Supporting Online Material

1 Definitions

A phylogenetic tree T is a graph $(V(T), E(T))$, with a set $V(T)$ of vertices and a set $E(T)$ of directed edges (v,u). Let $e(v)$ be the edge $(v, \rho(v))$, where $\rho(v)$ is the parent of node v. We denote the children of a node v as $c_1(v)$ and $c_2(v)$. The leaves of T are $L(T)$ and the internals nodes are $I(T)$. Also, let $t(v)$ be the length of branch $e(v)$ expressed in units of time (generations). We use $\tau(v)$ to represent the age of a node v (i.e. the length of any path from v to the leaves). We use the relation $v < w$ to mean that v is a descendant of w and we use $v \leq w$ to mean that v is either a descendant or is equal to w. We also define the relation for edges $e(v) < e(w)$ if $v < w$.

2 Relevant probabilities for the DLCoal model

2.1 Review of the coalescent model

The coalescent model describes the rate at which lineages within a population find a common ancestor (coalesce) as one goes backwards in time. For a diploid species with an effective population size N, the probability that any pair of k lineages coalesce at generation t is

$$P(t|k,N) = \binom{k}{2} \frac{1}{2N} \exp\left(-\binom{k}{2} \frac{1}{2N} t \right). \tag{1}$$

The process is repeated until all lineages coalesce into a single common ancestor, and the tree generated by this process is called a *coalescent tree*. Alternatively, the process can be terminated at some predetermined time, in which case it is possible that not all lineages fully coalesce. This truncated process is call the *censored coalescent* (Rannala and Yang 2003) and it has been derived (Saunders et al. 1984, Rosenberg 2002) that the probability of a lineages coalescing into b lineages in time t generations is

$$P(b|a,t,N) = \sum_{k=b}^{a} \exp\left(\frac{-k(k-1)}{4N} t \right) \frac{(2k-1)(-1)^{k-b}}{b!(k-b)!(k+b-1)} \prod_{y=0}^{k-1} \frac{(b+y)(a-y)}{a+y}. \tag{2}$$

An important special case is the probability that no coalescent occurs for k lineages before time t generations, which is

$$P(\text{no coal}|k,t,N) = P(b=k|a=k,t,N) = \exp\left(-\frac{k(k-1)}{4N} t \right). \tag{3}$$

Another important special case occurs in our bounded coalescent process, where we have $b = 1$,

$$P(b=1|a,t,N) = \sum_{k=1}^{a} \exp\left(-\frac{k(k-1)}{4N} t \right) (2k-1) \prod_{y=1}^{k-1} \frac{y-a}{a+y}. \tag{4}$$

2.2 The bounded coalescent

In this work, we introduce a new process called the bounded coalescent (Figure S1a). In this process, we imagine that we have a new allele (black dot) occurring at a known time t^*, and we are given k lineages at time $t = 0$ that also have the allele. For our purposes, the new allele represents the presence of a new duplicate and the old allele (white dots) represents its absence. In addition, we have no knowledge of the

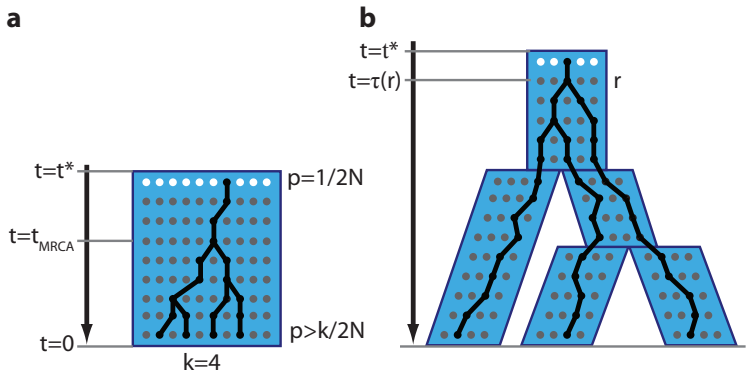

Figure S1: **Bounded coalescents.** (a) A bounded coalescent is a coalescent where the time of the MRCA t_{MRCA} is more recent than some deadline t^*. This is equivalent to conditioning the process on having a new mutation (black dots) occurring at time t^* and all k leaves have the mutation. The frequency of the mutation is unknown (grey dots) for all other times in the tree. (b) A bounded multispecies coalescent (BMC) is a multispecies coalescent with the condition that the root r of the gene tree has a time $\tau(r)$ more recent than some deadline t^*.

frequency of the allele at any other time (grey dots). Let the coalescent times of the k lineages be described by a new process called the bounded coalescent.

We can derive the distribution of the coalescent times in the bounded coalescent by making the following observation. Requiring that all k lineages have the new allele, implies that the k lineages must be descendants of the first individual with new allele at time t^*, and only coalescent trees whose most recent common ancestor has a time t_{MRCA} more recent than t^* satisfy this condition. Furthermore, given that a coalescent tree has $t_{MRCA} < t^*$, there is a $1/2N$ probability that root of the tree has the new allele. Notice that this probability is independent of the tree's topology or branch lengths. Therefore, a coalescent process conditioned on $t_{MRCA} < t^*$ is an equivalent definition of the bounded coalescent. The probability density of the time t of the next coalescent between k lineages in the bounded coalescent process is then

$$P(t|t_{MRCA} < t^*, k, N) = \frac{P(t, t_{MRCA} < t^*|k,N)}{P(t_{MRCA} < t^*|k,N)} \tag{5a}$$

$$= \begin{cases} \dfrac{P(t|k,N)}{P(t_{MRCA} < t^*|k,N)}, & \text{if } t_{MRCA} < t^* \\ 0, & \text{otherwise} \end{cases} \tag{5b}$$

Notice, that the term $P(t_{MRCA} < t^*|k,N)$ is equivalent to the probability that k lineages go to 1 lineage in time t^* (Equation 4).

2.3 The bounded multispecies coalescent (BMC)

Another useful process to define is the coalescence of the descendants of a duplication that occurs higher up in a species tree (Figure S1b). Using the same arguments, we can model these gene lineages as a multispecies coalescent with the condition that the age of their MRCA $\tau(r)$ is more recent than the time of the duplication t^*. We call this conditioned process, the *bounded multispecies coalescent* (BMC).

Let r be the root (MRCA) of the gene tree $\mathbb{G} = (T, \boldsymbol{t})$ with topology T and branch lengths $\boldsymbol{t}$. Let $\boldsymbol{n}$ be a vector of gene counts for each extant species, such that $n_u = |\{v : R(v) = u, v \in L(T)\}|$ for $u \in L(S)$.

Typically $n_u = 1$, unless multiple extant individuals are present per species in the data. The probability distribution of the gene tree is then

$$P(\mathbb{G}, R|\tau(r) < t^*, \boldsymbol{n}, \mathbb{S}, \boldsymbol{N}) = \frac{P(\mathbb{G}, R, \tau(r) < t^*|\boldsymbol{n}, \mathbb{S}, \boldsymbol{N})}{P(\tau(r) < t^*|\boldsymbol{n}, \mathbb{S}, \boldsymbol{N})} \tag{6a}$$

$$= \begin{cases} \dfrac{P(\mathbb{G}, R|\boldsymbol{n}, \mathbb{S}, \boldsymbol{N})}{P(\tau(r) < t^*|\boldsymbol{n}, \mathbb{S}, \boldsymbol{N})}, & \text{if } \tau(r) < t^* \\ 0, & \text{otherwise} \end{cases} \tag{6b}$$

The numerator is the probability of a gene tree in the multispecies coalescent, which has been derived by Rannala and Yang (Rannala and Yang 2003). The denominator has also been derived by Efromovich and Kubatko (Efromovich and Kubatko 2008) and we review its calculation using our own notation.

2.4 The age of the MRCA in a multispecies coalescent

We briefly review the computation of $P(\tau(r) < t^*|\boldsymbol{n}, \mathbb{S}, \boldsymbol{N})$, which is the cumulative distribution function (CDF) of the age of the MRCA of a gene tree. This probability can computed as

$$P(\tau(r) < t^*|\boldsymbol{n}, \mathbb{S}, \boldsymbol{N}) = \sum_{k=2}^{M} P(\tau(r) < t^*|k, \boldsymbol{n}, \mathbb{S}, \boldsymbol{N}) P(k|\boldsymbol{n}, \mathbb{S}, \boldsymbol{N}), \tag{7}$$

where k is the number of lineages present at $root(S)$ and $M = \sum_i n_i$. The first term is an application of Equation 4,

$$P(\tau(r) < t^*|k, \boldsymbol{n}, \mathbb{S}, \boldsymbol{N}) = P(b = 1|a = k, t = t^* - \tau(root(S)), N = N(r)). \tag{8}$$

The second term can be computed using dynamic programming. Let $a(u)$ be the number of lineages present at node $u \in V(S)$. Thus, $a(u) = n_u$ for $u \in L(S)$. Let $b(u)$ be the number of lineages present at the top of branch $e(u)$. Let $c_1(u)$ and $c_2(u)$ be the left and right children of u. Therefore,

$$a(u) = b(c_1(u)) + b(c_2(u)).$$

Using these definitions, we can express our desired term as

$$P(k|\boldsymbol{n}, \mathbb{S}, \boldsymbol{N}) = P(a(root(S)) = k|\boldsymbol{n}, \mathbb{S}, \boldsymbol{N}), \tag{9}$$

and we can compute it recursively. First, note the probability of seeing $b(u)$ lineages at the top of branch $e(u)$ in the specie stree is

$$P(b(u) = k|\boldsymbol{n}, \mathbb{S}, \boldsymbol{N}) = \sum_{i=1}^{M_u} P(b = k|a = i, t = t(u), N = N(u)) P(a(u) = i|\boldsymbol{n}, \mathbb{S}, \boldsymbol{N}), \tag{10}$$

where

$$M_u = \sum_{v \in L(S_u)} n_v,$$

and S_u is the subtree of S beneath node u. The probability of seeing $a(u)$ lineages at node u is

$$P(a(u) = k|\boldsymbol{n}, \mathbb{S}, \boldsymbol{N}) = \begin{cases} \displaystyle\sum_{i=1}^{k-1} P(b(c_1(u)) = i|\boldsymbol{n}, \mathbb{S}, \boldsymbol{N}) P(b(c_2(u)) = k - i|\boldsymbol{n}, \mathbb{S}, \boldsymbol{N}), & \text{if } u \in I(S) \\ \mathbb{I}_{k=n_u}, & \text{if } u \in L(S) \end{cases} \tag{11}$$

where $\mathbb{I}$ is an indicator function.

2.5 Review of the multispecies coalescent

The *multispecies coalescent* (Rannala and Yang 2003, Degnan and Rosenberg 2009) is a generalization of the coalescent for multiple populations connected into a tree representing the evolution of a set of species. The tree is called a species tree $\mathbb{S} = (S, \boldsymbol{t}(S))$ where S is a topology and $\boldsymbol{t}(S)$ is a vector of branch lengths expressed in generations. We initialize the process by specifying a number of gene lineages n_u present in each species leaf $u \in L(S)$. For each branch $e(u)$ of the species tree, we run a censored coalescent (Rannala and Yang 2003) process that coalesces $a(u)$ lineages into $b(u)$ lineages over time $t(u)$. Lineages that remain after this process proceed onto the parental branch $e(\rho(u))$.

For the root node $root(S)$ of the species tree, we run a normal coalescent process to coalesce $a(root(S))$ lineages into one lineage with no restriction on coalescent time.

This process generates a gene tree $\mathbb{G} = (T, \boldsymbol{t}(T))$, where T is the topology and $\boldsymbol{t}(T)$ is a vector of branch lengths expressed in number of generations, such that each branch $e(v) \in E(T)$ has length $t(v)$. The process also produces a reconciliation $R : V(T) \to E(S)$, which is a mapping of vertices in the gene tree to edges in the species tree.

The number of lineages starting and ending on each branch of the species tree are very useful for several calculations. If one has a reconciliation R, the lineage counts $a(u)$ and $b(u)$ for all $u \in V(S)$ can be computed using recursion. First, we can initialize the lineage counts at the leaves,

$$a(u) = |\{v : R(v) = e(u), v \in L(T)\}|, \text{ if } u \in L(S). \tag{12}$$

The lineage counts $b(u)$ present at the top of a branch $e(u)$ equals the count at the bottom $a(u)$ minus the number of coalescences in $e(u)$,

$$b(u) = a(u) - |\{v : R(v) = e(u), v \in I(T)\}|. \tag{13}$$

Lastly, the ending counts of two child branches of $e(u)$ sum together to give the starting count of branch $e(u)$. Therefore, for each internal node $u \in I(S)$ and its children $c_1(u)$ and $c_2(u)$ we have,

$$a(u) = b(c_1(u)) + b(c_2(u)), \text{ if } u \in I(S). \tag{14}$$

Another useful definition is to consider the subgraph T^u of the gene tree T that contains any edge that "crosses into" branch $e(u)$. This subgraph T^u is a forest of trees whose leaves represent the starting lineages, roots represent the ending lineages, and topology represent the particular pattern of coalesce. We define these subgraphs and their reconciliations R^u as,

$$T^u = (V(T^u), E(T^u)) \tag{15a}$$
$$V(T^u) = \{v : R(v) = e(u) \vee R(v) < e(u) \leq R(\rho(v))\} \tag{15b}$$
$$E(T^u) = \{(v, w) : v \in V(T^u), w \in V(T^u)\} \tag{15c}$$
$$R^u(v) = R(v), \forall v \in V(T^u). \tag{15d}$$

2.6 The probability of a reconciled topology in the multispecies coalescent

As currently formulated in the DLCoalRecon algorithm, we only consider the topology of the gene tree T^G. In the main text, we show how the probability of a gene tree topology can be computed. One of the terms needed for this calculation is the probability of a reconciled gene tree topology in the multispecies process, $P(T, R|S, \boldsymbol{t}, \boldsymbol{N})$. Here we show how this can be computed efficiently.

Degnan and Salter (Degnan and Salter 2005) were the first to introduce general equations for computing the probability of the gene tree topology with an arbitrary number of leaves for the coalescence process. A

major complexity in that work was summing over reconciliations. However, our current problem is much simpler since the reconciliation R is proposed.

Let T be the gene tree topology, R be the reconciliation, and $(S, \boldsymbol{t})$ be the species tree topology and branch lengths in time. Lastly, let $\boldsymbol{N}$ be a vector of population sizes. The coalescences are independent in each branch $e(u)$ of the species tree, so we can factor a gene tree T into its subgraphs T^u,

$$P(T,R|S,\boldsymbol{t},\boldsymbol{N}) = \prod_{u \in V(S)} P(T^u, b = b(u)|a = a(u), t = t(u), N = N(u)). \tag{16}$$

For each subgraph, we continue to factor,

$$\begin{aligned} P(T^u, b(u)|a(u), t(u), N(u)) =& P(T^u|a = a(u), b = b(u), t = t(u), N = N(u)) \times & \text{(17a)} \\ & P(b = b(u)|a = a(u), t = t(u), N = N(u)). & \text{(17b)} \end{aligned}$$

The second term is defined by Equation 2. Once we condition on going from a lineages to b lineages, we only need to compute the probability of the topology T^u. This can be done by working with *labeled histories*.

A labeled history is a labeled topology with an ordering defined on the internal nodes (representing the order of the coalescences). One convenient property of labeled histories is that for a given number of leaves a and roots b, each labeled history is equally likely. The total number of possible labeled histories H_{ab} for a leaves and b roots is

$$H_{ab} = \binom{a}{2}\binom{a-1}{2}\ldots\binom{b+1}{2} = \prod_{i=b+1}^{a}\binom{i}{2} \tag{18}$$

Next, we need to compute how many of these labeled histories have labeled topology T^u. In general, the number of labeled histories $H(T)$ that have labeled topology T is

$$H(T) = |I(T)|! \prod_{v \in I(T)} |I(T_v)|^{-1}. \tag{19}$$

where T_v is the subtree of T rooted at v. Thus, the probability of a topology T^u is

$$P(T^u|a, b, t(u), N(u)) = \frac{|I(T^u)|!}{H_{ab}} \prod_{v \in I(T^u)} |I(T^u_v)|^{-1} \tag{20}$$

For the basal species branch $u = root(S)$ we have

$$P(T^u|a(u), t(u), N(u)) = \frac{H(T^u)}{H_{a1}} \tag{21}$$

and

$$P(b = 1|a = a(u), t = \infty) = 1.$$

So in conclusion,

$$P(T,R|S,\boldsymbol{t},\boldsymbol{N}) = \prod_{u \in V(S)} P(b(u)|a(u), t(u), N(u)) \frac{|I(T^u)|!}{H_{ab}} \prod_{v \in I(T^u)} |I(T^u_v)|^{-1}. \tag{22}$$

2.7 Derivation of posterior probability of reconciliation

In the main text, we presented a new reconciliation algorithm called DLCoalRecon. The *reconciliation problem* is to determine the evolutionary events necessary for explaining a given gene tree $\mathbb{G} = (T^G, \boldsymbol{t}^G)$ and species tree $\mathbb{S} = (S, \boldsymbol{t}^S)$ (Goodman et al. 1979, Page 1994). Usually, a reconciliation is defined as a mapping from vertices in the gene tree to vertices and edges in the species tree, however, in the DLCoal model, the reconciliation $\mathbb{R}$ is a tuple

$$\mathbb{R} = (T^L, R^G, R^L, \delta^L), \tag{23}$$

where T^L is the locus tree, R^G is a mapping from the gene tree to the locus tree, R^L is a mapping from the locus tree to the species tree S, and δ^L is a set of daughter nodes. Given our model parameters,

$$\theta = (\boldsymbol{t}^S, \boldsymbol{N}, \lambda, \mu),$$

our goal is to compute the maximum *a posteriori* reconciliation, thus

$$\hat{\mathbb{R}} = \operatorname*{argmax}_{\mathbb{R}} P(\mathbb{R}|T^G, S, \theta) \tag{24a}$$

$$= \operatorname*{argmax}_{T^L, R^G, R^L, \delta^L} \frac{P(T^G, T^L, R^G, R^L, \delta^L|S, \theta)}{P(T^G|S, \theta)} \tag{24b}$$

$$= \operatorname*{argmax}_{T^L, R^G, R^L, \delta^L} P(T^G, T^L, R^G, R^L, \delta^L|S, \theta). \tag{24c}$$

Notice, that maximizing the posterior is the same as maximizing the joint probability when T^G is given. We currently assume that $\boldsymbol{t}^G$ is unknown, since in practice such times are not directly known without a molecular clock assumption. By introducing the locus tree branch lengths t^L, we can now separate the variables for the gene tree and locus tree.

$$P(T^G, R^G, T^L, R^L, \delta^L|S, \theta) = \int P(T^G, R^G, T^L, \boldsymbol{t}^L, R^L, \delta^L|S, \theta) d\boldsymbol{t}^L \tag{25a}$$

$$= \int P(T^G, R^G|T^L, \boldsymbol{t}^L, R^L, \delta^L, S, \theta) P(T^L, \boldsymbol{t}^L, R^L, \delta^L|S, \theta) d\boldsymbol{t}^L. \tag{25b}$$

Furthermore, we can factor the second term above into a probability for the locus tree branch lengths, daughter nodes, and topology, giving us

$$\int P(T^G, R^G|T^L, \boldsymbol{t}^L, \delta^L, \boldsymbol{N}^L) P(\boldsymbol{t}^L|T^L, R^L, S, \theta) P(\delta^L|T^L, R^L, S) P(T^L, R^L|S, \theta) d\boldsymbol{t}^L \tag{25c}$$

$$= P(\delta^L|T^L, R^L, S) P(T^L, R^L|S, \theta) \int P(T^G, R^G|T^L, \boldsymbol{t}^L, \delta^L, \boldsymbol{N}^L) P(\boldsymbol{t}^L|T^L, R^L, S, \theta) d\boldsymbol{t}^L. \tag{25d}$$

The term $P(T^L, R^L|S, \theta)$ has been derived (Arvestad et al. 2003; 2009) and for the daughters set δ^L, we have

$$P(\delta^L|T^L, R^L, S) = 2^{-|dup(T^L, R^L, S)|}, \tag{26}$$

where $dup(T^L, R^L, S)$ gives the number of duplications in the locus tree. This probability is derived from the fact that there are two ways to choose a daughter node for each duplication in the locus tree. We perform the integration by sampling as is done in (Arvestad et al. 2004, Rasmussen and Kellis 2010).

The probability of the reconciled topology T^G, R^G in the multilocus coalescent (MLC) process can be derived as follows. Since the MLC is the multispecies coalescent with the condition that coalescence is complete within each daughter edge, we can write the term as,

$$P(T^G, R^G | T^L, \boldsymbol{t}^L, \boldsymbol{N}^L, \tau(root(T^{G,v})) < \tau(\rho(v)) \; \forall v \in \delta^L) = \frac{P(T^G, R^G | T^L, \boldsymbol{t}^L, \boldsymbol{N}^L)}{\prod_{v \in \delta^L} P(\tau(root(T^{G,v})) < \tau(\rho(v)) | T^{L,v}, \boldsymbol{t}^{L,v}, \boldsymbol{N}^{L,v})}, \tag{27}$$

where $T^{G,v}$ are the subtrees of T^G that evolve inside daughter subtrees $T^{L,v}$. Here, we use the fact that if the reconciled topology T^G, R^G shows complete coalescence within each daughter edge, then the numerator simplifies to the probability of a reconciled topology in the multispecies coalescent (Equation 22). We factor the denominator into the probability that each daughter subtree $T^{G,v}$ completely coalesces before its deadline $\tau(\rho(v))$. Each term of the product can be computed using Equation 7 or Efromovich *et al.* (Efromovich and Kubatko 2008).

3 Simulation

For this work, we implemented a simulation program for our DLCoal model, where given a species tree S with divergence times $\boldsymbol{t}$, population sizes $\boldsymbol{N}$, and duplication-loss rates λ, μ we can sample a locus tree $(T^L, \boldsymbol{t}^L, \delta^L)$ and a gene tree $(T^G, \boldsymbol{t}^G)$. Sampling the locus tree is done with the birth-death process (Arvestad et al. 2004, Rasmussen and Kellis 2010). Here, we describe several algorithms needed for sampling the gene tree from the multilocus coalescent process.

3.1 Coalescent times conditioned on lineage counts

First, we describe a useful sub-procedure for sampling coalescent times conditioned on the starting lineage counts a and ending counts b over a time t'. The distribution of the waiting time x is then

$$P(x | a = k_1, b = k_2, t = t', N) = \frac{P(x, b = k_2 | a = k_1, t = t', N)}{P(b = k_2 | a = k_1, t = t', N)} \tag{28a}$$

$$= \frac{P(b = k_2 | a = k_1 - 1, t = t' - x, N) P(x | a = k_1, N)}{P(b = k_2 | a = k_1, t = t', N)}. \tag{28b}$$

For sampling, we use root finding on the cumulative distribution function (CDF) which is

$$P(x < x' | a = k_1, b = k_2, t = t') \tag{29a}$$

$$= \left[\frac{1}{k_2!} \sum_{k=k_2}^{k_1 - 1} \exp(\lambda t') \int_0^{x'} \exp((\lambda_1 - \lambda) x) dx \frac{2k-1}{k + k_2 - 1} C_{k-1} \right] (-\lambda_1) P(b = k_2 | a = k_1, t = t')^{-1} \tag{29b}$$

$$= \left[\frac{1}{k_2!} \sum_{k=k_2}^{k_1 - 1} \exp(\lambda t') \frac{1 - \exp((\lambda_1 - \lambda) x')}{\lambda_1 - \lambda} \frac{2k-1}{k + k_2 - 1} C_{k-1} \right] (-\lambda_1) P(b = k_2 | a = k_1, t = t')^{-1}. \tag{29c}$$

where

$$\lambda = \frac{-k(k-1)}{4N}, \;\; \lambda_1 = \frac{-k_1(k_1 - 1)}{4N}.$$

This sampling problem has also been investigated previously by Blum and Rosenberg using a slightly different approach (Blum and Rosenberg 2007).

3.2 Sampling the bounded coalescent

Once we have defined sampling with conditioned lineage counts, we can sample the bounded coalescent by specifying $t' = t^*$ and $b = 1$.

3.3 Sampling the multilocus coalescent (MLC)

We have implemented a sampling method of the multilocus coalescent (MLC) using a multi-step strategy: we first sample the gene lineage counts $a(u), b(u)$ in each branch of the locus tree $e(u)$, then we sample coalescence times conditioned on the beginning and ending lineages counts for each locus branch (using Section 3.1).

Sampling lineage counts. To sample lineage counts $a(u)$ and $b(u)$ we recursively choose counts from the root to the leaves of the locus tree.

First, consider a node $u \in V(T^L)$, where we are given the ending count $b(u)$. Given $b(u)$, we would like to sample the ending lineage counts of the children $b(c_1(u))$ and $b(c_2(u))$. The probability distribution is

$$P(b(c_1(u)) = k_1, b(c_2(u)) = k_2 | b(u), \theta) \tag{30a}$$

$$= \frac{P(b(c_1(u)) = k_1, b(c_2(u)) = k_2, b(u) | \theta)}{P(b(u)|\theta)} \tag{30b}$$

$$= \frac{P(b(u)|b(c_1(u)) = k_1, b(c_2(u)) = k_2, \theta) P(b(c_1(u)) = k_1, b(c_2(u)) = k_2 | \theta)}{P(b(u)|\theta)} \tag{30c}$$

$$= \frac{P(b(u)|a(u) = k_1 + k_2, t = t(u), N = N(u)) P(b(c_1(u)) = k_1|\theta) P(b(c_2(u)) = k_2|\theta)}{P(b(u)|\theta)}, \tag{30d}$$

where

$$\theta = (T^L, \boldsymbol{t}^L, \boldsymbol{N}, \boldsymbol{n}).$$

The first term of numerator is Equation 2, and each of the other terms are applications of Equation 10.

We sample the above distribution with a simple rejection sampling algorithm. We recursively apply the algorithm to sample lineage counts $a(u)$ and $b(u)$ for each branch $e(u)$ in the locus tree T^L. The sampling is initialized with $b(root(T^L))$ at the root of the locus tree. We then perform the following steps for each branch $e(u)$ in the locus tree in preorder traversal (top-down):

1. Sample $b(c_1(u)) = k_1$ and $b(c_2(u)) = k_2$ using the distributions $P(b(c_1(u)) = k_1|\theta)$ and $P(b(c_2(u)) = k_2|\theta)$. These are precomputed in the dynamic programming table of Equation 10.

2. With probability $P(b = b(u)|a = k_1 + k_2, t = t(u), N = N(u))$, accept the sample. Otherwise, go to step 1.

Once all lineage counts are sampled, use the distributions in Section 3.1 to sample times of all coalescences occurring on each branch $e(u)$ of the locus tree.

3.4 Relaxing the hemiplasy assumption

For this current model and reconciliation algorithm, we made the simplifying assumption that hemiplasy of duplication and loss events does not occur. Although it is more complicated, it is possible to implement a simulator that allows duplication-loss hemiplasy. Using this simulator, we have found that such events only occur in a small fraction of gene trees (5% of simulated fly gene trees with $N = 10^6, \lambda = \mu = 0.0012$).

The dup-loss hemiplasy simulator contains many of the same features as the DLCoal simulator described in this section, but differs in its sampling of the locus tree. Here, we briefly describe the generative process of the locus tree with dup-loss hemiplasy.

The process takes as input a species tree $\mathbb{S}$, a vector of population sizes $\boldsymbol{N}$, an initial allele frequency p_0, duplication λ and loss μ rates, and a time step t_Δ. A locus tree is initialized with a single gene lineage present at the root of the species tree and we recursively grow this lineage downwards. While growing the lineage, we keep track of its frequency p in the population. Along the way several events can effect a gene lineage.

Speciation. If a gene lineage reaches a speciation node in the species tree, the gene lineage bifurcates into two gene lineages that grow in each descendant species branch independently.

Duplication. At a constant rate $p\lambda$, a gene lineage can duplicate into two lineages: a *mother* and a *daughter*. The daughter lineage is initialized with a small frequency (e.g. 0.05). For efficiency, we choose a initial frequency greater than $1/2N$, since most duplications starting at such a small frequency would go extinct.

Loss. At a constant rate $p\mu$, a gene lineage can experience deletion. This is implemented as reducing the frequency p by some small amount (e.g. 0.05).

Time step. At regular time intervals of t_Δ, we update the frequency p of the gene lineage according to a diffusion approximation of the Wright-Fisher process. During this process, p may fix to 1, where it will stay unless a subsequent loss reduces it below 1. The lineage may also go *extinct* ($p = 0$), in which case we terminate its growth down the species tree.

Extant species. If a gene lineage encounters a leaf of the species tree, we terminate the gene lineage and mark its endpoint an extant gene.

Once all the gene lineages have terminated (i.e. become extant genes or go extinct), we next perform a pruning step that removes any extinct lineages. The resulting tree is then called a locus tree. One additional modification is that we define frequency dependent population sizes for the locus tree, such that

$$N^L(v) = p(v) \times N(R(v)),$$

where $p(v)$ is the allele frequency along branch $e(v)$. Lastly, we then run a multilocus coalescent process within this generated locus tree to produce a gene tree.

References

Arvestad L, Berglund A, Lagergren J and Sennblad B. 2004. Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. Proceedings of the Eighth Annual International Conference on Computational Molecular Biology 326–335.

Arvestad L, Berglund A.-C, Lagergren J and Sennblad B. 2003. Bayesian gene/species tree reconciliation and orthology analysis using mcmc. Bioinformatics **19 Suppl 1**:i7–15.

Arvestad L, Lagergren J and Sennblad B. 2009. The gene evolution model and computing its associated probabilities. Journal of the ACM (JACM) **56**:1–44.

Blum M. G. B and Rosenberg N. A. 2007. Estimating the number of ancestral lineages using a maximum-likelihood method based on rejection sampling. Genetics **176**:1741–1757.

Degnan J. H and Rosenberg N. A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol Evol **24**:332–340.

Degnan J. H and Salter L. A. 2005. Gene tree distributions under the coalescent process. Evolution **59**:24–37.

Efromovich S and Kubatko L. S. 2008. Coalescent time distributions in trees of arbitrary size. Stat Appl Genet Mol Biol **7**:Article2.

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

Page R. 1994. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. Systematic Biology **43**:58–77.

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

Rasmussen M. D and Kellis M. 2010. A bayesian approach for fast and accurate gene tree reconstruction. Mol Biol Evol .

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

Saunders I. W, Tavar S and Watterson G. A. 1984. On the genealogy of nested subsamples from a haploid population. Advances in Applied Probability **16**:pp. 471–491.