# Effect of Incomplete Lineage Sorting on Tree-Reconciliation-Based Inference of Gene Duplication

Yu Zheng and Louxin Zhang

Department of Mathematics, National University of Singapore,
10 Lower Kent Ridge, Singapore 119076

**Abstract.** Incomplete lineage sorting (ILS) gives rise to stochastic variation in the topology of a gene tree and hence introduces false duplication events when gene tree and species tree reconciliation method is used for inferring the duplication history of a gene family. We quantify the effect of ILS on inference of gene duplication by examining the expected number of false duplication events inferred from reconciling a random gene tree, which occurs with a probability predicted in coalescent theory, and the given species tree. We computationally analyze the relationships between the number of false duplication events inferred on a branch and its length in a species tree, and the relationships between the expected number of false duplication events in a species tree and its topological parameters. This study provides evidence that inference of gene duplication based on tree reconciliation was affected by ILS to a greater extent on an asymmetric species tree than on a symmetric one. Our findings also suggest that the bias caused by ILS in reconciliation-based inference of gene duplication might not be negligible. Hence, when gene duplication is inferred via tree reconciliation or any other method that takes gene tree topology into account, the ILS-induced bias should be examined cautiously.

## 1 Background

A gene tree is the phylogenetic tree of a family of homologous genes. A species tree is the phylogenetic tree of a collection of species. In population genomics and phylogenetics, it is important to distinguish gene trees and species trees, as a gene tree reconstructed from the DNA sequences of the given gene family is sometimes discordant with the species tree that contains it [31,33]. The incongruence of gene trees and species trees can be caused by gene duplication and loss, horizontal gene transfer, hybridization, or incomplete lineage sorting (ILS) [17,21,29]. Accordingly, the relationships between gene trees and species trees have been the focus of many studies over the past two decades [10,18]. Gene trees have been used to estimate the species trees [8,12,19,20,22], to estimate species divergence time [4] and ancestral population size [11,16,34], and to infer the history of gene duplication [1,3,5,6,13,23,32].

Z. Cai et al. (Eds.): ISBRA 2013, LNBI 7875, pp. 261–272, 2013.

One popular approach for gene duplication inference is gene tree and species tree reconciliation. It is formalized from the following fact: If the descendants of a node in a gene tree are distributed in the same set of species as one of its children, then the node corresponds to a gene duplication event [12,23]. Clearly, this approach takes gene tree topology into account. If incorrect gene trees are used, duplication events are often mis-inferred [14].

In a species tree, each internodal branch represents an ancestral population; each internal node represents a time point at which the ancestral population split into two subpopulations. It is assumed that there was no gene flow between the subpopulations after split. When the population of each species is large, the DNA sequences sampled from two species are more unlikely to have their common sequence ancestor living at the moment that the most recent common ancestor (MRCA) of the two species split; instead, the time back to the common sequence ancestor is uncertain and typically longer than the time back to the MRCA of the species. This evolutionary phenomenon is ILS or deep coalescence. Clearly, ILS gives rise to considerable stochastic variations in gene tree topology [27,31], implying that different unlinked loci might have different genealogical histories, and different samplings might also lead to different gene tree topologies for the same gene. Consider the two different gene tree topologies in Fig. 1. The gene topology in red is concordant to the species tree; reconciling this gene tree and the species tree does not infer any gene duplication events, whereas reconciliation with the gene tree in green gives one (false) duplication event. Hence, ILS affects gene duplication inference. To the best of our knowledge, the effects of ILS on gene duplication inference has not been examined quantitatively although they have been noticed for long time (see [21] for example).

The present paper examines quantitatively the effect of ILS on inference of gene duplication. Here, we assume that no genetic exchange has occurred between unrelated species and there is no sequence error to facilitate our quantitative study. Notice that the effects of horizontal gene transfer and hybridization events on gene duplication inference have been studied by proposing general evolutionary models to coordinate these events or by computational simulation [2,9,36].

## 2 Results and Discussion

In our study, we shall consider only gene trees over single-gene families. In other words, we assume only one gene is sampled from each species. Under such an assumption, any inferred gene duplication event is a false one, and the gene tree distribution can be computed using coalescent theory [7,27]. Accordingly, the assumption greatly simplifies our discussion and allows us to find out crucial connections between the effect of ILS and species tree topologies.

When calculating the probability that a gene tree is seen in the corresponding species tree, we consider a simple coalescent model each species has a constant diploid effective population size $N$ during its entire existence and evolutionary time of $t$ generations equals $T = t/(2N)$ coalescent time units [31].

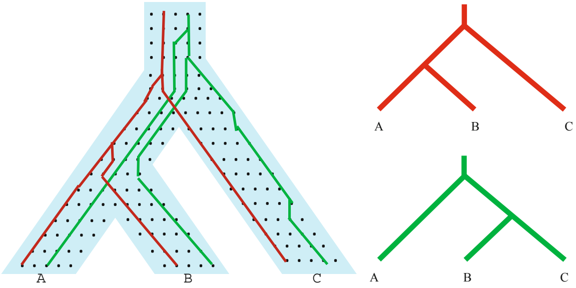

**Fig. 1. Schematic view of two different coalescent histories in a species tree.** The species tree (light blue) of three species is given in the left panel. DNA sequences sampled from different individuals within a species may give different collapsed gene trees (red and green, right panel) for a gene family. If the green gene tree is used, a gene duplication event is inferred in the branch entering the species tree root and a gene loss event is inferred in the lineage leading to $C$.

### 2.1 Measuring the Effect of ILS on Gene Duplication Inference

Consider a single-gene family $\mathcal{F}$ sampled from a set $X$ of species. Let $S$ be the phylogeny over $X$. If no gene duplication occurred to the gene family during the evolution of the species, the tree of the gene family has the same topology as $S$. If ILS events have occurred, however, the gene tree reconstructed from the gene sequences might be different from $S$. To quantify the effect of ILS on inference of gene duplication for a gene family on $S$, we use the expected number $D(S)$ (or $L(S)$) of false gene duplication (or loss) events output from the lca reconciliation of a random gene tree and $S$. For a gene tree $G$, we use $c_{\text{dup}}(G, S)$ (or $c_{\text{ loss}}(G, S)$) to denote the gene duplication (or loss) cost of the lca reconciliation between $G$ and $S$ (Materials and Methods). Since $c_{\text{dup}}(G, S) = c_{\text{ loss}}(G, S) = 0$ if $G = S$. $D(S)$ and $L(S)$ are simply:

$$D(S) = \sum_{G \in \mathcal{G}} c_{\text{dup}}(G, S)\Pr[G \mid S], \tag{1}$$

$$L(S) = \sum_{G \in \mathcal{G}} c_{\text{ loss}}(G, S)\Pr[G \mid S], \tag{2}$$

where $\mathcal{G}$ is the set of all possible gene tree topologies, and $\Pr[G \mid S]$ the probability that $G$ is the collapsed gene tree of a coalescent history of the sampled genes from the species belonging to $X$ (see Fig. 1 for an illustration of a coalescent history and its collapsed gene tree).

Let $\mathcal{H}(G)$ be the set of all possible coalescent histories that give the gene tree $G$. For each $H \in \mathcal{H}(G)$, we use $\Pr[H \mid S]$ to denote the probability that $H$ occurs in $S$. By definition, we compute $\Pr[G \mid S]$ by:

$$\Pr[G \mid S] = \sum_{H \in \mathcal{H}(G)} \Pr[H \mid S], \tag{3}$$

where $\Pr[H \mid S]$ can be computed efficiently given $H$ and $S$ [7,35].

### 2.2 The Case of Four Species

In the case of four species, there are only two different topologies

$$S_1 = ((A,B) : t_1, (C,D) : t_2),$$
$$S_2 = (((A,B) : \tau_1, C) : \tau_2, D),$$

in Newick phylogeny format (Fig. 2). For the sake of brevity, we use $p(x,y)$ to denote the parental node of two siblings $x$ and $y$ in each of these two species trees. In $S_1$, the evolutionary time of $p(A,B)$ is $t_1$ generations, whereas that of $p(C,D)$ is $t_2$ generations. Let $G$ be an arbitrary gene tree of a gene family. Consider the lca reconciliation between $G$ and $S_1$. Any gene tree node is mapped to $p(A,B)$ if and only if its two children are mapped to $A$ and $B$ respectively (Materials and Methods). This fact also holds for $p(C,D)$. Therefore, false duplication events can only be inferred on the branch entering the root in $S_1$. Set $T_i = t_i/(2N)$ for $i = 1,2$. By calculating the distribution of the gene trees [24,27], we obtain:

$$D(S_1) = \frac{2}{3}(e^{-T_1} + e^{-T_2}), \tag{4}$$

$$L(S_1) = 2(e^{-T_1} + e^{-T_2}) + \frac{2}{9}e^{-(T_1+T_2)} \tag{5}$$

from Eqn. (1) and (2).

Now, we switch to consider $S_2$. Setting $\bar{T}_i = \tau_i/(2N)$ for $i = 1,2$, we have:

$$D(S_2) = \frac{2}{3}\left(e^{-\bar{T}_1} + e^{-\bar{T}_2}\right) - \frac{1}{3}e^{-(\bar{T}_1+\bar{T}_2)} + \frac{5}{18}e^{-(\bar{T}_1+3\bar{T}_2)}, \tag{6}$$

$$L(S_2) = 2\left(e^{-\bar{T}_1} + e^{-\bar{T}_2}\right) - \frac{1}{3}e^{-(\bar{T}_1+\bar{T}_2)} + \frac{5}{6}e^{-(\bar{T}_1+3\bar{T}_2)}. \tag{7}$$

Since $e^{-\mathrm{x}} < 1$ for any $x > 0$, we have:

$$D(S_1) < 1\frac{1}{3} \text{ and } L(S_1) < 4\frac{2}{9};$$
$$D(S_2) < 1\frac{5}{18} \text{ and } L(S_2) < 4\frac{1}{2}.$$

If all branches have equal length (i.e. $T_1 = T_2 = \bar{T}_1 = \bar{T}_2 = T$), $D(S_1) \geq D(S_2)$ for any $T$. However, $L(S_1) \geq L(S_2)$ only if $T \geq \frac{1}{2}\ln(3/2)$.

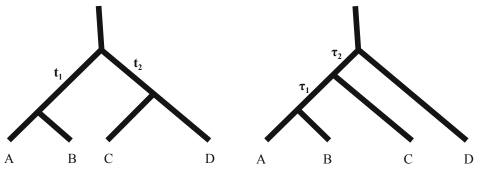

**Fig. 2.** Two topologies $S_1$ (left) and $S_2$ (right) of the species trees of 4 species

Assume $S_1$ and $S_2$ are ultrametric and have the same height. We further assume that $T_1 = T_2 = 2T$ and $\bar{T}_1 = \bar{T}_2 = T$, implying that $\tau_1 + \tau_2 = t_1 = t_2$ and that $A$ and $B$ diverged at the same time in both trees. Then,

$$D(S_1) = \frac{4}{3}e^{-2\mathrm{T}} \text{ and } L(S_1) = 4e^{-2\mathrm{T}} + \frac{2}{9}e^{-4\mathrm{T}};$$

$$D(S_2) = \frac{4}{3}e^{-\mathrm{T}} - \frac{1}{3}e^{-2\mathrm{T}} + \frac{5}{18}e^{-4\mathrm{T}} \text{ and } L(S_2) = 4e^{-\mathrm{T}} - \frac{1}{3}e^{-2\mathrm{T}} + \frac{5}{6}e^{-4\mathrm{T}}.$$

Using numerical computation, we obtained $D(S_1) < D(S_2)$ only if $T > 0.0649$, but $L(S_1) < L(S_2)$ for any $T$. This analysis suggests that the effect of ILS is closely related to species tree structure.

## 2.3 Effect Analysis on a *Drosophila* Species Tree

Genome-wide analysis provides strong evidence for the prevalence of ILS events in *Drosophila* evolution [25]. Here, we examined the expected number of false duplication events caused by ILS in the phylogeny of 12 *Drosophila* species [15], in which evolutionary time is dated for all branches. Since the effective population size $N$ for the *Drosophila* species is unknown, we considered four different effective population sizes ($2 \times 10^6, 6 \times 10^6, 10 \times 10^6$, and $14 \times 10^6$) and set the generation time to be 1/10 years [25]. The expected numbers of false duplication events caused by ILS for different effective population sizes are plotted (Fig. 3). Here, we point out that our conclusion does not depend on the specific effective population sizes we used.

Since only one gene is sampled from the population of each species, no gene duplication is inferred on branches connecting to the leaves. In other words,

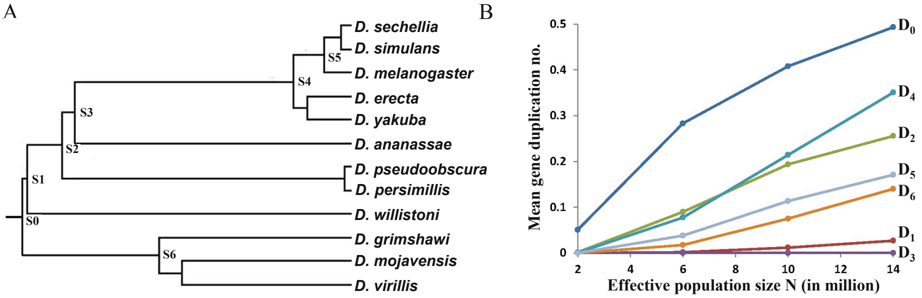

**Fig. 3. Effect analysis for a *Drosophila* species tree.** (A). A tree of 12 *Drosophila* species given in [15]. All the branches are drawn in proportion to evolutionary time. False duplication events caused by ILS can only be inferred on seven branches that enter S0–S6 respectively for single gene families. (B). The expected number $D_i$ of false gene duplication events on the branch entering S$i$ is plotted against the effective population size $N$, with the generation time being set to $\frac{1}{10}$ years. Four different effective population sizes ($2 \times 10^6$, $6 \times 10^6$, $10 \times 10^6$ and $14 \times 10^6$) were examined. It shows that the number of false gene duplication events on a branch correlates largely with its evolutionary time.

false gene duplication events can only be inferred on the seven branches that are denoted by their end nodes S$i$ ($0 \leq i \leq 6$) (Fig. 3A). The expected total number of false gene duplication events in the tree can range from 0.0534 to 1.7663 for each of the selected effective population sizes. Let $D_i$ be the expected number of false gene duplication events on the branch S$i$ for each $i$. Although the exact values of these $D_i$ are different for the different effective population sizes, their relative ranks remain almost the same, correlating well with the branches' evolutionary time. For instance, $D_0$ has the largest value for each effective population size. This is because the branch entering the root is assumed to be long enough that all the lineages coexisting at the moment that the MRCA of all the extant species split will coalesce on it. For the longest branch entering S4, $D_4$ is the second largest for effective population sizes of 10 and 14 million, and the third largest for other sizes.

Another finding is that on the branches close to the root, the expected number of false gene duplication events is relatively large. For example, for the shortest branch S1, $D_1$ is not the smallest; instead, it is larger than $D_3$, probably due to the closeness of $S_1$ to the tree root. Similarly, branch S6 is longer than S2, but $D_6$ is smaller than $D_2$ for each effective population size because S6 is closer to the tree root.

We now switch to 6698 gene trees in the *Drosophila* species tree [14]. We inferred gene duplication events for the corresponding gene families by reconciling the gene trees and the species tree (Fig. 3A). In total, we inferred 10,264 gene duplication events that are distributed on the seven branches as: 1.8% (S3), 6.5% (S0), 7.4% (S2), 8.0% (S5), 15.1% (S1), 20.5% (S6) and 40.6% (S4). Such a distribution is not quite consistent with the computational analysis presented above. The proportion of inferred duplication events on the branches entering S0 and S2 is significantly lower than what the analysis suggests, whereas those on branches entering S1 and S6 are much higher. Possible reasons for this are either because sequence sampling and alignment errors influenced gene tree reconstruction, leading to incorrect topology for some gene trees, or because effective population size varies for different ancestral species. At this stage, we are unable to assess the effect of these factors, as the estimation of ancestral population sizes remains as a challenging problem.

## 2.4 The Upper Bound of $D(S)$ and $L(S)$

To interrogate the impact of species tree topology on the effect of ILS for inference of gene duplication, we considered 10 ultrametric tree topologies over 10 species (Fig. 4). In each of the 5 asymmetric species trees, the two subtrees rooted at the children of the root are linear trees. In each of the 5 symmetric trees, the subtrees are balanced binary trees instead.

We define the height of a ultrametric species tree to be the coalescent time of a path from the root to a leaf, measured in coalescent time units. $D(S)$ and $L(S)$ for these 10 topologies with heights of 2 and 10 units are respectively presented in two panels in Fig. 4. Although each path from the root to a leaf has the same evolutionary time, the number of branches contained in each path

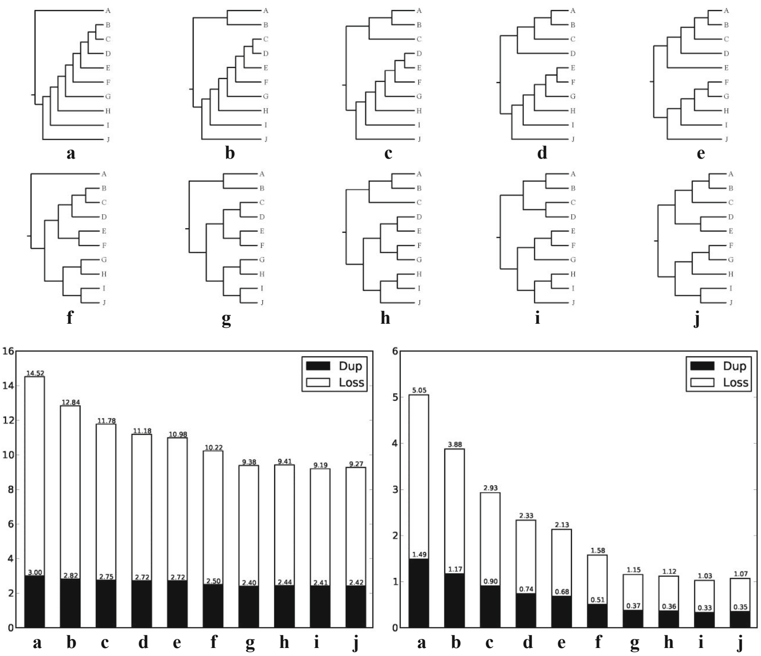

**Fig. 4.** **$D(S)$ and $L(S)$ for asymmetric topologies (first row) and symmetric topologies (second row)**. Branches in each of 10 ultrametric topologies are drawn in proportion to their length. In the bottom row, the left and right plots are drawn to different scales for the topologies of heights 2 and 10, respectively. In each plot, the white and black bars represent $L(S)$ and $D(S)$, respectively.

varies. For each leaf, we define its depth to be the number of branches in the unique path from it to the root. Although each path from the root to a leaf has the same evolutionary time, different leaves may have different depths. The Sackin index of a species tree is defined as the average depth of a leaf in the tree [28]. The ten tree topologies listed in the figure have the following Sackin indexes:

| Tree | a | b | c | d | e | f | g | h | i | j |
|---|---|---|---|---|---|---|---|---|---|---|
| Sackin index | 5.4 | 4.7 | 4.2 | 3.9 | 3.8 | 3.9 | 3.8 | 3.5 | 3.4 | 3.4 |

Hence, our experiments suggest that:

- $D(S)$ and $L(S)$ increase with the Sackin index of a species tree $S$;
- Asymmetric trees have a larger $D(S)$ and $L(S)$ than symmetric ones of the same height.

In [7], the authors studied the probability distribution of all the gene trees in a species tree over 5 species. Since our study focuses the mean duplication and gene loss costs of a gene tree defined in (1) and (2), the facts reported here are not direct consequences of those reported in [7].

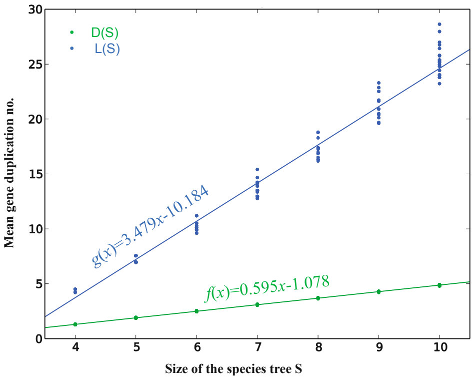

**Fig. 5. Regression of $D(S)$ and $L(S)$.** Given the size of $S$, $D(S)$ varies with the topology of $S$ in a narrow range, whereas $L(S)$ varies in a wide range. For each size, we generated 20 random species trees in the Yule model.

It is natural to ask to what extent ILS influences gene duplication inference. To answer this question, we compute the limit of $D(S)$ and $L(S)$ for an arbitrary ultrametric species tree $S$ by allowing the branches of $S$ to be extremely short. Fix the effective population size for each branch of $S$. When all branches of $S$ become very short, two lineages are unlikely to coalesce in any branch below the tree root; in other words, there is a high probability that any pair of lineages will coalesce in the branch entering the root. Therefore, in the limit case, for each gene tree $G$:

$$\Pr[G \mid S] \sim \sum_{H \in \mathcal{H}'(G)} \Pr[H \mid S],$$

where $\mathcal{H}'(G)$ is the set of the coalescent histories of $n$ lineages whose collapsed gene tree is $G$ in the root branch. Based on this fact, we computed the limit of $D(S)$ and $L(S)$ for 20 random species tree for each size (i.e. the number of species) from 4 to 10 (Fig. 5). We found that $D(S)$ varies in a narrow range for each tree size and linearly increases with species tree size. However, $L(S)$ changes in a different manner. First, $L(S)$ varies in a wide range for a fixed species tree size. Secondly, although $L(S)$ also fits a linear function for the tree size in the range of 4 to 10, it remains unclear if it grows linearly or not because of its wide range for a fixed tree size.

## 3 Conclusion

ILS introduces stochastic variation into the topology of the gene tree of a gene family. For the first time, we have quantified the effect of ILS on gene duplication inference by examining the expected number of false gene duplication events inferred from reconciling a random gene tree and the species tree that contains it. In this preliminary study, we have also analyzed the connection between the topological parameters of the corresponding species tree and the effect of ILS on gene duplication inference.

One of our findings is that inference of gene duplication based on tree reconciliation was affected by ILS to a greater extent on an asymmetric species tree than on a symmetric one. Considering gene duplication events arising from ILS on different species tree branches separately, we also found that the longer an internodal branch is, the more likely gene duplication events are to be mis-inferred on it. Additionally, gene duplication events are more likely to be mis-inferred on a branch close to the species tree root.

In analyzing the limit of $D(S)$ and $L(S)$ for a species tree $S$ when its branches are extremely short, we found that $D(S)$ increases linearly with the species tree size in the range of 4 to 10. This fact indicates that $D(S)$ increases with $|S|$, the size of $S$ in general and hence it is not bounded above. It also raises a theoretical problem: $D(S) \leq 0.6|S|$? Since $L(S) \geq 3D(S)$ for a species tree [37], $L(S)$ is not bounded above by a constant if $D(S)$ is not. Our findings imply that the bias caused by ILS in reconciliation-based gene duplication inference is not negligible. Therefore, when gene duplication is inferred via tree reconciliation or any other method that takes gene tree topology into account, the ILS-induced bias should be examined cautiously. Alternatively, one may use a unified reconciliation approach that considers gene duplication, loss and ILS simultaneously [26,30].

Finally, we remark that ILS also affect the gene trees for genes which are from different genera. How much ILS is expected to affect gene duplication inference for gene families cross different genera is definitely a research topic for future study.

# 4 Material and Methods

## 4.1 Computing the Gene Tree Distribution in a Species Tree

The probability that a gene tree occurs in a given species tree is computed by Eqn. (3). For the purpose of computing the gene tree distribution in a species tree, COAL is too slow to be used, although it has many useful features [7]. Our analysis used a home-made computer program implemented in C. It speeds up computation via the dynamic programming technique, which had also been used by Wu in STELLS [35]. Presently, it allows us to examine the effect of ILS on gene duplication inference for species trees of up to 12 species. For the case of 12 species, one needs to consider about 13.7 billion gene trees for the analysis.

## 4.2 Gene Duplication Inference

Consider a collection $X$ of extant species. The species tree of the given species is a rooted tree in which each leaf uniquely represents (and hence is labeled by) an extant species. Here, we further assume species trees are fully binary and branch-weighted. Therefore, in a species tree, each non-leaf node has exactly two children; each internodal branch represents an ancestral species and has the evolutionary time of the ancestral species as its length.

For a gene family sampled from $X$, its gene tree is a rooted tree in which each leaf represents a gene and is labelled by the species where the gene is found. Since the gene family is assumed to have one gene sampled from each species, the gene tree is uniquely leaf-labelled in our study.

Let $S$ be the species tree of $X$ and let $G$ be a binary gene tree for a gene family $\mathcal{F}$ over $X$. For any two nodes $x, y$ of $S$, we use $\text{lca}(x, y)$ to denote the MRCA of $x$ and $y$ in $S$. The lca reconciliation $\mathcal{R}$ between $S$ and $G$ is a node-to-node mapping from $V(G)$ to $V(S)$ defined as:

$$\mathcal{R}(g) = \begin{cases} \text{the unique leaf in } S \text{ that has the same label as } g, & \text{if } g \text{ is a leaf;} \\ \text{lca}\left(\mathcal{R}(g_1), \mathcal{R}(g_2)\right), & \text{otherwise,} \end{cases}$$

for any gene tree node $g$, where $g_1$ and $g_2$ are the children of $g$.

The duplication history of $\mathcal{F}$ can be inferred through the lca reconciliation $\mathcal{R}$ [12,23]. For a non-leaf node $g$ of $G$, if $\mathcal{R}(c(g)) = \mathcal{R}(g)$ for some child $c(g)$ of $g$, then a duplication event is inferred in the branch entering $\mathcal{R}(g)$ in $S$.

The number of the gene duplication events inferred by using the lca reconciliation is denoted by $c_{\text{dup}}(G, S)$. All the inferred gene duplication events form a putative duplication history of $\mathcal{F}$ in which some genes might become lost. The number of gene loss events assumed in the gene duplication history is computed as follows.

For any two nodes $s$ and $t$ such that $s$ is below $t$ in $S$, we write $s \subset h \subset t$ to denote that $h$ is a node in the path from $t$ to $s$ for a node $h$. We define:

$$l(s, t) = |\{h \in S \mid s \subset h \subset t\}|.$$

Note that $l(s, t)$ is equal to the number of lineages off the evolutionary path from $t$ to $s$. For a non-leaf node $g$ with children $g_1$ and $g_2$ of $G$, define:

$$l(g) = \begin{cases} 0, & \text{if } \mathcal{R}(g) = \mathcal{R}(g_1) = \mathcal{R}(g_2), \\ l(\mathcal{R}(g_1), \mathcal{R}(g_2)) + 1, & \text{if } \mathcal{R}(l_g) \subset \mathcal{R}(g) = \mathcal{R}(r_g), \\ l(\mathcal{R}(g_1), \mathcal{R}(g)) + l(\mathcal{R}(g_2), \mathcal{R}(g)), & \text{if } \mathcal{R}(g_1) \subset \mathcal{R}(g) \supset \mathcal{R}(g_2). \end{cases}$$

The number of genes that have to be assumed to be lost in the inferred duplication history is equal to $\sum_{g \in G} l(g)$, denoted by $c_{\text{loss}}(G, S)$ and called the *gene loss cost* of the lca reconciliation between $G$ and $S$.

In this work, we used our computer program to compute the gene duplication and loss costs for a gene tree and a species tree [38].

**Acknowledgements.** The work was supported by Singapore AcRF Tier-2 Grant R-146-000-134-112.

## References

1. Åkerborg, Ö., Sennblad, B., Arvestad, L., Lagergren, J.: Simultaneous bayesian gene tree reconstruction and reconciliation analysis. Proc. Natl. Acad. Sci. U. S. A. 106(14), 5714–5719 (2009)
2. Bansal, M.S., Alm, E.J., Kellis, M.: Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. Bioinformatics 28(12), i283–i291 (2012)
3. Berglund-Sonnhammer, A.C., Steffansson, P., Betts, M.J., Liberles, D.A.: Optimal gene trees from sequences and species trees using a soft interpretation of parsimony. J. Mol. Evol. 63(2), 240–250 (2006)
4. Cann, R.L., Stoneking, M., Wilson, A.C.: Mitochondrial DNA and human evolution. Nature 325(6099), 31–36 (1987)
5. Chauve, C., El-Mabrouk, N.: New perspectives on gene family evolution: Losses in reconciliation and a link with supertrees. In: Batzoglou, S. (ed.) RECOMB 2009. LNCS, vol. 5541, pp. 46–58. Springer, Heidelberg (2009)
6. Chen, K., Durand, D., Farach-Colton, M.: Notung: a program for dating gene duplications and optimizing gene family trees. J. Comput. Biol. 7(3-4), 429–447 (2000)
7. Degnan, J.H., Salter, L.A.: Gene tree distributions under the coalescent process. Evolution 59(1), 24–37 (2005)
8. Doyle, J.J.: Gene trees and species trees: molecular systematics as one-character taxonomy. Syst. Botany 17, 144–163 (1992)
9. Doyon, J.-P., Scornavacca, C., Gorbunov, K.Y., Szöllősi, G.J., Ranwez, V., Berry, V.: An Efficient Algorithm for Gene/Species Trees Parsimonious Reconciliation with Losses, Duplications and Transfers. In: Tannier, E. (ed.) RECOMB-CG 2010. LNCS, vol. 6398, pp. 93–108. Springer, Heidelberg (2010)
10. Edwards, S.V.: Is a new and general theory of molecular systematics emerging? Evolution 63(1), 1–19 (2008)
11. Edwards, S.V., Beerli, P.: Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. Evolution 54(6), 1839–1854 (2000)
12. Goodman, M., Czelusniak, J., Moore, G.W., Romero-Herrera, A.E., Matsuda, G.: Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Syst. Biol. 28(2), 132–163 (1979)
13. Górecki, P., Tiuryn, J.: DLS-trees: a model of evolutionary scenarios. Theor. Comput. Sci. 359(1), 378–399 (2006)
14. Hahn, M.W.: Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. Genome Biol. 8(7), R141 (2007)
15. Hahn, M.W., Han, M.V., Han, S.G.: Gene family evolution across 12 *Drosophila* genomes. PLoS Genetics 3(11), e197 (2007)
16. Hey, J., Nielsen, R.: Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. Genetics 167(2), 747–760 (2004)
17. Keeling, P.J., Palmer, J.D.: Horizontal gene transfer in eukaryotic evolution. Nat. Rev. Genet. 9(8), 605–618 (2008)
18. Knowles, L.L., Kubatko, L.S.: Estimating Species Trees: Practical and Theoretical Aspects. Wiley-Blackwel, New Jersey (2010)
19. Liu, L., Yu, L., Kubatko, L., Pearl, D.K., Edwards, S.V.: Coalescent methods for estimating phylogenetic trees. Mol. Phylogenet. Evol. 53(1), 320–328 (2009)

20. Ma, B., Li, M., Zhang, L.X.: From gene trees to species trees. SIAM J. Comput. 30(3), 729–752 (2000)
21. Maddison, W.P.: Gene trees in species trees. Syst. Biol. 46(3), 523–536 (1997)
22. Maddison, W.P., Knowles, L.L.: Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55(1), 21–30 (2006)
23. Page, R.D.M.: Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. Syst. Biol. 43(1), 58–77 (1994)
24. Pamilo, P., Nei, M.: Relationships between gene trees and species trees. Mol. Biol. Evol. 5(5), 568–583 (1988)
25. Pollard, D.A., Iyer, V.N., Moses, A.M., Eisen, M.B.: Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. PLoS Genet. 2(10), e173 (2006)
26. Rasmussen, M.D., Kellis, M.: Unified modeling of gene duplication, loss, and coalescence using a locus tree. Genome Research 22(4), 755–765 (2012)
27. Rosenberg, N.A.: The probability of topological concordance of gene trees and species trees. Theor. Popul. Biol. 61(2), 225–247 (2002)
28. Sackin, M.J.: Good and bad phenograms. Syst. Zool. 21, 225–226 (1972)
29. Sang, T., Zhong, Y.: Testing hybridization hypotheses based on incongruent gene trees. Syst. Biol. 49(3), 422–434 (2000)
30. Stolzer, M., Lai, H., Xu, M., Sathaye, D., Vernot, B., Durand, D.: Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. Bioinformatics 28(18), i409–i415 (2012)
31. Takahata, N.: Gene genealogy in three related populations: consistency probability between gene and population trees. Genetics 122(4), 957–966 (1989)
32. Wehe, A., Bansal, M.S., Burleigh, J.G., Eulenstein, O.: Duptree: a program for large-scale phylogenetic analyses using gene tree parsimony. Bioinformatics 24(13), 1540–1541 (2008)
33. Wong, K.M., Suchard, M.A., Huelsenbeck, J.P.: Alignment uncertainty and genomic analysis. Science 319(5862), 473–476 (2008)
34. Wu, C.I.: Inferences of species phylogeny in relation to segregation of ancient polymorphisms. Genetics 127(2), 429–435 (1991)
35. Wu, Y.: Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. Evolution 66, 763–775 (2012)
36. Yu, Y., Than, C., Degnan, J.H., Nakhleh, L.: Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. Syst. Biol. 60(2), 138–149 (2011)
37. Zhang, L.X.: From gene trees to species trees ii: Species tree inference by minimizing deep coalescence events. IEEE-ACM Trans. Comput. Biol. Bioinform. 8(6), 1685–1691 (2011)
38. Zheng, Y., Wu, T., Zhang, L.X.: Reconciliation of gene and species trees with polytomies. arXiv preprint, arXiv:1201.3995 (2012)