Syst. Biol. 71(3):649–659, 2022

DOI:10.1093/sysbio/syab097
Advance Access publication December 24, 2021

The Frequency and Topology of Pseudoorthologs

Megan L. Smith* and Matthew W. Hahn

Department of Biology and Department of Computer Science, Indiana University, Bloomington, IN 47405, USA
**Correspondence to be sent to: Department of Biology and Department of Computer Science, Indiana University, 1001 E 3rd St., Bloomington, IN 47405, USA;*
E-mail: mls16@indiana.edu.

Received 10 March 2021; reviews returned 15 December 2021; accepted 17 December 2021
Associate Editor: Stephen Smith

Abstract.—Phylogenetics has long relied on the use of orthologs, or genes related through speciation events, to infer species relationships. However, identifying orthologs is difficult because gene duplication can obscure relationships among genes. Researchers have been particularly concerned with the insidious effects of pseudoorthologs—duplicated genes that are mistaken for orthologs because they are present in a single copy in each sampled species. Because gene tree topologies of pseudoorthologs may differ from the species tree topology, they have often been invoked as the cause of counterintuitive results in phylogenetics. Despite these perceived problems, no previous work has calculated the probabilities of pseudoortholog topologies or has been able to circumscribe the regions of parameter space in which pseudoorthologs are most likely to occur. Here, we introduce a model for calculating the probabilities and branch lengths of orthologs and pseudoorthologs, including concordant and discordant pseudoortholog topologies, on a rooted three-taxon species tree. We show that the probability of orthologs is high relative to the probability of pseudoorthologs across reasonable regions of parameter space. Furthermore, the probabilities of the two discordant topologies are equal and never exceed that of the concordant topology, generally being much lower. We describe the species tree topologies most prone to generating pseudoorthologs, finding that they are likely to present problems to phylogenetic inference irrespective of the presence of pseudoorthologs. Overall, our results suggest that pseudoorthologs are unlikely to mislead inferences of species relationships under the biological scenarios considered here.[Birth–death model; orthologs; paralogs; phylogenetics.]

Phylogenetics aims to reconstruct evolutionary relationships among species. Recent advances in sequencing technologies have drastically increased the amount of data available for phylogenetic inference (Scornavacca et al. 2020), which has led in turn to increased concern about how to assemble and filter large genomic and transcriptomic data sets. Central to most data-generating pipelines is the identification of orthologs, or genes related through speciation events, to the exclusion of paralogs, or genes related through duplication events (Fitch 1970). Because orthologous gene trees reflect only the species history, it has been argued that solely orthologs are appropriate for phylogenetic inference (e.g., Fernández et al. 2020; Kapli et al. 2020). Methods to extract orthologs from large data sets have therefore proliferated (reviewed in Altenhoff et al. 2019a, e.g., Ebersberger et al. 2009; Altenhoff et al. 2011, 2013; Dunn et al. 2013; Yang and Smith 2014), but the task remains difficult, and pseudoorthologs (Koonin 2005) (or "hidden paralogs" Doolittle and Brown 1994), are thought to represent a particularly insidious problem. Pseudoorthologs are paralogs that are mistaken as orthologs because, due to patterns of differential duplication and loss, they are present in a single copy in each sampled species.

Pseudoortholog gene trees can differ from the species tree in their topology and branch lengths. Consider, for example, a scenario in which a duplication occurred in the ancestor of three species (A, B, and C), where species A and B are sister species (Fig. 1a,b). If one of the two copies is lost immediately, we can only sample genes with orthologous relationships (Fig. 1c). If one copy is retained in species A and species B, while the other is retained in species C, then we have a pseudoortholog that is topologically identical to the true ortholog, but which has a longer internal branch (Fig. 1d). Finally, if one copy is retained in species A (or B) and the other is retained in species B (or A) and C, then we have a pseudoortholog with a topology that differs from the species tree topology (Fig. 1e,f). Because discordant pseudoorthologs are difficult to identify—and may introduce both branch length and topological heterogeneity—they are often invoked as the culprits behind counterintuitive results in phylogenetics.

Multiple studies have attempted to assess the influence of paralogs (including pseudoorthologs) on phylogenetic inference, though they have generally done so by comparing results filtered using different ortholog detection methods (Fernández et al. 2020), none of which are likely to remove pseudoorthologs. The results of these analyses have been mixed, with some studies finding substantial differences in inferred species trees (Altenhoff et al. 2019b; Siu-Ting et al. 2019; Cheon et al. 2020) and others finding minimal differences (Fernández et al. 2018; Kallal et al. 2018; Cheon et al. 2020). Furthermore, and in contrast to the long-held opinion that orthologs, not paralogs, should be used to infer species relationships, recent methodological developments explicitly allow for the inclusion of paralogs in phylogenetic inference (reviewed in Smith and Hahn 2021). In particular, quartet-based gene tree methods are robust to the inclusion of paralogs because the concordant topology is expected to be the most common topology, in the limit of a very large number of genes (Legried et al. 2020; Markin and Eulenstein 2020; Zhang et al. 2020; Yan et al. 2021). However, branch-length estimates, concordance factors, and measures

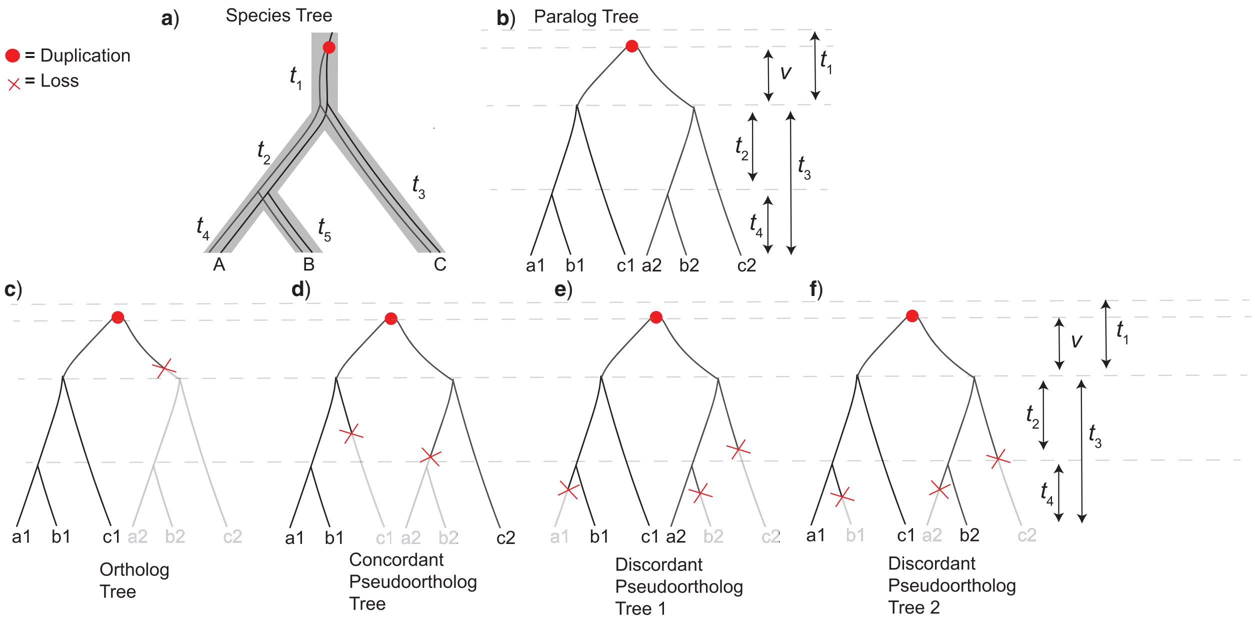

FIGURE 1. Orthologs and pseudoorthologs. a) A rooted three-taxon species tree showing the relationships between Species A, B, and C is depicted by the gray outline. Within the tree, the duplication is indicated by a dot, and the two daughter paralog trees are drawn. b) The full paralog tree depicting relationships between all gene copies. The lengths of all branches are shown on the right. c–f) The different relationships possible when a single copy in each species is present. X's indicate loss events. c) Orthologs require at least one loss, d) concordant pseudoorthologs require at least two losses, and e, f) discordant pseudoorthologs require at least three losses.

of nodal support may still be impacted by paralog inclusion.

For researchers who wish to only use orthologs for phylogenetic inference, current practices for excluding putative pseudoorthologs can be particularly restrictive because excluding these genes from phylogenetic data sets is difficult. When extracting putative orthologs, researchers typically rely on graph-based or tree-based approaches (Altenhoff et al. 2019a). Graph-based approaches are the most commonly used: they rely on the concept of reciprocal best hits (Li 2003) or length-normalized reciprocal best hits (Emms and Kelly 2015), followed by the application of clustering approaches to delineate orthologs. These methods assume that the two most closely related homologs between a pair of species should be orthologs. When pseudoorthologs are present they will be reciprocal best hits, despite not having an orthologous relationship, because true orthologs are absent. Tree-based approaches (e.g., Yang and Smith 2014) extract orthologs from the clusters identified using graph-based methods. Tree-based approaches are more computationally intensive, but in some cases may be able to identify and exclude pseudoorthologs by identifying excessively long branches. Even these approaches, though, will often fail to identify pseudoorthologs, particularly when duplication events were more recent. Another approach to removing putative pseudoorthologs relies on knowledge of a species tree, removing genes that show discordance between the gene and species tree for a set of predefined clades (Siu-Ting et al. 2019). This approach assumes that discordance between gene trees and the species tree with respect to *a priori* defined "uncontestable" relationships is due to gene duplication and loss, although many other factors including incomplete lineage sorting and introgression may also lead to gene tree heterogeneity (Maddison 1997).

Thus, options for excluding pseudoorthologs from phylogenetic data sets are limited and likely ineffectual at removing pseudoorthologs, and those that do exist may lead to a drastic reduction in the amount of data available. For example, when using their approach based on the monophyly of predefined clades (Siu-Ting et al. 2019) removed 637 of their 2656 putative orthologs. They found some differences between tree topologies and branch lengths inferred from these filtered and unfiltered data sets, but it is difficult to establish whether the removed genes were actually pseudoorthologs and how many pseudoorthologs remained after stringent filtering. A better understanding of when, and how stringently, we should filter our data to remove pseudoorthologs would clearly be helpful: it could prevent unnecessary filtering of informative genes from phylogenetic data sets and would guide researchers as to whether and when results should be interpreted with caution due to the potential presence of pseudoorthologs.

Despite long-standing concerns about the effects of pseudoorthologs on phylogenetic inference, no attempt has been made to calculate the probability of pseudoorthologs or to understand the regions of parameter space in which they may be most problematic. Here, we use a stochastic birth–death model to calculate the probabilities and branch lengths of orthologs and pseudoorthologs, including both concordant and discordant pseudoortholog topologies. In what follows, we first describe the model and then explore regions of parameter space that are most likely to produce pseudoorthologs. We show that the probability of orthologs is high relative to the probability of pseudoorthologs across parameter space, and that the ratio of concordant to discordant topologies is even higher. Our results should reassure researchers concerned about the effects of pseudoorthologs on phylogenetic inference.

The Model

Probabilities of Orthologs and Pseudoorthologs

To calculate the probabilities of orthologs and pseudoorthologs, we use a stochastic birth–death model (Bailey 1964). Previous work has applied birth–death models to gene trees with the aim of inferring orthology, reconciling gene and species trees, and accurately reconstructing gene trees (Arvestad et al. 2003, 2004; Rasmussen and Kellis 2011). Here, we evaluate a specific case by focusing on a rooted three-taxon species tree, considering scenarios that generate single-copy genes in order to estimate probabilities of orthologs and pseudoorthologs.

All calculations assume that there is one gene copy at the beginning of internal branch t_1 (Fig. 1a). When only a single duplication (and no loss) occurs on this branch, such that two copies exist at the most recent node, we treat each copy independently, generating two "daughter" gene trees (Fig. 1b). The independent evolution of each copy means that we can calculate probabilities of further gain and loss on all subsequent lineages, always beginning with a single copy at the base of the daughter gene trees. Since we always begin with a single copy, we use the following equations to calculate the probabilities of transitions along branches, where λ is the duplication rate and μ is the loss rate. The probability of starting with 1 copy and ending with n copies along a branch with length t can be calculated as (Bailey 1964):

$$p_n(t)=\begin{cases}(1-\alpha)(1-\beta)\beta^{n-1}, & n>0\\ \alpha, & n=0\end{cases}$$

where

$$\alpha=\frac{\mu(e^{(\lambda-\mu)t}-1)}{\lambda e^{(\lambda-\mu)t}-\mu},\beta=\frac{\lambda(e^{(\lambda-\mu)t}-1)}{\lambda e^{(\lambda-\mu)t}-\mu}$$

When $\lambda=\mu$, we use the following simplification (Bailey 1964):

$$p_n(t)=\begin{cases}\frac{(\lambda t)^{n-1}}{(1+\lambda t)^{n+1}}, & n>0\\ \frac{(\lambda t)}{(1+\lambda t)}, & n=0\end{cases}$$

Using the above equations, we can calculate the overall probabilities of different ortholog and pseudoortholog topologies. As an example, consider the concordant pseudoortholog in Figure 1d (see also Supplementary Fig. S2a available on Dryad at http://dx.doi.org/10.5061/dryad.573n5tb72). We calculate the overall probability of this topology by multiplying: the probability of transitioning from 1 to 2 copies on branch t_1, the probability of transitioning from 1 to 0 copies on branch t_2 in one copy and 1 to 0 copies on branch t_3 in the other copy, and the probability of no changes on any of the other branches (online Supplementary Appendix A available on Dryad). Note that there are two different arrangements that may lead to this outcome, depending on which daughter gene tree losses occur on. Similarly, we can calculate the probability of the type of ortholog shown in Fig. 1a by calculating the total probability that there are no transitions on any branch (i.e., that the state is 1 at all nodes; Supplementary Fig. S1a available on Dryad). In total, we consider six ortholog configurations (i.e., sets of events leading to orthologs; Supplementary Fig. S1 available on Dryad) that can each occur in from 1 to 6 different arrangements (depending on which exact copies are lost). We consider nine concordant pseudoortholog configurations (Supplementary Fig. S2 available on Dryad) and five configurations for each of the two discordant pseudoortholog topologies (Supplementary Fig. S3 available on Dryad), each of which can occur in 1–6 different arrangements. There are more configurations possible with more ancestral copies, but here we limit the number of copies at the end of branch t_1 to 3 (i.e., two duplications on branch t_1). Code to calculate the probabilities of orthologs, concordant pseudoorthologs, and discordant pseudoorthologs is given in online Supplementary Appendix A available on Dryad.

Our model makes several assumptions. Most notably, we assume that there are no more than three copies at the end of branch t_1, and we avoid nested duplication scenarios. To evaluate whether these assumptions lead to accurate predictions, we compared the exact probabilities calculated according to the equations above to the proportions of each scenario observed in simulations in SimPhy (Mallo et al. 2016). We calculated the proportion of observed orthologs, concordant pseudoorthologs, and discordant pseudoorthologs by evaluating topology and branch lengths. If our assumptions are reasonable, we expect a 1:1 relationship between calculations and observations. We drew 100 sets of parameters from uniform priors (Supplementary Table S1 available on Dryad) and performed 10,000 simulations under each set of parameters in SimPhy. While our model does not exhaust all possible scenarios that could lead to pseudoorthologs, these simulations show that the scenarios we consider lead to accurate predictions of the numbers of orthologs, as well as the numbers of concordant and discordant pseudoorthologs (Supplementary Fig. S4 available on Dryad).

Expectations for Branch Lengths of Pseudoorthologs

In addition to differing topologically from the species tree, pseudoorthologs differ from the species tree in terms of branch lengths. For orthologs, the single internal branch of a rooted three-taxon tree is length t_2; this branch determines the phylogenetic signal within each gene tree, and is the focus here. For the simplest concordant pseudoortholog (Fig. 1d), the internal branch length is equal to the sum of t_2 and the time until the duplication event occurs in branch t_1 (v in Fig. 1), while for the simplest discordant pseudoortholog the internal branch length is equal to v (Fig. 1e,f). Furthermore, average internal branch lengths for the two discordant pseudoorthologs are always equal.

The value of v is the expected time back to the duplication event on t_1 conditional on a duplication event occurring on branch t_1. Because waiting times for events in the birth–death process are also exponentially distributed (Gernhard 2008), we can use a model similar to that for the multispecies coalescent (Mendes and Hahn 2018) to calculate times here. To find the expectation for v, we need only convert from the coalescent units used in Mendes and Hahn (2018) to duplication units, where one coalescent unit is equal to $\frac{1}{\lambda}$ here. These considerations lead to the following expectation for the time back to the duplication event:

$$E[v]=\frac{1}{\lambda}-\frac{t_1}{e^{t_1\lambda}-1}$$

Although we have conditioned on a duplication event occurring in branch t_1, we have not conditioned on other events. For example, we have not conditioned on the absence of any subsequent duplication events or a subsequent loss on branch t_1. To evaluate whether the assumptions made here led to reasonable predictions of the internal branch length, we again used simulations in SimPhy (Mallo et al. 2016). We drew 50 sets of parameters from uniform priors; all priors were the same as in Supplementary Table S1 available on Dryad except μ and λ were drawn from U(0.004, 0.005) priors to ensure more pseudoorthologs. We performed 10,000 simulations under each set of parameters, and calculated the average internal branch lengths from simulations that produced either trees matching the concordant pseudoortholog shown in Fig. 1d or trees matching the discordant pseudoortholog shown in Fig. 1e. The expected branch lengths are a close match to simulated branch lengths (Supplementary Fig. S5 available on Dryad), and thus should provide accurate predictions of the internal branch lengths of pseudoorthologs.

The model presented here demonstrates that the expected internal branch length for concordant pseudoorthologs is always longer than the expected branch length for discordant pseudoorthologs, by the length of the internal branch t_2. Thus, even when pseudoorthologs are present, the total expected branch length supporting the concordant topology should exceed the expected branch length supporting the discordant topology. In other words, there is more phylogenetic signal in concordant trees than discordant ones. In addition, the internal branch supporting each of the two different discordant pseudoorthologs has the same expected length. This implies equal support for each of the two discordant topologies.

Probabilities of Orthologs and Pseudoorthologs

We used the model described above to explore how different parameters affected the probabilities of orthologs and pseudoorthologs, including both concordant and discordant topologies. We begin by describing our results in terms of unconditional probabilities, which consider all scenarios resulting from our model, including those that do not produce one gene copy per species. Considering the rates of gene duplication (λ) and loss (μ), we found that higher rates of each decreased the overall probability of orthologs (Supplementary Fig. S6a available on Dryad). The probability of orthologs decreases because there is a higher chance of duplication and loss events occurring: duplication creates additional copies, while loss means that no copy can be sampled from some species. A similar effect is generated by increasing all branch lengths.

By contrast, the probability of pseudoorthologs is maximized at intermediate values of λ and μ (Fig. 2a), because at least one duplication event and two loss events are required for pseudoorthologs (Fig. 1d–f). Values of these parameters that are too high decrease the probability of there being a single copy in each species; because pseudoorthologs require more losses than gains, slightly higher values of μ are possible. Similarly, the probability of pseudoorthologs is maximized at intermediate branch lengths of t_1 (Fig. 2b) and t_3 (Supplementary Fig. 6b available on Dryad), because at least one duplication is required on branch t_1 and at least one loss is required on branch t_3.

Probabilities of Concordant and Discordant Pseudoorthologs

To understand the relative probabilities of concordant and discordant pseudoortholog topologies, we examined many of the same rate and branch-length parameters. Increasing μ decreases the ratio of concordant pseudoorthologs to discordant pseudoorthologs, because discordant pseudoorthologs require at least three losses while concordant pseudoorthologs can occur with only two losses (Supplementary Fig. S6c available on Dryad). Increasing λ slightly increases the ratio of concordant pseudoorthologs to discordant pseudoorthologs (Supplementary Fig. S6c available on Dryad), as there are more possible configurations leading to concordant pseudoorthologs than discordant pseudoorthologs when there is more than one duplication event (Supplementary Figs. S2 and S3 available on Dryad). Changes to the lengths of branches t_2 and t_4/t_5 affect the relative probabilities of concordant and discordant pseudoorthologs: specifically, as branch t_2 gets longer and branches t_4/t_5 get shorter, concordant

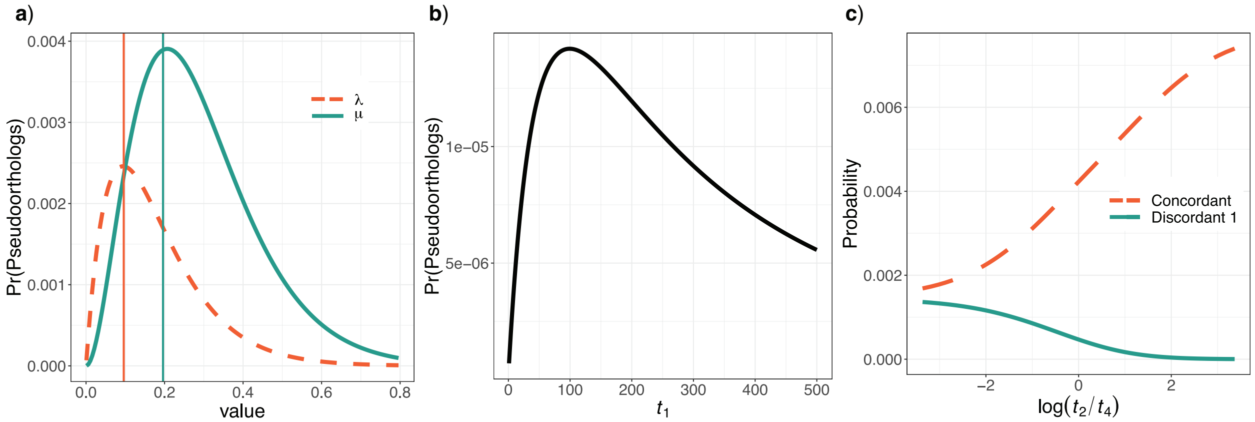

FIGURE 2. Effects of varying parameters on the probability of pseudoorthologs. a) The unconditional probability of pseudoorthologs is maximized at intermediate values of μ and λ. b) The unconditional probability of pseudoorthologs is maximized at intermediate values of branch length t_1. c) The relative unconditional probabilities of concordant and discordant pseudoorthologs depend on the ratio of branch lengths t_2 and t_4 (or t_5). As t_2 gets larger (x-axis becomes more positive) the unconditional probability of concordant pseudoorthologs increases and the unconditional probability of discordant pseudoorthologs decreases. The probability of one discordant pseudoortholog is shown, but the values are identical for the other discordant pseudoortholog.

pseudoorthologs become more likely and discordant pseudoorthologs become less likely (Fig. 2c). This occurs because concordant pseudoorthologs can be generated either by a loss on t_2 or by losses on both t_4 and t_5 (Fig. 1d; Supplementary Fig. S1a,b available on Dryad), while discordant pseudoorthologs require losses on branches t_4 and t_5 (Fig. 1e,f; Supplementary Fig. S3 available on Dryad). Both scenarios additionally require losses on t_3, and so the length of t_3 does not affect their relative frequencies. Note that results from the model presented here are also supported by results from simulations (Supplementary Fig. S4 available on Dryad).

We further explored the probabilities of all events conditional on a single copy being present in each species. These calculations directly address the chance that pseudoorthologs are mistaken for orthologs: the conditional probabilities represent the fraction of all single-copy genes that are orthologs, concordant pseudoorthologs, or discordant pseudoorthologs. We explored two general regions of parameter space, representing the range of values of λ and μ observed in empirical data sets: 0.002 and 0.005 per million years (Mendes et al. 2020). We considered a long length of branch t_3 (198.9 million years) across a range of lengths for branches t_1, t_2, t_4, and t_5. The large value of t_3 mirrors a potentially difficult region of tree space (see next section), coupled with moderate and high rates of duplication and loss.

The conditional probability of orthologs given that a single copy is present in each species is very high when rates of duplication and loss are moderate (0.002, Fig. 3a, Supplementary Fig. S7 available on Dryad; minimum conditional probability of orthologs $=0.955$), and is moderately high even when rates of duplication and loss approach the highest observed in empirical data sets (Fig. 3c; minimum conditional probability $=0.711$). Furthermore, the ratio of concordant to discordant topologies is very high when duplication and loss rates are moderate (Fig. 3b; minimum $=76.7{:}1$) and is still rather high even when rates of duplication and loss are high (Fig. 3d; minimum $=8.4{:}1$). These ratios include both orthologs and concordant pseudoorthologs in the "concordant" category. Note again that we chose t_3 to mirror the most problematic regions of parameter space for these results; Supplementary Fig. S8 available on Dryad shows results for different values of t_3, confirming the impression that the scenario shown here in the main text is a worst-case scenario with regards to this branch length.

Notably, the probability of either of the two discordant pseudoorthologs can never exceed the probability of the concordant pseudoortholog, because it is always possible to generate a concordant pseudoortholog with the same number of duplication and loss events (and often fewer events). For example, a discordant pseudoortholog can be generated by a duplication on branch t_1, a loss on branches t_4 and t_3 in one copy, and a loss on branch t_5 in the other copy (Fig. 1e). A concordant pseudoortholog could also be generated by this pattern, as long as the losses on branches t_4 and t_5 occurred in the same copy, while the loss on branch t_3 occurred in the other copy (Supplementary Fig. S2b available on Dryad). In reasonable regions of parameter space, the probability of concordant pseudoorthologs is much higher than the probability of either discordant pseudoortholog because most concordant pseudoorthologs require one fewer loss event (i.e., the scenario shown in Fig. 1d; Supplementary Figs. S2 and S3 available on Dryad). Moreover, the probabilities of the two discordant topologies are always equal, as these rely on the same events on the same branches, and only differ in terms of which branches are lost from which copy. Thus, one never expects either discordant topology to be significantly more frequent than the other. Finally, note again that in Fig. 3, we are showing the probabilities of orthologs

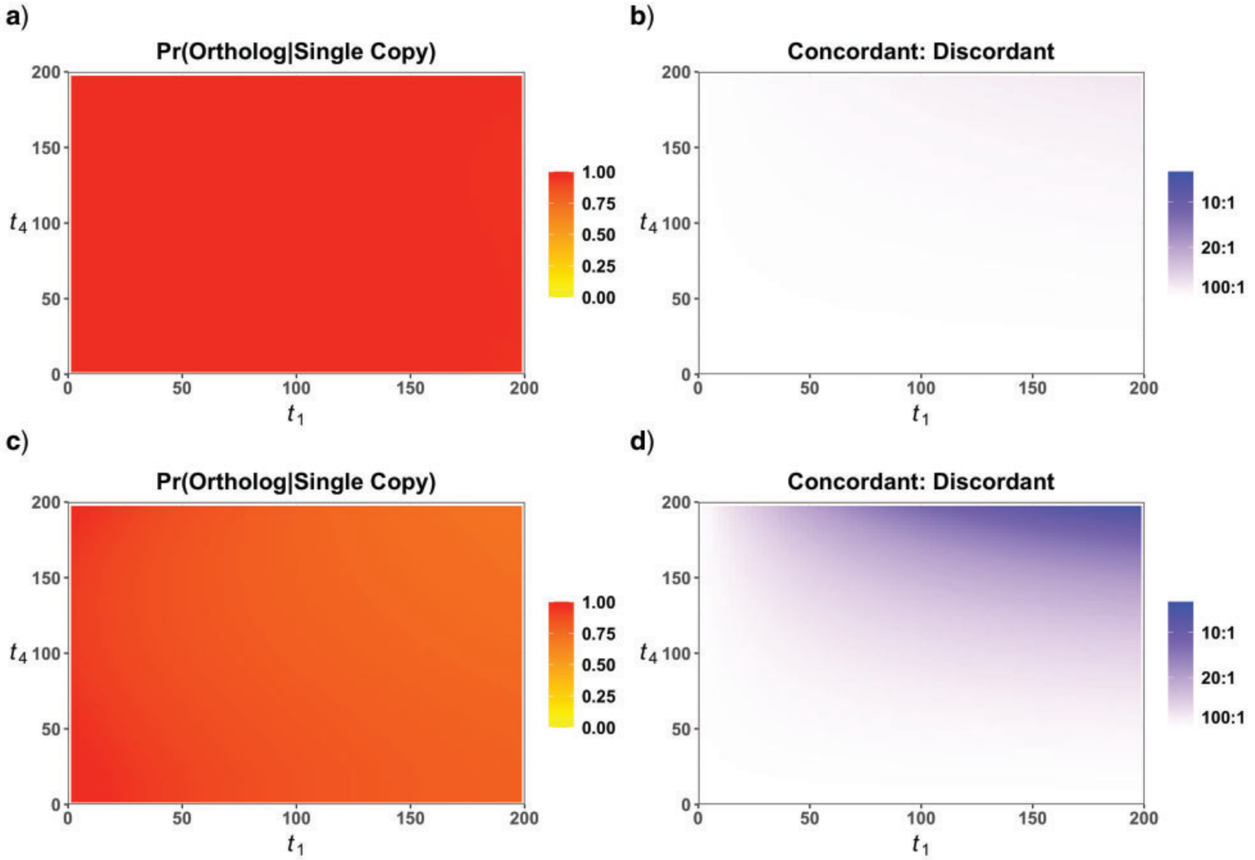

FIGURE 3. Probabilities of orthologs, pseudoorthologs, and discordance. Here, branch length $t_3 = 198.9$ mya. Branch length t_1 varies from 0.0001 to 200 mya, while branch length t_4 varies from 0.0001 to 198.8 mya; branch length t_2 is constrained such that the sum of t_2 and t_4 equals t_3. a) The conditional probability of orthologs given that a single copy is present in each species, with moderate rates of duplication and loss ($\lambda = 0.002$ per my, $\mu = 0.002$ per my). b) The ratio of the concordant topology (orthologs and concordant pseudoorthologs) to one discordant topology given that a single copy is present in each species, with moderate rates of duplication and loss ($\lambda = 0.002, \mu = 0.002$). c) The conditional probability of orthologs given that a single copy is present in each species with high rates of duplication and loss ($\lambda = 0.005, \mu = 0.005$). d) The ratio of the concordant topology (orthologs and concordant pseudoorthologs) to one discordant topology given that a single copy is present in each species with high rates of duplication and loss ($\lambda = 0.005, \mu = 0.005$).

and pseudoorthologs conditional on sampling a single copy per species. However, the absolute probability of sampling a single copy per species at all is lowest in the regions of parameter space that maximize the probability of pseudoorthologs and discordant topologies (Supplementary Fig. S9 available on Dryad).

WORST-CASE SCENARIOS

In order to find the species tree topologies most prone to producing pseudoorthologs (especially discordant ones), we searched parameter space for such trees. Specifically, we searched for regions of parameter space that a) maximized the probability of pseudoorthologs conditional on a single copy per species, b) maximized the probability of discordant pseudoorthologs conditional on a single copy per species, and c) minimized the ratio of concordant topologies to discordant topologies (Supplementary Table S2 available on Dryad). We set bounds on all parameters (Supplementary Table S3 available on Dryad), and constrained the species tree to be ultrametric by setting t_5 equal to t_4 and requiring that $t_4 + t_2 = t_3$. We changed each parameter in turn, increasing or decreasing the value at random by a value chosen from a uniform prior distribution U(0.000001, 0.001) for μ and λ, and U(0.0001, 20) for branch lengths. For each optimization, we accepted each change if it increased the probability (or decreased the ratio), and accepted the change one percent of the time if it decreased the probability (or increased the ratio). For each parameter we performed 100 optimization steps. We visited the parameters in the order: μ, λ, t_1, t_3, and t_2, and repeated the procedure ten times.

The maximum conditional probability of pseudoorthologs that we found was 0.285 (Supplementary Table S2 available on Dryad), and this value was only obtained with high values of λ and μ (Mendes et al. 2020). The highest conditional probability of either of the two discordant pseudoorthologs observed was 0.095, and again this involved high values of λ and μ (Supplementary Table S2 available on Dryad; Fig. 4). The minimum ratio of the concordant topology to either of the two discordant topologies was 8.5. This suggests that, even in the most problematic regions of parameter space, discordant pseudoorthologs will comprise fewer than 10% of all single-copy genes.

Equally importantly, our results show that the ratio of concordant to discordant topologies is lowest in a region of tree space in which discordance due to incomplete lineage sorting (ILS) is also likely to be a concern (Fig. 4)—when the internal branch of a three-species tree is very short. If we take units in millions of years and assume a species with a generation time of 29 years and an effective population size of 10,000, then the probability of either discordant topology for the worst-case species tree under the multispecies coalescent model is 0.225. By comparison, the conditional probability of either discordant topology under our model of duplication and loss in this same area of parameter space is 0.095, a value more than two times lower. Additionally, this region of parameter space involves very long branch lengths for t_1, t_3, and t_4/t_5 (nearly 200 million years) and high rates of duplication and loss. In such species trees, pseudoorthologs are not likely to be the biggest impediment to phylogenetic inference.

We also explored regions of parameter space that maximized the absolute probabilities of pseudoorthologs and discordant pseudoorthologs, rather than the probabilities conditional on a single copy per species (Supplementary Table S4 available on Dryad). The most notable difference was in the branch lengths that maximized the probability of pseudoorthologs. When absolute probabilities are considered, a long internal branch t_2 and short terminal branches t_4 and t_5 maximize this probability because they maximize the probability of the concordant pseudoortholog (Supplementary Table S4 available on Dryad). However, when conditional probabilities are considered, a shorter internal branch t_2 maximizes the probability of pseudoorthologs because, coupled with longer branches t_4 and t_5, a shorter branch t_2 decreases the probability of orthologs.

Whole-Genome Duplications

To evaluate the effects of whole-genome duplication events (WGDs), we used simulations in SimPhy (Mallo et al. 2016). Since WGDs cannot be specified in SimPhy, we simulated two locus trees per replicate for the rooted three-taxon tree (and an outgroup); these trees were treated as a pair of duplicates produced by WGD. These duplicates were identical in terms of their branch lengths, but all subsequent duplication or loss events were independent across the two copies. We simulated data under six conditions. We combined moderate (0.002) and high (0.005) duplication and loss rates with three branch length conditions that are described in Supplementary Table S5 available on Dryad. The final scenario was designed based on the worst-case results under the original model (Fig. 4). We recorded the proportion of single-copy genes and the proportion of those genes that were orthologs, concordant pseudoorthologs, and discordant pseudoorthologs, and compared these values to the predicted probabilities under the model without WGDs described above. In general, WGDs lead to fewer single-copy genes (Supplementary Table S5 available on Dryad). Conditional on a single copy per species, WGDs lead to a lower proportion of orthologs, and higher proportions of concordant pseudoorthologs and each discordant pseudoortholog. Despite this, the proportion of genes with the concordant topology is always higher than the proportion of genes with either discordant topology, and the proportion of genes with either discordant topology never exceeds 0.15 even in the worst-case scenario (Supplementary Table S5 available on Dryad). This suggests that, although polyploidy offers unique challenges, the expectation that the concordant topology should always be the most frequent holds, at least in the scenario considered here.

Extending the Model to Larger Trees

Thus far, we have considered the probability of orthologs and pseudoorthologs in a three-taxon species tree. While we might intuitively expect that the addition of more species would lower the relative probability of pseudoorthologs (because more losses would be required to mimic orthologs), we carried out additional analyses to evaluate slightly larger trees by adding a single extra taxon sister to Species A, Species C, or to internal branch t_1, assuming no discordance at the node uniting the new taxon and its sister species in the former two cases. These results support our prediction: adding taxa decreases the probability of all types of single-copy genes, including orthologs and concordant pseudoorthologs (Supplementary Figs. S10a,b, S11, and S12 available on Dryad). However, adding leaves disproportionately decreases the probability of discordant pseudoorthologs, particularly when branches are added as sister to Species A (Supplementary Fig. S10c,d available on Dryad). This outcome occurs because discordant pseudoorthologs require losses on branches t_4 and t_5, and, if these losses do not occur before the split between the two sister branches including Species A, then the number of losses required increases by one. While the same is true when a branch is added sister to Species C, since branch t_3 is longer than branches t_4 and t_5, there is more time for the loss to occur prior to the added speciation event. Adding a new lineage to internal branch t_1 has similar effects (Supplementary Text, Supplementary Figs. S11 and S12 available on Dryad), with a decrease in the absolute probabilities of orthologs and pseudoorthologs; the concordant topology remains the most probable. Overall, these limited extra analyses indicate that results

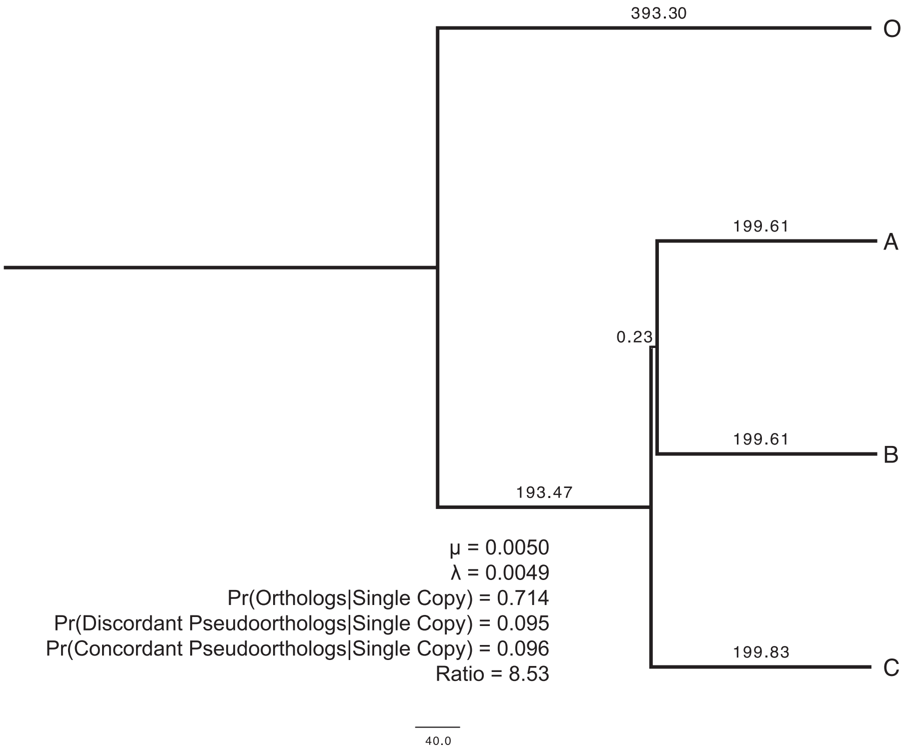
393.30
O
199.61
A
0.23
199.61
B
193.47
μ = 0.0050
λ = 0.0049
Pr(Orthologs|Single Copy) = 0.714
Pr(Discordant Pseudoorthologs|Single Copy) = 0.095
Pr(Concordant Pseudoorthologs|Single Copy) = 0.096
Ratio = 8.53
199.83
C
40.0

FIGURE 4. The species tree and parameters that minimize the ratio of the concordant topology to the two discordant topologies. Discordant probabilities and ratios refer only to discordant pseudoortholog 1, but the probabilities of the two discordant pseudoorthologs are equal. Probabilities are conditional on sampling a single copy per species. To facilitate visualization, the internal branch uniting Species A and B is not drawn to scale.

for a three-taxon tree represent a worst-case scenario for confusing pseudoorthologs with orthologs.

DISCUSSION

Pseudoorthologs have long been feared for their possible detrimental effects on species tree inference. Removing these genes is difficult, and methods have relied on removing long branches (e.g., Yang and Smith 2014) or on the monophyly of other clades defined *a priori* (e.g., Siu-Ting et al. 2019), both of which may remove a substantial fraction of the data, and neither of which is guaranteed to remove all (or only) pseudoorthologs. Our results suggest that pseudoorthologs are unlikely to mislead phylogenetic inferences, given the assumptions of the model presented here. We find that pseudoorthologs are rare overall (Fig. 3a), and that pseudoorthologs with discordant topologies are expected to be much less common than genes with concordant topologies (Fig. 3b). Thus, our results suggest that regardless of the particular method used to identify single-copy orthologs, discordant pseudoorthologs are unlikely to be mistakenly sampled; thus, they are unlikely to pose a challenge to phylogenetics.

Even in the most problematic regions of parameter space considered here, pseudoorthologs are unlikely to mislead topological inferences. First, orthologs are still substantially more likely than pseudoorthologs, and the concordant topology is still more than 8X as likely as either of the two discordant topologies. Second, topological heterogeneity is recognized as common across the tree of life due to many processes (Bravo et al. 2019), and in this particular region of parameter space we expect discordance to be high due to incomplete lineage sorting (Fig. 4). Since discordance is likely to be high irrespective of the presence of pseudoorthologs, the use of methods that are robust to discordance (e.g., ASTRAL, Mirarab et al. 2014; Zhang et al. 2018) should be of utmost importance. Moreover, our modeling results corroborate previous findings that quartet methods such as ASTRAL are statistically consistent under a model of gene duplication and loss (Legried et al. 2020; Markin and Eulenstein 2020). Such methods rely on the fact that, for a rooted three-taxon species tree, the concordant topology is always the most frequent, which is exactly what we find here among all single-copy genes. Additionally, the probabilities of the two discordant topologies are always equal, suggesting that methods for species tree inference (Chifman and Kubatko 2014) and tests of introgression (Huson et al. 2005; Vanderpool et al. 2020) based on symmetry in number of the two discordant topologies should not be misled by the inclusion of pseudoorthologs. Thus, even in the most problematic regions of parameter space, quartet-based methods should not be misled by the presence of pseudoorthologs. Finally, in these regions of parameter space—and when there are more than three species in a tree—single-copy genes shared across all species are particularly rare (e.g., Supplementary Fig. S9b available on Dryad). Thus, we expect that researchers would be unable to sample many single-copy genes in such cases, and should therefore consider explicitly including paralogs in their data set to gain more phylogenetic markers (Smith and Hahn 2021).

In addition to changes in gene tree topologies, pseudoorthologs have different branch lengths than orthologs. Concordant pseudoorthologs are expected to have the longest internal branch lengths, with the expected branch length converging on $t_2 + 1/\lambda$ (the latter term representing the expected time to the duplication event) as the length of branch t_1 increases (Fig. 1d; Mendes and Hahn 2018). Discordant pseudoorthologs will have longer terminal branch lengths (Fig. 1e,f), but a shorter internal branch length. The expected internal branch length of discordant pseudoortholog will converge to $1/\lambda$ as the length of branch length t_1 increases, which may be either shorter or longer than the internal branch length t_2 of true orthologs. Since concordant pseudoorthologs are never expected to occur at a lower frequency than either of the discordant pseudoorthologs—and the internal branch is always longer by t_2—the total expected internal branch length supporting the true topology should always exceed that supporting either discordant topology. These results suggest that, with enough data, concatenation-based methods are also unlikely to be misled by pseudoorthologs. However, the expected branch lengths depend on more assumptions than do the calculated probabilities, as these calculations are conditioned only on a duplication occurring on branch t_1 and not on the presence of other necessary events. These calculations also depend upon the presence of only a single copy at the beginning of branch t_1, and relaxations of these assumptions may alter the expected branch lengths.

While the incidental inclusion of pseudoorthologs seems unlikely to affect inferences of species tree topology, many phylogenetic studies also aim to estimate concordance factors, nodal support, and branch lengths, and sometimes to test for the presence of introgression. Pseudoorthologs could lead to biased branch length estimates. Specifically, since internal branch lengths of concordant pseudoorthologs and external branch lengths of discordant pseudoorthologs are always expected to be longer than the corresponding branch lengths of orthologs, the presence of pseudoorthologs should lead to overestimates of branch lengths for methods that estimate branch lengths in substitutions per site. For methods that estimate branch lengths in coalescent units, pseudoorthologs should lead to underestimated branch lengths since they should introduce additional discordance relative to expectations under the multispecies coalescent model. For the same reason, the presence of pseudoorthologs may decrease measures of nodal support and concordance factors. However, the rarity of pseudoorthologs across most of parameter space (Fig. 3a) should minimize their effects on estimates of branch lengths, concordance factors, and nodal support values. Notably, the inclusion of pseudoorthologs should not affect inferences of introgression that rely on symmetries in minor site patterns or topologies (e.g., Huson et al. 2005; Vanderpool et al. 2020) since each of the two discordant topologies is equally likely.

We stress that the results presented here make a number of assumptions about the process of gene duplication and loss. Here, we discuss some of these assumptions and the potential effects of their violations on the probabilities of observing orthologs and pseudoorthologs. Our model assumes a relatively simple process of gene duplication and loss, with constant rates through time and across lineages. Higher rates of gene duplication and loss during certain time intervals could change the relative probabilities of orthologs, discordant pseudoorthologs, and concordant pseudoorthologs. For example, if rates of gene loss are higher immediately after gene duplication, then we would expect that the probability of completely losing one copy would be higher, and thus that the probability of orthologs would increase. Alternatively, if rates of gene loss were higher near the tips of the tree, then we might expect an increased probability of pseudoorthologs and an increased ratio of discordant to concordant topologies. However, it is difficult to construct a scenario in which either of these model violations leads to more overall discordant than concordant topologies.

Most of the results presented here have also assumed that there is a single gene copy at the beginning of branch t_1. However, an alternative scenario for pseudoorthologs is polyploidy, a special case of gene duplication and loss in which the entire genome is duplicated (Otto 2007). Polyploidy can lead to increased probabilities of pseudoorthologs conditional on sampling a single gene per species, though the probability of sampling single-copy genes will also be much lower in this scenario. When the taxa considered are autopolyploids, or polyploid taxa for which both subgenomes come from the same species, then we need only condition on the polyploidy event having occurred at some point prior to branch t_1 for our model to apply. We used simulations to explore this scenario, finding that while WGDs decrease the overall probability of orthologs and increase the probability of pseudoorthologs conditional on sampling a single copy, the concordant topology is still always more likely than either discordant topology (Supplementary Table S5 available on Dryad). Notably, this is a relatively simple model of WGD, and more complex scenarios could include nested WGDs or biased gene retention across genome copies. Nested WGDs could lead both to increased probabilities of pseudoorthologs and to increased ratios of discordant to concordant pseudoorthologs. However, it is still difficult to construct a scenario in which the probability of discordant pseudoorthologs would exceed that of concordant pseudoorthologs, since either type of pseudoortholog can result from the same events on different copies. Even with biased gene retention, the only way to generate more discordant than concordant topologies is if gene retention varies across copies in a species-specific manner. In our simulations, we only explored autopolyploidy. Allopolyploids are polyploid taxa in which each subgenome comes from a different species (Otto 2007). In this case, there are two "concordant" trees, depending on which parental genome is considered. Paralogs from the different parental species are sometimes lost in a biased fashion (e.g., Chang et al. 2010). The main consequence of biased loss is that one set of orthologous species relationships will be retained over the other (Thomas et al. 2017). Of course, even more so in polyploids than in other taxa, excessive filtering to remove putative pseudoorthologs will decrease the amount of available data. Furthermore, the number of single-copy gene families will be limited in cases of polyploidy, and using multiple-copy gene families for phylogenetic inference is likely the ideal approach in this case (Smith and Hahn 2021).

Based on the results presented here, pseudoorthologs are unlikely to be the frequent cause of problems in phylogenetic inference. How then can we explain previous results that claim to demonstrate the negative effects of pseudoorthologs on phylogenetic inference? Some studies have found differences in trees inferred from data sets filtered using different ortholog detection methods (Altenhoff et al. 2019b; Siu-Ting et al. 2019; Cheon et al. 2020). However, most ortholog detection methods are unlikely to remove pseudoorthologs, and thus comparisons of these methods are not informative with respect to the effects of pseudoorthologs. In cases where researchers specifically aim to exclude pseudoorthologs (e.g., Siu-Ting et al. 2019) and inferences differ across filtered data sets, more stringent filtering may remove problematic sequences other than pseudoorthologs—for example, alternative isoforms or error-prone sequences. While such filtering may improve phylogenetic inference, this improvement cannot be attributed to the removal of pseudoorthologs. Paralogs included in data sets of putative single-copy orthologs may also not be true pseudoorthologs. In large genomic and transcriptomic data sets, putative pseudoorthologs may instead be paralogs for which, for technical reasons, different copies were assembled in each species (as appears to be the case in Brown and Thomson 2017). In our study, we assume that all pseudoorthologs are a result of true biological loss, rather than sampling artifacts; when random sampling occurs, the overall probability of pseudoorthologs will be higher than observed here. However, even in the extreme case, where a single paralog is sampled at random from each species, quartet-based methods appear to perform well when enough data are available (Legried et al. 2020; Markin and Eulenstein 2020; Yan et al. 2021). With respect to the results presented here, randomly sampling species in the present increases the probability of sampling pseudoorthologs. It may also change the proportions of discordant and concordant topologies, since the probability of losing any particular copy due to sampling will not depend on branch lengths. However, it remains true that the discordant topology should never be more probable than the concordant topology if sampling is random.

Overall, our results suggest that pseudoorthologs are not likely to mislead phylogenetic inference. Pseudoorthologs are rare across reasonable regions of parameter space, and even in the most extreme scenarios considered, the concordant topology is always expected to be the most common. These results should reassure researchers who are well aware of the difficulties of identifying and removing these genes from phylogenomic data sets and should encourage researchers to focus their filtering efforts elsewhere, for example, on detecting and removing assembly artifacts or on poorly aligned sequences.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.573n5tb72.

Funding

This work was supported by a National Science Foundation postdoctoral fellowship to MLS (DBI-2009989) and an National Science Foundation grant to MWH (DEB-1936187).

Acknowledgments

We thank Rafael Guerrero for helpful discussion, as well as three reviewers and Stephen Smith for constructive comments. This work was supported by a National Science Foundation postdoctoral fellowship to MLS (DBI-2009989) and an NSF grant to MWH (DEB-1936187).

Data Availability

All code for calculating the probabilities of orthologs and pseudoorthologs is reproduced and commented in Supplementary Appendix A available on Dryad.

References

Altenhoff A.M., Gil M., Gonnet G.H., Dessimoz C. 2013. Inferring hierarchical orthologous groups from orthologous gene pairs. PLoS One 8:e53786.

Altenhoff A.M., Glover N.M., Dessimoz C. 2019a. Inferring orthology and paralogy. In: Anisimova M., editor. Evolutionary genomics: statistical and computational methods. New York (NY): Springer. p. 149–175.

Altenhoff A.M., Levy J., Zarowiecki M., Tomiczek B., Warwick Vesztrocy A., Dalquen D.A., Müller S., Telford M.J., Glover N.M., Dylus D., Dessimoz C. 2019b. OMA standalone: orthology inference among public and custom genomes and transcriptomes. Genome Res. 29:1152–1163.

Altenhoff A.M., Schneider A., Gonnet G.H., Dessimoz C. 2011. OMA 2011: orthology inference among 1000 complete genomes. Nucleic Acids Res. 39:D289–D294.

Arvestad L., Berglund A.-C., Lagergren J., Sennblad B. 2003. Bayesian gene/species tree reconciliation and orthology analysis using MCMC. Bioinformatics 19:i7–i15.

Arvestad L., Berglund A.-C., Lagergren J., Sennblad B. 2004. Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. Proc. Eighth Annu. Int. Conf. Comput. Mol. Biol. - RECOMB 04:326–335.

Bailey N.T.J. 1964. The elements of stochastic processes with applications to the natural sciences. New York (NY): Wiley.

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

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

Chang P.L., Dilkes B.P., McMahon M., Comai L., Nuzhdin S.V. 2010. Homoeolog-specific retention and use in allotetraploid *Arabidopsis suecica* depends on parent of origin and network partners. Genome Biol. 11:R125.

Cheon S., Zhang J., Park C. 2020. Is phylotranscriptomics as reliable as phylogenomics? Mol. Biol. Evol. 37:3672–3683.

Chifman J., Kubatko L. 2014. Quartet inference from SNP data under the coalescent model. Bioinformatics 30:3317–3324.

Doolittle W.F., Brown J.R. 1994. Tempo, mode, the progenote, and the universal root. Proc. Natl. Acad. Sci. USA 91:6721–6728.

Dunn C.W., Howison M., Zapata F. 2013. Agalma: an automated phylogenomics workflow. BMC Bioinformatics 14:330.

Ebersberger I., Strauss S., von Haeseler A. 2009. HaMStR: profile hidden Markov model based search for orthologs in ESTs. BMC Evol. Biol. 9:157.

Emms D. M., Kelly S. 2015 OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 20:1–14.

Fernández R., Gabaldon T., Dessimoz C. 2020. Orthology: definitions, prediction, and impact on species phylogeny inference. In: Scornavacca C., Delsuc F., Galtier N., editors. Phylogenetics in the Genomic Era. Open access book. p. 2.4:1–2.4:14.

Fernández R., Kallal R.J., Dimitrov D., Ballesteros J.A., Arnedo M.A., Giribet G., Hormiga G. 2018. Phylogenomics, diversification dynamics, and comparative transcriptomics across the spider tree of life. Curr. Biol. 28:1489–1497.

Fitch W.M. 1970. Distinguishing homologous from analogous proteins. Syst. Zool. 19:99–113.

Gernhard T. 2008. The conditioned reconstructed process. J. Theor. Biol. 253:769–778.

Huson D.H., Klöpper T., Lockhart P.J., Steel M.A. 2005. Reconstruction of reticulate networks from gene trees. In: Miyano S., Mesirov J., Kasif S., Istrail S., Pevzner P.A., Waterman M., editors. Research in Computational Molecular Biology. Berlin, Heidelberg: Springer Berlin Heidelberg. p. 233–249.

Kallal R.J., Fernández R., Giribet G., Hormiga G. 2018. A phylotranscriptomic backbone of the orb-weaving spider family Araneidae (Arachnida, Araneae) supported by multiple methodological approaches. Mol. Phylogenet. Evol. 126:129–140.

Kapli P., Yang Z., Telford M.J. 2020. Phylogenetic tree building in the genomic age. Nat. Rev. Genet. 21:428–444.

Koonin E.V. 2005. Orthologs, paralogs, and evolutionary genomics. Annu. Rev. Genet. 39:309–338.

Legried B., Molloy E.K., Warnow T., Roch S. 2020. Polynomial-time statistical estimation of species trees under gene duplication and loss. J. Comput. Biol. 28:452–468.

Li L. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13:2178–2189.

Maddison W.P. 1997. Gene trees in species trees. Syst. Biol. 46:523–536.

Mallo D., de Oliveira Martins L., Posada D. 2016. SimPhy: phylogenomic simulation of gene, locus, and species trees. Syst. Biol. 65:334–344.

Markin A., Eulenstein O. 2021. Quartet-based inference is statistically consistent under the unified duplication-loss-coalescence model. Bioinformatics. 37:4064–4074.

Mendes F.K., Hahn M.W. 2018. Why concatenation fails near the anomaly zone. Syst. Biol. 67:158–169.

Mendes F.K., Vanderpool D., Fulton B., Hahn M.W. 2020. CAFE 5 models variation in evolutionary rates among gene families. Bioinformatics. 36:5516–5518.

Mirarab S., Reaz R., Bayzid Md.S., Zimmermann T., Swenson M.S., Warnow T. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30:i541–i548.

Otto S.P. 2007. The evolutionary consequences of polyploidy. Cell 131:452–462.

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

Scornavacca C., Delsuc F., Galtier N. 2020. Phylogenetics in the genomic era. Open access book. Available from: https://hal.inria.fr/PGE/.

Siu-Ting K., Torres-Sánchez M., San Mauro D., Wilcockson D., Wilkinson M., Pisani D., O'Connell M.J., Creevey C.J. 2019. Inadvertent paralog inclusion drives artifactual topologies and timetree estimates in phylogenomics. Mol. Biol. Evol. 36:1344–1356.

Smith M.L., Hahn M.W. 2021. New approaches for inferring phylogenies in the presence of paralogs. Trends Genet. 37:174–187.

Thomas G.W.C., Ather S.H., Hahn M.W. 2017. Gene-tree reconciliation with MUL-trees to resolve polyploidy events. Syst. Biol. 66:1007–1018.

Vanderpool D., Minh B.Q., Lanfear R., Hughes D., Murali S., Harris R.A., Raveendran M., Muzny D.M., Hibbins M.S., Williamson R.J., Gibbs R.A., Worley K.C., Rogers J., Hahn M.W. 2020. Primate phylogenomics uncovers multiple rapid radiations and ancient interspecific introgression. PLoS Biol. 18:e3000954.

Yan Z., Smith M.L. Du P., Hahn M.W., Nakhleh L. 2021. Species tree inference methods intended to deal with incomplete lineage sorting are robust to the presence of paralogs. Syst. Biol. 71:367—381.

Yang Y., 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.

Zhang C., Rabiee M., Sayyari E., Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics 19:153.

Zhang C., Scornavacca C., Molloy E.K., Mirarab S. 2020. ASTRAL-Pro: quartet-based species-tree inference despite paralogy. Mol. Biol. Evol. 37:3292–3307.