reconcILS: A gene tree-species tree reconciliation algorithm that allows for incomplete lineage sorting

Sarthak Mishra[1,*], Megan L. Smith[2], and Matthew W. Hahn[1,3]

[1] Department of Computer Science, Indiana University, Bloomington IN 47405, USA
[2] Department of Biological Sciences, Mississippi State University, Starkville MS 39762, USA
[3] Department of Biology, Indiana University, Bloomington IN 47405, USA

*Corresponding author. E-mail: samishr@iu.edu

Abstract. Reconciliation algorithms provide an accounting of the evolutionary history of individual gene trees given a species tree. Many reconciliation algorithms consider only duplication and loss events (and sometimes horizontal transfer), ignoring effects of the coalescent process, including incomplete lineage sorting (ILS). Here, we present a new algorithm for carrying out reconciliation that accurately accounts for ILS by treating it as a series of nearest neighbor interchange (NNI) events. For discordant branches of the gene tree identified by last common ancestor (LCA) mapping, our algorithm recursively chooses the optimal history by comparing the cost of duplication and loss to the cost of NNI and loss. We demonstrate the accuracy of our new method, which we call `reconcILS`, using a new simulation engine (`dupcoal`) that can accurately generate gene trees produced by the interaction of duplication, ILS, and loss. We show that `reconcILS` is much more accurate than models that ignore ILS, and at least as accurate or better than the leading method that can model ILS, duplication, and loss. We demonstrate the use of `reconcILS` by applying it to a dataset of 23 primate genomes, highlighting its accuracy compared to standard methods in the presence of large amounts of incomplete lineage sorting.
Availability: `reconcILS` is implemented in Python 3 and is available at `https://github.com/smishra677/reconcILS`. The `dupcoal` simulator is implemented in Python 3 and is available at `https://github.com/meganlsmith/dupcoal`

Keywords: reconciliation · duplication · loss · coalescent · hemiplasy · incomplete lineage sorting.

Introduction

The sequencing of a large number of genomes from many different species has highlighted the dynamic nature of genes: they are often duplicated, lost, or moved between species via introgression (no gain of an extra copy) or horizontal gene transfer (gain of an extra copy). One of the main goals of evolutionary and comparative genomics has been to quantify this dynamism, both in terms of the number of events of each type and their timing. A popular approach for quantifying both measures is gene tree-species tree reconciliation (reviewed in Boussau and Scornavacca 2020). Given a gene tree topology and a species tree topology, reconciliation algorithms attempt to infer the evolutionary events necessary to explain any discordance in the respective topologies. Many different reconciliation algorithms have been developed, most often considering only duplication and loss (so-called "DL" models; e.g Goodman et al. 1979; Page 1994; Guigo et al. 1996; Page and Charleston 1997; Chen et al. 2000; Zmasek and Eddy 2001; Durand et al. 2006), but also sometimes considering duplication, transfer, and loss ("DTL" models; e.g. Bansal et al. 2012; Szöllősi et al. 2013; Jacox et al. 2016; Morel et al. 2020).

Despite the success—and widespread use—of DL and DTL reconciliation algorithms, they do not consider gene tree discordance due to the coalescent process. The coalescent is a model that describes how different sequences within a population are related (Kingman 1982; Hudson 1983; Tajima 1983), and is relevant even when considering sequences from different species because these must also coalesce in shared ancestral populations. Importantly, gene trees and species trees can be discordant solely due to coalescence, in a process often called incomplete lineage sorting (ILS) or deep coalescence (Maddison 1997). Standard DL and DTL algorithms will incorrectly infer one duplication and three losses for every discordance only due to ILS (Hahn 2007), resulting in misleading results when applying reconciliation methods to trees with ILS (e.g. Stull et al. 2021). Duplication, loss, and coalescence can also interact in complex and non-intuitive ways (Rasmussen and Kellis 2012; Mallo et al. 2014; Li et al. 2021), making it almost impossible to simply exclude gene trees discordant due to ILS from an analysis.

A few different approaches to reconciliation have been used to deal with discordance due to coalescence. An early approach allowed short branches in the species tree—which can lead to ILS—to be collapsed (Vernot et al. 2008; Stolzer et al. 2012). The use of non-binary species trees means that such algorithms will not mistake ILS alone for duplication and loss; however, they do not account for interactions between events. Chan et al. (2017) introduced a reconciliation algorithm that included duplication, loss, transfer, and ILS, but it assumed that new duplicates were completely linked to their parental locus (i.e. no recombination between them). Given that the majority of new gene duplicates are not completely linked (Schrider and Hahn 2010), such an assumption limits the number of genes that can be analyzed. Wu et al. (2014) developed a parsimony-based algorithm (DLCpar) that models duplication, loss, and ILS; this approach is similar to the probabilistic DLCoalRecon algorithm (Rasmussen and Kellis 2012), but runs much faster and without the need to specify duplication and loss rates ahead of time.

Here, we introduce a new algorithm for finding the most parsimonious reconciliation of a gene tree and a species tree in the presence of duplication, loss, and ILS. Our goal is to improve upon existing methods by increasing speed and accuracy, as well as by incorporating more coalescent processes (i.e. ILS is not the only cause of discordance due to the coalescent). In addition, although DLCpar has been shown to perform well (Wu et al. 2014; Du et al. 2021), it is not straightforward to extend its algorithm to other tools that also have reconciliation steps (e.g. Thomas et al. 2017; Zhang et al. 2020). We hope that the algorithm introduced here will be easily adaptable to such scenarios. We demonstrate the use of this new algorithm by applying it to 23 whole primate genomes, a dataset that has previously been shown to contain a large amount of discordance due to ILS (Vanderpool et al. 2020).

Methods

Multilocus multispecies coalescent simulator

We begin with a description of the simulation program introduced here, `dupcoal`, as it allows us to explain the biological model underlying both the simulations and inferences.

The model we use closely resembles the multilocus multispecies coalescent (MLMSC) model of Li et al. (2021). This model treats each individual locus as evolving via the multispecies coalescent (MSC) model, in

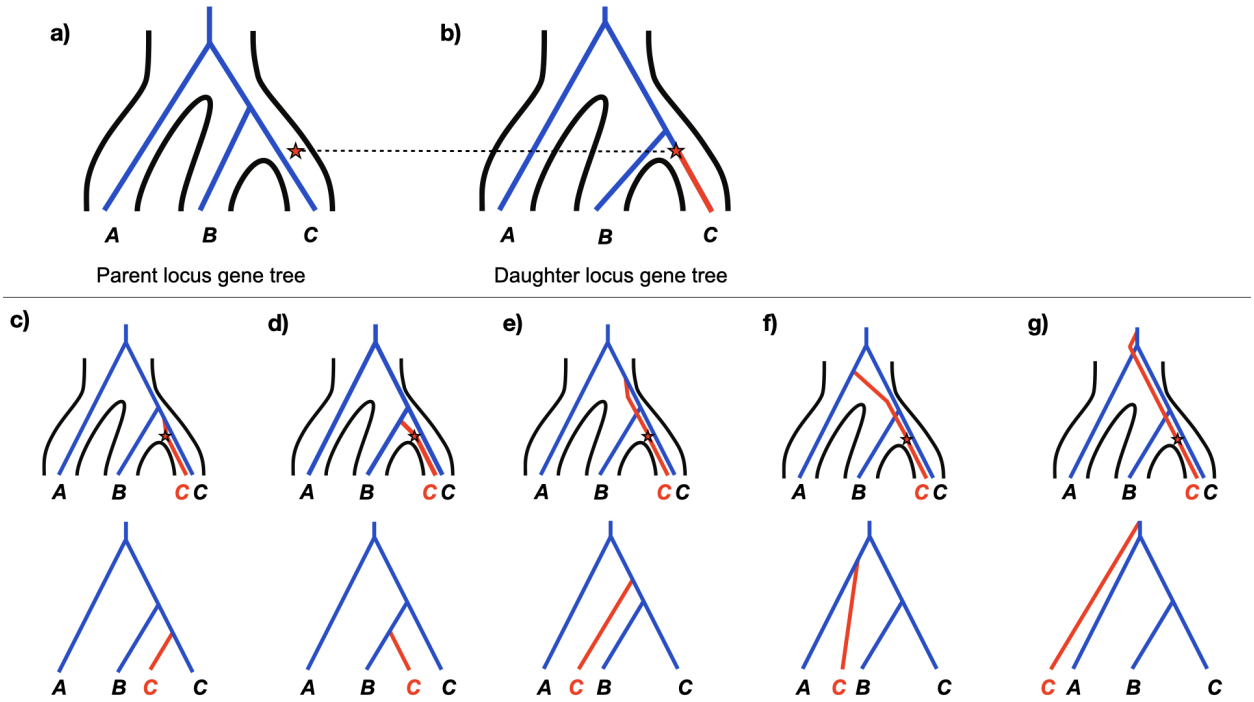

Fig. 1. The multilocus multispecies coalescent model generates full gene trees. a) A gene tree at the parent locus (blue) is drawn from the species tree (outlined in black) via the multispecies coalescent (MSC) model. A duplication event (red star) is drawn from a birth-death model over the species tree, in this case occurring in the population ancestral to species B and C. b) A gene tree at the daughter locus is drawn from the MSC. The duplication is placed on this gene tree at the time at which it occurred in the species tree (shown as a dotted line), choosing one of the two possible branches at random. The inserted DNA now takes on the gene tree branch labeled in red. c)-g) Possible ways that the parent and daughter loci coalesce to form the full gene tree. In all panels the star and red branch below it come from the daughter gene tree, and the blue tree comes from the parent gene tree. The coalescent process joining the two trees is shown as a red branch above the star. The trees in the bottom row are simply unfolded views of the full gene trees.

which discordance between individual gene trees and the species tree can arise only via incomplete lineage sorting. The MLMSC further models relationships among these loci, in addition to losses at any locus.

To illustrate how duplications and losses occur, and how relationships among loci are generated in `dupcoal`, we use a simplified example (a full description of the simulation algorithm can be found at `https://github.com/meganlsmith/dupcoal`). As a first step, a gene tree is generated via the MSC at the "parent" locus (Figure 1a); this tree exists regardless of whether subsequent duplications occur. Duplications and losses are then generated via a birth-death process on the species tree, with each event occurring at a specific time on a specific branch of the species tree (one such duplication is represented as a star in Figure 1a). For each duplication event, a gene tree is generated via the MSC at a "daughter" locus (Figure 1b). This tree represents the history of the locus at which the duplicated DNA will be inserted, and exists with or without this duplication event; for now we assume that this locus is unlinked from the parent locus. The mutation representing the duplication event is then placed on the gene tree at the daughter locus at the same time that it occurred on the species tree—if multiple branches of the daughter gene tree exist at this point in time on the relevant branch, the mutation is assigned to one at random (star in Figure 1b). If the mutation occurs on a discordant branch of the daughter gene tree (that is, one that does not exist in the species tree topology), this is referred to as "hemiplasy" or "copy-number hemiplasy." Such events are not simulated by SimPhy (Mallo et al. 2016), a popular platform for carrying out gene tree simulations with duplications and losses. Regardless of where exactly the mutation occurs on the daughter gene tree, the duplication now takes on the topology and branch lengths of the tree below this point (red branch in Figure 1b).

The process described thus far has produced two gene trees—one at the parent locus and one at the daughter locus—either or both of which may be discordant due to ILS. Because gene duplication is fundamentally a process by which an existing haplotype in a population at one locus is inserted into another location, the inserted DNA at the new daughter locus carries with it a history of coalescence with all of the other haplotypes at the parent locus. This coalescent process joining the two trees can cause additional discordance, even among gene trees that individually cannot be discordant (e.g. when the species tree has only two species). We illustrate this coalescent process here with a simple example.

Figures 1c-g show the five outcomes possible when joining the parent and daughter trees shown in Figures 1a and 1b. Figure 1c is the simplest outcome, as the daughter tree joins the parent tree on the same branch that the mutation occurred on (i.e. the one leading to species C); this is the tree that is assumed to be produced by duplication under standard DL reconciliation. Figures 1d and 1e represent cases where the daughter tree joins the parent tree on a different branch, each of which would require one NNI move to recover the tree topology in Figure 1c. Finally, Figures 1f and 1g represent more extreme cases, in which coalescence between parent and daughter further up the tree requires two NNI moves to recover the topology in Figure 1c.

It should be emphasized here that all of the topologies presented in Figure 1 involve only a single duplication and no losses, and neither the parent tree nor the daughter tree is discordant due to ILS. In fact, the examples in Figures 1d and 1e are not technically due to ILS either, since coalescence is occurring in the same population of the species tree (the ancestor of species B and C); in addition, in this scenario these two trees are actually equiprobable with the topology in Figure 1c. These simple examples demonstrate that the vagaries of the coalescent process lead to a thicket of different tree topologies.

Losses are also simulated on the species tree and placed on each gene tree. Loss events are placed on branches of gene trees in the same manner as duplications, with all branches below such mutations removed. Together, these processes make up the steps by which our simulator can generate full gene trees (i.e. ones with the parent and all daughter trees joined together). We used functions from DendroPy (Sukumaran and Holder 2010) to generate trees under the MSC and to manipulate parent and daughter trees; we used functions from NumPy (Harris et al. 2020) to draw coalescent times, duplication times, and loss times from exponential distributions.

`reconcILS`

Inputs and outputs. Given a rooted, bifurcating species tree, S, and a rooted, bifurcating gene tree, G, as input, `reconcILS` outputs the minimum number of duplications, losses, and NNIs required to reconcile G and S given event costs. Inferred events are all also assigned to nodes in the gene tree.

Main algorithm. An overview of `reconcILS` is given in Algorithm 1 (see Supplementary Materials). Our approach combines multiple sub-approaches that are commonly used in DL reconciliation and other phylogenetic inferences. As a first step, we use last common ancestor (LCA) mapping (Zmasek and Eddy 2001) to efficiently map each node in G to a node in S (Figure 2a).

Unlike in standard DL methods, LCA mapping here is used to initialize each round of our algorithm—and is sometimes used iteratively within a round—without representing the final reconciliation. Instead, this step identifies cases in which multiple nodes in the gene tree map to an identical node of the species tree. For instance, in Figure 2a nodes G_1 and G_2 in the gene tree both map to node S_1 in the species tree. We therefore call S_1 a "multiply mapped" node. Additionally, we refer to the branch in the gene tree bracketed by nodes G_1 and G_2 as a "discordant" branch (labeled K_1 in Figure 2a); such branches do not exist in the species tree. The focus of our method is in determining whether discordant branches should be resolved via NNI and loss (Figure 2b) or via duplication and loss (Figure 2c). Choosing among events (NNI, duplication, and loss) is carried out greedily for each branch, with no consideration of a globally optimal set of events among all discordant branches in the gene tree. Consequently, after choosing a set of events for one branch, the process repeats itself until there are no multiply mapped nodes and no discordant branches. Duplications implied by multiply mapped tip nodes in the species trees are inferred via standard methods; inferences of losses are explained below.

Here it may also be important to point out why we are using NNI rearrangements to model ILS. NNI has been implemented in reconciliation algorithms before, but in these cases it is used to correct gene trees that may have been inferred incorrectly (e.g. Chen et al. 2000; Chaudhary et al. 2011; Nguyen et al. 2013). We use NNI because the effects of ILS on gene trees can be modeled as a series of NNI rearrangements. For three species undergoing ILS, there are two possible discordant topologies, each of which is one NNI move away from the concordant topology. For four lineages undergoing ILS, there are 14 possible discordant topologies, each of which are between one and three NNI moves away from the concordant topology. (Note that it is the number of lineages undergoing ILS at a specific "knot" on the tree that matters, not the total number of tips in the tree.) Our algorithm takes advantage of this relationship between NNI and ILS in order to reconcile gene trees; other tree rearrangements (e.g. subtree-prune-regraft) could of course be used as an alternative, but would be much less efficient.

Traversing trees. The high-level description given above does not provide details about several important steps in our algorithm. In this section and the next two we provide some more information.

After the initial LCA mapping, our algorithm performs a preorder traversal of the species tree. For each species tree node, the algorithm checks how many gene tree nodes are mapped to it. If there are fewer than 2 mappings, the algorithm invokes a function that asks whether speciation or loss can be inferred. Nodes with 1 mapping are speciation events—these do not incur a cost. Nodes that have 0 mappings represent losses. Our function asks whether the parent node of a node with 0 mappings itself has 0 mappings. If not, a loss is inferred; if so, no event is recorded, because the loss must have occurred further up the tree. If there are 2 or more mappings the species tree node is multiply mapped, and a function to carry out "local reconciliation" is called. Local reconciliation is described in more detail in the next section, but here can simply be described as determining the events that have generated the multiple mappings to a single species tree node.

If there are only two adjacent nodes in the gene tree mapped to a node in the species tree (i.e. one discordant branch), then the method proceeds as described below. However, when there are more than two nodes mapped to an internal species tree node—and therefore more than one discordant branch in the gene tree—we must decide which branch to reconcile with the multiply mapped node first. We have found some cases where the order in which the branches are resolved matters, with only one of the choices leading to an optimal reconciliation. Algorithm 2 (see Supplementary Materials) describes how we choose which discordant branch to reconcile first. This algorithm calculates the "bipartition cost" of each possible branch—the number of bipartitions and losses induced by NNI moves applied to each branch (Figure S1). Such a calculation can be carried out quickly, and by choosing the branch with the lowest bipartition cost, we can more accurately proceed with reconciliation.

Once reconciliation for a multiply mapped node is carried out, our algorithm begins a new round. In each round we repeat LCA mapping and continue traversing the species tree, counting mappings at each node.

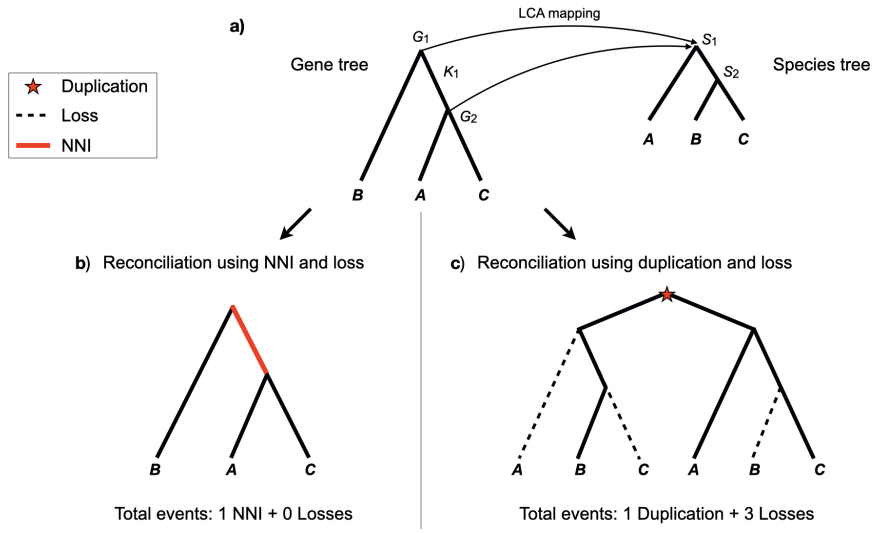

Fig. 2. Summary of `reconcILS`. a) The initial step carries out LCA mapping from the gene tree to the species tree. Here, nodes G_1 and G_2 in the gene tree both map to node S_1 in the species tree; we therefore refer to S_1 as a multiply mapped node. Additionally, G_1 and G_2 together define a discordant branch, K_1, in the gene tree. For clarity, all mappings involving tip nodes have been excluded. b) The algorithm attempts to reconcile the discordant branch using NNI and loss. Here, reconciliation can be carried out with 1 NNI and 0 losses. c) The algorithm attempts to reconcile the discordant branch using duplication and loss. Here, reconciliation can be carried out with 1 duplication and 3 losses. `reconcILS` must decide between the reconcilations in b) and c) for every discordant branch.

Choosing among events. In order for our algorithm to more efficiently and optimally choose among different evolutionary events (e.g. Figure 2b vs 2c), we have introduced the concept of a "local reconciliation" for a multiply mapped node. Here, a species tree node is locally reconciled if the number of gene tree nodes mapped to it, m, is reduced; for internal nodes of the species tree, this by definition also reduces the number of discordant branches in the gene tree.

The use of local reconciliation has several advantages. Most importantly, it provides a criterion for moving forward: we can determine when reconciliation has been achieved within a single round of the algorithm. It also makes it clear that there are some scenarios that can only be reconciled using duplication and not NNI (e.g. a multiply mapped tip node), in which case choosing events is straightforward. In a single round of our algorithm, we attempt to locally reconcile an identified multiply mapped node using both NNI+loss and duplication+loss (Figure 2).

Note that, given our definition, a species tree node can be locally reconciled by NNI while still remaining multiply mapped; e.g., if $m = 3$ is reduced to $m = 2$, the node still requires more reconciliation. In this case our algorithm will carry out up to $m - 1$ NNI moves within a single round, as long as each one results in a local reconciliation. For example, starting with $m = 3$, the first NNI move will locally reconcile a node if one of the two NNI topologies reduces the number of mappings to the species tree (we have never observed a case where both NNI topologies reduce this number). In this case, $m = 2$, and we now carry out another NNI move on the updated gene tree. Only one move needs to be examined at this point, however, since the back-NNI rearrangement would return us to the starting state. The cost of all of these NNI moves is then compared to the cost of duplication and loss at the end of the round. Duplication+loss naturally allows for more than one of each type of event for a local reconciliation, and is therefore not handled differently when $m > 2$.

As with all parsimony-based reconciliations, `reconcILS` requires a cost for each event type: one for duplications (C_D), one for losses (C_L), and one for NNIs (C_I). Reconciliations can be very sensitive to costs, such that maximum parsimony solutions under one set of costs can be very different than under another set of costs. Although there are few ways to determine costs independently (but see Discussion), there are some approaches for finding the Pareto-optimal costs within a reconciliation context (e.g. Libeskind-Hadas et al. 2014; Mawhorter et al. 2019). Here, as in most commonly used reconcilation algorithms, we allow the

user to set costs. The default costs within `reconcILS` are set to $C_D = 1.1$, $C_L = 1$, and $C_I = 1$. (The default costs within DLCpar are $C_D = 2$, $C_L = 2$, and $C_I = 1$.) It is also important to note that DLCpar is specifying a cost of ILS, while `reconcILS` is specifying a cost for each NNI induced by even a single ILS event. This distinction will become important when we test the accuracy of each method below, as we will be inferring slightly different events with the two methods (we use the same costs in both methods for these comparisons).

Finally, at the end of a round, `reconcILS` will sum the cost of all events in order to choose the optimal local reconciliation. If there are equally optimal reconciliations, `reconcILS` will choose the duplication+loss solution.

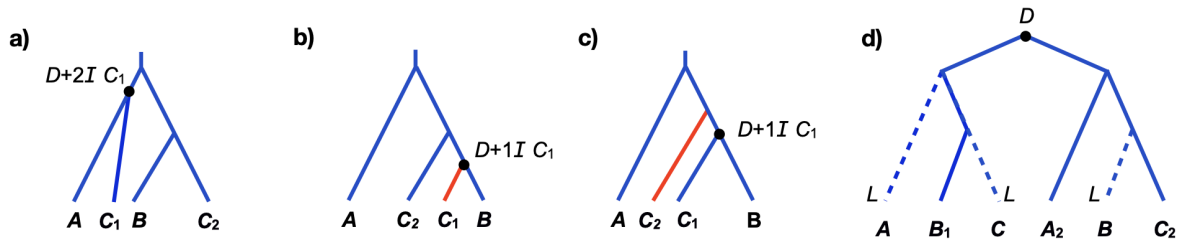

Fig. 3. Labeling reconciled gene trees. a) One possible reconciliation of the gene tree from Figure 1f labeled by `reconcILS`. The single ILS node (I) has been labeled with a branch that has been rearranged by NNI (C_1), the number of NNIs it has undergone (2), and the fact that it also represents a duplication (D). b) The gene tree from Figure 1d labeled by `reconcILS`. This tree is correctly labeled, but note that it is the same topology as in panel c (see main text for further explanation). c) The gene tree from Figure 1e labeled by `reconcILS`. Note that the program has labeled the bottom-most node flanking the discordant branch with D and I, although that is not the true history. d) A gene tree with only duplications and losses (L) labeled by `reconcILS`. Internal branches could also be labeled with L, though none are shown here.

Labeling a reconciled gene tree. A full reconciliation of a gene tree, G, and species tree, S, will have locally reconciled all multiply mapped nodes. The total number of duplications, losses, and NNIs required for this reconciliation is the sum of each event type across all branches; this number is reported by `reconcILS`. We also aim to label the gene tree with these events—identifying where each event occurred in history—as this is an important output of reconciliation algorithms. As `reconcILS` traversed the species tree finding optimal local reconciliations, it also recorded the number and location of all events. Labeling therefore consists of placing these events on the gene tree.

Traditional DL reconciliation algorithms label gene tree nodes as either D (for duplication) or S (for speciation); they may also label the location of losses, L. Algorithms including ILS should be able to additionally label nodes as I. However, there is no standard way to do this labelling, largely because there is no exact map between gene tree nodes and species tree nodes under ILS. In this sense, the labeling problem for algorithms including ILS differs fundamentally from labeling under DL and DTL algorithms. Here, we have created our own labeling system for ILS, one that we think is both understandable and easy to work with.

Our solution for placing the I label on nodes of gene trees in `reconcILS` is to denote both a daughter branch affected and the number of NNI moves that were associated with this branch. A few examples are shown in Figure 3, which also helps to illustrate why these choices were made. Figure 3a labels the gene tree from Figure 1f—with more complete tip names—with output from `reconcILS`. As can be seen, this tree has one node (denoted with a small black dot for clarity) labeled with both duplication and ILS. Further, the labeling tells you which daughter branch was reconciled with NNI (C_1), and how many NNI moves were required to reconcile it (2). Without such information, simply labeling the node with I would not have distinguished histories involving coalescent events along the A branch from those along the C_1 branch.

There are two important caveats to this approach, one perhaps more obvious than the other. First, a specific labeling is associated with a specific reconciliation. That is, if we had found a different optimal

reconciliation for the tree in Figure 3a (for instance, one duplication and two losses), we would have had different labeled events. So, as might be expected, a labeling is always reported in the context of a particular reconciliation.

Second, for some ILS events it is not always clear which of the two nodes flanking a discordant branch to label with I: this is why there is no exact map between gene tree node labels and species tree nodes under ILS. For a discordant branch, we have arbitrarily set `reconcILS` to always choose the node closest to the tips as I. In some cases this choice should not matter very much. Consider the toy reconciliation in Figure 2b—should we have placed the NNI event on node G_1 or on node G_2? A branch undergoes NNI, not a node, so it is not clear which node flanking such a branch should be chosen; in this case the choice does not seem to imply any difference in evolutionary histories. In contrast, consider the reconciliations shown in Figure 3b and 3c—these represent the output of `reconcILS` applied to the gene trees in Figure 1d and 1e, respectively. While we have left the daughter tree branch in red for clarity, it is important to note that these trees are identical topologically, which is why `reconcILS` has labelled them in exactly the same way. The algorithm has no way to "see" the red branches, and has consequently simply picked the bottom-most node in both cases as the one to label as having undergone ILS and duplication, even though this is not precisely correct in Figure 3c. At the moment this is a limitation of `reconcILS` (and all methods attempting to label gene trees with ILS events), but we mention other possible solutions in the Discussion.

`reconcILS` also labels duplications and losses in the absence of ILS. Figure 3d shows how both individual duplications and losses are labelled: observe that neither requires a label with a specified number of events (like the I label), nor does either need to specify which branch they act on. `reconcILS` places the label L on any node in the implied gene tree that is not represented by a gene copy; for example, generic tip nodes A, C, and B in Figure 3d do not exist in the input gene tree itself, but are labelled with an L. `reconcILS` is able to do this because every duplication event is represented internally by generating a new copy of the species tree below the duplicated node—if losses occur on these implied trees, we can track their timing as well. Note that internal nodes can also be labelled with an L, so that the timing of loss events must be inferred by comparing the state of parent and daughter nodes (i.e. losses on internal branches will lead the descendant internal node and all of its daughter nodes to be labelled with an L). Finally, any node without a label can be inferred to be a speciation node.

Results

Assessing accuracy using simulations

To test the accuracy and speed of `reconcILS` compared to other methods, we used our simulator (`dupcoal`) to generate gene trees evolving under ILS, duplication, and loss. Data were simulated on a species tree with topology (A:1.5,(B:0.5,C:0.5):1); branch lengths are measured in coalescent units. Given this tree, we expect single-copy gene trees to be concordant approximately 75% of the time. Simulated rates of duplication and loss were $\lambda = 0.3$ and $\mu = 0.3$, respectively, except where stated otherwise. The default costs of `reconcILS` were used for all analyses of simulated data for all methods ($C_D = 1.1$, $C_L = 1$, $C_I = 1$). Note that `reconcILS`, `dupcoal`, and ETE3 are written in Python 3, while DLCpar is written (and was run) in Python 2.

Accuracy of standard DL methods. As a simple test of DL reconciliation algorithms—which do not allow for ILS—we used the DL method implemented in ETE3 (Huerta-Cepas et al. 2016). We ran simulations with and without duplications, but always with the loss rate set to 0; this allows us to see incorrectly inferred losses more easily. As expected (Hahn 2007), in simulated trees with discordance due to ILS but no duplication or loss, ETE3 always inferred 1 extra duplication and 3 extra losses (Figure S2a). In cases with discordance due to both coalescence and duplication, but no loss, ETE3 again inferred a slight excess of duplications, but a large excess of losses (Figure S2b).

Accuracy of `reconcILS` and DLCpar. We ran two reconciliation methods that allow for ILS—`reconcILS` and DLCpar (Wu et al. 2014)—on 1000 gene trees simulated as described above. Overall, the accuracy of both methods was high and similar for inferences of ILS (Figure 4a) and duplication (Figure 4b). The correlation

(as measured by Spearman's ρ) between simulated and inferred ILS events was 0.719 for `reconcILS` and 0.580 for DLCpar (Figure 4a). It is important to note that these two methods are inferring slightly different kinds of events, which we account for in these results. `reconcILS` infers the number of NNI events needed to reconcile a gene tree; for example, the gene tree in Figure 3a requires two NNI events. The results reported for `reconcILS` therefore compare simulated NNI events vs. inferred events. In contrast, DLCpar infers the number of ILS events needed, regardless of the number of NNI rearrangements that occur within such an event. The tree in Figure 3a only counts as one ILS event using this method. Results for reported for DLCpar therefore compare simulated vs. inferred ILS events, to ensure that a fair comparison can be made. In both cases the methods are doing well in their respective inferences, but with a slight boost in accuracy for `reconcILS`.

Inferences about duplication events were also highly accurate and highly similar between `reconcILS` and DLCpar (correlations of 0.894 and 0.892, respectively). The slight undercounting in both methods is due to cases in which duplication events are immediately followed by losses (the same can happen with ILS events followed by loss). Such events cannot be inferred by any parsimony-based algorithm, and are likely not inferrable by probabilistic methods either (e.g. Rasmussen and Kellis 2012).

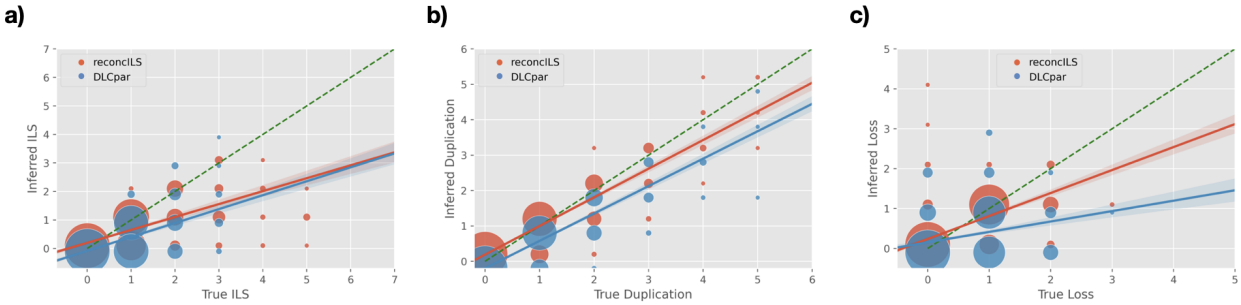

Fig. 4. Accuracy of `reconcILS` and DLCpar. a) Accuracy of inferred ILS events. Orange circles represent results from `reconcILS`. Blue circles represent results from DLCpar. Circle size is proportional to the number of reconciliations with each value, with jitter applied for clarity. Lines represent best-fit regressions (as calculated in Seaborn; Waskom 2021), plus confidence intervals. b) Accuracy of inferred duplication events. c) Accuracy of inferred loss events.

There is quite a difference between `reconcILS` and DLCpar in the inference of losses (correlations of 0.729 and 0.334, respectively; Figure 4c). Upon closer inspection, the lower accuracy in DLCpar is due to many loss events of one particular type not being counted. Specifically, DLCpar does not count losses that occur on either of the branches attaching to the root node of the species tree, as these will have occurred above the root of the resulting gene tree. This means that, for instance, DLCpar does not count the loss that occurred along the A branch in gene trees with the topology (B,C). (Gene trees with only a single tip, i.e. the topology (A), do not run at all in DLCpar, but were not counted here.) We confirmed this behavior on larger trees in DLCpar, as well as in all trees in ETE3. The difference in counting of losses among algorithms arises because `reconcILS` explicitly traverses the species tree, while the other methods do not. This enables our method to track losses that occur above the root of the input gene tree, but below the root of the input species tree. However, there may be some cases in which the behavior of DLCpar is preferred. For instance, since gene families have to originate on some branch of the species tree (e.g. Richter et al. 2018), forcing reconciliation algorithms to infer losses on branches that pre-date this appearance would be incorrect. It is not clear which approach to counting losses one should prefer when analyzing real data.

Analysis of primate genomes

To further demonstrate the power and accuracy of `reconcILS`, we applied it to a dataset of gene trees inferred from 23 primate genomes (Vanderpool et al. 2020; Smith et al. 2022). These genomes all come from the suborder Haplorhini, which includes all primates except the lemurs and lorises (i.e. suborder Strepsirhini).

Among these genomes, we analyzed two sets of gene trees that were previously inferred: First, we considered 1,820 single-copy orthologs ("Single-copy clusters" dataset in Smith et al. 2022). This dataset is not expected to contain any duplicates, though it does contain losses. However, because there are high levels of ILS among the primates sampled here, the dataset serves as an ideal comparator for reconciliation approaches with and without ILS. Second, 11,555 trees containing both orthologs and paralogs were used to infer duplications, losses, and ILS among the primates ("All paralogs" dataset in Smith et al. 2022). When using `reconcILS`, gene trees with polytomies were randomly resolved, and costs were set to $C_D = 2$, $C_L = 2$, and $C_I = 1$. We used both `reconcILS` and ETE3 to analyze the two datasets, assigning inferred events to each branch of the primate tree.

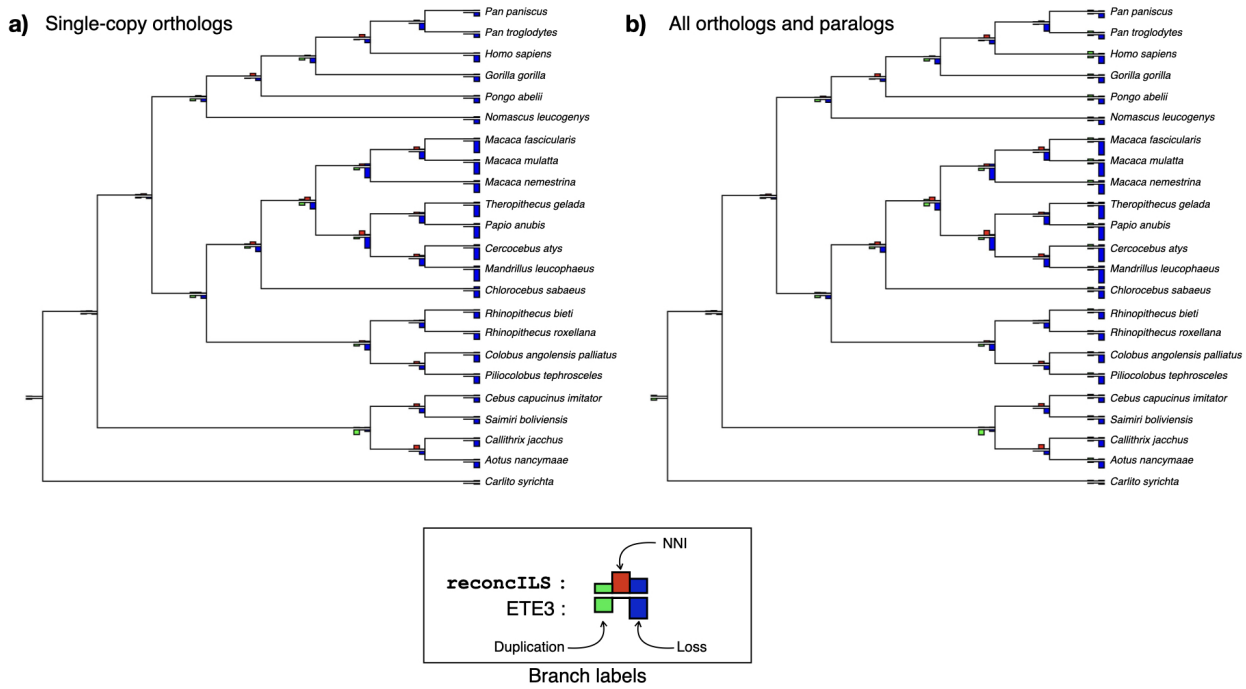

Fig. 5. Results from primates. a) Analysis of 1,820 single-copy orthologs. Both `reconcILS` and ETE3 were used to analyze the data, with duplication, NNI, and loss events mapped to each branch of the species tree. On each branch, the bar plots show results from `reconcILS` on top, with results from ETE3 mirrored below. ETE3 cannot infer NNI events, so this bar is always set to 0. b) Analysis of 11,555 gene trees containing both orthologs and paralogs. Results are presented in the same manner as in panel a. However, note that the height of all bars is normalized within a panel, so that heights of bars within a tree can be compared among branches, but the heights of bars between the trees cannot be compared. Figure S3 shows the absolute counts of all events from both panels, and Figure S4 shows species tree with branch lengths in substitution units rather than as a cladogram, as shown here.

Single-copy ortholog dataset. Results obtained from analysis of only the single-copy orthologs are shown graphically in Figure 5a (see Figure S3a for numerical values on each branch). The pattern that stands out from this graph is that ETE3 is vastly overestimating the number of duplications and, especially, the number of losses. Because ETE3 does not model ILS, all such events must be reconciled by one duplication (placed above the discordant branch) and three losses (placed below the discordant branch; Hahn 2007). We can see this effect in the data by looking at the correlation between the gene concordance factor (gCF) estimated for a branch with the number of duplicates inferred on its parent branch. Gene concordance factors quantify the fraction of loci that contain the same branch as the species tree; when single-copy genes are analyzed, low gCFs can indicate discordance due to ILS (Baum 2007; Lanfear and Hahn 2023). The correlation between gCFs for these data (estimated in Smith et al. 2022) and duplications inferred by ETE3 on parent branches

is $\rho = -0.69$. Lower gCFs lead to higher numbers of duplicates, as expected when ILS is not accounted for—this problem is not unique to ETE3, but should be expected for all standard DL and DTL approaches. Such errors can lead to incorrect inferences of bursts of duplication and loss.

Using the same reasoning, we can ask about the accuracy of `reconcILS` by examining the correlation between NNI events inferred by our algorithm and gCF values for single-copy genes. High NNI counts indicate more ILS, and therefore this measure should be negatively correlated with gCF values for the same branch. Indeed, the correlation between NNI counts and gCF values is $\rho = -0.86$, demonstrating the high accuracy of `reconcILS`. However, we can also see that `reconcILS` does sometimes infer a small number of duplicates (Figure 5a; Figure S3a). While it is possible that some of the single-copy genes we analyzed represent so-called pseudoorthologs—and therefore do contain actual duplication events—these are expected to be extremely rare (Smith and Hahn 2022). Instead, these false duplicates are likely due to ILS among four or more lineages. Recall that, although our algorithm has been developed to account for rearrangements due to the coalescent process, ILS among four lineages can result in gene trees that are three NNI moves away from the species tree; ILS among more lineages will produce trees that are even further away. Therefore, multiple NNI events are sometimes simply more costly than duplication and loss, even when ILS is the true cause of discordance.

Full gene tree dataset. We also examined patterns of ILS, duplication, and loss for the entire primate dataset. The results from the full set of gene trees (Figure 5b; Figure S3b) largely mirror the results from the single-copy orthologs (which are a subset of the entire dataset). Part of the apparent similarity is simply due to normalizing the two datasets separately, but several non-graphical factors also contribute to the pattern. First, although there are many true duplicates in the full dataset, the vast majority of gene trees only have events on tip branches: 7,693 out of 11,555 gene trees have duplicates specific to one species (Smith et al. 2022). Another ∼1,000 trees have duplicates specific to two species. Combined with the 1,820 single-copy genes, it becomes clear that there are in fact relatively few trees with duplicates on internal branches even in the full dataset.

Second, errors in inferring duplications and losses due to the coalescent process still dominate the results from standard DL reconciliation (i.e. ETE3), possibly even more so than in the single-copy dataset. Because ILS (and introgression) can affect all gene trees, trees with duplicates can experience twice as much discordance when there are twice as many branches. One example of this can be seen among the Papionini (macaques, baboons, geladas, and mandrills): ETE3 has to infer >25,000 losses on each of the tip branches in this clade, and more than 10,000 losses on most of the internal branches (Figure 5b; Figure S3b). There are also many thousands of duplicates inferred on internal branches of this subclade, though these species have the same number of total protein-coding genes as all the others analyzed here (Vanderpool et al. 2020). Although the inferred duplications and losses in this group were also relatively higher in the single-copy dataset, it is clear that even a small number of true duplicates can have an outsized effect on methods that cannot account for ILS.

The analysis using `reconcILS` is also highly similar between the two datasets, again largely because most duplicates are confined to tip branches. `reconcILS` is able to deal with discordance due to the coalescent, even in the presence of duplication: the correlation between NNI events inferred on each branch of the species tree from the single-copy orthologs and the full dataset is $\rho = 0.99$. Although the numbers of events are of course larger with the larger dataset (Figure S3b), the effect of ILS on any given branch of the species tree is experienced by all gene trees. Given its ability to deal with ILS, duplication events across the primate tree will be more reliably inferred with `reconcILS`.

Discussion

We have introduced a new reconciliation algorithm and associated software package (`reconcILS`) that considers discordance due to duplication, loss, and coalescence. To better demonstrate the accuracy of our method, we have also developed a new simulation program (`dupcoal`) that generates gene trees while allowing these processes to interact. Used together, our results demonstrate that `reconcILS` is a fast and accurate approach to gene tree reconciliation, one that offers multiple advantages over alternative approaches, even those that also consider ILS (e.g. DLCpar; Wu et al. 2014). Nevertheless, there are still several caveats to consider, as well as future directions to explore.

Despite our demonstration of good performance in practice, the biggest caveat to our approach is the lack of a theoretical proof that our method finds most parsimonious reconciliations. While such proofs often require strict assumptions—which can be relaxed in a simulation setting—they are very helpful in providing confidence across untested parameter space. However, our approach also differs fundamentally from DL and DTL algorithms with such proofs, as there is not a one-to-one labeling between gene trees and species trees when ILS is considered. As mentioned above, there is often ambiguity in how to label nodes in gene trees affected by ILS and duplication (e.g. Figures 3b and 3c). This is due in part to how `reconcILS` uses NNI rearrangements—it is not clear which node to pick if a branch is rearranged—and in part to the intrinsic lack of information contained only within tree topologies when coalescence and duplication interact. The difficulty in labeling is not confined to this work: DLCpar does not label nodes, and RecPhyloXML, a format for representing reconciled trees, does not include a way to label ILS events at all (Duchemin et al. 2018).

Nevertheless, labeling gene trees serves multiple important functions, and is often interpreted as biological truth. For example, multiple methods use inferred gene tree labels to identify orthologous genes (genes related by speciation events; Fitch 1970). The general approach taken is to label nodes using reconciliation or similar approaches, treating any node not labeled with a D as a speciation node; genes whose common ancestor is a speciation node are subsequently defined as orthologs. Both standard DL reconciliation and so-called "overlap" methods (e.g. Huerta-Cepas et al. 2007; Emms and Kelly 2019) will consistently label the top-most node flanking a discordant branch with a duplication event. This means that, in cases like the one shown in Figure 3b, these methods will label the wrong node with a D, and subsequently identify the wrong pair of genes as orthologs. (Figure 3c would be labeled correctly; recall that this outcome has an equal probability to the one in Figure 3b in the scenario considered here.) Cases such as these have been referred to as "unsorted homologs" (Mallo et al. 2014), and will occur any time duplication and coalescence interact, either before or after a speciation event. They can also occur even if only two species are analyzed (e.g. if only species B and C were considered in Figures 3b and 3c), which means that approaches that simply collapse short branches to avoid ILS will also be affected. Although `reconcILS` currently only labels one node with a D, in the future we could allow for multiple possible labelings in order to emphasize the inherent ambiguity of some labels.

The algorithm used by `reconcILS` deals at least in part with an additional interaction between duplication, loss, and coalescence: hemiplasy. Hemiplasy is the phenomenon by which a single mutational event on a discordant gene tree will be interpreted as multiple events ("homoplasy") when reconciled to the species tree (Avise and Robinson 2008; Hahn and Nakhleh 2016). In analyses of nucleotide substitutions, generic binary traits, and even continuous traits, hemiplasy requires that mutations have occurred on a discordant branch of a gene tree (Mendes and Hahn 2016; Guerrero and Hahn 2018; Hibbins et al. 2023). In the context of either duplication or loss, such mutations would have to occur on a discordant branch of the daughter gene tree to lead to "copy-number hemiplasy." Unfortunately, such events are still not accurately inferred by `reconcILS`: we have found that the algorithm consistently proposes an extra loss when hemiplasy occurs (`dupcoal` can track and report simulated hemiplasy events, allowing us to easily make such comparisons). However, there is also a broader definition of copy-number hemiplasy in use: a duplication or loss that does not fix in all descendant lineages (Rasmussen and Kellis 2012; Wu et al. 2014). While this definition encompasses the stricter one given above, it also includes cases such as the one shown in Figure 1. Because the mutation in that figure occurs in the common ancestral population of species B and C, but was only inherited by lineage C, this scenario would be considered to be hemiplasy by the broader definition (although it technically does not involve ILS if the outcome resembles the gene trees in Figure 1c-e, since coalescence is happening in the same population). The number of each type of event in such cases is accurately inferred using `reconcILS`; DLCpar assumes that this type of hemiplasy does not occur. Note, however, that if the duplication event (the star in Figure 1a) occurred a few generations later—after the split between species B and C, along a tip lineage—the scenario would not be considered hemiplasy by this broader definition, but would still produce the same range of gene trees (i.e. Figure 1c-g). (It would now involve ILS, however.) Regardless of the exact definitions and terms used, it should be clear that duplication, loss, and coalescence can interact to produce a wide array of topologies not considered by most reconciliation approaches.

One relatively straightforward improvement that could be made to `reconcILS` is in the cost of different events. As discussed earlier, it is difficult to determine biologically realistic costs for duplication, loss, and ILS. One approach to determining duplication and loss costs would be to infer the number of such events using an alternative, non-reconciliation approach (e.g. Mendes et al. 2020), setting the cost in `reconcILS`

to be inversely related to the inferred number of events of each type. Varying rates of duplication and loss can also be inferred along different branches of the species tree, and so the costs of these events could vary commensurately along the tree. Similarly, we can envision a simple way to scale the costs of NNI with the length of species tree branches: this is because the probability of discordance due to ILS is inversely proportional to internal branch lengths on a species tree (Hudson 1983). We could therefore make NNI events on longer branches of the species tree more costly, which would properly weight the relative probability of such events locally on the tree.

There are also several extensions that could be made to `dupcoal`, our new simulation engine. Currently, `dupcoal` draws a daughter gene tree from the multispecies coalescent (MSC) at random: this implicitly assumes that the parent and daughter trees are completely unlinked. It should be straightforward, however, to draw the daughter tree at a given recombination distance from the parent tree, ranging from fully linked to fully unlinked. In addition, though the software currently draws gene trees from the MSC, we could also include introgression as a source of discordance by drawing trees from the multispecies network coalescent (Meng and Kubatko 2009; Yu et al. 2011). In this case, users would simply input a species network to simulate from, rather than a species tree. We should also note the differences between the current version of `dupcoal` and alternative simulators. As mentioned previously, SimPhy (Mallo et al. 2016) does not simulate hemiplasy of any kind, as it uses the same three-tree model of Rasmussen and Kellis (2012) and Wu et al. (2014). However, this method does include gene transfers, which are not simulated by `dupcoal`. The MLMSC simulator of Li et al. (2021) simulates both hemiplasy and transfers. `dupcoal` is therefore most similar to this method, with two small differences. First, `dupcoal` does not change the rate of duplication and loss above speciation events to account for ancestral polymorphism (cf. Gillespie and Langley 1979); instead, users can simply extend all nodes $2N$ generations into the past to do so (where N is the effective population size). Second, `dupcoal` does not account for the possibility of further events occurring while duplications and losses are still polymorphic in a population. In order to streamline the simulation process, we have chosen to ignore this squared term.

Finally, we envision multiple ways in which the `reconcILS` algorithm could be used in conjunction with other methods. For instance, ASTRAL-Pro (Zhang et al. 2020) infers a species tree from gene trees containing paralogs by first labeling nodes in each gene tree as either duplication or speciation. The authors of this software rightfully recognize the problems with labeling discussed here, but nevertheless their labeling step might benefit from methods that include ILS. Likewise, GRAMPA (Thomas et al. 2017) allows for reconciliation of polyploidy events, but does not include coalescence as a source of discordance; their algorithm could easily accommodate this extension. There are also multiple approaches that can be used to "correct" gene trees via reconciliation (e.g. El-Mabrouk and Ouangraoua 2017; Christensen et al. 2020). These and many other tasks might benefit from approaches that model duplication, loss, and coalescence in a single framework.

Acknowledgements. We thank Yao-ban Chan for many helpful conversations about the MLMSC model, Jessica Wu for answering questions about DLCpar, and Ben Fulton for tips on best programming practices. Reviewers of a previous version of the manuscript also provided constructive comments. This work was supported by NSF grant DEB-1936187 to MWH.

References

Avise JC and Robinson TJ (2008) Hemiplasy: A new term in the lexicon of phylogenetics. Systematic Biology 57:503–507.

Bansal MS, Alm EJ and Kellis M (2012) Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. Bioinformatics 28:i283-i291

Baum DA (2007) Concordance trees, concordance factors, and the exploration of reticulate genealogy. Taxon 56:417–426.

Boussau B and Scornavacca C (2020) Reconciling gene trees with species trees. In: Scornavacca C, Delsuc F and Galtier N (eds) Phylogenetics in the Genomic Era. No commercial publisher — Authors open access book, p 3.2:1-3.2:23

Chan Y, Ranwez V and Scornavacca C (2017) Inferring incomplete lineage sorting, duplications, transfers and losses with reconciliations. Journal of Theoretical Biology 432:1–13.

Chaudhary R, Burleigh JG and Eulenstein O (2011) Algorithms for rapid error correction for the gene duplication problem. In: Chen J, Wang J and Zelikovsky A (eds) Bioinformatics Research and Applications. Springer, Berlin, Heidelberg, pp 227–239

Chen K, Durand D and Farach-Colton M (2000) NOTUNG: A program for dating gene duplications and optimizing gene family trees. Journal of Computational Biology 7:429–447.

Christensen S, Molloy EK, Vachaspati P, Yammanuru A and Warnow T (2020) Non-parametric correction of estimated gene trees using TRACTION. Algorithms for Molecular Biology 15:1.

Du H, Ong YS, Knittel M, Mawhorter R, Liu N, Gross G, Tojo R, Libeskind-Hadas R and Wu Y-C (2021) Multiple optimal reconciliations under the duplication-loss-coalescence model. IEEE/ACM Transactions on Computational Biology and Bioinformatics 18:2144–2156.

Duchemin W, Gence G, Arigon Chifolleau A-M, Arvestad L, Bansal MS, Berry V, Boussau B, Chevenet F, Comte N, Davín AA et al. (2018) RecPhyloXML: a format for reconciled gene trees. Bioinformatics 34:3646–3652.

Durand D, Halldórsson BV and Vernot B (2006) A hybrid micro–macroevolutionary approach to gene tree reconstruction. Journal of Computational Biology 13:320–335.

El-Mabrouk N and Ouangraoua A (2017) A general framework for gene tree correction based on duplication-loss reconciliation. WABI 2017: 8:1-8:14.

Emms DM and Kelly S (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biology 20:238.

Fitch WM (1970) Distinguishing homologous from analogous proteins. Systematic Zoology 19:99–113.

Gillespie JH and Langley CH (1979) Are evolutionary rates really variable? Journal of Molecular Evolution 13:27–34.

Goodman M, Czelusniak J, Moore GW, Romero-Herrera AE and Matsuda G (1979) Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Systematic Zoology 28:132–163.

Guerrero RF and Hahn MW (2018) Quantifying the risk of hemiplasy in phylogenetic inference. Proceedings of the National Academy of Sciences 115:12787–12792.

Guigo R, Muchnik I and Smith TF (1996) Reconstruction of ancient molecular phylogeny. Molecular Phylogenetics and Evolution 6:189–213.

Hahn MW (2007) Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. Genome Biology 8:R141.

Hahn MW and Nakhleh L (2016) Irrational exuberance for resolved species trees. Evolution 70:7–17.

Hibbins MS, Breithaupt LC and Hahn MW (2023) Phylogenomic comparative methods: Accurate evolutionary inferences in the presence of gene tree discordance. Proceedings of the National Academy of Sciences 120:e2220389120.

Huerta-Cepas J, Dopazo H, Dopazo J and Gabaldón T (2007) The human phylome. Genome Biology 8:R109.

Huerta-Cepas J, Serra F and Bork P (2016) ETE3: Reconstruction, analysis, and visualization of phylogenomic data. Molecular Biology and Evolution 33:1635–1638.

Jacox E, Chauve C, Szöllősi GJ, Ponty Y and Scornavacca C (2016) ecceTERA: comprehensive gene tree-species tree reconciliation using parsimony. Bioinformatics 32:2056–2058.

Kingman JFC (1982) The coalescent. Stochastic Processes and their Applications 13:235–248.

Lanfear R and Hahn M (2024) The meaning and measure of concordance factors in phylogenomics. EcoEvoRxiv doi: https://doi.org/10.32942/X27617

Li Q, Scornavacca C, Galtier N and Chan Y-B (2021) The multilocus multispecies coalescent: A flexible new model of gene family evolution. Systematic Biology 70:822–837

Maddison WP (1997) Gene trees in species trees. Systematic Biology 46:523–536.

Mawhorter R and Libeskind-Hadas R (2019) Hierarchical clustering of maximum parsimony reconciliations. BMC Bioinformatics 20:612.

Mendes FK and Hahn MW (2016) Gene tree discordance causes apparent substitution rate variation. Systematic Biology 65:711–721.

Mendes FK, Vanderpool D, Fulton B and Hahn MW (2021) CAFE 5 models variation in evolutionary rates among gene families. Bioinformatics 36:5516–5518.

Meng C and Kubatko LS (2009) Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: A model. Theoretical Population Biology 75:35–45.

Nguyen TH, Ranwez V, Pointet S, Chifolleau A-MA, Doyon J-P and Berry V (2013) Reconciliation and local gene tree rearrangement can be of mutual profit. Algorithms in Molecular Biology 8:12.

Page RDM (1994) Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. Systematic Biology 43:58–77.

Page RDM and Charleston MA (1997) From gene to organismal phylogeny: Reconciled trees and the gene tree/species tree problem. Molecular Phylogenetics and Evolution 7:231–240.

Rasmussen MD and Kellis M (2012) Unified modeling of gene duplication, loss, and coalescence using a locus tree. Genome Research 22:755–765.

Smith ML and Hahn MW (2022) The frequency and topology of pseudoorthologs. Systematic Biology 71:649–659.

Smith ML, Vanderpool D and Hahn MW (2022) Using all gene families vastly expands data available for phylogenomic inference. Molecular Biology and Evolution 39:msac112.

Stull GW, Qu X-J, Parins-Fukuchi C, Yang Y-Y, Yang J-B, Yang Z-Y, Hu Y, Ma H, Soltis PS, Soltis DE et al. (2021) Gene duplications and phylogenomic conflict underlie major pulses of phenotypic evolution in gymnosperms. Nature Plants 7:1015–1025.

Sukumaran J and Holder MT (2010) DendroPy: a Python library for phylogenetic computing. Bioinformatics 26:1569–1571.

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437–460.

Vanderpool D, Minh BQ, Lanfear R, Hughes D, Murali S, Harris RA, Raveendran M, Muzny DM, Hibbins MS, Williamson RJ et al. (2020) Primate phylogenomics uncovers multiple rapid radiations and ancient interspecific introgression. PLOS Biology 18:e3000954.

Waskom M (2021) seaborn: statistical data visualization. The Journal of Open Source Software 6:3021.

Wu Y-C, Rasmussen MD, Bansal MS and Kellis M (2014) Most parsimonious reconciliation in the presence of gene duplication, loss, and deep coalescence using labeled coalescent trees. Genome Research 24:475–486.

Yu Y, Than C, Degnan JH and Nakhleh L (2011) Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. Systematic Biology 60:138–149.