J. 2004. Inferring phylogenies. Sunderland, Massachusetts: Sinauer Associates, Inc.

Feuda R., Dohrmann M., Pett W., Philippe H., Rota-Stabelli O., Lartillot N., Wörheide G., Pisani D. 2017a. Improved modeling of compositional heterogeneity supports sponges as sister to all other animals. Curr. Biol. 27:3864-3870.e4.

Feuda R., Dohrmann M., Pett W., Philippe H., Rota-Stabelli O., Lartillot N., Wörheide G., Pisani D. 2017b. Data repository for “Improved modeling of compositional heterogeneity supports sponges as sister to all other animals”. Dataset. Bitbucket. Available from https://bitbucket.org/bzxdp/feuda_et_al_2017/src/master/.

Fitch W.M., Margoliash E. 1967. Construction of phylogenetic trees. Science. 155:279–284.

Flouri T., Jiao X., Rannala B., Yang Z. 2018. Species tree inference with BPP using genomic sequences and the multispecies coalescent. Mol. Biol. Evol. 35:2585–2593.

Folk R.A., Mandel J.R., Freudenstein J.V. 2017. Ancestral gene flow and parallel organellar genome capture result in extreme phylogenomic discord in a lineage of angiosperms. Syst. Biol. 66:320–337.

Foster P.G., de Oliveira T.M.P., Bergo E.S., Conn J.E., Sant’Ana D.C., Nagaki S.S., Nihei S., Lamas C.E., González C., Moreira C.C., Sallum M.A.M. 2017. Phylogeny of Anophelinae using mitochondrial protein coding genes. R. Soc. Open Sci. 4:170758.

Francis W.R., Canfield D.E. 2020. Very few sites can reshape the inferred phylogenetic tree. PeerJ. 8:e8865.

Frandsen P. 2015. fast-TIGER v0.0.2. Available from https://github.com/pbfrandsen/fast_TIGER (Accessed 12/01/2022).

Frandsen P.B., Calcott B., Mayer C., Lanfear R. 2015. Automatic selection of partitioning schemes for phylogenetic analyses using iterative k-means clustering of site rates. BMC Evol. Biol. 15:13.

Gascuel O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. 14:685–695.

Gatesy J. 2007. A tenth crucial question regarding model use in phylogenetics. Trends Ecol. Evol. 22:509–510.

Gatesy J., Springer M.S. 2013. Concatenation versus coalescence versus “concatalescence.” Proc. Natl. Acad. Sci. U. S. A. 110:E1179.

Gatesy J., Springer M.S. 2014. Phylogenetic analysis at deep timescales: unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. Mol. Phylogenet. Evol. 80:231–266.

Gazave E., Lapébie P., Ereskovsky A.V., Vacelet J., Renard E., Cárdenas P., Borchiellini C. 2012. No longer Demospongiae: Homoscleromorpha formal nomination as a fourth class of Porifera. In: Maldonado M., Turon X., Becerro M., Jesús Uriz M., editors. Ancient Animals, New Challenges: Developments in Sponge Research. Dordrecht: Springer Netherlands. p. 3–10.

Glazko G.V., Nei M. 2003. Estimation of divergence times for major lineages of primate species. Mol. Biol. Evol. 20:424–434.

Gogarten J.P., Townsend J.P. 2005. Horizontal gene transfer, genome innovation and evolution. Nat. Rev. Microbiol. 3:679–687.

Göker M., Grimm G.W. 2008. General functions to transform associate data to host data, and their use in phylogenetic inference from sequences with intra-individual variability. BMC Evol. Biol. 8:86.

Goldman N. 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. 36:182–198.

Gouy R., Baurain D., Philippe H. 2015. Rooting the tree of life: the phylogenetic jury is still out. Philos. Trans. R. Soc. B Biol. Sci. 370:20140329.

Gray R.D., Bryant D., Greenhill S.J. 2010. On the shape and fabric of human history. Philos. Trans. R. Soc. B Biol. Sci. 365:3923–3933.

Greenhill S.J. 2016. Phylogemetric: a Python library for calculating phylogenetic network metrics. J. Open Source Softw. 1:28.

Greenhill S.J. 2021. phylogemetric v1.0.0. Available from https://github.com/SimonGreenhill/phylogemetric (Accessed 19/01/2022).

Grimm G.W., Renner S.S. 2013. Harvesting Betulaceae sequences from GenBank to generate a new chronogram for the family. Bot. J. Linn. Soc. 172:465–477.

Grünewald S., Forslund K., Dress A., Moulton V. 2007. QNet: an agglomerative method for the construction of phylogenetic networks from weighted quartets. Mol. Biol. Evol. 24:532–538.

Gu X., Fu Y.X., Li W.H. 1995. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. Mol. Biol. Evol. 12:546–557.

Guindon S., Dufayard J.-F., Lefort V., Anisimova M., Hordijk W., Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59:307–321.

Hadfield J., Megill C., Bell S.M., Huddleston J., Potter B., Callender C., Sagulenko P., Bedford T., Neher R.A. 2018. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 34:4121–4123.

Haeckel E. 1866. Generelle morphologie der organismen. 2 vols. Berlin, Germany: Georg Reimer.

Haeckel E. 1872. Die kalkschwämme. Eine monographie. Berlin: Georg Reimer.

Halanych K.M., Whelan N.V., Kocot K.M., Kohn A.B., Moroz L.L. 2016. Miscues misplace sponges. Proc. Natl. Acad. Sci. U. S. A. 113:E946.

Hamilton N.E., Ferry M. 2018. ggtern: ternary diagrams using ggplot2. J. Stat. Softw. 87:1–17.

Hánová A., Konečný A., Nicolas V., Denys C., Granjon L., Lavrenchenko L.A., Šumbera R., Mikula O., Bryja J. 2021. Multilocus phylogeny of African striped grass mice (Lemniscomys): stripe pattern only partly reflects evolutionary relationships. Mol. Phylogenet. Evol. 155:107007.

Hasegawa M., Kishino H. 1989. Confidence limits of the maximum-likelihood estimate of the hominoid three from mitochondrial-DNA sequences. Evolution. 43:672–677.

Hasegawa M., Kishino H., Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160–174.

He D., Fiz-Palacios O., Fu C.-J., Fehling J., Tsai C.-C., Baldauf S.L. 2014. An alternative root for the Eukaryote tree of life. Curr. Biol. 24:465–470.

Heather J.M., Chain B. 2016. The sequence of sequencers: the history of sequencing DNA. Genomics. 107:1–8.

Heibl C. 2008. PHYLOCH: R language tree plotting tools and interfaces to diverse phylogenetic software packages. http://www.christophheibl.de/Rpackages.html (Accessed 05/10/2022).

Heikkilä M., Mutanen M., Wahlberg N., Sihvonen P., Kaila L. 2015. Elusive ditrysian phylogeny: an account of combining systematized morphology with molecular data (Lepidoptera). BMC Evol. Biol. 15:260.

Hejnol A., Obst M., Stamatakis A., Ott M., Rouse G.W., Edgecombe G.D., Martinez P., Baguñà J., Bailly X., Jondelius U., Wiens M., Müller W.E.G., Seaver E., Wheeler W.C., Martindale M.Q., Giribet G., Dunn C.W. 2009. Assessing the root of Bilaterian animals with scalable phylogenomic methods. Proc. R. Soc. B Biol. Sci. 276:4261–4270.

Hernández-López A., Chabrol O., Royer-Carenzi M., Merhej V., Pontarotti P., Raoult D. 2013. To tree or not to tree? Genome-wide quantification of recombination and reticulate evolution during the diversification of strict intracellular bacteria. Genome Biol. Evol. 5:2305–2317.

Hibbins M.S., Hahn M.W. 2021. The effects of introgression across thousands of quantitative traits revealed by gene expression in wild tomatoes. PLOS Genet. 17:e1009892.

Hibbins M.S., Hahn M.W. 2022. Distinguishing between histories of speciation and introgression using genomic data. bioRxiv. DOI: 10.1101/2022.09.07.506990

Ho J.W.K., Adams C.E., Lew J.B., Matthews T.J., Ng C.C., Shahabi-Sirjani A., Tan L.H., Zhao Y., Easteal S., Wilson S.R., Jermiin L.S. 2006. SeqVis: visualization of compositional heterogeneity in large alignments of nucleotides. Bioinformatics. 22:2162–2163.

Hoang D.T., Chernomor O., von Haeseler A., Minh B.Q., Vinh L.S. 2018a. UFBoot2: improving the ultrafast bootstrap approximation. Mol. Biol. Evol. 35:518–522.

Hoang D.T., Vinh L.S., Flouri T., Stamatakis A., von Haeseler A., Minh B.Q. 2018b. MPBoot: fast phylogenetic maximum parsimony tree inference and bootstrap approximation. BMC Evol. Biol. 18:11.

Hobolth A., Dutheil J.Y., Hawks J., Schierup M.H., Mailund T. 2011. Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. Genome Res. 21:349–356.

Höhna S., Hsiang A.Y. 2024. Sequential Bayesian phylogenetic inference. Syst. Biol. 73: 704–72.

Höhna S., Landis M.J., Heath T.A., Boussau B., Lartillot N., Moore B.R., Huelsenbeck J.P., Ronquist F. 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Syst. Biol. 65:726–736.

Holland B.R., Huber K.T., Dress A., Moulton V. 2002. δ plots: a tool for analyzing phylogenetic distance data. Mol. Biol. Evol. 19:2051–2059.

Holman E.W., Walker R., Rama T., Wichmann S. 2011. Correlates of reticulation in linguistic phylogenies. Lang. Dyn. Change. 1:205–240.

Holmes S. 2003. Bootstrapping phylogenetic trees: theory and methods. Stat. Sci. 18:241–255.

Huang H., He Q., Kubatko L.S., Knowles L.L. 2010. Sources of error inherent in species-tree estimation: impact of mutational and coalescent effects on accuracy and implications for choosing among different methods. Syst. Biol. 59:573–583.

Hudson R.R. 2002. Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics. 18:337–338.

Huelsenbeck J.P., Rannala B., Masly J.P. 2000. Accommodating phylogenetic uncertainty in evolutionary studies. Science. 288:2349–2350.

Huelsenbeck J.P., Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 17:754–755.

Hughes L.C., Ortí G., Huang Y., Sun Y., Baldwin C.C., Thompson A.W., Arcila D., Betancur-R. R., Li C., Becker L., Bellora N., Zhao X., Li X., Wang M., Fang C., Xie B., Zhou Z., Huang H., Chen S., Venkatesh B., Shi Q. 2018. Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. Proc. Natl. Acad. Sci. U. S. A. 115:6249–6254.

Hurvich C.M., Tsai C.-L. 1989. Regression and time series model selection in small samples. Biometrika. 76:297–307.

Huson D.H. 1998. SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics. 14:68–73.

Huson D.H., Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23:254-267. Software available from www.splitstree.org.

Huson D.H., Bryant D. 2022. SplitsTree4 v4.18.2. Available from https://software-ab.cs.uni-tuebingen.de/download/splitstree4/welcome.html (Accessed 08/04/2022).

Illergård K., Ardell D.H., Elofsson A. 2009. Structure is three to ten times more conserved than sequence — a study of structural response in protein cores. Proteins Struct. Funct. Bioinforma. 77:499–508.

James T.Y., Kauff F., Schoch C.L., Matheny P.B., Hofstetter V., Cox C.J., Celio G., Gueidan C., Fraker E., Miadlikowska J., Lumbsch H.T., Rauhut A., Reeb V., Arnold A.E., Amtoft A., Stajich J.E., Hosaka K., Sung G.-H., Johnson D., O'Rourke B., Crockett M., Binder M., Curtis J.M., Slot J.C., Wang Z., Wilson A.W., Schüßler A., Longcore J.E., O'Donnell K., Mozley-Standridge S., Porter D., Letcher P.M., Powell M.J., Taylor J.W., White M.M., Griffith G.W., Davies D.R., Humber R.A., Morton J.B., Sugiyama J., Rossman A.Y., Rogers J.D., Pfister D.H., Hewitt D., Hansen K., Hambleton S., Shoemaker R.A., Kohlmeyer J., Volkmann-Kohlmeyer B., Spotts R.A., Serdani M., Crous P.W., Hughes K.W., Matsuura K., Langer E., Langer G., Untereiner W.A., Lücking R., Büdel B., Geiser D.M., Aptroot A., Diederich P., Schmitt I., Schultz M., Yahr R., Hibbett D.S., Lutzoni F., McLaughlin D.J., Spatafora J.W., Vilgalys R. 2006. Reconstructing the early evolution of Fungi using a six-gene phylogeny. Nature. 443:818–822.

Jeffroy O., Brinkmann H., Delsuc F., Philippe H. 2006. Phylogenomics: the beginning of incongruence? Trends Genet. 22:225–231.

Jékely G., Paps J., Nielsen C. 2015. The phylogenetic position of Ctenophores and the origin(s) of nervous systems. EvoDevo. 6:1.

Jermiin L.S., Catullo R.A., Holland B.R. 2020. A new phylogenetic protocol: dealing with model misspecification and confirmation bias in molecular phylogenetics. NAR Genomics Bioinforma. 2:lqaa041.

Joly, S., McLenachan, P.A., Lockhart, P.J. 2009. A statistical approach for distinguishing hybridization and incomplete lineage sorting. Am. Nat. 174:E54-E70.

Jombart T., Kendall M., Almagro-Garcia J., Colijn C. 2017. treespace: statistical exploration of landscapes of phylogenetic trees. Mol. Ecol. Resour. 17:1385–1392.

Jukes T.H., Cantor C.R. 1969. Evolution of protein molecules. In: Munro H.N., editor. Mammalian Protein Metabolism. New York: Academic Press. p. 21–132.

Jumper J., Evans R., Pritzel A., Green T., Figurnov M., Ronneberger O., Tunyasuvunakool K., Bates R., Žídek A., Potapenko A., Bridgland A., Meyer C., Kohl S.A.A., Ballard A.J., Cowie A., Romera-Paredes B., Nikolov S., Jain R., Adler J., Back T., Petersen S., Reiman D., Clancy E., Zielinski M., Steinegger M., Pacholska M., Berghammer T., Bodenstein S., Silver D., Vinyals O., Senior A.W., Kavukcuoglu K., Kohli P., Hassabis D. 2021. Highly accurate protein structure prediction with AlphaFold. Nature. 596:583–589.

Juravel K., Porras L., Höhna S., Pisani D., Wörheide G. 2023. Exploring genome gene content and morphological analysis to test recalcitrant nodes in the animal phylogeny. PLOS ONE. 18:e0282444.

Kalyaanamoorthy S., Minh B.Q., Wong T.K.F., von Haeseler A., Jermiin L.S. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 14:587–589.

Kapli P., Flouri T., Telford M.J. 2021. Systematic errors in phylogenetic trees. Curr. Biol. 31:R59–R64.

Kapli P., Telford M.J. 2020. Topology-dependent asymmetry in systematic errors affects phylogenetic placement of Ctenophora and Xenacoelomorpha. Sci. Adv. 6:eabc5162.

Kapli P., Yang Z., Telford M.J. 2020. Phylogenetic tree building in the genomic age. Nat. Rev. Genet. 21:428–444.

Karimi N., Grover C.E., Gallagher J.P., Wendel J.F., Ané C., Baum D.A. 2019. Reticulate evolution helps explain apparent homoplasy in floral biology and pollination in baobabs (Adansonia; Bombacoideae; Malvaceae). Syst. Biol. 69:462–478.

Kassambara A. 2023. ggpubr: “ggplot2” based publication ready plots. R package version 0.6.0.999. Available at https://rpkgs.datanovia.com/ggpubr/.

Kawahara A.Y., Storer C., Carvalho A.P.S., Plotkin D.M., Condamine F.L., Braga M.P., Ellis E.A., St Laurent R.A., Li X., Barve V., Cai L., Earl C., Frandsen P.B., Owens H.L., Valencia-Montoya W.A., Aduse-Poku K., Toussaint E.F.A., Dexter K.M., Doleck T., Markee A., Messcher R., Nguyen Y.-L., Badon J.A.T., Benítez H.A., Braby M.F., Buenavente P.A.C., Chan W.-P., Collins S.C., Rabideau Childers R.A., Dankowicz E., Eastwood R., Fric Z.F.,

Gott R.J., Hall J.P.W., Hallwachs W., Hardy N.B., Sipe R.L.H., Heath A., Hinolan J.D., Homziak N.T., Hsu Y.-F., Inayoshi Y., Itliong M.G.A., Janzen D.H., Kitching I.J., Kunte K., Lamas G., Landis M.J., Larsen E.A., Larsen T.B., Leong J.V., Lukhtanov V., Maier C.A., Martinez J.I., Martins D.J., Maruyama K., Maunsell S.C., Mega N.O., Monastyrskii A., Morais A.B.B., Müller C.J., Naive M.A.K., Nielsen G., Padrón P.S., Peggie D., Romanowski H.P., Sáfián S., Saito M., Schröder S., Shirey V., Soltis D., Soltis P., Sourakov A., Talavera G., Vila R., Vlasanek P., Wang H., Warren A.D., Willmott K.R., Yago M., Jetz W., Jarzyna M.A., Breinholt J.W., Espeland M., Ries L., Guralnick R.P., Pierce N.E., Lohman D.J. 2023. A global phylogeny of butterflies reveals their evolutionary history, ancestral hosts and biogeographic origins. Nat. Ecol. Evol. 7:903–913.

Kayal E., Roure B., Philippe H., Collins A.G., Lavrov D.V. 2013. Cnidarian phylogenetic relationships as revealed by mitogenomics. BMC Evol. Biol. 13:5.

Kelchner S.A., Thomas M.A. 2007. Model use in phylogenetics: nine key questions. Trends Ecol. Evol. 22:87–94.

van Kempen M., Kim S.S., Tumescheit C., Mirdita M., Lee J., Gilchrist C.L.M., Söding J., Steinegger M. 2024. Fast and accurate protein structure search with Foldseek. Nat. Biotechnol. 42:243–246.

Kim J.I., Shin W., Triemer R.E. 2013. Phylogenetic reappraisal of the genus Monomorphina (Euglenophyceae) based on molecular and morphological data. J. Phycol. 49:82–91.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111–120.

King N., Rokas A. 2017. Embracing uncertainty in reconstructing early animal evolution. Curr. Biol. 27:R1081–R1088.

Kishino H., Hasegawa M. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J. Mol. Evol. 29:170–179.

Klimov P.B., OConnor B.M., Chetverikov P.E., Bolton S.J., Pepato A.R., Mortazavi A.L., Tolstikov A.V., Bauchan G.R., Ochoa R. 2018. Comprehensive phylogeny of acariform mites (Acariformes) provides insights on the origin of the four-legged mites (Eriophyoidea), a long branch. Mol. Phylogenet. Evol. 119:105–117.

Knowles L.L., Huang H., Sukumaran J., Smith S.A. 2018. A matter of phylogenetic scale: distinguishing incomplete lineage sorting from lateral gene transfer as the cause of gene tree discord in recent versus deep diversification histories. Am. J. Bot. 105:376–384.

Kohn A.B., Citarella M.R., Kocot K.M., Bobkova Y.V., Halanych K.M., Moroz L.L. 2012. Rapid evolution of the compact and unusual mitochondrial genome in the Ctenophore, Pleurobrachia bachei. Mol. Phylogenet. Evol. 63:203–207.

Kolaczkowski B., Thornton J.W. 2009. Long-branch attraction bias and inconsistency in Bayesian phylogenetics. PLOS ONE. 4:e7891.

Kosakovsky Pond S.L., Posada D., Gravenor M.B., Woelk C.H., Frost S.D.W. 2006. GARD: a genetic algorithm for recombination detection. Bioinformatics. 22:3096–3098.

Kosiol C., Goldman N. 2005. Different versions of the Dayhoff rate matrix. Mol. Biol. Evol. 22:193–199.

Kozak K.M., Wahlberg N., Neild A.F.E., Dasmahapatra K.K., Mallet J., Jiggins C.D. 2015. Multilocus species trees show the recent adaptive radiation of the mimetic Heliconius butterflies. Syst. Biol. 64:505–524.

Kozlov A.M., Darriba D., Flouri T., Morel B., Stamatakis A. 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics. 35:4453–4455.

Kruskal J.B. 1956. On the shortest spanning subtree of a graph and the traveling salesman problem. Proc. Am. Math. Soc. 7:48–50.

Kubatko L.S., Degnan J.H. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol. 56:17–24.

Kück P., Mayer C., Wägele J.-W., Misof B. 2012. Long branch effects distort maximum likelihood phylogenies in simulations despite selection of the correct model. PLOS ONE. 7:e36593.

Kumar S., Gadagkar S.R. 2001. Disparity index: a simple statistic to measure and test the homogeneity of substitution patterns between molecular sequences. Genetics. 158:1321–1327.

Kuritzin A., Kischka T., Schmitz J., Churakov G. 2016. Incomplete lineage sorting and hybridization statistics for large-scale retroposon insertion data. PLOS Comput. Biol. 12:e1004812.

Labate J.A., Robertson L.D., Strickler S.R., Mueller L.A. 2014. Genetic structure of the four wild tomato species in the Solanum peruvianum s.l. species complex. Genome. 57:169–180.

Lai J.-S., Rost B., Kobe B., Bodén M. 2020. Evolutionary model of protein secondary structure capable of revealing new biological relationships. Proteins Struct. Funct. Bioinforma. 88:1251–1259.

Lake J.A., Henderson E., Oakes M., Clark M.W. 1984. Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. Proc. Natl. Acad. Sci. U. S. A. 81:3786–3790.

Lam H.M., Ratmann O., Boni M.F. 2018. Improved algorithmic complexity for the 3SEQ recombination detection algorithm. Mol. Biol. Evol. 35:247–251.

Lanfear R., Calcott B., Ho S.Y.W., Guindon S. 2012. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol. Biol. Evol. 29:1695–1701.

Lanfear R., Frandsen P.B., Wright A.M., Senfeld T., Calcott B. 2017. PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. Mol. Biol. Evol. 34:772–773.

Lanfear R., Hahn M.W. 2024. The meaning and measure of concordance factors in phylogenomics. Mol. Biol. Evol. 41:msae214.

Langergraber K.E., Prüfer K., Rowney C., Boesch C., Crockford C., Fawcett K., Inoue E., Inoue-Muruyama M., Mitani J.C., Muller M.N., Robbins M.M., Schubert G., Stoinski T.S., Viola B., Watts D., Wittig R.M., Wrangham R.W., Zuberbühler K., Pääbo S., Vigilant L. 2012. Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. Proc. Natl. Acad. Sci. U. S. A. 109:15716–15721.

Lanier H.C., Knowles L.L. 2012. Is recombination a problem for species-tree analyses? Syst. Biol. 61:691–701.

Lapointe F.J., Kirsch J. 1995. Estimating phylogenies from lacunose distance matrices, with special reference to DNA hybridization data. Mol. Biol. Evol. 12:266.

Larget B., Simon D. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol. Biol. Evol. 16:750.

Lartillot N. 2020a. PhyloBayes: Bayesian phylogenetics using site-heterogeneous models. In: Scornavacca C., Delsuc F., Galtier N., editors. Phylogenetics in the Genomic Era. No commercial publisher | Authors open access book. p. 1.5:1-1.5:16.

Lartillot N. 2020b. The Bayesian approach to molecular phylogeny. In: Scornavacca C., Delsuc F., Galtier N., editors. Phylogenetics in the Genomic Era. No commercial publisher | Authors open access book. p. 1.4:1-1.4:17.

Lartillot N., Brinkmann H., Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol. Biol. 7:S4.

Lartillot N., Lepage T., Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics. 25:2286–2288.

Lartillot N., Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21:1095–1109.

Lartillot N., Philippe H. 2006. Computing Bayes factors using thermodynamic integration. Syst. Biol. 55:195–207.

Laumer C.E., Fernández R., Lemer S., Combosch D., Kocot K.M., Riesgo A., Andrade S.C.S., Sterrer W., Sørensen M.V., Giribet G. 2019a. Revisiting metazoan phylogeny with genomic sampling of all phyla. Proc. R. Soc. B Biol. Sci. 286:20190831.

Laumer C.E., Fernández R., Lemer S., Combosch D., Kocot K.M., Riesgo A., Andrade S.C.S., Sterrer W., Sørensen M.V., Giribet G. 2019b. Supplementary material from "Revisiting metazoan phylogeny with genomic sampling of all phyla". The Royal Society. Collection. https://doi.org/10.6084/m9.figshare.c.4552313.v3.

Laumer C.E., Fernández R., Lemer S., Combosch D., Kocot K.M., Riesgo A., Andrade S.C.S., Sterrer W., Sørensen M.V., Giribet G. 2019c. Data from: Revisiting metazoan phylogeny with genomic sampling of all phyla. Dataset. Dryad. Available from https://doi.org/10.5061/dryad.293kp3d.

Laumer C.E., Gruber-Vodicka H., Hadfield M.G., Pearse V.B., Riesgo A., Marioni J.C., Giribet G. 2018a. Support for a clade of Placozoa and Cnidaria in genes with minimal compositional bias. eLife. 7:e36278.

Laumer C.E., Gruber-Vodicka H., Hadfield M.G., Pearse V.B., Riesgo A., Marioni J.C., Giribet G. 2018b. Data from: Support for a clade of Placozoa and Cnidaria in genes with minimal compositional bias. Dataset. Dryad. Available from https://doi.org/10.5061/dryad.6cm1166.

Lavretsky P., McCracken K.G., Peters J.L. 2014. Phylogenetics of a recent radiation in the mallards and allies (Aves: Anas): inferences from a genomic transect and the multispecies coalescent. Mol. Phylogenet. Evol. 70:402–411.

Le Quesne W.J. 1969. A method of selection of characters in numerical taxonomy. Syst. Biol. 18:201–205.

Le S.Q., Dang C.C., Gascuel O. 2012. Modeling protein evolution with several amino acid replacement matrices depending on site rates. Mol. Biol. Evol. 29:2921–2936.

Le S.Q., Gascuel O. 2010. Accounting for solvent accessibility and secondary structure in protein phylogenetics is clearly beneficial. Syst. Biol. 59:277–287.

Le S.Q., Gascuel O., Lartillot N. 2008a. Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics. 24:2317–2323.

Le S.Q., Lartillot N., Gascuel O. 2008b. Phylogenetic mixture models for proteins. Philos. Trans. R. Soc. B Biol. Sci. 363:3965–3976.

Le V.S., Dang C.C., Le Q.S. 2017. Improved mitochondrial amino acid substitution models for metazoan evolutionary studies. BMC Evol. Biol. 17:136.

Lee S., Hasegawa T. 2013. Evolution of the Ainu language in space and time. PLOS ONE. 8:e62243.

Lee J.Y., Joseph, L., Edwards, S.V. 2012. A Species Tree for the Australo-Papuan Fairy-wrens and Allies (Aves: Maluridae). Sys. Biol. 61:253.

Leebens-Mack J.H., Barker M.S., Carpenter E.J., Deyholos M.K., Gitzendanner M.A., Graham S.W., Grosse I., Li Z., Melkonian M., Mirarab S., Porsch M., Quint M., Rensing S.A., Soltis D.E., Soltis P.S., Stevenson D.W., Ullrich K.K., Wickett N.J., DeGironimo L., Edger P.P., Jordon-Thaden I.E., Joya S., Liu T., Melkonian B., Miles N.W., Pokorny L., Quigley C., Thomas P., Villarreal J.C., Augustin M.M., Barrett M.D., Baucom R.S., Beerling D.J., Benstein R.M., Biffin E., Brockington S.F., Burge D.O., Burris J.N., Burris K.P., Burtet-Sarramegna V., Caicedo A.L., Cannon S.B., Çebi Z., Chang Y., Chater C., Cheeseman J.M., Chen T., Clarke N.D., Clayton H., Covshoff S., Crandall-Stotler B.J., Cross H., dePamphilis C.W., Der J.P., Determann R., Dickson R.C., Di Stilio V.S., Ellis S., Fast E., Feja N., Field K.J., Filatov D.A., Finnegan P.M., Floyd S.K., Fogliani B., García N., Gâteblé G., Godden G.T., Goh F., Greiner S., Harkess A., Heaney J.M., Helliwell K.E., Heyduk K., Hibberd J.M., Hodel R.G.J., Hollingsworth P.M., Johnson M.T.J., Jost R., Joyce B., Kapralov M.V., Kazamia E., Kellogg E.A., Koch M.A., Von Konrat M., Könyves K., Kutchan T.M., Lam V., Larsson A., Leitch A.R., Lentz R., Li F.-W., Lowe A.J., Ludwig M., Manos P.S., Mavrodiev E., McCormick M.K., McKain M., McLellan T., McNeal J.R., Miller R.E., Nelson M.N., Peng Y., Ralph P., Real D., Riggins C.W., Ruhsam M., Sage R.F., Sakai A.K., Scascitella M., Schilling E.E., Schlösser E.-M., Sederoff H., Servick S., Sessa E.B., Shaw A.J., Shaw S.W., Sigel E.M., Skema C., Smith A.G., Smithson A., Stewart C.N., Stinchcombe J.R., Szövényi P., Tate J.A., Tiebel H., Trapnell D., Villegente M., Wang C.-N., Weller S.G., Wenzel M., Weststrand S., Westwood J.H., Whigham D.F., Wu S., Wulff A.S., Yang Y., Zhu D., Zhuang C., Zuidof J., Chase M.W., Pires J.C., Rothfels C.J., Yu J., Chen C., Chen L., Cheng S., Li J., Li R., Li X., Lu H., Ou Y., Sun X., Tan X., Tang J., Tian Z., Wang F., Wang J., Wei X., Xu X., Yan Z., Yang F., Zhong X., Zhou F., Zhu Y., Zhang Y., Ayyampalayam S., Barkman T.J., Nguyen N., Matasci N., Nelson D.R., Sayyari E., Wafula E.K., Walls R.L., Warnow T., An H., Arrigo N., Baniaga A.E., Galuska S., Jorgensen S.A., Kidder T.I., Kong H., Lu-Irving P., Marx H.E., Qi X., Reardon C.R., Sutherland B.L., Tiley G.P., Welles S.R., Yu R., Zhan S., Gramzow L., Theißen G., Wong G.K.-S., One Thousand Plant Transcriptomes Initiative. 2019a. One thousand plant transcriptomes and the phylogenomics of green plants. Nature. 574:679–685.

Leebens-Mack J.H., Wong G.K.-S., One Thousand Plant Transcriptomes Initiative. 2019b. Data packages for One Thousand Plant transcriptomes and phylogenomics of green plants. Dataset. Data Commons. Available from https://doi.org/10.25739/8m7t-4e85.

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. Syst. Biol. 58:130–145.

Lemmon A.R., Moriarty E.C. 2004. The importance of proper model assumption in Bayesian phylogenetics. Syst. Biol. 53:265–277.

Li G., Figueiró H.V., Eizirik E., Murphy W.J. 2019. Recombination-aware phylogenomics reveals the structured genomic landscape of hybridizing cat species. Mol. Biol. Evol. 36:2111–2126.

Li T., Liu D., Yang Y., Guo J., Feng Y., Zhang X., Cheng S., Feng J. 2020a. Phylogenetic supertree reveals detailed evolution of SARS-CoV-2. Sci. Rep. 10:22366.

Li Y., Shen X.-X., Evans B., Dunn C.W., Rokas A. 2020b. Data repository for “Rooting the animal tree of life”. Dataset. Figshare. Available from https://doi.org/10.6084/m9.figshare.13122881.v1.

Li Y., Shen X.-X., Evans B., Dunn C.W., Rokas A. 2021. Rooting the animal tree of life. Mol. Biol. Evol. 38:4322–4333.

List J.-M. 2022. Correcting a bias in TIGER rates resulting from high amounts of invariant and singleton cognate sets. J. Lang. Evol. 7:53–58.

Liu K., Linder C.R., Warnow T. 2011. Multiple sequence alignment: a major challenge to large-scale phylogenetics. PLOS Curr. 2:RRN1198–RRN1198.

Liu L., Xi Z., Davis C.C. 2015a. Coalescent methods are robust to the simultaneous effects of long branches and incomplete lineage sorting. Mol. Biol. Evol. 32:791–805.

Liu L., Xi Z., Wu S., Davis C.C., Edwards S.V. 2015b. Estimating phylogenetic trees from genome-scale data. Ann. N. Y. Acad. Sci. 1360:36–53.

Louca S. 2023. castor: efficient phylogenetics on large trees. v1.7.8. Available from https://CRAN.R-project.org/package=castor.

Lughadha E.N., Govaerts R., Belyaeva I., Black N., Lindon H., Allkin R., Magill R.E., Nicolson N. 2016. Counting counts: revised estimates of numbers of accepted species of flowering plants, seed plants, vascular plants and land plants with a review of other recent estimates. Phytotaxa. 272:82–88.

Lundin D., Poole A.M., Sjöberg B.-M., Högbom M. 2012. Use of structural phylogenetic networks for classification of the Ferritin-like superfamily. J. Biol. Chem. 287:20565–20575.

Lutteropp S., Scornavacca C., Kozlov A.M., Morel B., Stamatakis A. 2022. NetRAX: accurate and fast maximum likelihood phylogenetic network inference. Bioinformatics. 38:3725–3733.

Ly-Trong N., Naser-Khdour S., Lanfear R., Minh B.Q. 2022. AliSim: a fast and versatile phylogenetic sequence simulator for the genomic era. Mol. Biol. Evol. 39:msac092.

Maddison W.P., Knowles L.L. 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55:21–30.

Malik A.J., Langer D., Verma C.S., Poole A.M., Allison J.R. 2023. Structome: a tool for the rapid assembly of datasets for structural phylogenetics. Bioinforma. Adv. 3:vbad134.

Malik A.J., Poole A.M., Allison J.R. 2020. Structural phylogenetics with confidence. Mol. Biol. Evol. 37:2711–2726.

Mallet J., Besansky N., Hahn M.W. 2016. How reticulated are species? BioEssays. 38:140–149.

Mao X., Tsagkogeorga G., Bailey S.E., Rossiter S.J. 2017. Genomics of introgression in the Chinese horseshoe bat (Rhinolophus sinicus) revealed by transcriptome sequencing. Biol. J. Linn. Soc. 121:698–710.

Martin D.P., Varsani A., Roumagnac P., Botha G., Maslamoney S., Schwab T., Kelz Z., Kumar V., Murrell B. 2020. RDP5: a computer program for analyzing recombination in, and removing signals of recombination from, nucleotide sequence datasets. Virus Evol. 7:veaa087–veaa087.

Martindale M.Q., Finnerty J.R., Henry J.Q. 2002. The Radiata and the evolutionary origins of the Bilaterian body plan. Mol. Phylogenet. Evol. 24:358–365.

Maynard Smith J. 1992. Analyzing the mosaic structure of genes. J. Mol. Evol. 34:126–129.

McCarthy C. 2022. chmccarthy/ATOLRootStudy. Dataset. https://github.com/chmccarthy/ATOLRootStudy (Accessed 17/08/20223).

McCarthy C.G.P., Mulhair P.O., Siu-Ting K., Creevey C.J., O'Connell M.J. 2023. Improving orthologous signal and model fit in datasets addressing the root of the animal phylogeny. Mol. Biol. Evol. 40:msac276.

McCormack J.E., Huang H., Knowles L.L. 2009. Maximum likelihood estimates of species trees: how accuracy of phylogenetic inference depends upon the divergence history and sampling design. Syst. Biol. 58:501–508.

McGuire J.A., Linkem C.W., Koo M.S., Hutchison D.W., Lappin A.K., Orange D.I., Lemos-Espinal J., Riddle B.R., Jaeger J.R. 2007. Mitochondrial introgression and incomplete lineage sorting through space and time: phylogenetics of Crotaphytid lizards. Evolution. 61:2879–2897.

Medina M., Collins A.G., Silberman J.D., Sogin M.L. 2001. Evaluating hypotheses of basal animal phylogeny using complete sequences of large and small subunit rRNA. Proc. Natl. Acad. Sci. U. S. A. 98:9707–9712.

Mendes F. K., Livera A. P., Hahn M. W. 2019. The perils of intralocus recombination for inferences of molecular convergence. Philos. Trans. R. Soc. B Biol. Sci. 374:20180244.

Mendes F.K., Hahn M.W. 2016. Gene tree discordance causes apparent substitution rate variation. Syst. Biol. 65:711–721.

Mendes F.K., Hahn M.W. 2018. Why concatenation fails near the anomaly zone. Syst. Biol. 67:158–169.

Minh B.Q., Dang C.C., Vinh L.S., Lanfear R. 2021. QMaker: fast and accurate method to estimate empirical models of protein evolution. Syst. Biol. 70: 1046–1060.

Minh B.Q., Hahn M.W., Lanfear R. 2020a. New methods to calculate concordance factors for phylogenomic datasets. Mol. Biol. Evol. 37:727–2733.

Minh B.Q., Nguyen M.A.T., von Haeseler A. 2013. Ultrafast approximation for phylogenetic bootstrap. Mol. Biol. Evol. 30:1188–1195.

Minh B.Q., Schmidt H.A., Chernomor O., Schrempf D., Woodhams M.D., von Haeseler A., Lanfear R. 2020b. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. 37:1530–1534.

Mirarab S. 2023. smirarab/ASTRAL. v5.7.8. Software. Available from https://github.com/smirarab/ASTRAL.

Mirarab S., Reaz R., Bayzid M.S., Zimmermann T., Swenson M.S., Warnow T. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics. 30:i541–i548.

Misof B., Meusemann K., von Reumont B.M., Kück P., Prohaska S.J., Stadler P.F. 2014. A priori assessment of data quality in molecular phylogenetics. Algorithms Mol. Biol. 9:22.

Mitchell K.J., Wood J.R., Scofield R.P., Llamas B., Cooper A. 2014. Ancient mitochondrial genome reveals unsuspected taxonomic affinity of the extinct Chatham duck (Pachyanas chathamica) and resolves divergence times for New Zealand and sub-Antarctic brown teals. Mol. Phylogenet. Evol. 70:420–428.

Miya M., Satoh T.P., Nishida M. 2005. The phylogenetic position of toadfishes (order Batrachoidiformes) in the higher ray-finned fish as inferred from partitioned Bayesian analysis of 102 whole mitochondrial genome sequences. Biol. J. Linn. Soc. 85:289–306.

Mo Y.K., Lanfear R., Hahn M.W., Minh B.Q. 2023. Updated site concordance factors minimize effects of homoplasy and taxon sampling. Bioinformatics. 39:btac741.

Moi D., Bernard C., Steinegger M., Nevers Y., Langleib M., Dessimoz C. 2023. Structural phylogenetics unravels the evolutionary diversification of communication systems in gram-positive bacteria and their viruses. bioRxiv. DOI: 10.1101/2023.09.19.558401.

Molloy E.K., Warnow T. 2018. To include or not to include: the impact of gene filtering on species tree estimation methods. Syst. Biol. 67:285–303.

Morales-Briones D.F., Liston A., Tank D.C. 2018. Phylogenomic analyses reveal a deep history of hybridization and polyploidy in the Neotropical genus Lachemilla (Rosaceae). New Phytol. 218:1668–1684.

Morales-Briones, D.F., Kadereit, G., Tefarikis, D.T., Moore, M.J., Smith, S.A., Brockington, S.F., Timoneda, A., Yim, W.C., Cushman, J.C., Yang, Y. 2021. Disentangling sources of gene tree discordance in phylogenomic data sets: testing ancient hybridizations in Amaranthaceae s.l.. Sys. Biol. 70:219-235.

Moroz L.L., Kocot K.M., Citarella M.R., Dosung S., Norekian T.P., Povolotskaya I.S., Grigorenko A.P., Dailey C., Berezikov E., Buckley K.M., Ptitsyn A., Reshetov D., Mukherjee K., Moroz T.P., Bobkova Y., Yu F., Kapitonov V.V., Jurka J., Bobkov Y.V., Swore J.J., Girardo D.O., Fodor A., Gusev F., Sanford R., Bruders R., Kittler E., Mills C.E., Rast J.P., Derelle R., Solovyev V.V., Kondrashov F.A., Swalla B.J., Sweedler J.V., Rogaev E.I., Halanych K.M., Kohn A.B. 2014. The Ctenophore genome and the evolutionary origins of neural systems. Nature. 510:109–114.

Morrison D.A. 2006. Multiple sequence alignment for phylogenetic purposes. Aust. Syst. Bot. 19:479–539.

Mossel E., Steel M. 2005. How much can evolved characters tell us about the tree that generated them? In: Gascuel O., editor. Mathematics of Evolution and Phylogeny. United Kingdom: Oxford University Press, Incorporated.

Murphy B., Forest F., Barraclough T., Rosindell J., Bellot S., Cowan R., Golos M., Jebb M., Cheek M. 2020. A phylogenomic analysis of Nepenthes (Nepenthaceae). Mol. Phylogenet. Evol. 144:106668.

Nadan S., Walter J.E., Grabow W.O.K., Mitchell D.K., Taylor M.B. 2003. Molecular characterization of astroviruses by reverse transcriptase PCR and sequence analysis: comparison of clinical and environmental isolates from South Africa. Appl. Environ. Microbiol. 69:747.

Naser-Khdour S., Minh B.Q., Zhang W., Stone E.A., Lanfear R. 2019. The prevalence and impact of model violations in phylogenetic analysis. Genome Biol. Evol. 11:3341–3352.

Naser-Khdour S., Quang Minh B., Lanfear R. 2021. Assessing confidence in root placement on phylogenies: an empirical study using nonreversible models for mammals. Syst. Biol. 71:959–972.

Natsidis P., Kapli P., Schiffer P.H., Telford M.J. 2021. Systematic errors in orthology inference and their effects on evolutionary analyses. iScience. 24:102110.

Nei M., Kumar S. 2001. Molecular evolution and phylogenetics. Oxford: Oxford University Press.

Neumann J.S., Desalle R., Narechania A., Schierwater B., Tessler M. 2021. Morphological characters can strongly influence early animal relationships inferred from phylogenomic data sets. Syst. Biol. 70:360–375.

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. Mol. Biol. Evol. 32:268–274.

Nickle D.C., Heath L., Jensen M.A., Gilbert P.B., Mullins J.I., Pond S.L.K. 2007. HIV-specific probabilistic models of protein evolution. PLOS ONE. 2:e503.

Nikolaev S., Montoya-Burgos J.I., Margulies E.H., Program N.C.S., Rougemont J., Nyffeler B., Antonarakis S.E. 2007. Early history of mammals Is elucidated with the ENCODE multiple species sequencing data. PLOS Genet. 3:e2.

Nosenko T., Schreiber F., Adamska M., Adamski M., Eitel M., Hammel J., Maldonado M., Müller W.E.G., Nickel M., Schierwater B., Vacelet J., Wiens M., Wörheide G. 2013a. Deep metazoan phylogeny: when different genes tell different stories. Mol. Phylogenet. Evol. 67:223–233.

Nosenko T., Schreiber F., Adamska M., Adamski M., Eitel M., Hammel J., Maldonado M., Müller W.E.G., Nickel M., Schierwater B., Vacelet J., Wiens M., Wörheide G. 2013b. Additional data to: Deep metazoan phylogeny: When different genes tell different stories. Dataset. Open Data LMU. Available from https://doi.org/10.5282/ubm/data.55.

Nute M., Chou J., Molloy E.K., Warnow T. 2018. The performance of coalescent-based species tree estimation methods under models of missing data. BMC Genomics. 19:286.

Ogilvie H.A., Bouckaert R.R., Drummond A.J. 2017. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. Mol. Biol. Evol. 34:2101–2114.

de Oliveira Martins, L., Mallo, D. and Posada, D. 2015. Phylogenetic likelihood. In: Encyclopedia of Life Sciences. Hoboken, New Jersey: John Wiley & Sons, Ltd.

Ou J. 2021. colorBlindness: safe color set for color blindness. R package v0.1.9. Available at https://CRAN.R-project.org/package=colorBlindness.

Oude Munnink B.B., Worp N., Nieuwenhuijse D.F., Sikkema R.S., Haagmans B., Fouchier R.A.M., Koopmans M. 2021. The next phase of SARS-CoV-2 surveillance: real-time molecular epidemiology. Nat. Med. 27:1518–1524.

Pagel M., Meade A. 2005. Mixture models in phylogenetic inference. In: Gascuel O., editor. Mathematics of evolution and phylogeny. Oxford, UK: University Press Oxford. p. 121–142.

Pandey A., Braun E.L. 2020. Phylogenetic analyses of sites in different protein structural environments result in distinct placements of the metazoan root. Biology. 9:64.

Pandey A., Braun E.L. 2021. The roles of protein structure, taxon sampling, and model complexity in phylogenomics: a case study focused on early animal divergences. Biophysica. 1:87–105.

Paradis E., Claude J., Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics. 20:289–290.

Paradis E., Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 35:526–528.

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. 192:1065.

Pease J.B. 2021. peaselab/mvftools. Software. Available from https://github.com/peaselab/mvftools.

Pease J.B., Haak D.C., Hahn M.W., Moyle L.C. 2016a. Phylogenomics reveals three sources of adaptive variation during a rapid radiation. PLOS Biol. 14:e1002379.

Pease J.B., Haak D.C., Hahn M.W., Moyle L.C. 2016b. Data from: Phylogenomics reveals three sources of adaptive variation during a rapid radiation. Dataset. Dryad. Available from https://doi.org/10.5061/dryad.182dv.

Pease J.B., Hahn M.W. 2015. Detection and polarization of introgression in a five-taxon phylogeny. Syst. Biol. 64:651–662.

Pease J.B., Rosenzweig B.K. 2018. Encoding data using biological principles: the Multisample Variant Format for phylogenomics and population genomics. IEEE/ACM Trans. Comput. Biol. Bioinform. 15:1231–1238.

Pedersen T.L. 2022. patchwork: the composer of plots. Package version 1.1.3. Available at https://CRAN.R-project.org/package=patchwork.

Pei J., Kim B.-H., Grishin N.V. 2008. PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res. 36:2295–2300.

Penny D. 1982. Towards a basis for classification: the incompleteness of distance measures, incompatibility analysis and phenetic classification. J. Theor. Biol. 96:129–142.

Peralta I.E., Knapp S., Spooner D.M. 2005. New species of wild tomatoes (Solanum section Lycopersicon: Solanaceae) from northern Peru. Syst. Bot. 30:424–434.

Peterson K.J., Lyons J.B., Nowak K.S., Takacs C.M., Wargo M.J., McPeek M.A. 2004. Estimating metazoan divergence times with a molecular clock. Proc. Natl. Acad. Sci. U. S. A. 101:6536–6541.

Pett W., Adamski M., Adamska M., Francis W.R., Eitel M., Pisani D., Wörheide G. 2019. The role of homology and orthology in the phylogenomic analysis of metazoan gene content. Mol. Biol. Evol. 36:643–649.

Pfeifer B., Kapan D.D. 2019. Estimates of introgression as a function of pairwise distances. BMC Bioinformatics. 20:207.

Philippe H., Brinkmann H., Copley R.R., Moroz L.L., Nakano H., Poustka A.J., Wallberg A., Peterson K.J., Telford M.J. 2011a. Acoelomorph flatworms are deuterostomes related to Xenoturbella. Nature. 470:255–258.

Philippe H., Brinkmann H., Lavrov D.V., Littlewood D.T.J., Manuel M., Wörheide G., Baurain D. 2011b. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Niology. 9:e1000602–e1000602.

Philippe H., Delsuc F., Brinkmann H., Lartillot N. 2005. Phylogenomics. Annu. Rev. Ecol. Evol. Syst. 36:541–562.

Philippe H., Derelle R., Lopez P., Pick K., Borchiellini C., Boury-Esnault N., Vacelet J., Renard E., Houliston E., Quéinnec E., Da Silva C., Wincker P., Le Guyader H., Leys S., Jackson D.J., Schreiber F., Erpenbeck D., Morgenstern B., Wörheide G., Manuel M. 2009. Phylogenomics revives traditional views on deep animal relationships. Curr. Biol. 19:706–712.

Philippe H., Laurent J. 1998. How good are deep phylogenetic trees? Curr. Opin. Genet. Dev. 8:616–623.

Phillips M.J., Delsuc F., Penny D. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol. Biol. Evol. 21:1455–1458.

Pick K.S., Philippe H., Schreiber F., Erpenbeck D., Jackson D.J., Wrede P., Wiens M., Alié A., Morgenstern B., Manuel M., Wörheide G. 2010. Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. Mol. Biol. Evol. 27:1983–1987.

Pisani D., Pett W., Dohrmann M., Feuda R., Rota-Stabelli O., Philippe H., Lartillot N., Wörheide G. 2015. Genomic data do not support comb jellies as the sister group to all other animals. Proc. Natl. Acad. Sci. U. S. A. 112:15402.

Planet P.J. 2006. Tree disagreement: measuring and testing incongruence in phylogenies. J. Biomed. Inform. 39:86–102.

Podar M., Haddock S.H., Sogin M.L., Harbison G.R. 2001. A molecular phylogenetic framework for the phylum Ctenophora using 18S rRNA genes. Mol. Phylogenet. Evol. 21:218–230.

Poormohammadi H., Zarchi M.S., Ghaneai H. 2020. NCHB: a method for constructing rooted phylogenetic networks from rooted triplets based on height function and binarization. J. Theor. Biol. 489:110144.

Posada D. 2002. Evaluation of methods for detecting recombination from DNA sequences: empirical data. Mol. Biol. Evol. 19:708–717.

Posada D., Crandall K.A. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics. 14:817–818.

Posada D., Crandall K.A. 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc. Natl. Acad. Sci. U. S. A. 98:13757.

Pozzi L., Hodgson J.A., Burrell A.S., Sterner K.N., Raaum R.L., Disotell T.R. 2014. Primate phylogenetic relationships and divergence dates inferred from complete mitochondrial genomes. Mol. Phylogenet. Evol. 75:165–183.

Prasanna A.N., Gerber D., Kijpornyongpan T., Aime M.C., Doyle V.P., Nagy L.G. 2020. Model choice, missing data, and taxon sampling impact phylogenomic inference of deep Basidiomycota relationships. Syst. Biol. 69:17–37.

Presnell J.S., Vandepas L.E., Warren K.J., Swalla B.J., Amemiya C.T., Browne W.E. 2016. The presence of a functionally tripartite through-gut in Ctenophora has implications for metazoan character trait evolution. Curr. Biol. 26:2814–2820.

Price M.N., Dehal P.S., Arkin A.P. 2010. FastTree 2 – approximately maximum-likelihood trees for large alignments. PLOS ONE. 5:e9490.

Puente-Lelievre C., Malik A.J., Douglas J., Ascher D., Baker M., Allison J., Poole A., Lundin D., Fullmer M., Bouckert R., Kim H., Steinegger M., Matzke N. 2024. Tertiary-interaction characters enable fast, model-based structural phylogenetics beyond the twilight zone. bioRxiv. DOI: 10.1101/2023.12.12.571181.

R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Ragan M.A. 2009. Trees and networks before and after Darwin. Biol. Direct. 4:43.

Rannala B., Edwards S.V., Leaché A., Yang Z. 2020. The multi-species coalescent model and species tree inference. In: Scornavacca C., Delsuc F., Galtier N., editors. Phylogenetics in the Genomic Era. No commercial publisher | Authors open access book. p. 3.3:1-3.3:21.

Rannala B., Yang Z. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J. Mol. Evol. 43:304–311.

Rannala B., Yang Z. 2017. Efficient Bayesian species tree inference under the multispecies coalescent. Syst. Biol. 66:823–842.

Ranwez V., Chantret N. 2020. Strengths and limits of multiple sequence alignment and filtering methods. In: Scornavacca C., Delsuc F., Galtier N., editors. Phylogenetics in the Genomic Era. No commercial publisher | Authors open access book. p. 2.2:1-2.2:36.

Redmond A.K., McLysaght A. 2021a. Evidence for sponges as sister to all other animals from partitioned phylogenomics with mixture models and recoding. Nat. Commun. 12:1783.

Redmond A.K., McLysaght A. 2021b. From: Evidence for sponges as sister to all other animals from partitioned phylogenomics with mixture models and recoding. Version 2. Dataset. Figshare. Available from https://doi.org/10.6084/m9.figshare.12746972.v2.

dos Reis M., Thawornwattana Y., Angelis K., Telford M.J., Donoghue P.C.J., Yang Z. 2015. Uncertainty in the timing of origin of animals and the limits of precision in molecular timescales. Curr. Biol. 25:2939–2950.

Revell L.J. 2012. phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol. 3:217–223.

Reynolds A.S. 2019. Ernst Haeckel and the philosophy of sponges. Theory Biosci. 138:133–146.

Rieppel O. 2020. Morphology and phylogeny. J. Hist. Biol. 53:217–230.

Robinson, D.F., Foulds, L.R. 1981, Comparison of phylogenetic trees, Math. Biosci., 53:131-147

Roch S., Steel M. 2015. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. Theor. Popul. Biol. 100:56–62.

Rodriguez F., Wu F., Ané C., Tanksley S., Spooner D.M. 2009. Do potatoes and tomatoes have a single evolutionary history, and what proportion of the genome supports this history? BMC Evol. Biol. 9:191.

Rokas A., Carroll S.B. 2005. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. Mol. Biol. Evol. 22:1337–1344.

Rokas A., Carroll S.B. 2006. Bushes in the tree of life. PLOS Biol. 4:e352.

Rokas A., Krüger D., Carroll S.B. 2005. Animal evolution and the molecular signature of radiations compressed in time. Science. 310:1933.

Romiguier J., Ranwez V., Delsuc F., Galtier N., Douzery E.J.P. 2013. Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. Mol. Biol. Evol. 30:2134–2144.

Romiguier J., Roux C. 2017. Analytical biases associated with GC-content in molecular evolution. Front. Genet. 8:16.

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. Syst. Biol. 61:539–542.

Rosenzweig B., Kern A., Hahn M. 2022. Accurate detection of incomplete lineage sorting via supervised machine learning. bioRxiv. DOI: 10.1101/2022.11.09.515828.

Roure B., Baurain D., Philippe H. 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. Mol. Biol. Evol. 30:197–214.

Rousselle M., Laverré A., Figuet E., Nabholz B., Galtier N. 2018. Influence of recombination and GC-biased gene conversion on the adaptive and nonadaptive substitution rate in mammals versus birds. Mol. Biol. Evol. 36:458–471.

Roycroft E.J., Moussalli A., Rowe K.C. 2020. Phylogenomics uncovers confidence and conflict in the rapid radiation of Australo-Papuan rodents. Syst. Biol. 69:431–444.

Ryan J.F., Chiodin M. 2015. Where is my mind? How sponges and Placozoans may have lost neural cell types. Philos. Trans. R. Soc. B Biol. Sci. 370:20150059.

Ryan J.F., Pang K., Schnitzler C.E., Nguyen A.-D., Moreland R.T., Simmons D.K., Koch B.J., Francis W.R., Havlak P., Smith S.A., Putnam N.H., Haddock S.H.D., Dunn C.W., Wolfsberg T.G., Mullikin J.C., Martindale M.Q., Baxevanis A.D. 2013. The genome of the Ctenophore Mnemiopsis leidyi and its implications for cell type evolution. Science. 342:1242592.

Saitou N., Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406–425.

Salzburger W., Martens J., Sturmbauer C. 2002. Paraphyly of the blue tit (Parus caeruleus) suggested from cytochrome b sequences. Mol. Phylogenet. Evol. 24:19–25.

Sanderson B.J., Gambhir D., Feng G., Hu N., Cronk Q.C., Percy D.M., Freaner F.M., Johnson M.G., Smart L.B., Keefover-Ring K., Yin T., Ma T., DiFazio S.P., Liu J., Olson M.S. 2023. Phylogenomics reveals patterns of ancient hybridization and differential diversification that contribute to phylogenetic conflict in willows, poplars, and close relatives. Syst. Biol. 72:1220–1232.

Sanderson M.J., McMahon M.M., Steel M. 2010. Phylogenomics with incomplete taxon coverage: the limits to inference. BMC Evol. Biol. 10:155.

Sanger F., Coulson A.R. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J. Mol. Biol. 94:441–448.

Sanger F., Nicklen S., Coulson A.R. 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U. S. A. 74:5463–5467.

Sawyer S. 1989. Statistical tests for detecting gene conversion. Mol. Biol. Evol. 6:526–538.

Sawyer S. 2000. GENECONV: a computer package for the statistical detection of gene conversion. Software. v1.81a. Distributed by the author, Department of Mathematics, Washington University in St. Louis, available at http://www.math.wustl.edu/~sawyer.

Sayyari E., Mirarab S. 2016. Fast coalescent-based computation of local branch support from quartet frequencies. Mol. Biol. Evol. 33:1654–1668.

Sayyari E., Whitfield J.B., Mirarab S. 2018. DiscoVista: interpretable visualizations of gene tree discordance. Mol. Phylogenet. Evol. 122:110–115.

Scharmann M., Wistuba A., Widmer A. 2021. Introgression is widespread in the radiation of carnivorous Nepenthes pitcher plants. Mol. Phylogenet. Evol.:107214.

Schierwater B., Eitel M., Jakob W., Osigus H.-J., Hadrys H., Dellaporta S.L., Kolokotronis S.-O., DeSalle R. 2009. Concatenated analysis sheds light on early metazoan evolution and fuels a modern “Urmetazoon” hypothesis. PLOS Biol. 7:e1000020.

Schierwater B., Osigus H.-J., Bergmann T., Blackstone N.W., Hadrys H., Hauslage J., Humbert P.O., Kamm K., Kvansakul M., Wysocki K., DeSalle R. 2021. The enigmatic Placozoa part 1: exploring evolutionary controversies and poor ecological knowledge. BioEssays. 43:2100080.

Schliep K., Potts A.J., Morrison D.A., Grimm G.W. 2017. Intertwining phylogenetic trees and networks. Methods Ecol. Evol. 8:1212–1220.

Schliep K.P. 2011. phangorn: phylogenetic analysis in R. Bioinformatics. 27:592–592.

Schmidt H.A. 2009. Testing tree topologies. In: Lemey P., Salemi M., Vandamme A.-M., editors. A phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing. Cambridge: Cambridge University Press. p. 381–405.

Schrago C.G., Seuánez H.N. 2019. Large ancestral effective population size explains the difficult phylogenetic placement of owl monkeys. Am. J. Primatol. 81:e22955.

Schultz D.T., Haddock S.H.D., Bredeson J.V., Green R.E., Simakov O., Rokhsar D.S. 2023. Ancient gene linkages support Ctenophores as sister to other animals. Nature. 618:1–8.

Schwarz G. 1978. Estimating the dimension of a model. Ann. Stat. 6:461–464.

Scornavacca C., Galtier N. 2017. Incomplete lineage sorting in mammalian phylogenomics. Syst. Biol. 66:112–120.

Sharma P.P., Fernández R., Esposito L.A., González-Santillán E., Monod L. 2015. Phylogenomic resolution of scorpions reveals multilevel discordance with morphological phylogenetic signal. Proc. R. Soc. B Biol. Sci. 282:20142953.

Shen X.-X., Hittinger C.T., Rokas A. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nat. Ecol. Evol. 1:0126.

Shen X.-X., Salichos L., Rokas A. 2016. A genome-scale investigation of how sequence, function, and tree-based gene properties influence phylogenetic inference. Genome Biol. Evol. 8:2565–2580.

Shi C.M., Yang Z. 2018. Coalescent-based analyses of genomic sequence data provide a robust resolution of phylogenetic relationships among major groups of gibbons. Mol. Biol. Evol. 35:159–179.

Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol. 51:492–508.

Shimodaira H., Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. 16:1114–1114.

Short D.P., O'Donnell K., Geiser D.M. 2014. Clonality, recombination, and hybridization in the plumbing-inhabiting human pathogen Fusarium keratoplasticum inferred from multilocus sequence typing. BMC Evol. Biol. 14:91.

Shulaev V., Sargent D.J., Crowhurst R.N., Mockler T.C., Folkerts O., Delcher A.L., Jaiswal P., Mockaitis K., Liston A., Mane S.P., Burns P., Davis T.M., Slovin J.P., Bassil N., Hellens R.P., Evans C., Harkins T., Kodira C., Desany B., Crasta O.R., Jensen R.V., Allan A.C., Michael T.P., Setubal J.C., Celton J.-M., Rees D.J.G., Williams K.P., Holt S.H., Rojas J.J.R., Chatterjee M., Liu B., Silva H., Meisel L., Adato A., Filichkin S.A., Troggio M., Viola R., Ashman T.-L., Wang H., Dharmawardhana P., Elser J., Raja R., Priest H.D., Bryant D.W., Fox S.E., Givan S.A., Wilhelm L.J., Naithani S., Christoffels A., Salama D.Y., Carter J., Girona E.L., Zdepski A., Wang W., Kerstetter R.A., Schwab W., Korban S.S., Davik J., Monfort A., Denoyes-Rothan B., Arus P., Mittler R., Flinn B., Aharoni A., Bennetzen J.L., Salzberg S.L., Dickerman A.W., Velasco R., Borodovsky M., Veilleux R.E., Folta K.M. 2011. The genome of woodland strawberry (Fragaria vesca). Nat. Genet. 43:109–116.

Silvestro D., Warnock R.C.M., Gavryushkina A., Stadler T. 2018. Closing the gap between palaeontological and neontological speciation and extinction rate estimates. Nat. Commun. 9:5237.

Simion P., Delsuc F., Philippe H. 2020. To what extent current limits of phylogenomics can be overcome? In: Scornavacca C., Delsuc F., Galtier N., editors. Phylogenetics in the Genomic Era. No commercial publisher | Authors open access book. p. 2.1:1-2.1:34.

Simion P., Philippe H., Baurain D., Jager M., Richter D.J., Di Franco A., Roure B., Satoh N., Quéinnec É., Ereskovsky A., Lapébie P., Corre E., Delsuc F., King N., Wörheide G., Manuel M. 2017a. A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. Curr. Biol. 27:958–967.

Simion P., Philippe H., Baurain D., Jager M., Richter D.J., Franco A.D., Roure B., Satoh N., Quéinnec É., Ereskovsky A., Lapébie P., Corre E., Delsuc F., King N., Wörheide G., Manuel M. 2017b. SuppData_Metazoa_2017. Available from GitHub https://github.com/psimion/SuppData_Metazoa_2017.

Simmons M.P., Gatesy J. 2015. Coalescence vs. concatenation: sophisticated analyses vs. first principles applied to rooting the angiosperms. Mol. Phylogenet. Evol. 91:98–122.

Simon C. 2022. An evolving view of phylogenetic support. Syst. Biol. 71:921–928.

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. Mol. Biol. Evol. 36:1344–1356.

Skaloud P., Peksa O. 2010. Evolutionary inferences based on ITS rDNA and actin sequences reveal extensive diversity of the common lichen alga Asterochloris (Trebouxiophyceae, Chlorophyta). Mol. Phylogenet. Evol. 54:36–46.

Smith M.L., Vanderpool D., Hahn M.W. 2022. Using all gene families vastly expands data available for phylogenomic inference. Mol. Biol. Evol. 39:msac112.

Smith M.R., Jonker R., Yang Y., Cao Y. 2023. TreeDist: calculate and map distances between phylogenetic trees. v2.7.0. Available from https://cran.r-project.org/web/packages/TreeDist/.

Smith M.R., Paradis E. 2023. TreeTools: create, modify and analyse phylogenetic trees. R package version 1.10.1. Available at https://CRAN.R-project.org/package=TreeTools.

Smith S.A., Walker-Hale N., Walker J.F. 2020. Intragenic conflict in phylogenomic data sets. Mol. Biol. Evol. 37:3380–3388.

Solís-Lemus C., Ané C. 2016. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. PLOS Genet. 12:e1005896.

Solís-Lemus C., Bastide P., Ané C. 2017. PhyloNetworks: a package for phylogenetic networks. Mol. Biol. Evol. 34:3292–3298.

Song S., Liu L., Edwards S.V., Wu S. 2012. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proc. Natl. Acad. Sci. U. S. A. 109:14942.

Soubrier J., Steel M., Lee M.S.Y., Der Sarkissian C., Guindon S., Ho S.Y.W., Cooper A. 2012. The influence of rate heterogeneity among sites on the time dependence of molecular rates. Mol. Biol. Evol. 29:3345–3358.

Sperling E.A., Peterson K.J., Pisani D. 2009. Phylogenetic-signal dissection of nuclear housekeeping genes supports the paraphyly of sponges and the monophyly of Eumetazoa. Mol. Biol. Evol. 26:2261–2274.

Sperling E.A., Pisani D., Peterson K.J. 2007. Poriferan paraphyly and its implications for Precambrian palaeobiology. Geol. Soc. Lond. Spec. Publ. 286:355–368.

Sperling E.A., Robinson J.M., Pisani D., Peterson K.J. 2010. Where's the glass? Biomarkers, molecular clocks, and microRNAs suggest a 200-Myr missing Precambrian fossil record of siliceous sponge spicules. Geobiology. 8:24–36.

Springer M.S., Gatesy J. 2016. The gene tree delusion. Mol. Phylogenet. Evol. 94:1–33.

Springer M.S., Gatesy J. 2018. Delimiting coalescence genes (c-genes) in phylogenomic data sets. Genes. 9:123.

Stadler T. 2011. Simulating trees with a fixed number of extant species. Syst. Biol. 60:676–684.

Stadler T. 2017. TreeSim: simulating phylogenetic trees. R package version 2.4. http://CRAN.R-project.org/package=TreeSim.

Stadler T., Bokma F. 2013. Estimating speciation and extinction rates for phylogenies of higher taxa. Syst. Biol. 62:220–230.

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

Steel M.A., Penny D. 1993. Distributions of tree comparison metrics—some new results. Syst. Biol. 42:126–141.

Steenwyk J.L., Li Y., Zhou X., Shen X.-X., Rokas A. 2023. Incongruence in the phylogenomics era. Nat. Rev. Genet. 14:1–17.

Steiner G., Dreyer H. 2003. Molecular phylogeny of Scaphopoda (Mollusca) inferred from 18S rDNA sequences: support for a Scaphopoda–Cephalopoda clade. Zool. Scr. 32:343–356.

Stenz N.W.M., Larget B., Baum D.A., Ané C. 2015. Exploring tree-like and non-tree-like patterns using genome sequences: an example using the inbreeding plant species Arabidopsis thaliana (L.) Heynh. Syst. Biol. 64:809–823.

Stiller J., Feng S., Chowdhury A.-A., Rivas-González I., Duchêne D.A., Fang Q., Deng Y., Kozlov A., Stamatakis A., Claramunt S., Nguyen J.M.T., Ho S.Y.W., Faircloth B.C., Haag J., Houde P., Cracraft J., Balaban M., Mai U., Chen G., Gao R., Zhou C., Xie Y., Huang Z., Cao Z., Yan Z., Ogilvie H.A., Nakhleh L., Lindow B., Morel B., Fjeldså J., Hosner P.A., da Fonseca R.R., Petersen B., Tobias J.A., Székely T., Kennedy J.D., Reeve A.H., Liker A., Stervander M., Antunes A., Tietze D.T., Bertelsen M.F., Lei F., Rahbek C., Graves G.R., Schierup M.H., Warnow T., Braun E.L., Gilbert M.T.P., Jarvis E.D., Mirarab S., Zhang G. 2024. Complexity of avian evolution revealed by family-level genomes. Nature. 629:851–860.

Strassert J.F.H., Irisarri I., Williams T.A., Burki F. 2021. A molecular timescale for eukaryote evolution with implications for the origin of red algal-derived plastids. Nat. Commun. 12:1879.

Strimmer K., von Haeseler A. 1997. Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. Proc. Natl. Acad. Sci. U. S. A. 94:6815–6819.

Sugiura N. 1978. Further analysis of the data by Akaike's information criterion and the finite corrections. Commun. Stat. - Theory Methods. 7:13–26.

Sullivan J., Joyce P. 2005. Model selection in phylogenetics. Annu. Rev. Ecol. Evol. Syst. 36:445–466.

Susko E., Roger A.J. 2021. Long branch attraction biases in phylogenetics. Syst. Biol. 70:838–843.

Suvorov A., Kim B.Y., Wang J., Armstrong E.E., Peede D., D'Agostino E.R.R., Price D.K., Waddell P.J., Lang M., Courtier-Orgogozo V., David J.R., Petrov D., Matute D.R., Schrider D.R., Comeault A.A. 2022. Widespread introgression across a phylogeny of 155 Drosophila genomes. Curr. Biol. 32:111-123.e5.

Swofford D.L. 2019. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.0a166. Sinauer Associates, Sunderland, Massachusetts. Available from https://paup.phylosolutions.com/.

Swofford D.L., Sullivan J. 2003. Phylogeny inference based on parsimony and other methods using PAUP*. In: Salemi M., Vandamme A.-M., editors. The Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny. Cambridge University Press.

Syrjänen K., Maurits L., Leino U., Honkola T., Rota J., Vesakoski O. 2021. Crouching TIGER, hidden structure: exploring the nature of linguistic data using TIGER values. J. Lang. Evol. 6:99–118.

Szánthó L.L., Lartillot N., Szöllősi G.J., Schrempf D. 2023. Compositionally constrained sites drive long-branch attraction. Syst. Biol. 72:767–780.

Tamura K., Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. 10:512–526.

Tan G., Muffato M., Ledergerber C., Herrero J., Goldman N., Gil M., Dessimoz C. 2015. Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference. Syst. Biol. 64:778–791.

Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. In: Miura R.M., editor. Some probabilistic and statistical problems in the analysis of DNA sequences. Providence: American Mathematical Society. p. 57–86.

Telford M.J., Budd G.E., Philippe H. 2015. Phylogenomic insights into animal evolution. Curr. Biol. 25:R876–R887.

Thawornwattana Y., Seixas F.A., Yang Z., Mallet J. 2022. Full-likelihood genomic analysis clarifies a complex history of species divergence and introgression: the example of the erato-sara group of Heliconius butterflies. Syst. Biol.71:1159–1177.

Thomson R.C., Brown J.M. 2022. On the need for new measures of phylogenomic support. Syst. Biol. 71:917–920.

Thorne J.L., Goldman N., Jones D.T. 1996. Combining protein evolution and secondary structure. Mol. Biol. Evol. 13:666–673.

Tonini J., Moore A., Stern D., Shcheglovitova M., Ortí G. 2015. Concatenation and species tree methods exhibit statistically indistinguishable accuracy under a range of simulated conditions. PLOS Curr. 7:ecurrents.tol.34260cc27551a527b124ec5f6334b6be.

Torres A., Goloboff P.A., Catalano S.A. 2021. Assessing topological congruence among concatenation-based phylogenomic approaches in empirical datasets. Mol. Phylogenet. Evol. 161:107086.

Tung J., Barreiro L.B. 2017. The contribution of admixture to primate evolution. Curr. Opin. Genet. Dev. 47:61–68.

Turakhia Y., Thornlow B., Hinrichs A.S., De Maio N., Gozashti L., Lanfear R., Haussler D., Corbett-Detig R. 2021. Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. Nat. Genet. 53:809–816.

Vanderpool D., Minh B.Q., Lanfear R., Hughes D., Murali S., Harris R.A., Raveendran M., Muzny D.M., Gibbs R.A., Worley K.C., Rogers J., Hahn M.W., Hibbins M.S., Williamson R.J. 2020a. Supplementary data for: Primate phylogenomics uncovers multiple rapid radiations and ancient interspecific introgression. Dryad. Available from doi.org/10.5061/dryad.rfj6q577d.

Vanderpool D., Minh B.Q., Lanfear R., Hughes D., Murali S., Harris R.A., Raveendran M., Muzny D.M., Hibbins M.S., Williamson R.J., Gibbs R.A., Worley K.C., Rogers J., Hahn M.W. 2020b. Primate phylogenomics uncovers multiple rapid radiations and ancient interspecific introgression. PLOS Biol. 18:e3000954.

Vanhove M.P.M., Pariselle A., Van Steenberge M., Raeymaekers J.A.M., Hablützel P.I., Gillardin C., Hellemans B., Breman F.C., Koblmüller S., Sturmbauer C., Snoeks J., Volckaert F.A.M., Huyse T. 2015. Hidden biodiversity in an ancient lake: phylogenetic congruence between Lake Tanganyika tropheine cichlids and their monogenean flatworm parasites. Sci. Rep. 5:13669.

Varadi M., Anyango S., Deshpande M., Nair S., Natassia C., Yordanova G., Yuan D., Stroe O., Wood G., Laydon A., Žídek A., Green T., Tunyasuvunakool K., Petersen S., Jumper J., Clancy E., Green R., Vora A., Lutfi M., Figurnov M., Cowie A., Hobbs N., Kohli P., Kleywegt G., Birney E., Hassabis D., Velankar S. 2022. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 50:D439–D444.

Vďačný P. 2017. Integrative taxonomy of ciliates: assessment of molecular phylogenetic content and morphological homology testing. Eur. J. Protistol. 61:388–398.

Wainright P.O., Hinkle G., Sogin M.L., Stickel S.K. 1993. Monophyletic origins of the Metazoa: an evolutionary link with Fungi. Science. 260:340–342.

Wang H.-C., Minh B.Q., Susko E., Roger A.J. 2018. Modeling site heterogeneity with Posterior Mean Site Frequency profiles accelerates accurate phylogenomic estimation. Syst. Biol. 67:216–235.

Wang H.-C., Spencer M., Susko E., Roger A.J. 2007. Testing for covarion-like evolution in protein sequences. Mol. Biol. Evol. 24:294–305.

Wang M., Cheng F. 2019. The complete mitochondrial genome of the Ctenophore Beroe cucumis, a mitochondrial genome showing rapid evolutionary rates. Mitochondrial DNA Part B. 4:3774–3775.

Weisrock D.W., Smith S.D., Chan L.M., Biebouw K., Kappeler P.M., Yoder A.D. 2012. Concatenation and concordance in the reconstruction of mouse lemur phylogeny: an empirical demonstration of the effect of allele sampling in phylogenetics. Mol. Biol. Evol. 29:1615–1630.

Wen D., Yu Y., Zhu J., Nakhleh L. 2018. Inferring phylogenetic networks using PhyloNet. Syst. Biol. 67:735–740.

Whelan N.V., Halanych K.M. 2017. Who let the CAT out of the bag? Accurately dealing with substitutional heterogeneity in phylogenomic analyses. Syst. Biol. 66:232–255.

Whelan N.V., Kocot K.M., Halanych K.M. 2015a. Employing phylogenomics to resolve the relationships among Cnidarians, Ctenophores, Sponges, Placozoans, and Bilaterians. Integr. Comp. Biol. 55:1084–1095.

Whelan N.V., Kocot K.M., Moroz L.L., Halanych K.M. 2015b. Error, signal, and the placement of Ctenophora sister to all other animals. Proc. Natl. Acad. Sci. U. S. A. 112:5773–5778.

Whelan N.V., Kocot K.M., Moroz L.L., Halanych K.M. 2016. Error, signal, and the placement of Ctenophora sister to all other animals. v3. Dataset. Figshare. Available from https://doi.org/10.6084/m9.figshare.1334306.v3.

Whelan N.V., Kocot K.M., Moroz T.P., Mukherjee K., Williams P., Paulay G., Moroz L.L., Halanych K.M. 2017a. Ctenophore relationships and their placement as the sister group to all other animals. Nat. Ecol. Evol. 1:1737–1746.

Whelan N.V., Kocot K.M., Moroz T.P., Mukherjee K., Williams P., Paulay G., Moroz L.L., Halanych K.M. 2017b. Ctenophora phylogeny datasets and core orthologs. Version 1. Dataset. Figshare. Available from doi.org/10.6084/m9.figshare.4484138.v1.

Whelan S., Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. 18:691–699.

White W.T., Hills S.F., Gaddam R., Holland B.R., Penny D. 2007. Treeness triangles: visualizing the loss of phylogenetic signal. Mol. Biol. Evol. 24:2029–2039.

Wickham H. 2007. Reshaping data with the reshape package. J. Stat. Softw. 21:1–20.

Wickham H. 2016. ggplot2: elegant graphics for data analysis. New York: Springer-Verlag.

Wickham H. 2023. stringr: simple, consistent wrappers for common string operations. R package. Version 1.5.1. Available at https://cran.r-project.org/package=stringr.

Wickham H., Bryan J. 2023. readxl: read Excel files. R package version 1.4.3. Available at https://CRAN.R-project.org/package=readxl.

Wickham H., François R., Henry L., Müller K. 2021. dplyr: a grammar of data manipulation. Package version 1.1.4. Available at https://CRAN.R-project.org/package=dplyr.

Wickham H., Seidel D. 2022. scales: scale functions for visualization. R package version 1.2.1. https://CRAN.R-project.org/package=scales. .

Wilke C.O. 2020. ggtext: improved text rendering support for “ggplot2”. Version 0.1.1. Available from https://CRAN.R-project.org/package=ggtext.

Wisniewski A.L., Lloyd G.T., Slater G.J. 2022. Extant species fail to estimate ancestral geographical ranges at older nodes in primate phylogeny. Proc. R. Soc. B Biol. Sci. 289:20212535.

Woese C.R., Fox G.E. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. U. S. A. 74:5088–5090.

Wolfe J.M., Ballou L., Luque J., Watson-Zink V.M., Ahyong S.T., Barido-Sottani J., Chan T.-Y., Chu K.H., Crandall K.A., Daniels S.R., Felder D.L., Mancke H., Martin J.W., Ng P.K.L., Ortega-Hernández J., Palacios Theil E., Pentcheff N.D., Robles R., Thoma B.P., Tsang L.M., Wetzer R., Windsor A.M., Bracken-Grissom H.D. 2023. Convergent adaptation of true crabs (Decapoda: Brachyura) to a gradient of terrestrial environments. Syst. Biol. 73:247–262.

Wong T.K.F., Cherryh C., Rodrigo A.G., Hahn M.W., Minh B.Q., Lanfear R. 2024. MAST: phylogenetic inference with Mixtures Across Sites and Trees. Syst. Biol. 73:375–391.

Woodhams M.D., Fernández-Sánchez J., Sumner J.G. 2015. A new hierarchy of phylogenetic models consistent with heterogeneous substitution rates. Syst. Biol. 64:638–650.

Wu S., Edwards S., Liu L. 2018. Genome-scale DNA sequence data and the evolutionary history of placental mammals. Data Brief. 18:1972–1975.

wwPDB consortium. 2019. Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res. 47:D520–D528.

Xi Z., Liu L., Rest J.S., Davis C.C. 2014. Coalescent versus concatenation methods and the placement of Amborella as sister to water lilies. Syst. Biol. 63:919–932.

Xu S., Li L., Luo X., Chen M., Tang W., Zhan L., Dai Z., Lam T.T.-Y., Guan Y., Yu G. 2022. ggtree: a serialized data object for visualization of a phylogenetic tree and annotation data. iMeta. 1:e56.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39:306–314.

Yang Z. 1995. A space-time process model for the evolution of DNA sequences. Genetics. 139:993–1005.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.

Yang Z., Rannala B. 2012. Molecular phylogenetics: principles and practice. Nat. Rev. Genet. 13:303.

Yu G. 2020. Using ggtree to visualize data on tree-like structures. Curr. Protoc. Bioinforma. 69:e96.

Yu G. 2022. Data integration, manipulation and visualization of phylogenetic trees. Florida: Chapman and Hall/CRC.

Yu G., Lam T.T.-Y., Zhu H., Guan Y. 2018. Two methods for mapping and visualizing associated data on phylogeny using ggtree. Mol. Biol. Evol. 35:3041–3043.

Yu G., Smith D.K., Zhu H., Guan Y., Lam T.T.-Y. 2017. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol. Evol. 8:28–36.

Zhang C., Ogilvie H.A., Drummond A.J., Stadler T. 2018a. Bayesian inference of species networks from multilocus sequence data. Mol. Biol. Evol. 35:504–517.

Zhang C., Rabiee M., Sayyari E., Mirarab S. 2018b. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics. 19:153.

Zhang J. 2017. phylotools: phylogenetic tools for eco-phylogenetics. R package version 0.2.2. Available at https://CRAN.R-project.org/package=phylotools.

Zhang S.-Q., Che L.-H., Li Y., Dan Liang, Pang H., Ślipiński A., Zhang P. 2018c. Evolutionary history of Coleoptera revealed by extensive sampling of genes and species. Nat. Commun. 9:205.

Zhao L., Li X., Zhang N., Zhang S.-D., Yi T.-S., Ma H., Guo Z.-H., Li D.-Z. 2016. Phylogenomic analyses of large-scale nuclear genes provide new insights into the evolutionary relationships within the rosids. Mol. Phylogenet. Evol. 105:166–176.

Zharkikh A. 1994. Estimation of evolutionary distances between nucleotide sequences. J. Mol. Evol. 39:315–329.

Zuckerkandl E., Pauling L. 1965. Molecules as documents of evolutionary history. J. Theor. Biol. 8:357–366.

Zuriaga E., Blanca J., Nuez F. 2009. Classification and phylogenetic relationships in Solanum section Lycopersicon based on AFLP and two nuclear gene sequences. Genet. Resour. Crop Evol. 56:663–678.