

Inferring Pareto-Optimal Reconciliations across Multiple Event Costs under the Duplication-Loss-Coalescence Model – Supplementary Material

Ross Mawhorter, Nuo Liu, Ran Libeskind-Hadas, and Yi-Chieh Wu

This supplement extends the work of Wu et al. (2014) and Du et al. (2019) to track Pareto-optimal event counts and descriptors. Significant new additions and changes are highlighted.

S1 DLC Reconciliation

S1.1 Preliminaries

Throughout this work, the term *tree* refers to a rooted binary tree. Given a tree T , let $V(T)$ denote its node set and $E(T)$ denote its branch set. Let $L(T) \subset V(T)$ denote its leaf set, $I(T) = V(T) \setminus L(T)$ denote its set of internal nodes, and $r(T) \in I(T)$ denote its root node. For node $v \in V(T)$, let $c(v)$ denote its set of children, $p(v)$ denote its parent, and $e(v)$ denote the branch $(p(v), v)$. Define $\leq_T$ ($<_T$) to be the partial order on $V(T)$, where given two nodes u and v of T , $u \leq_T v$ ($u <_T v$) if and only if u is on the unique path between $r(T)$ and v (and $u \neq v$). The partial order $\geq_T$ ($>_T$) is defined analogously. In such a case, u is said to be a (strict) *ancestor* of v and v a (strict) *descendant* of u .

Let a *species tree* S depict the evolutionary history of a set of species, and let a *gene tree* G depict the evolutionary history of a set of genes sampled from these species. To compare a gene tree with a species tree, let a *leaf map* $Le: L(G) \rightarrow L(S)$ label each leaf of the gene tree with the leaf of the species tree from which the gene was sampled.

S1.2 Species Maps and Speciation Nodes

Definition S1.1 (Species Map). Given G , S , and Le , let a *species map* $\mathcal{M}: V(G) \rightarrow V(S)$ map each node of G to the a node of S subject to the following constraints:

1. If $g \in L(G)$, then $\mathcal{M}(g) = Le(g)$.
2. If $g \in I(G)$, then for each $g' \in c(g)$, $\mathcal{M}(g) \leq_S \mathcal{M}(g')$.

Constraint 1 asserts that $\mathcal{M}$ extends the leaf map Le . Constraint 2 asserts that $\mathcal{M}$ satisfies the temporal constraints implied by S .

An internal gene node $g \in I(G)$ is said to be a *speciation node* with respect to species map $\mathcal{M}$ if for each child $g' \in c(g)$, $\mathcal{M}(g') \neq \mathcal{M}(g)$. Given a map $\mathcal{M}$, some nodes may initially be hidden in a gene tree due to losses and deep coalescence. Such “implied speciation nodes” are added to each gene branch that spans multiple branches of the species tree as follows: Locate each non-root internal node $g \in I(G) \setminus \{r(G)\}$ such that either (a) $p(\mathcal{M}(g)) \neq \mathcal{M}(p(g))$ or (b) $p(g)$ is not a speciation node and $\mathcal{M}(g) \neq \mathcal{M}(p(g))$. For each such gene node g , introduce a new node h and replace edge $(p(g), g)$ with the pair of edges $(p(g), h)$ and (h, g) , and define $\mathcal{M}(h) = p(\mathcal{M}(g))$. This process is repeated until there exists no node g that satisfies the conditions above. Additionally, if a speciation node has multiple children, then these children must map to different species. Note that if $\mathcal{M}$ is the lowest common ancestor (LCA) map, then this constraint is ensured by construction since the contrapositive is true. That is, if $\mathcal{M}$ is the LCA map, then for a gene node g with children g' and g'' , if $\mathcal{M}(g') = \mathcal{M}(g'') = s$, it must be that $\mathcal{M}(g) = s$, and g is not a speciation node. Otherwise, to ensure this constraint, implied speciation nodes are added as follows: Locate each internal node $g \in I(G)$ with children g' and g'' such that g is a speciation node and $\mathcal{M}(g') = \mathcal{M}(g'')$. For each child $g' \in c(g)$, introduce a new node h and replace edge (g, g') with the pair of edges (g, h) and (h, g') ,

Note that the sets $N(s, l)$ and $D(s, l)$ are disjoint. Now consider a total order on $D(s, l)$; this order introduces $|D(s, l)| + 1$ bins in which each node in $N(s, l)$ may occur. The total order on $D(s, l)$ and the partition of $N(s, l)$ represent the relative times of duplication nodes as well as the relative times of original nodes with respect to duplication nodes. Define $<_{\mathcal{O}}$ to be the partial order on $O(s, l)$, where given two nodes $g, g' \in O(s, l)$, $g \neq g'$, then $g <_{\mathcal{O}} g'$ if and only if g precedes g' in time. Note that no order is induced on nodes of $N(s, l)$ in the same bin.

The LCT is subject to the following constraints:

1. For each $g, g' \in L(G)$, $g \neq g'$, if $\mathcal{M}(g) = \mathcal{M}(g')$, then $\mathcal{L}(g) \neq \mathcal{L}(g')$.
2. For each $l \in \mathbb{L}$, there exists a $g \in V(G)$ such that $\mathcal{L}(g) = l$.
3. For each $l \in \mathbb{L}$, there exists exactly one $g \in V(G)$ such that $L(g) = l$ and either $g = r(G)$ or $\mathcal{L}(p(g)) \neq l$.
4. For each $s \in V(S)$, each $l \in \text{mother_loci}(s)$, and each $g, g' \in O(s, l)$, $g \neq g'$, if $g <_{\mathcal{O}} g'$, then $g \not\prec_G g'$.

Constraint 1 asserts that extant genes (leaves) mapped to the same extant species (leaves) belong to different loci. Constraint 2 asserts that $\mathbb{L}$ includes only loci used by at least one gene. Constraint 3 asserts that every locus is created only once. Constraint 4 asserts that $\mathcal{O}$ satisfies the temporal constraints implied by G .

Because the locus set $\mathbb{L}$ is defined by the locus map $\mathcal{L}$, we often represent an LCT using the reduced tuple $\langle \mathcal{M}, \mathcal{L}, \mathcal{O} \rangle$.

S1.4 Evolutionary Events in the LCT

For a set of gene nodes $W \subseteq V(G)$, let $\text{loci}(W) = \{ \mathcal{L}(g) \mid g \in W \}$ denote the set of associated loci.

Definition S1.3 (Events). Given G, S, Le , an LCT $\langle \mathcal{M}, \mathcal{L}, \mathcal{O} \rangle$ for $\langle G, S, Le \rangle$ induces the following events:

- **Speciation:** For each internal species node $s \in I(S)$, each locus $l \in \text{loci}(\text{bottoms}(s))$ induces a speciation event.

This speciation event is represented by a tuple $\langle s, A \rangle$, where A is the set of speciation nodes mapped to s and locus l .

- **Duplication:** For each species node $s \in V(S)$ and locus $l \in \text{mother_loci}(s)$, each gene node $d \in D(s, l)$ induces a duplication event.

This duplication event is represented by a tuple $\langle s, d, B \rangle$, where B is the set of gene lineages in species s at locus l contemporaneous with the duplication [see (Deep) Coalescence at Duplication]. Note that the contemporary lineages are not needed to include this duplication in a count, only to represent this duplication. Thus, different partial orders can result in different duplications despite the fact that d occurs along the same gene branch.

- **Loss:** For each species node $s \in V(S)$, each locus $l \in \text{loci}(\text{tops}(s) \cup \text{nodes}(s)) \setminus \text{loci}(\text{bottoms}(s))$ induces a loss event.

A MPR never loses a locus that is created via a duplication in the same species. Therefore, in practice, we must only consider each locus $l \in \text{loci}(\text{tops}(s)) \setminus \text{loci}(\text{bottoms}(s))$. This loss event is represented by a tuple $\langle s, A \rangle$, where A is the set of speciation nodes mapped to $p(s)$ and locus l .

- **(Deep) Coalescence at Speciation:** For each species node $s \in V(S)$ and each locus $l \in \text{loci}(\text{tops}(s))$, let $C(s, l)$ denote the set of gene lineages at locus l at the speciation creating s (from $p(s)$). That is, $C(s, l)$ is the subset of gene branches $(g, g') \in E(G)$ such that $g \in \text{tops}(s)$, $g' \in \text{nodes}(s)$, and $\mathcal{L}(g) = l$. If $|C(s, l)| > 1$, then $C(s, l)$ induces a coalescence at speciation event with $|C(s, l)| - 1$ extra lineages. Each extra lineage incurs the coalescence cost.

This coalescence at speciation event is represented by a tuple $\langle s, C(s, l) \rangle$.

- **(Deep) Coalescence at Duplication:** For each species node $s \in V(S)$ and each locus $l \in \text{mother_loci}(s)$, let $\text{start}(s, l)$ denote the set of “starting” gene lineages in species s at locus l . That is, $\text{start}(s, l)$ is the subset of gene branches $(g, g') \in E(G)$ such that $g \in \{r(G)\} \cup \text{tops}(s) \cup D(s, l)$, $g' \in \text{nodes}(s)$, and $\mathcal{L}(g) = l$. Then, for each duplication node $d \in D(s, l)$, let $K(s, l, d)$ denote the set of gene lineages in species s at locus l contemporaneous with duplication d , where a lineage is contemporaneous if it starts before and ends after the duplication node. This set is determined as follows: Start with $K(s, l, d) = \text{start}(s, l)$. Then process nodes $g \in O(s, l)$ in the order specified by $\mathcal{O}$: for each g , remove $e(g)$ from $K(s, l, d)$, and if $g \in N(s, l)$, then for each $g' \in c(g)$, add $e(g')$ to $K(s, l, d)$. Once $g = d$ (duplication reached), if $|K(s, l, d)| > 1$, then $K(s, l, d)$ induces a coalescence at duplication event with $|K(s, l, d)| - 1$ extra lineages. Each extra lineage incurs the coalescence cost.

This coalescence at duplication event is represented by a tuple $\langle s, K(s, l) \rangle$.

In practice, for a gene node g , rather than use its (possibly arbitrary) name, we report the leaves that are descendants of g .

S1.5 Maximum Parsimony Reconciliations

Let C_D , C_L , and C_C denote the positive real-number costs associated with duplication, loss, and coalescence events, respectively. The traditional goal of DLC reconciliation is to find a most parsimonious reconciliation:

Problem S1.1 (Most Parsimonious Reconciliation (MPR) Problem). Given G , S , Le , C_D , C_L , and C_C , find an LCT for $\langle G, S, Le \rangle$ with minimum reconciliation cost.

Problem S1.2 (Restricted Most Parsimonious Reconciliation (MPR) Problem). Given G , S , Le , C_D , C_L , and C_C , find an LCT for $\langle G, S, Le \rangle$ with minimum reconciliation cost *subject to the condition that the reconciliation between the locus tree and species tree is the LCA map*.

For a comparison of the problems, see Section S2.1 of Wu et al. (2014). In brief, using simulations, we found that the true (simulated) LCT satisfies the additional constraint in almost all gene families. For simplicity, in the main manuscript and henceforth, the term MPR refers to an LCT that solves the restricted MPR problem.

An MPR must satisfy certain properties.

Theorem S1.1 (Optimal Species Maps). *The species map $\mathcal{M}^*$ is optimal if and only if $\mathcal{M}^*$ is the lowest common ancestor (LCA) map.*

Theorem S1.2 (Optimal Locus Maps). *Given a species map $\mathcal{M}$, if the locus map $\mathcal{L}^*$ is optimal, then*

- Each gene branch $e(g) \in E(G)$ has at most one duplication.
- For each species node $s \in V(S)$ and each gene node $g \in \text{nodes}(s) \setminus \text{bottoms}(s)$ internal to the species branch, if g' and g'' denote the children of g , then at most one of the two children branches $e(g')$ or $e(g'')$ has a duplication.

Theorem S1.3 (Optimal Partial Orders). *Given a species map $\mathcal{M}$ and locus map $\mathcal{L}$, if the partial order $\mathcal{O}^*$ is optimal, then for each species $s \in V(S)$ and each locus $l \in \text{mother_loci}(s)$, duplications are placed as early in the species branch as possible. That is, for each original node $g \in N(s, l)$ and each duplication node $d \in D(s, l)$, $g <_{\mathcal{O}^*} d$ if and only if $g \leq_G d$.*

See Wu et al. (2014) and Du et al. (2019) for proofs.

S2 The DLCparETO Algorithm

In this section, we need to keep track of the LCT $\langle \mathcal{M}, \mathcal{L}, \mathcal{O} \rangle$ currently under consideration. In particular, many functions are parameterized by $\mathcal{M}$, $\mathcal{L}$, and/or $\mathcal{O}$. Because the optimal species map is the LCA map, the dependency on $\mathcal{M}$ is omitted. However, dependencies on $\mathcal{L}$ or $\mathcal{O}$ are denoted using subscripts.

S2.1 Definitions

In general, maps are denoted in calligraphic font (e.g. $\mathcal{A}$) and sets denoted in bold font (e.g. $\mathbf{A}$).

Given a tree T , define the following:

- **distance:** Given two nodes $u, v \in V(T)$, let the *distance* $d_T(u, v)$ from u to v be the length, in edges, of the unique path from u to v .
- **lowest common ancestor:** Given a non-empty set of nodes $W \subseteq V(T)$, let the *lowest common ancestor* $lca_T(W)$ be the unique shared ancestor of W of maximum distance from $r(T)$.

Given a gene tree G , species tree S , species map $\mathcal{M}$, and species node $s \in V(S)$, define the following:

- **subtrees:** Let $subtrees(s)$ denote the subtrees of G within $e(s)$. That is, each subtree $T \in subtrees(s)$ is a subtree of G , where the subtree root is a node $r \in \{r(G)\} \cup tops(s)$ and the subtree leaves are the nodes $g \in bottoms(s)$ that are descendants of r .
- **sub-locus maps:** Let $\mathcal{L}^s: tops(s) \cup bottoms(s) \rightarrow \mathbb{L}$ denote a *sub-locus map* that maps each gene node in species branch $e(s)$ to a locus in $\mathbb{L}$. Let $\mathbf{L}(s)$ denote the set of sub-locus maps for s .
- **local orders and sub-partial orders:** Given a set of sub-locus maps $\mathbf{L}(s)$:
 - Given a sub-locus map $\mathcal{L}^s \in \mathbf{L}(s)$ and a locus $l \in mother_loci_{\mathcal{L}^s}(s)$, let $\mathcal{O}_{\mathcal{L}^s}^{s,l}$ denote a *local order* that partially orders gene nodes $O_{\mathcal{L}^s}(s, l)$ in species branch $e(s)$.
 - Given a sub-locus map $\mathcal{L}^s \in \mathbf{L}(s)$, let $\mathcal{O}_{\mathcal{L}^s}^s$ denote a *sub-partial order* that, for each $l \in mother_loci_{\mathcal{L}^s}(s)$, partially orders gene nodes $O_{\mathcal{L}^s}(s, l)$ in species branch $e(s)$. Note that combining local orders $\mathcal{O}_{\mathcal{L}^s}^{s,l}$ for all $l \in mother_loci_{\mathcal{L}^s}(s)$ yields a sub-partial order $\mathcal{O}_{\mathcal{L}^s}^s$.
 - Given a sub-locus map $\mathcal{L}^s \in \mathbf{L}(s)$, let $\mathbf{O}(s, \mathcal{L}^s)$ denote the set of sub-partial orders for s and $\mathcal{L}^s$.
 - Given a sub-locus map $\mathcal{L}^s \in \mathbf{L}(s)$, let $\mathbf{D}(s, \mathcal{L}^s)$ denote the set of Pareto-optimal descriptors for s and $\mathcal{L}^s$ across the set $\mathbf{O}(s, \mathcal{L}^s)$ of sub-partial orders.
- **relative locus pairs, top loci, and bottom loci:** Given a set of sub-locus maps $\mathbf{L}(s)$:
 - Given a sub-locus map $\mathcal{L}^s \in \mathbf{L}(s)$, let $(\mathcal{TL}^s, \mathcal{BL}^s)$ denote a *relative locus pair* that compactly represents the locus assignments at $tops(s)$ and $bottoms(s)$ (see main text, Section 3 for details on the compact representation). That is, $\mathcal{TL}^s: tops(s) \rightarrow \mathbb{L}$ denotes a top-locus map (or *top loci*) that maps each gene node at the top of species branch $e(s)$ to a (relative) locus in $\mathbb{L}$, and $\mathcal{BL}^s: bottoms(s) \rightarrow \mathbb{L}$ denotes a bottom-locus map (or *bottom loci*) that maps each gene node at the bottom of species branch $e(s)$ to a (relative) locus. The sub-locus map $\mathcal{L}^s$ is said to be an *underlying sub-locus map* for $(\mathcal{TL}^s, \mathcal{BL}^s)$, and $\mathcal{L}^s$ is said to *induce* $(\mathcal{TL}^s, \mathcal{BL}^s)$.
 - Let $\mathbf{TL}(s)$ denote the set of top loci, $\mathbf{BL}(s)$ denote the set of bottom loci, and $\mathbf{RLP}(s)$ denote the set of relative locus pairs for s . That is, these are the sets of $\mathcal{TL}^s$, $\mathcal{BL}^s$, and $(\mathcal{TL}^s, \mathcal{BL}^s)$ for all $\mathcal{L}^s \in \mathbf{L}(s)$.
- **Pareto-optimal descriptors:** Let $\mathbb{P}$ denote the set of all Pareto-optimal descriptors. That is, the set $v(\mathbb{P})$ of event counts contains all vectors $v = \langle d, \ell, c \rangle$ such that $d, \ell, c \in \mathbb{N}$ and every event count v is Pareto-optimal with respect to $v(\mathbb{P})$.
 - Given a set of relative locus pairs $\mathbf{RLP}(s)$, let $C^s: \mathbf{RLP}(s) \rightarrow \mathbf{P} \subset \mathbb{P}$ map each relative locus pair to a set of Pareto-optimal descriptors. Specifically, for a relative locus pair $(\mathcal{TL}^s, \mathcal{BL}^s) \in \mathbf{RLP}(s)$, consider the set of LCTs with species map $\mathcal{M}$, a sub-locus map $\mathcal{L}^s \in \mathbf{L}(s)$ that induces $(\mathcal{TL}^s, \mathcal{BL}^s)$, and a sub-partial order $\mathcal{O}_{\mathcal{L}^s}^s \in \mathbf{O}(s, \mathcal{L}^s)$. $C^s(\mathcal{TL}^s, \mathcal{BL}^s)$ denotes the set of Pareto-optimal descriptors for this set of LCTs.
 - Given a set of relative locus pairs $\mathbf{RLP}(s)$, let $F^s: \mathbf{RLP}(s) \rightarrow \mathbf{P} \subset \mathbb{P}$ map each relative locus pair to a set of Pareto-optimal descriptors. Specifically, for a relative locus pair $(\mathcal{TL}^s, \mathcal{BL}^s) \in \mathbf{RLP}(s)$, $F^s(\mathcal{TL}^s, \mathcal{BL}^s)$ denotes the set of Pareto-optimal descriptors for the subtree of S rooted at s with relative locus pair $(\mathcal{TL}^s, \mathcal{BL}^s)$ for s . Furthermore, for top loci $\mathcal{TL}^s \in \mathbf{TL}(s)$, $F^s(\mathcal{TL}^s, \cdot)$ denotes the set of Pareto-optimal descriptors for the subtree of S rooted at s with top loci $\mathcal{TL}^s$ for s .

S2.2 Pseudocode

The main algorithm is shown in Algorithm S1, with helper algorithms in Algorithms S2–S3. For simplicity, details are omitted for cases when a species branch contains no gene tree nodes.

Algorithm S1 Main Algorithm

```

1: function DLCPARETO( $G, S, Le$ )
   input gene tree  $G$ , species tree  $S$ , leaf mapping  $Le$ 
   output set of Pareto-optimal descriptors for  $G$  and  $S$ 

   ▷ set optimal species map to LCA map (see Theorem S1.1), and decompose gene tree (Figure 2A)
2: for each gene node  $g \in V(G)$  in post-order do
3:   if  $g \in L(G)$  then
4:     Set  $\mathcal{M}^*(g) = Le(g)$ 
5:   else
6:     Set  $\mathcal{M}^*(g) = lca_S(\{\mathcal{M}^*(g'), \mathcal{M}^*(g'')\})$ , where  $g'$  and  $g''$  denote the children of  $g$ 
7:   Prune species tree  $S$  to the subtree rooted at  $\mathcal{M}^*(r(G))$  ▷ gene tree evolves within subtree of species tree
8:   Add implied speciation nodes to gene tree  $G$  using  $\mathcal{M}^*$  (Supplemental Section S1.2)
9:   Decompose gene tree  $G$  into disjoint subtrees  $subtrees(s)$ 

   ▷ construct tiles for each species (Figure 2B,D)
10: for each species node  $s \in V(S)$  in pre-order do
   ▷ get set of top loci
11:   if  $s = r(S)$  then
12:     Assign  $r(G)$  to an arbitrary locus
13:     Set  $\mathbf{TL}(s)$  to the set (of size 1) that includes this locus assignment
14:   else
15:     Set  $\mathbf{TL}(s) = \mathbf{BL}(p(s))$ 

   ▷ construct all tiles starting from set of top loci, find set of Pareto-optimal descriptors for tiles
16:   Set  $\mathbf{L}(s) = \text{GETALLLOCUSMAPS}(G, S, \mathcal{M}^*, s, subtrees(s), \mathbf{TL}(s))$  ▷ find set of sub-locus maps
17:   for each sub-locus map  $\mathcal{L}^s \in \mathbf{L}(s)$  do
18:     Initialize  $\mathbf{D}(s, \mathcal{L}^s) = \emptyset$  ▷ set of Pareto-optimal descriptors
19:     Set  $\mathbf{O}(s, \mathcal{L}^s) = \text{GETALLPARTIALORDERS}(G, \mathcal{M}^*, s, \mathcal{L}^s)$  ▷ find set of sub-partial orders
20:     for each sub-partial order  $\mathcal{O}_{\mathcal{L}^s}^s \in \mathbf{O}(s, \mathcal{L}^s)$  do
21:       Find set  $E$  of events and event count  $v = \langle d, \ell, c \rangle$  for  $\langle \mathcal{M}^*, \mathcal{L}^s, \mathcal{O}^s \rangle$ 
22:       Set  $a = \langle v, 1, \{(e, 1) \mid e \in E\} \rangle$  ▷ construct descriptor
23:       Update  $\mathbf{D}(s, \mathcal{L}^s) = \mathbf{D}(s, \mathcal{L}^s) \otimes \{a\}$ 

   ▷ compact tiles, find set of Pareto-optimal descriptors for compact tiles (Figure 2C,D)
24:   Initialize  $\mathbf{BL}(s) = \emptyset$ ,  $\mathbf{RLP}(s) = \emptyset$ 
25:   Initialize  $C^s$  to an empty map
26:   for each sub-locus map  $\mathcal{L}^s \in \mathbf{L}(s)$  do
27:     Find relative locus pair  $(\mathcal{TL}^s, \mathcal{BL}^s)$  for  $\mathcal{L}^s$ 
28:     Add  $\mathcal{BL}^s$  to  $\mathbf{BL}(s)$ , and  $(\mathcal{TL}^s, \mathcal{BL}^s)$  to  $\mathbf{RLP}(s)$ 
29:     if  $C^s(\mathcal{TL}^s, \mathcal{BL}^s)$  is empty then
30:       Initialize  $C^s(\mathcal{TL}^s, \mathcal{BL}^s) = \emptyset$ 
31:       Update  $C^s(\mathcal{TL}^s, \mathcal{BL}^s) = C^s(\mathcal{TL}^s, \mathcal{BL}^s) \otimes \mathbf{D}(s, \mathcal{L}^s)$ 

```

```

    ▷ merge Pareto-optimal descriptors across species (Figure 2E)
32: Initialize  $F^s$  to an empty map
33: for each species node  $s \in V(S)$  in post-order do
34:   for each top loci  $\mathcal{TL}^s \in \mathbf{TL}(s)$  do
35:     if  $s \in L(S)$  then
36:       Set  $\mathcal{BL}^s$  to single element of  $\mathbf{BL}(s)$ 
37:       Set  $F^s(\mathcal{TL}^s, \cdot) = C^s(\mathcal{TL}^s, \mathcal{BL}^s)$ 
38:     else
39:       Set  $s'$  and  $s''$  to be the children of  $s$ 
40:       for each bottom loci  $\mathcal{BL}^s \in \mathbf{BL}(s)$  do
41:         Set  $F^s(\mathcal{TL}^s, \mathcal{BL}^s) = F^{s'}(\mathcal{BL}^s, \cdot) \otimes F^{s''}(\mathcal{BL}^s, \cdot) \otimes C^s(\mathcal{TL}^s, \mathcal{BL}^s)$ 
42:       Set  $F^s(\mathcal{TL}^s, \cdot) = \oplus_{\mathcal{BL}^s: (\mathcal{TL}^s, \mathcal{BL}^s) \in \mathbf{RLP}(s)} F^s(\mathcal{TL}^s, \mathcal{BL}^s)$ 
43:   Set  $s = r(S)$ 
44:   Set  $\mathcal{TL}^s$  to single element of  $\mathbf{TL}(s)$ 
45: return  $F^s(\mathcal{TL}^s, \cdot)$ 

```

Algorithm S2 Enumerate Sub-Locus Maps

```

1: function GETALLLOCUSMAPS( $G, S, \mathcal{M}, s, subtrees(s), \mathbf{TL}(s)$ )
   input gene tree  $G$ , species tree  $S$ , species map  $\mathcal{M}$ , species node  $s$ 
       set  $subtrees(s)$  of subtrees for species  $s$ ,
       set  $\mathbf{TL}(s)$  of top loci for species  $s$ 
   output set of sub-locus maps  $\mathbf{L}(s)$  for species  $s$ 

2:   Find  $DP$  ▷ find all combinations of duplication placements over branches of  $subtrees(s)$  (see Theorem S1.2)

3:   Initialize  $\mathbf{L}(s) = \emptyset$ 
4:   for each sub-locus map  $\mathcal{TL}^s \in \mathbf{TL}(s)$  do ▷ combine set of top loci and set of duplication placements
5:     for each combination of duplication placements  $dp \in DP$  do
6:       Initialize an empty sub-locus map  $\mathcal{L}^s$ 
7:       for each subtree  $T \in subtrees(s)$  do
8:         for each gene node  $g \in V(T)$  in pre-order do ▷ “evolve” locus down subtree
9:           if  $g \in \{r(G)\} \cup tops(s)$  then ▷ set locus assignments for root nodes of subtrees
10:            Set  $\mathcal{L}^s(g) = \mathcal{TL}^s(g)$ 
11:           else ▷ set locus assignments for non-root nodes of subtrees
12:             if  $dp$  has a duplication on  $e(g)$  then
13:               Set  $\mathcal{L}^s(g)$  to a new locus ▷ a new locus is required by LCT Constraint 5
14:             else
15:               Set  $\mathcal{L}^s(g) = \mathcal{L}^s(p(g))$  ▷ locus is equal to parent locus
16:           if  $s \notin L(S) \vee (s \in L(S) \wedge \forall g, g' \in bottoms(s) \text{ s.t. } g \neq g', \mathcal{L}^s(g) \neq \mathcal{L}^s(g'))$  then ▷ LCT Constraint 3
17:             Add  $\mathcal{L}^s$  to  $\mathbf{L}(s)$ 
18:   return  $\mathbf{L}(s)$ 

```

Algorithm S3 Enumerate Sub-Partial Orders for a Sub-Locus Map

```
1: function GETALLPARTIALORDERS( $G, \mathcal{M}, s, \mathcal{L}^s$ )
   input gene tree  $G$ , species map  $\mathcal{M}$ , species node  $s$ 
         sub-locus map  $\mathcal{L}^s$  for species  $s$ 
   output set  $\mathbf{O}(s, \mathcal{L}^s)$  of sub-partial orders for species  $s$  and sub-locus map  $\mathcal{L}^s$ 

2:   for each locus  $l \in \text{mother\_loci}_{\mathcal{L}^s}(s)$  do                                      $\triangleright$  find local orders for this mother locus
3:     Initialize an empty set  $\mathbb{O}_{\mathcal{L}^s}^{s,l}$  of local orders
4:     for each local order  $\mathcal{O}_{\mathcal{L}^s}^{s,l} \in \text{ORDERHELPER}(\text{start}_{\mathcal{L}^s}(s, l), D_{\mathcal{L}^s}(s, l), O_{\mathcal{L}^s}(s, l), \text{empty ordered set})$  do
5:       Add  $\mathcal{O}_{\mathcal{L}^s}^{s,l}$  to  $\mathbb{O}_{\mathcal{L}^s}^{s,l}$ 

    $\triangleright$  combine local orders across mother loci
6:   Set  $\mathbf{O}(s, \mathcal{L}^s)$  by combining  $\mathbb{O}_{\mathcal{L}^s}^{s,l}$  for all  $l \in \text{mother\_loci}_{\mathcal{L}^s}(s)$ 
7:   return  $\mathbf{O}(s, \mathcal{L}^s)$ 

8: iterator ORDERHELPER( $G, SL, DN, ON, \hat{O}$ )
   input gene tree  $G$ , starting lineages  $SL$ , duplication nodes  $DN$ , original nodes  $ON$ , current local order  $\hat{O}$ 
   yield updated local order  $O$ 

9:   Set  $PDN = \hat{O} \cap DN$                                                           $\triangleright$  placed duplication nodes
10:  Set  $PON = \hat{O} \cap ON$                                                           $\triangleright$  placed original nodes

11:  if  $DN = PDN$  then                                                          $\triangleright$  all duplication nodes have been placed
12:    Add set  $U = ON \setminus PON$  to  $\hat{O}$                                             $\triangleright$  add unplaced original nodes
13:    yield  $\hat{O}$ 
14:  else
    $\triangleright$  find unplaced duplication nodes with minimum ancestor constraints (see Theorem S1.3)
15:    Set  $DN' = \arg \min_{g \in DN \setminus PDN} d_G(r, g)$ , where  $r \in V(G)$  such that  $e(r) \in SL$  and  $r \leq_G g$ 
16:    for each gene node  $g \in DN'$  do
17:      Create copy  $O'$  of  $\hat{O}$ 
18:      Find set  $U$  of gene nodes  $g' \in ON \setminus PON$  s.t.  $g' \leq_G g$             $\triangleright$  unplaced ancestors of duplication node
19:      Add  $U$  to  $O'$                                                           $\triangleright$  add partition of original nodes
20:      Add  $g$  to  $O'$                                                           $\triangleright$  add duplication node
21:      for each  $O \in \text{ORDERHELPER}(SL, DN, ON, O')$  do                          $\triangleright$  recur to add remaining nodes
22:        yield  $O$ 
23: end iterator
```

References

- Du H, Ong Y. S, Knittel M, Mawhorter R, Liu N, Gross G, Tojo R, Libeskind-Hadas R and Wu Y.-C. 2019. Multiple optimal reconciliations under the duplication-loss-coalescence model. In 17th Asia Pacific Bioinformatics Conference (APBC 2019). Wuhan, China.
- Wu Y.-C, Rasmussen M. D, Bansal M. S and Kellis M. 2014. Most parsimonious reconciliation in the presence of gene duplication, loss, and deep coalescence using labeled coalescent trees. *Genome Research* **24**:475–486.