MIT Libraries | DSpace@MIT

MIT Open Access Articles

Most parsimonious reconciliation in the presence of gene duplication, loss, and deep coalescence using labeled coalescent trees

Massachusetts Institute of Technology

Most parsimonious reconciliation in the presence of gene duplication, loss, and deep coalescence using labeled coalescent trees

Yi-Chieh Wu, Matthew D. Rasmussen, Mukul S. Bansal, et al.

Supplemental Material	http://genome.cshlp.org/content/suppl/2014/01/07/gr.161968.113.DC1.html
P<P	Published online December 5, 2013 in advance of the print journal.
Open Access	Freely available online through the *Genome Research* Open Access option.
Creative Commons License	This article, published in *Genome Research*, is available under a Creative Commons License (Attribution-NonCommercial 3.0 Unported), as described at http://creativecommons.org/licenses/by-nc/3.0/.
Email Alerting Service	Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**

To subscribe to *Genome Research* go to:
http://genome.cshlp.org/subscriptions

Method

Most parsimonious reconciliation in the presence of gene duplication, loss, and deep coalescence using labeled coalescent trees

Yi-Chieh Wu,[1,5] Matthew D. Rasmussen,[2,4] Mukul S. Bansal,[1,4] and Manolis Kellis[1,3,5]

[1]Department of Electrical Engineering and Computer Science, Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA; [2]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York 14853, USA; [3]Broad Institute, Cambridge, Massachusetts 02142, USA

Accurate gene tree-species tree reconciliation is fundamental to inferring the evolutionary history of a gene family. However, although it has long been appreciated that population-related effects such as incomplete lineage sorting (ILS) can dramatically affect the gene tree, many of the most popular reconciliation methods consider discordance only due to gene duplication and loss (and sometimes horizontal gene transfer). Methods that do model ILS are either highly parameterized or consider a restricted set of histories, thus limiting their applicability and accuracy. To address these challenges, we present a novel algorithm DLCpar for inferring a most parsimonious (MP) history of a gene family in the presence of duplications, losses, and ILS. Our algorithm relies on a new reconciliation structure, the labeled coalescent tree (LCT), that simultaneously describes coalescent and duplication-loss history. We show that the LCT representation enables an exhaustive and efficient search over the space of reconciliations, and, for most gene families, the least common ancestor (LCA) mapping is an optimal solution for the species mapping between the gene tree and species tree in an MP LCT. Applying our algorithm to a variety of clades, including flies, fungi, and primates, as well as to simulated phylogenies, we achieve high accuracy, comparable to sophisticated probabilistic reconciliation methods, at reduced run time and with far fewer parameters. These properties enable inferences of the complex evolution of gene families across a broad range of species and large data sets.

[Supplemental material is available for this article.]

Phylogenetic analysis forms the basis of many genomic studies (Zmasek and Eddy 2002; Hahn et al. 2005; Huerta-Cepas et al. 2007; Wapinski et al. 2007; Butler et al. 2009; Mi et al. 2013) by addressing a fundamental question of evolutionary biology: how do new genes and functions arise between species? Because evolutionary history exerts a strong influence on gene function (Ohno 1970; Lynch and Conery 2000; Koonin 2005; Peterson et al. 2009), accurate inference of gene histories is a crucial task with many important applications. For example, studies have used such histories to infer the number, age, and location of gene duplication and loss events across several genomes (Page 1994; Arvestad et al. 2004; Durand et al. 2006; Rasmussen and Kellis 2011) and to distinguish between orthologs and paralogs (Li et al. 2006; Datta et al. 2009; Vilella et al. 2009).

For a cluster of genes with detectable common ancestry, or a *gene family*, these histories are often inferred by comparing two kinds of phylogenetic trees: a *gene tree* that depicts the evolutionary relationships among the genes within the gene family, and a *species tree* that depicts the evolutionary relationships of a set of species. The gene tree can be thought of as evolving "inside" the species tree, and *reconciliation* methods infer this nesting to determine the evolutionary events that gave rise to a particular gene tree.

In eukaryotic organisms at sufficiently large evolutionary distances, discordance (topological differences) between the gene tree and species tree typically arises due to duplication and loss events (Fig. 1A). However, at smaller evolutionary distances, an evolutionary phenomenon known as deep coalescence or incomplete lineage sorting (ILS) (Maddison 1997; Wakeley 2009), in which polymorphisms survive several rapid speciations then eventually fix or go extinct in a pattern incongruent to the species tree, can have a potentially prominent effect (Fig. 1B). The simplest approach to resolving discordance is to seek a most parsimonious reconciliation (MPR) (Goodman et al. 1979; Page 1994; Maddison 1997), that is, one that minimizes the number of inferred events. It has been shown that the least common ancestor (LCA) mapping solves the MPR problem when minimizing duplications only (Górecki and Tiuryn 2006), duplications and losses (Górecki and Tiuryn 2006), and deep coalescence (Wu and Zhang 2011). However, these methods do not jointly model duplication, loss, and deep coalescence, thus limiting their applicability and accuracy.

To address this problem, more sophisticated methods accounting for incongruence due to duplication, loss, *and* deep coalescence have been proposed. For example, NOTUNG (Durand et al. 2006; Vernot et al. 2007) allows users to reconcile a gene tree with a nonbinary species tree, with the optimal reconciliation minimizing the duplication-loss cost while allowing for possible deep coalescence at unresolved nodes in the species tree. While this parsimony framework is simple, requiring only a known species tree
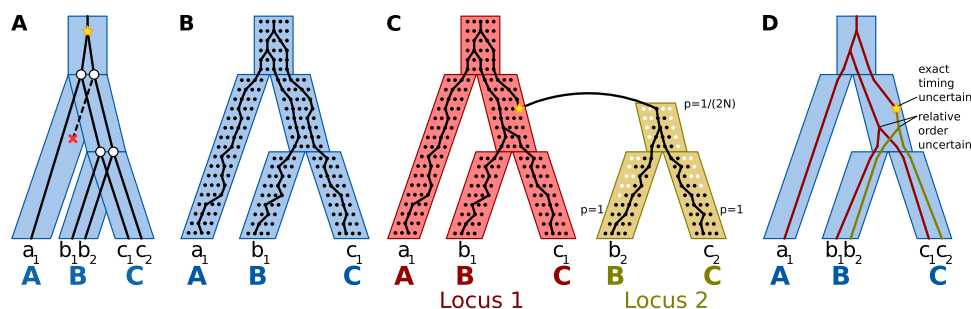
Figure 1. The three-tree model and the labeled coalescent tree. (*A*) In the duplication-loss model, incongruence between the gene tree (black) and species tree (blue) can be explained using gene duplications (yellow star) and gene losses (red "x"). (*B*) In a multispecies coalescent model, incongruence between the gene tree and species tree can be explained due to incomplete lineage sorting (ILS). Because no duplications or losses are allowed, this model is inapplicable to gene families in which multiple gene copies exist in at least one species. (*C*) The unified model proposed by Rasmussen and Kellis (2012) combines the multispecies coalescent and duplication-loss models. In this example, a duplication occurs in one chromosome [note the duplicate's frequency is initially $p = 1/(2N)$, where N is the effective population size, assuming a diploid genome] and creates a new locus, "locus 2," in the genome. At locus 2, the Wright-Fisher model dictates how the frequency p of the daughter duplicate (black dots) competes with the null allele (white dots) until it eventually fixates ($p = 1$). A gene tree is a "traceback" in this combined process. Note that the red and yellow trees form an intermediate locus tree (distinct from the gene tree and species tree) that describes how loci are created and destroyed. In this example, the gene tree has the same topology as that in *A*, but incongruence with the species tree is explained by duplication and deep coalescence. (*D*) The LCT combines the species tree, locus tree, gene tree, and reconciliations between them into a single structure. Each node of the gene tree is labeled with the species and locus to which it belongs, and gene tree nodes within the same species and locus are totally ordered in time. (Parts of this figure have been adapted with permission from Rasmussen and Kellis [2012].)

topology and (user-specified) costs for each evolutionary event, it cannot capture all possible evolutionary histories.

More recently, a three-tree model called DLCoal was proposed that introduced an intermediate *locus tree* describing how new loci are created and destroyed (Fig. 1C; Rasmussen and Kellis 2012). In the model, the reconciliation between the gene tree (or rather the *coalescent tree*) and the locus tree describes the history of coalescent events, including the presence of ILS, and the reconciliation between the locus tree and the species tree describes the history of duplication and loss events. The associated reconciliation algorithm DLCoalRecon seeks the maximum a posteriori reconciliation and shows dramatic improvement compared to ILS-unaware approaches. However, DLCoalRecon suffers from several drawbacks in practice: (1) It is highly parameterized, requiring divergence times, duplication and loss rates, population sizes, and generation times, all of which are difficult to estimate accurately; and (2) it relies on hill-climbing to search the space of locus trees and reconciliations and therefore may miss the optimal solution due to its limited search. These drawbacks limit the overall applicability, accuracy, and scalability of DLCoalRecon.

To address these shortcomings, we present a number of contributions to the reconciliation framework:

- We introduce the concept of a labeled coalescent tree (LCT), which simultaneously describes the species tree, locus tree, and coalescent tree, and the reconciliations between them (Fig. 1D). In the LCT, each node of the gene tree is labeled with both the species and locus to which it belongs, and gene tree nodes within the same species and locus are totally ordered in time. The LCT effectively combines the three-tree model into a single representation in which we can efficiently search over the space of locus trees and reconciliations.

- We present an algorithm DLCpar for inferring a most parsimonious (MP) LCT, that is, one that minimizes the total cost of inferred duplications, losses, and deep coalescence. This program is freely available at http://compbio.mit.edu/dlcpar/. As part of the development of DLCpar, we show that, assuming LCA reconciliation between the locus tree and species tree, the LCA mapping is an optimal solution for the species mapping between the gene tree and species tree in an MP LCT.

We have applied our DLCpar algorithm to both simulated and real data sets and find that it dramatically improves reconciliations compared to ILS-unaware approaches. Furthermore, its accuracy is comparable to DLCoalRecon while requiring far fewer parameters and having a faster run time.

Methods

Gene family evolution

In the reconciliation problem, we are given a gene tree and species tree (either inferred using existing phylogenetic algorithms or assumed known) and a leaf mapping that, for each extant gene, defines the extant species from which it was sampled. Note that both trees are full, rooted, and binary, and the leaf mapping indicates only the species, *not* the locus, to which each extant gene belongs. Our goal is to infer the evolutionary history of the gene family. We make the following assumptions:

1. Any incongruence between the gene tree and species tree topologies can be explained through duplication, loss, and incomplete lineage sorting. Each duplication creates a unique new locus that is unlinked with the original locus, and there is no gene conversion between duplicated loci.
2. Duplication and loss events do not undergo hemiplasy (Avise and Robinson 2008); that is, they do not fix differently in descendant species. Equivalently, all duplications and losses either always go extinct or never go extinct in all descendant lineages. This allows separation of the duplication-loss process from the coalescent process that affects gene family evolution.
3. Each extant species is represented by a single haploid sample; that is, within each gene family, multiple genes from the same extant species are sampled from multiple loci in a single individual (as opposed to being sampled from the same locus across multiple individuals).

Assumption 1 is applicable to evolution within eukaryotic species, and assumption 2 was shown to affect only a small number of gene trees in simulation with biologically realistic parameters (Rasmussen and Kellis 2012). A relaxation of assumption 3 is presented in the discussion.

To motivate the need for models of complex evolutionary histories in which duplication, loss, and ILS can occur, consider the

gene history illustrated in Figure 1C. Without duplications and losses, a multispecies coalescent process (Tajima 1983; Pamilo and Nei 1988; Rosenberg 2002; Rannala and Yang 2003; Degnan and Rosenberg 2009) would be sufficient to model the ancestry of genes a_1, b_1, and c_1. However, in this example, a duplication event occurs along the branch ancestral to species B and C, and this duplicate eventually fixes so that it is present in all descendant extant species, resulting in genes b_2 and c_2. In the *locus tree*, such a duplication appears as a bifurcation in which one child, the mother, continues to evolve at the original locus ("locus 1") and the other child, the daughter, evolves at a new locus ("locus 2"). However, molecular sequences evolve along the *gene tree*, which is affected by duplications, losses, *and* ILS. Continuing our example, the duplication has created an additional lineage within the locus 1 tree that must coalesce. Because there exist multiple chromosomes within the population, there likely exists a delay between the duplication event and the time at which the lineage with the duplicate coalesces with another lineage in locus 1. (This is analogous to the delay between a speciation event and the coalescent time of lineages from different species in a coalescent-only model.) In the gene tree, the duplication therefore appears as a locus "change" along a gene tree branch. If ILS occurs, two or more gene lineages may not coalesce at their first opportunity. In the example, despite the duplication event occurring along the branch ancestral to species B and C, the gene lineage with the duplicate does not coalesce with another lineage in locus 1 until after (looking backward in time) the speciation event that differentiated species A from B and C; this has caused incongruence between the gene tree and locus tree.

Originally proposed by Rasmussen and Kellis (2012), this unified three-tree model of gene tree-locus tree-species tree combines the multispecies coalescent and duplication-loss models for gene family evolution. Note that in addition to the gene tree, which represents the history of a set of gene sequences, and the species tree, which represents the history of a set of species, there exists a third tree, the locus tree. This locus tree shows the locus to which each sequence belonged at each point in its history and how these loci are related through duplication events. (While the reconciliation between the locus tree and species tree can additionally reveal "missing" locus tree branches due to loss events, such losses have no effect on the relationships between loci assigned to observable locus tree branches.) Importantly, under the three-tree model, only sequences within the same species *and* the same locus are allowed to coalesce. Therefore, the gene tree of this model can be viewed as a *generalized coalescent tree*, as it represents the history of gene sequences as they coalesce within the locus tree. As in the standard coalescent tree from coalescent-only models, internal nodes within the generalized coalescent tree capture coalescences; however, these coalescences can occur in a duplicate locus (for example, locus 2 in Fig. 1C), and furthermore, branches are allowed to "change" loci (for example, from locus 1 to locus 2 in Fig. 1C) due to duplication events. For simplicity, throughout the remainder of this manuscript, we will simply use *coalescent tree* to refer to the generalized coalescent tree. The problem with the term *gene tree* is that it is used in very different ways in the two underlying models: the gene branches present in one time slice in a species branch in the coalescent model (Fig. 1B) represent the chromosomes that are ancestral to the extant sequences; however, the same branches in the duplication-loss model (Fig. 1A) represent the loci that exist within the ancestral genome at that time. Therefore, the "gene tree" from the coalescent model corresponds to our coalescent tree, which evolves "inside" the locus tree, and the "gene tree" from the duplication-loss model corresponds to our locus tree, which evolves "inside" the species tree. Indeed, it is this two-step reconciliation that allows us to separate the multispecies coalescent

and duplication-loss processes responsible for gene family evolution. Finally, within the three-tree model, molecular sequences evolve along the coalescent tree; however, to be consistent with current phylogenetic literature, throughout this manuscript, we refer to a tree reconstructed using phylogenetic algorithms as a gene tree, with the understanding that this is equivalent to the coalescent tree.

The labeled coalescent tree

We now propose a new reconciliation structure, the *labeled coalescent tree* (LCT) (Fig. 1D), for modeling gene evolutionary histories. Here, we outline the basic structure of the LCT; a formal definition is provided in Supplemental Material S1. Given a gene tree, species tree, and leaf mapping, the LCT is specified by the following:

- a species map that labels each gene tree node with the species to which it belongs;
- a locus map that labels each gene tree node with the locus in which it evolves; and
- a partial order that defines a total order on gene tree nodes within the same species and locus.

The LCT implicitly models the locus tree *topology* and reconciliations between the species, locus, and gene trees; these are precisely the minimal information necessary for inferring evolutionary events. Note also that the species map effectively decomposes the gene tree into disjoint subtrees, one or more of which evolve within each species tree branch. Let a speciation node be any gene tree node that exists at a speciation event, that is, a node found at the "bottom" or "top" of a species tree branch. Some speciation nodes may be hidden in the gene tree due to gene loss or deep coalescence; we call such nodes "implied speciation nodes" and include them in the LCT.

The LCT allows for three sources of incongruence between the gene tree and species tree (Fig. 2):

- Duplication: the locus of a gene tree node differs from the locus of its parent node (Fig. 2A);
- Loss: a locus that exists within a species is not found within the set of loci at the bottom of the species tree branch (Fig. 2B). Note that we cannot infer when a locus is lost; thus, we assume that a locus is lost when it no longer exists within the species tree branch;
- Deep coalescence: deep coalescence is quantified by counting the number of implied extra lineages (Fig. 2C, top). As new lineages are created by speciation and duplication events, we must address both these sources:
  (1) ILS due to speciations: At a speciation, that is, at the top of a species tree branch, there exist multiple lineages within the same locus (Fig. 2C, middle); and
  (2) ILS due to duplications: At a duplication, there exist multiple contemporary lineages that belong to the same locus as the parent locus of the duplication (Fig. 2C, bottom).

Finally, note that the LCT is a simplified and condensed representation of the three-tree model of gene family evolution (for a comparison, see Supplemental Material S4).

Inferring a most parsimonious LCT

Using the LCT, we can now develop new methods for gene tree-species tree reconciliation. In this section, we assume that, in addition to the gene tree, species tree, and leaf mapping, we are given positive costs for duplications, losses, and deep coalescence (extra lineages). Our goal is to infer a most parsimonious (MP) LCT, that
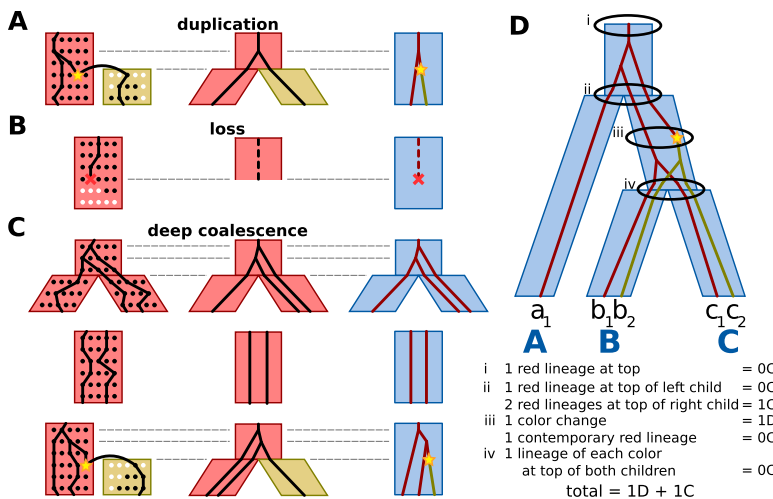
Figure 2. Duplications, losses, and deep coalescence. In *A–C*, evolutionary events are depicted using the unified model (*left*), the gene tree evolving "inside" the locus tree (*middle*), and the LCT (*right*). (*A*) Gene duplication was discussed in Figure 1. Evolution within a single species tree branch is shown. (*B*) Similar to duplication, a gene loss starts in one chromosome and drifts until it fixates or goes extinct. Evolution within a single species tree branch is shown. (*C*) Deep coalescence occurs when two lineages fail to coalesce before either can coalesce with a third lineage. This results in extra lineages at the top of a locus tree branch (*top*). Within a single species tree branch, multiple lineages can exist at the top of a locus tree branch created by a speciation (*middle*) or at the top of a locus tree branch created by a duplication (*bottom*). (*D*) Events are counted in the LCT of Figure 1D.

Our next task is to enumerate the set of species-specific locus maps for each species. We accomplish this through a pre-order traversal of the species tree. To start, we consider the (single) subtree (of the gene tree) that evolves within the root branch of the species tree. We label each gene tree branch within this subtree with a boolean variable that is true if the locus has changed along the branch and false otherwise. Next, we assign the root of the gene tree to an arbitrary locus and, for each combination of branch labels, recur down the subtree to determine a possible species-specific locus map. This yields the set of possible species-specific locus maps for the root species (Fig. 3B). For species that are not the root of the species tree, a similar approach is applied, except that multiple subtrees may exist within the species tree branch. Also, instead of assigning the root of the gene tree to an arbitrary locus, we consider all possible "root loci," that is, all possible locus assignments for the gene tree nodes at the top of the species tree branch. Similarly, "leaf loci" are the locus assignments for the gene tree nodes at the bottom of a species tree branch. Because we have performed a pre-order traversal of the species tree, the set of possible root loci for a species is equal to the set of leaf loci for its parent species, and in turn, the set of leaf loci is determined by the species-specific locus maps.

For each species-specific locus map, we must also determine an associated optimal order. One solution is to choose an order such that the implied duplications are as early in the species tree branch as possible. This is because the order only affects the number of extra lineages due to duplications, and such a selection minimizes this count (for full details, see Proposition S2.3). In essence, pushing duplications toward the top of a species tree branch effectively enforces LCA reconciliation between the gene tree and locus tree.

At this point, the reconciliation cost for each species-specific locus map (and associated optimal order) can be computed, but there is one last step to perform within the species tree branch. As previously mentioned, we need only transmit the loci at speciation nodes from one species tree branch to the next. In addition, later in our algorithm, we require a method for comparing the locus assignments at speciation nodes across different locus maps. In particular, even though there are multiple ways of arriving at the same leaf loci, we are only interested in the relative assignments, that is, which nodes are mapped to the same locus and which to different loci. Therefore, we remap the leaf loci to a set of relative loci; this is accomplished by arbitrarily ordering the speciation nodes, assigning the first to an arbitrary locus, then proceeding along the speciation nodes and, based on the species-specific locus map, assigning each to an existing locus or to a new locus. Once this has been performed for each species-specific locus map, we retain, for each unique *relative locus map*, only its optimal underlying (absolute) locus map (Fig. 3C). Finally, note that, to remove redundancy, the resulting relative leaf loci are used (instead of the absolute loci) when enumerating species-specific locus maps for the children species.

Once locus maps, orders, relative locus maps, and costs have been computed for all species tree branches (Fig. 3D), we use dy-

is, a (not necessarily unique) LCT with minimum reconciliation cost, subject to the condition that the reconciliation between the locus tree and species tree is the LCA mapping. Note that we have constrained the reconciliation between the locus tree and species tree in order to make our inference algorithm more efficient; we find that this assumption holds in the majority of gene trees in simulation with biologically realistic parameters.

We now present our algorithm DLCpar for inferring an MP LCT. Here, we outline the basics of our algorithm; technical details, including pseudocode and proofs, are provided in Supplemental Material S2 and S3, respectively. DLCpar adopts a multistage approach in which we first infer an optimal species map, then use this to infer an optimal locus map and order (Fig. 3). Our method makes use of the structure of the LCT to search over all possible MP reconciliations and reuse computations.

Our first step is to infer an optimal species map. As the leaf mapping of extant genes to extant species is known, our goal is to map each internal gene tree node to a species tree node. Under our assumptions, an optimal species map is the LCA mapping. We can prove this by contradiction: if the species map is not the LCA mapping, moving a gene tree node "down" (toward the leaves of) the species tree incurs a lower cost; thus, the species map cannot be optimal (for full details, see Theorem S2.1). Once the species map has been determined, implied speciation nodes are added, and the speciation nodes are used to decompose the gene tree into disjoint subtrees that evolve within each species tree branch (Fig. 3A).

To find an optimal locus map and order, note that knowledge of the loci at the speciation nodes would allow us to infer evolution within each species tree branch independently of one another. Therefore, rather than inferring the locus map and order for the entire gene tree at once, we consider the subproblem of determining species-specific locus maps and orders. In particular, for each species, a species-specific locus map is defined only on the gene tree nodes mapped to that species. By definition, a partial order for the LCT already separates gene tree nodes by their mapped species.
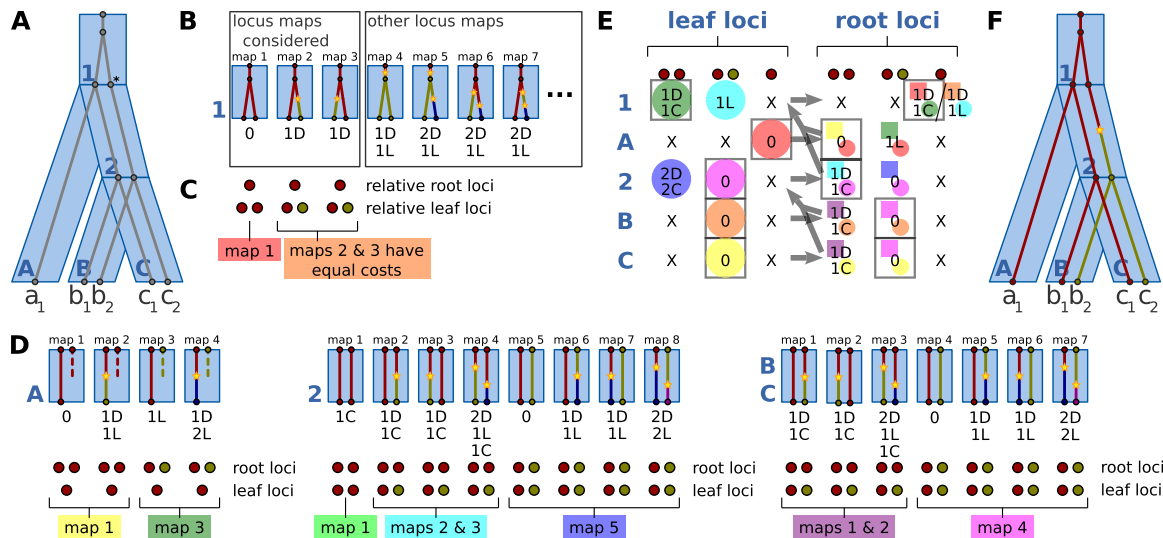
Figure 3. The DLCpar algorithm. In this example, we use equal costs for the evolutionary events. Furthermore, for illustrative purposes, the root of the gene tree has been extended so that a duplication may occur along the root branch; in general, this root branch is not necessary. (*A*) LCA mapping is used to map the gene tree (gray) within the species tree (blue), and implied speciation nodes (*) are added to gene tree branches that span multiple branches of the species tree. (*B*) Starting at the root branch of the species tree, DLCpar enumerates the locus maps and determines an optimal order and reconciliation cost for each. (In practice, some locus maps are not considered since they are never most parsimonious.) (*C*) The leaf loci are remapped to relative loci, and for each unique labeling of root loci and leaf loci, an optimal underlying locus map (and associated order) is selected. (*D*) This is repeated for all branches of the species tree in pre-order traversal, thereby enumerating all locus maps (along with associated optimal orders and reconciliation costs) for this gene tree and species tree. (*E*) For each species and each assignment of root loci and leaf loci, dynamic programming (DP) is used to determine the minimum total cost along all descendant species tree branches. The DP table is filled by post-order traversal of the species tree (arrows); see text for how these costs are computed. Colors indicate which leaf loci (circles) and which species-specific locus map (squares with colors corresponding to parts *C* and *D*) are used. At the root of the species tree, the optimal cost is selected (boxed), and traceback allows assignment of the loci for all speciation nodes (boxed). These can then be used to determine the species-specific locus maps and orders. (*F*) The inferred LCT is shown.

namic programming (DP) to determine optimal relative locus assignments for the speciation nodes (Fig. 3E). To accomplish this, we perform a post-order traversal of the species tree, and for each species tree branch and each assignment of (relative) root loci and (relative) leaf loci, use DP to determine the minimum total cost along all descendant species tree branches. The cost of assigning a particular leaf loci is either known (for extant species) or computed by adding the respective costs of assigning the same loci as root loci in the children species. The cost of assigning a particular root loci is computed by considering each possible leaf loci and adding the cost of the events within the species tree branch; the minimum among all possible choices is selected. At the root of the species tree, the optimal cost is selected, choosing randomly from among the optima if multiple optimal solutions exist. We then traceback (through a pre-order traversal of the species tree) to assign optimal root loci and leaf loci for each species.

Finally, because we have kept a mapping of relative locus maps to absolute locus maps, we can determine the optimal species-specific locus maps for each species, then look up the associated optimal order for each species-specific locus map. Combining the species-specific locus maps and orders yields an optimal locus map and order for the entire gene tree, and together with our previously inferred optimal species map, this constitutes the MP LCT (Fig. 3F).

Results

Simulated data sets

To evaluate the performance of DLCpar, we used the simulated data sets and metrics of Rasmussen and Kellis (2012) and compared DLCpar to the DLCoalRecon algorithm (Rasmussen and Kellis

2012) and the usual most parsimonious reconciliation (MPR) algorithm (Page 1994). To effectively model real data, these data sets used known species trees and parameters (divergence times, duplication and loss rates, population sizes, generation times) for two clades of 12 *Drosophila* species and 17 primates and other mammals (Supplemental Fig. S4A,B), then used the DLCoal model to simulate locus trees and gene trees. To reconcile the gene trees to the species trees, we ran DLCpar using the same event costs across all settings and ran DLCoalRecon using the true parameters used in the simulations (MPR requires no parameters).

For most simulation parameters, DLCpar performance either exceeds (fly data set) or is comparable to (primate data set) that of DLCoalRecon, with both of these programs showing substantial improvement over MPR (Fig. 4; Supplemental Fig. S5). As an example of the comparative performance of the various methods, for the *Drosophila* data set with a duplication-loss rate of 0.0012 events/gene/myr ($1\times$ the estimated real rate), a generation time of 0.1 yr, and an effective population of 25 million, our 500 simulated gene trees contain 232 duplications, 216 losses, and 33,182 pairs of orthologous genes. By confusing ILS events as duplications with compensating losses, MPR infers 1241 duplications (80.2% sensitivity, 15.0% precision), 2495 losses (97.7%, 6.0%), and 21,413 ortholog pairs (64.5%, 100%), with 8.4% (locus tree) topological accuracy. In comparison, DLCpar infers far fewer events at increased accuracy, sensitivity, and precision, with 232 duplications (96.6%, 96.6%), 213 losses (98.1%, 99.5%), 33,192 ortholog pairs (99.98%, 99.95%), and 98.0% topological accuracy. Impressively, DLCpar also outperforms DLCoalRecon, with the latter inferring 242 duplications (90.5%, 86.7%), 216 losses (98.6%, 98.6%), 33,285 ortholog pairs (99.7%, 99.4%), with 96.0% topological accuracy. This is despite the advantages that DLCoalRecon has over
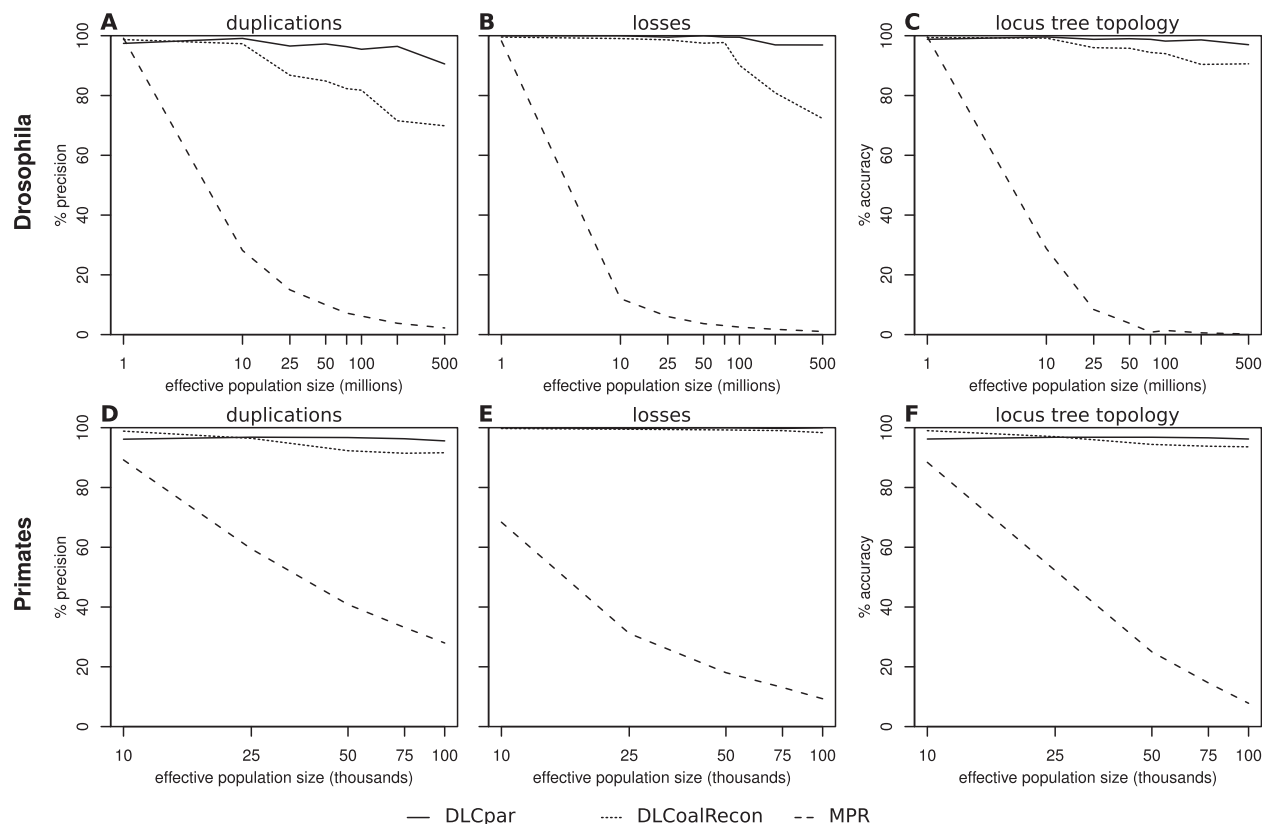
Figure 4. DLCpar performance on simulated fly and primate gene trees. DLCpar, DLCoalRecon, and MPR were used to reconcile simulated gene trees. Simulated data sets and DLCoalRecon and MPR results were obtained from Rasmussen and Kellis (2012), and DLCpar was run with costs of $D = 1$, $L = 1$, $C = 0.5$. Duplications and losses were simulated at rates estimated from real data (flies: 0.0012 events/gene/myr; primates: 0.0017 events/gene/myr), generation times for model organisms were obtained from the literature and assumed equal throughout the clade (flies: 0.1 yr; primates: 20 yr), a wide range of effective population sizes was used, and 500 gene trees were simulated per parameter setting. For the fly data set, DLCpar shows substantial improvement over DLCoalRecon in both the precision of inferring duplications and losses (*A,B*) as well as the accuracy of reconstructing the locus tree topology (*C*). For the primate data set, DLCpar and DLCoalRecon performance is comparable (*D–F*). Both ILS-aware methods dramatically outperform MPR.

DLCpar in this simulation study; in particular, the DLCoal model was used to generate the simulated gene trees, and DLCoalRecon both incorporates additional parameters and uses the true parameters used in the simulations.

We also find that with increasing ILS rate (due to increasing population size with a constant generation time), DLCpar performance remains relatively stable, whereas DLCoalRecon performance decreases. The decreased performance of DLCoalRecon is likely partially attributable to its heuristic search strategy Rasmussen and Kellis (2012): on the same simulated *Drosophila* data set as above (1× duplication-loss rate, $g = 0.1$ yr, $N = 25$ million), when the search is initialized on the correct locus tree, DLCoalRecon infers 232 duplications (97.4% sensitivity, 97.4% precision), 213 losses (98.6%, 100%), and 33,196 ortholog pairs (100%, 99.96%), with 99.2% locus tree topological accuracy. While this performance is slightly higher than that of DLCpar, it further highlights another advantage of DLCpar over DLCoalRecon, namely that DLCpar searches over the entire space of reconciliations compared to the heuristic search approach of DLCoalRecon (Fig. 5). While increasing the search space of DLCoalRecon could lead to performance increases, the total run time of DLCoalRecon already far exceeds that of DLCpar; for the data set above, DLCoalRecon took 5.6 h compared to DLCpar at 6.3 min, a slow down of 54.2×.

In addition to increased accuracy and reduced run time, DLCpar holds a major advantage over DLCoalRecon in terms of applicability. In DLCoalRecon, the maximum a posteriori reconciliation is inferred, thus requiring knowledge of speciation times, duplication and loss rates, generation times, and population sizes. For this evaluation, DLCoalRecon used the true parameters used in the simulation, but in practice, these parameters need to be estimated from genome-wide data and are often difficult, if not impossible, to obtain accurately. In contrast, DLCpar only requires the user to select costs for duplications, losses, and ILS. In this analysis, despite the differences in the number of implanted duplications and losses and varying ILS rate across the different species trees and simulation parameters, we used the same event costs for DLCpar throughout, yet DLCpar performance is still consistently high. Analysis using varying relative costs also shows that DLCpar performance is robust to the choice of these costs (Supplemental Material S6; Supplemental Fig. S6). To demonstrate that DLCpar is applicable to a wide range of data sets, we also conducted additional analyses using simulated species trees and gene trees and found that, almost universally, DLCpar shows dramatic improvement over DLCoalRecon and MPR (Supplemental Material S8; Supplemental Fig. S9).

Finally, we consider the performance of NOTUNG (Durand et al. 2006; Vernot et al. 2007), which allows for ILS by reconciling
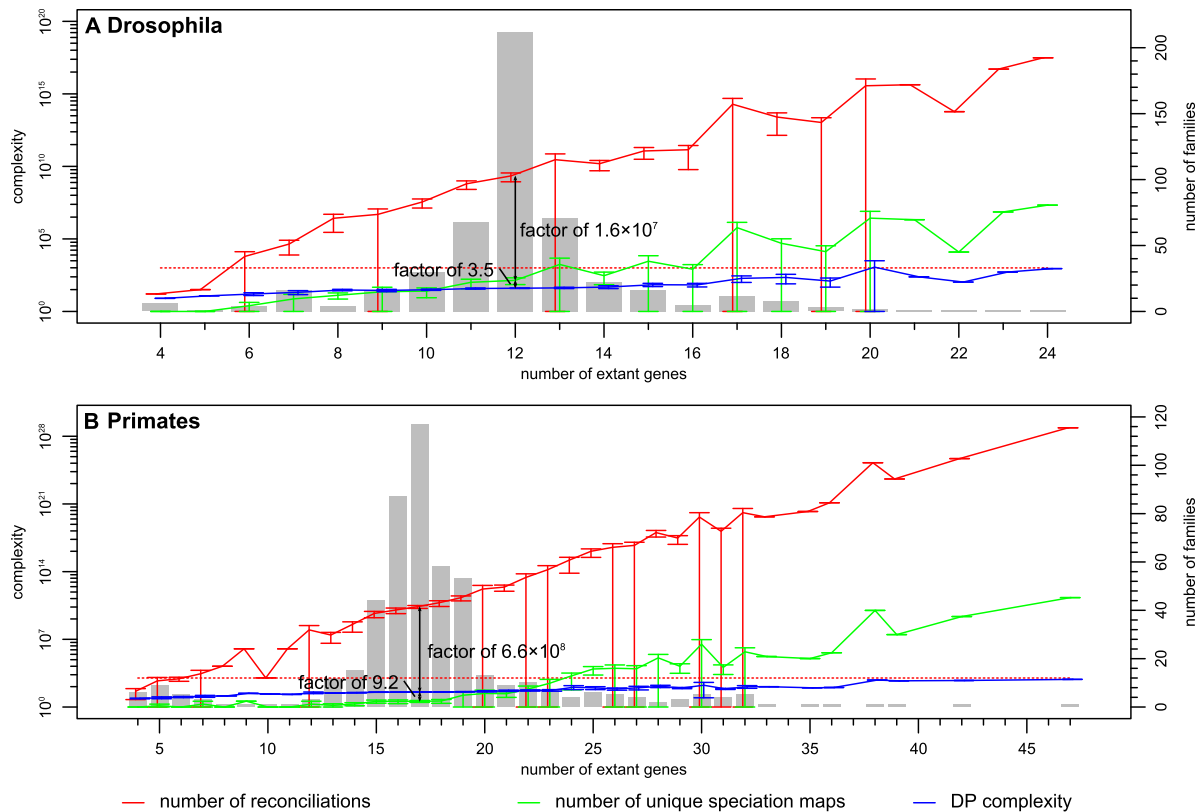
Figure 5. DLCpar complexity on simulated fly and primate gene trees. For each data set, gene trees were divided into classes based on the number of extant genes (counts shown as bars). For each gene tree, we determined the total number of reconciliations considered by DLCpar (red), the number of reconciliations if only relative locus maps for speciation nodes are considered (green), and the number of computations to complete the DP table (blue). The mean and 95% confidence interval (mean $\pm$ 1.96 $\times$ standard error) is shown for each class. Also shown are the number of reconciliations considered by DLCoalRecon (red dots), which was fixed at 1000 search iterations. (DLCoalRecon utilizes a prescreening approach so that 20 $\times$ 10^3 locus trees are proposed, but only 1000 of these are considered using the full probabilistic model.) Data are shown for flies with simulation parameters of 1$\times$ duplication-loss rate, g = 0.1 yr, N = 25 million; and for primates with simulation parameters of 1$\times$ duplication-loss rate, g = 20 yr, N = 25,000. Note that the number of reconciliations considered by DLCpar increases with increasing gene tree size (red), redundancy in the reconciliation search space is dramatically reduced by considering relative locus maps at speciation nodes (green), and dynamic programming further increases efficiency by reusing subproblems (blue).

gene trees to nonbinary species trees. This either requires species trees with well-studied polytomies or a procedure for collapsing species tree branches; a principled approach for the latter does not exist. Therefore, we ran NOTUNG using species tree topologies with a range of polytomies; these included collapsing the shortest branch or branches within the species tree, collapsing branches that have been found to be sources of ILS in previous studies, and collapsing all branches to create a "star" phylogeny. (Note that the "star" phylogeny is the only one that allows for ILS along all branches of the species tree.) In all cases, we find that DLCpar performs dramatically better than NOTUNG (Supplemental Fig. S8). As expected, NOTUNG performs worse than DLCpar and DLCoalRecon but better than MPR in terms of duplication precision, and precision increases with the number of polytomies. That is, by reconciling gene trees to nonbinary species trees, NOTUNG is able to correctly identify spurious duplications due to ILS. However, this comes at the cost of increasingly worse loss precision; in many cases, NOTUNG loss precision is worse than that of even MPR. This is likely because losses associated with a polytomy in the species tree can be assigned to multiple candidate gene tree edges depending on how the order of divergence within the polytomy is resolved; such ambiguities are not possible when both the gene tree and species tree are binary. We also find the duplication sensitivity of NOTUNG

to be similar to that of MPR (that is, worse than that of DLCpar and DLCoalRecon) and the loss sensitivity of NOTUNG to be worse than that of MPR. These results highlight the shortcomings of NOTUNG. In particular, the NOTUNG model cannot capture all possible evolutionary histories because it does not explicitly model the locus history of the gene family, for example, through a locus tree. Because of this, NOTUNG cannot model deep coalescence of a duplicated lineage with its parent lineage. Furthermore, NOTUNG can only capture pairwise relationships present as bifurcations in the gene tree; this is despite the gene tree being an inaccurate representation of the duplication-loss history of the gene family due to the presence of ILS.

Biological data set of 16 fungal genomes

To study whether our observed trends in simulated data sets translate to a real data set, we also assessed the performance of DLCpar on a biological data set of 5351 gene trees across 16 fungal genomes (Supplemental Fig. S4C; Butler et al. 2009); this data set has been used to evaluate numerous phylogenetic algorithms (Wapinski et al. 2007; Rasmussen and Kellis 2011, 2012; Wu et al. 2013). Compared algorithms include BIONJ (Gascuel 1997), PhyML (Guindon and Gascuel 2003), RAxML (Stamatakis 2006), MrBayes (Ronquist and

Huelsenbeck 2003), SYNERGY (Wapinski et al. 2007), TreeBeST (Vilella et al. 2009), PrIME-GSR or DLRS (Arvestad et al. 2004), SPIDIR (Rasmussen and Kellis 2007), SPIMAP (Rasmussen and Kellis 2011), tt (Górecki and Eulenstein 2011), NOTUNG (Chen et al. 2000; Durand et al. 2006), and TreeFix (Wu et al. 2013). As the truth is not known for real data, we used several informative metrics to assess the quality of reconciliations (Table 1).

For comparison, we evaluated several combinations of phylogenetic (gene tree) reconstruction programs and reconciliation programs. Phylogenetic programs included some of the best performing methods; we considered sequence-only (PhyML [Guindon and Gascuel 2003]), hybrid (RAxML [Stamatakis 2006] + TreeFix [Wu et al. 2013]), and species-tree-aware (SPIMAP [Rasmussen and Kellis 2011]) approaches, as well as approaches that integrated other sources of information such as synteny (SYNERGY [Wapinski et al. 2007]). Reconciliation programs included DLCpar, DLCoalRecon, and MPR. However, both SPIMAP and SYNERGY incorporate their own reconciliation model into gene tree reconstruction, and these models are incompatible with an ILS-aware reconciliation model; therefore, we applied DLCpar and DLCoalRecon only to PhyML and TreeFix gene trees.

The real data set included some very large gene families: the largest gene family contains 178 extant genes for an average of ~11 genes per extant species, and another family contains 42 genes in an extant species. We therefore bounded the search space of DLCpar by prescreening reconciliations. For the 0.2% (PhyML: 10/5351; TreeFix: 9/5351) of families that were too complicated using the bounded search (memory requirements exceeded 4 GB), we used DLCpar with a hill-climbing search heuristic. Evaluation of these heuristic approaches on the simulated fly and primate genomes shows only minor changes in performance when using biologically realistic simulation parameters (Supplemental Material S7; Supplemental Fig. S7).

First, we assessed the ability to recover syntenic orthologs (one-to-one homologs that are highly likely to be orthologous, given their surrounding conserved gene order). We find that when applied to trees reconstructed using the same phylogenetic program, DLCpar recovers slightly more syntenic orthologs than

DLCoalRecon (PhyML: 98.4% vs. 97.8%; TreeFix: 99.1% vs. 99.0%). Compared to reconciliations that do not take into account ILS, this is a dramatic improvement over sequence-only methods (64.2%) and higher than hybrid or species-tree-aware methods (96.4%–96.5%). These high sensitivities are also accompanied by more inferred orthologs and fewer inferred duplications and losses. This is, of course, expected, as ILS results in spurious duplications and losses in the gene tree, which are then removed by the ILS-aware reconciliation methods. We find that DLCpar and DLCoalRecon infer similar numbers of orthologs (0.5%–0.6% difference), duplications (0.4%–0.7%), and losses (1.4%–2.3%), and other than SYNERGY, which infers more orthologs and a comparable number of duplications, DLCpar and DLCoalRecon infer at least 16.1% fewer duplications and 35.5% fewer losses than other methods.

Next, we analyzed the duplication consistency score (Vilella et al. 2009), which measures the plausibility of inferred duplications. For each duplication node, this score computes the percentage of species overlap in the two child subtrees; the assumption is that erroneous duplications are often followed by compensating losses and therefore yield a low score. We find that DLCoalRecon slightly outperforms DLCpar, as evidenced by its score distribution (Supplemental Fig. S10). Compared to other methods, the ILS-aware methods have an average duplication consistency score at least 13.1% higher than the next best performing method (SYNERGY) under this metric.

For our last metric, we evaluated the ability to recover more recent duplications due to gene conversion events. When paired with a sequence-only method, DLCpar recovers slightly more recent gene-converted paralogs than DLCoalRecon (91.9% vs. 86.5%), but more impressively, when paired with a hybrid method that accounts for gene tree errors due to statistical uncertainty, DLCpar and DLCoalRecon show 100% recovery. This suggests that even though the alignment is not passed to either DLCpar or DLCoalRecon, and despite neither algorithm modeling gene conversion, the gene tree topology still adequately constrains the underlying locus tree topology so that species tree information does not override sequence evidence in the reconciliation.

Table 1. Evaluation on real fungal data set

Phylo prog[a]	Recon prog[a]	% Orths[b]	# Orths[c]	# Dups[c]	# Losses[c]	DCS[d]	% GC[e]	Phylo run time[f]	Recon run time[g]
PhyML	DLCpar	98.4	581,017	4498	6699	0.823	91.9	45.3 (20.0) sec	74.4 (0.9) sec
PhyML	DLCoalRecon	97.8	575,374	4533	6398	0.883	86.5	45.3 (20.0) sec	54.1 (52.2) sec
PhyML	MPR	64.2	464,479	21,264	64,391	0.153	86.5	45.3 (20.0) sec	–
TreeFix	DLCpar	99.1	590,113	4535	5535	0.899	100.0	25.7 (8.0) min	11.7 (0.8) sec
TreeFix	DLCoalRecon	99.0	583,490	4472	5378	0.927	100.0	25.7 (8.0) min	47.6 (48.6) sec
TreeFix	MPR	96.4	574,946	6062	10,981	0.649	97.3	25.7 (8.0) min	–
SPIMAP	–	96.5	557,981	5407	10,384	0.650	83.8	21.9 (12.2) min	–
SYNERGY	–	99.2	595,289	4604	8179	0.692	2.7	–	–

[a]Several combinations of phylogenetic reconstruction (phylo) programs and reconciliation (recon) programs were evaluated. PhyML+DLCoal, PhyML+MPR, SPIMAP, and SYNERGY results were taken from Rasmussen and Kellis (2012). TreeFix+MPR results were taken from Wu et al. (2013), where TreeFix used as input RAxML trees (with 100 bootstraps) and long search parameters. PhyML+DLCpar, TreeFix+DLCpar, and TreeFix+DLCoalRecon were evaluated using PhyML trees from Rasmussen and Kellis (2012), or TreeFix (long) trees from Wu et al. (2013), and running DLCpar with costs of $D = L = C = 1$, or DLCoalRecon with parameters from Rasmussen and Kellis (2012).
[b]Percentage of 183,374 syntenic orthologs recovered.
[c]Number of pairwise orthologs, duplications, and losses inferred across all gene trees.
[d]Average duplication consistency score. Scores range from 0 to 1, with a higher score indicating more consistent duplications.
[e]Percentage of 37 recent gene-converted paralogs recovered.
[f]Average (median) run time for reconstructing each gene tree. TreeFix run times include times for reconstructing initial RAxML trees. Note that depending on program parameterization, shorter run times may be possible to achieve similar performance. Since SYNERGY trees were downloaded, no run time was estimated.
[g]Average (median) run time for reconciling each gene tree. MPR run times are not included, but on average took less than a second. Note that, depending on program parameterization, shorter run times may be possible to achieve similar performance.

In addition to studying reconstruction quality, we compared the run time performance of DLCpar and DLCoalRecon. We find that while DLCpar has a longer run time than DLCoalRecon for PhyML trees (74.4 sec vs. 54.1 sec), its average run time is highly skewed by large families. If we exclude the longest 0.25% of run times for either program, DLCpar run time dramatically decreases (29.4 sec) while DLCoalRecon run time is unaffected (53.2 sec). (For many gene families, DLCoalRecon run time could likely be decreased by reducing the number of search iterations. However, reducing the search space yields lower accuracy, and there is no principled way of balancing this trade-off.) Additionally, for TreeFix trees, the average run times of DLCpar (DLCoalRecon) reduced by a factor of 6.4 (1.1)× to 11.7 (47.6) sec. Again, the average run time for DLCpar is highly skewed by large families, as 90.0% of families ran in under 1 sec. The decreased run time of DLCpar can likely be attributed to the smaller degree of gene tree-species tree incongruence for the TreeFix trees, which translates to smaller search spaces for the LCT. Finally, we find that the run times for the full phylogenomic pipeline (including both phylogenetic reconstruction and reconciliation) are comparable for DLCpar (PhyML: 2.0 min; TreeFix: 25.9 min) and DLCoalRecon (PhyML: 1.7 min; TreeFix: 26.5 min). However, the total run time for pipelines with TreeFix is dominated by the phylogenetic reconstruction and can be greatly reduced. Our TreeFix trees used long search parameters and were built on RAxML gene trees with 100 bootstraps to yield the highest accuracy, but using a smaller search space and no bootstraps would result in only a minor decrease in performance (Rasmussen and Kellis 2011; Wu et al. 2013) while requiring much less run time.

Discussion

In this work, we have presented a novel algorithm DLCpar for inferring a most parsimonious gene evolutionary history in the presence of duplication, loss, and ILS. To develop our algorithm, we have also introduced the labeled coalescent tree (LCT) and shown it to be an efficient data structure for representing reconciliations. Our analysis of simulated and biological data sets shows that DLCpar achieves performance comparable to more sophisticated probabilistic reconciliation methods at a fraction of the run time. In addition, with its parsimony approach, DLCpar is more applicable to a broad range of species and large data sets.

A core novelty of the LCT is that it labels gene tree branches with the species to which it belongs *and* the locus in which it evolves. This allows the LCT to simultaneously model the duplication-loss history and coalescent history of a gene family. In contrast, current reconciliation structures that map gene tree nodes to species tree nodes are only appropriate for capturing coalescent-only histories (in which case, the reconciliation traces coexisting chromosomes) *or* duplication-loss histories (in which case, the reconciliation traces coexisting loci). The only existing reconciliation framework that accounts for duplication, loss, and ILS is the three-tree (DLCoal) model, which relies on a two-step reconciliation structure, with one (the reconciliation between the gene tree and locus tree) describing the coalescent history and the other (the reconciliation between the locus tree and species tree) describing the duplication-loss history. Thus, the LCT bridges the gap between previously disjoint reconciliation structures and allows us to simultaneously capture multiple sources of incongruence. Furthermore, while conceptually identical to the three-tree model, the LCT enables the development of an efficient inference algorithm based on dynamic programming.

Similarly, a major feature of DLCpar is its simple yet powerful model. DLCpar relies on the three-tree model of evolution and the assumption that the most parsimonious reconciliation is the most accurate. This model has a strong biological foundation and captures the most relevant phenomena responsible for eukaryotic evolution, and parsimony assumptions are used in the simplest, most popular reconciliation methods. Currently, only two other methods exist that simultaneously model duplications, losses, and ILS: NOTUNG and DLCoalRecon. However, NOTUNG, while simple, can only capture a subset of the possible reconciliations and therefore has limited accuracy, and DLCoalRecon, while powerful, is highly parameterized, making it difficult to use in many analyses.

Additionally, by using the LCT structure, DLCpar is able to search over the entire reconciliation space. While there exists previous work on exploring and summarizing the reconciliation space, they consider only duplications and losses (Arvestad et al. 2004; Doyon et al. 2008, 2009, 2012) or additionally horizontal gene transfer (Scornavacca et al. 2013), or, if ILS is addressed, they model only a subset of the evolutionary histories that are possible in our model (Vernot et al. 2007). That is, this work presents the first approach for fully exploring the reconciliation space while accounting for duplications, losses, and ILS. As we have shown, this is a major advantage over the hill-climbing approach of DLCoalRecon, likely resulting in the latter's lower performance in our simulations. Furthermore, DLCpar's dynamic programming approach allows it to reuse computations across different reconciliations. In contrast, DLCoalRecon recomputes the posterior probability for every proposed reconciliation. In addition to the simpler task of computing a parsimony-based reconciliation cost rather than a probability, this likely accounts for the dramatic run time improvement of DLCpar over DLCoalRecon.

Our results demonstrate that a parsimony approach is capable of distinguishing between gene tree-species tree incongruence due to duplication-loss and ILS. This is in spite of the fact that simple formulas exist for relating the number of duplication-loss events and the number of deep coalescence events (Zhang 2011), which, at first glance, might suggest that parsimony methods would always infer one event over the other (for further discussion, see Supplemental Material S5). Furthermore, while we have demonstrated that DLCpar performs well under a variety of species tree parameters, more study is needed in the "anomaly zone" of species histories in which anomalous gene trees (AGTs) exist, that is, where the gene tree *most likely* to evolve under a coalescent model differs from the species tree topology (Degnan and Rosenberg 2006; Degnan et al. 2012; Degnan 2013). Within an anomaly zone, it has been shown that species trees reconstructed using democratic vote (Degnan and Rosenberg 2006), greedy consensus (Degnan et al. 2009), maximum likelihood from concatenated sequences (Kubatko and Degnan 2007), and minimizing-deep-coalescence (Than and Rosenberg 2011) methods are inconsistent, meaning that the estimated species tree topology is incorrect in the limit as the number of sampled gene trees goes to infinity. However, the risk of AGTs are rarely realized in practice (Huang and Knowles 2009), and a Bayesian approach for estimating species trees is consistent (Liu and Edwards 2009). Altogether, this suggests that a more careful treatment of parsimonious reconciliations within this anomaly zone may be required, with particular focus on investigating whether the species tree and gene tree can mislead the locus tree.

In our simulations, we reconciled the simulated gene trees to the species trees, thus assuming that gene trees are accurate and fully resolved. However, in practice, gene tree reconstruction is

complicated by the fact that many gene families lack enough information to confidently support a single gene tree topology. This is an important point as phylogenetic analyses depend strongly on the accuracy of reconstructed gene trees (Hahn 2007). Here, we must distinguish between topological errors due to lack of phylogenetic signal and due to ILS. If we are interested in the duplication-loss (locus) history of a gene family (for example, if the goal is to infer duplications, losses, orthologs, and paralogs), then the tree of interest is the locus tree, not the gene tree. However, to infer the locus tree from molecular sequences involves inferring the gene tree from the sequences, then inferring the locus tree from the gene tree. Phylogenetic signal (or lack thereof) affects the former, and ILS affects the latter. That is, though ILS often biases reconciliation-based inferences to yield false duplications due to stochastic variations in gene tree topology (Zheng and Zhang 2013), there could be sufficient phylogenetic signal to accurately infer the gene tree topology, but this gene tree is still an inaccurate representation of the duplication-loss history due to the presence of ILS. Our analysis of the biological data set demonstrates that while many ILS events can be accounted for by correcting for gene tree errors due to lack of phylogenetic signal, substantial improvement is achieved in locus tree accuracy by accounting for errors due to a lack of phylogenetic signal *and* ILS. To address this need, we suggest a powerful phylogenetic pipeline for gene tree (or more accurately, locus tree) reconstruction and event and homolog inference: reconstruct gene trees using a ML method, correct for topological errors due to statistical uncertainty using TreeFix, and account for ILS using DLCpar. These methods are based on simple models and have few requirements: aside from some user-defined thresholds and costs (to which the algorithms are robust), we require only a sequence alignment and species tree topology, and these are precisely the minimal requirements for any phylogenetic pipeline. Compare this to other methods that require parameters such as speciation times, duplication and loss rates, generation times, or population sizes, which must be estimated from genome-wide data and are often noisy, and even with the additional information, achieve, at most, comparable performance.

In this work, we have presented a method for reconciling binary gene trees to binary species tree, but in practice, inferring fully resolved trees is a major issue due to lack of phylogenetic signal or conflicting sources of information. For example, for the PhyML (RAxML) gene trees reconstructed on our biological data set, 1.4% (0.1%) of gene tree branches are unresolved (have zero branch lengths), and 15.7% (1.7%) of gene trees had at least one unresolved branch. While this might indicate a lack of gene tree polytomies, we also find that for the RAxML gene trees reconstructed on our biological data set, 29.5% of gene tree branches are poorly supported (have bootstrap < 80%), and 84.3% of gene trees have at least one poorly supported branch. (No estimates for PhyML trees are available, as PhyML trees were reconstructed without bootstrapping.) Accounting for gene tree error resolves some polytomies (for example, TreeFix gene trees have 23.7% of gene tree branches and 55.8% of gene trees poorly supported), but, altogether, these results suggest a need for ILS-aware reconciliation methods capable of handling multifurcating gene trees. One possible approach is to resolve gene tree polytomies by expanding them and selecting the expansion with minimum reconciliation cost; such a method has already been presented for the duplication-loss problem (Chang and Eulenstein 2006; Durand et al. 2006; Lafond et al. 2012). Future work could also consider reconciliations with multifurcating species trees. NOTUNG (Vernot et al.

2007) addresses this problem by allowing for ILS at unresolved nodes in the species tree, but as we have shown in our analysis on simulated data sets, this approach greatly diminishes loss sensitivity and precision. In addition, more complex models that differentiate between hard and soft polytomies, that is, polytomies that represent simultaneous divergence versus polytomies that replace a binary branching process that could not be fully resolved, may be required.

While we have shown that DLCpar is robust to the choice of reconciliation costs for the different evolutionary events, we suggest that users try multiple event costs, then either take a conservative approach and use only the intersection of the inferred orthologs and events, or alternatively, find the consensus tree among the inferred locus trees, and from this, infer orthologs and events. A possible future direction is to incorporate knowledge of species tree parameters to estimate these costs. For example, under the DLCoal model, short species tree branches should induce fewer duplications and losses and more deep coalescence, resulting in higher duplication and loss costs and a lower deep coalescence cost. Determining an equation that takes into account such species tree information is not straightforward, however, due to the complex relationship between duplications, losses, and deep coalescence; in particular, ILS increases with more duplications and decreases with more losses (Rasmussen and Kellis 2012). Another option is to use an iterative approach to incorporate the inferred counts of duplications, losses, and extra lineages to assign event costs, but we wished to avoid such circular dependencies in our approach. We note also that using branch-specific costs would require us to search over species maps rather than assuming the LCA mapping.

We also envision the framework presented here to be useful in the future development of methods for understanding gene family evolution. For example, in addition to modeling phylogeny across genomes, the LCT could be extended to capture phylogeny across multiple individuals per genome. This has the benefit of incorporating both population genetic and phylogenomic data into a single analysis. There have also been recent developments in combining models of sequence evolution and duplication-loss (Arvestad et al. 2004; Dubb 2005; Vilella et al. 2009); incorporating substitution rate variation (Rasmussen and Kellis 2007, 2011; Åkerborg et al. 2009), synteny (Wapinski et al. 2007), or horizontal gene transfer (David and Alm 2011; Doyon et al. 2011; Tofigh et al. 2011; Bansal et al. 2012); jointly inferring species trees and gene trees (Boussau et al. 2013) or gene trees and sequence alignments (Liu et al. 2009); reconstructing species trees (Liu and Pearl 2007); and estimating population statistics (Rannala and Yang 2003; Hahn et al. 2005; Drummond and Rambaut 2007). However, so far, all of these methods rely on the multispecies coalescent or duplication-loss (possibly duplication-transfer-loss) model of evolution, and we believe that incorporating joint models is an exciting area for future study.

In conclusion, as we sequence ever denser clades, deep coalescence will only increase, requiring ILS-aware reconciliation methods that are both accurate and efficient. The LCT structure and DLCpar algorithm presented here enable us for the first time to exhaustively and efficiently search the space of reconciliations and infer gene evolutionary histories on a genome-wide scale without additional knowledge of the species tree beyond its topology. Thus, we recommend that studies use DLCpar in place of the standard MPR reconciliation algorithm, and we believe that its application will have many important implications for future investigations of gene evolution.

Data access

The DLCpar software and Supplemental Material are freely available for download at http://compbio.mit.edu/dlcpar.

Acknowledgments

References

Åkerborg Ö, Sennblad B, Arvestad L, Lagergren J. 2009. Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proc Natl Acad Sci* **106:** 5714–5719.

Arvestad L, Berglund A-C, Lagergren J, Sennblad B. 2004. Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. In *Proceedings of the eighth annual international conference on research in computational molecular biology, RECOMB '04*, pp. 326–335. ACM, New York.

Avise JC, Robinson TJ. 2008. Hemiplasy: A new term in the lexicon of phylogenetics. *Syst Biol* **57:** 503–507.

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

Boussau B, Szöllosi GJ, Duret L, Gouy M, Tannier E, Daubin V. 2013. Genome-scale coestimation of species and gene trees. *Genome Res* **23:** 323–330.

Butler G, Rasmussen MD, Lin MF, Santos MA, Sakthikumar S, Munro CA, Rheinbay E, Grabherr M, Forche A, Reedy JL, et al. 2009. Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. *Nature* **459:** 657–662.

Chang W-C, Eulenstein O. 2006. Reconciling gene trees with apparent polytomies. In *Lecture notes in computer science* (ed. Chen DZ, Lee D), Vol. 4112, pp. 235–244. Springer, Berlin.

Chen K, Durand D, Farach-Colton M. 2000. NOTUNG: A program for dating gene duplications and optimizing gene family trees. *J Comput Biol* **7:** 429–447.

Datta RS, Meacham C, Samad B, Neyer C, Sjölander K. 2009. Berkeley phog: Phylofacts orthology group prediction web server. *Nucleic Acids Res* **37:** W84–W89.

David LA, Alm EJ. 2011. Rapid evolutionary innovation during an archaean genetic expansion. *Nature* **469:** 93–96.

Degnan JH. 2013. Anomalous unrooted gene trees. *Syst Biol* **62:** 574–590.

Degnan JH, Rosenberg NA. 2006. Discordance of species trees with their most likely gene trees. *PLoS Genet* **2:** e68.

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol Evol* **24:** 332–340.

Degnan JH, DeGiorgio M, Bryant D, Rosenberg NA. 2009. Properties of consensus methods for inferring species trees from gene trees. *Syst Biol* **58:** 35–54.

Degnan JH, Rosenberg NA, Stadler T. 2012. A characterization of the set of species trees that produce anomalous ranked gene trees. *IEEE/ACM Trans Comput Biol Bioinform* **9:** 1558–1568.

Doyon J-P, Chauve C, Hamel S. 2008. Algorithms for exploring the space of gene tree/species tree reconciliations. In *Lecture notes in computer science* (ed. Nelson CE, Vialette S), Vol. 5267, pp. 1–13. Springer, Berlin.

Doyon J-P, Chauve C, Hamel S. 2009. Space of gene/species trees reconciliations and parsimonious models. *J Comput Biol* **16:** 1399–1418.

Doyon J-P, Scornavacca C, Gorbunov KY, Szöllősi GJ, Ranwez V, Berry V. 2011. An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. In *Lecture notes in computer science* (ed. Tannier E), Vol. 6398, pp. 93–108. Springer, Berlin.

Doyon J-P, Chauve C, Hamel S. 2012. An efficient method for exploring the space of gene tree/species tree reconciliations in a probabilistic framework. *IEEE/ACM Trans Comput Biol Bioinfor* **9:** 26–39.

Drummond A, Rambaut A. 2007. Beast: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* **7:** 214.

Dubb L. 2005. "A likelihood model of gene family evolution." PhD thesis, University of Washington, Seattle.

Durand D, Hallórsson BV, Vernot B. 2006. A hybrid micro-macroevolutionary approach to gene tree reconstruction. *J Comput Biol* **13:** 320–335.

Gascuel O. 1997. BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol* **14:** 685–695.

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

Górecki P, Eulenstein O. 2011. A linear time algorithm for error-corrected reconciliation of unrooted gene trees. In *Lecture notes in computer science* (ed. Chen J, et al.), Vol. 6674, pp. 148–159. Springer, Berlin.

Górecki P, Tiuryn J. 2006. Dls-trees: A model of evolutionary scenarios. *Theor Comput Sci* **359:** 378–399.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52:** 696–704.

Hahn M. 2007. Bias in phylogenetic tree reconciliation methods: Implications for vertebrate genome evolution. *Genome Biol* **8:** R141.

Hahn MW, De Bie T, Stajich JE, Nguyen C, Cristianini N. 2005. Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Res* **15:** 1153–1160.

Huang H, Knowles LL. 2009. What is the danger of the anomaly zone for empirical phylogenetics? *Syst Biol* **58:** 527–536.

Huerta-Cepas J, Dopazo H, Dopazo J, Gabaldon T. 2007. The human phylome. *Genome Biol* **8:** R109.

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

Kubatko LS, Degnan JH. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst Biol* **56:** 17–24.

Lafond M, Swenson KM, El-Mabrouk N. 2012. An optimal reconciliation algorithm for gene trees with polytomies. In *Lecture notes in computer science* (ed. Raphael B, Tang J), Vol. 7534, pp. 106–122. Springer, Berlin.

Li H, Coghlan A, Ruan J, Coin LJ, Hériché JK, Osmotherly L, Li R, Liu T, Zhang Z, Bolund L, et al. 2006. TreeFam: A curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res* **34:** D572–D580.

Liu L, Edwards SV. 2009. Phylogenetic analysis in the anomaly zone. *Syst Biol* **58:** 452–460.

Liu L, Pearl DK. 2007. Species trees from gene trees: Reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst Biol* **56:** 504–514.

Liu K, Raghavan S, Nelesen S, Linder CR, Warnow T. 2009. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science* **324:** 1561–1564.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290:** 1151–1155.

Maddison WP. 1997. Gene trees in species trees. *Syst Biol* **46:** 523–536.

Mi H, Muruganujan A, Thomas PD. 2013. Panther in 2013: Modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res* **41:** D377–D386.

Ohno S. 1970. *Evolution by gene duplication*. Springer-Verlag, New York.

Page RD. 1994. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Syst Biol* **43:** 58–77.

Pamilo P, Nei M. 1988. Relationships between gene trees and species trees. *Mol Biol Evol* **5:** 568–583.

Peterson ME, Chen F, Saven JG, Roos DS, Babbitt PC, Sali A. 2009. Evolutionary constraints on structural similarity in orthologs and paralogs. *Protein Sci* **18:** 1306–1315.

Rannala B, Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* **164:** 1645–1656.

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

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

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

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19:** 1572–1574.

Rosenberg NA. 2002. The probability of topological concordance of gene trees and species trees. *Theor Popul Biol* **61:** 225–247.

Scornavacca C, Paprotny W, Berry V, Ranwez V. 2013. Representing a set of reconciliations in a compact way. *J Bioinform Comput Biol* **11:** 1250025.

Stamatakis A. 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22:** 2688–2690.

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

Than CV, Rosenberg NA. 2011. Consistency properties of species tree inference by minimizing deep coalescences. *J Comput Biol* **18:** 1–15.

Tofigh A, Hallett M, Lagergren J. 2011. Simultaneous identification of duplications and lateral gene transfers. *IEEE/ACM Trans Comput Biol Bioinform* **8:** 517–535.

Vernot B, Stolzer M, Goldman A, Durand D. 2007. Reconciliation with non-binary species trees. *J Comput Biol* **15:** 981–1006.

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res* **19:** 327–335.

Wakeley J. 2009. *Coalescent theory: An introduction.* Roberts & Co., Greenwood Village, CO.

Wapinski I, Pfeffer A, Friedman N, Regev A. 2007. Natural history and evolutionary principles of gene duplication in fungi. *Nature* **449:** 54–61.

Wu T, Zhang L. 2011. Structural properties of the reconciliation space and their applications in enumerating nearly-optimal reconciliations between a gene tree and a species tree. *BMC Bioinformatics* **12:** S7.

Wu Y-C, Rasmussen MD, Bansal MS, Kellis M. 2013. Treefix: Statistically informed gene tree error correction using species trees. *Syst Biol* **62:** 110–120.

Zhang L. 2011. From gene trees to species trees II: Species tree inference by minimizing deep coalescence events. *IEEE/ACM Trans Comput Biol Bioinform* **8:** 1685–1691.

Zheng Y, Zhang L. 2013. Effect of incomplete lineage sorting on tree-reconciliation-based inference of gene duplication. In *Lecture notes in computer science* (ed. Cai Z, et al.), Vol. 7875, pp. 261–272. Springer, Berlin.

Zmasek CM, Eddy SR. 2002. Rio: Analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics* **3:** 14.