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

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

[1]Department of Computer Science, Indiana University, Bloomington, Indiana 47405

[2]Department of Biological Sciences, Mississippi State University, Starkville MS 39762

[3]Department of Biology, Indiana University, Bloomington, Indiana 47405

*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 heuristic 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, loss, and ILS. Despite being a heuristic method, we show that `reconcILS` is much more accurate than models that ignore ILS, and at least as accurate or better than leading methods that can model ILS, while also being able to handle much larger datasets. 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 ILS.

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. Hallett and Lagergren 2001; Hallett et al. 2004; Bansal et al. 2012; Stolzer 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 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 due only 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 accommodate 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 all effects of coalescence on incorrect inferences of duplication and loss (see below). 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 based on the probabilistic three-tree model of Rasmussen and Kellis (2012), but runs faster and without the need to specify duplication and loss rates ahead of time.

Here, we introduce a new heuristic method 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 on smaller trees (Wu et al. 2014; Du et al. 2021), it is not straightforward to include its algorithm within other tools that also have reconciliation steps (e.g. Thomas et al. 2017; Zhang et al. 2020) or to run it on larger datasets. 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 genomes from primates, a dataset that has previously been shown to contain a large amount of discordance due to ILS (Vanderpool et al. 2020).

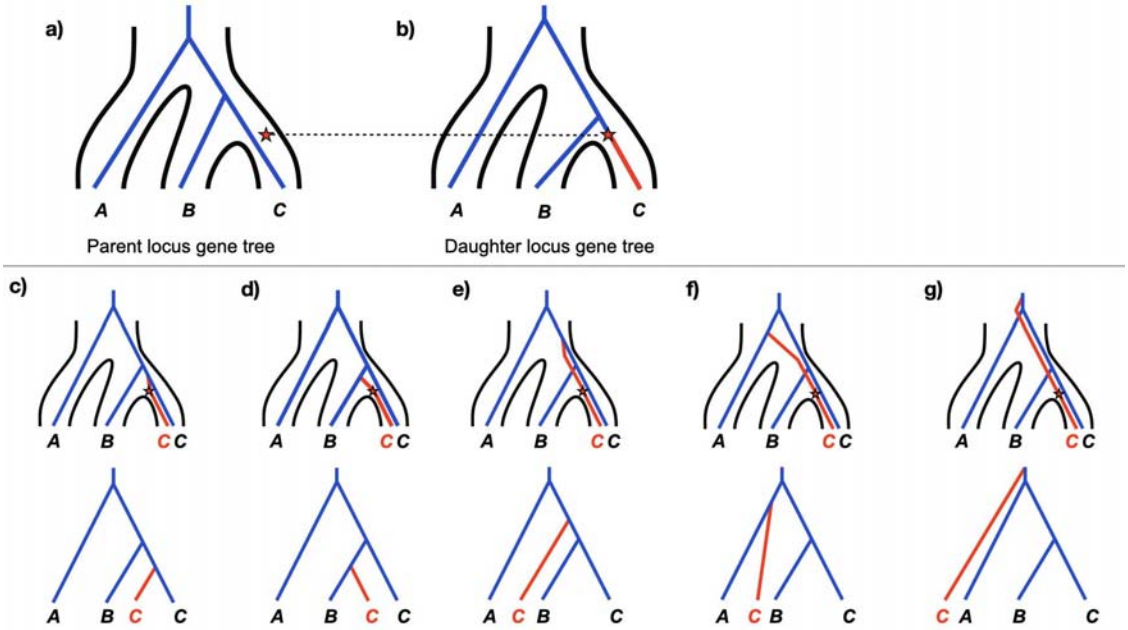


Figure 1. The multilocus multispecies coalescent model generates full gene trees. *a) A gene tree at the parent locus (blue) is drawn from the species tree (outlined in black) via the multispecies coalescent (MSC) model. A duplication event (red star) is drawn from a birth-death model over the species 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.*

New Approaches

Simulation using `dupcoal`

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

102 inferences. The explanation also makes clear which common processes are not modeled by
103 DL and DTL methods. Our simulations are based on the multilocus multispecies coalescent
104 (MLMSC) model of Li et al. (2021). This model treats each individual locus as evolving via
105 the multispecies coalescent (MSC) model, in which discordance between individual gene
106 trees and the species tree can arise only via incomplete lineage sorting. The MLMSC further
107 models relationships among these loci via duplication and coalescence, in addition to losses
108 at any locus.

109 To illustrate how duplications and losses occur, and how relationships among loci
110 are generated in `dupcoal`, we use a simplified example (a full description of the
111 simulation algorithm can be found at https://github.com/meganlsmith/dupcoal). As a first
112 step, a gene tree is generated via the MSC at the "parent" locus (Figure 1a); this tree exists
113 regardless of whether subsequent duplications occur. Duplications and losses are then
114 generated via a birth-death process on the species tree, with each event occurring at a
115 specific time on a specific branch of the species tree (one such duplication is represented as
116 a star in Figure 1a). For each duplication event, an additional gene tree is generated via the
117 MSC at a "daughter" locus (Figure 1b). This tree represents the history of the locus at which
118 the duplicated DNA will be inserted, and exists with or without this duplication event. In
119 other words, each locus has its own gene tree topology, and any sequence duplicated into a
120 locus will follow its topology. For now, we assume that this locus is unlinked from the
121 parent locus, so that the topologies are completely independent. The mutation representing
122 the duplication event is then placed on the gene tree at the daughter locus at the time at
123 which it occurred on the species tree—if multiple branches of the daughter gene tree exist
124 at this point in time on the relevant branch, the mutation is assigned to one at random (star
125 in Figure 1b). Regardless of where exactly the mutation occurs on the daughter gene tree,
126 the duplication now takes on the topology and branch lengths of the tree below this point
127 (red branch in Figure 1b). Losses are also simulated on the species tree and placed on each
128 gene tree in a similar manner (though no losses are shown in Figure 1, for simplicity).
129 When loss events occur, they are placed on branches of gene trees (either parent or
130 daughter trees), with all branches below such mutations removed. We use both a forward
131 algorithm (the birth-death process) and a backward algorithm (the coalescent) in order to

accurately model the full gene tree containing all duplications, losses, and coalescent events.

The process described thus far in our example has produced two gene trees—one at the parent locus (Figure 1a) and one at the daughter locus (Figure 1b). Either or both of these two trees may be concordant (as shown) or 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 additional key coalescent process with an example in Figure 1.

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 coalesces with the parent tree on the same branch on which the mutation occurred (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 in which the daughter tree joins the parent tree on a different branch, each of which would require one nearest neighbor interchange (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 itself 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 ancestral population of species B and C). This same dynamic in the ancestral population also arises when the species tree only contains two species—this is why approaches that simply collapse internal branches of a species tree will not eliminate incorrect reconciliations due to coalescence. In addition, note that, in the scenario shown here, the trees in Figures 1d and 1e are equiprobable with the topology in Figure 1c,

161 making even probabilistic reconciliation difficult. These simple examples demonstrate that
162 the vagaries of the coalescent process lead to a thicket of different tree topologies.

163 Parent and daughter trees can also be discordant if any of the locus gene trees do
164 not match the species tree. In such cases, a further complication can arise: "hemiplasy" or
165 "copy-number hemiplasy" (Avise and Robinson 2008; Rasmussen and Kellis 2012; Hahn
166 and Nakhleh 2016). Copy-number hemiplasy occurs when either a duplication or a loss
167 mutation occurs on a discordant branch of a gene tree (i.e. a branch that does not exist in
168 the species tree). While loss events can cause copy-number hemiplasy when they occur on
169 a discordant parent tree (even in the absence of duplication), duplication events can only
170 cause copy-number hemiplasy on discordant daughter trees. Such events are not simulated
171 by SimPhy (Mallo et al. 2016), a popular platform for carrying out gene tree simulations
172 with duplications and losses, but are simulated by `dupcoal`. Copy-number hemiplasy can
173 cause further complications for reconciliation, and so are an important phenomenon to
174 include (see Results).

175 Together, the processes described here make up the steps by which our simulator
176 can generate full gene trees: i.e. ones with the parent and all daughter trees joined together.

177 Reconciliation using `reconcILS`

178 *Overview of main algorithm*

179 Our reconciliation approach uses nearest-neighbor interchange (NNI)
180 rearrangements to model ILS and coalescence events. NNI has been used previously in
181 reconciliation algorithms to correct gene trees that may have been inferred incorrectly (e.g.
182 Chen et al. 2000; Chaudhary et al. 2011; Nguyen et al. 2013). Here, we use NNI because the
183 effects of ILS on gene trees can be modeled as a series of NNI rearrangements (Than and
184 Rosenberg 2013). For three species undergoing ILS, there are two possible discordant
185 topologies, each of which is one NNI move away from the species tree topology (e.g. Figure
186 2). For four lineages undergoing ILS, there are 14 possible discordant topologies, each of
187 which are between one and three NNI moves away from the species tree topology. (Note
188 that it is the number of lineages undergoing ILS at a specific point on the tree that matters,

not the total number of tips in a tree.) Our algorithm takes advantage of this relationship between NNI and ILS in order to reconcile gene trees, while also considering duplication and loss.

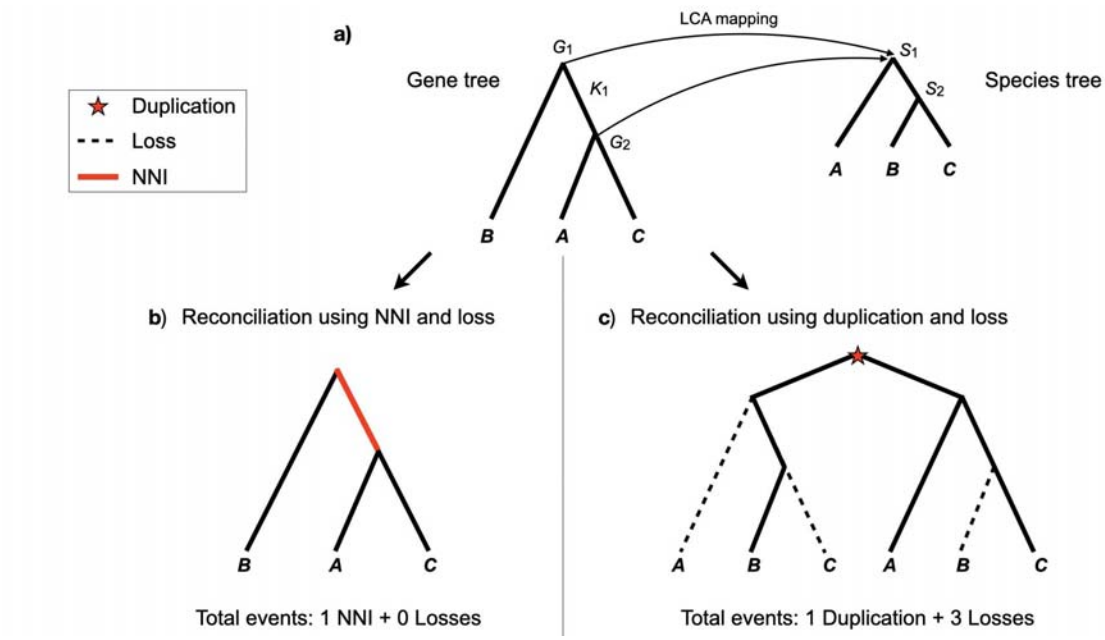


Figure 2. Summary of `reconcILS`. *a) The initial step generates an LCA mapping function between the gene tree and the species tree. Here, nodes G_1 and G_2 in the gene tree both map to node S_1 in the species tree; we therefore refer to S_1 as a multiply mapped node. Additionally, G_1 and G_2 together 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 reconciliations in b) and c) for every discordant branch.*

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

207 ancestor (LCA) mapping (Zmasek and Eddy 2001) to create a mapping function between
208 each node in G and a node in S (Figure 2a). We refer to each of these pairwise relationships
209 (arrows in Figure 2a) as "mappings" from here on.

210 Unlike in standard DL methods, LCA mapping here does not represent a final
211 reconciliation. Instead, this step identifies cases in which multiple nodes in the gene tree
212 map to an identical node in the species tree. For instance, in Figure 2a nodes G_1 and G_2 in
213 the gene tree both map to node S_1 in the species tree. We therefore call S_1 a "multiply
214 mapped" node. Additionally, we refer to the branch in the gene tree bracketed by nodes G_1
215 and G_2 as a "discordant" branch (labeled K_1 in Figure 2a); such branches do not exist in the
216 species tree. The focus of our method is in determining whether discordant branches
217 should be resolved via NNI and loss (Figure 2b) or via duplication and loss (Figure 2c).
218 Choosing among events (NNI, duplication, and loss) is carried out greedily for each branch,
219 with no consideration of a globally optimal set of events among all discordant branches in
220 the gene tree. Consequently, after choosing a set of events for one branch (given specified
221 costs), the process repeats itself until there are no multiply mapped nodes and no
222 discordant branches. Duplications implied by multiply mapped tip nodes in the species tree
223 are inferred via standard methods; inferences of losses are explained in the Methods.

224 As with all parsimony-based reconciliations, `reconcILS` requires a cost for each
225 event type: one for duplications (C_D), one for losses (C_L), and one for NNIs (C_I).
226 Reconciliations can be very sensitive to costs, such that maximum parsimony solutions
227 under one set of costs can be very different than under another set of costs. Although there
228 are few ways to determine costs independently (but see Discussion), there are some
229 approaches for finding the Pareto-optimal costs within a reconciliation context (e.g.
230 Libeskind-Hadas et al. 2014; Mawhorter et al. 2019). Here, as in most commonly used
231 reconcilation algorithms, we allow the user to set costs. The default costs within
232 `reconcILS` are set to $C_D = 1.1$, $C_L = 1$, and $C_I = 1$.

233 ***Labeling a reconciled gene tree***

234 A full reconciliation of a gene tree, G, and species tree, S, will have reconciled all
235 multiply mapped nodes, leaving no discordant branches. The total number of duplications,

losses, and NNIs required for this reconciliation is the sum of each event type across all branches; these numbers are 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` traverses the species tree finding optimal local reconciliations for each branch, it also records 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 (for duplication) or (for speciation); they may also label the location of losses, . Here, we additionally aim to label the NNI events induced by ILS (*I*).

Importantly, there is no standard way to label ILS events on a tree, largely because there is no exact map between gene tree nodes and species tree nodes under ILS. A branch undergoes NNI (and ILS), not a node, so it is not clear which node flanking such a branch should be chosen to be labelled. This labeling problem therefore differs fundamentally from labeling under DL and DTL algorithms because of the lack of a one-to-one map between nodes in the two trees. Here, we have created our own labeling system for ILS, one that we think is both understandable and easy to work with.

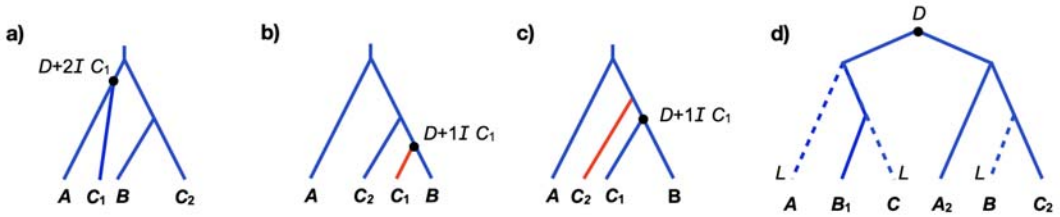


Figure 3. Labeling reconciled gene trees. *a) One possible reconciliation of the gene tree from Figure 1f labeled by* `reconcILS`*. The single ILS node ( ) has been labeled with a branch that has been rearranged by NNI ( ), the number of NNIs it has undergone ( ), and the fact that it also represents a duplication ( ). 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 Methods 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 and , although that is not the true history. d) A gene tree with only*

duplications and losses (L) labeled by `reconcILS`*. Internal branches on a tree can also be labeled with L, though none are shown here.*

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. Further details of this labeling are provided in the Methods, and caveats are covered 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 on which branch they act. `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` can 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.

288 # Results

289 ## Accuracy and advantages of `dupcoal`

290 We assessed the accuracy of `dupcoal` by first simulating 1000 gene trees along a
291 40-tip species tree with no duplication and no loss (Supplementary Figure S1). These
292 simulations therefore only model ILS as a cause of discordance. We compared the degree of
293 discordance in these simulations to ones carried out using the same species tree with
294 SimPhy (after appropriate branch-length scaling; see Methods). We found a very high
295 correlation between concordance factors on each branch calculated from gene trees
296 produced by the two methods (ρ=0.99, P=1.13 x 10^{-30}), indicating that we are accurately
297 simulating ILS. We also carried out simulations in `dupcoal` with both duplication and
298 loss, again comparing our results to SimPhy (after scaling duplication and loss rates
299 appropriately; see Methods). Again, the average number of duplications and losses per
300 simulated gene tree (n=1000) were highly similar and not significantly different between
301 the two methods: 1.81 vs. 1.77 duplications and 1.78 vs. 1.84 losses for `dupcoal` and
302 SimPhy, respectively (P=0.13 for duplications and P=0.29 for losses; t-tests), indicating that
303 we are accurately simulating duplications and losses in `dupcoal`.

304 To highlight additional features contained within `dupcoal`, we also tracked the
305 number of simulated gene trees in which copy-number hemiplasy occurred (this is a
306 standard output reported by `dupcoal`). Out of 1000 gene trees simulated on the 40-tip
307 tree with duplication, loss, and ILS, 246 contained at least one hemiplasy event. As a
308 straightforward way to demonstrate the effect of such events, we separately inferred the
309 number of duplications and losses on gene trees with and without copy-number hemiplasy
310 using ETE3 (Huerta-Cepas et al. 2016). While ETE3 implements a DL reconciliation
311 algorithm that does not model any type of ILS (see next section), this contrast shows the
312 effects of hemiplasy above and beyond the effects of ILS. As expected, both the number of
313 inferred duplications (1.7 excess duplications; P=6.3 x 10^{-6}, t-test) and the number of
314 inferred losses (5.4 excess losses; P=6.7 x 10^{-7}, t-test) were higher among the trees
315 containing hemiplasy. These results emphasize the importance of having a simulation
316 engine that can simulate hemiplasy.

317 ## Accuracy and advantages of `reconcILS`

318 As a simple test of DL reconciliation algorithms—which do not allow for ILS—we
319 used the DL method implemented in ETE3 (Huerta-Cepas et al. 2016). We initially ran
320 simulations in `dupcoal` without duplication or loss (i.e. ILS only); this allows us to see
321 incorrectly inferred duplications and losses more easily. As expected, on a species tree with
322 only 3 tips, ETE3 always inferred 1 extra duplication and 3 extra losses on gene trees with
323 discordance due to ILS (not shown). On a larger species tree with 40 tips, ETE3 inferred a
324 large excess of duplications and losses when only ILS is acting (9.5 duplications and 65.0
325 losses per gene tree, on average). On datasets simulated on a 40-tip tree with duplication,
326 loss, and ILS, the accuracy of ETE3 is low for losses (ρ=-0.04) and fairly good for
327 duplications (ρ=0.73) (Supplementary Figure S2).

328 In contrast, `reconcILS` accurately infers the numbers of duplications and losses
329 (Figure 4a and 4b), as well as ILS events (Figure 4c) on the 40-tip tree. The correlation
330 between inferred and simulated duplications and losses were high (ρ=0.95 and 0.65,
331 respectively). We also observed a high correlation between the inferred and simulated
332 number of NNI events needed to reconcile a gene tree (ρ=0.76). ILS immediately followed
333 by loss will not leave a signal in the resulting gene trees, leading to a slight undercounting
334 of NNI events caused by ILS.

335 As we did with ETE3 in the previous section, we also contrasted the number of
336 duplications and losses inferred by `reconcILS` on simulated gene trees with and without
337 copy-number hemiplasy. Unfortunately, `reconcILS` also does not deal perfectly with
338 hemiplasy, though in a slightly different way than ETE3. Rather than overestimating both
339 the number of duplications and losses, `reconcILS` appears to generally convert true ILS
340 events into losses. This occurs because (for instance) a duplication on a discordant
341 daughter gene tree branch will duplicate a non-sister pair of lineages (e.g. *A* and *C*), which
342 `reconcILS` interprets as a duplication and a loss (of *B*, in this case) rather than as a
343 duplication and an NNI event. Consistent with this, we observed very few excess
344 duplications in the simulations with hemiplasy (0.26 extra duplications, on average), but

significantly more losses (1.45; P=3.8 x 10^{-26}, t-test) and significantly fewer NNI events (1.66 P=3.5 x 10^{-7}, t-test) compared to simulations that did not have hemiplasy.

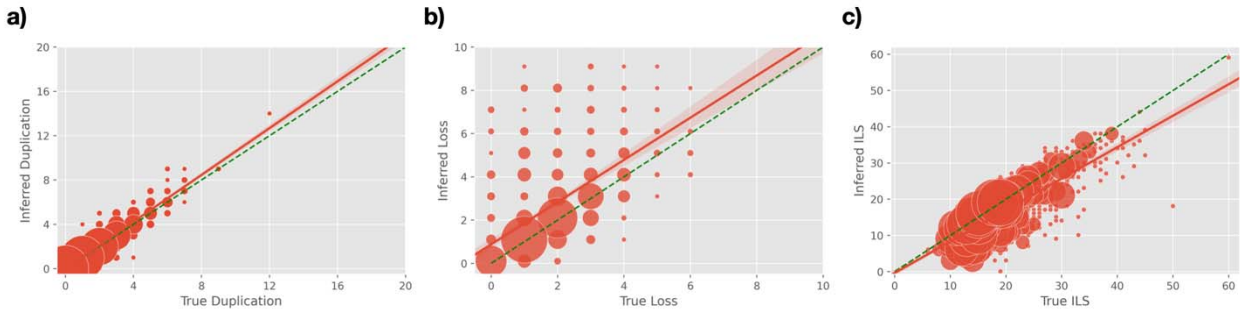


Figure 4. Accuracy of `reconcILS` on data simulated from the 40-tip tree. *a) Number of inferred versus true duplication events. b) Number of inferred versus true loss events. c) Number of inferred versus true ILS events. Circle size is proportional to the number of reconciliations with each value, with jitter applied for clarity. Orange line represent best-fit regression (as calculated in Seaborn; Waskom 2021), plus confidence intervals. The dashed line represents y=x. See Supplementary Figure S2 for results from ETE3 on the same data.*

We also compared the results from `reconcILS` with another reconciliation method that allows for ILS, DLCpar (Wu et al. 2014). DLCpar was written in Python 2, and—possibly because we had to run it in a virtual environment—could not run on the larger gene trees produced by the 40-tip tree. Therefore, we compared the performance of `reconcILS` and DLCpar on 1000 gene trees simulated on the 3-tip species tree. Overall, the accuracy of both methods was high for duplication (`reconcILS`, ρ=0.89; DLCpar, ρ=0.89; Supplementary Figure S3a). The correlation between simulated and inferred ILS events was ρ= for `reconcILS` and ρ= for DLCpar (Supplementary Figure S3b). It is important to note that DLCpar infers the number of ILS events, rather than the number of NNI rearrangements that occur within such an event (which are what is inferred by `reconcILS`). To ensure that a fair comparison can be made, results reported for DLCpar therefore are simulated vs. inferred ILS events. In both cases the methods are doing well in their respective inferences, but with slightly higher accuracy for `reconcILS`.

368 There is quite a difference between `reconcILS` and DLCpar in the inference of
369 losses (correlations of ρ=0.67 and ρ=0.27, respectively; Supplementary Figure S3c). Upon
370 closer inspection, the lower accuracy in DLCpar is due to many loss events of one particular
371 type not being counted. Specifically, DLCpar does not count losses that occur above the root
372 of the gene tree being considered. This means that, for instance, in gene trees with the
373 topology (B,C), DLCpar does not count the loss that occurred along the A branch (assuming
374 the species tree in Figure 1). We confirmed this behavior on slightly larger trees in DLCpar,
375 as well as in all trees in ETE3—no losses above the root of a gene tree are considered. The
376 difference in counting of losses among algorithms arises because `reconcILS` explicitly
377 traverses the species tree, while the other methods do not. This enables our method to
378 track losses that occur above the root of the input gene tree, but below the root of the input
379 species tree. However, there may be some cases in which the behavior of DLCpar and ETE3
380 is preferred. For instance, since gene families must originate on some branch of the species
381 tree (e.g. Richter et al. 2018), forcing reconciliation algorithms to infer losses on branches
382 that pre-date this appearance would be incorrect. It is not clear which approach to counting
383 losses one should prefer when analyzing real data.

384 Finally, we assessed the runtime and memory requirements of `reconcILS`
385 compared to ETE3 on a single node of the Indiana University Research Desktop, running an
386 Intel Xeon processor with 100 GB of RAM (DLCpar can only be run in Python 2, which
387 makes comparisons highly unfair). On a 3-species tree, both programs are extremely fast
388 (average time of 0.015 seconds per gene tree for both; Supplementary Figure S4a) and both
389 use very little memory (average peak memory per gene tree of 0.17 MB for `reconcILS`
390 and 0.075 MB for ETE3; Supplementary Figure S4b). Both programs remain very fast on
391 data produced from a 40-species tree, though the faster speed of ETE3 is also evident (5.02
392 seconds per gene tree for `reconcILS` and 1.60 seconds for ETE3; Supplementary Figure
393 S4a). Regardless, this means that the entire 1000-gene-tree simulated dataset on the 40-
394 species tree took less than 84 minutes to analyze using `reconcILS` on a moderately
395 powerful computer. Similarly, the memory requirements of `reconcILS` increase on the
396 40-tip tree (16.88 MB per gene tree for `reconcILS` and 2.93 MB for ETE3; Supplementary
397 Figure S4a), but are not onerous for modern computers. As the 40-species tree can contain

398 many more events (duplication, loss, or ILS), we also plotted the average time and memory
399 required for trees containing different numbers of total events (Supplementary Figure
400 S4c,d). Both programs show linear increases in time and memory with the number of
401 events, which is promising for analyzing larger datasets with larger species trees.

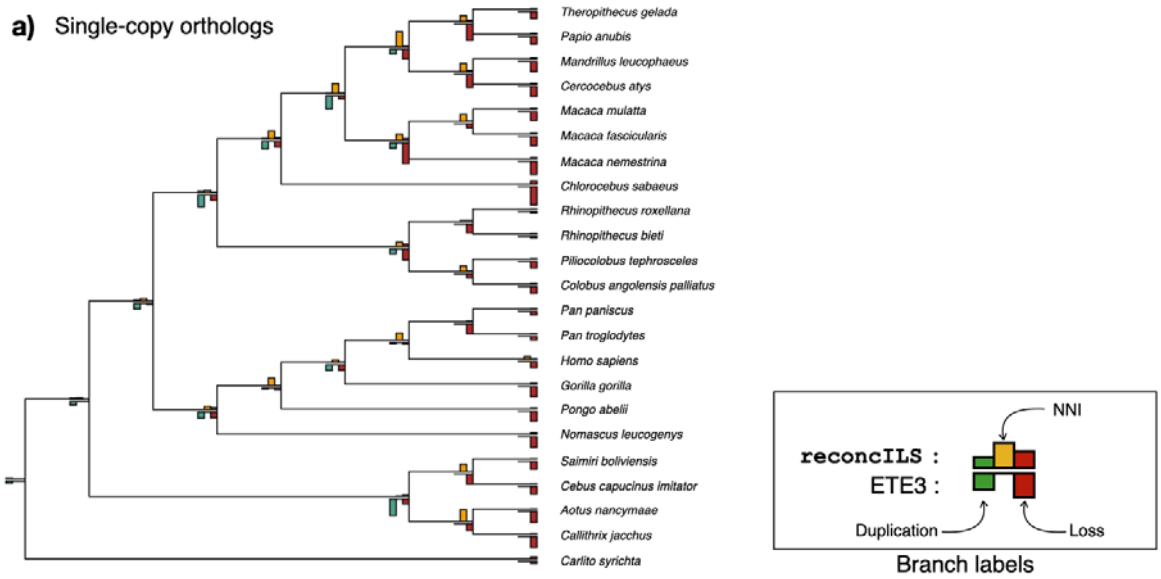
402 Analysis of primate genomes

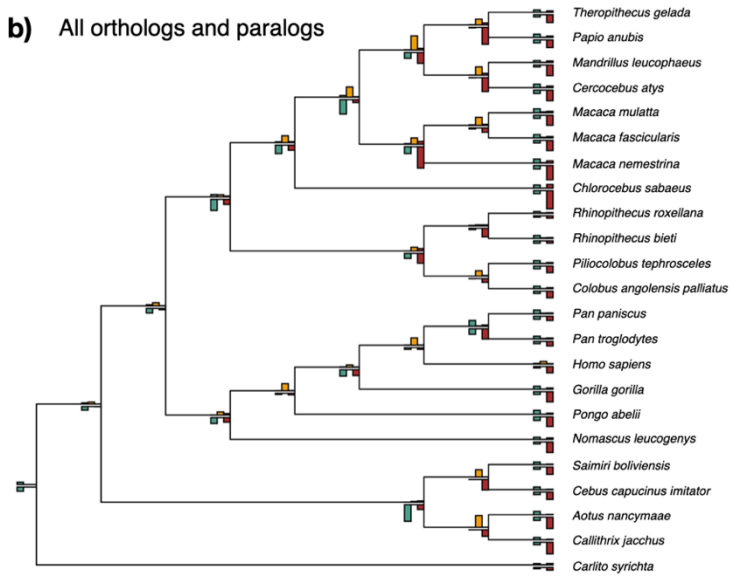
403 To further demonstrate the power and accuracy of `reconcILS`, we applied it to a
404 dataset of gene trees inferred from 23 primate genomes (Vanderpool et al. 2020; Smith et
405 al. 2022). These genomes all come from the suborder Haplorhini, which includes all
406 primates except the lemurs and lorises (suborder Strepsirhini). Among these genomes, we
407 analyzed two sets of gene trees that were previously inferred: First, we considered 1,820
408 single-copy orthologs. This dataset is not expected to contain any duplicates, though it does
409 contain losses. However, because there are high levels of ILS among the primates sampled
410 here, the dataset serves as an ideal comparator for reconciliation approaches with and
411 without ILS. Second, 11,555 trees containing both orthologs and paralogs were used to
412 infer duplications, losses, and ILS among the primates. We used both `reconcILS` and
413 ETE3 to analyze the two datasets, assigning inferred events to each branch of the primate
414 tree.

415 *Single-copy ortholog dataset*

416 Results obtained from analysis of only the single-copy orthologs are shown
417 graphically in Figure 5a (see Supplementary Figure S5a for numerical values on each
418 branch). The pattern that stands out from this graph is that ETE3 is vastly overestimating
419 the number of duplications and, especially, the number of losses. Because ETE3 does not
420 model ILS, all such events must be reconciled by one duplication (placed above the
421 discordant branch) and three losses (placed below the discordant branch; Hahn 2007). We
422 can see this effect in the data by looking at the correlation between the gene concordance
423 factor (gCF) estimated for a branch with the number of duplicates inferred on its parent
424 branch. Gene concordance factors quantify the fraction of loci that contain the same branch
425 as the species tree; when single-copy genes are analyzed, low gCFs can indicate

discordance due to ILS (Baum 2007; Lanfear and Hahn 2024). The correlation between gCFs for these data (estimated in Smith et al. 2022) and duplications inferred by ETE3 on parent branches is ρ=-0.56. 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.





434 ***Figure 5. Results from primates.*** *a) Analysis of 1,820 single-copy orthologs. Both*
435 *`reconcILS` and ETE3 were used to analyze the data, with duplication, NNI, and loss events*
436 *mapped to each branch of the species tree. On each branch, the bar plots show results from*
437 *`reconcILS` on top, with results from ETE3 mirrored below. ETE3 cannot infer NNI events, so*
438 *this bar is always set to 0. b) Analysis of 11,555 gene trees containing both orthologs and*
439 *paralogs. Results are presented in the same manner as in panel a. However, note that the*
440 *height of all bars is normalized within a panel, so that heights of bars within a tree can be*
441 *compared among branches, but the heights of bars between the trees cannot be compared.*
442 *Supplementary Figure S5 shows the absolute counts of all events from both panels and*
443 *Supplementary Figure S6 shows species tree with branch lengths in substitution units rather*
444 *than as a cladogram, as shown here.*

445

446 Using the same reasoning, we can assess the accuracy of `reconcILS` by examining
447 the correlation between NNI events inferred by our algorithm and gCF values for single-
448 copy genes. High NNI counts indicate more ILS, and therefore this measure should be
449 negatively correlated with gCF values for the same branch. Indeed, the correlation between
450 NNI counts and gCF values is ρ=-0.88, demonstrating the high accuracy of `reconcILS`.
451 However, `reconcILS` does sometimes infer a small number of duplications (Figure 5a;
452 Supplementary Figure S5a). While it is possible that some of the single-copy genes we
453 analyzed represent so-called pseudoorthologs—and therefore do contain actual
454 duplication events—these are expected to be rare (Smith and Hahn 2022). Instead, these
455 false duplicates are likely due to ILS among four or more lineages. Recall that, although our
456 algorithm has been developed to account for rearrangements due to the coalescent process,
457 ILS among four lineages can result in gene trees that are three NNI moves away from the
458 species tree; ILS among more lineages will produce trees that are even further away.
459 Therefore, multiple NNI events are sometimes simply more costly than duplication and
460 loss, even when ILS is the true cause of discordance.

461 ## Full gene tree dataset

462 We also examined patterns of ILS, duplication, and loss for the entire primate
463 dataset. The results from the full set of gene trees (Figure 5b; Supplementary Figure S5b)
464 largely mirror the results from the single-copy orthologs (which are a subset of the entire
465 dataset). Part of the apparent similarity is simply due to normalizing the two datasets
466 separately, but several non-graphical factors also contribute to the pattern. First, although
467 there are many true duplicates in the full dataset, the vast majority of gene trees only have
468 events on tip branches: 7,693 out of 11,555 gene trees have duplicates specific to one
469 species (Smith et al. 2022). Another ~1,000 trees have duplicates specific to two species.
470 Combined with the 1,820 single-copy genes, it becomes clear that there are in fact
471 relatively few trees with duplicates on internal branches even in the full dataset.

472 Second, errors in inferring duplications and losses due to the coalescent process still
473 dominate the results from standard DL reconciliation (i.e. ETE3), possibly even more so
474 than in the single-copy dataset. Because ILS (and introgression) can affect all gene trees,
475 trees with duplicates can experience twice as much discordance when there are twice as
476 many branches. One example of this can be seen among the Papionini (macaques, baboons,
477 geladas, and mandrills): ETE3 has to infer >6,000 losses on each of the tip-branches in this
478 clade, and more than 10,000 losses on most of the internal branches (Figure 5b;
479 Supplementary Figure S5b). There are also many thousands of duplicates inferred on
480 internal branches of this subclade, though these species have the same number of total
481 protein-coding genes as all the others analyzed here (Vanderpool et al. 2020). Although the
482 inferred duplications and losses in this group were also relatively higher in the single-copy
483 dataset, it is clear that even a small number of true duplicates can have an outsized effect
484 on methods that cannot account for ILS.

485 The analysis using `reconcILS` is also highly similar between the two datasets,
486 again largely because most duplicates are confined to tip branches. `reconcILS` is able to
487 deal with discordance due to the coalescent, even in the presence of duplication: the
488 correlation between NNI events inferred on each branch of the species tree from the single-
489 copy orthologs and the full dataset is $\rho = 0.99$. Although the numbers of events are of

490 course larger with the larger dataset (Supplementary Figure S5b), the effect of ILS on any
491 given branch of the species tree is experienced by all gene trees. Given its ability to deal
492 with ILS, duplication events across the primate tree are more reliably inferred with
493 `reconcILS`.

494 Discussion

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

503 Despite our demonstration of good performance in practice, our method is heuristic:
504 we have no theoretical proof that our method finds most parsimonious reconciliations.
505 While such proofs often require strict assumptions—e.g. the proofs for DL algorithms
506 assume there is no ILS—they are very helpful in providing confidence under the specified
507 conditions. However, given the high accuracy of `reconcILS` demonstrated with both
508 simulations and empirical data from single-copy genes, we hope the heuristic nature of our
509 algorithm is not a barrier to its utility.

510 As mentioned earlier, there is often ambiguity in how to label nodes in gene trees
511 affected by ILS and duplication (e.g. Figures 3b and 3c). This is due in part to the nature of
512 ILS—it is not clear which node to label because there is not a one-to-one map between
513 gene tree nodes and species tree nodes—and in part to the intrinsic lack of information
514 contained only within tree topologies when coalescence and duplication interact. The
515 difficulty in labeling is not confined to this work: DLCpar does not label nodes, and
516 RecPhyloXML, a format for representing reconciled trees, does not include a way to label
517 ILS events at all (Duchemin et al. 2018).

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

537 The algorithm used by `reconcILS` deals, in part, with an additional interaction
538 between duplication, loss, and coalescence: hemiplasy. Hemiplasy is the phenomenon by
539 which a single mutational event on a discordant gene tree will be interpreted as multiple
540 events ("homoplasy") when reconciled to the species tree (Avise and Robinson 2008; Hahn
541 and Nakhleh 2016). In analyses of nucleotide substitutions, generic binary traits, and even
542 continuous traits, hemiplasy requires that mutations have occurred on a discordant branch
543 of a gene tree (Mendes and Hahn 2016; Guerrero and Hahn 2018; Hibbins et al. 2023). In
544 the context of either duplication or loss, such mutations would have to occur on a
545 discordant branch of a gene tree to lead to copy-number hemiplasy. The simulation engine
546 introduced here, `dupcoal`, simulates hemiplasy events and reports any gene trees
547 affected. Our results show that these events lead to increased error in reconciliation, and

548 are therefore an important process to include. Interestingly, such events have little effect
549 on species tree inference when paralogs are used for this purpose (Li et al. 2024).

550 Unfortunately, most copy-number hemiplasy events as defined above are still not
551 accurately inferred by `reconcILS`: we have found that our algorithm consistently turns
552 NNIs into losses when hemiplasy occurs. While the total number of events does not seem to
553 increase, it is not ideal that NNIs are mistaken for losses. However, there is also a broader
554 definition of copy-number hemiplasy in use, one that `reconcILS` does deal with
555 accurately (and is also simulated by `dupcoal`). This broader definition includes any
556 duplication or loss that does not fix in all descendant lineages, regardless of underlying
557 gene tree topologies (Rasmussen and Kellis 2012; Wu et al. 2014). While this definition
558 encompasses the stricter one given above, it also includes cases such as the one shown in
559 Figure 1, which the stricter definition does not include. Because the mutation in that figure
560 occurs in the common ancestral population of species B and C, but was only inherited by
561 lineage C, this scenario would be considered to be hemiplasy by the broader definition
562 (although it technically does not even involve ILS if the outcome resembles the gene trees
563 in Figure 1c-e, since coalescence is happening in the same population). The number of each
564 type of event in such cases is accurately inferred using `reconcILS`; DLCpar assumes that
565 this type of hemiplasy does not occur, and SimPhy does not simulate this type of hemiplasy,
566 either. Note, however, that if the duplication event (the star in Figure 1a) occurred a few
567 generations later—after the split between species B and C, along a tip lineage—the
568 scenario would not be considered hemiplasy by this broader definition, but would still
569 produce the same range of gene trees (i.e. Figure 1c-g). (This would now involve ILS,
570 however.) Regardless of the exact definitions and terms used, it should be clear that
571 duplication, loss, and coalescence can interact to produce a wide array of topologies not
572 considered by most reconciliation approaches.

573 The algorithm used by `reconcILS` deals with ILS differently than does DLCpar.
574 While DLCpar treats ILS as an event, `reconcILS` considers NNIs as events induced by ILS.
575 It is not clear that either approach should necessarily be preferred on first principles. In
576 simple cases involving three species and a single discordant branch (e.g. Figure 2), ILS and
577 NNI are more or less equivalent, though it should be noted that the process of ILS can also

578 lead to concordant trees and is not synonymous with discordance. However, in cases with
579 more than one discordant branch there is no simple relationship between the two. For
580 instance, in a 4-tip tree undergoing ILS, lineages may undergo ILS across two branches, but
581 result in between 0 and 3 NNI events (Than and Rosenberg 2013). We think that directly
582 modeling NNI—as is done in `reconcILS`—is a natural way to express the events needed
583 to reconcile a gene tree with a species tree.

584 One relatively straightforward improvement that could be made to `reconcILS` is
585 in the cost of different events. Costs must be specified in `reconcILS` ahead of time, but it
586 is difficult to determine biologically realistic costs for duplication, loss, and ILS. One
587 approach to determining duplication and loss costs would be to infer the number of such
588 events using an alternative, non-reconciliation approach (e.g. Mendes et al. 2020), setting
589 the cost in `reconcILS` to be inversely related to the inferred number of events of each
590 type. Varying rates of duplication and loss can also be inferred along different branches of
591 the species tree, and so the costs of these events could vary commensurately along the tree.
592 Similarly, we can envision a simple way to scale the costs of NNI with the length of species
593 tree branches: this is because the probability of discordance due to ILS is inversely
594 proportional to internal branch lengths on a species tree (Hudson 1983). We could
595 therefore make NNI events on longer branches of the species tree more costly, which would
596 properly weight the relative probability of such events locally on the tree.

597 There are also several extensions that could be made to `dupcoal`, our new
598 simulation engine. Currently, `dupcoal` draws a daughter gene tree from the multispecies
599 coalescent (MSC) at random: this implicitly assumes that the parent and daughter trees are
600 completely unlinked. It should be straightforward, however, to draw the daughter tree at a
601 given recombination distance from the parent tree, ranging from fully linked to fully
602 unlinked. In addition, though the software currently draws gene trees from the MSC, we
603 could also include introgression as a source of discordance by drawing trees from the
604 multispecies network coalescent (Meng and Kubatko 2009; Yu et al. 2011). In this case,
605 users would simply input a species network to simulate from, rather than a species tree.
606 We should also note the differences between the current version of `dupcoal` and
607 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 the 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; its 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.

Methods

Building and testing `dupcoal`

The simulator `dupcoal` is written in Python 3 and uses functions from DendroPy (Sukumaran and Holder 2010) to generate trees under the MSC and to manipulate parent

635 and daughter trees. It also uses functions from NumPy (Harris et al. 2020) to draw
636 coalescent times, duplication times, and loss times from exponential distributions.

637 To test the accuracy of `dupcoal` we simulated gene trees along a species tree with
638 40 tips (Supplementary Figure S1). The species tree topology was drawn at random using
639 the program ape (Paradis and Schliep 2018), with the "Grafen" method for drawing branch
640 lengths (scaled up by 10); this tree was used as input to both `dupcoal` and SimPhy.
641 Because `dupcoal` specifies branch lengths in coalescent units, while SimPhy uses time,
642 branch lengths were converted by assuming a population size of N=10 and a generation
643 time of 1 year. Simulations were run on this tree with only ILS (setting duplication and loss
644 rates to $\lambda = 0$ and $\mu = 0$, respectively). When using positive duplication and loss rates (see
645 next paragraph), they were scaled to the appropriate branch lengths to make them
646 comparable in `dupcoal` and SimPhy. Comparisons between `dupcoal` and SimPhy with no
647 duplication and loss were carried out by calculating gene concordance factors (gCFs) on the
648 species tree using IQ-TREE2 (Minh et al. 2020a).

649 To test the accuracy of inferences using `reconcILS` and ETE3, the same 40-taxon
650 tree was used for simulations in `dupcoal` with ILS, duplication, and loss ($\lambda = 0.03$ and
651 $\mu = 0.03$). DLCpar is written in Python 2 and could not be run in reasonable time on the 40-
652 taxon tree. We therefore also simulated data on a 3-taxon tree with topology
653 (A:1.5,(B:0.5,C:0.5):1), where branch lengths are measured in coalescent units and
654 duplication and loss rates were the same as on the larger tree. The same costs were used
655 for all analyses of simulated data for all methods ($C_D = 1.1$, $C_L = 1$, $C_I = 1$).

656 ## Implementation of `reconcILS`

657 ### *Inputs and outputs*

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

662 ### *Traversing trees*

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

675 If there are only two adjacent nodes in the gene tree mapped to a node in the species
676 tree (i.e. one discordant branch, as in Figure 2), then the method proceeds as described
677 below. However, when there are more than two nodes mapped to an internal species tree
678 node—and therefore more than one discordant branch in the gene tree—we must decide
679 which branch to reconcile with the multiply mapped node first. We have found cases where
680 the order in which the branches are resolved matters, with only one of the choices leading
681 to an optimal reconciliation. Algorithm 2 (see Supplementary Materials) describes an
682 algorithm for choosing which discordant branch to locally reconcile first. This algorithm
683 calculates the "bipartition cost" of each possible branch—the number of bipartitions and
684 losses induced by NNI moves applied to each branch (Supplementary Figure S7). Such a
685 calculation can be carried out quickly, and by choosing the branch with the lowest
686 bipartition cost, we can more accurately proceed with reconciliation.

687 Once reconciliation for a single multiply mapped node in the species tree is carried
688 out (i.e. it is no longer multiply mapped), our algorithm begins again: a new round. In each
689 round we generate an LCA mapping function and continue traversing the species tree,
690 counting gene tree nodes mapped to each species tree node.

691 ### *Choosing among events*

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

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

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

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

720 ### *Labeling events*

721 Our approach for labeling ILS on reconciled gene trees is described above in the
722 "New Approaches" section. Here, we consider two important caveats to this approach, one
723 perhaps more obvious than the other. First, a specific labeling is associated with a specific
724 reconciliation. That is, if we had found a different optimal reconciliation for the tree in
725 Figure 3a (for instance, one duplication and two losses), we would have had different
726 labeled events. Therefore, a labeling is always reported in the context of a particular
727 reconciliation.

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

744 ### Primate data

745 We used two sets of gene trees that were previously constructed in Smith et al.
746 (2022) using IQ-TREE2 (Minh et al. 2020b), based on data from Vanderpool et al. (2020).
747 The first set consists of 1,820 single-copy orthologs; this set was denoted as "Single-copy
748 clusters" in Smith et al. (2022). The second set consists of 11,555 trees containing both

749 orthologs and paralogs; this set was denoted "All paralogs" in Smith et al. (2022). We
750 randomly resolved all polytomies. When using `reconcILS`, costs were set to $C_D = 2$,
751 $C_L = 2$, and $C_I = 1$.

752

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 and DBI-2146866 to MWH.

Data availability

`reconcILS` is available at https://github.com/smishra677/reconcILS. `dupcoal` is available at https://github.com/meganlsmith/dupcoal.

766 ## References

767 Avise, J. C., and T. J. Robinson. 2008. Hemiplasy: A new term in the lexicon of phylogenetics.
768 Systematic Biology **57**:503-507.

769 Bansal, M. S., E. J. Alm, and M. Kellis. 2012. Efficient algorithms for the reconciliation
770 problem with gene duplication, horizontal transfer and loss. Bioinformatics **28**:i283-
771 i291.

772 Baum, D. A. 2007. Concordance trees, concordance factors, and the exploration of reticulate
773 genealogy. Taxon **56**:417-426.

774 Boussau, B., and C. Scornavacca. 2020. Reconciling gene trees with species trees. Pp. 3.2:1-
775 3.2:23 *in* C. Scornavacca, F. Delsuc, and N. Galtier, eds. Phylogenetics in the Genomic
776 Era. No commercial publisher | Authors open access book.

777 Chan, Y.-b., V. Ranwez, and C. Scornavacca. 2017. Inferring incomplete lineage sorting,
778 duplications, transfers and losses with reconciliations. Journal of Theoretical Biology
779 **432**:1-13.

780 Chaudhary, R., J. G. Burleigh, and O. Eulenstein. 2012. Efficient error correction algorithms
781 for gene tree reconciliation based on duplication, duplication and loss, and deep
782 coalescence. BMC Bioinformatics **13**:S11.

783 Chen, K., D. Durand, and M. Farach-Colton. 2000. NOTUNG: A program for dating gene
784 duplications and optimizing gene family trees. Journal of Computational Biology
785 **7**:429-447.

786 Christensen, S., E. K. Molloy, P. Vachaspati, A. Yammanuru, and T. Warnow. 2020. Non-
787 parametric correction of estimated gene trees using TRACTION. Algorithms for
788 Molecular Biology **15**:1.

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

793 Duchemin, W., G. Gence, A.-M. Arigon Chifolleau et al. 2018. RecPhyloXML: a format for
794 reconciled gene trees. Bioinformatics **34**:3646-3652.

795 Durand, D., B. V. Halldórsson, and B. Vernot. 2006. A hybrid micro-macroevolutionary
796 approach to gene tree reconstruction. Journal of Computational Biology **13**:320-335.
797 El-Mabrouk, N., and A. Ouangraoua. 2017. A general framework for gene tree correction
798 based on duplication-loss reconciliation. Pp. 8:1-8:14. WABI.
799 Emms, D. M., and S. Kelly. 2019. OrthoFinder: phylogenetic orthology inference for
800 comparative genomics. Genome Biology **20**:238.
801 Fitch, W. M. 1970. Distinguishing homologous from analogous proteins. Systematic Zoology
802 **19**:99-113.
803 Gillespie, J. H., and C. H. Langley. 1979. Are evolutionary rates really variable? Journal of
804 Molecular Evolution **13**:27-34.
805 Goodman, M., J. Czelusniak, G. W. Moore, A. E. Romero-Herrera, and G. Matsuda. 1979.
806 Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by
807 cladograms constructed from globin sequences. Systematic Zoology **28**:132-163.
808 Guerrero, R. F., and M. W. Hahn. 2018. Quantifying the risk of hemiplasy in phylogenetic
809 inference. Proceedings of the National Academy of Sciences **115**:12787-12792.
810 Guigo, R., I. Muchnik, and T. F. Smith. 1996. Reconstruction of ancient molecular phylogeny.
811 Molecular Phylogenetics and Evolution **6**:189-213.
812 Hahn, M. W. 2007. Bias in phylogenetic tree reconciliation methods: implications for
813 vertebrate genome evolution. Genome Biology **8**:R141.
814 Hahn, M. W., and L. Nakhleh. 2016. Irrational exuberance for resolved species trees.
815 Evolution **70**:7-17.
816 Hallett, M., J. Lagergren, and A. Tofigh. 2004. Simultaneous identification of duplications
817 and lateral transfers. Pp. 347-356. Proceedings of the eighth annual international
818 conference on Research in computational molecular biology (RECOMB '04).
819 Hallett, M. T., and J. Lagergren. 2001. Efficient algorithms for lateral gene transfer
820 problems. Pp. 149-156. Proceedings of the fifth annual international conference on
821 Computational biology (RECOMB '01).
822 Harris, C. R., K. J. Millman, S. J. Van Der Walt et al. 2020. Array programming with NumPy.
823 Nature **585**:357-362.

824 Hibbins, M. S., L. C. Breithaupt, and M. W. Hahn. 2023. Phylogenomic comparative methods:
825 Accurate evolutionary inferences in the presence of gene tree discordance.
826 Proceedings of the National Academy of Sciences **120**:e2220389120.
827 Hudson, R. R. 1983. Testing the constant-rate neutral allele model with protein sequence
828 data. Evolution **37**:203-217.
829 Huerta-Cepas, J., H. Dopazo, J. Dopazo, and T. Gabaldón. 2007. The human phylome.
830 Genome Biology **8**:R109.
831 Huerta-Cepas, J., F. Serra, and P. Bork. 2016. ETE 3: Reconstruction, analysis, and
832 visualization of phylogenomic data. Molecular Biology and Evolution **33**:1635-1638.
833 Jacox, E., C. Chauve, G. J. Szöllősi, Y. Ponty, and C. Scornavacca. 2016. ecceTERA:
834 comprehensive gene tree-species tree reconciliation using parsimony. Bioinformatics
835 **32**:2056-2058.
836 Kingman, J. F. C. 1982. The coalescent. Stochastic Processes and their Applications **13**:235-
837 248.
838 Lanfear, R., and M. W. Hahn. in press. The meaning and measure of concordance factors.
839 Molecular Biology and Evolution.
840 Li, Q., Y.-b. Chan, N. Galtier, and C. Scornavacca. 2024. The effect of copy number hemiplasy
841 on gene family evolution. Systematic Biology **73**:355-374.
842 Li, Q., C. Scornavacca, N. Galtier, and Y.-B. Chan. 2021. The multilocus multispecies
843 coalescent: A flexible new model of gene family evolution. Systematic Biology **70**:822-
844 837.
845 Libeskind-Hadas, R., Y.-C. Wu, M. S. Bansal, and M. Kellis. 2014. Pareto-optimal phylogenetic
846 tree reconciliation. Bioinformatics **30**:i87-i95.
847 Maddison, W. P. 1997. Gene trees in species trees. Systematic Biology **46**:523-536.
848 Mallo, D., L. De Oliveira Martins, and D. Posada. 2014. Unsorted homology within locus and
849 species trees. Systematic Biology **63**:988-992.
850 Mallo, D., L. De Oliveira Martins, and D. Posada. 2016. *SimPhy*: Phylogenomic simulation of
851 gene, locus, and species trees. Systematic Biology **65**:334-344.
852 Mawhorter, R., and R. Libeskind-Hadas. 2019. Hierarchical clustering of maximum
853 parsimony reconciliations. BMC Bioinformatics **20**:612.

854 Mendes, F. K., and M. W. Hahn. 2016. Gene tree discordance causes apparent substitution
855 rate variation. Systematic Biology **65**:711-721.
856 Mendes, F. K., D. Vanderpool, B. Fulton, and M. W. Hahn. 2020. CAFE 5 models variation in
857 evolutionary rates among gene families. Bioinformatics **36**:5516-5518.
858 Meng, C., and L. S. Kubatko. 2009. Detecting hybrid speciation in the presence of incomplete
859 lineage sorting using gene tree incongruence: A model. Theoretical Population Biology
860 **75**:35-45.
861 Minh, B. Q., M. W. Hahn, and R. Lanfear. 2020a. New methods to calculate concordance
862 factors for phylogenomic datasets. Molecular Biology and Evolution **37**:2727-2733.
863 Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, Lanfear
864 R. 2020b. IQ-TREE 2: New models and efficient methods for phylogenetic inference in
865 the genomic era. Molecular Biology and Evolution **37**:1530-1534.
866 Morel, B., A. M. Kozlov, A. Stamatakis, and G. J. Szöllősi. 2020. GeneRax: A tool for species-
867 tree-aware maximum likelihood-based gene family tree inference under gene
868 duplication, transfer, and loss. Molecular Biology and Evolution **37**:2763-2774.
869 Nguyen, T. H., V. Ranwez, S. Pointet, A.-M. A. Chifolleau, J.-P. Doyon, and V. Berry. 2013.
870 Reconciliation and local gene tree rearrangement can be of mutual profit. Algorithms
871 for Molecular Biology **8**:12.
872 Page, R. D. M. 1994. Maps between trees and cladistic analysis of historical associations
873 among genes, organisms, and areas. Systematic Biology **43**:58-77.
874 Page, R. D. M., and M. A. Charleston. 1997. From gene to organismal phylogeny: Reconciled
875 trees and the gene tree/species tree problem. Molecular Phylogenetics and Evolution
876 **7**:231-240.
877 Paradis, E., and K. Schliep. 2018. ape 5.0: an environment for modern phylogenetics and
878 evolutionary analyses in R. Bioinformatics **35**:526-528.
879 Rasmussen, M. D., and M. Kellis. 2012. Unified modeling of gene duplication, loss, and
880 coalescence using a locus tree. Genome Research **22**:755-765.
881 Richter, D. J., P. Fozouni, M. B. Eisen, and N. King. 2018. Gene family innovation,
882 conservation and loss on the animal stem lineage. eLife **7**:e34226.

883 Schrider, D. R., and M. W. Hahn. 2010. Lower linkage disequilibrium at CNVs is due to both
884 recurrent mutation and transposing duplications. Molecular Biology and Evolution
885 **27**:103-111.
886 Smith, M. L., and M. W. Hahn. 2022. The frequency and topology of pseudoorthologs.
887 Systematic Biology **71**:649-659.
888 Smith, M. L., D. Vanderpool, and M. W. Hahn. 2022. Using all gene families vastly expands
889 data available for phylogenomic inference. Molecular Biology and Evolution
890 **39**:msac112.
891 Stolzer, M., H. Lai, M. Xu, D. Sathaye, B. Vernot, and D. Durand. 2012. Inferring duplications,
892 losses, transfers and incomplete lineage sorting with nonbinary species trees.
893 Bioinformatics **28**:i409-i415.
894 Stull, G. W., X.-J. Qu, C. Parins-Fukuchi et al. 2021. Gene duplications and phylogenomic
895 conflict underlie major pulses of phenotypic evolution in gymnosperms. Nature Plants
896 **7**:1015-1025.
897 Sukumaran, J., and M. T. Holder. 2010. DendroPy: a Python library for phylogenetic
898 computing. Bioinformatics **26**:1569-1571.
899 Szöllősi, G. J., W. Rosikiewicz, B. Boussau, E. Tannier, and V. Daubin. 2013. Efficient
900 exploration of the space of reconciled gene trees. Systematic Biology **62**:901-912.
901 Tajima, F. 1983. Evolutionary relationships of DNA sequences in finite populations.
902 Genetics **105**:437-460.
903 Than, C. V., and N. A. Rosenberg. 2013. Mathematical properties of the deep coalescence
904 cost. IEEE/ACM Transactions on Computational Biology and Bioinformatics **10**:61-72.
905 Thomas, G. W. C., S. H. Ather, and M. W. Hahn. 2017. Gene-tree reconciliation with MUL-
906 trees to resolve polyploidy events. Systematic Biology **66**:1007-1018.
907 Vanderpool, D., B. Q. Minh, R. Lanfear et al. 2020. Primate phylogenomics uncovers multiple
908 rapid radiations and ancient interspecific introgression. PLoS Biology **18**:e3000954.
909 Vernot, B., M. Stolzer, A. Goldman, and D. Durand. 2008. Reconciliation with non-binary
910 species trees. Journal of Computational Biology **15**:981-1006.
911 Waskom, M. 2021. Seaborn: Statistical data visualization. Journal of Open Source Software
912 **6**:3021.

913 Wu, Y.-C., M. D. Rasmussen, M. S. Bansal, and M. Kellis. 2014. Most parsimonious
914 reconciliation in the presence of gene duplication, loss, and deep coalescence using
915 labeled coalescent trees. Genome Research **24**:475-486.
916 Yu, Y., C. Than, J. H. Degnan, and L. Nakhleh. 2011. Coalescent histories on phylogenetic
917 networks and detection of hybridization despite incomplete lineage sorting.
918 Systematic Biology **60**:138-149.
919 Zhang, C., C. Scornavacca, E. K. Molloy, and S. Mirarab. 2020. ASTRAL-Pro: Quartet-based
920 species-tree inference despite paralogy. Molecular Biology and Evolution **37**:3292-
921 3307.
922 Zmasek, C. M., and S. R. Eddy. 2001. A simple algorithm to infer gene duplication and
923 speciation events on a gene tree. Bioinformatics **17**:821-828.
924