Trends in
Genetics

CellPress

Review

New Approaches for Inferring Phylogenies in the Presence of Paralogs

Megan L. Smith[1],* and Matthew W. Hahn[1]

The availability of whole genome sequences was expected to supply essentially unlimited data for phylogenetics. However, strict reliance on single-copy genes for this purpose has drastically limited the amount of data that can be used. Here, we review several approaches for increasing the amount of data used for phylogenetic inference, focusing on methods that allow for the inclusion of duplicated genes (paralogs). Recently developed methods that are robust to high levels of incomplete lineage sorting also appear to be robust to the inclusion of paralogs, suggesting a promising way to take full advantage of genomic data. We discuss the pitfalls of these approaches, as well as further avenues for research.

Highlights

Despite the increased availability of whole genome sequences, the data available for phylogenetic studies are extremely limited. This is because only single-copy genes present in most sampled species are used to infer phylogenies. In this review, we discuss several approaches for increasing the amount of data that can be used in phylogenetic inference.

Recent work suggests that the inclusion of loci missing data for some taxa should not mislead phylogenetic inference with several popular methods.

Even if orthologs are required, researchers need not limit themselves to single-copy orthologs, as paralogs specific to a single lineage or to a pair of sister lineages should not lead to topological errors in any approach to phylogeny inference.

Several recent methods for species-tree inference that are robust to high levels of incomplete lineage sorting also appear to be robust to the inclusion of paralogs.

The Search for Orthologs

The business of phylogeny-building has been transformed by the availability of whole genome sequences (reviewed in [1]). Indeed, the promise of 'phylogenomics' was access to many thousands of loci [2]. However, the data requirements of most phylogenetic inference methods – single-copy genes present in almost all species sampled (Figure 1A, Key Figure) – have meant that a growing number of phylogenomic studies have actually used very small amounts of data. For instance, in their dataset of 76 arthropod genomes, Thomas *et al.* [3] found no genes that were single-copy and present in all species. This study is not unique: even with whole-genome data, as the number of species sampled goes up, the number of single-copy genes found in all taxa goes down [4].

Phylogeny estimation has long relied on the identification of single-copy orthologous genes, filtering out paralogous genes found in multiple copies in one or more species (Box 1). Indeed, when Fitch [5] introduced the terms **orthologs** and **paralogs** (see Glossary), it was in the context of species phylogeny estimation: 'Phylogenies require orthologous, not paralogous, genes.' This sentiment is echoed repeatedly in the literature [6,7], based on the belief that, since orthologous genes are related by speciation events alone, their relationships should more accurately reflect the species phylogeny. Similar claims are made about the privileged use of orthologs in protein function prediction [8–10].

However, algorithms that enable species trees inference using both orthologs and paralogs were proposed more than 40 years ago [11], and efficient software implementing these approaches has been around for at least 20 years [12]. Methods using orthologs and paralogs work because gene trees containing duplication events also include all of the speciation events that follow (Figure 1B). While each duplication event does add a branch not found in the species tree, it also doubles the amount of information contained about subsequent speciation events (in the absence of subsequent losses). Most significantly, recent methods developed for phylogeny inference using orthologs [13,14] turn out to be highly accurate and extremely efficient when applied to datasets including paralogs [15,16]. Although the application of these approaches to such datasets is just beginning, their promise for phylogenomics is clear.

[1]Department of Biology and Department of Computer Science, Indiana University, Bloomington, IN 47405, USA

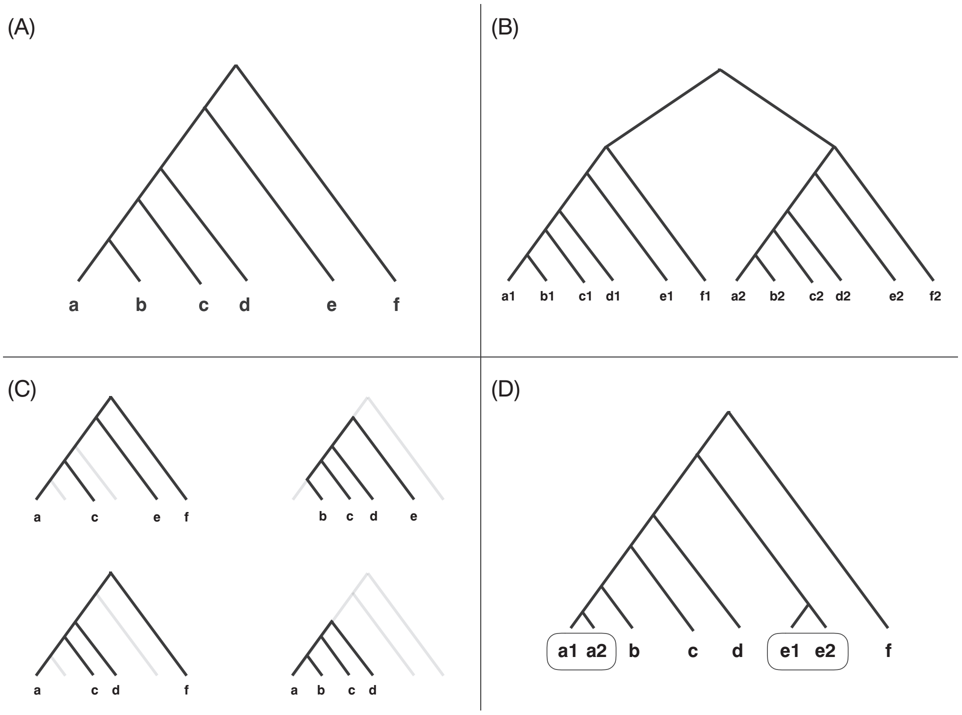
*Correspondence:
mls16@indiana.edu (M.L. Smith).

https://doi.org/10.1016/j.tig.2020.08.012

© 2020 Elsevier Ltd. All rights reserved.

Key Figure

Sampling Orthologs and Paralogs



Trends in Genetics

Figure 1. There are several potential sampling strategies in phylogenetic inference. Here, we illustrate a few of these, although these categories are not mutually exclusive. (A) Phylogenies can be constructed from complete sampling of single-copy orthologs. (B) Phylogenies can be reconstructed from sets of paralogs. The tree shown has a single duplication event in the ancestor of all species. (C) Phylogenies can be constructed from genes with missing data, either owing to incomplete sampling or to gene loss. (D) Phylogenies can be constructed from loci with lineage-specific duplications. Duplications in lineages ‘a’ and ‘e’ result in two copies in each of these species in the tree shown. Sampling a single copy from each species should not affect phylogenetic inference.

In this review, we discuss ways to combat the limitation of single-copy orthologs by increasing the amount of data that can be used in phylogenomics, while still maintaining a high degree of accuracy. We first discuss the problem of **gene tree heterogeneity**, and how it affects the accuracy of species trees. Next, we review two broad approaches for increasing the amount of data used in phylogenomic inference: one that still includes only orthologs and one that includes both orthologs and paralogs. We also describe the newly developed phylogenetic methods that make both of these approaches possible. Finally, we identify some key topics to consider when inferring phylogenies in the presence of paralogs, including promising future areas of research on this topic.

Gene Tree Heterogeneity and the Problem of ‘Hidden Paralogy’

Gene tree heterogeneity is now recognized as common in phylogenetics [17]. Topological heterogeneity may be due to a number of biological factors, including **incomplete lineage sorting (ILS)**, **introgression**, and **gene duplication and loss (GDL)** [18], in addition to technical factors such as error in gene tree reconstruction. This heterogeneity has important consequences for species tree inference, as if it is not accounted for it can lead to an incorrect phylogeny. Methods

Glossary

Bayesian inference: in phylogenetic inference, an approach to estimate the posterior distribution of tree topologies and branch lengths.

Gene duplication and loss (GDL): the process by which genes are duplicated and lost. Loss can occur with or without a previous duplication.

Gene tree heterogeneity: a mismatch between the topology of a single region and the topology of a species, or between different genomic regions.

Homologs: genes that share a common ancestor.

Incomplete lineage sorting (ILS): the failure of two lineages to coalesce within a population, which may lead to gene trees that disagree with the species tree.

Introgression: gene flow between divergent lineages, referred to as horizontal gene transfer in asexual species.

Orthologs: homologous genes that share a common ancestor owing to speciation.

Paralogs: homologous genes that share a common ancestor owing to duplication.

Pseudo-orthologs (or hidden paralogs): regions with a history of duplication for which only a single copy is retained per species owing to differential loss of duplicate copies across species.

Quartet-based methods: algorithms for inferring species trees from a collection of unrooted four-taxon (or rooted three-taxon) trees.

Statistically consistent: A method is statistically consistent under a particular model if, when given an unlimited amount of data, it would arrive at the correct answer under the model.

Box 1. Types of Homologous Relationships and Implications for Phylogenetic Inference

Homologous loci share a common ancestor. Orthologous loci share a common ancestor owing to speciation (e.g., a1 and b1 in Figure I), while paralogous loci share a common ancestor owing to duplication (e.g., a1 and a3; [5]). Orthology relationships can be classified as 'one-to-one', 'one-to-many', and 'many-to-many' based on whether speciation was followed by duplication in neither, either, or both lineages [115]. For example, b1 and c1, are 'one-to-one' orthologs. These are the orthologs that are typically used in phylogenetic inference. Specifically, researchers target single-copy orthologs, which exist in only a single copy in all species considered. However, 'many-to-one' or 'many-to-many' orthologs may also be useful. Since the duplication event leading to paralogs a1 and a3 occurred after the speciation event with b1, they have a 'many-to-one' orthologous relationship. Such lineage-specific duplications should not affect phylogenetic inference because a1 and a3 are 'co-orthologous' to b1 and c1, meaning that either copy has an orthologous relationship with b1 and c1. Similarly, a2 and a4 have a 'many-to-one' orthologous relation to c2. The large numbers of complex 'many-to-many' relationships that can arise (for instance, the relationship between a1, a2, e1, and e2 in Figure 1D in main text) make ortholog group delimitation a difficult task, although these loci can still be used in many types of phylogenetic inference.

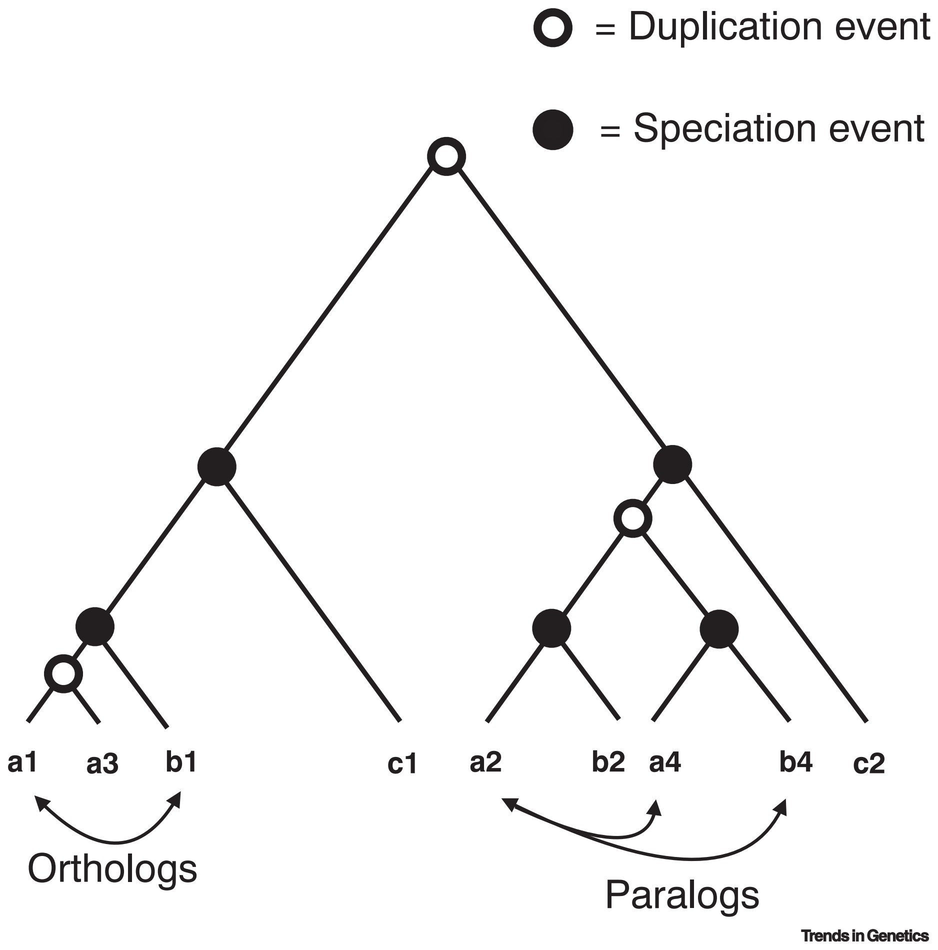
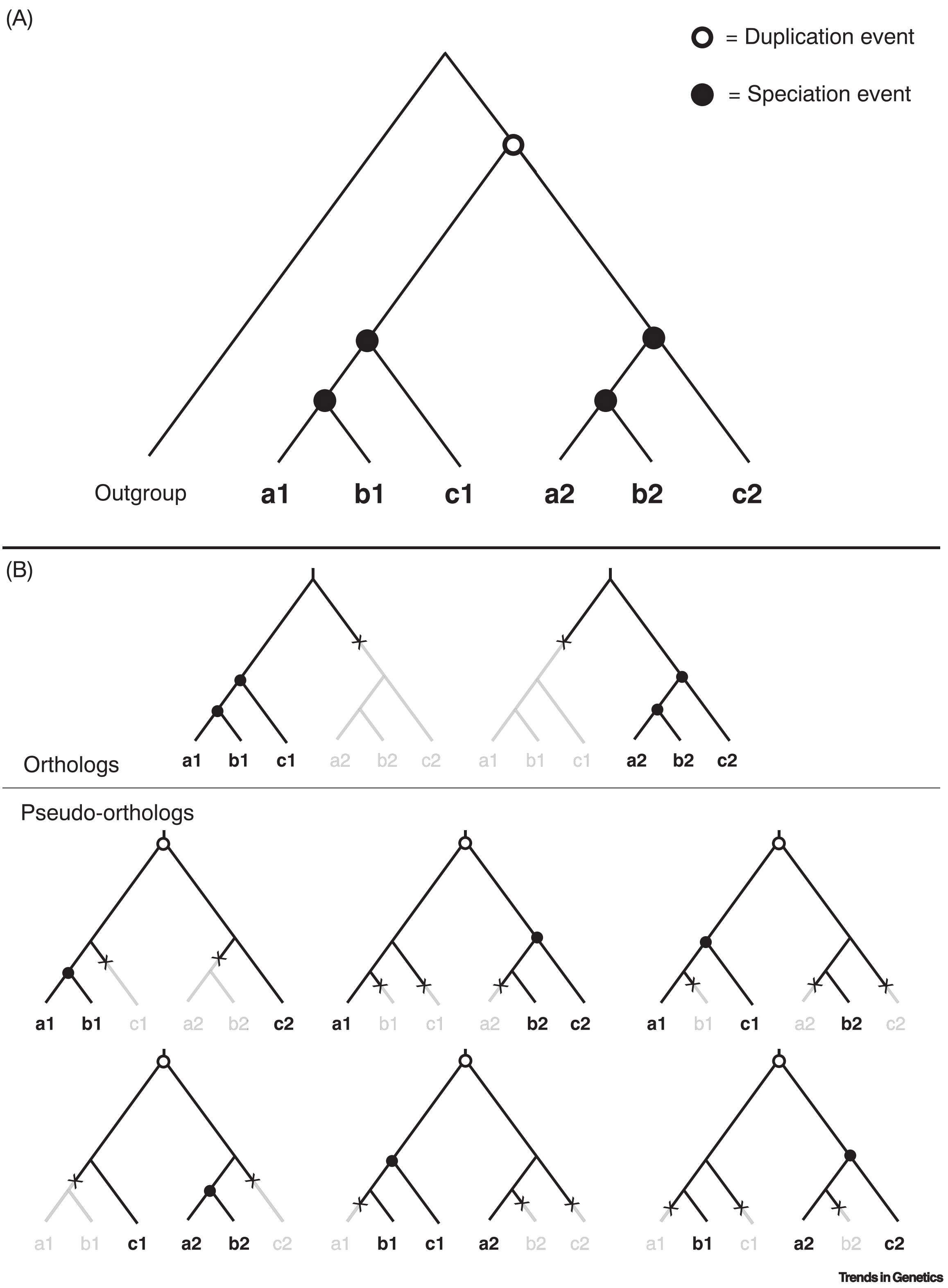


Figure I. Types of Homologous Relationships.

developed to deal with multiple causes of heterogeneity can also help us to infer phylogenies from a broader set of loci.

In particular, high levels of ILS can mislead many species tree methods, whether they apply maximum likelihood methods to concatenated alignments of all loci [19,20] or use the most common



(See figure legend at the bottom of the next page.)

gene tree topology as an estimate of the species phylogeny [21]. Partly because of these issues, methods that account for ILS when estimating species phylogenies have proliferated [13,14,22–27]. Several of these methods require that gene trees be estimated separately for each locus and then combine these in a principled way to infer a species tree [13,14,22,24,27], while others either bypass gene tree reconstruction [23,26] or jointly infer gene and species trees [25]. As with most phylogenetic approaches, these methods were initially designed to use datasets consisting of only single-copy orthologs, as they were intended to account only for ILS as a source of gene tree heterogeneity. Importantly, however, many of these methods also deal naturally with missing data; this will be key for several of the new approaches described below.

Gene duplication and loss leads to gene tree heterogeneity by adding duplication events to gene trees (Box 1). Such events are not expected in histories that follow only the species tree, so trees that contain more than one copy of a gene are generally removed from phylogenetic datasets. More insidiously, '**hidden paralogs**' [28], or '**pseudo-orthologs**' [29], contain only a single copy per species owing to differential loss of duplicate copies across species (Figure 2) and can be mistaken for single-copy orthologs. The topologies inferred from pseudo-orthologs can differ from the species tree via a process that is rarely modeled by phylogenetic methods.

Although they are much feared, few studies have actually evaluated the effects of including pseudo-orthologs on phylogenetic inference, and these found mixed results. Brown and Thomson [30] suggested that outlier loci supporting a contentious placement of turtles were paralogs, and that these had an extreme effect on **Bayesian inference** applied to a concatenated dataset. Many other studies have shown differences in the species tree inferred from datasets assembled using different orthology detection tools, differences that are possibly due to the inclusion of pseudo-orthologs (reviewed in [6]). Some of these studies found substantial differences in the inferred trees [31], while others found minimal effects [32,33].

What is clear from the work briefly summarized here is that there are many causes of gene tree heterogeneity that have the capacity to mislead phylogenetic inference. With respect to increasing the types of loci that can be used in phylogenomics, we would like any approaches using these loci to be robust to the known problems caused by both ILS and hidden paralogy.

Increasing Data Availability Without Including Paralogs

If only orthologous genes are required, there are multiple ways to increase the total number of loci used in phylogenetic inference. Here we discuss two such approaches that can increase the amount of available data: relaxing filters for missing data (Figure 1C) and sampling lineage-specific duplicates (Figure 1D).

Sampling Single-Copy Orthologs with Missing Data

Often, researchers require that all or most of their taxa are sampled for a locus to be included in phylogenetic inference. However, the actual effects of including missing data – that is loci for

Figure 2. Orthologs, Pseudo-orthologs, and Quartet Frequencies. (A) The full history of a locus in three species and an outgroup, including one duplication event and two speciation events (which are shown separately for each set of orthologs). (B) Scenarios where only a single gene copy is sampled per species; the outgroup is assumed to be sampled in each, but is not shown for clarity. The single copies may be present because of gene losses (shown here as Xs), or simply because a single copy is randomly chosen per species. The latter case is also what would happen if there were no missing copies but quartets were sampled from the full gene tree as input to ASTRAL [15,16]. There are four quartets that match the species tree: the two orthologs and the two left-most pseudo-orthologs ('hidden paralogs'). The remaining pseudo-orthologs either place lineages 'b' and 'c' sister to one another (center) or 'a' and 'c' sister to one another (right). Therefore, quartet methods should perform well even when paralogs are included, because the most common set of relationships should still match the species tree. Note that if genes are single-copy because of gene losses, the species tree relationship is likely to become even more common: the orthologs require only one loss in their history and the matching pseudo-orthologs require two losses. Pseudo-orthologs not matching the species tree can only be generated when there are three separate loss events.

which no sequences exist in one or more species – remain unclear. In concatenated analyses, simulation studies have demonstrated that there are limited negative effects of missing data [34]. Other studies have argued that the issue is a lack of informative data rather than missing data *per se* [35,36]. Many empirical studies show little effect of including missing data [37,38], and often the positive effects of including a larger number of loci or sites seem to outweigh the negative effects of missing data [39,40].

There has been a lot of recent work on the effects of missing data on gene-tree-based methods that can account for ILS [14,22,24,27]. Because these methods combine individual gene trees from each locus, they can naturally accommodate missing taxa in a subset of trees. Studies have shown that ILS methods can be robust to substantial levels of missing data, whether these are randomly or nonrandomly distributed [41–43]. Note, however, that these results may break down in cases of extreme branch lengths [44,45].

Based on these considerations, one simple way to drastically increase the amount of data that can be used for phylogenetic inference is to relax missing data thresholds. For **quartet-based methods** such as ASTRAL [13], the minimum number of taxa required from each locus is four (Figure 1C), as a four-taxon unrooted tree is all that is needed to specify phylogenetic relationships. Empirically, results of relaxing these thresholds can be dramatic. For example, Eaton and Ree [46] found that requiring a minimum of four taxa increased the number of loci available in a group of flowering plants nearly ninefold compared with requiring that all taxa be sampled. The relative advantage gained by using these methods can only go up as more taxa are included in a dataset, although researchers should try to ensure that species are represented approximately evenly across loci to avoid cases where most of the signal for some branches comes from a small number of genes (e.g., [47]).

Sampling Orthologs that Have Lineage-Specific Duplication Events

The requirement that only orthologs be sampled for phylogenetic inference does not mean that we must only include single-copy orthologs. Notably, there is no theoretical reason to exclude loci that have undergone lineage-specific duplications, as they can have many-to-one orthologous relationships with single-copy genes (Box 1). For example, in Figure 1D, species-specific duplications have occurred in lineages a and e. Since the two copies in each species are both orthologous to the gene copies in all other lineages, if we chose a single gene from each species the resulting gene tree would include only speciation events. There can be no gene tree heterogeneity induced by such a sampling scheme, even when there are more than two copies in each species.

Surprisingly, this approach has rarely been used in phylogenetics research. The number of loci that could be included would greatly increase, but the computational burden would increase slightly, as well (Box 2). These numbers could be increased even further, too: there should be no negative effect on the inferred topology of including duplications specific to a pair of sister species. In other words, if one or more duplication events occur in the ancestor of a pair of species, sampling a single copy from each of these species cannot induce gene tree heterogeneity. This occurs because there is only a single way this pair can be related, while such gene tree invariance cannot be ensured for duplicates ancestral to three or more species. Although the inclusion of duplicates specific to a pair of sister species should not affect the inferred topology, it could affect estimates of terminal branch lengths (see section on Branch Lengths later). Broadening sampling to include these genes would lead to a further increase in the number of loci available for phylogenetic inference.

Box 2. Identifying Orthologous Genes and Sampling Lineage-Specific Paralogs

Owing to interest in identifying orthologs both for phylogeny reconstruction and for functional prediction, several methods for ortholog detection have been developed (reviewed in [115]). The most commonly used approaches for ortholog detection are graph-based approaches, which rely on the identification of reciprocal best hits (RBHs). These methods are based on the assumption that the two most closely related **homologs** between a pair of species should be orthologs. After RBHs are identified, some approach must be used to construct groups of orthologous sequences. One such approach, implemented in the software OrthoMCL [116], uses a Markov clustering algorithm to identify orthogroups consisting of orthologs and recent paralogs. Typically, for downstream phylogenomic inference, single-copy orthologs present in most species are extracted from these results. While lineage-specific duplicates need not be excluded from datasets for phylogenetic inference (see main text), it is not straightforward to extract these automatically from the output of most graph-based approaches. Instead, the most obvious way to identify and include these genes is by reconstructing gene trees for all orthogroups, identifying lineage-specific duplicates, and selecting one copy per species for downstream inference. Some recently introduced branch-cutting methods can also sample such genes from orthogroups containing duplicates. Yang and Smith [117] consider several different branch-cutting algorithms to extract orthologs appropriate for phylogeny estimation, showing that these methods considerably increase the number of genes available for phylogenetic inference. For example, in a Hymenoptera dataset analyzed by these authors, the number of orthologs present in at least eight taxa increased from 4937 using only single-copy-orthologs to 9128 under one branch-cutting technique [117]. Thus, even when including paralogs is not desirable, orthologs can be extracted from many datasets not traditionally considered in phylogenetic inference.

Estimating Species Trees in the Presence of Paralogs

In the methods described thus far we have still limited ourselves to analyses involving only orthologous loci. If we relax this restriction even more, we can again greatly increase the number of loci to be used. We review five general approaches for reconstructing species trees in the presence of paralogs. We largely go through these methods in the chronological order in which they appeared in the literature, spending the most time at the end on promising new methods.

Gene Tree Parsimony

The earliest methods to infer species trees in the presence of gene duplication and loss used gene tree parsimony (GTP) [11,12,48,49]. In these approaches the aim is to find the species tree with the minimum 'reconciliation' cost [50] to a collection of input gene trees; that is the species tree that minimizes the distance to all gene trees. Reconciliation costs are calculated based on explicit biological causes of gene tree heterogeneity, including, but not limited to, GDL. Some algorithms calculate reconciliation costs based on minimizing duplications and losses [11,49,51,52], while others focus completely on minimizing the number of differences induced by ILS [53,54], or allow users to choose among these reconciliation costs [55]. Recognizing that these processes do not act in isolation, recent approaches consider both GDL and ILS [56] or GDL and introgression [57], with some incorporating all three processes [58,59]. Although these approaches appear to deal with ILS, they do not completely account for very high levels of ILS when inferring the species tree [60], and therefore may give misleading results in such cases.

Robinson-Foulds-Based Methods

The Robinson-Foulds (RF) distance between two trees measures the number of branches that must be removed, and the number of subsequent branches that must be added, to make them have the same topology [61]. RF species tree methods try to find the species tree that minimizes the RF distance to a collection of input gene trees [62]. Although this is a similar approach to gene tree parsimony, RF-based approaches make no assumptions about the biological processes leading to heterogeneity between the gene trees and the species tree, and there are therefore no options to apply different costs to different processes.

Although RF-based methods as originally described were applicable only to input trees with no duplicates, interest in applying these methods to multicopy gene trees (i.e., those with both orthologs and paralogs) led to several advancements that permitted the calculation of RF distances between them [63,64]. Chaudhary *et al.* [65,66] then introduced an approach for finding a species tree using multicopy gene trees as input. Their method, MulRF, compares favorably with GTP approaches [67], and has recently been improved by Molloy and Warnow [68]. RF methods appear to perform well under general conditions [67,68], although, like GTP methods, they are not accurate under high levels of ILS [69].

Probabilistic Methods

Several stochastic models of duplication and loss have been described (e.g., [70–74], reviewed in [75]), as well as stochastic models that together consider duplication and loss and ILS [76,77] or introgression [78,79]. These models pave the way for probabilistic methods to infer a species tree. However, while they have often been used to infer gene trees (e.g., [80,81]), the development of methods to infer species trees based on these models has been much slower (but see [71,79]). One possible reason for this is that probabilistic approaches are more computationally intensive than GTP and RF methods. PHYLDOG is one such method that jointly estimates gene trees, species trees, and the number of duplications and losses under a model of GDL by maximizing their likelihood given a set of alignments [82]. However, PHYLDOG does not consider other sources of gene tree incongruence (e.g., ILS) and the computational costs are high, preventing its application to large genomic datasets [67].

De Oliveira Martins *et al.* [83] introduced ‘guenomu’, a probabilistic supertree approach to infer species trees in the presence of both ILS and GDL. Guenomu implements a hierarchical Bayesian model: it takes as input a posterior distribution of gene trees and uses a multivariate distance metric based on ILS and GDL to infer a posterior distribution of species trees. However, like PHYLDOG, guenomu is computationally intensive, and therefore neither approach truly expands the number of loci one could use in phylogenomics.

Methods Based on Neighbor Joining (and Other Clustering Approaches)

Neighbor Joining (NJ; [84]) and other distance-based approaches are popular methods for species tree inference using orthologs. Newer application of these approaches can accommodate ILS by calculating a distance matrix from a collection of gene trees inferred from separate loci, and then using NJ or another clustering algorithm to estimate a species tree from this distance matrix. Distance methods applicable to gene trees can broadly be divided into two classes: those that construct distance matrices based on sequence distances and those that construct distance matrices based on internode distances. The former approach includes the methods implemented in STEAC [85] and METAL [86]. Methods based on internode distances include STAR [85], $\mathrm{NJ}_{\mathrm{st}}$ [14], and ASTRID [24]. Distance-based approaches are **statistically consistent** under the multispecies coalescent model, meaning that, given enough data, they should return the correct species tree when ILS is the only source of discordance [86–88].

Extending distance methods to cases including paralogs is straightforward, because distance matrices can be calculated as averages over multiple samples from a species. Application to datasets containing orthologs and paralogs has already been done using $\mathrm{NJ}_{\mathrm{st}}$ [15,67] and ASTRID [16]. STAG [4] is another distance method introduced specifically to estimate species trees from multicopy gene trees, though it requires that loci have no missing species. Testing the accuracy of distance methods using orthologs and paralogs, Chaudhary *et al.* [67] found that $\mathrm{NJ}_{\mathrm{st}}$ was outperformed by methods based on GTP, RF distances, and probabilistic models. By contrast, Yan *et al.* [15] found that $\mathrm{NJ}_{\mathrm{st}}$ performed comparably with quartet-based methods, and Legried *et al.* [16] found that ASTRID had similar or higher accuracy than all other methods evaluated. Overall, distance-based methods appear to be generally accurate and efficient for inferring species trees using paralogs.

Quartet-Based Methods

Methods to build species trees from quartet sub-trees have been around for some time [89–93], but have found renewed popularity owing to the introduction of more accurate, more efficient algorithms. These methods scale well to genomic datasets and are robust to both high levels of ILS [94,95], and, as mentioned earlier, large amounts of missing data. ASTRAL [13,22,94] is among the most popular of these methods: it infers a species tree from a set of input gene trees, extracting quartets from them automatically, and finding the phylogeny that maximizes the number of shared quartet trees. ASTRAL was designed for use with single-copy orthologs, but can accommodate multiple haplotypes sampled within species (ASTRAL-multi [96]). In these cases, ASTRAL-multi effectively averages over haplotypes by sampling quartets with at most one sequence per species.

Gene trees with paralogs in them take advantage of the same sampling scheme used by ASTRAL-multi, and perform very well because the most common quartet in multicopy gene trees is still the quartet that matches the species tree (Figure 2; [15,16]). ASTRAL-multi is statistically consistent under the multispecies coalescent model [96] and a model of duplication and loss [16], and simulation studies have also demonstrated its accuracy [15,16]. Most recently, a version of the software explicitly built for the inclusion of paralogs, ASTRAL-Pro, outperformed ASTRAL-multi, MulRF, and GTP methods [69].

Quartet-based methods appear to be robust to the hidden paralog problem, as can be illustrated by an extreme example. Yan *et al.* [15] suggested that such methods should be accurate even if a single gene is randomly selected from each species for each gene tree and used as input to ASTRAL (a sampling scheme that has been referred to as 'ASTRAL-ONE' [15,16]). In such a scenario, there are more combinations of sampled genes that result in pseudo-orthologs than in true orthologs (Figure 2B). However, one-third of the pseudo-ortholog combinations match the species tree topology, and the other two-thirds are split evenly between the two alternative topologies. Because of the orthologs and the pseudo-orthologs that match the species tree, it appears that the quartet matching the species tree will be the most common [16,97]. Legried *et al.* [16] demonstrated that ASTRAL-ONE is statistically consistent under a model of duplication and loss, and Markin and Eulenstein [97] demonstrated consistency of this approach under the DLCoal model, which incorporates both ILS and duplication and loss. While statements of consistency deal with the unrealistic scenario of unlimited data, simulations show that with even a few hundred loci accurate species trees can be recovered using this approach [15]. Although the numbers of tree topologies given here only involve four species (including the outgroup) and one duplication event, they would likely hold for all larger trees since these can be deconstructed into quartets (cf., [98]). In biological scenarios involving similarly extreme gene loss, both orthologs and pseudo-orthologs matching the species tree should be more likely to be sampled because they require fewer losses to produce them than the pseudo-ortholog trees that do not match (Figure 2B). This suggests that the species tree may be even more likely to be accurately inferred using quartet methods.

Because of their relative simplicity, ease-of-use, speed, accuracy, and robustness to multiple issues that confound other phylogenetic methods, quartet methods have become a mainstay of standard phylogenetic inference using single-copy orthologs. For all of the same reasons, they are likely to become widely used when sampling both orthologs and paralogs. We also suspect that other quartet-based methods that have not yet been evaluated under the inclusion of paralogs (e.g., [23]) will perform equally well under these conditions.

Considerations When Inferring Phylogenies with Paralogs

Although many of the methods discussed here ensure accurate inference of species tree topologies when paralogs are used, there are important caveats and implications that merit specific consideration. In the following sections, we discuss several of these.

Branch Lengths

Although topology estimates should not be biased by the inclusion of paralogs, the same is not true for branch lengths. When branch lengths are estimated as substitutions per site (e.g., [99,100]), the inclusion of pseudo-orthologs will force branches to be longer than they actually are (e.g., Figure 2; [98]). Conversely, when branch lengths are estimated in coalescent units (e.g., [13,27]), the additional gene tree heterogeneity introduced by paralogs (hidden or not) will result in the underestimation of branch lengths. No matter what type of branch lengths are to be estimated, we recommend that the dataset used be restricted to orthologs. Thus, a reasonable approach would be to estimate a species tree topology using all genes, and then to estimate branch lengths on this topology with a dataset including only orthologs (allowing for sampling among species-specific paralogs; Figure 1D).

Alignment

One of the most error-prone, but underappreciated, steps in phylogenomics is alignment [101,102]. Automated alignment of thousands of loci means that many errors can creep in, especially when nonhomologous (alternative) exons are sampled from different species. Fortunately, there are good methods for identifying regions with low alignment quality (e.g., GUIDANCE2; [102]). A related problem involves deciding how to choose among lineage-specific paralogs (Figure 1D) in order to maximize alignment length while minimizing alignment error. One promising approach would be to co-opt methods designed to choose among alternative isoforms at a single locus: some of these try to pick the set of genes that are most similar in length across species to avoid the inclusion of nonhomologous exons [103]. Combining such methods with tools that identify and filter unreliable portions of alignments [102,104–107] should minimize error.

Polyploidy

Polyploidy is a special case of gene duplication and loss in which the whole genome is duplicated, and offers a particular challenge both to methods for identifying orthologs and to species tree inference. In autopolyploidy both sets of chromosomes come from the same species, and gene copies are paralogs that behave in much the same manner as the smaller duplication events described earlier. Therefore, the gene tree methods discussed here should not be misled by autopolyploidy.

Allopolyploidy occurs when the chromosome number doubles via hybridization between species; the resulting gene copies are referred to as homeologs [108]. Since gene copies found in the same allopolyploid genome are related through speciation between the parental species, homeologs are not paralogs in the traditional sense. Similarly, there is not a single bifurcating species tree that describes relationships involving allopolyploids. While this makes it difficult to evaluate the effect of including homeologs on traditional species tree inference, gene-tree-based methods should usually identify one of the two potentially correct species tree topologies (e.g., [109]).

Detecting Introgression

Much less consideration has been given to the effect of including paralogs when attempting to detect introgression. The most commonly used phylogenetic methods for detecting introgression are based on the expectation that, for any quartet of species, the two minor topologies (i.e., the topologies that do not match the species tree) should occur at the same frequency; therefore, asymmetries between topologies can provide evidence for introgression [110–113]. We suggest here that, for methods that depend on the frequencies of minor topologies to detect introgression, the inclusion of paralogs should not bias inference (see Outstanding Questions). Consider the example shown in Figure 2: as discussed earlier, the most common topology matches the species tree. However, four topologies do not match the species tree. These four potential trees all require three lineage-specific losses (one in each taxon), and should occur at equal frequency under a model of GDL in the absence of introgression, similarly to under cases without duplication. Thus, methods for detecting introgression based on asymmetry in minor topologies should perform well in the presence of paralogs. This proposal merits additional consideration, however, as does the effect of paralogs on additional methods for detecting introgression not discussed here.

Concatenation

To carry out a concatenated analysis, one gene copy must be sampled per species per locus and put into a single alignment. If the intention is to include only orthologs (whether single-copy or not), a small number of pseudo-orthologs can have an extreme, negative influence on phylogenetic relationships [30,114]. This occurs because pseudo-orthologs – some of which have topologies that do not match the species tree – have internal branches that are longer than those of true orthologs (Figure 2B), giving them more phylogenetically informative changes. To minimize these potential problems, it may in fact help to instead include all of the data, rather than attempting to include only orthologs. We imagine a sampling scheme similar to the approach taken in [15], where a single copy is randomly sampled per species (i.e., 'ASTRAL-ONE'). Not only are more underlying tree topologies guaranteed to match the species tree topology, but also the pseudo-orthologs matching the species tree have longer internal branches than those matching alternate topologies (Figure 2B). Thus, with enough data, the topology matching the species tree should be favored by concatenated analyses, even in the presence of pseudo-orthologs. While certainly not a standard phylogenetic analysis, we suggest that this may be a fruitful way forward in the future.

Concluding Remarks

Despite the massive amount of genomic data being collected across the tree of life, phylogeny inference is often restricted to a small portion of this data owing to filtering for single-copy orthologs and minimal missing data. Recent work has demonstrated that several leading methods for species tree inference perform well in the presence of paralogs, suggesting a source of additional data for phylogenomic inference. Additionally, recent work has shown that missing data may not be as much of an issue as feared. Thus, the amount of data available for phylogenomic inference may be much larger than previously thought. Future work should consider branch length estimation when paralogs are present, as well as the potential effects of paralog inclusion on inferences of introgression (see Outstanding Questions).

Outstanding Questions

What are the relative advantages and disadvantages of different approaches for inferring species trees in the presence of paralogs?

Can we use paralogs to estimate branch lengths accurately under models of duplication and loss?

How does the inclusion of paralogs in phylogenomic datasets impact the alignment process, and how well do existing filtering techniques both remove problematic regions and optimize the amount of information used?

How do various methods for detecting introgression behave when paralogs are included?

How does the inclusion of paralogs affect concatenation-based approaches to phylogeny inference? Is it necessary to filter outlier regions that may have an undue effect on inference? Do differences in gene length affect which paralogs are outliers?

Can information about gene order (synteny) be used in combination with sequences from paralogous and orthologous genes to infer species trees?

Acknowledgments

We thank Rob Lanfear, Erin Molloy, and Dannie Durand for helpful comments. This work was supported by a National Science Foundation (NSF) postdoctoral fellowship to M.L.S. (DBI-2009989) and an NSF grant to M.W.H. (DEB-1936187).

References

1. Scornavacca, C. *et al.* (2020) *Phylogenetics in the genomic era,* Open access book. https://hal.inria.fr/PGE/
2. Delsuc, F. *et al.* (2005) Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.* 6, 361–375
3. Thomas, G.W.C. *et al.* (2020) Gene content evolution in the arthropods. *Genome Biol.* 21, 15
4. Emms, D.M. and Kelly, S. (2018) STAG: species tree inference from all genes. *bioRxiv.* Published online February 19, 2018. https://doi.org/10.1101/267914
5. Fitch, W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.* 19, 99–113
6. Fernández, R. *et al.* (2020) Orthology: definitions, prediction, and impact on species phylogeny inference. In *Phylogenetics in the Genomic Era* (Scornavacca, C. *et al.*, eds), pp. 2.4: 1–2.4:14. https://hal.inria.fr/PGE/
7. Kapli, P. *et al.* (2020) Phylogenetic tree building in the genomic age. *Nat. Rev. Genet.* 21, 428–444
8. Nehrt, N.L. *et al.* (2011) Testing the ortholog conjecture with comparative functional genomic data from mammals. *PLoS Comput. Biol.* 7, e1002073
9. Studer, R.A. and Robinson-Rechavi, M. (2009) How confident can we be that orthologs are similar, but paralogs differ? *Trends Genet.* 25, 210–216
10. Stamboulian, M. *et al.* (2020) The ortholog conjecture revisited: The value of orthologs and paralogs in function prediction. *Bioinformatics* 36, i219–i226

11. Goodman, M. *et al.* (1979) Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst. Biol.* 28, 132–163

12. Page, R.D. (1998) GeneTree: comparing gene and species phylogenies using reconciled trees. *Bioinformatics* 14, 819–820

13. Zhang, C. *et al.* (2018) ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* 19, 153

14. Liu, L. and Yu, L. (2011) Estimating species trees from unrooted gene trees. *Syst. Biol.* 60, 661–667

15. Yan, Z. *et al.* (2018) Species tree inference under the multispecies coalescent on data with paralogs is accurate. *bioRxiv.* Published online March 24, 2020. https://doi.org/10.1101/498378

16. Legried, B. *et al.* (2020) Polynomial-time statistical estimation of species trees under gene duplication and loss. In *Proceedings of RECOMB 2020: the 24th Annual International Conference Research in Computational Molecular Biology*, pp. 120–135, Springer, Cham

17. Bravo, G.A. *et al.* (2019) Embracing heterogeneity: coalescing the tree of life and the future of phylogenomics. *PeerJ* 7, e6399

18. Maddison, W.P. (1997) Gene trees in species trees. *Syst. Biol.* 46, 523–536

19. Kubatko, L.S. and Degnan, J.H. (2007) Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst. Biol.* 56, 17–24

20. Roch, S. and Steel, M. (2015) Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theor. Popul. Biol.* 100, 56–62

21. Degnan, J.H. and Rosenberg, N.A. (2006) Discordance of species trees with their most likely gene trees. *PLoS Genet.* 2, e68

22. Mirarab, S. and Warnow, T. (2015) ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* 31, i44–i52

23. Chifman, J. and Kubatko, L. (2014) Quartet inference from SNP data under the coalescent model. *Bioinformatics* 30, 3317–3324

24. Vachaspati, P. and Warnow, T. (2015) ASTRID: accurate species trees from internode distances. *BMC Genomics* 16, S3

25. Heled, J. and Drummond, A.J. (2010) Bayesian inference of species trees from multilocus data using *BEAST. *Mol. Biol. Evol.* 27, 570–580

26. Bryant, D. *et al.* (2012) Inferring species trees directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. *Mol. Biol. Evol.* 29, 1917–1932

27. Liu, L. *et al.* (2010) A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evol. Biol.* 10, 302

28. Doolittle, W.F. and Brown, J.R. (1994) Tempo, mode, the progenote, and the universal root. *Proc. Natl. Acad. Sci. U. S. A.* 91, 6721–6728

29. Koonin, E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.* 39, 309–338

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

31. Altenhoff, A.M. *et al.* (2019) OMA standalone: orthology inference among public and custom genomes and transcriptomes. *Genome Res.* 29, 1152–1163

32. Kallal, R.J. *et al.* (2018) A phylotranscriptomic backbone of the orb-weaving spider family Araneidae (Arachnida, Araneae) supported by multiple methodological approaches. *Mol. Phylogenet. Evol.* 126, 129–140

33. Fernández, R. *et al.* (2018) Phylogenomics, diversification dynamics, and comparative transcriptomics across the spider tree of life. *Curr. Biol.* 28, 1489–1497

34. Wiens, J.J. (2008) Missing data and the accuracy of Bayesian phylogenetics. *J. Syst. Evol.* 46, 307–314

35. Wiens, J.J. (2003) Missing data, incomplete taxa, and phylogenetic accuracy. *Syst. Biol.* 52, 528–538

36. Wiens, J.J. (2006) Missing data and the design of phylogenetic analyses. *Syst. Biol.* 39, 34–42

37. Philippe, H. *et al.* (2004) Phylogenomics of eukaryotes: impact of missing data on large alignments. *Mol. Biol. Evol.* 21, 1740–1752

38. Driskell, A.C. *et al.* (2004) Prospects for building the tree of life from large sequence databases. *Science* 306, 1172–1174

39. Hosner, P.A. *et al.* (2016) Avoiding missing data biases in phylogenomic inference: an empirical study in the landfowl (Aves: Galliformes). *Mol. Biol. Evol.* 33, 1110–1125

40. Wiens, J.J. and Morrill, M.C. (2011) Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. *Syst. Biol.* 60, 719–731

41. Nute, M. *et al.* (2018) The performance of coalescent-based species tree estimation methods under models of missing data. *BMC Genomics* 19, 286

42. Xi, Z. *et al.* (2016) The impact of missing data on species tree estimation. *Mol. Biol. Evol.* 33, 838–860

43. Molloy, E.K. and Warnow, T. (2018) To include or not to include: the impact of gene filtering on species tree estimation methods. *Syst. Biol.* 67, 285–303

44. Rhodes, J.A. *et al.* (2020) NJ_{st} and ASTRID are not statistically consistent under a random model of missing data. *arXiv.* Published online January 22, 2020. https://arxiv.org/abs/2001.07844

45. Nute, M. *et al.* (2020) Correction to: the performance of coalescent-based species tree estimation methods under models of missing data. *BMC Genomics* 21, 133

46. Eaton, D.A.R. and Ree, R.H. (2013) Inferring phylogeny and introgression using RADseq data: an example from flowering plants (*Pedicularis*: Orobanchaceae). *Syst. Biol.* 62, 689–706

47. Gatesy, J. *et al.* (2019) Partitioned coalescence support reveals biases in species-tree methods and detects gene trees that determine phylogenomic conflicts. *Mol. Phylogenet. Evol.* 139, 106539

48. Page, R.D.M. (1994) Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Syst. Biol.* 43, 58–77

49. Guigo, R. *et al.* (1996) Reconstruction of ancient molecular phylogeny. *Mol. Phylogenet. Evol.* 6, 189–213

50. Boussau, B. and Scornavacca, C. (2020) Reconciling gene trees with species trees. In *Phylogenetics in the Genomic Era*, pp. 3.2:1–3.2:23. https://hal.inria.fr/PGE/

51. Wehe, A. *et al.* (2008) DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. *Bioinformatics* 24, 1540–1541

52. Bayzid, M.S. and Warnow, T. (2018) Gene tree parsimony for incomplete gene trees: addressing true biological loss. *Algorithms Mol. Biol.* 13, 1

53. Maddison, W.P. and Knowles, L.L. (2006) Inferring phylogeny despite incomplete lineage sorting. *Syst. Biol.* 55, 21–30

54. Than, C. and Nakhleh, L. (2009) Species tree inference by minimizing deep coalescences. *PLoS Comput. Biol.* 5, e1000501

55. Chaudhary, R. *et al.* (2010) iGTP: a software package for large-scale gene tree parsimony analysis. *BMC Bioinformatics* 11, 574

56. Wu, Y.-C. *et al.* (2014) Most parsimonious reconciliation in the presence of gene duplication, loss, and deep coalescence using labeled coalescent trees. *Genome Res.* 24, 475–486

57. Hallett, M. *et al.* (2004) Simultaneous identification of duplications and lateral transfers. In *Proceedings of RECOMB 2004: The 8th Annual International Conference Research in Computational Molecular Biology*, pp. 347–356, Association for Computing Machinery, NY

58. Chan, Y. *et al.* (2017) Inferring incomplete lineage sorting, duplications, transfers and losses with reconciliations. *J. Theor. Biol.* 432, 1–13

59. Stolzer, M. *et al.* (2012) Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics* 28, i409–i415

60. Than, C.V. and Rosenberg, N.A. (2011) Consistency properties of species tree inference by minimizing deep coalescences. *J. Comput. Biol.* 18, 1–15

61. Robinson, D.F. and Foulds, L.R. (1981) Comparison of phylogenetic trees. *Math. Biosci.* 53, 131–147

62. Bansal, M.S. *et al.* (2010) Robinson-Foulds supertrees. *Algorithms Mol. Biol.* 5, 18

63. Puigbo, P. *et al.* (2007) TOPD/FMTS: a new software to compare phylogenetic trees. *Bioinformatics* 23, 1556–1558

64. Marcet-Houben, M. and Gabaldón, T. (2011) TreeKO: a duplication-aware algorithm for the comparison of phylogenetic trees. *Nucleic Acids Res.* 39, e66

65. Chaudhary, R. *et al.* (2013) Inferring species trees from incongruent multi-copy gene trees using the Robinson-Foulds distance. *Algorithms Mol. Biol.* 8, 28

66. Chaudhary, R. *et al.* (2015) MulRF: a software package for phylogenetic analysis using multi-copy gene trees. *Bioinformatics* 31, 432–433

67. Chaudhary, R. *et al.* (2015) Assessing approaches for inferring species trees from multi-copy genes. *Syst. Biol.* 64, 325–339

68. Molloy, E.K. and Warnow, T. (2020) FastMulRFS: fast and accurate species tree estimation under generic gene duplication and loss models. *Bioinformatics* 36, i57–i65

69. Zhang, C. *et al.* (2019) ASTRAL-Pro: quartet-based species tree inference despite paralogy. *bioRxiv*. Published online December 15, 2019. https://doi.org/10.1101/2019.12.12.874727

70. Arvestad, L. *et al.* (2003) Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics* 19, i7–i15

71. Arvestad, L. *et al.* (2004) Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. In *Proceedings of RECOMB 2004: The 8th Annual International Conference Research in Computational Molecular Biology*, pp. 326–335, Association for Computing Machinery, NY

72. Akerborg, O. *et al.* (2009) Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proc. Natl. Acad. Sci. U. S. A.* 106, 5714–5719

73. Rasmussen, M.D. and Kellis, M. (2007) Accurate gene-tree reconstruction by learning gene- and species-specific substitution rates across multiple complete genomes. *Genome Res.* 17, 1932–1942

74. Górecki, P. *et al.* (2011) Maximum likelihood models and algorithms for gene tree evolution with duplications and losses. *BMC Bioinformatics* 12, S15

75. Szöllősi, G.J. *et al.* (2015) The Inference of gene trees with species trees. *Syst. Biol.* 64, e42–e62

76. Rasmussen, M.D. and Kellis, M. (2012) Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Res.* 22, 755–765

77. Li, Q. *et al.* (2020) The multilocus multispecies coalescent: a flexible new model of gene family evolution. *bioRxiv*. Published online July 14, 2020. https://doi.org/10.1101/2020.05.07.081836

78. Sjöstrand, J. *et al.* (2014) A Bayesian method for analyzing lateral gene transfer. *Syst. Biol.* 63, 409–420

79. Szollosi, G.J. *et al.* (2012) Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proc. Natl. Acad. Sci. U. S. A.* 109, 17513–17518

80. Rasmussen, M.D. and Kellis, M. (2011) A Bayesian approach for fast and accurate gene tree reconstruction. *Mol. Biol. Evol.* 28, 273–290

81. Morel, B. *et al.* (2020) GeneRax: a tool for species tree-aware maximum likelihood based gene family tree inference under gene duplication, transfer, and loss. *Mol. Biol. Evol.* Published online June 5, 2020. https://doi.org/10.1093/molbev/msaa141

82. Boussau, B. *et al.* (2013) Genome-scale coestimation of species and gene trees. *Genome Res.* 23, 323–330

83. De Oliveira Martins, L. *et al.* (2016) A Bayesian supertree model for genome-wide species tree reconstruction. *Syst. Biol.* 65, 397–416

84. Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425

85. Liu, L. *et al.* (2009) Estimating species phylogenies using coalescence times among sequences. *Syst. Biol.* 58, 468–477

86. Dasarathy, G. *et al.* (2015) Data requirement for phylogenetic inference from multiple loci: a new distance method. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 12, 422–432

87. Allman, E.S. *et al.* (2013) Species tree inference by the STAR method and its generalizations. *J. Comput. Biol.* 20, 50–61

88. Allman, E.S. *et al.* (2018) Species tree inference from gene splits by unrooted STAR methods. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 15, 337–342

89. Graur, D. *et al.* (1996) Phylogenetic position of the order Lagomorpha (rabbits, hares, and allies). *Nature* 379, 333–335

90. Bryant, D. and Steel, M. (2001) Constructing optimal trees from quartets. *J. Algorithms* 38, 237–259

91. Strimmer, K. and von Haeseler, A. (1996) Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* 13, 964–969

92. Snir, S. and Rao, S. (2012) Quartet MaxCut: a fast algorithm for amalgamating quartet trees. *Mol. Phylogenetic. Evol.* 62, 1–8

93. Reaz, R. *et al.* (2014) Accurate phylogenetic tree reconstruction from quartets: a heuristic approach. *PLoS One* 9, e104008

94. Mirarab, S. *et al.* (2014) ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* 30, i541–i548

95. Wascher, M. and Kubatko, L. (2020) Consistency of SVDQuartets and maximum likelihood for coalescent-based species tree estimation. *Syst. Biol.* Published online May 16, 2020. https://doi.org/10.1093/sysbio/syaa039

96. Rabiee, M. *et al.* (2019) Multi-allele species reconstruction using ASTRAL. *Mol. Phylogenet. Evol.* 130, 286–296

97. Markin, A. and Eulenstein, O. (2020) Quartet-based inference methods are statistically consistent under the unified duplication-loss-coalescence model. *arXiv*. Published online April 8, 2020. https://arxiv.org/abs/2004.04299

98. Siu-Ting, K. *et al.* (2019) Inadvertent paralog inclusion drives artifactual topologies and timetree estimates in phylogenomics. *Mol. Biol. Evol.* 36, 1344–1356

99. Kozlov, A.M. *et al.* (2019) RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 35, 4453–4455

100. Huelsenbeck, J.P. and Ronquist, F. (2001) MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics* 17, 754–755

101. Thompson, J.D. *et al.* (2011) A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. *PLoS ONE* 6, e18093

102. Sela, I. *et al.* (2015) GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Res.* 43, W7–W14

103. Villanueva-Cañas, J.L. *et al.* (2013) Improving genome-wide scans of positive selection by using protein isoforms of similar length. *Genome Biol. Evol.* 5, 457–467

104. Capella-Gutiérrez, S. *et al.* (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973

105. Castresana, J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 17, 540–552

106. Dress, A.W. *et al.* (2008) Noisy: identification of problematic columns in multiple sequence alignments. *Algorithms Mol. Biol.* 3, 7

107. Landan, G. and Graur, D. (2007) Heads or tails: a simple reliability check for multiple sequence alignments. *Mol. Biol. Evol.* 24, 1380–1383

108. Glover, N.M. *et al.* (2016) Homoeologs: what are they and how do we infer them? *Trends Plant Sci.* 21, 609–621

109. Thomas, G.W.C. *et al.* (2017) Gene-tree reconciliation with MUL-trees to resolve polyploidy events. *Syst. Biol.* 66, 1007–1018

110. Huson, D.H. *et al.* (2005) Reconstruction of reticulate networks from gene trees. In *Proceedings of RECOMB 2005: The 9th Annual International Conference Research in Computational Molecular Biology*, pp. 233–249, Springer, Berlin

111. Vanderpool, D. *et al.* (2020) Primate phylogenomics uncovers multiple rapid radiations and ancient interspecific introgression. *bioRxiv*. Published online April 16, 2020. https://doi.org/10.1101/2020.04.15.043786

112. Yu, Y. and Nakhleh, L. (2015) A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics* 16, S10

113. Solís-Lemus, C. and Ané, C. (2016) Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genet.* 12, e1005896

114. Shen, X.-X. *et al.* (2017) Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nat. Ecol. Evol.* 1, 126

115. Altenhoff, A.M. *et al.* (2019) Inferring orthology and paralogy. In *Evolutionary Genomics: Statistical and Computational Methods* (Anisimova, M., ed.), pp. 149–175, Springer, Berlin

116. Li, L. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 1, 2178–2189

117. Yang, Y. and Smith, S.A. (2014) Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. *Mol. Biol. Evol.* 31, 3081–3092