Species Tree and Reconciliation Estimation under a Duplication-Loss-Coalescence Model

Peng Du
Department of Computer Science
Houston, Texas
peng.du@rice.edu

Luay Nakhleh
Department of Computer Science
Houston, Texas
nakhleh@rice.edu

ABSTRACT

Gene duplication and loss are two evolutionary processes that occur across all three domains of life. These two processes result in different loci, across a set of related genomes, having different gene trees. Inferring the phylogeny of the genomes from data sets of such gene trees is a central task in phylogenomics. Furthermore, when the evolutionary history of the genomes includes short branches, deep coalescence, or incomplete lineage sorting (ILS), could be at play, in addition to duplication and loss, further adding to the complexity of gene/genome relationships. Recently, researchers have developed methods to infer these evolutionary processes by simultaneously modeling gene duplication, loss, and incomplete lineage sorting with respect to a given (fixed) species tree.

In this work, we focused on the task of inferring species trees, as well as locus and gene trees, from sequence data in the presence of all three processes. We developed a search heuristic for estimating the maximum a posteriori species/locus/gene tree triad, as well as their associated parameters, from the sequence data of independent gene families.

We demonstrate the performance of our method on simulated data and a data set of 200 gene families from six yeast genomes. Our work enables new statistical phylogenomic analyses, particularly when hidden paralogy and incomplete lineage sorting could be simultaneously at play.

ACM Reference Format:
Peng Du and Luay Nakhleh. 2018. Species Tree and Reconciliation Estimation under a Duplication-Loss-Coalescence Model. In *ACM-BCB'18: 9th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, August 29-September 1, 2018, Washington, DC, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3233547.3233600

1 INTRODUCTION

A species tree models how species split from common ancestors and diverge over time. Gene trees grow within (and, sometimes across) the branches of the species tree. Each gene tree represents how a single gene family evolves from its common ancestor. The relationship between the species tree and the gene trees within its branches is often very complex due to a variety of processes that could act on the genomes, such as horizontal gene transfer (HGT), incomplete lineage sorting (ILS), and gene duplication and loss [12].

The complex relationship between species and gene trees has given rise to two challenging computational problems: The inference problem and the reconciliation problem. The goal of the inference problem is to estimate a species tree from sequence data of multiple gene families (which are typically assumed to be independent). The reconciliation problem seeks to estimate the evolutionary processes that acted on the individual gene families by reconciling gene family trees with a given species tree.

Several methods have been developed for inferring species trees assuming only ILS; e.g., [8, 10, 11, 13, 14, 24]. All these methods make use of the multispecies coalescent model [6] and assume all sequences within a gene family are orthologs. An array of methods that focus only on gene duplication and loss were also introduced; e.g., [1, 2, 4, 16, 17, 20]. In these methods, incomplete lineage sorting is ignored.

In [18], the authors developed the first unified model that took both duplication/loss and incomplete lineage sorting into consideration by introducing a new tree called a " locus tree." The model that Rasmussen and Kellis introduced consists of three trees and a well-defined reconciliation between them: A locus tree grows within the branches of the species tree assuming only duplication and loss, and a gene tree grows within the branches of the locus tree assuming only coalescence events. The resulting model accounts for gene trees whose disagreement with the species tree are due to potential gene duplication, loss, and incomplete lineage sorting events. The authors developed a Bayesian framework for estimating the duplication, loss, and incomplete lineage sorting events from inputs that consist of species and gene trees. That is, in their work, neither gene trees nor species trees were inferred; they were assumed to have been estimated.

More recently, Zhang and Wu [25] extended the work of [18] by developing a method for co-estimation of gene trees and reconciliations directly from DNA sequences. Still, in their work, the species tree was assumed to be given. Thus, the task of inferring the species and gene trees, along with their parameters, directly from the sequence data, remained unsolved. In this paper we address this problem. In particular, we develop a set of operations that allow for searching for the maximum a posteriori (MAP) estimate of the 3-tree model parameters from sequences of multiple, independent gene families.

We implemented our method and studied its performance on simulated data and a biological data set. For the simulated data, we generated gene trees under the 3-tree model and evolved sequences down these gene trees under a given model of sequence evolution. We then inferred the model parameters using our method

ACM-BCB'18, August 29-September 1, 2018, Washington, DC, USA

ACM ISBN 978-1-4503-5794-4/18/08…$15.00
https://doi.org/10.1145/3233547.3233600

directly from the sequence data. Results show good performance and a promising method for estimating evolutionary histories while accounting simultaneously for gene duplication and loss, as well as incomplete lineage sorting. Our work provides another step towards inferences that utilize as much of the available genomic data as possible without throwing out much of it due to lack of models that account for biological complexity. The major bottleneck of the method is its computational complexity. Work on speeding up the method and improving its computational performance is imperative for it to be scalable to larger data sets.

2 METHOD

In this paper, we use the 3-tree model of [18] and design a set of operators on this model to enable inference of the species tree (along with the other trees). Sections 2.1 and 2.2 provide a review of the model. The main contribution of this paper is the material of Section 2.3.

2.1 3-Tree Model

In the 3-tree model, we have a species tree, a locus tree and a gene tree and a reconciliation between them.

2.1.1 Species Tree. Similar to the definition in [18], we define species tree as $\mathbb{S} = (V(S), E(S), \tau^S)$ with vertex set $V(S)$, branch set $E(S)$ and branch lengths τ^S. Further, $V(S)$ and $E(S)$ are collectively denoted as S. For a vertex x, we define its parent as $pa(x)$ and the set of its children as $c(x)$. A branch $e = (pa(x), x)$ is abbreviated as $e(x)$. The definition of parent, children and branch are also applied to the following two trees. The population sizes are given and we denote the population size on branch $e(x)$ as $N(x)$.

2.1.2 Locus Tree and Locus Tree to Species Tree Reconciliation. A locus tree $\mathbb{L} = (L, \tau^L)$ is generated by applying duplication and loss onto the species tree. It is generated by a top-down birth-death process [1, 2, 20] with the idea that genes can get duplicated and lost in the genomes as the population evolves. When a duplication is encountered, the branch is bifurcated to two branches, and when a loss happens, the branch is terminated. We thus have a reconciliation R^L from the locus tree to the species tree, where the vertices on the locus tree can be mapped to either the vertices or the branches of the species tree. If $u \in V(L)$ is mapped to a species tree vertex, then we call it a speciation vertex, and if it is mapped to a species tree branch, we call it a duplication vertex. Branches with no existing leaf vertices are pruned out. For a duplication, a new locus is generated, so we have $\delta^L((u, v)) = 0$ and $\delta^L((u, w)) = 1$ to indicate (u, w) leads to the newly created (daughter) locus and (u, v) is the mother branch where u is the duplication vertex. The population size of branch $e = (u, v)$ in the locus tree is the population size of the branch $e' = (x, y)$ where $R^L(v) = y$ or $R^L(v) = (x, y)$. Fig. 1 shows the species tree, the locus tree evolving inside of the species tree (Fig. 1(a)), and the pruned locus tree (Fig. 1(b)).

2.1.3 Gene Tree and Gene Tree to Locus Tree Reconciliation. A gene tree $\mathbb{G} = (G, \tau^G)$ coalesces within the branches of the locus tree, and the reconciliation from the gene tree to the locus tree is denoted by R^G. The two reconciliations R^L and R^G are collectively denoted by R. For each branch $e = (u, w)$ with $\delta^L((u, w)) = 1$, all the gene vertices mapped to the leaf vertices under w must coalesce more recently than u when going backward in time (from leaves toward the root). Also, we define M as the mapping from the gene tree leaf vertex-set to the locus tree leaf vertex-set. M indicates what gene is from what locus in the locus tree. Fig. 2 shows a locus tree and a gene tree within its branches.

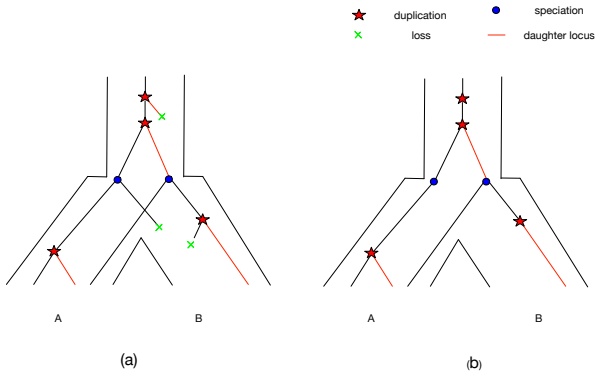

Figure 1: A gene duplication and loss scenario inside of a species tree. (a) The "raw" duplication and loss events given by a birth-death process. (b) Loci that go extinct are deleted.

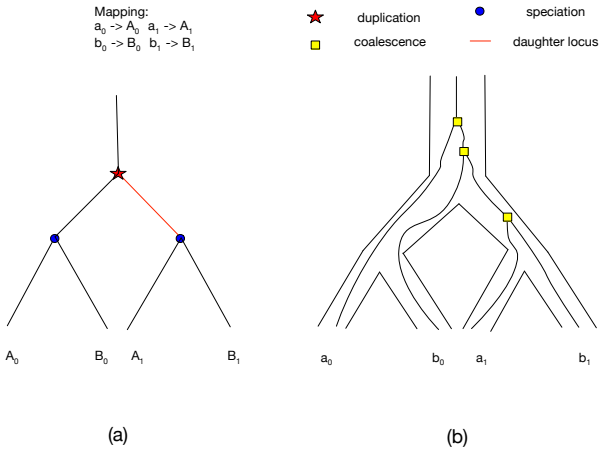

Figure 2: The coalescent events of genes within the branches of a locus tree. (a) Locus tree generated by the birth-death process. (b) Gene lineages coalesce within the branches of the locus tree. The mapping indicates that gene a_0 is mapped to locus A_0, a_1 is mapped to A_1, etc. The coalescence of gene lineages of duplicated locus must happen more recently than the duplication event when going backward in time: a_1 and b_1 coalesce more recently than the duplication event.

Fig. 3 shows the reconciliations from the gene tree to the locus tree (R^G) and from the locus tree to the species tree (R^L).

2.1.4 Assumptions. In this model, some assumptions are made as were made in [18] and [25] .

(1) After the duplication, the daughter locus is "moved" to an unlinked position on the genome and the evolution of the

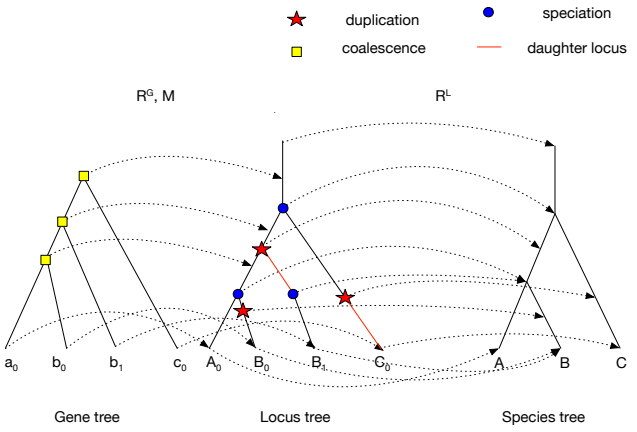

Figure 3: Reconciliation of a gene tree with a locus tree, and then of a locus tree with a species tree. The dashed lines show the reconciliations.

daughter and mother loci become independent. In this way, we can calculate the coalescent probability of the mother and daughter loci independently with respect to each other.

(2) No hemiplasy [3] is allowed. That is, all the duplication and loss events are all fixed to the species under that event; they either all go extinct or all are kept in all the descendent species. This allows to explain all unobserved loci by means of gene loss.

(3) Exactly one individual per species is sampled.

2.2 Probability Distribution

For a species tree $\mathbb{S}$ and a set of gene families $\mathbb{GF}$ with each member $\mathbb{GF}_i = (\mathbb{L}_i, \mathbb{G}_i, R_i, M_i, \delta_i^L)$, and parameters θ, the posterior $p(\mathbb{S}, \mathbb{GF}, \theta|D)$ given observed DNA sequences D is

$$p(\mathbb{S}, \mathbb{GF}, \theta|D) = \prod_i p(\mathbb{GF}_i|\mathbb{S}, \theta) \times p(D_i|\mathbb{GF}_i) \quad (1)$$

$$\times p(\mu) \times p(\lambda) \times p(N^S) \quad (2)$$

$$\times p(\mathbb{S}) \quad (3)$$

$$/p(D) \quad (4)$$

where D_i is the DNA sequences for $\mathbb{GF}_i$ and $\theta = \{\mu, \lambda, \gamma, N^S\}$ which are the duplication rate, loss rate, substitution rate and population size respectively. The first line can further be decomposed as (we drop the subscript i for the individual term for better readability):

$$p(\mathbb{GF}_i|\mathbb{S}, \theta) \times p(D|\mathbb{GF}_i) = p(G, \tau^G, R^G|L, \tau^L, \delta^L, M, N^S) \quad (5)$$

$$\times p(M|\mathbb{L}, R^L, \delta^L) \quad (6)$$

$$\times p(L, \tau^L, R^L, \delta^L|S, \tau^S, \mu, \lambda) \quad (7)$$

$$\times p(D|G, \tau^G, \gamma). \quad (8)$$

The term $p(G, \tau^G, R^G|L, \tau^L, \delta^L, M, N^S)$ is the probability of the gene tree coalescing in the locus tree and is derived in [18, 25]. The term $p(M|\mathbb{L}, R^L, \delta^L)$ is the probability of the labeling of gene tree leaves to the locus tree leaves. Since we assume no prior knowledge of locus information of each sampled gene copy from a certain species, the mapping has a uniform distribution based on the number of possible permutations:

$$p(M|\mathbb{L}, R^L, \delta^L) = \prod_{x \in L(S)} \frac{1}{|u : \sigma(u) = x|!}. \quad (9)$$

The term $p(L, \tau^L, R^L, \delta^L|S, \tau^S, \mu, \lambda)$ is the probability of the locus tree generated inside of the species tree with duplication rate μ and loss rate λ and is derived in [1, 2, 20]. The term $p(S, \tau^S)$ is the prior of the species tree which is a compound prior with uniform prior on the topology and exponential prior on branch lengths as in [8, 21].

2.3 MAP Inference

Our goal is to find the maximum a posteriori (MAP) estimate of the parameters; that is,

$$(\mathbb{S}^*, \mathbb{GF}^*, \theta^*) = \text{argmax}_{(\mathbb{S}, \mathbb{GF}, \theta)} p(\mathbb{S}, \mathbb{GF}, \theta|D). \quad (10)$$

Given the 3-tree model, search for the MAP estimate must operate on all the three trees and the reconciliations between them. There is another angle of looking at this model which is to break the 3-tree model into three components: speciation events, duplication/loss events, and coalescent events. The speciation events determine the topology and branch lengths of the species tree and locus tree while the duplication/loss events determine the topology and branch lengths of the locus tree and gene tree and, finally, the coalescent events determine the topology and branch lengths of the gene tree. The complexity of search under this model is that operating on one component of the 3-tree model without regard to the other two can easily result in invalid instances of the model. Therefore, operations to search the space of parameters must be designed carefully. We have three groups of operators for the three different types of trees and a fourth group of operators to change the continuous parameters. Species tree operators inspect and modify the locus and gene trees in conjunction with modifying the species tree. Locus tree operators inspect and potentially modify the gene tree while modifying the locus tree. Finally, gene tree operators only make changes to the gene trees. Some of the operators are described below.

2.3.1 Species Tree Operators. Here we describe the species tree operators.

Species-Tree-SPR works by cutting a subtree (donor) and pasting to another (receiver) (Fig. 4). First, we randomly pick an internal vertex x_1 on the species tree as the donor vertex and randomly pick up one of its children x_2. This planted subtree rooting at x_1 leading to the clade x_2 will be pruned and attached to a new location. We randomly pick up a branch $e = (y_1, y_2)$ such that $\tau(y_1) > \tau(x_1)$ and $\tau(y_2) < \tau(x_1)$of species tree as the branch to receive the planted subtree. A new speciation vertex y_3 is created on (y_1, y_2). New loci will be created if the donor subtree has more loci than on the receiver branch. There is only one loci on the branch (y_1, y_2), but there are 3 loci from the planted tree that need to be added to (y_1, y_2). So, two new loci need to be created. The creation is done by adding two duplications u_2 and u_3 and now we have 3 loci u_4, u_5 and u_6. At the same time, the planted subtree rooting at v_1 leading to v_3 will be empty since the loci leading to BC are all removed, so (v_1, v_3) and the subtree below it will be deleted. In the gene tree, the

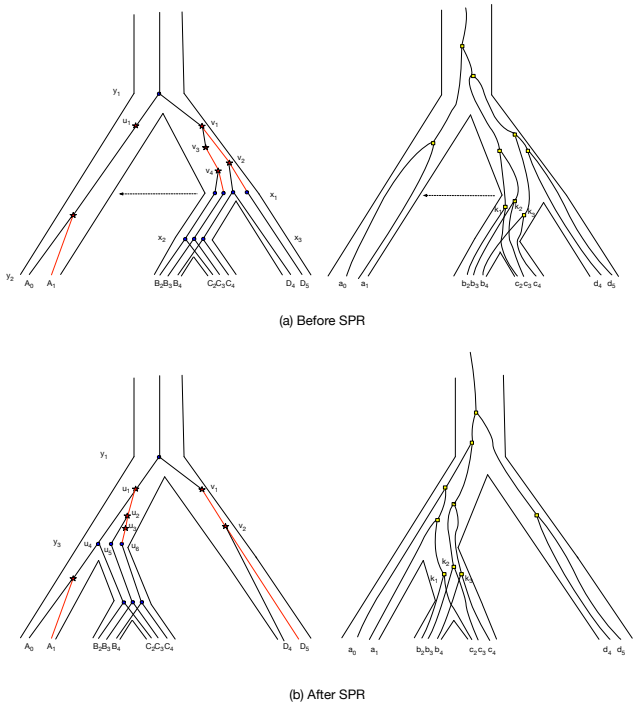

Figure 4: The Species-Tree-SPR operator. The subtree leading to B and C in the species tree is cut and pasted to be with A. New loci are created to accommodate the 6 loci from B and C. The old loci will be deleted and finally the new locus tree is constructed. The gene tree is rebuilt by cutting the affected gene lineages and reattaching to proper locations in the new gene tree by simulating coalescent events inside of the locus tree.

3 lineages k_1, k_2 and k_3 leading to the 6 leaf nodes b_2, b_3, b_4, c_2, c_3 and c_4 will be pruned. As now, we have 3 vertices on the locus tree mapped to y_3 and 3 gene lineages k_1, k_2 and k_3 that need to be handled. We randomly assign the 3 lineages to the u_4, u_5 and u_6 and will rebuild the gene tree starting from the 3 loci on the locus tree by simulating coalescence events. For example, a_1 can coalesce with k_1 inside (u_1, u_4) and k_2 and k_3 can coalesce in (u_1, u_2) etc. Also, sometimes, with a certain probability, the whole locus tree and gene tree will be totally reinitiated.

Scale-Time. In this operator, the speciation times τ of all internal vertices of the tree $\mathbb{S}$ are scaled by a scale factor r and modified into $\tau' = r\tau$. r is drawn from Uniform$(f, \frac{1}{f})$ where f is a small value close to 1. Times of all the speciation vertices on the locus tree will also be changed accordingly. This operator can be invalid and thusly rejected if temporal constraints are violated.

Scale-Time-All. In this operator, the times τ of all internal vertices of all the 3 trees $\mathbb{S}$, $\mathbb{L}$ and $\mathbb{G}$ are scaled by a factor u and modified into $\tau' = r\tau$. r is drawn from Uniform$(f, \frac{1}{f})$ where f is a small value close to 1.

Change-Time. An internal vertex b is selected uniformly at random and the time τ of the vertex is modified into $\tau' \sim$ Uniform(l, h) where l and h are the lower and higher bound of time τ_b respectively. The lower bound and upper bound are determined by the times of $c(b)$ and all duplication and coalescence events on the subtree rooting at b in the locus trees and gene trees and by the minimum time of $pa(b)$ and times of all the duplication and coalescence events above b. Times of the affected speciation vertices on the locus trees will also be changed accordingly.

2.3.2 Operators on the Locus Trees. We describe operators that change a locus tree and also the gene tree.

Move-Up-Down-Duplications. In this operator, we define two complementary operators, one is to move the duplication event down to the two child species, called as **Dup-Move-Down** while the other one is to move the duplications in the two child species up into the mother species denoted as **Dup-Move-Up**. With this operator, we are able to change the location and number of duplications (Fig. 5). If the state is in Fig. 5.a, i.e. there is only one duplication event u mapped the the species tree branch $(pa(x), x)$ that give rise to four leaf nodes, we can move the duplication events down and map them to (x, y) and (x, z) respectively to form u_1 and u_2. The gene tree needs to be changed too. In Fig. 5.a, a_1 and a_2 coalesce and a_3 and a_4 coalesce, we need to change it so that a_1 and a_3 coalesce and a_2 and a_4 coalesce. The new duplication event times and coalescence times are randomized. **Dup-Move-Up** can be used to restore to Fig. 5.a from Fig. 5.b.

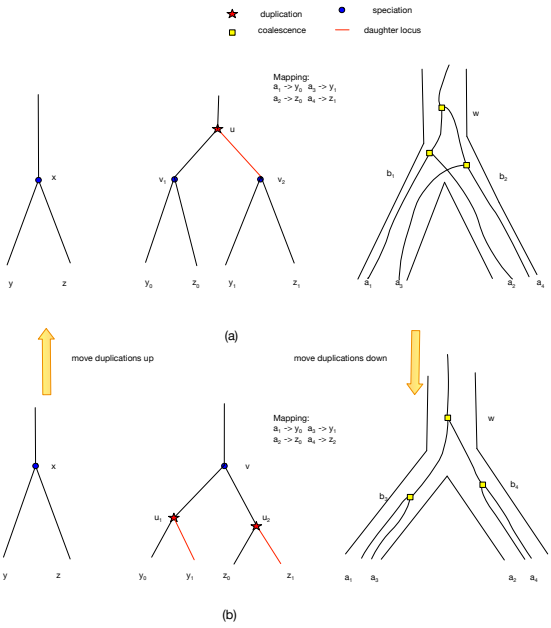

Figure 5: The duplication is on the mother branch giving 4 loci with 2 for each child species in (a). It can be moved down to two independent duplications on the two child species to form (b). Also, if the state is in (b), the two duplications can be "merged" to be on the mother species branch. The duplication heights are randomized and the lineages in the gene tree are swapped.

Add-Remove-Single-Locus. If a duplication vertex has one child,

we call it a single locus (the top one in Fig. 1.b), it can be removed or added back. The operator includes two operators **Remove-Single-Locus** and **Add-Single-Locus**. **Add-Single-Locus** is used to add single locus onto the locus tree and the single locus can be deleted by **Remove-Single-Locus**.

Relocate-Duplications. We can relocate the duplication subject to being bounded by its upper and lower bound. The upper bound is the minimum of the times of coalescence events, duplication events or species events above this vertex and the lower bound is the maximum time of coalescence events, duplication events or species events below this vertex. We denote the current time of the duplication as τ. For a duplication, h denotes its upper bound, l denotes the lower bound, a random time $\tau' \sim \text{Uniform}(l, h)$ is sampled and is assigned to τ.

Locus-Tree-SPR. This operator will cut a locus to another position. For example, in Fig. 6.a, the locus C_2 is the child of v_1, the coalescence event of c_2 is at x_1. v_2 is one potential location (among others) for relocating the locus. With a probability, we can then cut off C_2 and reattach to v_2 by extending the branches down to the leaf vertex, (see the path from v_2 to C_2). Because we allow single locus vertex on the locus tree, sometimes, we randomly add single locus vertices and results in v_3 in this case. In Fig. 6.a, c_2 is coalesced at x_1, we now find possible locations for attaching c_2. In this case, it is coalesced on branch $(pa(v_2), v_2)$ and form a new gene tree vertex x_2.

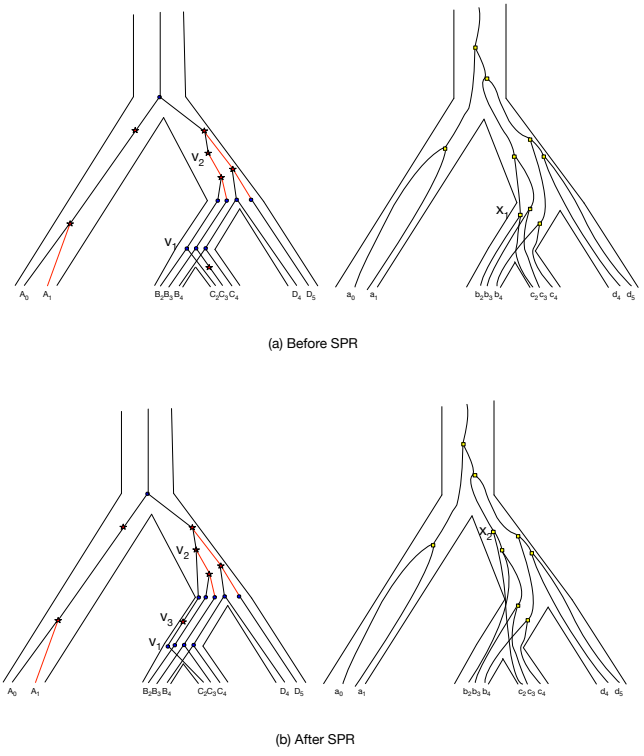

Figure 6: The Locus-Tree-SPR operator. Each leaf locus is attached a certain vertex in the locus tree, we randomly picked one that can be attached to other places.

2.3.3 Operators on the Gene Trees. **Scale-Time.** In this operator, the coalescent times τ of all internal vertices of the tree $\mathbb{G}$ are scaled by a scale factor r and modified into $\tau' = r\tau$. r is drawn from $\text{Uniform}(f, \frac{1}{f})$ where f is a value close to 1. This operator can be invalid and thusly rejected if temporal constraints are violated.

Change-Time. An internal vertex b is selected uniformly at random and the time τ of the vertex is modified into $\tau' \sim \text{Uniform}(l, h)$ where l and h are the lower and higher bound of time τ respectively. The upper bound and lower bound are determined by the minimum of coalescence times of the $pa(b)$ or the duplication vertex if the locus branch b is mapped to is the daughter locus branch of that duplication and the coalescence times of child vertices of b or the temporal bound on the locus tree.

Gene-Tree-SPR. This operator will cut a subtree in the gene tree and reattach to another branch. This operator is only done between gene lineages reconciled to the same locus tree branch.

Shuffle-Labeling. In this operator, we shuffle the labeling function M from gene tree leaves to locus tree leaves. In this operator, we select a subset of species and then randomly shuffle the labeling of genes sampled from these species respectively.

2.3.4 Operators on the Non Tree parameters. We can also search for the duplication and loss rate using **Change-Duplication-Loss-Rate** by making small local changes to them. Suppose the current duplication rate is μ. The new value μ' can be proposed by adding a small random value r where $r \sim \text{Uniform}(-t, t)$ where t is another small value. The similar is done for loss rate.

2.3.5 Optimization. We use a simple hill-climbing algorithm for searching for the optimal point as given by Eq. (10). We run multiple worker chains to perform the search. We have a main thread that oversees all the worker threads each of which searches locus trees, gene trees and reconciliations on a given species tree topology with random initiations. After a certain number of iterations (40,000 in our algorithm), which we call one epoch, the worker threads are suspended and their current posteriors are ranked. Then, from the top ranked chains, we apply **Species-Tree-SPR** to create new species tree topologies to replace the lower ranked chains and proceed to the next epoch. The newly created species tree topologies should not be identical with existing topologies and are awarded a small amount of extra iterations in the first several epochs. Unlike in [25] where the locus trees and gene trees are optimized iteratively, in each chain, we propose changes on all model parameters except for the species tree topology altogether each with a fixed probability. We accept each new state according to how good the new state is relative to the old state (similar to the Metropolis-Hastings algorithm [7]). We use a coefficient to flatten the posterior space when the moves are to change the locus tree topologies to achieve higher acceptance probability and help jump out of local optima (similar to Metropolis coupling and the use of temperate in simulated annealing). Gradually, we reduce the number of chains as confidence that the top chains may have the best topologies increases.

3 RESULTS

3.1 Simulations

To generate the sequence data for multiple gene families under a duplication-loss-coalescence model, we wrote a gene tree simulator

that operates in two phases. In phase I, the simulator uses a given species tree (topology and divergence times in generations of its nodes) and specified duplication and loss rates to generate a locus tree. In phase II, the simulator uses the locus tree along with a specified population size to generate gene trees under the multispecies coalescent. Once the gene trees are generated, the program Seq-Gen [15] was used to simulate the evolution of DNA sequences down the gene trees under a specified model of evolution. In all simulations reported here, we used the Jukes-Cantor model of evolution [9] and set the sequence length to 1,000 sites. For experiments 1, 2, and 3 below, we used the tree of Fig. 7 as the model species tree. The species were assumed to be diploid, and specified duplication/loss rates and population size were assumed to be the same across all branches of the model tree.

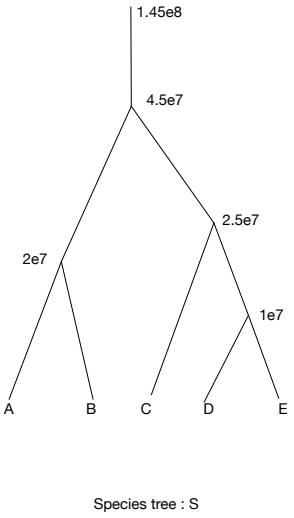

Figure 7: The model tree used to simulate data for experiments 1, 2, and 3. The values associated with the nodes correspond to divergence times in terms of number of generations.

3.1.1 Experiment 1: Testing the effects of duplication/loss rates and population sizes. In this experiment, we simulated data under different settings of duplication/loss rates and population sizes to test how these different parameters would affect the accuracy of inferences. The duplication and loss rates (both were equal) used were 10^{-10}, 10^{-9}, 5×10^{-9}, and 10^{-8} and the population sizes were 5×10^5, 10^6, 5×10^6, and 10^7. For each of the 16 different settings of duplication/loss rates and population size, we generated 5 replica each with 64 gene families. For each of the 80 64-locus data sets, we ran the aforementioned heuristic for 400 epochs to produce the species/locus/gene trees. To assess the accuracy of these trees, we calculated the Robinson-Foulds (RF) distances [19] between the estimated species trees, locus trees and gene trees and their corresponding true trees. The results are shown in Fig. 8. As the results show, the method has very good accuracy (indicated by RF distances close to 0), though the accuracy suffers as the duplication/loss rates and/or the population size get larger. When the duplication/loss rates are high, the relationships between the gene trees and the species tree becomes more complex due to extensive incongruence. Furthermore, when the population size is large, incomplete lineage sorting becomes more extensive, further adding to the hardness of estimating trees from the data set. Indeed, as Fig. 9 shows, the posteriors of the true trees are always better than those of the ones found by our method within the allotted number of search epochs (all values in the figure are smaller than or equal to 0). These results indicate that accurate inferences from "hard" data sets (ones with high duplication/loss rates and/or large population sizes) require long runs of the search heuristic.

Finally, we assessed the method's performance in terms of estimating the duplication and loss rates. The results are shown in Fig. 10. As the results show, the method performs well at estimating the two rates, though it has a poor performance for the lowest duplication/loss rates. One explanation for this is that the true data set has very little signal for the duplication and loss at these low rates (e.g., gene families could mostly consist of single copies of each gene in each species).

3.1.2 Experiment 2: Testing the effect of the number of gene families. A desired property of an inference method is that its accuracy improves as more data (more gene families in our case) are used. To study whether our method has this property, we varied the number of gene families (8, 16, 32, and 64) under one setting of a relatively high duplication/loss rate (8×10^{-9}) and a large population size (8×10^6). Five replica for each number of gene families were simulated and we ran the inference method for 300 epochs. Fig. 11 shows the RF distances between the inferred trees and the true ones as the number of gene families used in the inference increases. As the results show, the accuracy of the species tree improves significantly as the number of gene families increases from 8 to 64. The accuracy of the locus tree improves as well, but not at the same rate as that of the species tree. The gene trees have very good accuracy even when only 8 gene families are used, and not much improvement is seen across the different settings.

3.1.3 Experiment 3: Inference under a duplication-loss-coalescent model vs. a coalescent-only model. In this experiment we set out to test how a method that accounts only for incomplete lineage sorting but ignores duplication and loss would perform as compared to our method here. To achieve this, we ran the Bayesian MCMC species tree inference method (the `mcmc_seq` command, with the maximum number of reticulations set to 0 so as to sample the posterior of trees only) in PhyloNet [23] which implements the method of [22]. We simulated 5 replica under duplication and loss rates of 6×10^{-9} and population size 10^6 and 96 gene families for each data set. To run the coalescent-only method, for each gene family we randomly selected one gene copy for each species if there was at least one (as a result, between 50.1% and 56.2% of the sequences in the gene families were used as a result of this pruning of the data sets). We fed the sequences to both methods and ran our method for 600 epochs (24million iterations) and the coalescent-only method in PhyloNet for 24 million iterations. Both methods were able to identify the species tree topology in all five replica. Fig. 12 shows the divergence time estimates obtained by the two methods for the four internal nodes of the model species tree. As the results show, even when at most a single copy of a gene is present in each gene

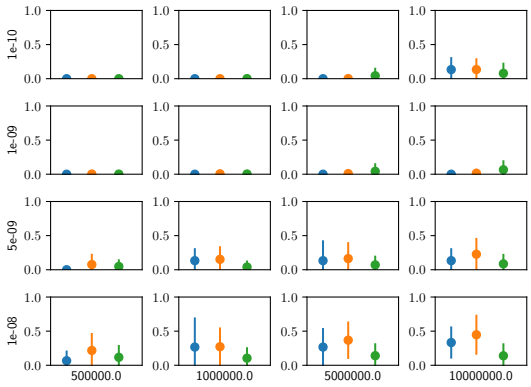

Figure 8: The average RF distances between the estimated tree and its corresponding true tree. Each row of panels corresponds to one setting of duplication/loss rate (shown to the left of the row), and each column of panels corresponds to a population size value (shown at the bottom of the column). RF values for the species trees, locus trees, and gene trees are shown in blue, orange, and green, respectively.

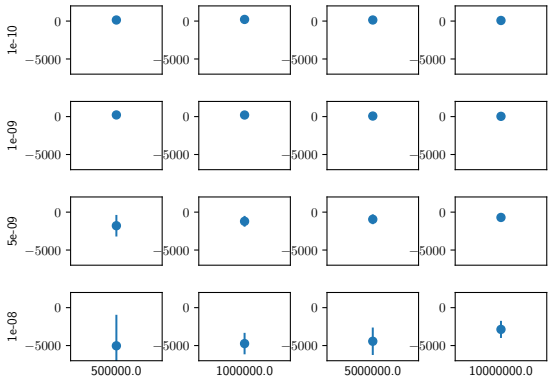

Figure 9: Average difference between the log **posterior of the inferred trees and the true trees, with standard deviation in vertical bars for each setting. Rows correspond to the shown duplication/loss rates and columns correspond to the show population sizes.**

family, a method that accounts for duplication and loss, in addition to coalescence, performs better than a method that accounts only for coalescence events. This is because our method handles hidden paralogy properly by counting some copies as paralogs, rather than orthologs as a coalescent-only method would do. In particular, our method found that in the five replica 19.7%, 6.25%, 10.4%, 13.5%, and 8.33% of the gene families had at least one duplications identified (again, despite the fact that no gene family in the input had more than a single gene copy in any of the five species).

3.1.4 Experiment 4: A slightly larger data set. In this experiment, we used the 7-taxon model species tree shown in Fig. 13 and used duplication/loss rates of 10^{-9} and population size of 2×10^6 (diploid genomes were assumed). While the data set has only two more taxa as compared to the previous one, the number of possible species tree topologies is now 10,395 compared to 105 for the case of 5 species. Furthermore, the numbers of locus trees and gene trees are even much larger. We simulated 5 replica each with 128 gene families and ran our method on these data sets. We report the RF distances for the inferred trees in Fig. 14. As the results show, the method performs very well at estimating all three trees, with that performance improving significantly within the first 60 epochs.

3.2 Biological Data

We used the yeast genome data set with duplications reported in [5] for 200 epochs with 5×10^{-9} as duplication and loss rate and 2×10^{-10} as mutation rate and 10^7 as population size, which

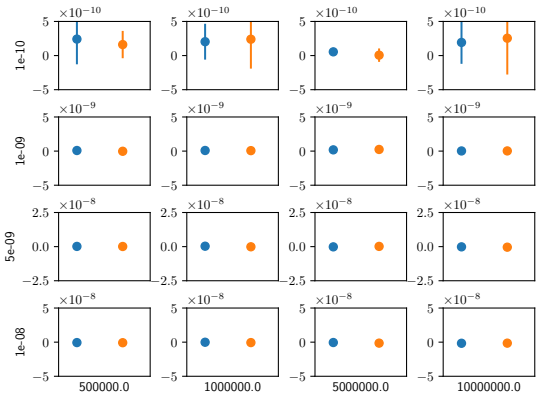

Figure 10: The average difference of estimated duplication (blue) and loss (orange) rates from true ones for each setting. Rows correspond to the shown duplication/loss rates and columns correspond to the show population sizes. Standard deviation is represented as vertical bar.

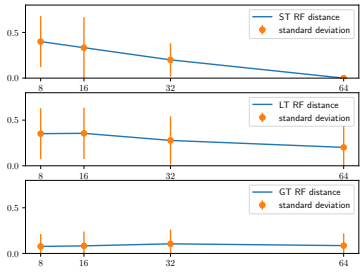

Figure 11: The average RF distances between the inferred and true trees. Top: Species tree error. Middle: Locus tree error. Bottom: Gene tree error. The number of gene families used as input to the inference method is shown on the x-axis.

are comparable with the settings used in [17]. Given that the full data set is very large and computationally too demanding for our method, we selected six genomes and randomly sampled 200 gene families rather than using all 706 gene families. In total, 1,397 DNA sequences were used. The tree obtained by our method is shown in Fig. 15. The tree that our method produced is consistent with that used in [5, 17].

4 DISCUSSION

Species phylogenies (trees and networks alike) play an important role in biology. They are the main framework with which trait evolution is understood. They guide genome sequencing and conservation efforts. And they help us understand the evolutionary processes that acted on and diversified the genomes of all species on Earth. Advances in sequencing technologies have enables biologists to collect genome-wide data from species of interest and, consequently, are promising to provide more power for more accurate inference of species phylogenies. However, with this power comes complexity. In particular, accounting for the various evolutionary processes that act on different regions of the genomes is a very challenging task. In this paper we focused on the task of species tree inference from sequence data under a model that accounts simultaneously for gene duplication and loss events, as well as coalescence effects (mainly incomplete lineage sorting).

We developed a hill-climbing heuristic for obtaining MAP estimates of the species tree, locus trees, and gene trees from DNA sequences of independent gene families. To the best of our knowledge, this is the first method to incorporate gene duplication, loss and ILS and search for species tree topology at the same time, as existing methods in this domain either do inference under a duplication/loss model or an ILS model, or infer reconciliation scenarios assuming a given species tree. One significant advantage of a method that incorporates all these processes is that it does not require an a priori orthology assignment as the orthology assignment can be either integrated out (summation over all possible orthology assignments is taken) or sampled (the posterior distribution is not marginalized over the orthology assignments; rather, the orthology assignment is sampled and reported along with the other parameters) in this framework.

Our simulation studies demonstrated a good performance of the method in terms of the accuracy of inferences it makes. Furthermore, we showed that when hidden paralogy could be at play (e.g., in cases where single-copy genes are mistakenly assumed to be orthologs by choice of the data), this simultaneous accounting for duplication, loss, and ILS results in better estimates of evolutionary parameters. Our analysis of a biological data set provided results in agreement with the species tree in the literature.

Our simulations are, by choice, conducted on small data sets owing mainly to the prohibitive computational complexity of the method, in particular that of computing the likelihood of a point

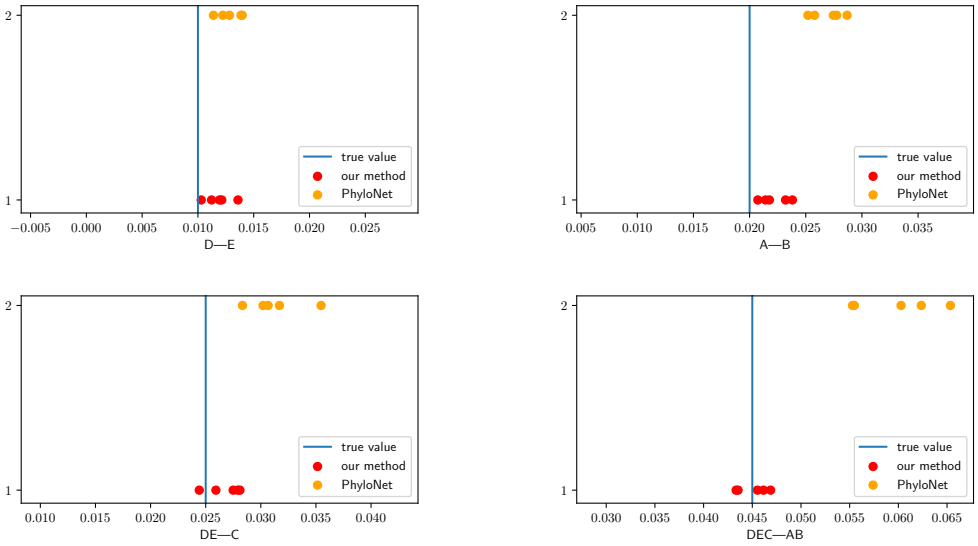

Figure 12: The divergence times in population mutation rates for the ancestors of D and E (D|E), of A and B (A|B), of D, E, and C (DE|C), and for the root (DEC|AB) estimated by our method and the species tree inference method in PhyloNet. Vertical bars are the true divergence times.

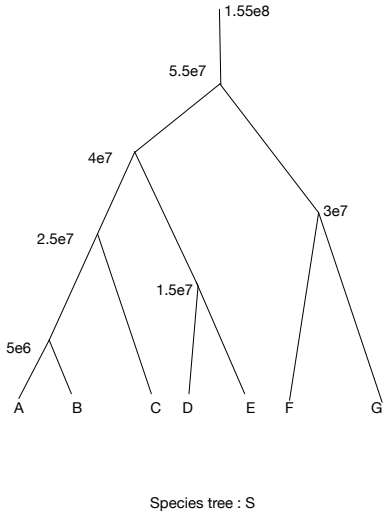

Figure 13: A 7-taxon species tree for experiment 4. The values associated with the nodes correspond to divergence times in terms of number of generations.

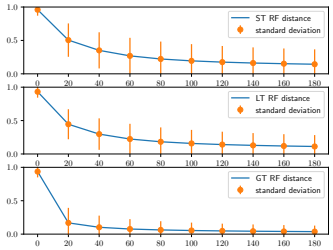

Figure 14: The average RF distances between the inferred and true trees. Top: Species tree accuracy. Middle: Locus tree accuracy. Bottom: Gene tree accuracy. The number of epochs that the method ran is shown on the x-axis. Standard deviation is represented as vertical bar.

in the parameter space. For example, analysis of each of the aforementioned data sets took about 15 to 20 hours. Taming the complexity of these computations is a major task to pursue for this method to scale up to larger data sets. Furthermore, replacing the hill-climbing search for the MAP estimate by a principled Markov chain Monte Carlo technique for sampling the posterior distribution of the parameter space would be preferable, as it provides estimates of confidence in the inference. Finally, a major direction for future research is to extend the model so that it incorporates reticulation in addition to duplication, loss, and incomplete lineage sorting.

ACKNOWLEDGMENTS

This work has been supported by grants CCF-1514177, DBI-1355998 and CCF-1302179 from the National Science Foundation.

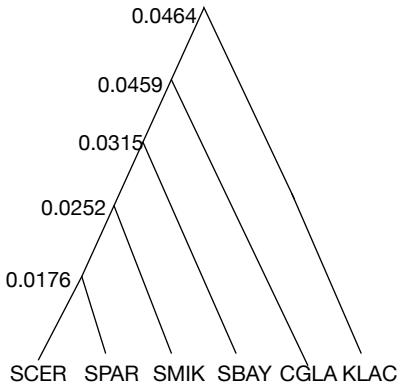

Figure 15: The species tree inferred by our method on the 200 gene families from the six yeast species. Divergence times shown are in units of population mutation rate.

REFERENCES

[1] Örjan Åkerborg, Bengt Sennblad, Lars Arvestad, and Jens Lagergren. 2009. Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proceedings of the National Academy of Sciences* 106, 14 (2009), 5714–5719.

[2] Lars Arvestad, Jens Lagergren, and Bengt Sennblad. 2009. The gene evolution model and computing its associated probabilities. *Journal of the ACM (JACM)* 56, 2 (2009), 7.

[3] John C Avise and Terence J Robinson. 2008. Hemiplasy: a new term in the lexicon of phylogenetics. *Systematic Biology* 57, 3 (2008), 503–507.

[4] Bastien Boussau, Gergely J Szöllősi, Laurent Duret, Manolo Gouy, Eric Tannier, and Vincent Daubin. 2013. Genome-scale coestimation of species and gene trees. *Genome research* 23, 2 (2013), 323–330.

[5] Geraldine Butler, Matthew D Rasmussen, Michael F Lin, Manuel AS Santos, Sharadha Sakthikumar, Carol A Munro, Esther Rheinbay, Manfred Grabherr, Anja Forche, Jennifer L Reedy, et al. 2009. Evolution of pathogenicity and sexual reproduction in eight Candida genomes. *Nature* 459, 7247 (2009), 657.

[6] James H Degnan and Noah A Rosenberg. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in ecology & evolution* 24, 6 (2009), 332–340.

[7] W Keith Hastings. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57, 1 (1970), 97–109.

[8] Joseph Heled and Alexei J Drummond. 2009. Bayesian inference of species trees from multilocus data. *Molecular biology and evolution* 27, 3 (2009), 570–580.

[9] Thomas H Jukes, Charles R Cantor, et al. 1969. Evolution of protein molecules. *Mammalian protein metabolism* 3, 21 (1969), 132.

[10] Laura S Kubatko, Bryan C Carstens, and L Lacey Knowles. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics* 25, 7 (2009), 971–973.

[11] Liang Liu, Lili Yu, and Scott V Edwards. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC evolutionary biology* 10, 1 (2010), 302.

[12] Wayne P Maddison. 1997. Gene trees in species trees. *Systematic biology* 46, 3 (1997), 523–536.

[13] Siavash Mirarab, Rezwana Reaz, Md S Bayzid, Théo Zimmermann, M Shel Swenson, and Tandy Warnow. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* 30, 17 (2014), i541–i548.

[14] Siavash Mirarab and Tandy Warnow. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* 31, 12 (2015), i44–i52.

[15] Andrew Rambaut and Nicholas C Grass. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics* 13, 3 (1997), 235–238.

[16] Matthew D Rasmussen and Manolis Kellis. 2007. Accurate gene-tree reconstruction by learning gene-and species-specific substitution rates across multiple complete genomes. *Genome research* 17, 12 (2007), 1932–1942.

[17] Matthew D Rasmussen and Manolis Kellis. 2011. A Bayesian approach for fast and accurate gene tree reconstruction. *Molecular Biology and Evolution* 28, 1 (2011), 273–290.

[18] Matthew D Rasmussen and Manolis Kellis. 2012. Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Research* 22, 4 (2012), 755–765.

[19] David F Robinson and Leslie R Foulds. 1981. Comparison of phylogenetic trees. *Mathematical biosciences* 53, 1-2 (1981), 131–147.

[20] Joel Sjöstrand, Bengt Sennblad, Lars Arvestad, and Jens Lagergren. 2012. DLRS: gene tree evolution in light of a species tree. *Bioinformatics* 28, 22 (2012), 2994–2995.

[21] Cuong Than, Derek Ruths, and Luay Nakhleh. 2008. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC bioinformatics* 9, 1 (2008), 1.

[22] Dingqiao Wen and Luay Nakhleh. 2018. Co-estimating Reticulate Phylogenies and Gene Trees from Multi-locus Sequence Data. *Systematic Biology* 67, 3 (2018), 439–457.

[23] Dingqiao Wen, Yun Yun, Jiafan Zhu, and Luay Nakhleh. 2018. Inferring Phylogenetic Networks Using PhyloNet. *Systematic Biology* 67, 4 (2018), 735–740.

[24] Yufeng Wu. 2012. Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. *Evolution* 66, 3 (2012), 763–775.

[25] Bo Zhang and Yi-Chieh Wu. 2017. Coestimation of Gene Trees and Reconciliations Under a Duplication-Loss-Coalescence Model. In *International Symposium on Bioinformatics Research and Applications*. Springer, 196–210.