

# Supplementary Materials for “Distribution of Gene Tree Topologies with Duplication, Loss, and Coalescence”

## 1 Insertion and Attachment in the 1 Duplication Case

### 1.1 Number of possible branches duplicates can insert into (Daughter locus)

Let  $N$  be the number of species. Under the MSC-DL model a duplication insertion can occur on any gene-tree branch that subtends a non-empty subset of taxa. We here show that the total number of such branches is:

$$\text{Possible Duplication Branches}(N) = \sum_{k=1}^N \binom{N}{k} (2k - 3)!! \quad (1)$$

#### 1.1.1 Branches being counted

A duplication insertion can occur on all branch with non zero probability under the MSC-DL. Hence, we are counting all possible branches in all the gene tree topologies across all epochs of time in the species tree.

#### 1.1.2 All size subsets can appear in a clade

Under the multispecies coalescent (MSC) model, any subset of taxa can appear at any given time before the most recent common ancestor(MRCA) of a gene tree with non-zero probability (Degnan and Salter 2005). That means a branch where the duplication on a species tree landed can represent any subset of taxa below it. The number of such subsets is:

$$\sum_{k=1}^N \binom{N}{k} \quad (2)$$

#### 1.1.3 Shape of clades for each subset

For any fixed subset of size  $k$  under the MSC, we can generate  $(2k - 3)!!$  number of rooted binary gene tree topologies. In all of these trees there is exactly one branch subtending these clades whose descendant leaves correspond to the chosen subset. The number of such trees is:

$$(2k - 3)!! \quad (3)$$

#### 1.1.4 No double counting

Two different subsets of leaves (in 1.1.2) correspond to different descendant sets of lineage. Hence, they have different and distinct associated branches. Within each subset of lineages each binary tree topology counts exactly one branch of the set in that shape, making equation 1 unique.

#### 1.1.5 Final expression

Summing over all subsets of lineage of all sizes and all the corresponding trees (clade shapes) that can be formed from them:

$$\text{Possible Duplication Branches}(N) = \sum_{k=1}^N \underbrace{\binom{N}{k}}_{\text{subsets of lineages}} \cdot \underbrace{(2k-3)!!}_{\text{shapes contributing one branch}} \quad (4)$$

## 1.2 Number of possible branches duplicated lineages can attach to (Parent locus)

### 1.2.1 Number of possible attachment branches

We now show that the number of possible branches where the duplicated lineage can attach in the parent locus is:

$$\text{Possible Attachment Branches}(N) = (2N-3)!!(2N-1) \quad (5)$$

### 1.2.2 MSC generates all gene tree topologies

Under the MSC all rooted binary gene tree topology on sets of taxa occurs with non-zero probability. The number of such gene trees for  $N$  species is:

$$(2N-3)!! \quad (6)$$

### 1.2.3 Each topology contains branches

A rooted binary tree with  $N$  leaves has exactly  $(2N-1)$  branches.

### 1.2.4 All branches can serve as an attachment point

Given a duplication event at time  $t_d$ , the duplicated lineage can coalesce with all branch of the parent locus. This is because:

1. Under the MSC all topologies occurs with non-zero probability.
2. All branches in that topology exists for some time interval ( $> t_d$ ) with non-zero probability.
3. The MSC-DL does not imposes any restrictions on coalescence between loci.
4. Hence, all branches in all MSC-possible topologies are a valid attachment location.

### 1.2.5 No double counting

Two different topologies represent two unique set of branches. Given a topology, all branches are also unique. This makes lineages being counted by equation 5 unique.

### 1.2.6 Final expression

The total number of branches in the daughter locus where a duplicate can attach is all possible branches in all possible gene trees .

$$\text{Possible Attachment Branches}(N) = \underbrace{(2N-3)!!}_{\text{possible topologies}} \times \underbrace{(2N-1)}_{\text{branches per topology}} \quad (7)$$

## 2 Extra Trees Generated in the 1D + 1L Case

We derive the number of additional gene tree topologies that are possible with one duplication and one loss (1D + 1L + Coalescence), but that cannot occur with 1D + Coalescence .

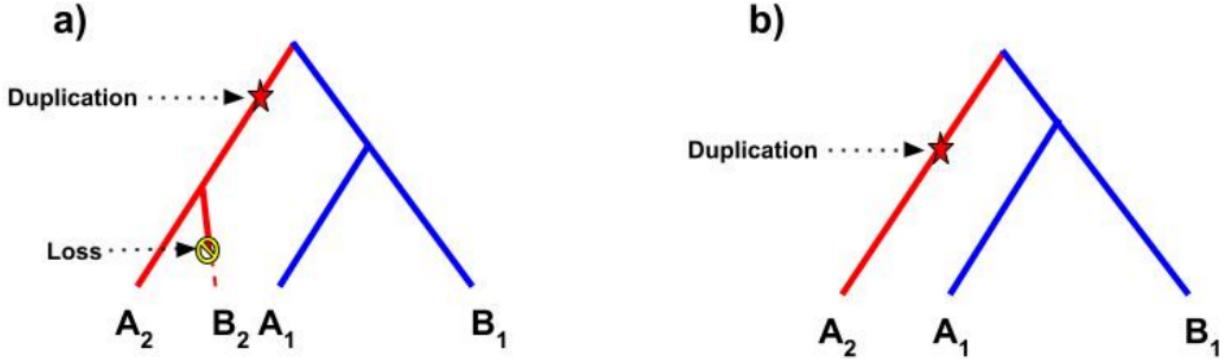


Figure 1: **Loss at the daughter locus can be explained by duplication and coalescence:** a) A duplication on the  $(A_2, B_2)$  lineage and loss of lineage  $B_2$  at the daughter locus, making the history of this gene tree one duplication, one loss, and coalescence. b) The same tree topology can also be produced by just one duplication and coalescence. The duplication of the remaining lineage ( $A_2$ ) in this case will produce the same tree topology.

## 2.1 Observations

A loss in the daughter copy does not produce any new gene trees that cannot already be obtained by 1D+Coalescence. As an example, let loss occur on a branch in the daughter copy on lineage  $B_2$  (Supplementary Figure 1). The resulting tree can also be explained with duplication along just the  $A_2$  lineage plus coalescence to the branch subtending both copies at the parent locus. Every loss at the daughter locus (except loss on the root branch) can be explained this way. Hence, only losses landing on the parent locus create genuinely new topologies. Therefore, adding a single loss after a duplication will only change the attachment branches (parent locus) and will not affect the insertion branches (daughter locus).

## 2.2 Possible size of the remaining parent lineage

On any parent locus with  $N$  species there are three kind of branches, with losses landing on each of these branches producing different classes of lineages based on clade size,  $k$ .

1. A loss on an external branch will give a remaining clade of size  $k = N - 1$ .
2. A loss on an internal branch subtending  $M$  taxa will give a remaining clade of size  $k = N - M$ .
3. A loss on a root branch will result in a clade of size  $k = 0$ . We will ignore counting this case as it is the loss of the entire parent locus.

Thus the parent copy can produce clade of sizes  $S \in \{0, \dots, N - 1\}$ .

## 2.3 Extra attachment branches created by a loss

For each surviving clade in a parent locus of size  $k$  there are  $(2k - 3)!!$  rooted gene trees. Each tree contains  $(2k - 1)$  branches excluding the root, and 1 root branch. Thus, the total number of extra attachment branches in the parent locus created by a single loss after a duplication is:

$$\sum_{k=1}^{N-1} \binom{N}{k} (2k - 3)!! (2k - 1) \quad (8)$$

This equation is counting all the possible subsets that can be formed with one loss (explained in section 2.2), along with all the trees that can be formed from those subsets and all the branches within those trees.

## 2.4 Number of possible insertion branches

Since a loss landing on a daughter copy does not add any new trees that 1D+Coalescence cannot produce, the number of possible insertion branches remains the same (see section 2.1)

## 2.5 Loss of the daughter copy

Extra lineages are also created when the loss lands on the root branch of the daughter locus. This creates no lineages at the daughter locus, resulting in the tree at the parent locus being the final gene tree. Thus for  $N$  species we add  $(2N - 3)!!$  extra gene trees from parent locus alone if there is loss of the entire daughter copy.

## 2.6 Final expression

For each attachment branch generated by loss, we must multiply it with the number of possible insertion branches to get the total number of extra trees generated by adding 1 loss:

$$\text{ExtraGeneTrees}(N) = \left( \sum_{k=1}^{N-1} \binom{N}{k} (2k-3)!! (2k-1) \right) \left( \sum_{k=1}^N \binom{N}{k} (2k-3)!! \right) + (2N-3)!! \quad (9)$$

### 2.6.1 Example when $N = 2$

When  $N = 2$ , we obtain the following extra trees:

$$\left( \underbrace{\{A_1, B_1\}}_{\text{losses in leaves}} \right) \times \left( \underbrace{\{A_2, B_2, (A_2, B_2)\}}_{\text{all duplication lineages}} \right) + \underbrace{\emptyset}_{\text{when daughter copy is lost}} \times \underbrace{\{(A_1, B_1)\}}_{\text{all gene tree topologies in parent copy}} = 2 \times 3 + 1 = 7$$

#### 1. When duplication happens in root and loss happens in leaves and root

- $(A_1, (A_2, B_2))$
- $(B_1, (A_2, B_2))$

#### 2. When duplication happens in $A_2$ and loss happens in leaves and root

- $(A_1, A_2)$
- $(B_1, A_2)$

#### 3. When duplication happens in $B_2$ and loss happens in leaves and root

- $(A_1, B_2)$
- $(B_1, B_2)$

#### 4. When loss of entire daughter copy

- $(A_1, B_1)$

Table 1: Number of extra trees generated by 1D+1L+Coalescence  $N = 2$  to  $N = 6$

$N$	Extra Attachment Branches	Number of Possible Duplication Branches	Loss of Daughter Copy	Total Trees
2	2	3	1	7
3	12	9	3	111
4	82	37	15	3049
5	710	225	105	159855
6	7596	1881	945	14289021

Table 2: Extra trees when  $N = 2$  grouped by loss location.

Duplication Location	Loss Location	Extra Tree
$(A_2, B_2)$	$B_1$	$(A_1, (A_2, B_2))$
$(A_2, B_2)$	$A_1$	$(B_1, (A_2, B_2))$
$A_2$	$B_1$	$(A_1, A_2)$
$A_2$	$A_1$	$(B_1, A_2)$
$B_2$	$B_1$	$(A_1, B_2)$
$B_2$	$A_1$	$(B_1, B_2)$
Any lineage	Loss of daughter copy	$(A_1, B_1)$

### 3 Symmetries and Reduction of Complexity

In the section above we have shown that the MSC-DL generates an enormous number of topologies. Fortunately, many of these topologies share identical probability expressions. This leads to a reduction of complexity in the number of probability expressions. In this section we explore the different ways in which events can give rise to the same probability expressions.

We define two kinds of symmetries—duplication (insertion) symmetries and attachment symmetries—and argue that in order for two gene trees to have the same probability expression they should be duplication symmetrical (belong to same duplication class) and attachment symmetrical (belong to same attachment class). Throughout this section we justify why many distinct gene tree topologies produced by MSC-DL have identical probability expressions and why the number of unique probability equations is smaller than the total number of topologies.

#### 3.1 Duplication Symmetries

Two lineages are duplication symmetrical if they have same probabilities of duplicating (inserting) at all times in a species tree. This is highly dependent on the shape of the species tree and co-occurrence of the species lineages across different time intervals of the species tree. For example In Supplementary Figure 2 we show two different kinds of duplication symmetries. In Supplementary Figure 2a lineages  $A$  and  $B$  are duplication symmetrical: the probability of duplication inserting at lineage  $A$  or lineage  $B$  is same throughout the tree. But lineages  $(A, B)$  and  $C$  do not have equal probability of duplicating given the species tree, and thus are not duplication symmetrical. This is because the probability of duplication landing on lineage  $C$  is always higher than  $(A, B)$ . This can be reasoned in different regimes (explained below), for instance when the duplication happens before first speciation ( $t_d < t_{s_1}$ ) the probability of duplication inserting into  $C$  lineage is non-zero where as probability of it entering into lineage  $(A, B)$  is zero. Similarly in Supplementary Figure 2b,c we show different kinds of duplication symmetries bases on the shape of species tree. Duplication symmetries mean that duplications on symmetrical lineages always contribute the same probability expression.

##### 3.1.1 Definition

Two lineages  $A$  and  $B$  in the species tree are duplication symmetrical if:

1. In every epoch of the species tree  $A$  and  $B$  have identical probabilities of being observed under the MSC.
2. A duplication event occurring on lineage  $A$  has exactly the same probability as a duplication event occurring on lineage  $B$  (function of time of duplication  $t_d$  and time of speciation  $t_s$ ).

##### 3.1.2 Duplication Symmetry Lemma

If lineages  $A$  and  $B$  are duplication symmetrical then duplications placed on  $A$  and duplications placed on  $B$  contribute identically to the probability expressions of all resulting gene trees.

**Proof sketch** Since their coalescence histories are identical in distribution in all epochs of the species tree, all duplication events modeled by MSC-DL land on them with identical probabilities, therefore contributing equally to resulting gene tree expressions.

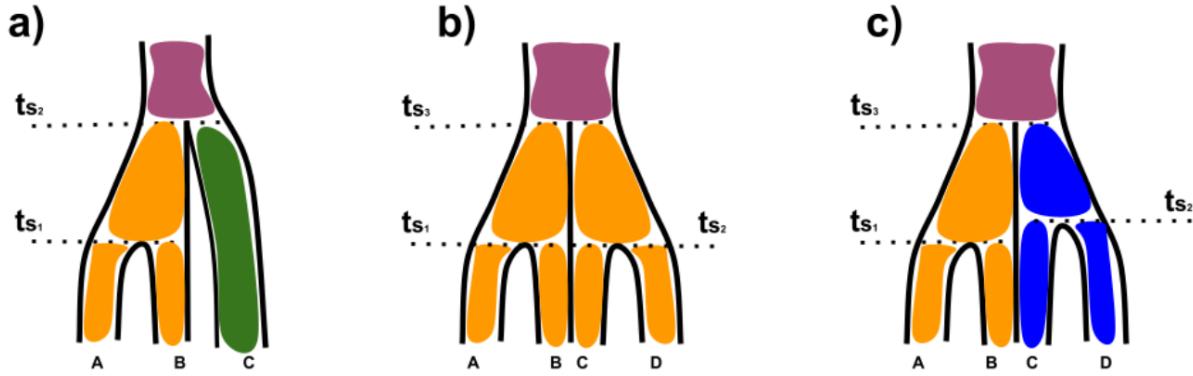


Figure 2: **Duplication Symmetries**

- a)** Here a three-species tree is given with two speciation events marked by  $t_{s_1}$  and  $t_{s_2}$  giving rise three regimes ( $< t_{s_1}$ ,  $t_{s_1} < t_{s_2}$  and  $t_{s_2} <$ ). The probability of seeing lineages **A** and **B** in all regimes is the same, hence they are duplication symmetrical which is marked with orange color. Lineage **C** does not have a duplication symmetrical pair in  $< t_{s_1}$ ,  $t_{s_1} < t_{s_2}$  and  $t_{s_2} <$  regimes.
- b)** Here a four-species tree is given with three speciation events:  $t_{s_1}$ ,  $t_{s_2}$ , and  $t_{s_3}$ , where  $t_{s_1} = t_{s_2}$ . Again this gives three distinct regimes( $< t_{s_1}$ ,  $t_{s_1} < t_{s_3}$  and  $t_{s_3} <$ ). Here, all four lineages **A**, **B**, **C** and **D** have equal probability of being seen in all three regimes (colored orange) making them duplication symmetrical.
- c)** Here a four-species tree is given with three speciation events:  $t_{s_1}$ ,  $t_{s_2}$ , and  $t_{s_3}$ , where  $t_{s_1} \neq t_{s_2}$  giving four distinct regimes ( $< t_{s_1}$ ,  $t_{s_1} < t_{s_2}$ ,  $t_{s_2} < t_{s_3}$  and  $t_{s_3} <$ ). Here lineages **A** and **B** are duplication symmetrical (colored orange) and **C** and **D** are duplication symmetrical (colored blue) but **A,C** or **B,D** are not duplication symmetrical.

### 3.1.3 Example of duplication symmetries

Here we show all the duplication symmetrical branches for the species tree given in Supplementary Figure 2, the probabilities of which are given in Supplementary section 4. Here are all the possible duplication branches in the daughter locus.

$$\{A_2, B_2, C_2, (A_2, B_2), (A_2, C_2), (C_2, B_2), (A_2, (B_2, C_2)), (B_2, (A_2, C_2)), (C_2, (B_2, C_2))\}$$

#### 1. Duplication at Regime 1 ( $t_d > t_{s_2}$ ):

The lineage of daughter locus seen in this regime are:

$\{A_2, B_2, C_2, (A_2, B_2), (A_2, C_2), (C_2, B_2), (A_2, (B_2, C_2)), (B_2, (A_2, C_2)), (C_2, (B_2, C_2))\}$  where  $\{A_2 \text{ and } B_2\}$ ,  $\{(A_2, C_2) \text{ and } (C_2, B_2)\}$  and  $\{(A_2, (B_2, C_2)) \text{ and } (B_2, (A_2, C_2))\}$  are symmetrical (i.e. duplication landing on either of lineage is the same).

#### 2. Duplication at Regime 2 ( $t_{s_2} > t_d > t_{s_1}$ ):

The lineage of daughter locus seen in this regime are:

$\{A_2, B_2, C_2, (A_2, B_2)\}$  where  $\{A_2 \text{ and } B_2\}$  are symmetrical (i.e. duplication landing on either of lineage is the same).

#### 3. Duplication at Regime 3 ( $t_{s_1} > t_d$ ):

The lineage of daughter locus seen in this regime are:

$\{A_2, B_2, C_2\}$  where  $\{A_2, B_2 \text{ and } C_2\}$

## 3.2 Attachment Symmetries

When the duplication occurs the duplicated lineage can attach back to any branch at the parent locus that exists above the duplication time. Some of these branches are also symmetrical: any duplicated lineage can have the same probabilities of attaching with any two branches in the parent locus. We call these branches attachment symmetrical branches or Attachment Symmetries. Attachment symmetries also depends on where the duplication occurred. A duplication on a particular lineage may break some symmetries, while deeper internal branches often remain symmetric regardless

of duplication placement. For example in Supplementary Figure 3, duplication in lineage **A** has different attachment symmetries than duplication in lineage **C**.

### 3.2.1 Definition

Two branches  $\text{Branch}_1$  and  $\text{Branch}_2$  of the parent gene tree are attachment symmetrical if, for a given duplication lineage  $A$ , the duplicated lineage attaches to  $\text{Branch}_1$  and  $\text{Branch}_2$  with equal probability.

$$\Pr((\text{Branch}_1, A_1) | A_1) = \Pr((\text{Branch}_2, A_1) | A_1) \quad (10)$$

### 3.2.2 Attachment Symmetry Lemma

If branches  $\text{Branch}_1$  and  $\text{Branch}_2$  are attachment symmetrical for a given duplication lineage  $A$ , then they contribute identically to the probability expressions of all resulting gene trees.

**Proof** Attachment on a branch  $\text{Branch}_1$  depends on the descendant tree topology and branch lengths below  $\text{Branch}_1$ . If tree topologies and branch lengths are identical below branch  $\text{Branch}_1$  and  $\text{Branch}_2$  down to point  $t_d$ , this means that any lineage will have identical attachment probabilities to both branches.

### 3.2.3 Example of Attachment Symmetries

Here we show all the attachment symmetrical branches when duplication lands on lineage  $A_2$  in species tree given in Supplementary Figure 3. Here are all the possible attachment branches in the parent gene tree:

$\{A_1, B_1, C_1, (A_1, B_1), (A_1, C_1), (C_1, B_1), (A_1, (B_1, C_1)), (B_1, (A_1, C_1)), (C_1, (B_1, C_1))\}$   
the probabilities of which are given in section 4.

#### 1. Attachment at Regime 1 ( $t_d > t_{s_2}$ ):

The lineage of parent gene tree seen in this regime are:

$\{A_1, B_1, C_1, (A_1, B_1), (A_1, C_1), (C_1, B_1), (A_1, (B_1, C_1)), (B_1, (A_1, C_1)), (C_1, (B_1, C_1))\}$ . Where  $\{A_1 \text{ and } B_1\}$ ,  $\{(A_1, C_1) \text{ and } (C_1, B_1)\}$  and  $\{(A_1, (B_1, C_1)) \text{ and } (B_1, (A_1, C_1))\}$  are symmetrical (i.e. attachment of each lineage is the same).

#### 2. Attachment at Regime 2 ( $t_{s_2} > t_d > t_{s_1}$ ):

The lineage of parent gene tree seen in this regime are:

$\{A_1, B_1, C_1, (A_1, B_1)\}$ . Where  $\{A_1 \text{ and } B_1\}$  are symmetrical (i.e. attachment of each lineage is the same).

#### 3. Attachment at Regime 3 ( $t_{s_1} > t_d$ ):

The lineage of parent gene tree seen in this regime are:

$\{A_1, B_1, C_1\}$  where none of the lineage are attachment symmetrical given duplication at  $A_2$ . This is because if attachment occurred ( $t_{s_1} > t_d$ ) then duplication must have occurred ( $t_{s_1} > t_d$ ) too. This means that only the probability of  $A_2$  attaching with  $A_1$  is non-zero at  $t_{s_1} > t_d$ .

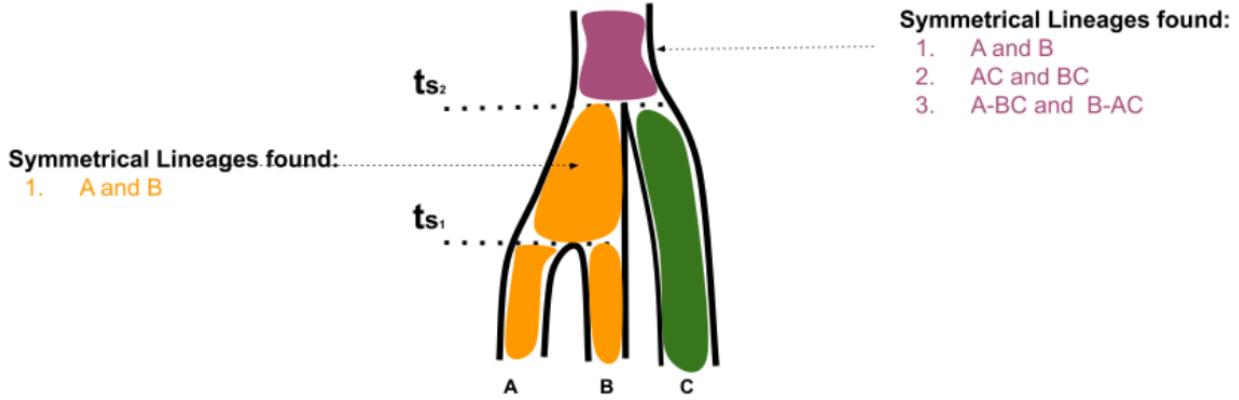


Figure 3: **Attachment Symmetries**

A three-species tree is shown with two speciation events,  $ts_1$  and  $ts_2$  giving rise to three regimes ( $< t_{s_1}$ ,  $t_{s_1} < t_{s_2}$  and  $< t_{s_2}$ ). In some regimes there are symmetrical attachment lineages. For instance, in the purple regime there are three different symmetrical lineages: *A* and *B* are symmetrical with each other, *AC* and *BC* lineages (which can only occur in  $> t_{s_2}$ ) are symmetrical with each other, and  $(A(B, C))$  and  $(B(A, C))$  are also symmetrical with each other.

### 3.3 Impact on Probability Expressions

Two gene tree topologies generated from the MSC-DL share the same probability expression only if their duplications belong to the same duplication symmetry class and their attachments belong to the same attachment symmetry class for that duplication class. This reduces the number of unique gene tree expressions. For example, for two species there were 9 possible topologies, which reduced to 5 unique probability expressions due to symmetry in both duplication and attachment. Similarly, for three species (Supplementary Figure 3) there are 135 possible topologies, which reduces to 63 unique probability expressions. Hence, symmetries can be key in understanding gene trees generated by MSC-DL. It can explain why many topologies behave identically in probability and also reduces the analytical complexity of probability classes without losing any biological reality.

## 4 Probability of Gene Tree Insertion and Attachment

In this section we give an overview of the different types of events that can occur (insertion and attachment) and their probabilities under different scenarios.

## 4.1 Insertion probabilities

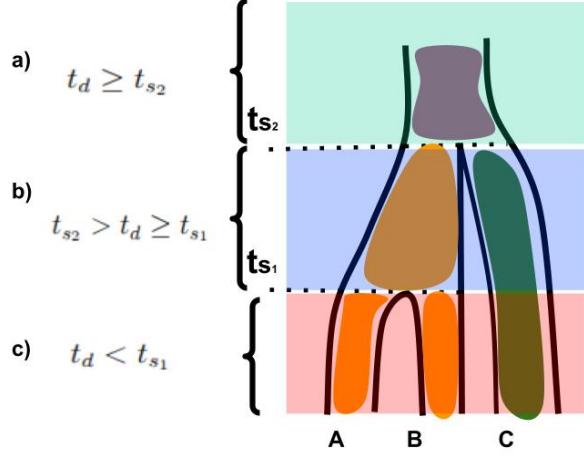


Figure 4: **Probability of Duplication Insertion on a Lineage at Different Times**

We start with a three-species tree, which is defined by two speciation times,  $t_{s_1}$  and  $t_{s_2}$ . **a)** This is the regime in which the time of duplication is greater than the time of the second speciation ( $t_d > t_{s_2}$ ). If a duplication occurs in this regime, it can insert into any lineage on the daughter gene tree:  $\{A_2, B_2, (A_2, B_2), (B_2, C_2), (A_2, C_2), C_2, (C_2, (A_2, B_2)), (A_2, (B_2, C_2)), (B_2, (A_2, C_2))\}$ . The corresponding probabilities are provided in the table below (**Regime 3**).

**b)** This is the regime in which the time of duplication lies between the two speciation events ( $t_{s_2} > t_d > t_{s_1}$ ). If a duplication occurs in this regime, it can insert on lineages  $\{A_2, B_2, (A_2, B_2), C_2\}$  at the daughter locus. The corresponding probabilities are provided in the table below (**Regime 2**).

**c)** This is the regime in which the time of duplication is less than the time of the first speciation ( $t_{s_1} > t_d$ ). If a duplication occurs in this regime, it can insert on  $\{A_2, B_2, C_2\}$  in the daughter locus. The corresponding probabilities are provided in the table below (**Regime 1**).

**Notation.** For the regime  $t_d \geq t_{s_1}$  let

$$s := t_d - t_{s_2}, \quad \Delta := t_{s_2} - t_{s_1}, \quad z := t_d - t_{s_1}$$

$$\Pr(\text{two coalescence happening by } s) = 1 - \frac{3}{2} \cdot e^{-s} + \frac{1}{2} \cdot e^{-3s}$$

$$\Pr(\text{just one coalescence happening by } s) = \frac{3}{2} \cdot (e^{-s} - e^{-3s})$$

### 4.1.1 Regime 1: $t_d < t_{s_1}$

Event	Probability
Duplicate inserting on $A_2$	$\frac{1}{3}$
Duplicate inserting on $B_2$	$\frac{1}{3}$
Duplicate inserting on $C_2$	$\frac{1}{3}$

#### 4.1.2 Regime 2: $t_{s_2} > t_d \geq t_{s_1}$

Event	Probability
Duplicate inserting on $C_2$	$\frac{1}{2}$
Duplicate inserting on $A_2$	$\frac{1}{4} \cdot e^{-z}$
Duplicate inserting on $B_2$	$\frac{1}{4} \cdot e^{-z}$
Duplicate inserting on $(A_2, B_2)$	$\frac{1}{2} \cdot (1 - e^{-z})$

#### 4.1.3 Regime 3: $t_d \geq t_{s_2}$

Event	Probability
Duplicate inserting on $C_2$	$(1 - e^{-\Delta}) \cdot \frac{1}{2} \cdot e^{-s} + e^{-\Delta} \cdot \frac{1}{3} \cdot e^{-3s} + e^{-\Delta} \cdot \frac{1}{2} \cdot \frac{1}{3} \cdot \frac{3}{2} \cdot (e^{-s} - e^{-3s})$
Duplicate inserting on $A_2$	$e^{-\Delta} \cdot \frac{1}{3} \cdot e^{-3s} + e^{-\Delta} \cdot \frac{1}{2} \cdot \frac{1}{3} \cdot \frac{3}{2} \cdot (e^{-s} - e^{-3s})$
Duplicate inserting on $B_2$	$e^{-\Delta} \cdot \frac{1}{3} \cdot e^{-3s} + e^{-\Delta} \cdot \frac{1}{2} \cdot \frac{1}{3} \cdot \frac{3}{2} \cdot (e^{-s} - e^{-3s})$
Duplicate inserting on $(A_2, B_2)$	$(1 - e^{-\Delta}) \cdot \frac{1}{2} e^{-s} + e^{-\Delta} \cdot \frac{1}{2} \cdot \frac{1}{3} \cdot \frac{3}{2} \cdot (e^{-s} - e^{-3s})$
Duplicate inserting on $(C_2, B_2)$	$e^{-\Delta} \cdot \frac{1}{2} \cdot \frac{1}{3} \cdot \frac{3}{2} \cdot (e^{-s} - e^{-3s})$
Duplicate inserting on $(C_2, A_2)$	$e^{-\Delta} \cdot \frac{1}{2} \cdot \frac{1}{3} \cdot \frac{3}{2} \cdot (e^{-s} - e^{-3s})$
Duplicate inserting on $((A_2, B_2), C_2)$	$(1 - e^{-\Delta}) \cdot (1 - e^{-s}) + e^{-\Delta} \cdot \frac{1}{3} \cdot \left(1 - \frac{3}{2} \cdot e^{-s} + \frac{1}{2} \cdot e^{-3s}\right)$

<b>Event</b>	<b>Probability</b>
Duplicate inserting on $((C_2, B_2), A_2)$	$e^{-\Delta} \cdot \frac{1}{3} \cdot \left(1 - \frac{3}{2} \cdot e^{-s} + \frac{1}{2} \cdot e^{-3s}\right)$
Duplicate inserting on $((C_2, A_2), B_2)$	$e^{-\Delta} \cdot \frac{1}{3} \cdot \left(1 - \frac{3}{2} \cdot e^{-s} + \frac{1}{2} \cdot e^{-3s}\right)$

Table 6: Probability that a duplication inserts on  $C_2$  in Regime 3.

<b>Event</b>	<b>Probability</b>
$A_2$ and $B_2$ attached by $t_{s_2}$ , AND $C_2$ not attached with $(A_2, B_2)$ by $t_d$ , AND duplication inserts on $C_2$ with probability $\frac{1}{2}$ .	$(1 - e^{-\Delta}) \cdot \frac{1}{2} \cdot e^{-s}$
$A_2$ and $B_2$ not attached in between $t_{s_1}$ to $t_{s_2}$ , AND $A_2, B_2$ and $C_2$ not attached by $t_d$ , AND with probability $\frac{1}{3}$ it inserts on $C_2$ .	$e^{-\Delta} \cdot \frac{1}{3} \cdot e^{-3s}$
$A_2$ and $B_2$ not attached in between $t_{s_1}$ to $t_{s_2}$ , AND there was precisely one coalescence(attachment) before $t_d$ , AND with probability $\frac{1}{3}$ , $C_2$ is not involved in first attachment, AND with probability $\frac{1}{2}$ it inserts on $C_2$ .	$e^{-\Delta} \cdot \frac{1}{2} \cdot \frac{1}{3} \cdot \frac{3}{2} \cdot (e^{-s} - e^{-3s})$

## 4.2 Attachment probabilities

We provide detailed calculations for the case in which insertion has occurred on branch  $C_2$  and it attaches to branch  $C_1$  when  $t_d < t_{s_1}$ . This example demonstrates all of the dynamics that can occur during attachment.

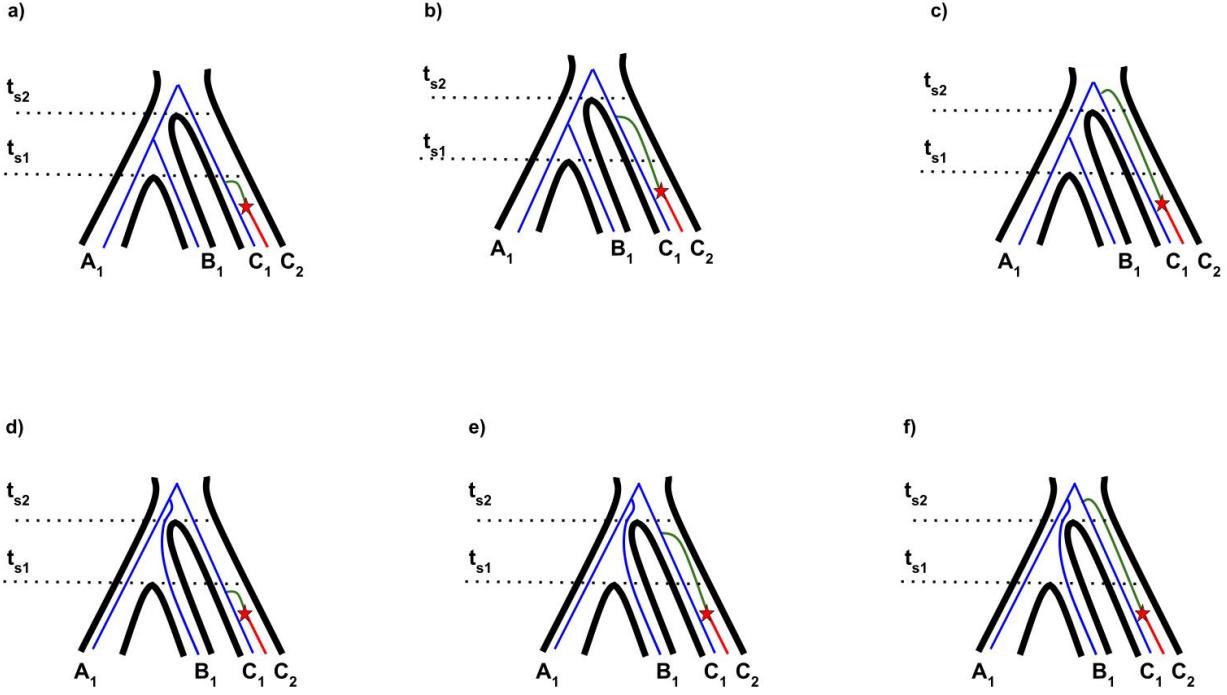


Figure 5: Given that duplication inserts on  $C_2$  in regime 1 ( $t_d < t_{s_1}$ ) we show the probability of generating  $((A_1, B_1), (C_1, C_2))$ :

- a,d) Case where  $C_2$  attaches to  $C_1$  in regime 1. The probability of this happening is given in Table 4.2.1.
- b,e) Case where  $C_2$  attaches to  $C_1$  in regime 2. The probability of this happening is given in Table 4.2.2.
- c,f) Case where  $C_2$  attaches to  $C_1$  in regime 3. The probability of this happening is given in Table 4.2.3.

### Notation.

Let  $\Delta := t_{s_2} - t_{s_1}$ . For Regime 1 let  $u := t_{s_1} - t_d$ .

#### 4.2.1 Attachment at Regime 1

Figure 5	Event	Probability
a)	$C_1$ and $C_2$ attached by $t_{s_1}$ AND $A_1$ and $B_1$ attached by $t_{s_2}$	$(1 - e^{-u}) \cdot (1 - e^{-\Delta})$
d)	$C_1$ and $C_2$ attached by $t_{s_1}$ AND $A_1$ and $B_1$ not attached by $t_{s_2}$ AND $A_1$ and $B_1$ attached in regime 3.	$(1 - e^{-u}) \cdot e^{-\Delta} \cdot \frac{1}{\binom{3}{2}}$

#### 4.2.2 Attachment at Regime 2

Figure 5	Event	Probability
b)	$C_1$ and $C_2$ not attached by $t_{s_1}$ AND $C_1$ and $C_2$ attached between $t_{s_1}$ and $t_{s_2}$ AND $A_1$ and $B_1$ attached by $t_{s_2}$	$e^{-u} \cdot (1 - e^{-\Delta}) \cdot (1 - e^{-\Delta})$
e)	$C_1$ and $C_2$ not attached before $t_{s_1}$ AND $C_1$ and $C_2$ attached between $t_{s_1}$ and $t_{s_2}$ AND $A_1$ and $B_1$ not attached by $t_{s_2}$ AND $A_1$ and $B_1$ attached in regime 3.	$e^{-u} \cdot (1 - e^{-\Delta}) \cdot e^{-\Delta} \cdot \frac{1}{\binom{3}{2}}$

#### 4.2.3 Attachment at Regime 3

Figure 5	Event	Probability
c)	$C_1$ and $C_2$ not attached by $t_{s_2}$ AND $C_1$ and $C_2$ attached after $t_{s_2}$ when $A_1$ and $B_1$ have attached by $t_{s_2}$	$e^{-u} \cdot e^{-\Delta} \cdot (1 - e^{-\Delta}) \cdot \frac{1}{\binom{3}{2}}$
f)	$C_1$ and $C_2$ not attached before $t_{s_2}$ AND $C_1$ and $C_2$ attached after $t_{s_2}$ AND $A_1$ and $B_1$ not attached by $t_{s_2}$ WHEN the first attachment is $(A_1, B_1)$ and second attachment is $(C_1, C_2)$ or vice-versa.	$e^{-u} \cdot e^{-\Delta} \cdot \frac{2}{\binom{4}{2}} \cdot \frac{1}{\binom{3}{2}}$

### 4.3 Probability of tree $((A_1, B_1), (C_1, C_2))$ when $t_d < t_{s_1}$

$$\Pr(\text{Duplication inserting at } C_2 | t_d < t_{s_1}) * \Pr(\text{Attachment at } C_1 | \text{ Duplication at } C_2) \quad (11)$$

$$\Pr(\text{Duplication inserting at } C_2 | t_d < t_{s_1}) = \frac{1}{3} \quad (12)$$

$$\begin{aligned} \Pr(\text{Attachment at } C_1 | \text{ Duplication at } C_2) &= (1 - e^{-u}) e^{-\Delta} \frac{1}{\binom{3}{2}} + (1 - e^{-u})(1 - e^{-\Delta}) \\ &\quad + e^{-u}(1 - e^{-\Delta})(1 - e^{-\Delta}) + e^{-u}(1 - e^{-\Delta})e^{-\Delta} \frac{1}{\binom{3}{2}} \\ &\quad + e^{-u}e^{-\Delta} \frac{2}{\binom{4}{2}} \frac{1}{\binom{3}{2}} + e^{-u}e^{-\Delta}(1 - e^{-\Delta}) \frac{1}{\binom{3}{2}}. \end{aligned} \quad (13)$$

$$\Pr(((A_1, B_1), (C_1, C_2)) | t_d < t_{s_1} \text{ AND attachment at } C_1) = \text{equation 12} \times \text{equation 13} \quad (14)$$

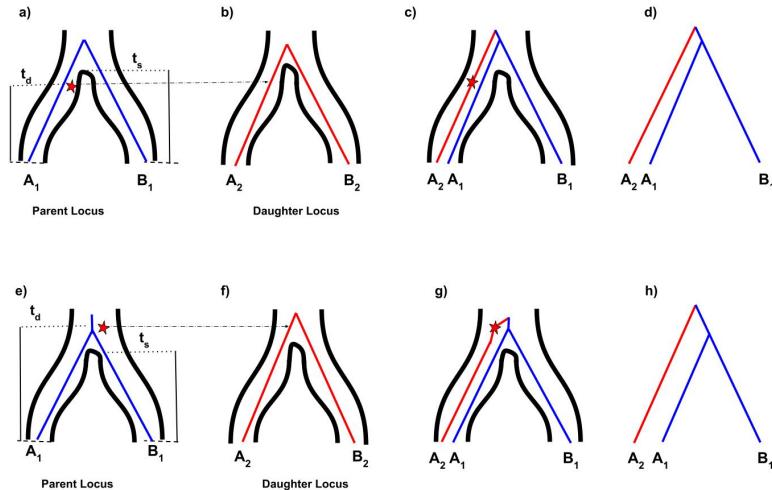


Figure 6: Ways to generate  $(A_2, (A_1, B_1))$ . **a-d**) When the time of duplication is less than the time of speciation ( $t_d < t_s$ ) and the duplication inserts on the  $A$  lineage, the only way to obtain the tree  $(A_2, (A_1, B_1))$  is if the daughter copy  $A_2$  fails to coalesce with the parent copy  $A_1$  before  $t_s$  and enters the ancestral population. Once  $A_2$  enters the ancestral population, the daughter copy can attach with  $(A_1, B_1)$  generating the required tree. **e-h**) When the time of duplication is more than the time of speciation ( $t_d \geq t_s$ ) and the duplication occurs in the  $(A, B)$  population, the only way to obtain the tree  $(A_2, (A_1, B_1))$  is if the duplication inserts on daughter branch  $A_2$ . When the daughter copy attaches with the parent copy above the coalescence time at that locus (i.e. the branch subtending  $(A_1, B_1)$ ), the required tree is generated.

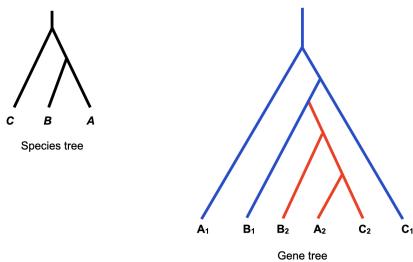


Figure 7: **One duplication and coalescence in a three-species tree.** Here, a single duplicate has occurred and attached to the branch subtending lineages  $A_2, B_2, C_2$  at the daughter locus, with the gene tree at this locus being discordant with the species tree (shown to the left). The duplication has attached to the  $B_1$  lineage at the parent locus, which also has a discordant gene tree. In total, standard DL reconciliation would infer 4 duplications and 8 losses given this gene tree and species tree.