Methods in Molecular Biology 1910

Springer Protocols

Maria Anisimova *Editor*

Evolutionary Genomics

Statistical and Computational Methods

Second Edition

OPEN

Humana Press

METHODS IN MOLECULAR BIOLOGY

Series Editor
John M. Walker
School of Life and Medical Sciences
University of Hertfordshire
Hatfield, Hertfordshire, AL10 9AB, UK

For further volumes:
http://www.springer.com/series/7651

Evolutionary Genomics

Statistical and Computational Methods

Second Edition

Edited by

Maria Anisimova

Institute of Applied Simulations, School of Life Sciences and Facility Management, Zurich University of Applied Sciences (ZHAW), Wädenswil, Switzerland

Swiss Institute of Bioinformatics, Lausanne, Switzerland

OPEN Humana Press

Editor
Maria Anisimova
Institute of Applied Simulations
School of Life Sciences and Facility Management
Zurich University of Applied Sciences (ZHAW)
Wädenswil, Switzerland

Swiss Institute of Bioinformatics
Lausanne, Switzerland

ISSN 1064-3745 ISSN 1940-6029 (electronic)
Methods in Molecular Biology
ISBN 978-1-4939-9073-3 ISBN 978-1-4939-9074-0 (eBook)
https://doi.org/10.1007/978-1-4939-9074-0

This book is an open access publication.

© The Editor(s) (if applicable) and The Author(s) 2012, 2019.
Open Access This book is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.
The images or other third party material in this book are included in the book's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the book's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Humana imprint is published by the registered company Springer Science+Business Media, LLC, part of Springer Nature.
The registered company address is: 233 Spring Street, New York, NY 10013, U.S.A.

Chapter 5

Inferring Orthology and Paralogy

Adrian M. Altenhoff, Natasha M. Glover, and Christophe Dessimoz

Abstract

The distinction between orthologs and paralogs, genes that started diverging by speciation versus duplication, is relevant in a wide range of contexts, most notably phylogenetic tree inference and protein function annotation. In this chapter, we provide an overview of the methods used to infer orthology and paralogy. We survey both graph-based approaches (and their various grouping strategies) and tree-based approaches, which solve the more general problem of gene/species tree reconciliation. We discuss conceptual differences among the various orthology inference methods and databases and examine the difficult issue of verifying and benchmarking orthology predictions. Finally, we review typical applications of orthologous genes, groups, and reconciled trees and conclude with thoughts on future methodological developments.

Key words Orthology, Paralogy, Tree reconciliation, Orthology benchmarking

1 Introduction

The study of genetic material almost always starts with identifying, within or across species, *homologous* regions—regions of common ancestry. As we have seen in previous chapters, this can be done at the level of genome segments [1], genes [2], or even down to single residues, in sequence alignments [3]. Here, we focus on genes as evolutionary and functional units. The central premise of this chapter is that it is useful to distinguish between two classes of homologous genes: *orthologs*, which are pairs of genes that started diverging via evolutionary speciation, and *paralogs*, which are pairs of genes that started diverging via gene duplication [4] (Fig. 1, Box 1). Originally, the terms and their definition were proposed by Walter M. Fitch in the context of species phylogeny inference, i.e., the reconstruction of the tree of life. He stated "Phylogenies require orthologous, not paralogous, genes" [4]. Indeed, since orthologs arise by speciation, any set of genes in which every pair is orthologous has by definition the same evolutionary history as the

Adrian M. Altenhoff and Natasha M. Glover are the Joint first authors

Maria Anisimova (ed.), *Evolutionary Genomics: Statistical and Computational Methods*, Methods in Molecular Biology, vol. 1910, https://doi.org/10.1007/978-1-4939-9074-0_5, © The Author(s) 2019

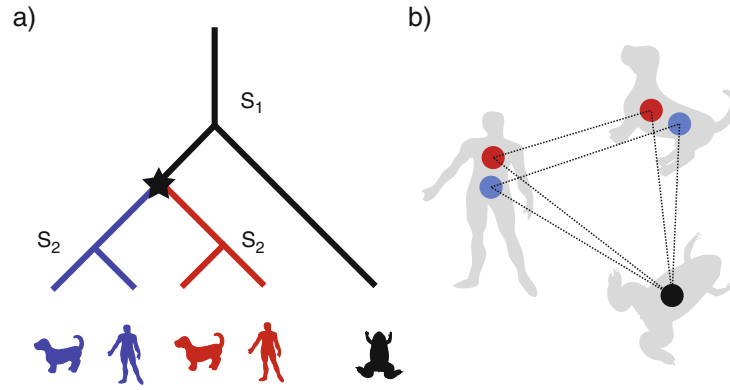
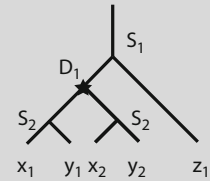


Fig. 1 (**a**) Simple evolutionary scenario of a gene family with two speciation events (S_1 and S_2) and one duplication event (star). The type of events completely and unambiguously define all pairs of orthologs and paralogs: The frog gene is orthologous to all other genes (they coalesce at S_1). The red and blue genes are orthologs between themselves (they coalesce at S_2), but paralogs between each other (they coalesce at star). (**b**) The corresponding orthology graph. The genes are represented here by vertices and orthology relationships by edges. The frog gene forms *one-to-many* orthology with both the human and dog genes, because it is orthologous to more than one sequence in each of these organisms. In such cases, the *bi-directional best-hit* approach only recovers one of the relations (the highest scoring one). Note that in contrary to BBH, the nonsymmetric BeTs approach—simply taking the best genome-wide hit for each gene regardless of reciprocity—would in the situation of a lost blue human gene infer an incorrect orthologous relation between the blue dog and red human gene

underlying species. These days, however, the most frequent motivation for the orthology/paralogy distinction is to study and predict gene function: it is generally believed that orthologs—because they were the same gene in the last common ancestor of the species involved—are likely to have similar biological function. By contrast, paralogs—because they result from duplicated genes that have been retained, at least partly, over the course of evolution—are believed to often differ in function. Consequently, orthologs are of interest to infer function computationally, while paralogs are commonly used to study function innovation.

Box 1: Terminology

Homology is a relation between a pair of genes that share a common ancestor. All pairs of genes in the below figure are homologous to each other.



(continued)

Box 1: (continued)

Orthology is a relation defined over a pair of homologous genes, where the two genes have emerged through a speciation event [4]. Example pairs of orthologs are (x_1, y_1) or (x_2, z_1). Orthologs can be further subclassified into one-to-one, one-to-many, many-to-one, and many-to-many orthologs. The qualifiers *one* and *many* indicate for each of the two involved genes whether they underwent an additional duplication after the speciation between the two genomes. Hence, the gene pair (x_1, y_1) is an example of a one-to-one orthologous pair, whereas (x_2, z_1) is a many-to-one ortholog relation.

Paralogy is a relation defined over a pair of homologous genes that have emerged through a gene duplication, e.g., (x_1, x_2) or (x_1, y_2).

In-Paralogy is a relation defined over a triplet. It involves a pair of genes and a speciation event of reference. A gene pair is an in-paralog if they are paralogs and duplicated *after* the speciation event of reference [5]. The pair (x_1, y_2) are in-paralogs with respect to the speciation event S_1.

Out-Paralogy is also a relation defined over a pair of genes and a speciation event of reference. This pair is out-paralogs if the duplication event through which they are related to each other *predates* the speciation event of reference. Hence, the pair (x_1, y_2) are out-paralogs with respect to the speciation event S_2.

Co-orthology is a relation defined over three genes, where two of them are in-paralogs with respect to the speciation event associated to the third gene. The two in-paralogous genes are said to be *co-orthologous* to the third (out-group) gene. Thus, x_1 and y_2 are co-orthologs with respect to z_1.

Homoeology is a specific type of homologous relation in a polyploid species, which thus contain multiple "sub-genomes." This relation describes pairs of genes that originated by speciation and were brought back together in the same genome by allopolyploidization (hybridization) [6]. Thus, in the absence of rearrangement, homoeologs can be thought of as orthologs between sub-genomes.

In this chapter, we first review the main methods used to infer orthology and paralogy, including recent techniques for scaling up algorithms to big data. We then discuss the problem of benchmarking orthology inference. In the last main section, we focus on various applications of orthology and paralogy.

2 Inferring Orthology

Most orthology inference methods can be classified into two major types: graph-based methods and tree-based methods [7]. Methods of the first type rely on graphs with genes (or proteins) as nodes and evolutionary relationships as edges. They infer whether these edges represent orthology or paralogy and build clusters of genes on the basis of the graph. Methods of the second type are based on gene/species tree reconciliation, which is the process of annotating all splits of a given gene tree as duplication or speciation, given the phylogeny of the relevant species. From the reconciled tree, it is trivial to derive all pairs of orthologous and paralogous genes. All pairs of genes which coalesce in a speciation node are orthologs and paralogs if they split at a duplication node. In this section, we present the concepts and methods associated with the two types and discuss the advantages, limitations, and challenges associated with them.

2.1 Graph-Based Methods

Graph-based approaches were originally motivated by the availability of complete genome sequences and the need for efficient methods to detect orthology. They typically run in two phases: a graph construction phase, in which pairs of orthologous genes are inferred (implicitly or explicitly) and connected by edges, and a clustering phase, in which groups of orthologous genes are constructed based on the structure of the graph.

2.1.1 Graph Construction Phase: Orthology Inference

In its most basic form, the graph construction phase identifies orthologous genes by considering pairs of genomes at a time. The main idea is that between any given two genomes, the orthologs tend to be the homologs that diverged least. Why? Because assuming that speciation and duplication are the only types of branching events, the orthologs branched by definition at the latest possible time point—the speciation between the two genomes in question. Therefore, using sequence similarity score as surrogate measure of closeness, the basic approach identifies the corresponding ortholog of each gene through its genome-wide best hit (*BeT*)—the highest scoring match in the other genome [8]. To make the inference symmetric (as orthology is a symmetric relation), it is usually required that BeTs be reciprocal, i.e., that orthology be inferred for a pair of genes g_1 and g_2 if and only if g_2 is the BeT of g_1 and g_1 is the BeT of g_2 [9]. This symmetric variant, referred to as *bi-directional best hit* (*BBH*), has also the merit of being more robust against a possible gene loss in one of the two lineages (Fig. 1).

Inferring orthology from BBH is computationally efficient, because each genome pair can be processed independently and high-scoring alignments can be computed efficiently using dynamic programming [10] or heuristics such as BLAST [11]. Overall, the

time complexity scales quadratically in terms of the total number of genes (Box 2). Furthermore, the implementation of this kind of algorithm is simple.

Box 2: Computational Considerations for Scaling to Many Genomes

Time complexity—the amount of time for an algorithm to run as a function of the input—is an important consideration when dealing with big data. This is relevant for inferring orthologs and paralogs due to the massive amounts of sequence data. Thus, it is necessary to consider the time complexity of the inference algorithms, especially when scaling for large and multiple genomes. In computer science, this is commonly denoted in terms of "Big O" notation, which expresses the scaling behavior of the algorithm, up to a constant factor. Below are listed the common time complexities for aspects of some orthology inference algorithms, in order of most efficient to least efficient.

Linear time

- $O(n)$: Optimal algorithm to reconcile rooted, fully resolved gene tree and species tree [12]; Hieranoid algorithm, which recursively merges genomes along the species tree to avoid all-against-all computation [13].

Quadratic time

- $O(n^2)$: The all-against-all stage central to many orthology algorithms scales quadratically, where n is total number of genes.

Cubic time

- $O(n^3)$: The COG database's graph-based clustering merge triplets of homologs which share a common face until no more can be added.

NP-complete

- "Nondeterministic polynomial time," a large class of algorithms for which no solution in polynomial time is known, (e.g. scaling exponentially with respect to the input size), and thus are impractical. NP-complete problems are typically solved approximately, using heuristics. For instance, maximum likelihood gene tree estimation is NP-complete [14].

However, orthology inference by BBH has several limitations, which motivated the development of various improvements (Table 1).

Table 1
Overview of graph-based orthology inference methods and their main properties

Method	In-paralogs	Based on	Grouping strategy	Database	Extra	Available algorithm/DB	References
BBH (best bi-directional hit)	No	BLAST scores	n.a.	–		–/–	[9]
COG	Yes	BLAST scores	Merged adjacent triangles of BeTs	COG/KOG		✓/✓	[8]
EggNOG	Yes	Smith Waterman scores	Hierarchical orthologous groups	EggNOG	Computed at several levels of taxonomic tree	–/✓	[15–17]
Hieranoid	Yes	BLAST scores and HMM profiles	Hierarchical orthologous groups	HieranoiDB		✓/✓	[13, 18]
InParanoid	Yes	BLAST scores	Orthologous groups between pairs of species	InParanoid		✓/✓	[5, 19, 20]
OMA GETHOGS	Yes	ML distance estimates	Hierarchical orthologous groups	OMA Browser	Computed at all levels of the taxonomic tree	✓/✓	[21, 22]
OMA Pairs	Yes	ML distance estimates	Every pair is orthologous	OMA Browser	Detects differential gene loss	✓/✓	[23, 24]
OrthoDB	Yes	Smith Waterman scores	Hierarchical orthologous groups	OrthoDB	Computed at any level of taxonomic tree	✓/✓	[25, 26])
OrthoInspector	Yes	BLAST scores	Only between pairs of species	OrthoInspector		✓/✓	[27, 28]
OrthoMCL	Yes	BLAST scores	MCL clusters	OrthoMCL-DB		✓/✓	[29, 30]
RSD (reciprocal smallest distance)	No	ML distance estimates	Deterministic single-linkage clustering	–		✓/✓	[31–33]

Allowing for More Than One Ortholog

Some genes can have more than one orthologous counterpart in a given genome. This happens whenever a gene undergoes duplication *after* the speciation of the two genomes in question. Since BBH only picks the best hit, it only captures part of the orthologous relations (Fig. 1). The existence of multiple orthologous counterparts is often referred to as *one-to-many* or *many-to-many* orthology, depending whether duplication took place in one or both lineages. To designate the copies resulting from such duplications occurring *after* a speciation of reference, Remm et al. coined the term *in-paralogs* and introduced a method called *InParanoid* that improves upon BBH by potentially identifying all pairs of many-to-many orthologs [5]. In brief, their algorithm identifies all paralogs within a species that are evolutionarily closer (more similar) to each other than to the BBH gene in the other genome. This results in two sets of in-paralogs—one for each species—where all pairwise combinations between the two sets are orthologous relations. Alternatively, it is possible to identify many-to-many orthology by relaxing the notion of "best hit" to "group of best hits." This can be implemented using a score tolerance threshold or a confidence interval around the BBH [23, 34].

Evolutionary Distances

Instead of using sequence similarity as a surrogate for evolutionary distance to identify the closest gene(s), Wall et al. proposed to use direct and proper maximum likelihood estimates of the evolutionary distance between pairs of sequences [31]. This estimate of evolutionary distance is based on the number and type of amino acid substitutions between the two sequences. Indeed, previous studies have shown that the highest scoring alignment is often not the nearest phylogenetic neighbor [35]. Building upon this work, Roth et al. showed how statistical uncertainties in the distance estimation can be incorporated into the inference strategy [36].

Differential Gene Losses

As discussed above, one of the advantages of BBH over BeT is that by virtue of the bi-directional requirement, the former is more robust to gene losses in one of the two lineages. But if gene losses occurred along both lineages, it can happen that a pair of genes mutually closest to one another is in fact paralogs, simply because both their corresponding orthologs were lost—a situation referred to as "differential gene losses." Dessimoz et al. [37] presented a way to detect some of these cases by looking for a third species in which the corresponding orthologs have not been lost and thus can act as *witnesses of non-orthology*.

2.1.2 Clustering Phase: From Pairs to Groups

The graph construction phase yields orthologous relationships between pairs of genes. But this is often not sufficient. Conceptually, information obtained from multiple genes or organisms is often more powerful than that obtained from pairwise comparisons

only. In particular, as the use of a third genome as potential witness of non-orthology suggests, a more global view can allow identification and correction of inconsistent/spurious predictions. Practically, it is more intuitive and convenient to work with groups of genes than with a list of gene pairs. Therefore, it is often desirable to cluster orthologous genes into groups.

Tatusov et al. [8] introduced the concept of clusters of orthologous groups (COGs). COGs are computed by using triangles (triplets of genes connected to each other) as seeds and then merging triangles which share a common face, until no more triangle can be added. This clustering can be computed relatively efficient in time $O(n^3)$, where n is the number of genomes analyzed [38]. The stated objective of this clustering procedure is to group genes that have diverged from a single gene in the last common ancestor of the species represented [8]. Practically, they have been found to be useful by many, most notably to categorize prokaryotic genes into broad functional categories.

A different clustering approach was adopted by *OrthoMCL*, another well-established graph-based orthology inference method [29]. There, groups of orthologs are identified by Markov Clustering [39]. In essence, the method consists in simulating a random walk on the orthology graph, where the edges are weighted according to similarity scores. The Markov Clustering process gives rise to probabilities that two genes belong to the same cluster. The graph is then partitioned according to these probabilities and members of each partition form an orthologous group. These groups contain orthologs and "recent" paralogous genes, where the recency of the paralogs can be somewhat controlled through the parameters of the clustering process.

A third grouping strategy consists in building groups by identifying fully connected subgraphs (called "cliques" in graph theory) [23]. This approach has the merits of straightforward interpretation (groups of genes which are all orthologous to one another) and high confidence in terms of orthology within the resulting groups, due to the high consistency required to form a fully connected subgraph. But it has the drawbacks of being hard to compute (clique finding belongs to the NP-complete class of problems, for which no polynomial time algorithm is known; see Box 2) and being excessively conservative for many applications.

As emerges from these various strategies, there is more than one way orthologous groups can be defined, each with different implications in terms of group properties and applications [40]. In fact, there is an inherent trade-off in partitioning the orthology graph into clusters of genes, because orthology is a non-transitive relation: if genes A and B are orthologs and genes B and C are orthologs, genes A and C are not necessarily orthologs, e.g., consider in Fig. 1 the blue human gene, the frog gene, and the red dog

gene. Therefore, if groups are defined as sets of genes in which all pairs of genes are orthologs (as with OMA groups), it is not possible to partition A, B, and C into groups capturing all orthologous relations while leaving out all paralogous relations.

2.1.3 Hierarchical Clustering

More inclusive grouping strategies necessarily lead to orthologs and paralogs within the same group. Nevertheless, it can be possible to control the nature of the paralogs included. For instance, as seen above, OrthoMCL attempts at including only "recent" paralogs in its groups. This idea can be specified more precisely by defining groups with respect to a particular speciation event of interest, e.g., the base of the mammals. Such *hierarchical groups* are expected to include orthologs and in-paralogs with respect to the reference speciation—in our example all copies that have descended from a single common ancestor gene in the last mammalian common ancestor. Conceptually, hierarchical orthologous groups can be defined as groups of genes that have descended from a single common ancestral gene within a taxonomic range of interest.

Several resources provide hierarchical clustering of orthologous groups. EggNOG [15] and OrthoDB [25], for example, both implement this concept by applying a COG-like clustering method for various taxonomic ranges. Another example, Hieranoid, produces hierarchical groups by using a guide tree to perform pairwise orthology inferences at each node from the leaves to the root—inferring ancestral genomes at each node in the tree [13, 18]. Similarly, OMA GETHOGs is an approach based on an orthology graph of pairwise orthologous gene relations, where hierarchical orthologous groups are formed starting with the most specific taxonomy and incrementally merges them toward the root [21, 22]. Another method, COCO-CL, identifies hierarchical orthologous groups recursively, using correlations of similarity scores among homologous genes [41] and, interestingly, without relying on a species tree. By capturing part of the gene tree structure in the group hierarchies, these methods try in some way to bridge the gap between graph-based and tree-based orthology inference approaches. We now turn our attention to the latter.

2.2 Tree-Based Methods

At their core, tree-based methods infer orthologs on the basis of gene family trees whose internal nodes are labeled as speciation or duplication nodes. Indeed, once all nodes of the gene tree have been inferred as a speciation or duplication event, it is trivial to establish whether a pair of genes is orthologous or paralogous, based on the type of the branching where they coalesce. Such labeling is traditionally obtained by reconciling gene and species trees. In most cases, gene and species trees have different topologies, due to evolutionary events acting specifically on genes such as duplications, losses, lateral transfers, or incomplete lineage sorting [42]. Goodman et al. [43] pioneered research to resolve these

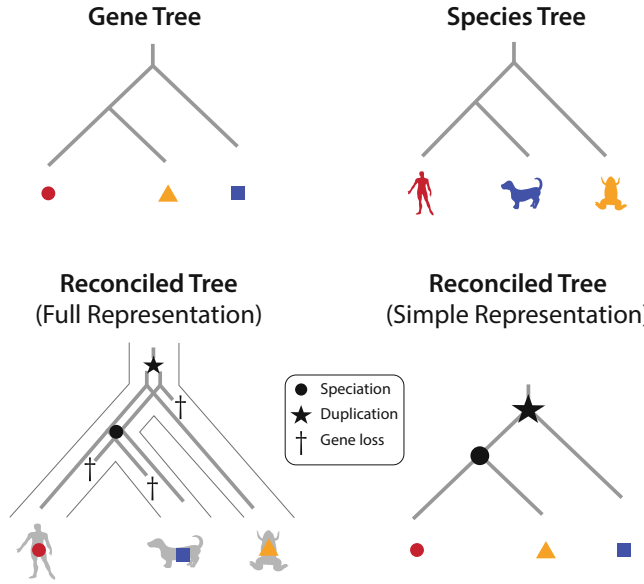


Fig. 2 Schematic example of the gene/species tree reconciliation. The gene tree and species tree are not compatible. Reconciliation methods resolve the incongruence between the two by inferring speciation, duplication, and losses events on the gene tree. The reconciled tree indicates the most parsimonious history of this gene, constrained to the species tree. The simple representation (bottom right) suggests that the human and frog genes are orthologs and that they are both paralogous to the dog gene

incongruences. They showed how the incongruences can be explained in terms of speciation, duplication, and loss events on the gene tree (Fig. 2) and provided an algorithm to infer such events.

Most tree reconciliation methods rely on a parsimony criterion: the most likely reconciliation is the one which requires the least number of gene duplications and losses. This makes it possible to compute reconciliation efficiently and is tenable as long as duplication and loss events are rare compared to speciation events. In their seminal article, Goodman et al. [43] had already devised their reconciliation algorithm under a parsimony strategy. In the subsequent years, the problem was formalized in terms of a map function between the gene and species trees [44], whose computational cost was conjectured [45], and later proved [12, 46] to coincide with the number of gene duplication and losses. These results yielded highly efficient algorithms, either in terms of asymptotic time complexity [12] or in terms of runtimes on typical problem sizes [47]. With these near-optimal solutions, one might think that the tree reconciliation problem has long been solved. As we shall see in the rest of this section, however, the original formulation of the tree reconciliation problem has several limitations in practice, which have stimulated the development of various refinements to overcome them (Table 2).

Table 2
Overview of gene/species tree reconciliation methods and their main properties

Method	Species tree[a]	Rooting[b]	Gene tree uncertainty[c]	Framework[d]	Available Algo/DB	References
BranchClust	Species overlap	Min number of clusters	None	n.a.	–/✓	[48]
DLRSOrthology	Fully resolved	n.a.	n.a.	Probabilistic	✓/–	[49–51]
Ensembl/TreeBeST	Partially resolved	Min dupl + min loss	None	MP	–/✓	[52–54]
HOGENOM	Partially resolved	Min dupl	Multifurcate	MP	✓/✓	[55, 56]
LOFT	Species overlap	Min dupl	None	MP	✓/–	[57]
Orthostrapper	Fully resolved	Min dupl	Bootstrap	MP	✓/–	[58]
PhylomeDB	Species overlap	Outgroup	None	MP	–/✓	[59, 60]
Softparsmap	Partially resolved	Min dupl + min loss	None	MP	✓/–	[61]
Speciation vs. duplication inference (SDI)	Fully resolved	n.a.	None	MP	✓/–	[47]

[a]Required species tree: Fully resolved, multifurcations allowed, computed from species overlap
[b]Approach to root gene tree (n.a. indicates that the initial rooting is assumed to be correct)
[c]Approach taken to handle reconstruction uncertainties of the gene tree (bootstrap, reconcile every bootstrap sample; multifurcate, splits in the gene tree with low support are collapsed)
[d]Used optimization framework (MP, maximum parsimony)

2.2.1 Unresolved Species Tree

A first problem ignored by most early reconciliation algorithms lies in the uncertainty often associated with the species tree, which these methods assume as correct and heavily rely upon.

One way of dealing with the uncertainties is to treat unresolved parts of the species tree as multifurcating nodes (also known as *soft polytomies*). By doing so, the reconciliation algorithm is not forced to choose for a specific type of evolutionary event in ambiguous regions of the tree. This approach is, for instance, implemented in *TreeBeST* [52] and used in the *Ensembl Compara* project [53].

Alternatively, Heijden et al. [57] demonstrated that it is often possible to infer speciation and duplication events on a gene tree without knowledge of the species tree. Their approach, which they call *species overlap*, identifies for a given split the species represented in the two subtrees induced by the split. If at least one species has genes in both subtrees, a duplication event is inferred; else a speciation event is inferred. In fact, this approach is a special case of soft polytomies where all internal nodes have been collapsed. Thus, the only information needed for this approach is a rooted gene tree. Since then, this approach has been adopted in other projects, such as PhylomeDB [59].

2.2.2 Rooting

The classical reconciliation formulation requires both gene and species trees to be rooted. But most models of sequence evolution are time reversible and thus do not allow to infer the rooting of the reconstructed gene tree. One sensible solution is to root a gene tree so that it minimizes the number of duplication events [62]. Thus, this method uses the parsimony principle for both rooting and reconciliation. For cases of multiple optimal rootings, ties can be broken by selecting the tree that minimizes the tree height [63] or by picking the rooting which minimizes the number of gene losses [61].

Another approach is to place the root at the "center of the tree"—also known as "midpoint rooting" [58]. The idea of this method goes back to Farris [64] and is motivated by the concept of a molecular clock. But for most gene families, assuming a constant rate of evolution is inappropriate [65, 66], and thus this approach is not used widely. A newly introduced refinement based on minimizing average deviations among children nodes holds promise of being more robust [67] but still relies on a molecular clock assumption.

For the species tree, the most common and reliable way of rooting trees is by identifying an outgroup species. PhylomeDB uses genes from outgroup species to root gene trees [59]. One main potential problem with this approach is that in many situations, it can be difficult to identify a suitable outgroup. For example, in analysis covering all kingdoms of life, an outgroup species may not be available, or the relevant genes might have been lost

[68]. A suitable out-group needs to be close enough to allow for reliable sequence alignment, yet it must have speciated clearly before any other species separated. Furthermore, ancient duplications can cause outgroup species to carry *in-group* genes. These difficulties make this approach more challenging for automated, large-scale analysis [69].

2.2.3 Gene Tree Uncertainty

Another assumption made in the original tree reconciliation problem is the (topological) correctness of the gene tree. But it has been shown that this assumption is commonly violated, often due to finite sequence lengths, taxon sampling [70, 71], or gene evolution model violations [72]. On the other hand, techniques of expressing uncertainties in gene tree reconstruction via support measures, e.g., bootstrap values, have become well established. Storm and Sonnhammer [58] as well as Zmasek and Eddy [63] independently suggested to extend the bootstrap procedure to reconciliation, thereby reducing the dependency of the reconciliation procedure on any one gene tree while providing a measure of support of the inferred speciation/duplication events. The downsides of using the bootstrap are the high computational costs and interpretation difficulties associated with it [73].

Similarly to how unresolved species tree can be handled, unresolved parts of the gene tree can also be collapsed into multifurcating nodes. For instance, HOGENOM [55] and *Softparsmap* [61] collapse branches with low bootstrap support values.

A third way of tackling this problem consists in simultaneously solving both the gene tree reconstruction and reconciliation problems [74]. They use the parsimony criterion of minimizing the number of duplication events to improve on the gene tree itself. This is achieved by rearranging the local gene tree topology of regions with low bootstrap support such that the number of duplications and losses is further reduced.

2.2.4 Parsimony vs. Likelihood

All the approaches mentioned so far try to minimize the number of gene duplication events. This is generally justified by a parsimony argument, which assumes that gene duplications and losses are rare events. But what if this assumption is frequently violated? Little is known about duplication and loss rates in general [75], but there is strong evidence for historical periods with high gene duplication occurrence rates [76] or gene families specifically prone to massive duplications (e.g., olfactory receptor, opsins, serine/threonine kinases, etc.)

Motivated by this reasoning, Arvestad et al. introduced the idea of a probabilistic model for tree reconciliation [49]. They used a Bayesian approach to estimate the posterior probabilities of a reconciliation between a given gene and species tree using Markov chain Monte Carlo (MCMC) techniques. Arvestad et al. [49]

modeled gene duplication and loss events through a *birth-death process* [77]. In the subsequent years, they refined their method to also model sequence evolution and substitution rates in a unified framework called *gene sequence evolution model with iid rates* (*GSR*) [49, 50].

Perhaps the biggest problem with the probabilistic approach is that it is not clear how well the assumptions of their model (the *birth-death process* with fixed parameters) relate to the true process of gene duplication and gene loss. Doyon et al. [78] compared the maximum parsimony reconciliation trees from 1278 fungi gene families to the probabilistically reconciled trees using gene birth/death rates fitted from the data. They found that in all but two cases, the maximum parsimony scenario corresponds to the most probable one. This remarkably high level of consistency indicates that in terms of the accuracy of the "best" reconciliation, there is little to gain from using a likelihood approach over the parsimony criterion of minimizing the number of duplication events. But how this result generalizes to other datasets has yet to be investigated.

2.3 Graph-Based vs. Tree-Based: Which Is Better?

Given the two fundamentally different paradigms in orthology inference that we reviewed in this section, one can wonder which is better. Conceptually, tree reconciliation methods have several advantages. In terms of inference, by considering all sequences from all species at the same time, it can also be expected that they can extract more information from the sequences. This in turn should translate into higher statistical power. In terms of their output, reconciled gene trees provide the user more information than pairs or groups of orthologs. For example, the trees display the order of duplication and speciation events, as well as evolutionary distances between these events. In practice, however, these methods have the disadvantage of having much higher computational complexity than their graph-based counterparts. Furthermore, the two approaches are in practice often not that strictly separated. Tree-based methods often start with a graph-based clustering step to identify families of homologous genes. Conversely, several hierarchical grouping algorithms also rely on species trees in their inference.

Thus, it is difficult to make general statements about the relative performance of the two classes of inference methods. One solution that can leverage the unique abilities of both tree-based and graph-based methods is to combine several independent orthology inference methods into one. We discuss this technique in the next section.

3 Meta-methods

In recent years a new class of orthology inference tools has emerged which attempts to make the most out of multiple orthology prediction algorithms—*meta-methods*. These are approaches which combine several individual and distinct methods in order to produce more robust orthology predictions. These meta-methods are able to take advantage of the standardized formats of output which has been a goal of the orthology community [79], as well as the many new and well-established methods out there.

Generally, meta-methods assign a confidence score to a given predicted orthologous relation. In its most basic form, more weight is given to orthologs predicted by the most methods. Some examples include methods which simply take the intersection of several methods, such as GET_HOMOLOGUES [80], COMPARE [81], HCOP [82], and DIOPT [83]. These methods maintain a high level of precision, but since they are based on intersections, they necessarily have a lower recall.

Additionally, post-processing techniques can be used to build upon the base of orthologs found by several methods—thus assigning more sequences as orthologs and improving performance. For example, MOSAIC (Multiple Orthologous Sequence Analysis and Integration by Cluster optimization) [84] uses an iterative graph-based optimization approach that works on ortholog sets predicted by several independent methods. MOSAIC captures orthologs which are missed by some individual methods, producing a 1.6-fold increase in the number of orthologs detected. Another example is the MARIO software, which looks for the intersection of several different orthology methods as seed groups and then progressively adds unassigned proteins to the groups based on HMM profiles [85]. MetaPhOrs' approach integrates phylogenetic and homology information derived from different databases [86]. They demonstrate that the number of independent sources from which an orthology prediction is made, as well as the level of consistency across predictions, can be used as confidence scores.

So far the previously mentioned meta-methods combine independent orthology prediction algorithms and give a higher score based on the more algorithms which predict a given orthologous relation. However, another emerging approach is to use machine learning techniques to recognize patterns among several different orthology inference methods. With this, one can predict previously unknown high-confidence orthologs. WORMHOLE is a tool which uses the information from 17 different orthology prediction methods to train support vector machine classifiers for predicting least diverged orthologs [87]. WORMHOLE was able to strongly re-predict least diverged orthologs in the reference set and also predict previously unclassified orthologous genes.

The type of meta-approach and its associated stringency depends on what the user is going after. For example, if the goal is to get very-high-confidence groups, methods which only combine for the intersection without trying to add more orthologs may be preferable. Studies requiring both high precision and recall may be better suited to use the meta-methods which use post-processing or machine learning to predict orthologs. And as with all methods, it is important to understand which clades the method has been benchmarked in and which orthology tools have been combined. For example, if several methods have the same bias, one will just propagate the bias and end up with a false sense of security because the methods are not independent.

4 Scaling to Many Genomes

In terms of orthology inference, the abundance of genomes now available has resulted in an emphasis on driving down computational processing time via efficient algorithms. When inferring orthology for many genomes, the bottleneck is generally the all-against-all computations—aligning the proteins in every genome against the proteins in every other genome. This is the first step of nearly all graph-based methods. The all-against-all computation has an $O(n^2)$ runtime, meaning it scales quadratically with the number of genomes analyzed (Box 2).

So far, two main techniques for scaling orthology prediction to many genomes have emerged. The first approach is by making the all-against-all comparisons faster. Because comparisons are independent of each other, the most obvious way of doing this is by taking advantage of a high-performance computing cluster, as this is an embarrassingly parallel computing problem. Many methods have implemented this, such as Hieranoid [13], PorthoMCL [88], or OMA [22]. Another way to save time on the all-against-all comparisons is by using very fast algorithms for the homology search. For example, preliminary results of SonicParanoid showed 160–750× speedup of orthology inference compared to InParanoid [89]. Innovations in alignment algorithms with methods such as DIAMOND [90] or MMSeq2 [91] have the potential to greatly reduce the time to do the all-against-all comparisons.

A second approach to efficiently scale up orthology inference to many genomes is by simply avoiding doing the entire all-against-all comparisons. This makes sense, since a significant amount of time is spent comparing unrelated gene pairs. For example, it is possible to avoid aligning many unrelated pairs by exploiting the transitive property of homology. Wittwer et al. [92] did this by first building clusters of homologous sequences with one representative sequence per cluster and subsequently performing the all-against-all within each cluster. Hieranoid avoids unnecessary all-against-all

comparisons by using a species tree as a guide, reducing the number of comparisons to $N-1$ for N genomes, scaling linearly rather than quadratically [18]. Another way to avoid all-by-all comparison is by using a mapping strategy, whereby new proteomes are mapped onto precomputed orthologous groups. This strategy has been successfully implemented with the eggNOG database—each sequence in a new proteome is mapped to a precomputed orthologous cluster based on hidden Markov models. Then, orthology relations and function are transferred to the new sequence from the best matching sequence in the database [93].

5 Benchmarking Orthology

Assessing the quality of orthology predictions is important but difficult. The main challenge is that the precise evolutionary history of entire genomes is largely unknown and thus, predictions can only be validated indirectly, using surrogate measures. To be informative, such measures need to strongly correlate with orthology/paralogy. At the same time, they should be independent from the methods used in the orthology inference process. Concretely, this means that the orthology inference is not based on the surrogate measure and the surrogate measure is not derived from orthology/paralogy.

5.1 Benchmarking Approaches

Several ways of benchmarking orthology inference have been developed in the past years. In the next sections, we go over the main approaches, bringing attention to the advantages and limitations to each.

5.1.1 Functional Conservation

The first surrogate measures proposed revolved around conservation of function [94]. This was motivated by the common belief that orthologs tend to have conserved function, while paralogs tend to have different functions. Indeed, orthologs tend to be more conserved than paralogs in terms of GO annotation similarity [95]. Thus, "for a given evolutionary distance, more accurate orthology inference is likely to be correlated with more functionally similar gene pairs." Hulsen et al. [94] assessed the quality of ortholog predictions in terms of conservation of co-expression levels, domain annotation, and protein-protein interaction partners. Additionally, Altenhoff et al. [96] used similarity of experimentally validated GO annotations as well as Enzyme Commission (EC) numbers as a functional benchmark. Functional benchmarks have an advantage in that many researchers are interested in orthology because they want to find functionally conserved genes, thus making functional tests important for assessing different inference methods. The main limitation of these measures is that it is not so clear how much they correlate with orthology/paralogy. Indeed, it

has been argued that the difference in function conservation trends between orthologs and paralogs might be much smaller than commonly assumed and indeed many examples are known of orthologs that have dramatically different functions [97].

5.1.2 Gene Neighborhood Conservation

The fraction of orthologs that have neighboring genes being orthologs themselves is an indicator of consistency and therefore to some extent also of quality of orthology predictions [94]. Although synteny has been used as part of the orthology inference for several algorithms, to date it has not been used as part of large-scale benchmarking efforts. One possible problem is that gene neighborhood can be conserved among paralogs, such as those resulting from whole-genome duplications. Furthermore, some methods use gene neighborhood conservation to help in their inference process, which can bias the assessment done on such measures (principle of independence stated above).

5.1.3 Species Tree Discordance Test

The quality of ortholog predictions can also be assessed based on phylogeny. By definition, the tree relating a set of genes all orthologous to one another only contains speciation splits and has the same topology as the underlying species. We introduced a benchmarking protocol that quantifies how well the predictions from various orthology inference methods agree with undisputed species tree topologies [96, 98]. Thus, the species tree discordance test judges the accuracy of ortholog predictions based on the correctness of the species tree which can be constructed from them. The advantage of this measure is that by virtue of directly ensuing from the definition of orthology, it correlates strongly with it and thus satisfies the first principle. However, the second principle, independence from the inference process, is not satisfied with methods relying on the species tree—typically all reconciliation methods but also most graph-based methods producing hierarchical groups. In such cases, interpretation of the results must be done carefully.

5.1.4 Gold Standard Gene Tree Test

High-quality reference gene trees can also be used to assess orthology inferences. For this, one compares the pairs of orthologs from a given method to pairs of orthologs derived from these expertly curated gene trees [40, 99]. One drawback of this benchmark is that it is limited by the ability to curate the phylogeny—if the evolutionary history of the gene family is ambiguous, the resulting reference tree will unavoidably have mistakes. Another limitation is the small size of most benchmarks of this type. This casts doubts on their generalizability and makes them prone to overfitting.

5.1.5 Subtree Consistency Test

For inference methods based on reconciliation between gene and species trees, Vilella et al. [53] proposed a different phylogeny-based assessment scheme. For any duplication node of the labeled gene tree, a consistency score is computed, which captures the balance of the species found in the two subtrees. Unbalanced nodes correspond to an evolutionary scenario involving extensive gene losses and therefore, under the principle of parsimony, are less likely to be correct. Given that studies to date tend to support the adequacy of the parsimony criterion in the context of gene family dynamics (Subheading 2.2.4), it can be expected that this metric correlates highly with correct orthology/paralogy assignments. However, since virtually all tree-based methods themselves incorporate this very criterion in their objective function (i.e., minimizing the number of gene duplications and losses), the principle of independence is violated, and thus the adequacy of this measure is questionable.

5.1.6 Latent Class Analysis

Chen et al. [100] proposed a purely statistical benchmark based on *latent class analysis* (*LCA*). Given the absence of a definitive answer on whether two given genes are orthologs, the authors argue that by looking at the agreement and disagreement of predictions made by several inference methods on a common dataset, one can estimate the reliability of individual predictors. More precisely, LCA is a statistical technique that computes maximum likelihood estimates of sensitivity and specificity rates for each orthology inference methods, given their predictions and given an error model. This is attractive, because it does not depend on any surrogate measure. However, the results depend on the error model assumed. Thus, we are of the opinion that LCA merely shifts the problem of assessing orthology to the problem of assessing an error model of various orthology inference methods.

5.1.7 Simulated Genomes

Finally, simulated data can be used in benchmarking. By this, the precise evolutionary history of a genome can be validated, in terms of gene duplication, insertion, deletion, and lateral gene transfer [101]. Knowing for certain all aspects of the simulated genomes gives an advantage over assessments based on empirical data, where the true evolutionary history is unknown. On the other hand, how well the simulated data reflect "real" data is debatable.

5.2 Orthology Benchmarking Service

The orthology benchmarking service is a web-based platform for which users can upload their ortholog predictions and run them through a variety of benchmarks. The user must use *quest for orthologs* (QFO) reference proteome set, which is a set of 66 genomes that covers a diverse set of species across all domains [79], to infer pairwise or groups of orthologs. Several phylogenetic and function-based benchmarks are automatically run on the uploaded data, and then summary statistics of the results of each benchmark

are reported. The user can compare their method's performance with that of other well-known orthology inference algorithms and choose to make theirs public as well. For each benchmark, a precision-recall curve is reported, allowing for ease of comparison and evaluation of individual inference techniques. Because of the range of benchmarking tests and publicly available methods for comparison, the benchmarking service is useful for both users, who can check which methods work well for their particular problem and for method developers. The orthology benchmarking service can be accessed at http://orthology.benchmarkservice.org.

5.3 Conclusions on Benchmarking

Overall, it becomes apparent that there is no "magic bullet" strategy for orthology benchmarking, as each approach discussed here has its limitations (though some limitations are more serious than others). Nevertheless, comparative studies based on these various benchmarking measures have reported surprisingly consistent findings [40, 94, 96, 98, 100]: these assessments generally observe that there is a trade-off between accuracy and coverage and most common databases are situated on a Pareto frontier. The various assessments concur that the "best" orthology approach is highly dependent on the various possible applications of orthology.

6 Applications

As we have seen so far, there is a large diversity in the methods for orthology inference. The main reason is that, although the methods discussed here all infer orthology as part of their process, many of them have been developed for different reasons and have different ultimate goals. Unfortunately, this is often not mentioned explicitly and tends to be a source of confusion. In this section, we review some of these ultimate goals and discuss which methods and representation of orthology are better suited to address them and why.

As mentioned in the introduction, most interest for orthology is in the context of function prediction and is largely based on the belief that orthologs tend to have conserved function. A conservative approach consists in propagating function between one-to-one orthologs, i.e., pairs of orthologous genes that have not undergone gene duplication since they diverged from one another. Several orthology databases directly provide one-to-one orthology predictions. But even with those that do not, it might still be possible to obtain such predictions, for instance, by selecting hierarchical groups containing at most one sequence in each species or by extracting from reconciled trees' subtrees with no duplication. A more sophisticated approach consists in propagating gene function annotations across genomes on the basis of the full reconciled gene tree. Thomas et al. [102], for instance, proposed a way to assign

gene function to uncharacterized proteins using a gene tree and a hidden Markov model (HMM) among gene families. Engelhardt et al.. [103] developed a Bayesian model of function change along reconciled gene trees and showed that their approach significantly improves upon several methods based on pairwise gene function propagation. Ensembl Compara [53] and Panther [102] are two major databases providing reconciled gene trees.

Since Darwin, one traditional question in biology has always been how species are related to each other. As we recall in the introduction of this chapter, Fitch's original motivation for defining orthology was phylogenetic inference. Indeed, the gene tree reconstructed from a set of genes which are all orthologous to each other should by definition be congruent to the species tree. OMA Groups (OMA) have this characteristic and, crucially, are constructed without help of a species tree.

Yet another application associated with orthology are general alignments between genomes, e.g., protein-protein interaction (PPI) network alignments or whole-genome alignments. Finding an optimal PPI network alignment between two genomes on the basis of the network topology alone is a computationally hard problem (i.e., it is an instance of the subgraph isomorphism problem which is NP-complete [104]). Orthology is often used as heuristic to constrain the mapping of the corresponding genes between the two networks and thus to reduce the problem complexity of aligning networks [105]. For whole-genome alignments, people most often use homologous regions and use orthologs as anchor points [106]. These types of application typically rely on ortholog predictions between pairs of genomes, as provided, e.g., by InParanoid [5] or OMA [23].

7 Conclusions and Outlook

The distinction between orthologs and paralogs is at the heart of many comparative genomic studies and applications. The original and generally accepted definition of orthology is based on the evolutionary history of pairs of genes. By contrast, there is a considerable diversity in how groups of orthologs are defined. These differences largely stem from the fact that orthology is a non-transitive relation and therefore, dividing genes into orthologous groups will either miss or wrongly include orthologous relations. This makes it important and worthwhile to identify the type of orthologous group best suited for a given application.

Regarding inference methods, while most approaches can be ordered into two fundamental paradigms—graph-based and tree-based—the difference between the two is shrinking, with graph-based methods increasingly striving to capture more of the evolutionary history. On the other hand, the rapid pace at which new

genomes are sequenced limits the applicability of tree-based methods, computationally more demanding.

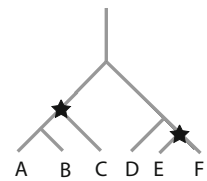
Benchmarking this large variety of methods remains a hard problem—from a conceptual point as described above but also because of very practical challenges such as heterogeneous data formats, genome versions, or gene identifiers. This has been recognized by the research community and has led to the development of the QFO consortium benchmarking service [96].

Looking forward, we see potential in extending the current model of gene evolution, which is limited to speciation, duplication, and loss events. Indeed, nature is often much more complicated. For instance, lateral gene transfer (LGT) is believed to be a major mode of evolution in prokaryotes. While there has been several attempts at extending tree reconciliation algorithms to detecting LGT [107, 108], the problem is largely unaddressed in typical orthology resources [109]. Another relevant evolutionary process omitted by most methods is whole-genome duplications (WGD). Even though WGD events act jointly on all gene families, with few exceptions [110, 111], most methods consider each gene family independently.

Overall, the orthology/paralogy dichotomy has proved to be useful but also inherently limited. Reducing the whole evolutionary history of homologous genes into binary pairwise relations is bound to be a simplification—and at times an oversimplification. The shift toward hierarchical orthologous groups is thus a promising step toward capturing more features of the evolutionary history of genes. Yet further development will still be needed, as we are nowhere close to grasp the formidable complexity of gene evolution across the full diversity of life.

8 Exercises

Assume the following evolutionary scenario



where duplications are depicted as star and all other splits are speciations.

Problem #1: Draw the corresponding orthology graph, where the vertices correspond to the observed genes and the edges indicate orthologous relations between them.

Problem #2: Apply the following two clustering methods on your orthology graph. First, reconstruct all the maximal fully

connected subgraphs (cliques) that can be found. Second, reconstruct the COGs. COGs are built by merging triangles of orthologs whenever they share a common face. Remember that in both methods, a gene can only belong to a one cluster.

Acknowledgments

We thank Stefan Zoller for helpful feedback on an earlier version of the manuscript. We gratefully acknowledge support by the Swiss National Science Foundation grant PP00P3_150654 to CD. Adrian M. Altenhoff and Natasha M. Glover contributed equally to this work.

References

1. Dewey CN (2012) Whole-genome alignment. Methods Mol Biol 855:237–257
2. Alioto T (2012) Gene prediction. In: Anisimova M (ed) Evolutionary genomics: statistical and computational methods, vol 1. Humana, Totowa, NJ, pp 175–201
3. Löytynoja A (2012) Alignment methods: strategies, challenges, benchmarking, and comparative overview. In: Anisimova M (ed) Evolutionary genomics: statistical and computational methods, vol 1. Humana, Totowa, NJ, pp 203–235
4. Fitch WM (1970) Distinguishing homologous from analogous proteins. Syst Zool 19:99–113
5. Remm M, Storm CEV, Sonnhammer ELL (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol 314:1041–1052
6. Glover NM, Redestig H, Dessimoz C (2016) Homoeologs: what are they and how do we infer them? Trends Plant Sci 21:609–621
7. Kuzniar A, van Ham RCHJ, Pongor S et al (2008) The quest for orthologs: finding the corresponding gene across genomes. Trends Genet 24:539–551
8. Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278:631–637
9. Overbeek R, Fonstein M, D'Souza M et al (1999) The use of gene clusters to infer functional coupling. Proc Natl Acad Sci U S A 96:2896–2901
10. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147:195–197
11. Altschul SF, Madden TL, Schäffer AA et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402
12. Zhang L (1997) On a Mirkin-Muchnik-Smith conjecture for comparing molecular phylogenies. J Comput Biol 4:177–187
13. Schreiber F, Sonnhammer ELL (2013) Hieranoid: hierarchical orthology inference. J Mol Biol 425:2072–2081
14. Chor B, Tuller T (2005) Maximum likelihood of evolutionary trees is hard. In: Proceedings of the 9th annual international conference on research in computational molecular biology. Springer, Berlin, pp 296–310
15. Jensen LJ, Julien P, Kuhn M et al (2008) eggNOG: automated construction and annotation of orthologous groups of genes. Nucleic Acids Res 36:D250–D254
16. Muller J, Szklarczyk D, Julien P et al (2010) eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. Nucleic Acids Res 38:D190–D195
17. Huerta-Cepas J, Szklarczyk D, Forslund K et al (2016) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res 44: D286–D293
18. Kaduk M, Sonnhammer E (2017) Improved orthology inference with Hieranoid 2. Bioinformatics 33:1154–1159
19. Ostlund G, Schmitt T, Forslund K et al (2010) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res 38:D196–D203

20. Sonnhammer ELL, Östlund G (2015) InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. Nucleic Acids Res 43:D234–D239
21. Altenhoff AM, Gil M, Gonnet GH et al (2013) Inferring hierarchical orthologous groups from orthologous gene pairs. PLoS One 8:e53786
22. Train C-M, Glover NM, Gonnet GH et al (2017) Orthologous Matrix (OMA) algorithm 2.0: more robust to asymmetric evolutionary rates and more scalable hierarchical orthologous group inference. Bioinformatics 33:i75–i82
23. Dessimoz C, Cannarozzi G, Gil M et al (2005) OMA, a comprehensive, automated project for the identification of orthologs from complete genome data: introduction and first achievements. In: Comparative genomics. Springer, Berlin, pp 61–72
24. Altenhoff AM, Schneider A, Gonnet GH et al (2011) OMA 2011: orthology inference among 1000 complete genomes. Nucleic Acids Res 39:D289–D294
25. Kriventseva EV, Rahman N, Espinosa O et al (2008) OrthoDB: the hierarchical catalog of eukaryotic orthologs. Nucleic Acids Res 36: D271–D275
26. Zdobnov EM, Tegenfeldt F, Kuznetsov D et al (2017) OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. Nucleic Acids Res 45:D744–D749
27. Linard B, Thompson JD, Poch O et al (2011) OrthoInspector: comprehensive orthology analysis and visual exploration. BMC Bioinform 12:11
28. Linard B, Allot A, Schneider R et al (2015) OrthoInspector 2.0: software and database updates. Bioinformatics 31:447–448
29. Li L, Stoeckert CJ Jr, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13:2178–2189
30. Chen F, Mackey AJ, Stoeckert CJ Jr et al (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res 34:D363–D368
31. Wall DP, Fraser HB, Hirsh AE (2003) Detecting putative orthologs. Bioinformatics 19:1710–1711
32. DeLuca TF, Wu I-H, Pu J et al (2006) Roundup: a multi-genome repository of orthologs and evolutionary distances. Bioinformatics 22:2044–2046
33. DeLuca TF, Cui J, Jung J-Y et al (2012) Roundup 2.0: enabling comparative genomics for over 1800 genomes. Bioinformatics 28:715–716
34. Fulton DL, Li YY, Laird MR et al (2006) Improving the specificity of high-throughput ortholog prediction. BMC Bioinform 7:270
35. Koski LB, Golding GB (2001) The closest BLAST hit is often not the nearest neighbor. J Mol Evol 52:540–542
36. Roth ACJ, Gonnet GH, Dessimoz C (2008) Algorithm of OMA for large-scale orthology inference. BMC Bioinform 9:518
37. Dessimoz C, Boeckmann B, Roth ACJ et al (2006) Detecting non-orthology in the COGs database and other approaches grouping orthologs using genome-specific best hits. Nucleic Acids Res 34:3309–3316
38. Kristensen DM, Kannan L, Coleman MK et al (2010) A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. Bioinformatics 26:1481–1487
39. Van Dongen SM (2001) Graph clustering by flow simulation. PhD thesis, University of Utrecht
40. Boeckmann B, Robinson-Rechavi M, Xenarios I et al (2011) Conceptual framework and pilot study to benchmark phylogenomic databases based on reference gene trees. Brief Bioinform 12:423–435
41. Jothi R, Zotenko E, Tasneem A et al (2006) COCO-CL: hierarchical clustering of homology relations based on evolutionary correlations. Bioinformatics 22:779–788
42. Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York
43. Goodman M, Czelusniak J, Moore GW et al (1979) Fitting the gene lineage into its species lineage, a Parsimony strategy illustrated by cladograms constructed from globin sequences. Syst Zool 28:132–163
44. Page RDM (1994) Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. Syst Biol 43:58–77
45. Mirkin B, Muchnik I, Smith TF (1995) A biologically consistent model for comparing molecular phylogenies. J Comput Biol 2:493–507
46. Eulenstein O (1997) A linear time algorithm for tree mapping. Arbeitspapiere der GMD No. 1046, St
47. Zmasek CM, Eddy SR (2001) A simple algorithm to infer gene duplication and speciation events on a gene tree. Bioinformatics 17:821–828

48. Poptsova MS, Gogarten JP (2007) BranchClust: a phylogenetic algorithm for selecting gene families. BMC Bioinform 8:120
49. Arvestad L, Berglund A-C, Lagergren J et al (2003) Bayesian gene/species tree reconciliation and orthology analysis using MCMC. Bioinformatics 19(Suppl 1):i7–i15
50. Åkerborg Ö, Sennblad B, Arvestad L et al (2009) Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. Proc Natl Acad Sci U S A 106:5714–5719
51. Ullah I, Sjöstrand J, Andersson P et al (2015) Integrating sequence evolution into probabilistic orthology analysis. Syst Biol 64:969–982
52. Li H, Coghlan A, Ruan J et al (2006) TreeFam: a curated database of phylogenetic trees of animal gene families. Nucleic Acids Res 34: D572–D580
53. Vilella AJ, Severin J, Ureta-Vidal A et al (2009) EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. Genome Res 19:327–335
54. Herrero J, Muffato M, Beal K et al (2016) Ensembl comparative genomics resources. Database 2016:bav096
55. Dufayard J-F, Duret L, Penel S et al (2005) Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. Bioinformatics 21:2596–2603
56. Penel S, Arigon A-M, Dufayard J-F et al (2009) Databases of homologous gene families for comparative genomics. BMC Bioinform 10(Suppl 6):S3
57. van der Heijden RTJM, Snel B, van Noort V et al (2007) Orthology prediction at scalable resolution by phylogenetic tree analysis. BMC Bioinform 8:83
58. Storm CEV, Sonnhammer ELL (2002) Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. Bioinformatics 18:92–99
59. Huerta-Cepas J, Dopazo H, Dopazo J et al (2007) The human phylome. Genome Biol 8: R109
60. Huerta-Cepas J, Capella-Gutiérrez S, Pryszcz LP et al (2014) PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. Nucleic Acids Res 42:D897–D902
61. Berglund-Sonnhammer A-C, Steffansson P, Betts MJ et al (2006) Optimal gene trees from sequences and species trees using a soft interpretation of parsimony. J Mol Evol 63:240–250
62. Hallett MT, Lagergren J (2000) New algorithms for the duplication-loss model. In: Proceedings of the fourth annual international conference on computational molecular biology. ACM, New York, NY, pp 138–146
63. Zmasek CM, Eddy SR (2002) RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. BMC Bioinform 3:14
64. Farris JS (1972) Estimating phylogenetic trees from distance matrices. Am Nat 106:645–668
65. Avise JC, Bowen BW, Lamb T et al (1992) Mitochondrial DNA evolution at a turtle's pace: evidence for low genetic variability and reduced microevolutionary rate in the Testudines. Mol Biol Evol 9:457–473
66. Ayala FJ (1999) Molecular clock mirages. Bioessays 21:71–75
67. Tria FDK, Landan G, Dagan T (2017) Phylogenetic rooting using minimal ancestor deviation. Nat Ecol Evol 1:193
68. Huelsenbeck JP, Bollback JP, Levine AM (2002) Inferring the root of a phylogenetic tree. Syst Biol 51:32–43
69. Tarrío R, Rodríguez-Trelles F, Ayala FJ (2000) Tree rooting with outgroups when they differ in their nucleotide composition from the ingroup: the Drosophila saltans and Willistoni groups, a case study. Mol Phylogenet Evol 16:344–349
70. Graybeal A (1998) Is it better to add taxa or characters to a difficult phylogenetic problem? Syst Biol 47(1):9–17
71. Rokas A, Williams BL, King N et al (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804
72. Yang Z, Goldman N, Friday A (1994) Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol Biol Evol 11:316–324
73. Anisimova M, Gascuel O (2006) Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. Syst Biol 55:539–552
74. Durand D, Halldórsson BV, Vernot B (2006) A hybrid micro-macroevolutionary approach to gene tree reconstruction. J Comput Biol 13:320–335
75. Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155
76. Robinson-Rechavi M, Marchand O, Escriva H et al (2001) Euteleost fish genomes are characterized by expansion of gene families. Genome Res 11:781–788

77. Kendall DG (1948) On the generalized "birth-and-death" process. Ann Math Stat 19:1–15
78. Doyon J-P, Hamel S, Chauve C (2012) An efficient method for exploring the space of gene tree/species tree reconciliations in a probabilistic framework. IEEE/ACM Trans Comput Biol Bioinform 9:26–39
79. Gabaldón T, Dessimoz C, Huxley-Jones J et al (2009) Joining forces in the quest for orthologs. Genome Biol 10:403
80. Contreras-Moreira B, Vinuesa P (2013) GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pan-genome analysis. Appl Environ Microbiol 79:7696–7701
81. Salgado D, Gimenez G, Coulier F et al (2008) COMPARE, a multi-organism system for cross-species data comparison and transfer of information. Bioinformatics 24:447–449
82. Eyre TA, Wright MW, Lush MJ et al (2007) HCOP: a searchable database of human orthology predictions. Brief Bioinform 8:2–5
83. Hu Y, Flockhart I, Vinayagam A et al (2011) An integrative approach to ortholog prediction for disease-focused and other functional studies. BMC Bioinform 12:357
84. Maher MC, Hernandez RD (2015) Rock, paper, scissors: harnessing complementarity in ortholog detection methods improves comparative genomic inference. G3 5:629–638
85. Pereira C, Denise A, Lespinet O (2014) A meta-approach for improving the prediction and the functional annotation of ortholog groups. BMC Genomics 15(Suppl 6):S16
86. Pryszcz LP, Huerta-Cepas J, Gabaldón T (2011) MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. Nucleic Acids Res 39:e32
87. Sutphin GL, Mahoney JM, Sheppard K et al (2016) WORMHOLE: novel least diverged ortholog prediction through machine learning. PLoS Comput Biol 12:e1005182
88. Tabari E, Su Z (2017) PorthoMCL: parallel orthology prediction using MCL for the realm of massive genome availability. Big Data Anal 2:4
89. Cosentino S, Iwasaki W (2018) SonicParanoid: extremely fast, accurate, and easy orthology inference. Bioinformatics. https://doi.org/10.1093/bioinformatics/bty631
90. Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. Nat Methods 12:59–60
91. Steinegger M, Söding J (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol 35(11):1026–1028
92. Wittwer LD, Piližota I, Altenhoff AM et al (2014) Speeding up all-against-all protein comparisons while maintaining sensitivity by considering subsequence-level homology. PeerJ 2:e607
93. Huerta-Cepas J, Forslund K, Coelho LP et al (2017) Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. Mol Biol Evol 34:2115–2122
94. Hulsen T, Huynen MA, de Vlieg J et al (2006) Benchmarking ortholog identification methods using functional genomics data. Genome Biol 7:R31
95. Altenhoff AM, Studer RA, Robinson-Rechavi M et al (2012) Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. PLoS Comput Biol 8:e1002514
96. Altenhoff AM, Boeckmann B, Capella-Gutierrez S et al (2016) Standardized benchmarking in the quest for orthologs. Nat Methods 13:425–430
97. Studer RA, Robinson-Rechavi M (2009) How confident can we be that orthologs are similar, but paralogs differ? Trends Genet 25:210–216
98. Altenhoff AM, Dessimoz C (2009) Phylogenetic and functional assessment of orthologs inference projects and methods. PLoS Comput Biol 5:e1000262
99. Trachana K, Larsson TA, Powell S et al (2011) Orthology prediction methods: a quality assessment using curated protein families. BioEssays 33:769–780
100. Chen F, Mackey AJ, Vermunt JK et al (2007) Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLoS One 2:e383
101. Dalquen DA, Altenhoff AM, Gonnet GH et al (2013) The impact of gene duplication, insertion, deletion, lateral gene transfer and sequencing error on orthology inference: a simulation study. PLoS One 8:e56925
102. Thomas PD, Campbell MJ, Kejariwal A et al (2003) PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 13:2129–2141
103. Engelhardt BE, Jordan MI, Muratore KE et al (2005) Protein molecular function prediction by Bayesian phylogenomics. PLoS Comput Biol 1:e45

104. Cook SA (1971) The complexity of theorem-proving procedures. In: Proceedings of the third annual ACM symposium on theory of computing. ACM, New York, NY, pp 151–158
105. Sharan R, Ideker T (2006) Modeling cellular machinery through biological network comparison. Nat Biotechnol 24:427–433
106. Dewey CN, Pachter L (2006) Evolution at the nucleotide level: the problem of multiple whole-genome alignment. Hum Mol Genet 15 Spec No 1:R51–RR6
107. Górecki P (2004) Reconciliation problems for duplication, loss and horizontal gene transfer. In: Proceedings of the eighth annual international conference on research in computational molecular biology. ACM, New York, NY, pp 316–325
108. Hallett M, Lagergren J, Tofigh A (2004) Simultaneous identification of duplications and lateral transfers. In: Proceedings of the eighth annual international conference on Research in computational molecular biology. ACM, New York, NY, pp 347–356
109. Forslund K, Pereira C, Capella-Gutierrez S et al (2017) Gearing up to handle the mosaic nature of life in the quest for orthologs. Bioinformatics. https://doi.org/10.1093/bioinformatics/btx542
110. Guigó R, Muchnik I, Smith TF (1996) Reconstruction of ancient molecular phylogeny. Mol Phylogenet Evol 6:189–213
111. Bansal MS and Eulenstein O (2008) The multiple gene duplication problem revisited. Bioinformatics 24:i132–i13i138

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Part III

Phylogenomics and Genome Evolution

Chapter 7

Modern Phylogenomics: Building Phylogenetic Trees Using the Multispecies Coalescent Model

Liang Liu, Christian Anderson, Dennis Pearl, and Scott V. Edwards

Abstract

The multispecies coalescent (MSC) model provides a compelling framework for building phylogenetic trees from multilocus DNA sequence data. The pure MSC is best thought of as a special case of so-called "multispecies network coalescent" models, in which gene flow is allowed among branches of the tree, whereas MSC methods assume there is no gene flow between diverging species. Early implementations of the MSC, such as "parsimony" or "democratic vote" approaches to combining information from multiple gene trees, as well as concatenation, in which DNA sequences from multiple gene trees are combined into a single "supergene," were quickly shown to be inconsistent in some regions of tree space, in so far as they converged on the incorrect species tree as more gene trees and sequence data were accumulated. The anomaly zone, a region of tree space in which the most frequent gene tree is different from the species tree, is one such region where many so-called "coalescent" methods are inconsistent. Second-generation implementations of the MSC employed Bayesian or likelihood models; these are consistent in all regions of gene tree space, but Bayesian methods in particular are incapable of handling the large phylogenomic data sets currently available. Two-step methods, such as MP-EST and ASTRAL, in which gene trees are first estimated and then combined to estimate an overarching species tree, are currently popular in part because they can handle large phylogenomic data sets. These methods are consistent in the anomaly zone but can sometimes provide inappropriate measures of tree support or apportion error and signal in the data inappropriately. MP-EST in particular employs a likelihood model which can be conveniently manipulated to perform statistical tests of competing species trees, incorporating the likelihood of the collected gene trees on each species tree in a likelihood ratio test. Such tests provide a useful alternative to the multilocus bootstrap, which only indirectly tests the appropriateness of competing species trees. We illustrate these tests and implementations of the MSC with examples and suggest that MSC methods are a useful class of models effectively using information from multiple loci to build phylogenetic trees.

Key words Introgression, Hybridization, Coalescent, Recombination, Neutrality, Molecular evolution

1 Introduction

The concept of a phylogeny or "species tree," a bifurcating dendrogram graphically depicting the relationships among a group species, is one of the oldest and most powerful icons in all of biology. After Charles Darwin sketched the first species tree

Maria Anisimova (ed.), *Evolutionary Genomics: Statistical and Computational Methods*, Methods in Molecular Biology, vol. 1910, https://doi.org/10.1007/978-1-4939-9074-0_7, © The Author(s) 2019

(in *Transmutation of Species*, Notebook B, 1837), he remained fascinated by the image for 22 years, eventually including a species tree as the only figure in *On the Origin of Species* [1]. Though species trees reached their aesthetic apogee with Ernst Haeckel's *Tree of Life* in 1886, the pursuit of ever-more scientifically accurate trees has kept phylogenetics a vibrant discipline for the 150 years since.

Because the direct evolution of species in the past is not observable (not even in the fossil record), relationships among species are often inferred by shared characteristics among extant taxa. Until the 1970s, this effort took place almost exclusively by using morphological characters. Although this approach had many successes, the paucity of characters and the challenges of comparing species with no obvious morphological homologies were persistent problems [2, 3]. When molecular techniques were developed in the late 1960s, it soon became clear that the sheer volume of molecular data that could be collected would represent a vast improvement. When DNA sequences became widely available for a range of species [4], molecular comparisons quickly became *de rigueur* [5–8]. Nonetheless, it was recognized early on that molecular phylogenies had their own suite of problems; the concept that not all gene tree topologies would match the true species tree topology (i.e., would not be speciodendric sensu Rosenberg [9]) was implicit in early empirical allozyme and mitochondrial DNA studies [10, 11]. However, it was generally assumed that the idiosyncratic genealogical history of any one gene, as reconstructed from extant mutations, was an acceptable approximation for the true history of the species given the potentially overwhelming quantity and seductive utility of molecular data [12–15]. Indeed, this assumption is still prevalent in the thinking of those who favor concatenation or supermatrix approaches as a means of combining information from multiple genes that may still differ in their genealogy from each other and from the species tree [16, 17]. In the meantime, the term "phylogeny" frequently became conflated with "gene tree," the entity produced by many of the leading phylogenetic packages of the day. The term "species tree," in use since the late 1970s to emphasize the distinction between lineage histories and gene histories (reviewed in [11, 18]), was only gradually acknowledged, despite the fact that species trees are the rightful heirs to the term "phylogeny" and better encapsulate the true goals of molecular and morphological systematics [19].

1.1 Stopgap Approaches to Gene Tree Heterogeneity

By and large, the ensuing decades of molecular phylogenetics has fulfilled much of its potential, revolutionizing taxonomies and resolving conundrums previously considered intractable. However, as the amount of genetic data per species becomes ever-more voluminous, it has become clear that the conflicts between individual genes with each other and with the overarching species tree,

both in topology and branch lengths, can have practical consequences for phylogenetic analysis if not dealt with properly [18–23]. At first, some researchers treated this phenomenon as though it were an information problem: when working with only a few mutations, you were bound to occasionally get unlucky and sequence a gene whose random signal of evolution did not match that of the taxa being studied. The reasoning was surely more and/or longer sequences would fix that problem and cause gene trees to converge [16]. However, as more genes were sequenced, and as the properties of gene lineages within populations were studied in detail [24, 25], the twin realities of gene tree heterogeneity and "incomplete lineage sorting" [11] (ILS) became clear (Figs. 1 and 2). The probability of an event such as incomplete lineage sorting, which if considered alone would lead to inferring the wrong species tree, was worked out theoretically for the four allele/two species case first [26], followed by the three allele/three species case [7, 13] and more general cases [12, 27]. Pamilo and Nei [12] were among those that proposed that the solution was to simply acquire more gene sequences, after which the central tendency of this gene set would point to the correct relationships, a "democratic vote" method, where each gene was allowed to propose its own tree, and the topology with the most "votes" was declared the winner and therefore the true phylogeny. Though generally true for three-species case, it can sometimes produce the wrong topology with four or more species [28]. In fact, we now know that with four or more species, there is an "anomaly zone" for species trees with short branch lengths as measured in coalescence units, in which the addition of more genes for sampled taxa is guaranteed to lead to the wrong species tree topology for the democratic vote method [29, 30]. (Coalescent time units, equivalent to t/Ne where t is the number of generations since divergence and Ne is the effective population size of the lineage, are a convenient unit for discussions of gene tree/species tree heterogeneity. For a clear explanation, see Box 2 of Degnan and Rosenberg [28].) Such anomaly zones may be rare empirically [31], but empirical examples are emerging [32, 33], and even the theoretical possibility remains disconcerting. In addition, because the number of possible tree topologies increases as the double factorial of the number of tips, for species trees with more than four tips, a very large number of genes are required to determine which gene tree is in fact the most frequent. Advanced consensus methods [34] can circumvent some of the problems of the democratic vote by using novel assembly methods, such as rooted triple consensus [35], greedy consensus [36], or supertree methods [37]. However, although such methods suffer from lack of a biological model motivating the method of consensus, approaches such as that proposed by Steel and Rodrigo [38] might help approximate the dynamics of biological models while allowing for faster and more flexible extensions and should be further developed.

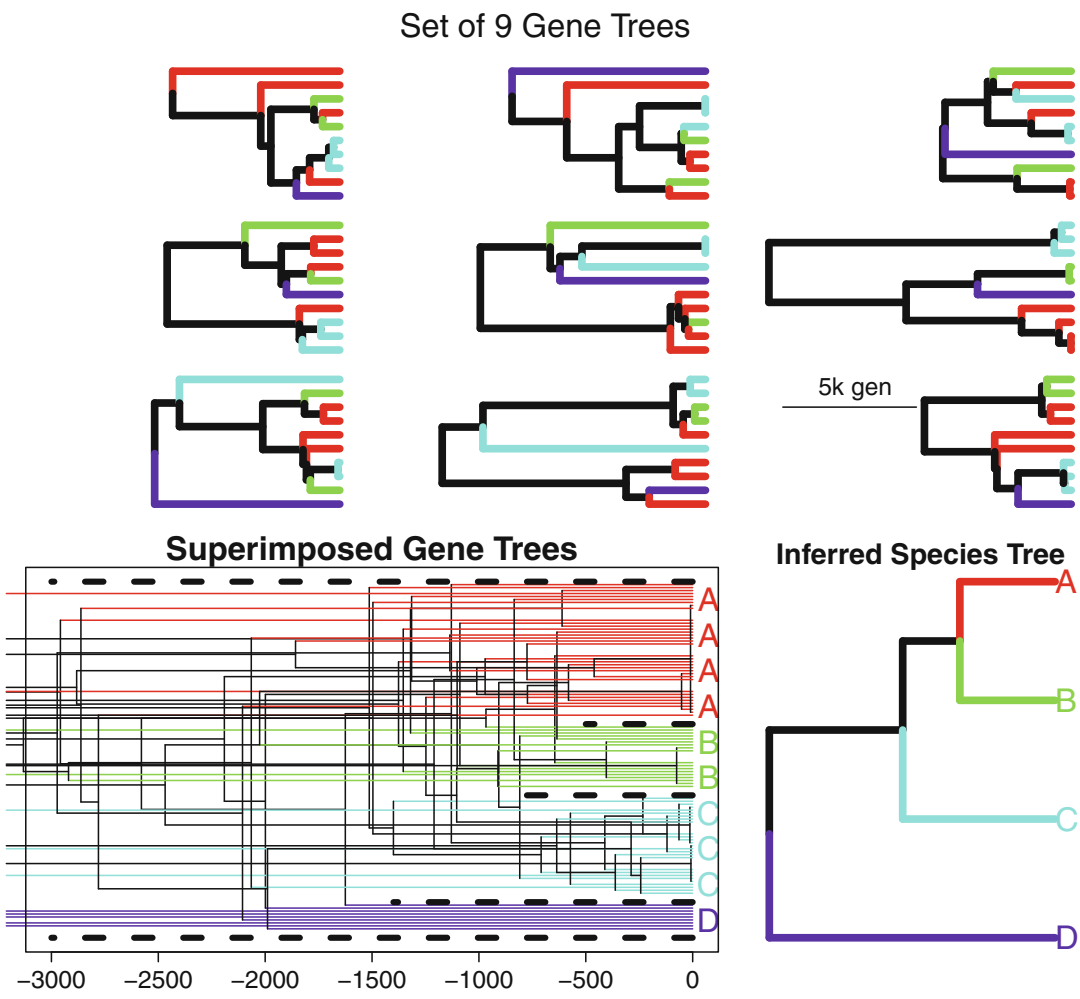


Fig. 1 An example showing the utility of multiple gene trees in producing species tree topologies. (**a**) Nine unlinked loci are simulated (or inferred without error) from a species group with substantial amounts of incomplete lineage sorting. Note that no single gene recovers the correct relationship between clades. Furthermore, despite identical conditions for all nine simulations, no two genes agree on the correct topology, let alone the correct divergence times. (**b**) Superimposing the nine gene trees on top of each other clarifies the relationships. It can be (correctly) inferred that the true tree is perfectly ordered, with (ABC) diverging from D about 1500 generations ago, the (AB)-C split occurring at 800, and A diverging from B about 600 generations ago. Also, the amount of crossbreeding within the recently diverged taxa implies (correctly) that C has the effective smallest population size

The second empirical approach to the problem of conflicting gene trees was to bypass it altogether. Concatenation methods appended the sequence of one gene to that of the next, to create long alignments or supermatrices [39], a technique that in some situations was superior to standard consensus methods in resolving discordance or achieving statistical consistency [40]. But some researchers, including those who questioned the "total-evidence"

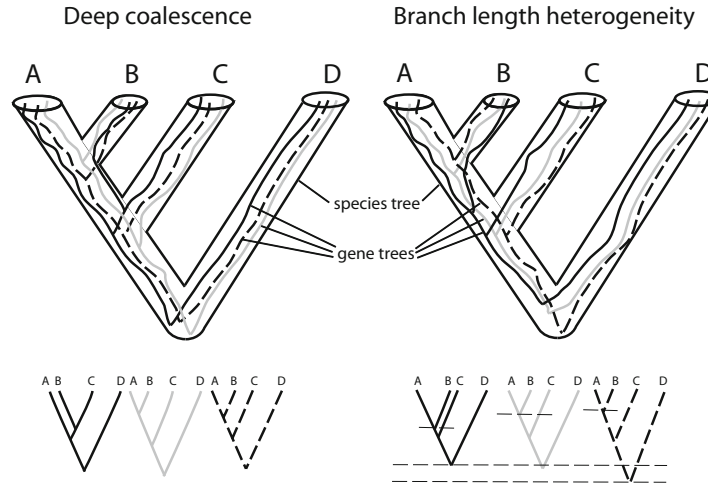


Fig. 2 The relationship between gene trees and species trees. Lines within the species trees indicate gene lineages. Simplified gene trees are shown below each species tree. Whereas gene trees on the left vary due to deep coalescence, gene trees on the right are topologically concordant but vary slightly in branch lengths due to the coalescent. Modified with permission from [19]

approach to systematics (e.g., [41]), advocated against concatenation when, for whatever reason, gene trees appeared to conflict with one another. One problem with the concatenation approach was that it assumed full linkage across the supermatrix, a situation that would obviously not be the case if genes were on different chromosomes. Even when the lineage lengths in a species tree are long in coalescent units, such that gene tree topologies are congruent, the branch lengths of trees of genes on different chromosomes will differ subtly from one another due to the stochasticity of the coalescent process. The early implementations of this method also assumed the same distribution of mutation rates across the sequence, which was clearly not the case if the matrix included coding and noncoding regions. Like democratic vote methods, concatenation of many genes was sometimes defended as sufficient to override the conflicting signal across genes [42, 43], despite widespread acknowledgment that gene tree heterogeneity is ubiquitous and that concatenation can sometimes give the wrong answer, especially although not exclusively in the anomaly zone [44, 45].

Concatenation as a method of combining phylogenomic data still remains popular by default [16, 46], particularly among phylogenetic studies of higher taxa where incomplete lineage sorting is assumed to be rare. However, this logic suffers from two flaws frequently seen in the literature. First, "deep" phylogenetic studies among higher taxa are no more immune to the problems of ILS

than are studies among closely related species, because it is the *length* of a given branch, not its *depth* in the tree, that is relevant to probability of gene tree discordance [28]. Detecting such ILS and ruling out gene tree congruence will indeed be more challenging in deep phylogenomic studies, but it should not be assumed that ILS will be less prevalent at deep scales than at shallow scales. Second, current implementations of concatenation represent only one way of species tree construction in which each gene is forced to have the same topology. The real distinction between concatenation and coalescent models is not the presence or absence of ILS but rather the possibility of conditional independence of gene trees as mediated by recombination between genes [47]. Even if all gene trees in an analysis are topologically identical, physically connecting different genes in a single supermatrix does not capture variation in branch lengths that recombination will allow in nature. More effort should be devoted to "supermatrix-like" methods that constrain gene trees to the same topology but allow recombination between genes and conditional independence of branch lengths, since these qualities will influence how signal is accumulated as more genes are added [47]. A final problem with concatenation is that, in a strict sense, concatenation also does not generate species trees, in so far as the method treats all nucleotides as if they were part of a single non-recombining gene, and thus does not distinguish between gene and species trees [19]. In the end, concatenation is best thought of as a special case of more general models of phylogenetic inference that acknowledge gene tree heterogeneity and conditional independence of genes. One such model is the multispecies coalescent model [23, 28, 48]. It is this model that provides the basis for a recent flurry of promising methods that permit efficient and consistent estimation of species trees under a variety of conditions.

2 The Multispecies Coalescent Model

A plausible probabilistic model for analyzing multilocus sequences should involve not only the phylogenetic relationship of species (species tree) but also the genealogical history of each gene (gene tree) and allow different genes to have different histories. Unlike concatenation, such a multispecies coalescent model (MSC) explains the evolutionary history of multilocus sequences through two levels of biological hierarchy, the gene tree and the species tree, rather than just one [23, 49]. Models acknowledging these two levels require an explicit description of how sequences evolve on gene trees, the traditional likelihood equation of Felsenstein [50] and others, as well as how gene trees evolve in the species tree, the likelihood for which was first described by Rannala and Yang [48]. With a few exceptions (described below), the genealogical

relationship (gene tree) of neutral alleles can be simply depicted by a coalescence process in which lineages randomly coalesce with each other backward in time. The MSC is a simple application of the single population coalescent model to each branch in a species tree [28]. It holds the standard assumptions found in many neutral coalescent models: no natural selection or gene flow among populations, no recombination within loci but free recombination between loci, random mating and a Wright-Fisher model of inheritance down each branch of the species tree. Despite these seemingly oversimplified assumptions, the pure coalescent model is fundamental in explaining the gene tree-species tree relationship because it forms a baseline for incorporating additional evolutionary forces on top of random drift [28, 49]. More importantly, the pure coalescent model provides an analytic tool to detect the evolutionary forces responsible for the deviation of the observed data (molecular sequences) from those expected from the model.

The coalescent process works, in effect, by randomly choosing ancestors with replacement from the population backward through time for each sequence in the original sample. Eventually, two of these lineages will share a common ancestor, and the lineages are said to "coalesce." The process continues until all lineages coalesce at the most recent common ancestor (MRCA). Multispecies coalescence works the same way but places constraints on how recently the coalescences occur, corresponding to the species' divergence times. Translating this model into computer algorithms for inferring species trees has led to a plethora of models [51–55], some of which first build gene trees by traditional methods and then combine them into a species tree with the highest likelihood or other criteria ("two-step" methods, e.g., [56] or [57]), others of which, particularly Bayesian methods [58–60], simultaneously estimate gene trees and species tree. In general for likelihood or Bayesian approaches, a species tree has been proposed, and the likelihood of each gene tree is evaluated using the MSC, with or without various priors, to evaluate the likelihood of the data (DNA sequences in the case of Bayesian methods or gene trees in the case of likelihood methods like MP-EST [56]) given the species tree or the posterior probability of the species tree. In this way, traditional multispecies coalescent methods are the converse of consensus methods; rather than each locus proposing a potentially divergent species tree, a common species tree is assumed and evaluated, given the sometimes divergent patterns observed among multiple loci.

A number of implementations of this idea have been developed (reviewed by Edwards [19, 54]). Several "two-step" packages are available for moving from independently built gene trees to species trees, including minimization of deep coalescence [61], STEM [62], JIST [63], GLASS [64], STAR, STEAC [65], NJst [66], and ASTRAL [57, 67]. Three methods to date utilize "one-step" Bayesian methods to infer gene trees and the species tree, with the

input data being DNA sequences: BEST [58, 68, 69], *BEAST2 [59], and a new model (A00) in the Bayesian Phylogenetics and Phylogeography (bpp) package [70–72]. An additional "one-step" method, SVD Quartets [73], derives species trees directly from aligned, unlinked single-nucleotide polymorphisms using the method of invariants in a coalescent framework. Species tree methods exhibit a number of attractive advantages over concatenation methods in terms of performance. These advantages are not restricted to the anomaly zone, occur across broad regions of tree space, and include less susceptibility to long-branch attraction [74] and missing data [75]. Another attractive aspect of species tree methods and multispecies coalescent models is that they deliver more appropriate estimated levels of confidence that are more evenly spread across genes and appear to be less susceptible to the inflation of posterior probabilities that was early on attributed to Bayesian analyses (e.g., [76, 77]) but may also be due to model misspecification due to concatenation [53]. Bayesian methods are generally agreed to be the most efficient and accurate, capturing all details of the MSC model seamlessly [52]. However, one drawback is that the estimation of larger numbers of parameters (population sizes and divergence times in addition to topologies) can slow computation, may not be relevant in some situations [78], and is generally not possible with the large data sets that are routinely seen today in phylogenomics [59]. Thus far, two-step methods such as ASTRAL, STAR, NJst, and MP-EST have proven the most widely used for large-scale phylogenomic studies, such as the Avian Phylogenomics Project [79] and large-scale phylogenomics of fish [80] and plants [81].

2.1 Sources of Gene Tree/Species Tree Discordance and Violations of the Multispecies Coalescent Model

2.1.1 Population Processes

The "standard" and most common reason why gene trees are not speciodendritic is incomplete lineage sorting, i.e., lineages have not yet been reproductively isolated for long enough for drift to cause complete genetic divergence in the form of reciprocal monophyly of gene trees ([82]; Figs. 1 and 2). This source of gene tree heterogeneity is guaranteed to be ubiquitous, if only because it arises from the finite size populations of all species that have ever come into existence. Almost all the techniques and software packages discussed above are designed to approximate uncertainties in species tree topology arising from this phenomenon.

Delimitation of Species and Diverging Lineages

For recent divergences, the definition of "species" can become problematic for species tree methods [63], and the challenge of delimiting species has, if anything, increased now that the overly conservative strictures of gene tree monophyly as a delimiter of species have been mostly abandoned [82]. This fundamental issue in a phylogenetic study—whether the extent of divergence among lineages warrants species status—has not gone away in the genomic era. However, traditional species tree methods using the MSC need

not use "good" species as OTUs; they will work perfectly well on lineages that have recently diverged, so long as they have ceased exchanging genes. The key issue is not whether the OTUs in species tree analyses are in fact species but rather whether they have ceased exchanging genes, which has been shown to compromise traditional MSC methods [83, 84] (see below).

The problem of species delimitation may ultimately be solved by data other than genetics, and today few species concepts use strictly genetic criteria [85]. Some have suggested that the line between a population-level difference and a species-level difference can be drawn empirically and with consistency in well-studied taxa such as birds, using morphological, environmental, and behavioral data simultaneously [86]. Thus, there is some hope that species delimitation can be performed rigorously a priori in many cases. Researchers who opt for delimiting species primarily with molecular data have a wide array of techniques and prior examples available to them, although not all without controversy [71, 87–93]. Recent progress in species delimitation is motivated by the conceptual transition from "biological/reproductive isolation species" to the "lineage species concept," which defines species not in terms of monophyly of gene lineages but as population lineage segments in the species tree [93]. Under that expanded concept, boundaries of species (i.e., lineages in the species tree) can be facilitated by collection and analysis of gene trees in the framework of the multispecies coalescent model [72]. The recent suggestion that coalescent species delimitation methods define only structure but not species [90] was, in our view, already well-established, with confusion stemming largely from the term "species delimitation," as opposed to "delimitation of populations between which gene flow has ceased."

Gene Flow

There are a number of other situations in which the assumptions of the coalescent are violated. MSC models involve a series of isolation events unaccompanied by gene flow. In this regard, they are like the isolation-migration models of phylogeography [94, 95] but without the migration. The assumption of no gene flow naturally restricts their utility, but gene flow of course compromises other methods of phylogenetic inference, including concatenation methods, as well. Additionally, situations in which gene flow yields a prominent molecular signal often are detectable primarily among very closely related species in the realm of phylogeography [96]. If some substantial gene flow continues between species after divergence, then the multispecies coalescent can quickly destabilize, especially for a small number of loci and as the rate of genetic introgression increases (Fig. 6 in [87, 97–99]). We recommend model comparison algorithms like PHRAPL [87] for determining whether a given data set conforms to the assumptions of the MSC.

2.1.2 Molecular Processes

In addition to species delimitation and gene flow, there are at least three mechanisms that generate discordance on the molecular level (Fig. 3). These include horizontal gene transfer (HGT), which can pose a serious risk to phylogenetic analysis; gene duplication, whose risks can be avoided by certain models; and natural selection, which generally poses no direct threat but, depending on its mode of action and consequences for DNA and protein sequences, can be the most challenging of all.

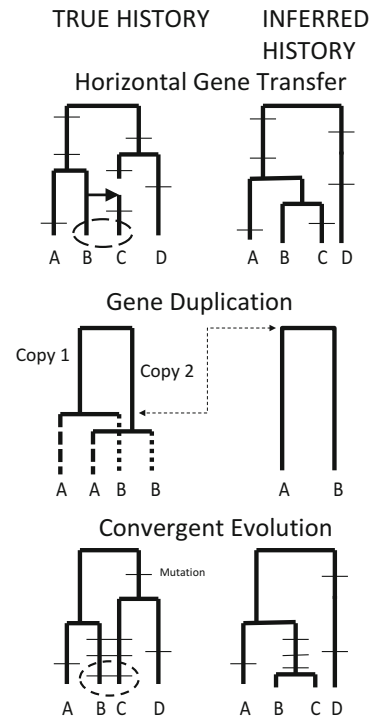


Fig. 3 Three examples of gene histories that depart from the standard multispecies coalescent model. (**a**) A duplication event that precedes a speciation event can lead to incorrect inference of divergence times in the species tree if copy 1 is compared to copy 2. This can be particularly difficult if one of the gene copies has been lost or not sequenced by the researcher. (**b**) Convergent evolution can occur at the molecular level, for example, in certain genes under strong natural selection or highly biased mutational processes. These processes will tend to bring together distantly related taxa in the phylogenetic tree and are likely to be given additional false support by morphological data. (**c**) Horizontal gene transfer causes difficulties in some current species tree methods, because it establishes a spurious lower bound to divergence times. Though rare in eukaryotes, it is by no means unknown and is likely to become a more difficult problem in the future when species trees are based on tens of thousands of loci

Horizontal Gene Transfer

HGT is now known to be so widespread across the Tree of Life, especially in prokaryotes, that some have suggested a web of life may be a more appropriate paradigm for phylogenetic change [100–102]. Growing evidence shows that even eukaryotic genomes contain substantial amounts of "uploaded" genetic material from bacteria, archaea, viruses, and even fellow eukaryotes [103–105]. Even though effective techniques are not yet widely available for detecting HGT in eukaryotes, enough individual cases have been "accidentally" discovered that researchers have given up trying to list them all [103].

The implications of HGT for species tree construction vary depending on the method used. For example, following the standard assumption in coalescent theory that allelic divergences must occur earlier in time than the divergences of species harboring those alleles, some species tree techniques [48, 58], as well as classical approaches (e.g., [13]), assume that the gene tree exhibiting the most recent divergence between taxon A and taxon B establishes a hard upper limit on the divergence time of those species in the species tree. For small sets of genes in taxa where HGT is rare, a researcher would need to be quite unlucky to choose a horizontally transferred gene for analysis. However, as the genomic era advances, it becomes more likely that at least one of the thousands of genes studied will have been transferred horizontally and thus establish a spurious upper bound for clade divergence at the species level. When selective introgression of genes from one species to another is considered, this number of genes coalescing recently between species will increase [106]. Although HGT is clearly a problem for some current methodologies, if transferred genes can first be identified, then they could be extremely useful as genomic markers for monophyletic groups that have inherited such genes and would otherwise be difficult to resolve [107]. However, for other species tree methods that calculate averages of coalescence times, such as STAR [65], HGT events will have less of an impact. Liu et al. [56] examined the effect of HGT on the pseudo-likelihood method MP-EST and predicted that, mathematically, species tree branch lengths may be biased by HGT but that topologies were fairly robust. Davidson et al. [108] found that quartet-based methods, such as ASTRAL-II, were fairly robust to HGT in the presence of ILS. Removal of genes suspected to be transferred via HGT prior to species tree analysis would be warranted; however, some methods to detect such events rely both on having the true species tree already in hand and also on the absence of other mechanisms causing gene tree discordance [109–112]. Recent work aims to incorporate HGT into other mechanisms of gene tree incongruence (reviewed in [113]); how much we need to invest in such synthetic methods will likely depend on the prevalence of HGT in particular taxonomic groups.

Gene Duplication

Gene duplication presents another violation of the basic MSC model (Fig. 3); like HGT, its potential problems are worst when they go unrecognized [49]. Imagine a taxon where a gene of interest duplicated 10 Mya into copy α and copy β; the taxon then split 5 Mya into species 1 and 2. A researcher investigating the daughter species would therefore sequence four orthologous genes, with the potential to compare α1 to β2 and β1 to α2 and thus generate two gene trees where the estimated split time was 10 Mya, rather than 5 Mya. Such a situation will be easily recognized if copy α and β have diverged sufficiently by the time of their duplication, and a number of methods of coalescent analysis have incorporated gene duplication (e.g., [114, 115]; reviewed in [116]). Additionally, failure to recognize the situation may not have drastic consequences for phylogenetic analysis if the paralogs have coalesced very recently or are species-specific, in which case the estimated gene coalescence would be approximately correct no matter which comparison was made. However, if one of the copies has been lost and only one of the remaining copies is sequenced, then the chances of inferring an inappropriately long period of genetic isolation are larger and will increase as the size of the family of paralogs increases. Assessing paralogs in phylogenomic data is a major challenge, particularly in groups like plants and fish, and a growing number of dedicated methods ([117]; assessed in [118]) or filtering protocols [119] for doing so exist. This problem will tend to overestimate gene coalescence times, and some species tree methods depend on minimum isolation times among a large set of genes. These deep coalescences might spuriously increase inferred ancestral population sizes. A systematic search for biases incurred by species tree methods due to gene duplication is needed.

Natural Selection

Natural selection causes yet another violation of the multispecies coalescent model. Selection can cause serious problems in some cases, although in other circumstances it is predicted not to cause problems of phylogenetic analysis [47, 120]. The usual stabilizing selection can be helpful to taxonomists working at high levels because it slows the substitution rate; likewise selective sweeps, directional selection, and genetic surfing [121] tend to clarify phylogenetic relationships by accelerating reciprocal monophyly for genes in rapidly diverging clades. However, challenges to phylogenetic inference are posed by any evolutionary force that may bias the reconstruction of gene trees, including convergent neutral mutations (homoplasy), balancing selection, and selection-driven convergent evolution (e.g., [122]). Balancing selection tends to preserve beneficial alleles at a gene for long periods of time and is probably the most insidious form of selection with respect to accurately reconstructing gene trees and species trees.

2.2 More About Violations and Model Fit of the Multispecies Coalescent Model

Many of the instances of violations of the coalescent model will occur at individual genes and usually will not dominate the signal of the entire suite of genes sampled for phylogenetic analysis. Reid et al. [123] conducted one of the few tests of the fit of the MSC to multilocus phylogenetic data. Although the title of their article suggests that the MSC overall provides a "poor fit" to empirical data, we suggest that their results provide a more hopeful picture. The most important thing is that they investigated the fit of the MSC to individual loci in phylogenetic data sets and were able to identify loci that failed to fit the MSC. They were less successful at identifying the causes of departure from the MSC for individual loci.

More common but still rare are efforts to determine which models of phylogenetic inference, the MSC or concatenation, provide a better fit to empirical phylogenomic data. Edwards et al. [124] and Liu and Pearl [58] both used the Bayesian species tree method BEST [68] to ask using Bayes factors whether the MSC or concatenation fits empirical data sets better. Uniformly, they found that the MSC fit empirical data sets better than concatenation, often by a large margin. However, further work in this area is still needed. Most discussions in the literature have focused on the perceived failings or violations of the MSC by empirical data sets—such as evidence for recombination within loci—even when such failings or assumptions also apply to concatenation [47]. Given that all models are approximations of reality, a better focus would be to ask which model better fits empirical data sets better. The limited research that has been done suggests overwhelmingly that the MSC provides a better fit to empirical data sets than concatenation.

Are there better models for phylogenomics than the MSC? Depending on the data set, almost surely there are (Fig. 4). Several authors working with phylogenomic data sets have suggested that gene flow is detectable, even among lineages that diverged a long time ago (e.g., [129, 130]). The increasing number of reports of hybridization and introgression among phenotypically distinct species suggests that hybridization may be a typical component of speciation and that even phylogenetic models can be improved by incorporating such reticulation (e.g., [47, 106, 131]). The pure MSC is best thought of as a special case of so-called "multispecies network coalescent" models, or MSNC [127, 132–134] (Fig. 4), in which gene flow connects some branches of the species tree. In the end, empiricists will need to decide what level of model fit they are willing to tolerate and which software packages can accommodate the large data sets that are now routine in phylogenomics.

2.2.1 Phylogenetic Outlier Loci

Genes whose phylogenetic signal differs significantly from that of the remainder of data set can be thought of as phylogenetic outliers. These loci are conceptually similar to outliers in population genetics, which have been the focus of many studies (reviewed in

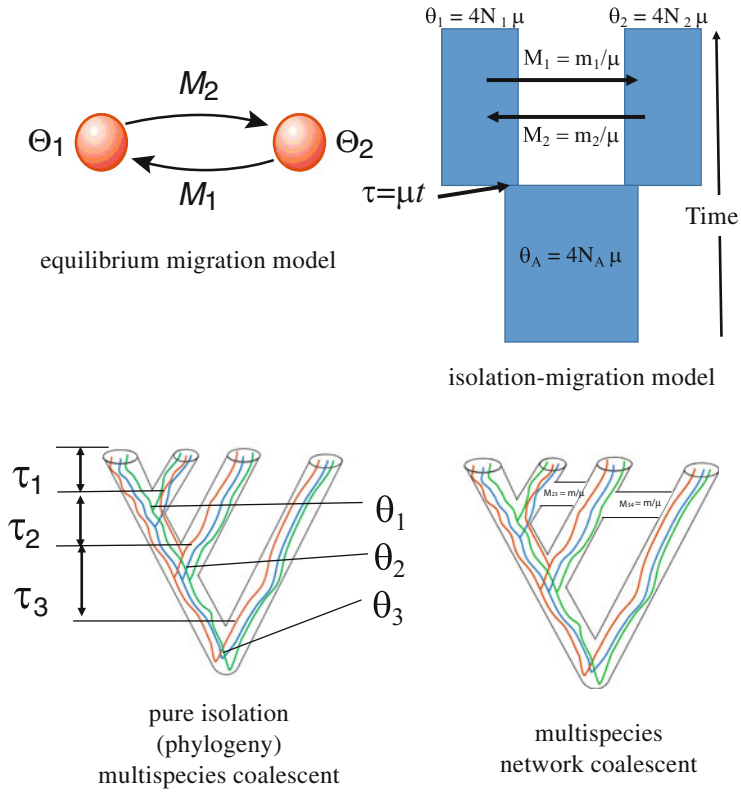


Fig. 4 Diversity of phylogeographic models. Species trees estimated by the multispecies coalescent are naturally related to previous phylogeographic models by their shared demographic parameters, usually measured in units of mutation rate or substitutions per site (μ), including genetic diversity or effective population size ($4N\mu$, where $N =$ effective population size; gene flow M/μ, where $M =$ the scaled migration rate; $4Nm$, where m is the number of migrants per generation; and divergence time $\tau = \mu t$, where t is the divergence time in generations). (**a**) Equilibrium migration models as envisioned by early versions of the software MIGRATE [125]. (**b**) Isolation-migration models envisioned by Hey and coworkers [48, 95, 126]. Subscript A indicates ancestral population size. (**c**) Species tree models estimated by the multispecies coalescent [28]. (**d**) Multispecies network coalescent models or phylogenetic network models including divergence and gene flow [127, 128]

[135–137]). However, there has been little work in detecting phylogenomic outliers. Much attention has been paid to particular sites in a data set that differ from the majority and therefore exhibit homoplasy or incongruence with the rest of the data set [76, 138]. The sources of such incongruence are many and can

include mutational processes (e.g., gene duplication), HGT, as well homoplasy (e.g., [139, 140]). Incongruence of particular sites, or entire loci, may also be due to technical issues such as contamination, misassembly, mistaken paralogy, annotation mistakes, and alignment errors (e.g., [119]). Here, in an analogy with work in population genetics, we will focus primarily on entire loci that deviate from the expected distribution governed by neutral processes due to natural selection. Understanding the distribution of gene tree topologies expected under the neutral multispecies coalescent [25] is a good starting point for identifying loci that may be targets of natural selection.

2.2.2 Genomic Signals of Phylogenetic Outliers

When faced with a surprising or nonconvergent species tree, one possibility is that an unusual gene tree is to blame. Though techniques for dealing with violations of the coalescent model are in their infancy, researchers do have a few options. Below we list several ideas, some borrowed from classical phylogenetics or from methods used in bioinformatics. It is likely that the several tests constructed to detect phylogenetic outliers in classical phylogenetics can be extended slightly to incorporate the additional variation among genes expected due to the coalescent process. Of course, with larger data sets, at least with some coalescent methods, single anomalous genes may have little effect on the resulting species tree, particularly in species tree methods utilizing summary statistics [65]. However, as pointed out above, species tree methods such as BEST that relies on "hard" boundaries for the species tree by individual genes could be derailed due to the anomalous behavior of even a single gene.

Jackknifing: A straightforward approach to detecting phylogenetic outliers under the multispecies coalescent model is to rerun the analysis n times, where n is the number of loci in the study, leaving one locus out each time. An outlier can then be identified if the analysis that does not include that gene differs from the remaining analyses in which that gene is included. This approach has been applied successfully in fruit flies by Wong et al. [21], who considered their problem resolved when the elimination of one of the ten genes unambiguously resolved a polytomy. There may be other metrics of success that are more robust or sensitive or do not depend as strongly on a priori beliefs about the relationships among taxa. Because some duplications or horizontal transfers may affect only one taxon, whole-tree topology summary statistics are unlikely to be sensitive enough to detect recent events. However, the cophenetic distance of each taxon to its nearest neighbor in the complete species tree could be compared across jackknife results. This procedure will produce a distribution of "typical" distances, and significance can therefore be assigned to highly divergent results. The drawback to such an approach is the

computational demand. Species tree analyses on their own can be extremely time consuming to run even once, so jackknifing may prove intractable for studies involving many species and loci (*see* ref. 141).

2.2.3 Simulation Approaches to Detecting Phylogenetic Outliers

Simulating gene trees from a species tree is another method for identifying gene trees that differ from the majority of loci in the data set. Several species tree methods yield estimate of the phylogeny that include branch lengths in coalescent units [56, 57, 70], which are required to simulate gene trees from a species tree. Branch lengths in the estimated species tree can be decomposed into a number of substitutions per site and an estimate of $\theta = 4N\mu$ that are compatible with the original branch length in coalescent units. For example, using any number of algorithms, including maximum likelihood or Bayesian methods, the length of species tree branch lengths in substitutions per site can be approximated by fitting the concatenated alignment of genes to the estimated species tree topology, yielding a tree with the same topology but branch lengths in substitutions per site (μt, where t is the time span of the branch in either generations or years). With these branch lengths in hand, estimates of θ can then be applied to each branch so that the original coalescent units $t/2N \approx \mu t/\theta$ from the species tree are retained. Care needs to be taken to preserve the appropriate ploidy units when simulating gene trees from an estimated species tree. Packages such as MP-EST yield estimates of species tree branch lengths in coalescent units of $4N$ generations, appropriate for diploids, whereas packages such as Phybase [142] simulate gene trees from a species tree in estimates of $2N$ units, appropriate for haploids. Another issue that is important to be aware of is the distinction between gene coalescence times and species tree branch lengths [143, 144]. Whereas species tree branch lengths are estimates of lineage or population branch lengths in the species tree, the DNA sequence alignment that is fitted to the species tree will yield branch lengths reflecting the coalescence time of genes in ancestral species. This discrepancy occurs because gene coalescence times by necessity predate and record a more ancient event than do species divergence times. The discrepancy may represent a small fraction of the branch length if species divergence times are large, but Angelis and dos Reis [143] have suggested that the discrepancy can be quite large even in comparisons of distantly related species, such as exemplars of mammalian orders. There is a great need for methods of molecular dating and combining fossils and DNA data that distinguish between gene coalescence times and speciation times, the latter of which is usually of primary interest.

Once the branch lengths of the species tree are prepared for simulation, gene trees can be simulated using a number of packages (Phybase, [142]; TreeSim, [145]; CoMus, [146]). Even packages traditionally used in phylogeography can be used to simulated gene

trees on species trees, given the close relationship between species trees and phylogeographic models like isolation migration [147, 148]. One can then compare the distribution of gene tree topologies and branch lengths observed in one's data set with those simulated under the neutral coalescent model. A common approach is to calculate the distribution of Robinson-Foulds [149] distances among simulated gene trees and compare these to those observed in the original data set. Such approaches have been used to determine if a data set is consistent with the MSC or the percent of the observed gene tree variation that is explained by the MSC. Other statistics, such as the similarity in number of minority gene tree triplets produced by a given species tree at each node, can also be compared to the observed distribution. Song et al. [150] used coalescent simulations using Phybase to propose that the MSC could explain a large (>75%) fraction of the observed gene tree variation in a mammalian data set. Such simulations assume that the gene tree variation observed is biological in origin and not due to errors in reconstruction. They also noted that the near equivalence in frequency of minority triplets in gene trees at various nodes in the mammal tree suggested broad applicability of the neutral coalescent without gene flow or other complicating factors. Still, many papers observe some level of departure of the patterns in the observed data set from those expected under simulation. Usually the source of this departure is unknown. Natural selection or any other force such as HGT or anomalous mutation might be culprits in these cases. Heled et al. [151] proposed a simulation regime that incorporates gene flow between species and thus can be used to test for the effects of migration on gene trees and species tree estimation.

To detect possible phylogenetic outliers, Edwards et al. [152] applied a recently proposed method of detecting gene tree outliers, KDEtrees [153], to a series of phylogenomic data sets. KDEtrees uses the kernel density distribution of gene tree distances to estimate the 95% confidence limits on gene tree topologies in a given data set. Surprisingly, using default parameters, Edwards et al. [152] could not detect a higher-than-expected number of gene tree outliers in any data set, despite the fact that the data sets in several cases contained hundreds of loci. No data set possessed more than the expected 5% of outliers given the test implemented in KDEtrees. Clearly further work is needed to understand the pros and cons of various tests of phylogenetic outliers. For the time being, we can note the robustness of various species tree methods to phylogenetic outliers. One attractive prospect of algorithms for species tree construction that use summary statistics, such as STAR and STEAC, is that these methods are powerful and fast, yet they appear less susceptible to error due to deviations of single genes from neutral expectations. These methods do not utilize all the information in the data and hence can be less efficient than Bayesian or likelihood methods [52], yet they perform well with moderate amounts of gene tree outliers due to processes like HGT.

3 Hypothesis Testing Using the Multispecies Coalescent Model

Hypothesis testing is a cornerstone of phylogenetic analysis but has received little attention in the context of the MSC (*see* ref. 154). Bayesian species tree inference [58, 59, 68–70] provides perhaps the most seamless approach to hypothesis testing. One can relatively easily assess the fit of the collected data to alternative tree topologies and compare the fit using Bayes factors or other approaches. One can also assess the fit of various models of analysis to the collected data [155]. Liu and Pearl [58] and Edwards et al. [124] used Bayes factors to determine whether concatenation or the MSC was a more appropriate model for several data sets; in all cases tested thus far, the MSC provides a far better fit to multilocus data ($BF > 10$) than does concatenation, in which all gene trees among loci are identical. Further work is needed to apply Bayes factors and likelihood ratio tests to multilocus data.

The bootstrap, introduced to phylogenetics by Felsenstein [156], is the most common statistic applied to phylogenetic trees [157]. In the era of multilocus phylogenetics, the "multilocus bootstrap" of Seo [158] has been recommended as a more suitable approach to assessing confidence limits than the traditional bootstrap. In the traditional bootstrap, sites within a locus, or a series of concatenated loci, are resampled with replacement to create pseudomatrices, which are then subjected to phylogenetic analysis, after which a majority rule consensus tree is usually made. By contrast, in the multilocus bootstrap, sites within loci and the loci themselves are resampled with replacement. In the context of the MSC, resampled pseudomatrices of the same number of loci as the original data set, which may contain duplicates of specific loci due to the random nature of the bootstrap, are then made into gene trees, from which a species tree can be made. The bootstrap and various other measures of branch-specific support [159] have been proposed as a means of assessing confidence in species trees made using the multilocus coalescent. Care should be taken in the comparison of different studies using different measures of support, since not all measures can be directly compared to one another. For example, as pointed out by Liu et al. [160], the measure of posterior support for ASTRAL trees proposed by Sayyari and Mirarab [159] is not the same as traditional bootstrap supports, and we do not yet know how they will scale under different conditions compared to the bootstrap. Edwards [161] summarized knowledge about the use of phylogenomic subsampling, in which data sets of increasing size or signal are analyzed so as to understand the stability and speed of approach to certainty of phylogenetic estimates under the MSC and under concatenation. He found that MSC methods tended to approach phylogenomic certainty more smoothly and monotonically than do concatenation methods, which jump around erratically in their certainty for sometimes conflicting topologies,

especially when sampling smaller numbers of genes. Although we cannot simply translate many conclusions from the gene tree era of phylogenetics to the MSC era—for example, contrary to gene tree conclusions, it is not clear for MSC models that more taxa are always better than more loci [74]—many of these discussions about hypothesis testing echo early comparisons of posterior probabilities and bootstrap proportions used in the gene tree era of phylogenetics.

The bootstrap has always provided a means of hypothesis testing that is very indirect with respect to comparing alternative phylogenetic hypotheses. Aside from the tests allowed by Bayesian approaches, there have been few discussions of testing of alternative phylogenetic trees in the era of the multispecies coalescent. In this regard, the pseudo-likelihood model provided by MP-EST [56] provides a convenient framework for hypothesis testing using species trees. This framework is not available in most other species tree methods, including ASTRAL, STAR, and STEAC, since these methods do not employ a likelihood model. MP-EST takes advantage of the likelihood model of Rannala and Yang [48] to assess the fit of a species tree to a collection of gene trees and can thus be used to compare alternative species tree topologies and branch lengths directly.

To conduct a direct comparison of species trees using the likelihood ratio test, we first compare the likelihoods of two trees to find the most probable species tree that can explain the empirical set of gene trees. The likelihood of a set of gene trees given a species tree with branch lengths can be ascertained using functions in Phybase [142]. Let Tree 1 be the null tree and Tree 2 be the alternative tree. The likelihood ratio test statistic is $t = 2(L_{\text{Tree2}} - L_{\text{Tree1}})$, in which L_{Tree1} and L_{Tree2} are the log-likelihoods of the null and alternative hypotheses. The log-likelihood of the null hypothesis can be obtained from the output of the program MP-EST by fitting the branch lengths and topology of Tree 1 to the set of empirical gene trees. Similarly, we can find the log-likelihood of the alternative tree Tree 2 using MP-EST. The null distribution of the test statistic t is approximated by a parametric bootstrap. Specifically, we generate 100 or more bootstrap samples of gene trees under the null tree Tree 1. For each sample of these bootstrapped trees, we calculate the log-likelihoods of the null and alternative trees using the procedure described above. The null distribution of the test statistic t is approximated by the test statistics of the bootstrap samples. If t for the null and alternative species trees is outside the expected distribution of the bootstrap sample statistics, then the result can be considered significant.

We applied this approach to assessing alternative phylogenetic hypotheses to an example from birds (fairy wrens; [162]; Fig. 5). This data set consists of 18 genes and 26 taxa, with loci coming

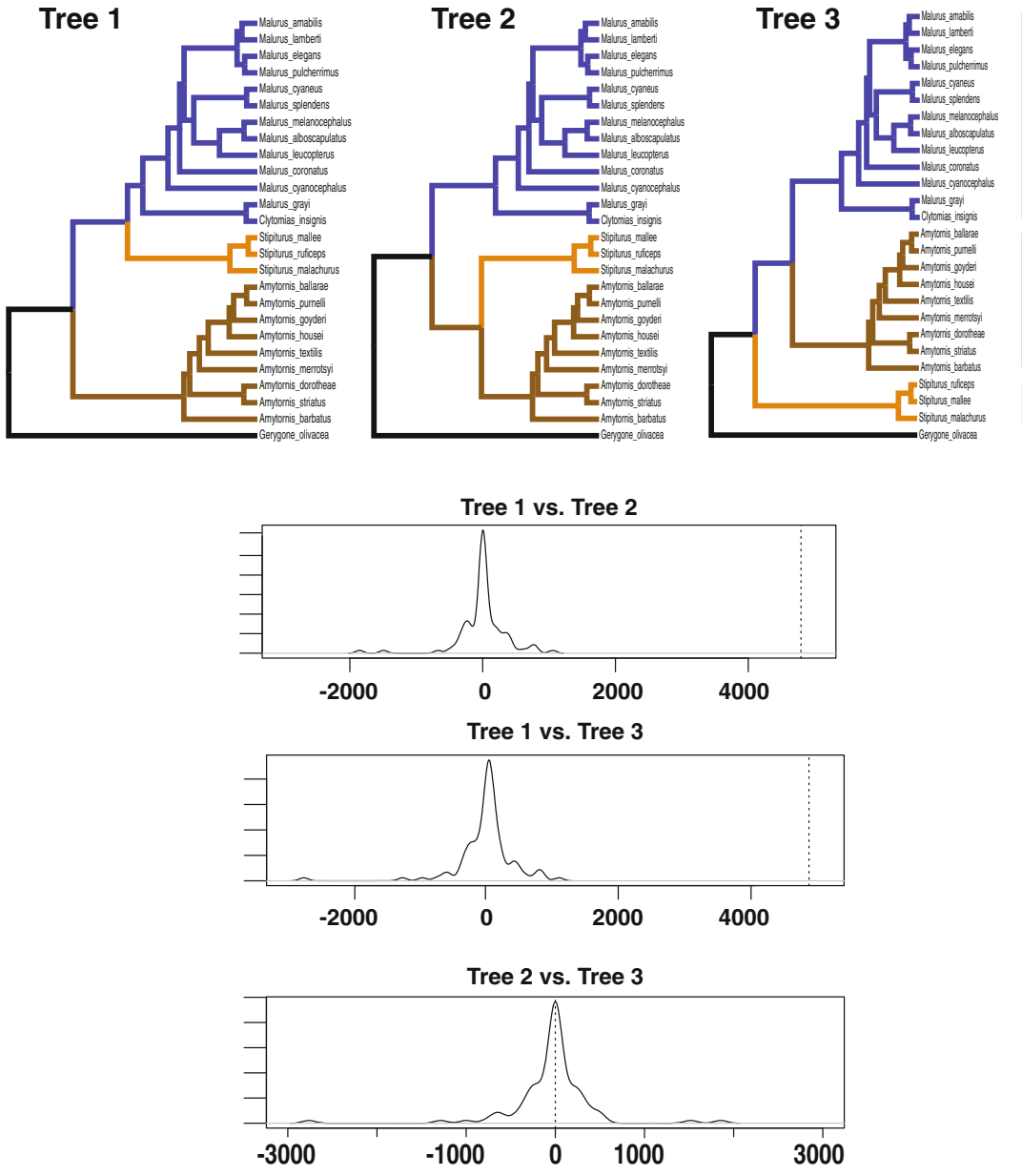


Fig. 5 Example of hypothesis testing of alternative phylogenetic trees under the multispecies coalescent model. Top: alternative phylogenetic hypotheses involving the rearrangement of major groups of Australo-Papuan fairy wrens based on Lee et al. [162]. The three alternative phylogenetic trees are colored to indicate the three major groups whose relationships are being tested. Bottom: results of the likelihood ratio test (LRT) and estimates of confidence limits on the test statistic *t* using parametric bootstrapping. The plots show the distributions of the test statistic *t* resulting from gene trees built from resampled, bootstrapped sequence data. Despite the use of sequence data to generate the bootstrap gene tree distributions, the LRT is only an indirect test of the signal in the sequence data and instead is best thought of as a test of the fit of the estimated gene tree distribution on alternative phylogenies. See main text for further details

from a variety of marker types (exons, introns, anonymous loci). Lee et al. [162] applied a number of MSC approaches to this data set but did not compare alternative trees directly, having only used bootstrap approaches. Here, we consider three-species trees generated from the rearrangement of the three major clades of wrens: the core fairy wrens (*Malurus*), emu-wrens (*Stipitirus*), and grasswrens (*Amytornis*; Fig. 5). Rearranging these major clades results in three alternative rooted species trees. Based on traditional taxonomy and because the gene trees in this data set were highly variable, even among the three major clades, we consider these three alternative hypotheses true alternatives and not "straw men." Rooted maximum likelihood gene trees were built from the alignments of each locus using RaxML [163] and then used as input data for the likelihood ratio test described above. The LRT was applied first to Tree 1 (null) versus Tree 2 and was also applied to Tree 1 versus Tree 3 and Tree 2 versus Tree 3. The results indicate that Tree 1 fits to the empirical gene trees significantly better than does Tree 2 or Tree 3 does ($p < 0.01$), and there is no significant difference between Trees 2 and 3 in their fit to the empirical gene trees ($p = 0.52$). Thus, the LRTs strongly favor Tree 1 over both Tree 2 and Tree 3.

It is important to note that the LRT described above is not a direct test of the phylogenetic signal in the DNA sequence data. Rather, it is a test of the distribution of gene trees inferred from the sequence data and assumes that the gene trees provided as data are without error. It does indirectly test the signal in the sequence data, because if the DNA sequences provide strong and consistent support of the gene trees, then the bootstrapped set of gene trees will be highly similar to one another, and the confidence limits on t will be very tight. By contrast, if the DNA sequence data does not have a strong signal, then the confidence limits on t will be very wide, and it will be difficult to reject alternative species trees. The LRT described here does not involve nested models. If the gene trees are known without error, then the value of t itself can be used to assess significance, assuming a chi-square distribution with 2 degrees of freedom. Further research is needed on methods for comparing and testing alternative species trees in the context of the MSC.

4 Future Directions

Species tree methods are likely to continue to gain ascendancy as the strongest evidence of taxonomic relationship in phylogenetic research. As with any form of evidence, the conclusions of a species tree analysis are fallible, with each method susceptible to biases in the input data. For example, Xi et al. [164] showed that Phyml [165] yields biased gene trees when there is little information in the DNA sequences and can therefore result in biased species trees. This issue is particularly problematic when using MP-EST v. 1.5,

which, unlike ASTRAL or MP-EST v. 2.0, does not randomly resolve or appropriately accommodate gene trees with polytomies or 0 or near 0-length branches. This bias may have affected the performance of MP-EST in previous side-by-side comparisons with ASTRAL. In the future, further work should be devoted to discovering and quantifying additional biases in inference of species trees. With the size of phylogenomic data sets increasing, even small biases can be amplified and result in poorly estimated species trees.

Many in the field agree that the most appealing statistical models for species tree inference using the MSC include Bayesian and full-likelihood models [52]. But it is still clear, at least to empiricists, not only that "two-step" methods of species tree inference work quite well in general but also that the large phylogenomic data sets available today prohibit the use of full-likelihood methods. Regardless, we now know that both types of models clearly outperform concatenation across wide swaths of parameter space, especially if one also evaluates the reliability of the confidence limits on the estimate of phylogeny and not only the point estimate of the topology. The major directions for future research in the field of species tree inference therefore include increasing the scalability of computational inference of species trees, further development of frameworks for hypothesis testing using the MSC, developing additional models of divergence with gene flow and network coalescent models (Fig. 4), and improvement in the estimation of gene trees and species trees from SNP data [166]. Linking mutations in species trees and heterogeneous gene trees to diverse phenotypic and ecological data will be another important avenue for the future [167, 168]. We view the MSC, with its application of population genetic models to higher-level systematics, as a key component of the long-term goal of uniting microevolution and macroevolution. Even if it proves incomplete in the long term, the neutral MSC provides a powerful null model for the understanding of genetic diversity across time and space.

5 Practice Problems

1. Consider the following discordant set of gene trees. {Gene 1 = (A:10,(B:8,C:8):2); Gene 2 = (B:9,(A:6,C:6):3); Gene 3 = ((A:4,B:4):4,C:8)}. Assuming that these genes perfectly reflect the time of genetic divergence, and the only cause of discordance is incomplete lineage sorting or deep coalescence, what is the most likely species tree? *Answer: ((A:4,B:4):2,C:6)*
2. Find the data set for 30 noncoding loci from 4 species of Australian grass finches (3 *Poephila*, plus out-group *Taeniopygia*) from Jennings and Edwards [169]. It can be found in the web page for Liang Liu's BEST program: http://faculty.franklin.uga.edu/lliu/content/BEST. Use the Bayesian program

BEST [68] or BPP [70] and the nonparametric method in STAR [65] to estimate the species tree for the four species, using *Taeniopygia* as the out-group. Do you estimate the same topology with both methods? What about the support for the single internal branch? If the support is not the same, what could be causing the difference? *Answer: The BEST or BPP tree should have higher support than the STAR tree, but they both should have the same topology. The STAR tree might have lower support because in the data set about half of the gene trees have a topology differing from the species tree; whereas the full Bayesian model accommodates this variation accurately, nonparametric "two-step" methods interpret this type of gene tree variation as discordance, in conflict with the majority of the gene trees and with the species tree.*

3. For the above data set, make individual gene trees using RaXml [170], and use the likelihood functions and bootstrap capabilities of Phybase [142] to conduct a likelihood ratio test of the two alternative species tree topologies for the four grass finches. Alternatively, you could use the posterior distribution of gene trees generated in BEST to estimate the confidence limits on the test statistic *t*. Is the tree estimated in question 2 significantly better than alternative trees? *Answer: The LRT indicates that the tree estimated in question 2 is significantly better than alternative trees.*

References

1. Darwin C (1859) On the origin of species, vol Facsimile of 1st Edition. Harvard University Press, Cambridge, p 513
2. Hillis DM (1987) Molecular versus morphological approaches to systematics. Annu Rev Ecol Syst 18:23–42
3. Scotland RW, Olmstead RG, Bennett JR (2003) Phylogeny reconstruction: the role of morphology. Syst Biol 52:539–548
4. Kocher TD, Thomas WK, Meyer A, Edwards SV, Pääbo S, Villablanca FX, Wilson AC (1989) Dynamics of mitochondrial DNA evolution in animals: amplification and sequencing with conserved primers. Proc Natl Acad Sci U S A 86:6196–6200
5. Miyamoto MM, Cracraft J (1991) Phylogeny inference, DNA sequence analysis, and the future of molecular systematics. In: Miyamoto MM, Cracraft J (eds) Phylogenetic analysis of DNA sequences. Oxford University Press, New York, NY, pp 3–17
6. Swofford DL, Olsen GJ, Waddell PJ, Hillis DM (1996) Phylogenetic inference. In: Hillis DM, Moritz C, Mable BK (eds) Molecular systematics. Sinauer, Sunderland, MA
7. Nei M (1987) Molecular evolutionary genetics, vol 512. Columbia University Press, New York
8. Nei M, Kumar S (2000) Molecular evolution and phylogenetics. Oxford University Press, New York
9. Rosenberg NA (2002) The probability of topological concordance of gene trees and species trees. Theor Popul Biol 61:225–247
10. Cavalli-Sforza LL (1964) Population structure and human evolution. Proc R Soc Lond Series B 164:362–379
11. Avise JC, Arnold J, Ball RM, Bermingham E, Lamb T, Neigel JE, Reeb CA, Saunders NC (1987) Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. Annu Rev Ecol Syst 18:489–522
12. Pamilo P, Nei M (1988) Relationships between gene trees and species trees. Mol Biol Evol 5:568–583

13. Takahata N (1989) Gene genealogy in three related populations: consistency probability between gene and population trees. Genetics 122:957–966
14. Avise JC (1994) Molecular markers, natural history and evolution. Chapman and Hall, New York
15. Wollenberg K, Avise JC (1998) Sampling properties of genealogical pathways underlying population pedigrees. Evolution 52:957–966
16. Gatesy J, Springer MS (2014) Phylogenetic analysis at deep timescales: unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. Mol Phylogenet Evol 80:231–266
17. de Queiroz A, Gatesy J (2007) The supermatrix approach to systematics. Trends Ecol Evol 22:34–41
18. Maddison WP (1997) Gene trees in species trees. Syst Biol 46:523–536
19. Edwards SV (2009) Is a new and general theory of molecular systematics emerging? Evolution 63:1–19
20. Carstens BC, Knowles LL (2007) Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from melanoplus grasshoppers. Syst Biol 56:400–411
21. Wong A, Jensen JD, Pool JE, Aquadro CF (2007) Phylogenetic incongruence in the Drosophila melanogaster species group. Mol Phylogenet Evol 43:1138–1150
22. Knowles LL, Kubatko LS (2010) Estimating species trees: an introduction to concepts and models. In: Knowles LL, Kubatko LS (eds) Estimating species trees: practical and theoretical aspects. Wiley-Blackwell, New York, pp 1–14
23. Liu L, Yu L, Kubatko L, Pearl DK, Edwards SV (2009) Coalescent methods for estimating phylogenetic trees. Mol Phylogenet Evol 53:320–328
24. Neigel JE, Avise JC (1986) Phylogenetic relationships of mitochondrial DNA under various demographic models of speciation. In: Karlin S, Nevo E (eds) Evolutionary processes and theory. Academic, New York, pp 515–534
25. Degnan JH, Salter L (2005) Gene tree distributions under the coalescent process. Evolution 59:24–37
26. Tajima F (1983) Evolutionary relationships of DNA sequences in finite populations. Genetics 105:437–460
27. Mehta RS, Bryant D, Rosenberg NA (2016) The probability of monophyly of a sample of gene lineages on a species tree. Proc Natl Acad Sci U S A 113:8002–8009
28. Degnan JH, Rosenberg NA (2009) Gene tree discordance, phylogenetic inference, and the multispecies coalescent. Trends Ecol Evol 24:332–340
29. Degnan JH, Rosenberg NA (2006) Discordance of species trees with their most likely gene trees. Public Lib Sci Genet 2:762–768
30. Rosenberg NA, Tao R (2008) Discordance of species trees with their most likely gene trees: the case of five taxa. Syst Biol 57:131–140
31. Huang HT, Knowles LL (2009) What is the danger of the anomaly zone for empirical phylogenetics? Syst Biol 58:527–536
32. Sackton TB et al (2018) Convergent regulatory evolution and the origin of flightlessness in palaeognathous birds. bioRxiv. https://doi.org/10.1101/262584
33. Linkem CW, Minin VN, Leaché AD (2016) Detecting the anomaly zone in species trees and evidence for a misleading signal in higher-level skink phylogeny (Squamata: Scincidae). Syst Biol 65:465–477
34. Bryant D (2003) A classification of consensus methods for phylogenetics. In: Janowitz M et al (eds) BioConsensus. American Mathematical Society, Providence, RI, pp 163–183
35. Ewing GB, Ebersberger I, Schmidt HA, von Haeseler A (2008) Rooted triple consensus and anomalous gene trees. BMC Evol Biol 8:118
36. Degnan JH, DeGiorgio M, Bryant D, Rosenberg NA (2009) Properties of consensus methods for inferring species trees from gene trees. Syst Biol 58:35–54
37. Ranwez V, Criscuolo A, Douzery EJ (2010) SuperTriplets: a triplet-based supertree approach to phylogenomics. Bioinformatics 26:i115–i123
38. Steel M, Rodrigo A (2008) Maximum likelihood supertrees. Syst Biol 57:243–250
39. Wiens JJ (2003) Missing data, incomplete taxa, and phylogenetic accuracy. Syst Biol 52:528–538
40. Gadagkar SR, Rosenberg MS, Kumar S (2005) Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. J Exp Zool B Mol Dev Evol 304:64–74
41. Bull JJ, Huelsenbeck JP, Cunningham CW, Swofford DL, Waddell PJ (1993) Partitioning and combining data in phylogenetic analysis. Syst Biol 42:384–397
42. Rokas A, Williams B, King N, Carroll S (2003) Genome-scale approaches to resolving

incongruence in molecular phylogenies. Nature 425:798–804
43. Driskell AC, Ane C, Burleigh JG, McMahon MM, O'Meara BC, Sanderson MJ (2004) Prospects for building the tree of life from large sequence databases. Science 306:1172–1174
44. Rokas A (2006) Genomics. Genomics and the tree of life. Science 313:1897–1899
45. Kubatko LS, Degnan JH (2007) Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst Biol 56:17–24
46. Wu M, Eisen JA (2008) A simple, fast, and accurate method of phylogenomic inference. Genome Biol 9:R151
47. Edwards SV et al (2016) Implementing and testing the multispecies coalescent model: a valuable paradigm for phylogenomics. Mol Phylogenet Evol 94:447–462
48. Rannala B, Yang Z (2003) Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164:1645–1656
49. Bravo GA et al (2019) Embracing heterogeneity: coalescing the Tree of Life and the future of phylogenomics. PeerJ 7:e6399. https://doi.org/10.7717/peerj.6399
50. Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–376
51. Rannala B, Yang ZH (2008) Phylogenetic inference using whole genomes. Annu Rev Genomics Hum Genet 9:217–231
52. Xu B, Yang Z (2016) Challenges in species tree estimation under the multispecies coalescent model. Genetics 204:1353–1368
53. Liu L, Xi Z, Wu S, Davis CC, Edwards SV (2015) Estimating phylogenetic trees from genome-scale data. Ann N Y Acad Sci 1360:36–53
54. Edwards SV (2016) Inferring species trees. In: Kliman R (ed) Encyclopedia of evolutionary biology. Elsevier Inc., New York, pp 236–244
55. Castillo-Ramírez S, Liu L, Pearl D, Edwardsm SV (2010) Bayesian estimation of species trees: a practical guide to optimal sampling and analysis. In: Knowles LL, Kubatko LS (eds) Estimating species trees: practical and theoretical aspects. Wiley-Blackwell, New Jersey, pp 15–33
56. Liu L, Yu L, Edwards S (2010) A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol Biol 10:302
57. Mirarab S, Warnow T (2015) ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31:i44–i52
58. Liu L, Pearl DK (2007) Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst Biol 56:504–514
59. Ogilvie HA, Bouckaert RR, Drummond AJ (2017) StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. Mol Biol Evol 34(8):2101–2114
60. Heled J, Drummond AJ (2010) Bayesian inference of species trees from multilocus data. Mol Biol Evol 27:570–580
61. Maddison WP, Knowles LL (2006) Inferring phylogeny despite incomplete lineage sorting. Syst Biol 55:21–30
62. Kubatko LS, Carstens BC, Knowles LL (2009) STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25:971–973
63. O'Meara BC (2010) New heuristic methods for joint species delimitation and species tree inference. Syst Biol 59:59–73
64. Mossel E, Roch S (2010) Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. IEEE/ACM Trans Comput Biol Bioinform 7:166–171
65. Liu L, Yu L, Pearl DK, Edwards SV (2009) Estimating species phylogenies using coalescence times among sequences. Syst Biol 58:468–477
66. Liu L, Yu L (2011) Estimating species trees from unrooted gene trees. Syst Biol 60:661–667
67. Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T (2014) ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30:i541–i548
68. Liu L (2008) BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics 24:2542–2543
69. Liu L, Pearl DK, Brumfield RT, Edwards SV (2008) Estimating species trees using multiple-allele DNA sequence data. Evolution 62:2080–2091
70. Rannala B, Yang Z (2017) Efficient Bayesian species tree inference under the multispecies coalescent. Syst Biol 66:823–842
71. Yang Z (2015) The BPP program for species tree estimation and species delimitation. Curr Zool 61:854–865

72. Yang Z, Rannala B (2010) Bayesian species delimitation using multilocus sequence data. Proc Natl Acad Sci U S A 107:9264–9269
73. Chifman, J Kubatko L (2014) Quartet inference from SNP data under the coalescent model. Bioinformatics 30:3317–3324
74. Liu L, Xi ZX, Davis CC (2015) Coalescent methods are robust to the simultaneous effects of long branches and incomplete lineage sorting. Mol Biol Evol 32:791–805
75. Xi ZX, Liu L, Davis CC (2016) The impact of missing data on species tree estimation. Mol Biol Evol 33:838–860
76. Shen X-X, Hittinger CT, Rokas A (2017) Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nat Ecol Evol 1:0126
77. Suzuki Y, Glazko GV, Nei M (2002) Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc Natl Acad Sci U S A 99:16138–16143
78. Huang HT, He QI, Kubatko LS, Knowles LL (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
79. Jarvis ED et al (2014) Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346:1320–1331
80. Hughes LC et al (2018) Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. Proc Natl Acad Sci U S A 115:6249–6254
81. Wickett NJ et al (2014) Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc Natl Acad Sci U S A 111: E4859–E4868
82. Avise JC, Ball RMJ (1990) Principles of genealogical concordance in species concepts and biological taxonomy. Oxf Surv Evol Biol 7:45–67
83. Solis-Lemus C, Yang M, Ane C (2016) Inconsistency of species tree methods under gene flow. Syst Biol 65:843–851
84. Stenz NW, Larget B, Baum DA, Ane 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
85. Hudson RR, Coyne JA (2002) Mathematical consequences of the genealogical species concept. Evolution 56:1557–1565
86. Tobias JA, Seddon N, Spottiswoode CN, Pilgrim JD, Fishpool LDC, Collar NJ (2010) Quantitative criteria for species delimitation. Ibis 152:724–746
87. Jackson ND, Carstens BC, Morales AE, O'Meara BC (2017) Species delimitation with gene flow. Syst Biol 66:799–812
88. Leache AD, Zhu T, Rannala B, Yang Z (2018) The spectre of too many species. Syst Biol 66:379
89. Solis-Lemus C, Knowles LL, Ane C (2015) Bayesian species delimitation combining multiple genes and traits in a unified framework. Evolution 69:492–507
90. Sukumaran J, Knowles LL (2017) Multispecies coalescent delimits structure, not species. Proc Natl Acad Sci U S A 114:1607–1612
91. Carstens BC, Pelletier TA, Reid NM, Satler JD (2013) How to fail at species delimitation. Mol Ecol 22:4369–4383
92. Carstens BC, Dewey TA (2010) Species delimitation using a combined coalescent and information-theoretic approach: an example from North American Myotis bats. Syst Biol 59:400–414
93. De Queiroz K (2007) Species concepts and species delimitation. Syst Biol 56:879–886
94. Pinho C, Hey J (2010) Divergence with gene flow: models and data. Annu Rev Ecol Evol Syst 41:215–230
95. Hey J, Nielsen R (2007) Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. Proc Natl Acad Sci U S A 104:2785–2790
96. Carstens BC, Morales AE, Jackson ND, O'Meara BC (2017) Objective choice of phylogeographic models. Mol Phylogenet Evol 116:136–140
97. Wakeley J (2001) The effects of subdivision on the genetic divergence of populations and species. Evolution 54:1092–1101
98. Solís-Lemus C, Yang M, Ané C (2016) Inconsistency of species tree methods under gene flow. Syst Biol 65:843–851
99. Eckert AJ, Carstens BC (2008) Does gene flow destroy phylogenetic signal? The performance of three methods for estimating species phylogenies in the presence of gene flow. Mol Phylogenet Evol 49(3):832–842
100. Doolittle WF, Bapteste E (2007) Pattern pluralism and the Tree of Life hypothesis. Proc Natl Acad Sci U S A 104:2043–2049
101. Boto L (2010) Horizontal gene transfer in evolution: facts and challenges. Proc Biol Sci 277:819–827
102. Soucy SM, Huang J, Gogarten JP (2015) Horizontal gene transfer: building the web of life. Nat Rev Genet 16:472–482

103. Keeling PJ, Palmer JD (2008) Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet 9:605–618
104. Soanes D, Richards TA (2014) Horizontal gene transfer in eukaryotic plant pathogens. Annu Rev Phytopathol 52:583–614
105. Thomas J, Schaack S, Pritham EJ (2010) Pervasive horizontal transfer of rolling-circle transposons among animals. Genome Biol Evol 2:656–664
106. Mallet J, Besansky N, Hahn MW (2015) How reticulated are species? BioEssays 38:140–149
107. Huang J, Gogarten JP (2006) Ancient horizontal gene transfer can benefit phylogenetic reconstruction. Trends Genet 22:361–366
108. Davidson R, Vachaspati P, Mirarab S, Warnow T (2015) Phylogenomic species tree estimation in the presence of incomplete lineage sorting and horizontal gene transfer. BMC Genomics 16(Suppl 10):S1
109. Linz S, Semple C, Stadler T (2010) Analyzing and reconstructing reticulation networks under timing constraints. J Math Biol 61:715–737
110. Linz S, Radtke A, von Haeseler A (2007) A likelihood framework to measure horizontal gene transfer. Mol Biol Evol 24:1312–1319
111. Rasmussen MD, Kellis M (2011) A Bayesian approach for fast and accurate gene tree reconstruction. Mol Biol Evol 28:273–290
112. Rasmussen MD, Kellis M (2007) Accurate gene-tree reconstruction by learning gene- and species-specific substitution rates across multiple complete genomes. Genome Res 17:1932–1942
113. Szöllosi GJ, Tannier E, Daubin V, Boussau B (2015) The inference of gene trees with species trees. Syst Biol 64:e42–e62
114. Sanderson MJ, McMahon MM (2007) Inferring angiosperm phylogeny from EST data with widespread gene duplication. BMC Evol Biol 7(Suppl 1):S3
115. Thomas PD (2010) GIGA: a simple, efficient algorithm for gene tree inference in the genomic age. BMC Bioinform 11:312
116. Boussau B, Szollosi GJ, Duret L, Gouy M, Tannier E, Daubin V (2013) Genome-scale coestimation of species and gene trees. Genome Res 23:323–330
117. Conte MG, Gaillard S, Droc G, Perin C (2008) Phylogenomics of plant genomes: a methodology for genome-wide searches for orthologs in plants. BMC Genomics 9:183
118. Altenhoff AM, Gil M, Gonnet GH, Dessimoz C (2013) Inferring hierarchical orthologous groups from orthologous gene pairs. PLoS One 8:e53786
119. Irisarri I et al (2017) Phylotranscriptomic consolidation of the jawed vertebrate timetree. Nat Ecol Evol 1:1370–1378
120. Edwards SV (2009) Natural selection and phylogenetic analysis. Proc Natl Acad Sci U S A 106:8799–8800
121. Ray N, Excoffier L (2009) Inferring past demography using spatially explicit population genetic models. Hum Biol 81:141–157
122. Castoe TA, de Koning APJ, Kim H-M, Gu W, Noonan BP, Naylor G, Jiang ZJ, Parkinson CL, Pollock DD (2009) Evidence for an ancient adaptive episode of convergent molecular evolution. Proc Natl Acad Sci U S A 106:8986–8991
123. Reid NH, Hird SM, Brown JM, Pelletier TA, McVay JD, Satler JD, Carstens BC (2014) Poor fit to the multispecies coalescent is widely detectable in empirical data. Syst Biol 63:322–333
124. Edwards SV, Liu L, Pearl DK (2007) High-resolution species trees without concatenation. Proc Natl Acad Sci U S A 104:5936–5941
125. Beerli P, Felsenstein J (2001) Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. Proc Natl Acad Sci U S A 98:4563–4568
126. Hey J, Wakeley J (1997) A coalescent estimator of the population recombination rate. Genetics 145:833–846
127. Solis-Lemus C, Bastide P, Ane C (2017) PhyloNetworks: a package for phylogenetic networks. Mol Biol Evol 34:3292–3298
128. Than C, Ruths D, Nakhleh L (2008) PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinform 9:322
129. Hallström BM, Janke A (2010) Mammalian evolution may not be strictly bifurcating. Mol Biol Evol 27:2804–2816
130. Kutschera VE, Bidon T, Hailer F, Rodi JL, Fain SR, Janke A (2014) Bears in a forest of gene trees: phylogenetic inference is complicated by incomplete lineage sorting and gene flow. Mol Biol Evol 31:2004–2017
131. Mavárez J, Salazar CA, Bermingham E, Salcedo C, Jiggins CD, Linares M (2006) Speciation by hybridization in Heliconius butterflies. Nature 441:868–871
132. Wen D, Nakhleh L (2017) Coestimating reticulate phylogenies and gene trees from multilocus sequence data. Syst Biol 67 (3):439–457

133. Yu Y, Nakhleh L (2015) A maximum pseudo-likelihood approach for phylogenetic networks. BMC Genomics 16(Suppl 10):S10
134. Stenz NW, Larget B, Baum DA, 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(5):809–823
135. Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. Mol Ecol 13:969–980
136. Storz JF (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. Mol Ecol 14:671–688
137. Barrett RDH, Hoekstra HE (2011) Molecular spandrels: tests of adaptation at the genetic level. Nat Rev Genet 12:767–780
138. Swofford DL (1991) When are phylogeny estimates from molecular and morphological data incongruent? In: Miyamoto MM, Cracraft J (eds) Phylogenetic analysis of DNA sequences. Oxford University Press, Oxford, pp 295–333
139. Dufraigne C, Fertil B, Lespinats S, Giron A, Deschavanne P (2005) Detection and characterization of horizontal transfers in prokaryotes using genomic signature. Nucleic Acids Res 33:e6
140. Roettger M, Martin W, Dagan T (2009) A machine-learning approach reveals that alignment properties alone can accurately predict inference of lateral gene transfer from discordant phylogenies. Mol Biol Evol 26:1931–1939
141. Zimmermann T, Mirarab S, Warnow T (2014) BBCA: improving the scalability of *BEAST using random binning. BMC Genomics 15(Suppl 6):S11
142. Liu L, Yu L (2010) Phybase: an R package for species tree analysis. Bioinformatics 26:962–963
143. Angelis K, dos Reis M (2015) The impact of ancestral population size and incomplete lineage sorting on Bayesian estimation of species divergence times. Curr Zool 61:874–885
144. Edwards SV, Beerli P (2000) Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. Evolution 54:1839–1854
145. Stadler T (2011) Simulating trees with a fixed number of extant species. Syst Biol 60:676–684
146. Papadantonakis S, Poirazi P, Pavlidis P (2016) CoMuS: simulating coalescent histories and polymorphic data from multiple species. Mol Ecol Resour 16:1435–1448
147. Anderson CNK, Ramakrishnan U, Chan YL, Hadly EA (2005) Serial SimCoal: a population genetics model for data from multiple populations and points in time. Bioinformatics 21:1733–1734
148. Excoffier L, Foll M (2011) fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. Bioinformatics 27:1332–1334
149. Robinson DF, Foulds LR (1981) Comparison of phylogenetic trees. Math Biosci 53:131–147
150. Song S, Liu L, Edwards SV, Wu SY (2012) Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proc Natl Acad Sci U S A 109:14942–14947
151. Heled J, Bryant D, Drummond AJ (2013) Simulating gene trees under the multispecies coalescent and time-dependent migration. BMC Evol Biol 13:44
152. Edwards SV, Potter S, Schmitt CJ, Bragg JG, Moritz C (2016) Reticulation, divergence, and the phylogeography–phylogenetics continuum. Proc Natl Acad Sci U S A 113:8025–8032
153. Weyenberg G, Huggins PM, Schardl CL, Howe DK, Yoshida R (2014) KDETREES: non-parametric estimation of phylogenetic tree distributions. Bioinformatics 30:2280–2287
154. Gaither, J Kubatko L (2016) Hypothesis tests for phylogenetic quartets, with applications to coalescent-based species tree inference. J Theor Biol 408:179–186
155. McVay JD, Carstens BC (2013) Phylogenetic model choice: justifying a species tree or concatenation analysis. J Phylogenet Evol Biol 1:114
156. Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791
157. Lemoine F, Domelevo Entfellner JB, Wilkinson E, Correia D, Davila Felipe M, De Oliveira T, Gascuel O (2018) Renewing Felsenstein's phylogenetic bootstrap in the era of big data. Nature 556:452–456
158. Seo TK (2008) Calculating bootstrap probabilities of phylogeny using multilocus sequence data. Mol Biol Evol 25:960–971
159. Sayyari E, Mirarab S (2016) Fast coalescent-based computation of local branch support from quartet frequencies. Mol Biol Evol 33:1654–1668
160. Liu L et al (2017) Reply to Gatesy and Springer: Claims of homology errors and

zombie lineages do not compromise the dating of placental diversification. Proc Natl Acad Sci U S A 114:E9433–E9434

161. Edwards SV (2016) Phylogenomic subsampling: a brief review. Zool Scr 45:63–74
162. Lee JY, Joseph L, Edwards SV (2012) A species tree for the Australo-Papuan Fairy-wrens and Allies (Aves: Maluridae). Syst Biol 61:253–271
163. Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313
164. Xi Z, Liu L, Davis CC (2015) Genes with minimal phylogenetic information are problematic for coalescent analyses when gene tree estimation is biased. Mol Phylogenet Evol 92:63–71
165. Guindon S, Dufayard JF, Hordijk W, Lefort V, Gascuel O (2009) PhyML: fast and accurate phylogeny reconstruction by maximum likelihood. Infect Genet Evol 9:384–385
166. Leaché AD, Oaks JR (2017) The utility of single nucleotide polymorphism (SNP) data in phylogenetics. Annu Rev Ecol Evol Syst 48:69–84
167. Pease JB, Haak DC, Hahn MW, Moyle LC (2016) Phylogenomics reveals three sources of adaptive variation during a rapid radiation. PLoS Biol 14:e1002379
168. Hahn MW, Nakhleh L (2016) Irrational exuberance for resolved species trees. Evolution 70:7–17
169. Jennings WB, Edwards SV (2005) Speciational history of Australian grass finches (*Poephila*) inferred from 30 gene trees. Evolution 59:2033–2047
170. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.