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

Sarthak R. 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 infer 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 generates gene trees produced by the interaction of duplication, loss, and ILS under the MSC-DL model. Despite being a heuristic method, `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 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 or horizontal gene transfer. 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. Rasmussen and Kellis (2012) introduced a probabilistic model considering duplication, loss, and ILS and an associated reconciliation algorithm. While promising, DLCoalRecon is highly parameterized, requiring users to specify rates of duplication and loss, population sizes, and more prior to inference, and is computationally intensive. Wu et al. (2014) developed a parsimony-based algorithm (DLCpar) based on this same model that 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—see section below on "Complex coalescence"). 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 run it on larger datasets or include its algorithm within 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 genomes from primates, a dataset that has previously been shown to contain a large amount of discordance due to ILS (Vanderpool et al. 2020).

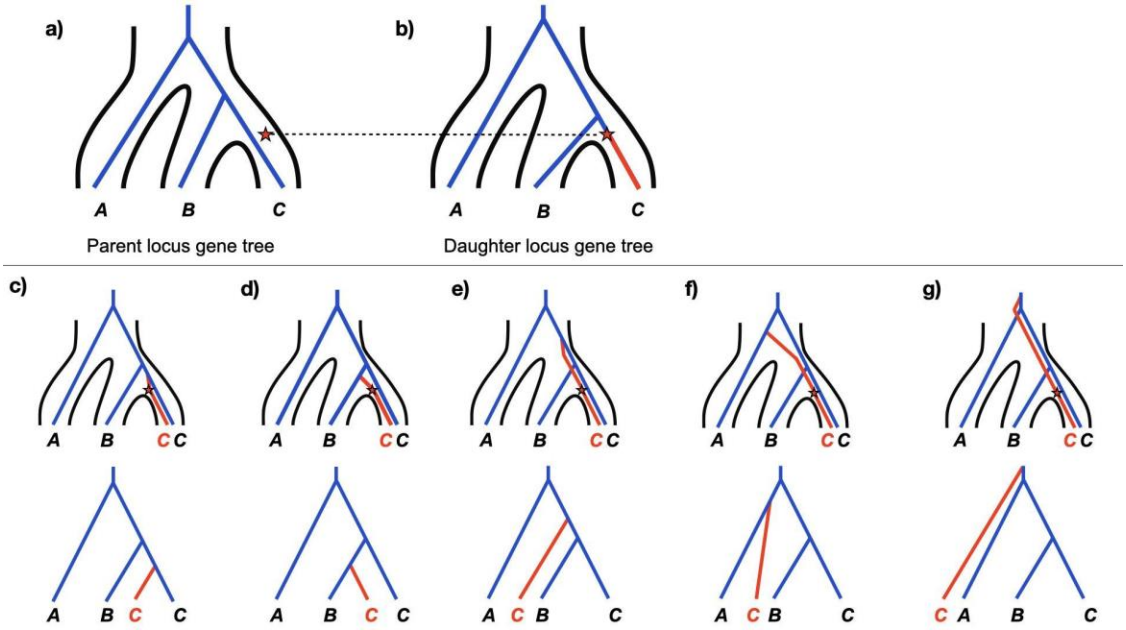

Figure 1. The MSC-DL model generates full gene trees. *a) A gene tree at the parent locus (blue) is drawn from the species tree (outlined in black) via the multispecies coalescent (MSC) model. A duplication event (red star) is drawn from a birth-death model, in this case occurring in the population ancestral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

The multispecies coalescent model with duplication and loss

We begin with a description of the multispecies coalescent model with duplication and loss (MSC-DL). Our model is very similar to the multilocus multispecies coalescent (MLMSC) model of Li et al. (2021). The description of the MLMSC model and an algorithm for generating trees under this model were first given in Li et al. (2021), with a mathematical model and further algorithmic updates in Li et al. (2024). The Appendix goes over the conceptual and algorithmic differences between the MSC-DL and MLMSC. Below, we describe the most important biological parts of our model, leaving algorithmic details to the next section. This explanation is intended to make clearer which common biological processes are modeled by the MSC-DL, but not by standard DL and DTL methods.

The MSC-DL model treats each individual locus (i.e. non-recombing position in the genome) as evolving via the multispecies coalescent (MSC) model, in which discordance can arise only via incomplete lineage sorting. The MSC-DL further models relationships among these loci via duplication and coalescence, in addition to losses at any locus. In particular, the MSC-DL specifies:

i. Every locus in the genome has a history (sometimes called a “locus tree”) that follows the species tree.
ii. Each locus in the genome has a gene tree topology drawn from the species tree according to the MSC model. This gene tree (or “genealogy”) exists at this locus whether or not a duplication or loss has occurred.

121 iii. Duplication events copy genetic material—i.e. DNA sequence—from one locus (the
122 "parent" locus) and insert it at another locus (the "daughter" locus). At the parent
123 locus, duplicated genetic material descends from a lineage that existed at the time of
124 duplication, and so must coalesce above this point with other lineages at the parent
125 locus according to the MSC (Figure 1a). At the daughter locus, this material inserts
126 into (is inherited by) a randomly chosen lineage of the gene tree at the time of
127 duplication (e.g. lineage *C* in Figure 1b), following the gene tree of this lineage after
128 this point. Nucleotide mutations occur in the sequence that was duplicated, allowing
129 us to reconstruct gene trees connecting one or more loci.
130 iv. Parent and daughter loci can be any recombination distance apart. If they are
131 completely linked they share the same gene tree topology, while if they are unlinked
132 each locus has an independent topology.
133 v. Duplications and losses are generated at rates μ and λ via a birth-death process,
134 starting at the original locus. Each duplication event adds a sub-tree from a new
135 daughter locus, consisting only of lineages descending from the duplication (see
136 point iii). Losses remove descendant branches of gene trees below their occurrence
137 (see point vii).
138 vi. Duplications and losses can occur on any gene tree lineage at any locus that contains
139 genetic material descended from the original parent locus, regardless of whether
140 they were the original parent or daughter locus.
141 vii. Loss events do not remove a locus, but rather do not allow lineages descended from
142 a loss to extend to the present (Supplementary Figure S1). These lineages are
143 therefore not present in the full gene tree.
144 viii. The gene tree at each daughter locus coalesces with the gene tree at its parent locus
145 via the MSC (Figure 1c-g; Supplementary Figure S2). Note that even if parent and
146 daughter loci are completely linked, they still must coalesce together at the parent
147 locus—linkage only determines whether the two marginal gene trees have the same
148 topology or not.
149 ix. The "full" gene tree consists of the combined genealogies at all parent and daughter
150 loci together, including all duplication and loss events.

Simulation using `dupcoal`

We created a new simulator—`dupcoal`—for generating gene trees under the MSC-DL model. This simulator was originally needed because it was not possible to track events using the simulator of Li et al. (2021; the only one available at the time), and to evaluate the performance of `reconcILS` we require detailed output. `dupcoal` generates data under the MSC-DL model, which differs in some ways from the MLMSC model (see Appendix). Our algorithm works as follows (code can be found at https://github.com/meganlsmith/dupcoal):

1. Generate the original parent gene tree from the species tree according to the MSC model.
2. Simulate duplications along the branches of existing gene trees (only once for any tree) at a rate λ.
3. For each duplication, generate a daughter tree at a new locus according to the MSC model. Then, place the duplication on the daughter tree by randomly selecting a lineage available at the time of duplication on the appropriate branch of the species tree (Figure 1b). Discard all other lineages from the daughter locus, and truncate the sub-tree at the time of duplication.
4. Return to step 2 and recursively generate new duplications (and subsequent daughter trees) on each new daughter tree. Proceed to step 5 when no additional duplications have been generated on the last daughter tree.
5. Coalesce all trees together. Begin at the present and work backwards in time. First, attempt to coalesce a daughter tree with its parent, which could itself be a duplicated tree. If coalescence fails to occur before the tree originates, attempt to coalesce the daughter tree with the parent of its parent, and so on. Repeat until all trees have been joined together.
6. Generate losses on the final, full gene family tree at a rate μ.

In general, this algorithm is similar to the updated "MLMSC-II" simulator of Li et al. (2024) and is capable of generating copy-number hemiplasy, along with several other interactions between ILS, duplication, and loss not modeled by DL and DTL approaches

180 (see next section). However, our algorithm differs in a few ways from those of Li et al.
181 (2021; 2024); these differences are discussed in the Appendix.

182 While our simulator does not currently include the ability to simulate linked
183 duplications, introgression, or horizontal gene transfers, it is straightforward to add these
184 types of events in future iterations of our software. Linked duplications simply require that
185 the topology of the daughter gene tree is correlated with the topology of the parent gene
186 tree (proportionally to recombination distance), while both introgression and transfer can
187 be accommodated by drawing gene trees from a different locus history (e.g. a history of
188 introgression) in step 1 or step 3.

189 Complex coalescence: ILS, unsorted homologs, and copy-number hemiplasy

190 The process we are modeling—the interaction of duplication, loss, and
191 coalescence—can be quite complex, leading to both conceptual and linguistic confusion.
192 Here we briefly attempt to clarify some important terms that have been used in different
193 ways by different researchers. While we try to avoid introducing new language, we do want
194 to clearly define how terms are being used in this paper and their relation to previous uses.
195 *Incomplete lineage sorting (ILS)*. ILS occurs when two or more lineages do not
196 coalesce in the population in which they co-exist. When this happens, these lineages enter
197 an ancestral population, at which time they may again coalesce, either with each other or
198 with additional lineages that are present. If ILS results in lineages from the same original
199 population coalescing first with a lineage from a different population, gene trees will be
200 discordant with the species tree. This process can result in discordant gene trees at the
201 parent and daughter loci considered here.

202 Importantly for the MSC-DL model, the probability of ILS for two lineages is
203 normally expressed as e^{-T}, where T represents the length of time during which the lineages
204 can coalesce in the same population, and is usually taken as the length of the relevant
205 branch in the species tree (see Rosenberg 2002; Bryant and Hahn 2020). However, when
206 considering ILS between a pair of duplicates, the amount of time available for coalescence
207 before ILS has occurred is limited to the time between the duplication event and the next
208 speciation event (going backward in time). This time is always shorter than the length of

209 the entire branch of the species tree on which the duplication occurred, which can lead to
210 ILS even when all branches of the species tree are long. The importance of this distinction
211 will hopefully become clearer in the next sub-section.

212 *Unsorted homology*. Gene duplication is fundamentally a process by which an
213 existing haplotype in a population at one locus is inserted into another location. Because of
214 this, the inserted DNA at the new daughter locus carries with it a history of coalescence
215 with all the other haplotypes at the parent locus, and the full gene tree will reflect this
216 history. In other words, all duplication events also involve coalescence events. Mallo et al.
217 (2014) described a common phenomenon that can occur when duplicates coalesce
218 together, which they called "unsorted homology." The general idea is that standard DL/DTL
219 methods assume that duplication along a lineage also requires the joining of parent and
220 daughter trees along that same lineage; if this does not occur, there is unsorted homology.

221 As an example, duplication of a gene along the C lineage (Figure 1b) is assumed to
222 only lead to gene trees with the two copies in species C as sister to one another (Figure 1c).
223 Any alternative joining of these trees will lead DL/DTL methods to incorrect
224 reconciliations, which will in turn lead to incorrect homology assignments (i.e. orthology
225 and paralogy). However, due to the coalescent process, many other trees can be formed
226 (Figure 1d-g). Because of the mismatch between the lineage containing the duplicate and
227 the lineage to which it coalesces in Figures 1d-g, all of these trees would be considered
228 unsorted homologs. Note that, in the scenario shown here, the trees in Figures 1d and 1e
229 are equiprobable with the topology in Figure 1c, making unsorted homology the most likely
230 outcome.

231 It should be emphasized that all of the topologies presented in Figure 1 involve only
232 a single duplication and no losses, and neither the parent tree nor the daughter tree is itself
233 discordant due to ILS. In fact, the example in Figure 1d is not technically due to ILS either,
234 since coalescence occurs in the same population of the species tree in which the duplication
235 appeared (the ancestral population of species B and C).

236 This example highlights an additional complication of unsorted homology: it can
237 occur with or without ILS, and even when the species tree only contains two species—and

238 therefore neither parent nor daughter gene trees can themselves be discordant.
239 Supplementary Figure S3 shows examples of unsorted homology in a two-species tree with
240 ILS, while Supplementary Figure S4 shows an example without ILS. One can see that there
241 is very little time for coalescence (i.e. T) before ILS occurs in Supplementary Figure S3, as it
242 is the timing of duplication relative to speciation, not the length of the branch in the species
243 tree, that determines the probability of ILS. Some approaches to reduce the effects of ILS on
244 reconciliation have required the presence of "apparent duplications" (Lafond et al. 2014) or
245 "observable duplications" (Scornavacca et al. 2011)—essentially, duplication nodes that
246 have more than one copy of the gene in a descendant species. While such approaches do
247 prevent the inference of duplication when none has occurred, as shown by the examples of
248 unsorted homology here (Figure 1d-f, Supplementary Figure S3, Supplementary Figure S4),
249 they do not ensure the correct identification of duplication nodes, nor do they prevent
250 excess losses from being inferred.

251 *Hemiplasy*. Gene trees at any locus in the genome can be discordant with the species
252 tree. In such cases, the term "hemiplasy" (Avise and Robinson 2008) describes the
253 incorrect inference of homoplasy that can arise when mutations occur on a discordant
254 branch of a gene tree (i.e. an internal branch that does not exist in the species tree; see
255 Figure 1 in Hahn and Nakhleh 2016). In the context of either duplication or loss, such
256 mutations must occur on a discordant branch of a gene tree to lead to "copy-number
257 hemiplasy" (CNH). Loss events can cause CNH when they occur either on a discordant
258 parent tree (Supplementary Figure S5a) or a discordant daughter tree (Supplementary
259 Figure S5b). In contrast, duplication events can only cause CNH on discordant daughter
260 trees (Supplementary Figure S5c,d).

261 However, there is also a broader definition of copy-number hemiplasy in use, which
262 includes any duplication or loss that does not fix in all descendant lineages, regardless of
263 underlying gene tree topologies (Rasmussen and Kellis 2012; Wu et al. 2014). While this
264 definition of CNH encompasses the stricter one given above, it also includes cases such as
265 the one shown in Figure 1 (and Supplementary Figure S4), which the stricter definition
266 does not include. Because the mutation in Figure 1 occurs in the common ancestral
267 population of species B and C, but was only inherited by lineage C, this scenario would be

268 considered to be hemiplasy by the broader definition. A similar event occurs in
269 Supplementary Figure S4, even though there is no ILS, and no gene tree discordance is
270 possible. Importantly, unlike the strict definition of CNH (and unlike unsorted homology),
271 there is not necessarily a genealogical signal left by events under this broader definition of
272 CNH. For instance, if the duplication event in Figure 1a occurred a few generations later—
273 after the split between species B and C, along a tip lineage—the scenario would not be
274 considered hemiplasy by this broader definition, but would still produce the same range of
275 gene trees (i.e. Figure 1c-g). (This would now involve ILS, however.)

276 Regardless of the exact definitions and terms used, many of these important
277 processes are not modeled by common simulators (e.g. SimPhy does not simulate either
278 type of hemiplasy) and are not considered by common reconciliation algorithms (e.g.
279 DLCpar assumes that neither type of hemiplasy occurs). `Dupcoal` simulates all of these
280 phenomena—as does MLMSC-II—but also reports the occurrence of each in generating the
281 underlying gene trees.

282 Reconciliation using `reconcILS`

283 *Overview of main algorithm*

284 Our reconciliation approach uses nearest-neighbor interchange (NNI)
285 rearrangements to model ILS and coalescence events, and to avoid invoking unnecessary
286 duplications and losses. An NNI move on a rooted tree takes the three lineages defined by a
287 single internal branch, for instance (A,(B,C)), and rearranges them into the two alternative
288 trees, either (B,(A,C)) or (C,(B,A)). Any two bifurcating trees with the same set of labels at
289 the tips can always be transformed one into the other by a series of NNI rearrangements
290 (Robinson 1971).

291 NNI has been used previously in reconciliation algorithms to correct gene trees that
292 may have been inferred incorrectly (e.g. Chen et al. 2000; Chaudhary et al. 2011; Górecki
293 and Eulenstein 2011; Nguyen et al. 2013). Here, we use NNI because the effects of ILS on
294 gene trees can be modeled as a series of NNI rearrangements (Than and Rosenberg 2013),
295 not because there is necessarily a biological link between NNI and ILS. For instance, for

three species undergoing ILS, there are two possible discordant topologies, each of which is one NNI move away from the species tree topology (e.g. Figure 2). For four lineages undergoing ILS, there are 14 possible discordant topologies, each of which are between one and three NNI moves away from the species tree topology. (Note that it is the number of lineages undergoing ILS at a specific point on the tree that determines the NNI distance between trees, not the total number of tips in a tree.) Our algorithm takes advantage of this connection between NNI and ILS in order to reconcile gene trees, but can also accommodate gene tree error using the same NNI events (see Results).

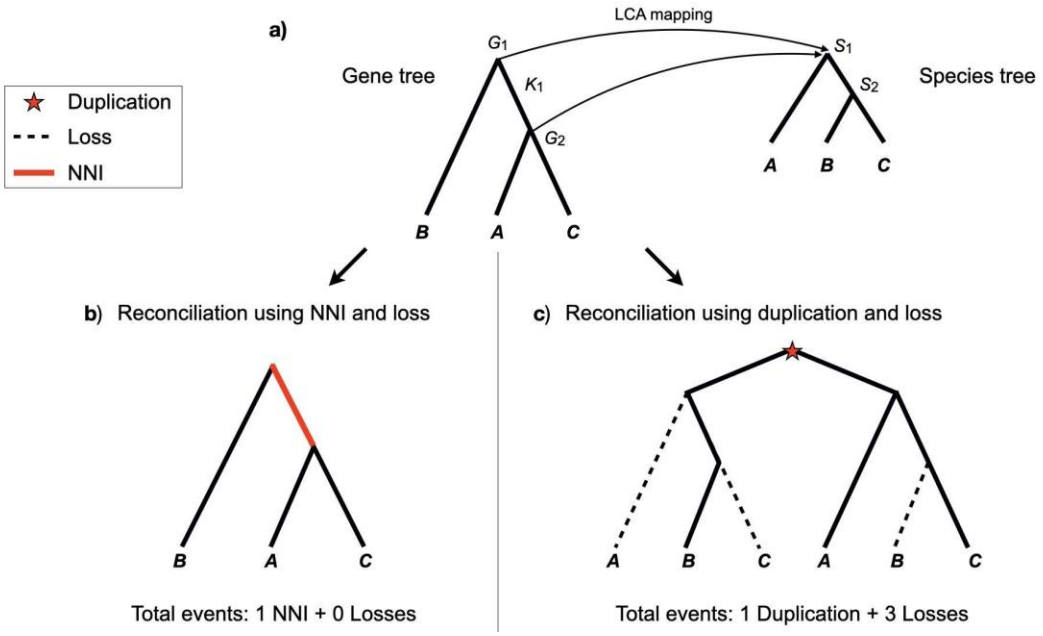

Figure 2. Summary of `reconcILS`*. a) The initial step generates an LCA mapping function between the gene tree and the species tree. Here, nodes G_1 and G_2 in the gene tree both map to node S_1 in the species tree; we therefore refer to S_1 as a multiply mapped node. Additionally, G_1 and G_2 together 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 the last common ancestor (LCA) algorithm (Zmasek and Eddy 2001) to create a mapping function between each node in a gene tree, G, and nodes in a species tree, S (Figure 2a). We refer to each of these pairwise relationships (arrows in Figure 2a) as "mappings" from here on.

Unlike in standard DL methods, the LCA mapping here does not represent a final reconciliation. Instead, this step identifies cases in which multiple nodes in the gene tree map to an identical node in 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 (given specified costs), 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 tree are inferred via standard DL methods; inferences of losses are explained in the Methods.

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 $C_I = 1$.

Labeling a reconciled gene tree

A full reconciliation of a gene tree, G, and species tree, S, will have reconciled all 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 D (for duplication) or S (for speciation); they may also label the location of losses, L. 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.

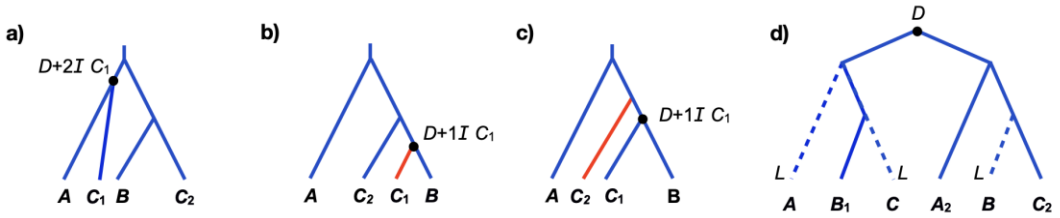

Figure 3. Labeling reconciled gene trees. *a) One possible reconciliation of the gene tree from Figure 1f labeled by* `reconcILS`*. The single ILS node (I) has been labeled with a branch*

367 *that has been rearranged by NNI (C_1), the number of NNIs it has undergone (2), and the fact*
368 *that it also represents a duplication (D). b) The gene tree from Figure 1d labeled by*
369 *`reconcILS`. This tree is correctly labeled, but note that it is the same topology as in panel c*
370 *(see Methods for further explanation). c) The gene tree from Figure 1e labeled by*
371 *`reconcILS`. Note that the program has labeled the bottom-most node flanking the*
372 *discordant branch with D and I, although that is not the true history. d) A gene tree with only*
373 *duplications and losses (L) labeled by `reconcILS`. Internal branches on a tree can also be*
374 *labeled with L, though none are shown here.*

375

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

387 `reconcILS` also labels duplications and losses in the absence of ILS. Figure 3d
388 shows how both individual duplications and losses are labelled: observe that neither
389 requires a label with a specified number of events (like the I label), nor does either need to
390 specify on which branch they act. `reconcILS` places the label L on any node in the implied
391 gene tree that is not represented by a gene copy; for example, generic tip nodes A, C, and B
392 in Figure 3d do not exist in the input gene tree itself, but are labelled with an L.
393 `reconcILS` can do this because every duplication event is represented internally by
394 generating a new copy of the species tree below the duplicated node—if losses occur on
395 these implied trees, we can track their timing as well. Note that internal nodes can also be

396 labelled with an L, so that the timing of loss events must be inferred by comparing the state
397 of parent and daughter nodes (i.e. losses on internal branches will lead the descendant
398 internal node and all of its daughter nodes to be labelled with an L). Finally, any node
399 without a label can be inferred to be a speciation node.

400 Results

401 Accuracy and advantages of `dupcoal`

402 We assessed the accuracy of `dupcoal` by first simulating 1000 gene trees along a
403 40-tip species tree with no duplication and no loss (Supplementary Figure S6). These
404 simulations therefore only model ILS as a cause of discordance. We compared the degree of
405 discordance in these simulations to ones carried out using the same species tree with
406 SimPhy (after appropriate branch-length scaling; see Methods). We found a very high
407 correlation between concordance factors on each branch calculated from gene trees
408 produced by the two methods (ρ=0.99, P=3.53 x 10^{-31}), indicating that we are accurately
409 simulating ILS. We also carried out simulations in `dupcoal` with both duplication and
410 loss, again comparing our results to SimPhy (after scaling duplication and loss rates
411 appropriately; see Methods). Again, the average number of duplications and losses per
412 simulated gene tree (n=1000) were highly similar and not significantly different between
413 the two methods: 3.82 vs. 3.84 duplications and 3.08 vs. 3.12 losses for `dupcoal` and
414 SimPhy, respectively (P=0.88 for duplications and P=0.64 for losses; t-tests), indicating that
415 we are accurately simulating duplications and losses in `dupcoal`.

416 To highlight additional features contained within `dupcoal`, we also tracked the
417 number of simulated gene trees in which copy-number hemiplasy occurred (this is a
418 standard output reported by `dupcoal`). Out of 1000 gene trees simulated on the 40-tip
419 tree with duplication, loss, and ILS, 511 contained at least one (strict) hemiplasy event. As a
420 straightforward way to demonstrate the effect of such events, we separately inferred the
421 number of duplications and losses on gene trees with and without copy-number hemiplasy
422 using ETE3 (Huerta-Cepas et al. 2016). While ETE3 implements a DL reconciliation

423 algorithm that does not model any type of ILS (see next section), this contrast shows the
424 effects of hemiplasy above and beyond the effects of ILS. As expected, both the number of
425 inferred duplications (2.61 excess duplications; P=6.44 x 10^{-11}, t-test) and the number of
426 inferred losses (8.31 excess losses; P=1.59 x 10^{-12}, t-test) were higher among the trees
427 containing hemiplasy. These results emphasize the importance of having a simulation
428 engine that can simulate hemiplasy.

429 **Accuracy and advantages of `reconcILS`**

430 As a simple test of DL reconciliation algorithms—which do not allow for ILS—we
431 used the DL method implemented in ETE3 (Huerta-Cepas et al. 2016). We initially ran
432 simulations in `dupcoal` without duplication or loss (i.e. ILS only); this allows us to see
433 incorrectly inferred duplications and losses more easily. As expected, on a species tree with
434 only 3 tips, ETE3 always inferred 1 extra duplication and 3 extra losses on gene trees with
435 discordance due to ILS (not shown). On a larger species tree with 40 tips, ETE3 inferred a
436 large excess of duplications and losses when only ILS is acting (9.5 duplications and 34.4
437 losses per gene tree, on average). On datasets simulated on a 40-tip tree with duplication,
438 loss, and ILS, the accuracy of ETE3 is low for losses (ρ=0.23) but is still fairly good for
439 duplications (ρ=0.79) (Supplementary Figure S7).

440 In contrast, `reconcILS` accurately infers the numbers of duplications and losses
441 (Figure 4a and 4b), as well as ILS events (Figure 4c) on the 40-tip tree. The correlation
442 between inferred and simulated duplications and losses were high (ρ=0.95 and 0.63,
443 respectively). We also observed a high correlation between the inferred and simulated
444 number of NNI events needed to reconcile a gene tree (ρ=0.67). ILS immediately followed
445 by loss will not leave a signal in the resulting gene trees, leading to a slight undercounting
446 of NNI events caused by ILS. It should also be noted that all inferences here are carried out
447 with the true gene trees, as we have not included any (error-prone) steps to infer the trees
448 from sequence data. Later, we relax this assumption.

449 We also examined results from `reconcILS` when using different costs (the
450 numbers in the previous paragraph all used the default costs: $C_D = 1.1$, $C_L = 1$, and $C_I = 1$).

When increasing the cost of duplication and loss ($C_D = 4$, $C_L = 3$, and $C_I = 1$), the results remained largely the same, with high correlations between inferred and true duplication, loss, and NNI: ρ=0.92, 0.61, and 0.69, respectively. The same is true when using the default costs from DLCpar ($C_D = 2$, $C_L = 2$, and $C_I = 1$): all correlations remain high (ρ=0.98, 0.64, and 0.68, respectively). However, when increasing the cost of ILS ($C_D = 1$, $C_L = 1$, and $C_I = 5$), the results are much worse, closely resembling those from ETE3: ρ=0.79, 0.23, and 0.25, respectively. This match between `reconcILS` and ETE3 should not be surprising, as having a high cost of ILS is similar to ignoring ILS.

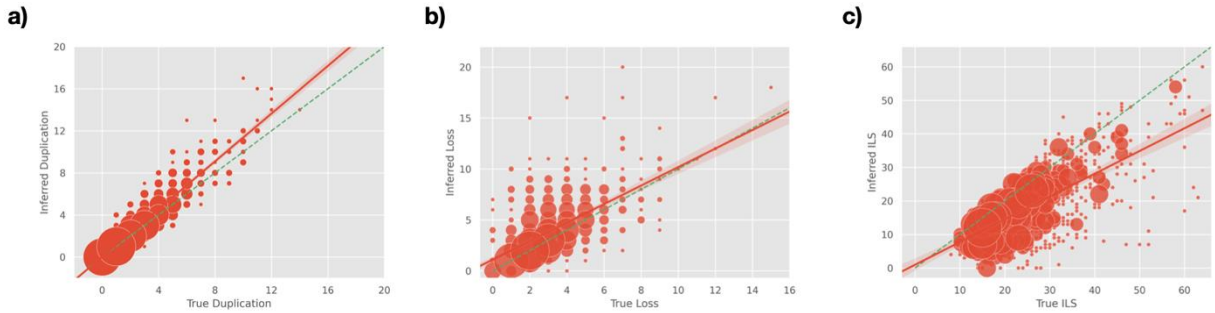

Figure 4. Accuracy of `reconcILS` on data simulated from the 40-tip tree. *a) Number of inferred versus true duplication events. b) Number of inferred versus true loss events. c) Number of inferred versus true ILS events. Circle size is proportional to the number of reconciliations with each value, with 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 S7 for results from ETE3 on the same data.*

As we did with ETE3 in the previous section, we also contrasted the number of duplications and losses inferred by `reconcILS` on simulated gene trees with and without copy-number hemiplasy (using default costs). Unfortunately, `reconcILS` also does not deal perfectly with strict hemiplasy, though in a slightly different way than ETE3. Rather than overestimating both the number of duplications and losses, `reconcILS` appears to generally infer a true ILS event as a loss event instead. This occurs because (for instance) a duplication on a discordant daughter gene tree branch will duplicate a non-sister pair of lineages (e.g. *A* and *B*), which `reconcILS` interprets as a duplication and a loss (of *C*, in this case) rather than as a duplication and an NNI event (Supplementary Figure S5).

475 Consistent with this, we observed no excess duplications in the simulations with hemiplasy
476 (0 extra duplications, on average), but significantly more losses (1.36; P=2.2 x 10^{-25}, t-test)
477 and significantly fewer NNI events (3.63; P=4.1 x 10^{-13}, t-test) compared to simulations that
478 did not have hemiplasy.

479 For completeness, we compared `reconcILS` to ETE3 under two additional
480 conditions. First, we used the same 40-tip tree topology as above to simulate gene trees
481 with duplication and loss, but very little ILS (Supplementary Table 1). Under these
482 conditions `reconcILS` remained highly accurate for inferences of duplication and loss
483 (ρ=0.99 and 0.85, respectively), though the inferences of very rare ILS events (much fewer
484 than 1 per tree) meant that the accuracy of these inferences was low (ρ=0.08). As expected
485 under these conditions, ETE3 was highly accurate in its inferences of duplications and
486 losses (ρ=0.99 and 0.94, respectively), with even greater accuracy than `reconcILS` for
487 losses. Second, we added error to the trees generated by the original 40-tip tree with
488 duplication, loss, and ILS (Methods). At the lowest level of error (mean of 5 errors per tree),
489 both `reconcILS` and ETE3 show slight reductions in accuracy (Supplementary Table 1).
490 As more error is added, both methods continue to become less accurate. At extreme levels
491 of error (mean of 15 errors per tree) `reconcILS` and ETE3 have almost equivalent
492 accuracy, with only a slight advantage in inferring loss for `reconcILS` (Supplementary
493 Table 1).

494 We also compared the results from `reconcILS` with another reconciliation method
495 that allows for ILS, DLCpar (Wu et al. 2014). DLCpar was written in Python 2, and—
496 possibly because we had to run it in a virtual environment—could not run on the larger
497 gene trees produced by the 40-tip tree. Therefore, we compared the performance of
498 `reconcILS` and DLCpar on 1000 gene trees simulated on both a 3-tip and a 16-tip species
499 tree (we still had to remove all gene trees with 25 or more tips, as these would not run in
500 DLCpar). On the 3-tip tree, the accuracy of both methods was high for duplication
501 (`reconcILS`, ρ=0.97; DLCpar, ρ=0.97; Supplementary Figure S8a). The correlation
502 between simulated and inferred ILS events was also high for both (`reconcILS`, ρ=0.85;
503 DLCpar, ρ=0.76; Supplementary Figure S8b). All results were highly similar on the 16-tip

504 tree, using the default costs from both programs (Supplementary Table 2). It is important
505 to note that DLCpar infers the number of ILS events, rather than the number of NNI
506 rearrangements that occur within such an event (as in `reconcILS`). To ensure that a fair
507 comparison can be made, results reported for DLCpar are simulated vs. inferred ILS events,
508 rather than NNI events. In both cases the methods are doing well in their respective
509 inferences, but with slightly higher accuracy for `reconcILS`.

510 There is quite a difference between `reconcILS` and DLCpar in the inference of loss
511 on the 3-tip tree (correlations of ρ=0.84 and ρ=0.48, respectively; Supplementary Figure
512 S8c). Upon closer inspection, the lower correlation in DLCpar is due to many loss events of
513 one particular type not being counted, which are much more common on the 3-tip tree;
514 indeed, on the 16-tip tree the correlations are highly similar (Supplementary Table 2).
515 Specifically, DLCpar does not count losses that occur above the root of the gene tree being
516 considered. This means that, for instance, in gene trees with the topology (B,C), DLCpar
517 does not count the loss that occurred along the A branch (assuming the species tree in
518 Figure 1). We confirmed this behavior on larger trees in DLCpar, as well as in all trees in
519 ETE3—no losses above the root of a gene tree are considered. The difference in counting of
520 losses among algorithms arises because `reconcILS` explicitly traverses the species tree,
521 while the other methods do not. This enables our method to track losses that occur above
522 the root of the input gene tree, but below the root of the input species tree. However, there
523 may be some cases in which the behavior of DLCpar and ETE3 is preferred. For instance,
524 since gene families must originate on some branch of the species tree (e.g. Richter et al.
525 2018), forcing reconciliation algorithms to infer losses on branches that pre-date this
526 appearance would be incorrect. It is not clear which approach to counting losses one
527 should prefer when analyzing real data.

528 Finally, we assessed the runtime and memory requirements of `reconcILS`
529 compared to ETE3 on a single node of the Indiana University Research Desktop, running an
530 Intel Xeon processor with 100 GB of RAM (DLCpar can only be run in Python 2, which
531 makes comparisons highly unfair). On a 3-species tree, both programs are extremely fast
532 (average time of 0.0086 seconds per gene tree for reconcILS and 0.013 seconds per gene

533 tree for ETE3; Supplementary Figure S9a) and both use very little memory (average peak
534 memory per gene tree of 0.075 MB for `reconcILS` and 0.054 MB for ETE3;
535 Supplementary Figure S9b). Both programs remain very fast on data produced from a 40-
536 species tree, though the faster speed of ETE3 is also evident (4.9 seconds per gene tree for
537 `reconcILS` and 1.3 seconds for ETE3; Supplementary Figure S9a). Regardless, this means
538 that the entire 1000-gene-tree simulated dataset on the 40-species tree took less than 84
539 minutes to analyze using `reconcILS` on a moderately powerful computer. Similarly, the
540 memory requirements of `reconcILS` increase on the 40-tip tree (18.79 MB per gene tree
541 for `reconcILS` and 3.19 MB for ETE3; Supplementary Figure S9a), but are not onerous for
542 modern computers. As the 40-species tree can contain many more events (duplication, loss,
543 or ILS), we also plotted the average time and memory required for trees containing
544 different numbers of total events (Supplementary Figure S9c,d). Both programs show
545 approximately linear increases in time and memory with the number of events, which is
546 promising for analyzing larger datasets with larger species trees, though we have not
547 tested it on much larger trees.

548 **Analysis of primate genomes**

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

561 ***Single-copy ortholog dataset***

562 Results obtained from analysis of only the single-copy orthologs are shown
563 graphically in Figure 5a (see Supplementary Figure S10a for numerical values on each
564 branch). While this is a somewhat contrived example—a set of only single-copy genes
565 would not be analyzed on its own using reconciliation methods—we think the comparison
566 with the full dataset is highly informative. For instance, the pattern that stands out from
567 this graph is that ETE3 is vastly overestimating the number of duplications and, especially,
568 the number of losses. Because ETE3 does not model ILS, all such events must be reconciled
569 by one duplication (placed above the discordant branch) and three losses (placed below
570 the discordant branch; Hahn 2007). We can see this effect in the data by looking at the
571 correlation between the gene concordance factor (gCF) estimated for a branch with the
572 number of duplicates inferred on its parent branch. Gene concordance factors quantify the
573 fraction of loci that contain the same branch as the species tree; when single-copy genes
574 are analyzed, low gCFs can indicate discordance due to ILS or other factors, like gene tree
575 estimation error (Baum 2007; Lanfear and Hahn 2024). The correlation between gCFs for
576 these data (estimated in Smith et al. 2022) and duplications inferred by ETE3 on parent
577 branches is ρ=-0.56. Lower gCFs lead to higher numbers of duplicates, as expected when
578 ILS is not accounted for—this problem is not unique to ETE3, but should be expected for all
579 standard DL and DTL approaches. Such errors can lead to incorrect inferences of bursts of
580 duplication and loss.

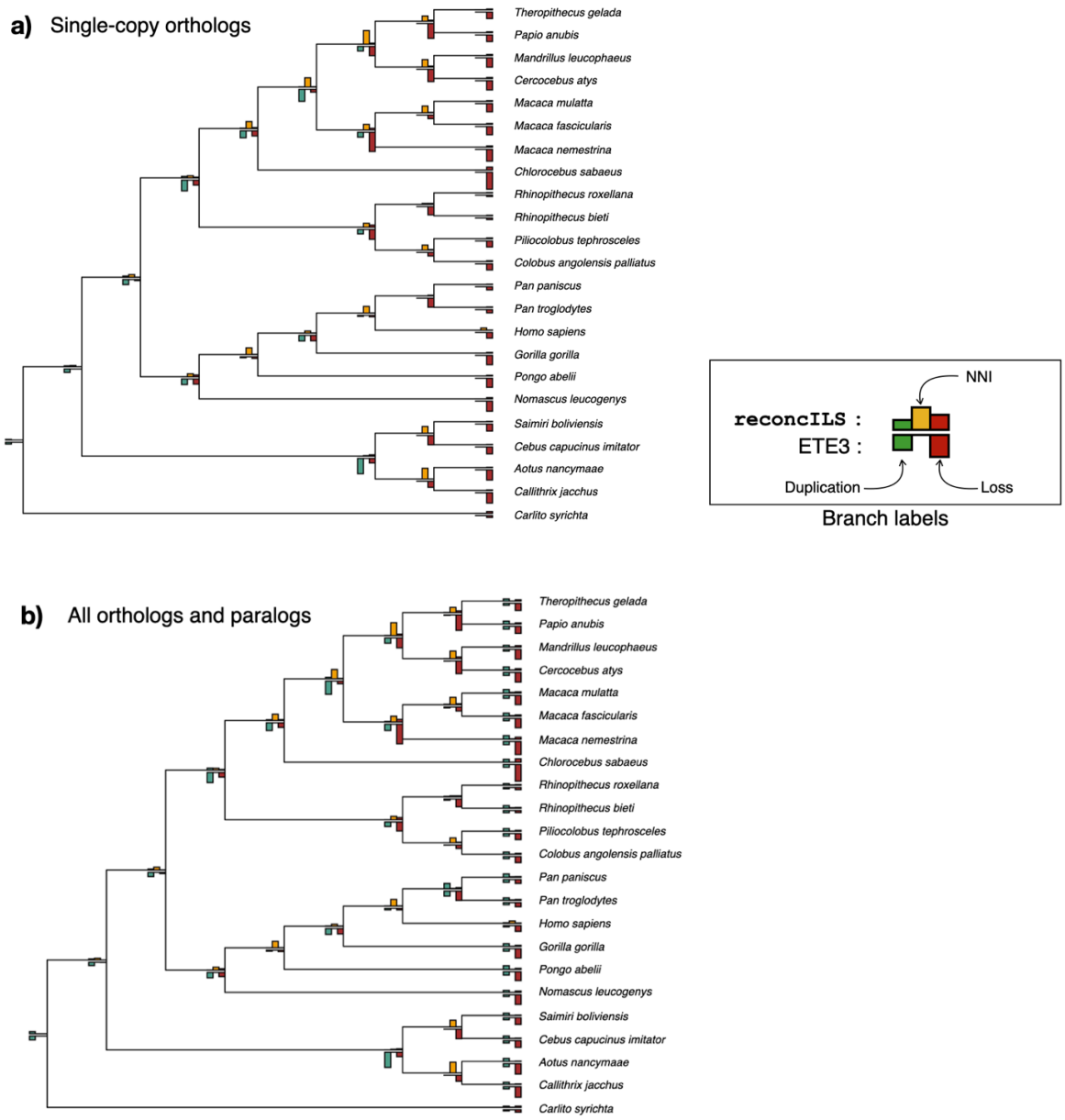

Figure 5. Results from primates. *a) Analysis of 1,820 single-copy orthologs. Both* `reconcILS` *and ETE3 were used to analyze the data, with duplication, NNI, and loss events mapped to each branch of the species tree. On each branch, the bar plots show results from* `reconcILS` *on top, with results 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*

590 *compared among branches, but the heights of bars between the trees cannot be compared.*
591 *Supplementary Figure S10 shows the absolute counts of all events from both panels and*
592 *Supplementary Figure S11 shows species tree with branch lengths in substitution units rather*
593 *than as a cladogram, as shown here. Trees were drawn with ggtree (Yu et al. 2017).*

594

595 Using the same reasoning, we can assess the accuracy of `reconcILS` by examining
596 the correlation between NNI events inferred by our algorithm and gCF values for single-
597 copy genes. High NNI counts indicate more ILS, and therefore this measure should be
598 negatively correlated with gCF values for the same branch. Indeed, the correlation between
599 NNI counts and gCF values is ρ=-0.83, demonstrating the high accuracy of `reconcILS` in
600 quantifying ILS (or in correcting for gene tree error). However, `reconcILS` does
601 sometimes infer a small number of duplications (Figure 5a; Supplementary Figure S10a).
602 While it is possible that some of the single-copy genes we analyzed represent so-called
603 pseudoorthologs—and therefore do contain actual duplication events—these are expected
604 to be rare (Smith and Hahn 2022). Instead, these false duplicates are likely due to ILS
605 among four or more lineages. Recall that, although our algorithm has been developed to
606 account for rearrangements due to the coalescent process, ILS among four lineages can
607 result in gene trees that are three NNI moves away from the species tree; ILS among more
608 lineages will produce trees that are even further away. Therefore, multiple NNI events may
609 sometimes simply be equally or more costly than duplication and loss, even when ILS is the
610 true cause of discordance (`reconcILS` will report the duplication+loss solution to a
611 reconciliation when there are equally optimal reconciliations).

612 **Full gene tree dataset**

613 We also examined patterns of ILS, duplication, and loss for the entire primate
614 dataset. The results from the full set of gene trees (Figure 5b; Supplementary Figure S10b)
615 largely mirror the results from the single-copy orthologs (which are a subset of the entire
616 dataset). Part of the apparent similarity is simply due to normalizing the two datasets
617 separately, but several non-graphical factors also contribute to the pattern. First, although
618 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 >6,000 losses on each of the tip-branches in this clade, and more than 10,000 losses on most of the internal branches (Figure 5b; Supplementary Figure S10b). 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 coalescence, 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 (Supplementary Figure S10b), 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 are more reliably inferred with `reconcILS`.

645 ## Discussion

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

655 Despite our demonstration of good performance in practice, our method is heuristic:
656 we have no theoretical proof that our method finds most parsimonious reconciliations.
657 While such proofs often require strict assumptions—e.g. the proofs for DL algorithms
658 assume there is no ILS—they are very helpful in providing confidence under the specified
659 conditions. However, given the high accuracy of `reconcILS` demonstrated with both
660 simulations and empirical data from single-copy genes, we hope the heuristic nature of our
661 algorithm is not a barrier to its utility. Future theoretical work on this or similar
662 approaches using NNIs (e.g. Górecki and Eulenstein 2011) would of course be of benefit to
663 the field. Under any such model, it must also be recognized that errors in gene tree
664 reconstruction will have a detrimental effect on the results (e.g. Supplementary Table 1).
665 Both theoretical work and new algorithms must take such limits into account.

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

674 Nevertheless, labeling gene trees serves multiple important functions, and is often
675 interpreted as biological truth. For example, multiple methods use inferred gene tree labels
676 to identify orthologous genes (genes related by speciation events; Fitch 1970). The general
677 approach taken is to label nodes using reconciliation or similar approaches, treating any
678 node not labeled with a *D* as a speciation node; genes whose common ancestor is a
679 speciation node are subsequently defined as orthologs. Both standard DL reconciliation
680 and so-called "overlap" methods (e.g. Huerta-Cepas et al. 2007; Emms and Kelly 2019) will
681 consistently label the top-most node flanking a discordant branch with a duplication event.
682 This means that, in cases like the one shown in Figure 3b, these methods will label the
683 wrong node with a *D*, and subsequently identify the wrong pair of genes as orthologs.
684 (Figure 3c would be labeled correctly; recall that this outcome has an equal probability to
685 the one in Figure 3b in the scenario considered here.) Cases such as these have been
686 referred to as "unsorted homologs" (Mallo et al. 2014), and will occur any time duplication
687 and coalescence interact, either before or after a speciation event. They can also occur even
688 if only two species are analyzed (e.g. Supplementary Figures S3 and S4), which means that
689 approaches that simply collapse short branches to avoid ILS will also be affected. Although
690 `reconcILS` currently only labels one node with a *D*, in the future we could allow for
691 multiple possible labelings in order to emphasize the inherent ambiguity of some labels.

692 The algorithm used by `reconcILS` deals, in part, with an additional interaction
693 between duplication, loss, and coalescence: hemiplasy. Hemiplasy is the phenomenon by
694 which a single mutational event on a discordant gene tree will be interpreted as multiple
695 events ("homoplasy") when reconciled to the species tree (Avise and Robinson 2008; Hahn
696 and Nakhleh 2016). In analyses of nucleotide substitutions, generic binary traits, and even
697 continuous traits, hemiplasy requires that mutations have occurred on a discordant branch
698 of a gene tree (Mendes and Hahn 2016; Guerrero and Hahn 2018; Hibbins et al. 2023). In
699 the context of either duplication or loss, such mutations would have to occur on a
700 discordant branch of a daughter gene tree to lead to copy-number hemiplasy. The
701 simulation engine introduced here, `dupcoal`, simulates hemiplasy events and reports any
702 gene trees affected. Our results show that these events lead to increased error in
703 reconciliation, and are therefore an important process to include. Interestingly, such events

704 have little effect on species tree inference when paralogs are used for this purpose (Li et al.
705 2024).

706 Unfortunately, strict copy-number hemiplasy events as defined above are still not
707 accurately inferred by `reconcILS`: we have found that our algorithm consistently turns
708 NNIs into losses when hemiplasy occurs. While the total number of events does not seem to
709 increase, it is not ideal that NNIs are mistaken for losses. However, `reconcILS` does deal
710 accurately with copy-number hemiplasy under the broader definition in use (see section on
711 "Complex coalescence…"), and `dupcoal` also simulates and records such events. In
712 contrast, DLCpar assumes that this type of hemiplasy does not occur, and SimPhy does not
713 simulate this type of hemiplasy.

714 The algorithm used by `reconcILS` deals with ILS differently than does DLCpar.
715 While DLCpar treats ILS as an event, `reconcILS` considers NNIs as events induced by ILS.
716 It is not clear that either approach should necessarily be preferred on first principles. In
717 simple cases involving three species and a single discordant branch (e.g. Figure 2), ILS and
718 NNI are more or less equivalent, though it should be noted that the process of ILS can also
719 lead to concordant trees and is not synonymous with discordance. However, in cases with
720 more than one discordant branch there is no simple relationship between the two. For
721 instance, in a 4-tip tree undergoing ILS, lineages may undergo ILS across two branches, but
722 result in between 0 and 3 NNI events (Than and Rosenberg 2013). We think that directly
723 modeling NNIs—as is done in `reconcILS`—is a natural way to express the events needed
724 to reconcile a gene tree with a species tree, including discordance due to gene tree error. In
725 fact, NNI and other moves have been used previously in reconciliation algorithms to correct
726 gene trees that may have been inferred incorrectly (e.g. Chen et al. 2000; Chaudhary et al.
727 2011; Górecki and Eulenstein 2011; Nguyen et al. 2013). Though these algorithms
728 minimize DL reconciliation costs using NNIs, they could also have been used within
729 `reconcILS` for each local reconciliation step. Regardless of the exact algorithm used, the
730 utility of such methods also provides a motivation for using `reconcILS` to obtain more
731 accurate inferences of duplications and losses in the presence of estimation error, not just
732 ILS.

733 One relatively straightforward improvement that could be made to `reconcILS` is
734 in the cost of different events. Costs must be specified in `reconcILS` ahead of time, but it
735 is difficult to determine biologically realistic costs for duplication, loss, and ILS. One
736 approach to determining duplication and loss costs would be to infer the number of such
737 events using an alternative, non-reconciliation approach (e.g. Mendes et al. 2020), setting
738 the cost in `reconcILS` to be inversely related to the inferred number of events of each
739 type. Varying rates of duplication and loss can also be inferred along different branches of
740 the species tree, and so the costs of these events could vary commensurately along the tree.
741 Similarly, we can envision a simple way to scale the costs of NNI with the length of species
742 tree branches: this is because the probability of discordance due to ILS is inversely
743 proportional to internal branch lengths on a species tree (Hudson 1983). We could
744 therefore make NNI events on longer branches of the species tree more costly, which would
745 properly weight the relative probability of such events locally on the tree. Regardless of the
746 method used, researchers should ensure that biological conclusions are qualitatively and
747 quantitatively similar across a range of costs.

748 There are also several extensions that could be made to `dupcoal`, our new
749 simulation engine. Currently, `dupcoal` draws a daughter gene tree from the multispecies
750 coalescent (MSC) at random: this implicitly assumes that the parent and daughter trees are
751 completely unlinked. It is straightforward, however, to draw the daughter tree at a given
752 recombination distance from the parent tree, ranging from fully linked to fully unlinked. In
753 addition, though the software currently draws gene trees from the MSC, we could also
754 include introgression as a source of discordance by drawing trees from the multispecies
755 network coalescent (Meng and Kubatko 2009; Yu et al. 2011). In this case, users would
756 simply input a species network to simulate from, rather than a species tree. We should also
757 note the differences between the current version of `dupcoal` and alternative simulators.
758 As mentioned previously, SimPhy (Mallo et al. 2016) does not simulate hemiplasy of any
759 kind, as it uses the same three-tree model of Rasmussen and Kellis (2012) and Wu et al.
760 (2014). However, this method does include gene transfers, which are not simulated by
761 `dupcoal`. The MLMSC-II simulator of Li et al. (2024) simulates hemiplasy under a similar
762 model to the one used here; `dupcoal` is therefore most similar to this method, with a few

differences (see Appendix). Despite these minor differences, both methods are capable of generating a wide variety of gene trees impacted by duplication, loss, coalescence, and the interactions of these processes. Unlike MLMSC-II, `dupcoal` records the exact counts of several different event types, a necessity when evaluating the performance of reconciliation approaches.

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

To test the accuracy of `dupcoal` we simulated gene trees along a species tree with 40 tips (Supplementary Figure S6). The species tree topology was drawn at random using the program ape (Paradis and Schliep 2018), with the "Grafen" method for drawing branch lengths (scaled up by 10); this tree was used as input to both `dupcoal` and SimPhy. The

790 amount of ILS in this tree, as measured by gene concordance factors (Minh et al. 2020a),
791 ranged from 29.1 to 99.6; in comparison, the values estimated from the primate tree in
792 Vanderpool et al. (2020) ranged from 27 to 96, so the two are quite similar. Because
793 `dupcoal` specifies branch lengths in coalescent units, while SimPhy uses time, branch
794 lengths were converted by assuming a population size of N=10 and a generation time of 1
795 year. Simulations were run on this tree with only ILS (setting duplication and loss rates to
796 $\lambda = 0$ and $\mu = 0$, respectively). When using non-zero duplication and loss rates (see next
797 paragraph), they were scaled to the appropriate branch lengths to make them comparable
798 in `dupcoal` and SimPhy. Comparisons between `dupcoal` and SimPhy with no duplication
799 and loss were carried out by calculating gene concordance factors (gCFs) on the species
800 tree using IQ-TREE2 (Minh et al. 2020a).

801 To test the accuracy of inferences using `reconcILS` and ETE3, the same 40-taxon
802 tree was used for simulations in `dupcoal` with ILS, duplication, and loss ($\lambda = 0.03$ and $\mu =$
803 0.03). These parameter values were chosen so that the simulations produced a large
804 number of duplications and losses to assess accuracy with, not because they are
805 biologically motivated. To add error, we specified three levels of gene tree errors, with a
806 Poisson number of errors drawn for each tree. The three levels corresponded to a mean of
807 5, 10, or 15 errors per gene tree; all errors were modeled as an NNI rearrangement.
808 Changes on each branch of each tree were drawn independently and sequentially, such that
809 the same branch could be rearranged multiple times. We also carried out simulations on
810 the 40-tip tree with very low levels of ILS by scaling all branch lengths up by 1000; to avoid
811 extremely large gene trees under this scenario, we also scaled the duplication and loss
812 rates down by 10.

813 DLCpar is written in Python 2 and could not be run in reasonable time on the 40-
814 taxon tree. We therefore also simulated data on a 3-taxon tree with topology
815 (A:1.5,(B:0.5,C:0.5):1), where branch lengths are measured in coalescent units and
816 duplication and loss rates were the same as on the larger tree. We also simulated data on a
817 16-taxon tree (based on the fungal tree in Butler et al. 2009), though DLCpar was unable to
818 analyze any gene tree with 25 or more tips. All reported results are therefore only from
819 those trees that could be analyzed by both methods. The default costs ($C_D = 1.1$, $C_L = 1$,

820 $C_I = 1$) were used for all analyses of simulated data for all methods, except where noted.
821 Note that it is the relative value of costs that matters for analysis, and the absolute values
822 do not correspond to any biological parameter.

823 **Implementation of `reconcILS`**

824 ***Inputs and outputs***

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

829 ***Traversing trees***

830 After generating an initial LCA mapping between a gene tree and the species tree,
831 `reconcILS` performs a preorder traversal of the species tree. For each species tree node,
832 the algorithm checks how many gene tree nodes are mapped to it. If there are fewer than 2
833 mappings, the algorithm asks whether speciation or loss can be inferred. As is standard in
834 reconciliation, species tree nodes mapped to one gene tree node imply a speciation event at
835 that gene tree node—these do not incur a cost. Nodes mapped to zero gene tree nodes
836 imply at least one loss in the gene tree. Our program asks whether the parent node of a
837 node with 0 mappings itself has 0 mappings. If not, a loss is inferred; if so, no event is
838 recorded, because the loss must have occurred further up the tree. If there are 2 or more
839 mappings the species tree node is multiply mapped, and a function to carry out "local
840 reconciliation" is called. Local reconciliation is described in more detail in the next section,
841 but here can simply be described as determining the events that have generated the
842 multiple mappings to a single species tree node. It should also be noted that if a duplication
843 node is inferred for a gene tree, both sub-trees induced by the duplication are further
844 locally reconciled (Algorithm 1); this means that `reconcILS` can detect cases in which
845 multiple independent losses have occurred along the same lineage (e.g. Supplementary
846 Figure 1).

847 If there are only two adjacent nodes in the gene tree mapped to a node in the species
848 tree (i.e. one discordant branch, as in Figure 2), then the method proceeds as described
849 below. However, when there are more than two nodes mapped to an internal species tree
850 node—and therefore more than one discordant branch in the gene tree—we must decide
851 which branch to reconcile with the multiply mapped node first. We have found cases where
852 the order in which the branches are resolved matters, with only one of the choices leading
853 to an optimal reconciliation. Algorithm 2 (see Supplementary Materials) describes an
854 algorithm for choosing which discordant branch to locally reconcile first. This algorithm
855 calculates the "bipartition cost" of each possible branch to reconcile. We calculate the
856 bipartition cost of a branch by examining the two trees induced by an NNI move applied to
857 the branch. For each of these new topologies, we calculate costs using two measures, both
858 of which are applied independently to the two sub-trees extending from the root of these
859 topologies (denoted by the red vertical line in Supplementary Figure 12b,c). First, we count
860 the number of bipartitions present in these sub-trees but not present in the species tree
861 ("bipartition" cost in Supplementary Figure 12); second, we count the number of lineages
862 missing relative to the node in the species tree the root of the topology is mapped to ("lost"
863 cost in Supplementary Figure 12). The bipartition cost of a single branch therefore requires
864 eight values to be calculated, four for each of the trees created by NNI (e.g. Supplementary
865 Figure 12b). Such calculations are simple and can be carried out quickly, and by choosing
866 the branch with the lowest bipartition cost we can more accurately proceed with
867 reconciliation.

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

872 ***Choosing among events***

873 In order for our algorithm to more efficiently and optimally choose among different
874 evolutionary events (e.g. Figure 2b vs 2c), we introduce the concept of a "local
875 reconciliation" for a multiply mapped node. Here, a species tree node is locally reconciled if

876 the number of gene tree nodes mapped to it, m, is reduced. For internal nodes of the
877 species tree, this, by definition, also reduces the number of discordant branches in the gene
878 tree.

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

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

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

901 ***Labeling events***

902 Our approach for labeling ILS on reconciled gene trees is described above in the
903 "New Approaches" section. Here, we consider two important caveats to this approach, one

904 perhaps more obvious than the other. First, a specific labeling is associated with a specific
905 reconciliation. That is, if we had found a different optimal reconciliation for the tree in
906 Figure 3a (for instance, one duplication and two losses), we would have had different
907 labeled events. Therefore, a labeling is always reported in the context of a particular
908 reconciliation.

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

925 **Primate data**

926 We used two sets of gene trees that were previously constructed in Smith et al.
927 (2022) using IQ-TREE2 (Minh et al. 2020b), based on data from Vanderpool et al. (2020).
928 The first set consists of 1,820 single-copy orthologs; this set was denoted as "Single-copy
929 clusters" in Smith et al. (2022). The second set consists of 11,555 trees containing both
930 orthologs and paralogs; this set was denoted "All paralogs" in Smith et al. (2022). We
931 randomly resolved all polytomies. When using `reconcILS`, costs were set to $C_D = 1.1$,
932 $C_L = 1$, and $C_I = 1$.

Acknowledgements

We thank Yao-ban Chan for many helpful conversations about the MLMSC and MSC-DL models, Jessica Wu for answering questions about DLCpar, and Ben Fulton for tips on best programming practices. Reviewers of the manuscript also provided many 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.

946 ## References

947 Avise, J. C., and T. J. Robinson. 2008. Hemiplasy: A new term in the lexicon of phylogenetics.
948 Systematic Biology **57**:503-507.
949 Bansal, M. S., E. J. Alm, and M. Kellis. 2012. Efficient algorithms for the reconciliation
950 problem with gene duplication, horizontal transfer and loss. Bioinformatics **28**:i283-
951 i291.
952 Baum, D. A. 2007. Concordance trees, concordance factors, and the exploration of reticulate
953 genealogy. Taxon **56**:417-426.
954 Boussau, B., and C. Scornavacca. 2020. Reconciling gene trees with species trees. Pp. 3.2:1-
955 3.2:23 *in* C. Scornavacca, F. Delsuc, and N. Galtier, eds. Phylogenetics in the Genomic
956 Era. No commercial publisher | Authors open access book.
957 Butler G, Rasmussen MD, Lin MF, Santos MAS, Sakthikumar S, Munro CA, Rheinbay E,
958 Grabherr M, Forche A, Reedy JL, et al. 2009. Evolution of pathogenicity and sexual
959 reproduction in eight *Candida* genomes. Nature 459:657-662.
960 Chan, Y.-b., V. Ranwez, and C. Scornavacca. 2017. Inferring incomplete lineage sorting,
961 duplications, transfers and losses with reconciliations. Journal of Theoretical Biology
962 **432**:1-13.
963 Chaudhary, R., J. G. Burleigh, and O. Eulenstein. 2012. Efficient error correction algorithms
964 for gene tree reconciliation based on duplication, duplication and loss, and deep
965 coalescence. BMC Bioinformatics **13**:S11.
966 Chen, K., D. Durand, and M. Farach-Colton. 2000. NOTUNG: A program for dating gene
967 duplications and optimizing gene family trees. Journal of Computational Biology
968 **7**:429-447.
969 Christensen, S., E. K. Molloy, P. Vachaspati, A. Yammanuru, and T. Warnow. 2020. Non-
970 parametric correction of estimated gene trees using TRACTION. Algorithms for
971 Molecular Biology **15**:1.
972 Du, H., Y. S. Ong, M. Knittel, R. Mawhorter, N. Liu, G. Gross, R. Tojo, R. Libeskind-Hadas, and
973 Y.-C. Wu. 2021. Multiple optimal reconciliations under the duplication-loss-
974 coalescence model. IEEE/ACM Transactions on Computational Biology and
975 Bioinformatics **18**:2144-2156.

976 Duchemin, W., G. Gence, A.-M. Arigon Chifolleau et al. 2018. RecPhyloXML: a format for
977 reconciled gene trees. Bioinformatics **34**:3646-3652.
978 Durand, D., B. V. Halldórsson, and B. Vernot. 2006. A hybrid micro-macroevolutionary
979 approach to gene tree reconstruction. Journal of Computational Biology **13**:320-335.
980 El-Mabrouk, N., and A. Ouangraoua. 2017. A general framework for gene tree correction
981 based on duplication-loss reconciliation. Pp. 8:1-8:14. WABI.
982 Emms, D. M., and S. Kelly. 2019. OrthoFinder: phylogenetic orthology inference for
983 comparative genomics. Genome Biology **20**:238.
984 Fitch, W. M. 1970. Distinguishing homologous from analogous proteins. Systematic Zoology
985 **19**:99-113.
986 Gillespie, J. H., and C. H. Langley. 1979. Are evolutionary rates really variable? Journal of
987 Molecular Evolution **13**:27-34.
988 Górecki P, Eulenstein O. 2011. A linear time algorithm for error-corrected reconciliation of
989 unrooted gene trees. ISBRA 2011: Springer Berlin Heidelberg. p. 148-159.
990 Goodman, M., J. Czelusniak, G. W. Moore, A. E. Romero-Herrera, and G. Matsuda. 1979.
991 Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by
992 cladograms constructed from globin sequences. Systematic Zoology **28**:132-163.
993 Guerrero, R. F., and M. W. Hahn. 2018. Quantifying the risk of hemiplasy in phylogenetic
994 inference. Proceedings of the National Academy of Sciences **115**:12787-12792.
995 Guigo, R., I. Muchnik, and T. F. Smith. 1996. Reconstruction of ancient molecular phylogeny.
996 Molecular Phylogenetics and Evolution **6**:189-213.
997 Hahn, M. W. 2007. Bias in phylogenetic tree reconciliation methods: implications for
998 vertebrate genome evolution. Genome Biology **8**:R141.
999 Hahn, M. W., and L. Nakhleh. 2016. Irrational exuberance for resolved species trees.
1000 Evolution **70**:7-17.
1001 Hallett, M. T., and J. Lagergren. 2001. Efficient algorithms for lateral gene transfer
1002 problems. Pp. 149-156. Proceedings of the fifth annual international conference on
1003 Computational biology (RECOMB '01).
1004 Hallett, M., J. Lagergren, and A. Tofigh. 2004. Simultaneous identification of duplications
1005 and lateral transfers. Pp. 347-356. Proceedings of the eighth annual international
1006 conference on Research in computational molecular biology (RECOMB '04).

1007 Harris, C. R., K. J. Millman, S. J. Van Der Walt et al. 2020. Array programming with NumPy.
1008 Nature **585**:357-362.
1009 Hibbins, M. S., L. C. Breithaupt, and M. W. Hahn. 2023. Phylogenomic comparative methods:
1010 Accurate evolutionary inferences in the presence of gene tree discordance.
1011 Proceedings of the National Academy of Sciences **120**:e2220389120.
1012 Hudson, R. R. 1983. Testing the constant-rate neutral allele model with protein sequence
1013 data. Evolution **37**:203-217.
1014 Huerta-Cepas, J., H. Dopazo, J. Dopazo, and T. Gabaldón. 2007. The human phylome.
1015 Genome Biology **8**:R109.
1016 Huerta-Cepas, J., F. Serra, and P. Bork. 2016. ETE 3: Reconstruction, analysis, and
1017 visualization of phylogenomic data. Molecular Biology and Evolution **33**:1635-1638.
1018 Jacox, E., C. Chauve, G. J. Szöllősi, Y. Ponty, and C. Scornavacca. 2016. ecceTERA:
1019 comprehensive gene tree-species tree reconciliation using parsimony. Bioinformatics
1020 **32**:2056-2058.
1021 Kingman, J. F. C. 1982. The coalescent. Stochastic Processes and their Applications **13**:235-
1022 248.
1023 Lafond M, Chauve C, Dondi R, El-Mabrouk N. 2014. Polytomy refinement for the correction
1024 of dubious duplications in gene trees. Bioinformatics **30**:i519-i526.
1025 Lanfear R, Hahn MW. 2024. The meaning and measure of concordance factors. Molecular
1026 Biology and Evolution **41**:msae214.
1027 Li, Q., C. Scornavacca, N. Galtier, and Y.-B. Chan. 2021. The multilocus multispecies
1028 coalescent: A flexible new model of gene family evolution. Systematic Biology **70**:822-
1029 837.
1030 Li, Q., Y.-b. Chan, N. Galtier, and C. Scornavacca. 2024. The effect of copy number hemiplasy
1031 on gene family evolution. Systematic Biology **73**:355-374.
1032 Libeskind-Hadas, R., Y.-C. Wu, M. S. Bansal, and M. Kellis. 2014. Pareto-optimal phylogenetic
1033 tree reconciliation. Bioinformatics **30**:i87-i95.
1034 Maddison, W. P. 1997. Gene trees in species trees. Systematic Biology **46**:523-536.
1035 Mallo, D., L. De Oliveira Martins, and D. Posada. 2014. Unsorted homology within locus and
1036 species trees. Systematic Biology **63**:988-992.

1037 Mallo, D., L. De Oliveira Martins, and D. Posada. 2016. *SimPhy*: Phylogenomic simulation of
1038 gene, locus, and species trees. Systematic Biology **65**:334-344.
1039 Mawhorter, R., and R. Libeskind-Hadas. 2019. Hierarchical clustering of maximum
1040 parsimony reconciliations. BMC Bioinformatics **20**:612.
1041 Mendes, F. K., and M. W. Hahn. 2016. Gene tree discordance causes apparent substitution
1042 rate variation. Systematic Biology **65**:711-721.
1043 Mendes, F. K., D. Vanderpool, B. Fulton, and M. W. Hahn. 2020. CAFE 5 models variation in
1044 evolutionary rates among gene families. Bioinformatics **36**:5516-5518.
1045 Meng, C., and L. S. Kubatko. 2009. Detecting hybrid speciation in the presence of incomplete
1046 lineage sorting using gene tree incongruence: A model. Theoretical Population Biology
1047 **75**:35-45.
1048 Minh, B. Q., M. W. Hahn, and R. Lanfear. 2020a. New methods to calculate concordance
1049 factors for phylogenomic datasets. Molecular Biology and Evolution **37**:2727-2733.
1050 Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, Lanfear
1051 R. 2020b. IQ-TREE 2: New models and efficient methods for phylogenetic inference in
1052 the genomic era. Molecular Biology and Evolution **37**:1530-1534.
1053 Morel, B., A. M. Kozlov, A. Stamatakis, and G. J. Szöllősi. 2020. GeneRax: A tool for species-
1054 tree-aware maximum likelihood-based gene family tree inference under gene
1055 duplication, transfer, and loss. Molecular Biology and Evolution **37**:2763-2774.
1056 Nguyen, T. H., V. Ranwez, S. Pointet, A.-M. A. Chifolleau, J.-P. Doyon, and V. Berry. 2013.
1057 Reconciliation and local gene tree rearrangement can be of mutual profit. Algorithms
1058 for Molecular Biology **8**:12.
1059 Page, R. D. M. 1994. Maps between trees and cladistic analysis of historical associations
1060 among genes, organisms, and areas. Systematic Biology **43**:58-77.
1061 Page, R. D. M., and M. A. Charleston. 1997. From gene to organismal phylogeny: Reconciled
1062 trees and the gene tree/species tree problem. Molecular Phylogenetics and Evolution
1063 **7**:231-240.
1064 Paradis, E., and K. Schliep. 2018. ape 5.0: an environment for modern phylogenetics and
1065 evolutionary analyses in R. Bioinformatics **35**:526-528.
1066 Rasmussen, M. D., and M. Kellis. 2012. Unified modeling of gene duplication, loss, and
1067 coalescence using a locus tree. Genome Research **22**:755-765.

1068 Richter, D. J., P. Fozouni, M. B. Eisen, and N. King. 2018. Gene family innovation,
1069 conservation and loss on the animal stem lineage. eLife **7**:e34226.
1070 Robinson DF. 1971. Comparison of labeled trees with valency three. Journal of
1071 Combinatorial Theory, Series B **11**:105-119.
1072 Schrider, D. R., and M. W. Hahn. 2010. Lower linkage disequilibrium at CNVs is due to both
1073 recurrent mutation and transposing duplications. Molecular Biology and Evolution
1074 **27**:103-111.
1075 Scornavacca C, Berry V, Ranwez V. 2011. Building species trees from larger parts of
1076 phylogenomic databases. Information and Computation **209**:590-605.
1077 Smith, M. L., and M. W. Hahn. 2022. The frequency and topology of pseudoorthologs.
1078 Systematic Biology **71**:649-659.
1079 Smith, M. L., D. Vanderpool, and M. W. Hahn. 2022. Using all gene families vastly expands
1080 data available for phylogenomic inference. Molecular Biology and Evolution
1081 **39**:msac112.
1082 Stolzer, M., H. Lai, M. Xu, D. Sathaye, B. Vernot, and D. Durand. 2012. Inferring duplications,
1083 losses, transfers and incomplete lineage sorting with nonbinary species trees.
1084 Bioinformatics **28**:i409-i415.
1085 Stull, G. W., X.-J. Qu, C. Parins-Fukuchi et al. 2021. Gene duplications and phylogenomic
1086 conflict underlie major pulses of phenotypic evolution in gymnosperms. Nature Plants
1087 **7**:1015-1025.
1088 Sukumaran, J., and M. T. Holder. 2010. DendroPy: a Python library for phylogenetic
1089 computing. Bioinformatics **26**:1569-1571.
1090 Szöllősi, G. J., W. Rosikiewicz, B. Boussau, E. Tannier, and V. Daubin. 2013. Efficient
1091 exploration of the space of reconciled gene trees. Systematic Biology **62**:901-912.
1092 Tajima, F. 1983. Evolutionary relationships of DNA sequences in finite populations.
1093 Genetics **105**:437-460.
1094 Than, C. V., and N. A. Rosenberg. 2013. Mathematical properties of the deep coalescence
1095 cost. IEEE/ACM Transactions on Computational Biology and Bioinformatics **10**:61-72.
1096 Thomas, G. W. C., S. H. Ather, and M. W. Hahn. 2017. Gene-tree reconciliation with MUL-
1097 trees to resolve polyploidy events. Systematic Biology **66**:1007-1018.

1098 Vanderpool, D., B. Q. Minh, R. Lanfear et al. 2020. Primate phylogenomics uncovers multiple
1099 rapid radiations and ancient interspecific introgression. PLoS Biology **18**:e3000954.
1100 Vernot, B., M. Stolzer, A. Goldman, and D. Durand. 2008. Reconciliation with non-binary
1101 species trees. Journal of Computational Biology **15**:981-1006.
1102 Waskom, M. 2021. Seaborn: Statistical data visualization. Journal of Open Source Software
1103 **6**:3021.
1104 Wu, Y.-C., M. D. Rasmussen, M. S. Bansal, and M. Kellis. 2014. Most parsimonious
1105 reconciliation in the presence of gene duplication, loss, and deep coalescence using
1106 labeled coalescent trees. Genome Research **24**:475-486.
1107 Yu, Y., C. Than, J. H. Degnan, and L. Nakhleh. 2011. Coalescent histories on phylogenetic
1108 networks and detection of hybridization despite incomplete lineage sorting.
1109 Systematic Biology **60**:138-149.
1110 Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. 2017. ggtree: an r package for visualization and
1111 annotation of phylogenetic trees with their covariates and other associated data.
1112 Methods in Ecology and Evolution **8**:28-36.
1113 Zhang, C., C. Scornavacca, E. K. Molloy, and S. Mirarab. 2020. ASTRAL-Pro: Quartet-based
1114 species-tree inference despite paralogy. Molecular Biology and Evolution **37**:3292-
1115 3307.
1116 Zmasek, C. M., and S. R. Eddy. 2001. A simple algorithm to infer gene duplication and
1117 speciation events on a gene tree. Bioinformatics **17**:821-828.
1118