The Multilocus Multispecies Coalescent: A Flexible New Model of Gene Family Evolution

Qiuyi Li[1], Nicolas Galtier[2], Celine Scornavacca[2,†], and Yao-ban Chan[1,*]

[1] *School of Mathematics and Statistics / Melbourne Integrative Genomics, The University of Melbourne, Melbourne, 3010, Australia*

[2] *Institut des Sciences de l'Evolution, Université Montpellier, CNRS, IRD, EPHE, Montpellier, 34095, France*

[*] *yaoban@unimelb.edu.au*, [†] *celine.scornavacca@umontpellier.fr*

Abstract

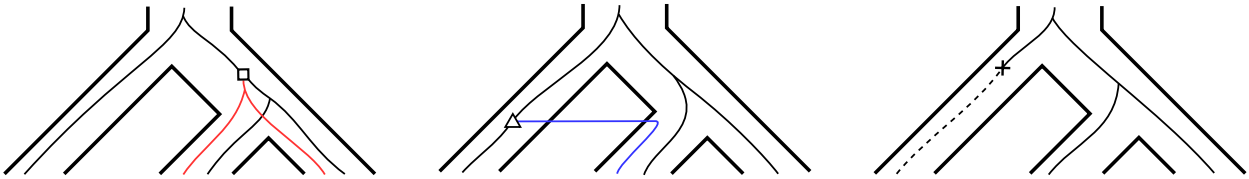
Incomplete lineage sorting (ILS), the interaction between coalescence and speciation, is a process which generates incongruences between gene trees and species trees, as do gene duplication (D), transfer (T) and loss (L). These processes are usually modelled independently, but in reality, ILS can affect gene copy number polymorphism, i.e., interfere with DTL. This has been previously recognised, but not treated in a satisfactory way, mainly because DTL events are naturally modelled forward-in-time, while ILS requires backward-in-time coalescent modelling. Here we consider the joint action of ILS and DTL on the gene tree/species tree problem in all its complexity. In particular, we show that the interaction between ILS and duplications/transfers only can result in patterns usually interpreted as resulting from gene loss, and that the realised rate of D, T and L becomes non-homogeneous in time when ILS is taken into account. We introduce algorithmic solutions to these problems. Our new model, the *multilocus multispecies coalescent* (MLMSC), which also accounts for any level of linkage between loci, generalises the multispecies coalescent model and offers a versatile, powerful framework for proper simulation and inference of gene family evolution.

Key words: Gene duplication, gene loss, horizontal gene transfer, incomplete lineage sorting, multispecies coalescent, hemiplasy, recombination

Species trees and gene trees are two important kinds of phylogenetic trees which are key to the study of gene and genome evolution. A species tree depicts the evolutionary history of a set of organisms, whereas a gene tree depicts the evolutionary history of a gene family within a set of organisms. Species trees have a straightforward interpretation, with internal nodes simply representing speciation events, and branch lengths representing divergence times/amounts. Gene trees, in contrast, reflect a complex evolution potentially involving many diverse processes, such as variation in gene copy number or gene transfers between species, in addition to speciations. For this reason, gene trees often differ from species trees, which can be seen as a problem — if the goal is to infer the species tree — or a source of information — if the goal is to study molecular evolution.

These evolutionary processes can include several 'gene-range' events such as gene duplications, gene losses, and horizontal gene transfers. A gene duplication (D) is an event in which a single gene copy gives rise to two copies at distinct loci: the parent locus and a new (child) locus. In contrast, a gene loss (L) removes a gene from the genome. Horizontal gene transfer (T) occurs when a gene from one species enters the genome of another contemporary species, which can occur (frequently in bacteria, for example) through a number of biological mechanisms such as transformation, transduction and conjugation. Collectively, we refer to these processes as 'DTL'. These events can occur multiple times, allowing the gene tree to possibly differ greatly from the species tree. Examples of these events are given in Figure 1.

In addition to these events, a gene tree can also be incongruent to the species tree due to a phenomenon known as incomplete lineage sorting (ILS, Maddison, 1997). When a population of individuals undergoes several speciations in a relatively short time, polymorphism (different alleles) maintained throughout this time may eventually fix in different descendant lineages. This can produce discrepancies between the gene tree and species tree. ILS is more likely to occur in branches of the species tree (i.e., ancestral species)



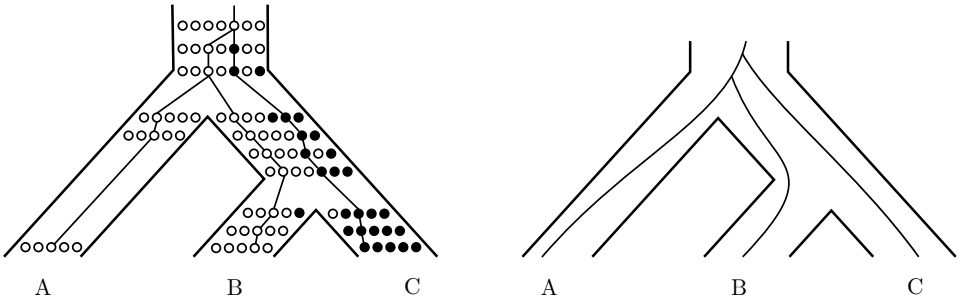
(a) Gene duplication. The duplication is represented by a white square, and the new gene lineage created by the duplication is coloured red.

(b) Horizontal gene transfer. The transfer origin is represented by a white triangle, and the new gene lineage created by the transfer is coloured blue.

(c) Gene loss. The loss event is represented by a cross, and the lost lineage is dashed.

Fig. 1: Tree representations of a gene duplication, a horizontal gene transfer, and a gene loss, respectively. The gene lineages (finer lines) evolve within a species tree (outer 'tubes').

that represent small time spans and/or large population sizes (Pamilo and Nei, 1988). An example of ILS is given in Figure 2. Hemiplasy (Avise and Robinson, 2008) is a term used to refer to the species tree/gene tree conflicts that result from incomplete lineage sorting.



(a) Population-genetic view

(b) Gene evolution representation

Fig. 2: Example of incomplete lineage sorting (ILS). The original population contains a single white allele for the gene of interest. First, a mutation leads to a new black allele at the locus, then the first speciation takes place, rapidly followed by a second one. As the white and black alleles still coexist when the second speciation takes place, both alleles may be fixed in separate descendant species, resulting in a gene history which differs from the species history.

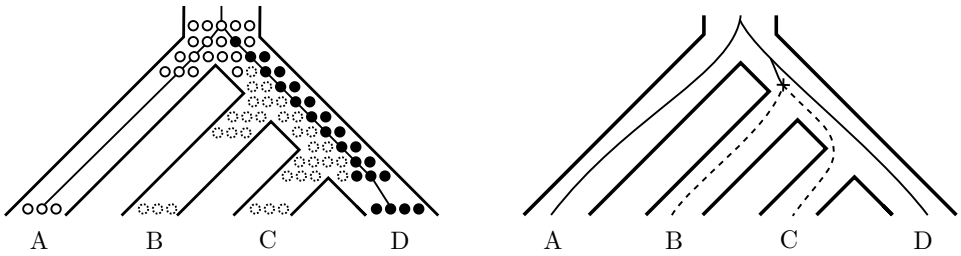
Coalescent theory (Kingman, 1982) provides a genealogical interpretation of ILS that helps connect this phenomenon to gene tree-species tree discordance. A key point is that the age of the common ancestor to two gene copies sampled in two different species is older (in the absence of DTL) than the time of speciation between the two species. This is due to the existence of polymorphism in the ancestral species. When speciations occur far apart in time from each other, ancestral polymorphism only creates differences in branch lengths, not in

topology, between gene trees and species trees. If, however, two or more speciations occur in a time interval of the order of coalescence times, then coalescence and speciation may interact. This can cause not only branch lengths, but also topology, to differ between gene trees and species trees, as in Figure 2.

The multispecies coalescent model (MSC, Rannala and Yang, 2003) predicts the effect of ILS on gene tree branch lengths and topology as a function of the effective population size and timing of speciations. ILS and the MSC have received much attention after the discovery that classical phylogenetic methods are inconsistent in a subset of the parameter space (Degnan and Rosenberg, 2006), and in the context of the study of convergent evolution (Guerrero and Hahn, 2018).

So far ILS has been mainly considered separately from other sources of conflict between gene and species trees. However, ILS can interact in complex and often unintuitive ways with the processes of gene duplication, loss, and transfer, as first noted by Rasmussen and Kellis (2012). This is because DTL events spend some time in a polymorphic stage in a population, when individuals differ from each other in terms of gene copy number, before they become fixed. If speciations occur during this transient period of polymorphism, then the issue of lineage sorting becomes even more complex than in the one-locus case. For example, an allele which is not yet fixed can be lost, as shown in Figure 3. It is also possible that a newly created locus does not fix in all descendant species, as shown in Figure 4. Thus modelling ILS together with DTL requires greater flexibility than simply modelling each process individually. We refer to the discrepancies in gene copy number that result from the interaction between ILS and DTL as 'copy number hemiplasy' (or CNH for short).

Discrepancies between a gene tree and a species tree due to DTL and/or ILS are frequently analysed via mappings from the gene tree into the species tree, called *reconciliations* (Goodman et al., 1979). More formally, a reconciliation between a gene tree G and species tree S is a mapping of the nodes of G into the nodes (or a sequence of nodes) of S (Doyon et al., 2011), respecting some constraints that depend on the evolutionary



(a) Population-genetic view (b) Gene evolution representation

Fig. 3: Example of ILS interacting with loss. With only one loss event, two descendant species may end up with an empty locus due to the presence of ILS.

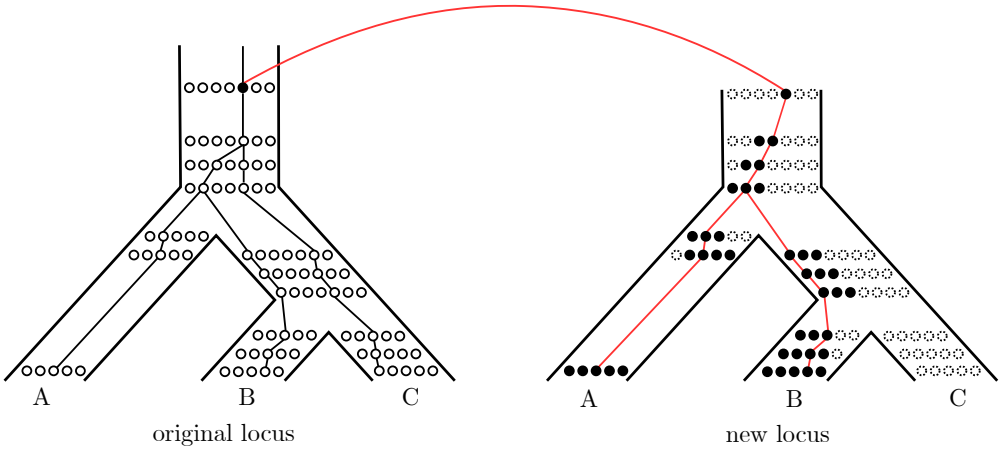


Fig. 4: Example of copy number hemiplasy. A duplication arises in an ancestral species, and then two successive speciation events occur. This results in some species retaining two copies of the gene, while others have only one, without any gene loss event. Here, the original gene is sorted in all descendant species (A, B and C) in the original locus, but the duplicated gene only fixes in species A and B. Species B ends up with the same number of gene copies as A, while being more closely related to C.

model chosen. This gives rise to the problem of reconciliation inference, where we seek to reconstruct the 'true' reconciliation from the gene and species trees.

There are two main paradigms for reconciliation inference: the parsimonious and the probabilistic (Doyon et al., 2011). In the parsimonious approach, we assign a cost to each evolutionary event and search for the most parsimonious reconciliation, i.e., the one which induces the lowest total cost. In the probabilistic approach, a stochastic model of evolution is assumed, and either the reconciliation with the maximum likelihood under this model is

found, or a Bayesian approach is used to sample the posterior reconciliation space. In general, the probabilistic approach is more accurate, but less time efficient.

In our discussion, it is important to distinguish between three concepts:

- The *model* of evolution — a specification of what can (and cannot) happen in the evolution of a gene family;
- A statistical specification of the model, which we call here the *implementation* of the model, which is necessary to support a probabilistic method; and
- A *method* to choose an optimal reconciliation under some criterion (parsimonious or probabilistic).

Every reconciliation method assumes a particular model of gene family evolution, limiting the potential sources of gene tree-species tree discrepancy. However, the underlying model is not always explicitly specified, which is necessary for a proper understanding and comparison of reconciliation methods.

The first reconciliation models only considered gene duplications and losses as the source of discrepancy (Goodman et al., 1979; Zhang, 1997; Arvestad et al., 2004; Durand et al., 2006; Rasmussen and Kellis, 2010). More recently, horizontal gene transfer has been included in these models (Arvestad et al., 2009; Doyon et al., 2010; David and Alm, 2011; Tofigh et al., 2011; Sjöstrand et al., 2013), increasing both their complexity and realism. Statistically, these models are usually implemented by birth-death processes running inside the species tree. On the other hand, the multispecies coalescent (Rannala and Yang, 2003) implements a model where discrepancies between gene and species trees are only due to ILS.

However, only a few models (and corresponding methods to infer reconciliations) exist which attempt to unify these processes. The DLCoal and DLCpar methods (Rasmussen and Kellis, 2012; Wu et al., 2014; Du et al., 2019; Mawhorter et al., 2019) consider ILS together with duplications and losses, overlaid on a model called the *locus tree model*. SimPhy (Mallo et al., 2015) is a simulator based on the locus tree model that additionally considers

transfers. Schrempf and Szöllősi (2018) also consider these possible events, but use a different model that we call here the *haplotype tree model*. Lastly, the IDTL method (Chan et al., 2017) and Notung (Stolzer et al., 2012) incorporate both ILS and transfers with duplications and losses, but again using different underlying models. Each of these models are slightly different from each other, and some of the papers (Chan et al., 2017; Du et al., 2019) compare the various models. However, they all have their limitations, which are discussed in Section "Existing Models of Evolution". In particular, none of them can appropriately model copy number hemiplasy, although CNH is at the heart of the ILS/DTL interaction.

Modelling ILS together with DTL furthermore implies considering the issue of genetic linkage between gene copies. A new gene copy arising from a duplication often appears close to the parent gene on the chromosome, in which case the two loci are expected to follow correlated coalescent processes. The strength of the correlation is controlled by the amount of recombination between the two loci; in the absence of recombination, the two genealogies will be the same. The joint coalescent process of partly linked loci in a single population is well characterised (Hudson, 1983), but the connection with speciation and ILS has only been rarely considered so far, the model of Slatkin and Pollack (2006) being one notable exception. (See Section "Generating a Haplotype Forest" for details about their model.) A realistic model of gene family evolution should account for the possible existence of linkage between duplicated gene copies.

In this paper, we propose a new gene family evolution model, called the *multilocus multispecies coalescent* (MLMSC). This model generalises the multispecies coalescent to gene families, and is designed to capture all possible scenarios that can arise through ILS, DTL, and any interaction between these processes. The MLMSC combines forward- and backward-in-time modelling in order to properly account for copy number hemiplasy and linkage between loci. Importantly, we show that the realised rates of D, T and L become inhomogeneous as these processes interact with ILS, and we introduce a solution to this problem. The MLMSC model is more flexible and predicts a more diverse range of biological

patterns than existing models of gene family evolution.

Existing Models of Evolution

We consider the problem of modelling gene family evolution in a phylogenetic context. Given a set of species, we assume that we have sampled exactly one haploid genome per species, and collected in these genomes all the gene copies that descend from one particular gene present at the root of the species tree, i.e., a complete gene family. We aim at modelling the genealogical relationships between these gene copies, assuming that it has been shaped by speciations, duplications, transfers, losses and ILS. In this section, we first review the existing models which have previously addressed this problem.

Locus Tree Model

The reconciliation methods DLCoal and DLCpar are both based on the locus tree model (Rasmussen and Kellis, 2012). In this model, when a gene duplication occurs, the child and parent gene copies evolve independently of each other, i.e., the two loci are unlinked. Biologically, this arises when there is a sufficient level of recombination between the two loci that they can be considered to evolve completely independently.

To implement this model statistically, given a species tree, we first perform a birth-death process for duplications and losses on the species tree. At each duplication, the species tree is copied from that point onwards and attached to the original species tree (where it may be subject to further duplications), whereas the species tree is truncated at each loss. This produces the so-called *locus tree*, which depicts the bifurcating evolution of all loci containing a copy of the gene.

The gene tree is then constructed by applying a multispecies coalescent process within the locus tree, with the caveat that only one gene copy can 'travel' along the branch connecting a duplicated locus back to its parent. In other words, all the lineages in one particular locus must coalesce into one lineage before coalescing with a copy in its sister

locus — the so-called *bounded coalescent process*. This does not permit copy number hemiplasy — a duplicated gene is transmitted to all descendant species at the new locus, barring gene loss. Under this model, ILS does not interfere with gene duplication and loss in generating variation in gene copy number among species.

An example of this model is given in Figure 5, in which a new locus (with descendants in B and C) is created by duplication. The duplicated gene is then fixed in all species at the new locus.

It is easy to extend this formulation to include gene transfers, and indeed SimPhy (Mallo et al., 2015) is a simulator that implements the locus model, extended to transfers (note that SimPhy does not provide a method to find an optimal reconciliation).

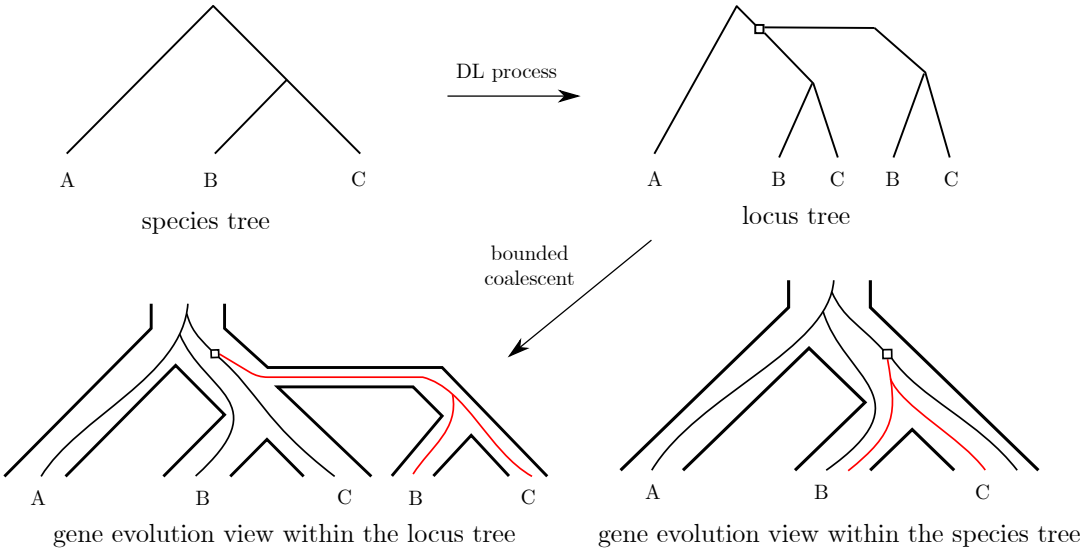


Fig. 5: An example of the locus tree model.

Haplotype Tree Model

The haplotype tree model was introduced by Schrempf in a talk at the SMBE 2018 conference (Schrempf and Szöllősi, 2018). In contrast to the locus tree model, the haplotype tree model assumes that there is no recombination. This implies that a duplicated gene must undergo the exact same genealogical history as its parent gene — a strong assumption.

To implement this model statistically, we first perform a multispecies coalescent process on the species tree, obtaining a so-called *haplotype tree*. The gene tree is then obtained by performing a duplication-loss birth-death process on the haplotype tree. An example is given in Figure 6. Note that, while some sort of copy number hemiplasy is allowed in this model, it is restricted: a gene which duplicates must be sorted into exactly the same descendants as its parent gene. For instance, it is not allowed for a duplicated gene to be sorted into all descendant species, if the parent gene is not.

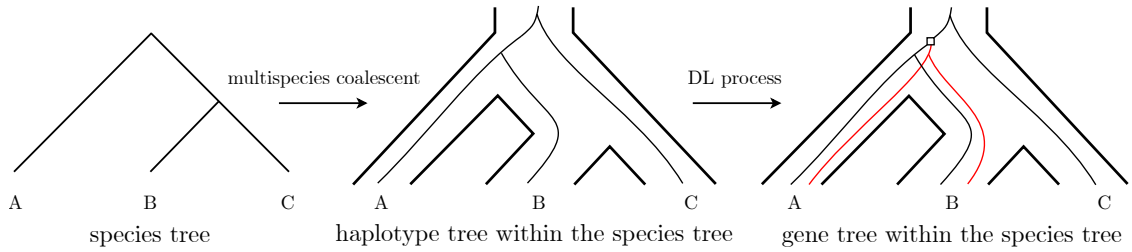


Fig. 6: An example of the haplotype tree model.

IDTL Model

The IDTL reconciliation method (Chan et al., 2017) takes a parsimony approach to the reconciliation problem. It defines events forward in time and assumes an underlying gene family evolution model which shares some elements with both the locus and haplotype tree models. Note that, while the model is specified, no statistical implementation is given in (Chan et al., 2017), since this is not necessary for a parsimony method.

One possible implementation of the underlying model is to perform a multispecies coalescent on the species tree, followed by a DTL birth-death process, as is done in the haplotype tree model. However, for each new locus, a new multispecies coalescent is performed as in the locus tree model. As in the haplotype tree model, a limited form of copy number hemiplasy is allowed where a duplicate gene must be sorted into exactly the same descendant species as the parent gene. An example of this is given in Figure 7.

Note that this model deals with recombination in an inconsistent way; this will be

discussed further in Section "MLMSC vs IDTL Model".

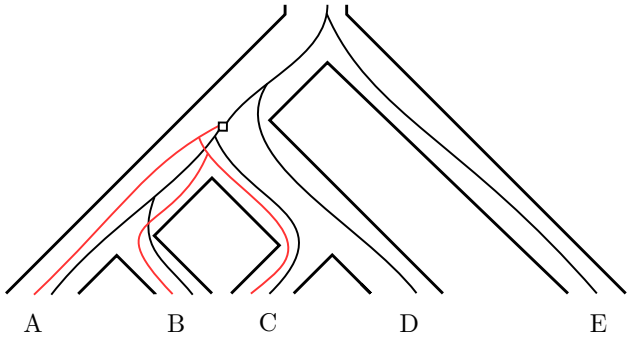


Fig. 7: An example of the IDTL model. An allele which is sorted into species A, B and C is duplicated. The duplicated gene must then be sorted into the same species, but the coalescent process may be different from the parent gene, resulting in a subtree (colored red) with differing topology.

Notung

The Notung model (Vernot et al., 2008; Stolzer et al., 2012) was one of the earliest models to unify the processes of DTL and ILS. It is primarily intended to be used with non-binary species trees, i.e., situations where the branching order is unclear. Because of this, ILS in Notung is only allowed at a polytomy (node with more than two descending lineages) in the species tree, and each possible sorting of genes at the polytomy is considered to be equally likely. Thus there is no explicit modelling of alleles co-existing inside a species branch. As with the IDTL method, Notung was introduced as a parsimony method, and does not have a formal statistical implementation.

The MLMSC Model

The models of gene family evolution in the literature do not appropriately model copy number hemiplasy or linked loci. Here we introduce a new gene family evolution model, the *multilocus multispecies coalescent* (MLMSC), for this purpose. This model, which generalises the multispecies coalescent model to gene families, is an exact coalescent-based simulation of

the genealogies that arise from a Wright-Fisher model in each species branch with constant rates of duplications, transfers, and losses. The major innovations in this model are:

1. Modelling copy number hemiplasy with the *incomplete coalescent*;
2. Modelling linked loci with recombination with the *linked coalescent*;
3. Simulating surviving DTL events under the correct rates with the *coalescent-rate process*.

In this section, we outline each of these ideas in turn. A full specification of the model is left to Section "Full Model Specification". Firstly, we introduce the concepts of *unilocus trees* and *haplotype forests and trees*.

Unilocus trees model the history of duplications, transfers, and speciations. We assume that each duplication or transfer event creates a new locus. This new locus may be inherited by the descendants of the ancestral species, S, in which it was created, and only by these. For each newly created locus, we define a *unilocus* tree as the subtree of the species tree rooted at S, i.e., including S and its descendants. The history of gene duplications, gene transfers, and speciations in the MLMSC is therefore stored as a collection of disconnected unilocus trees. This contrasts with models based on the locus tree (see above), and offers more flexibility in terms of how lineages may be sorted.

Haplotype forests and trees model the genealogies of lineages within a unilocus tree. We consider that we have sampled a single haploid genome per species, meaning that, for a given locus, the number of gene copies carried by a given species is zero or one. We model the genealogical relationships of gene copies across species, for a given locus, via a so-called *haplotype forest*, which we define as a collection of *haplotype trees*, where a haplotype tree depicts the genealogy of gene copies for a subset of species. A haplotype forest contains either one haplotype tree or a set of disconnected haplotype trees, i.e., haplotype trees whose sets of leaves are disjoint. Generally, only one of the trees in the haplotype forest will

actually carry a copy of the gene, in which case we refer to it as *the* haplotype tree for the locus. Haplotype forests and trees will be used to correctly model the presence of different alleles in the populations.

Modelling Copy Number Hemiplasy

To model the interaction between D, T and ILS, and allow for copy number hemiplasy, we introduce a new process called the *incomplete multispecies coalescent*. In this model, when generating the haplotype tree for a given locus, we do not require that all extant genes have to coalesce to their most recent common ancestor (MRCA) by the time of origination, i.e., at the root of the unilocus tree. Rather, we simply stop the coalescent process at the time of origination of the locus, thus producing a haplotype forest. One of these trees is then randomly chosen to be the haplotype tree. The root of this tree represents the first individual carrying the new copy of the gene. In this way, a duplicated/transferred gene copy does not have to be transmitted to all descendant species, and CNH can occur. An example of this process is shown in Figure 8.

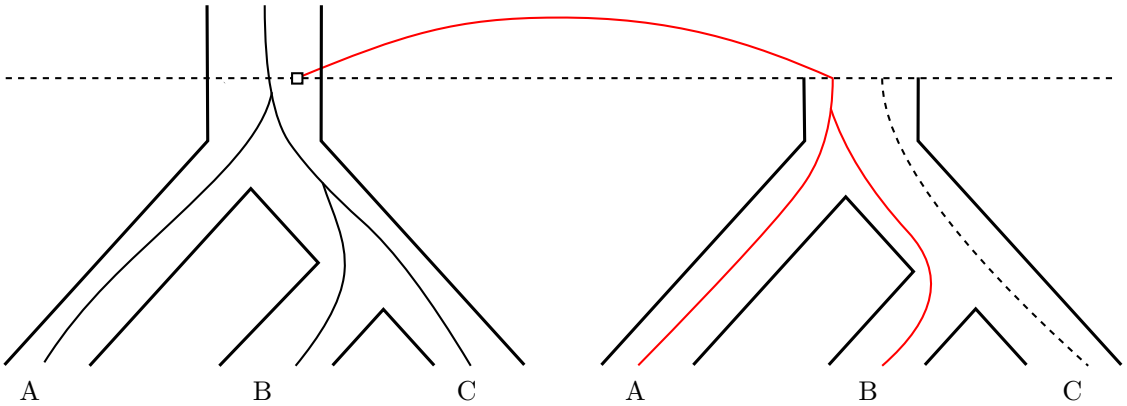


Fig. 8: An example of the incomplete multispecies coalescent. Two trees are generated in the new locus using the multispecies coalescent, and the red tree is randomly chosen to be the haplotype tree for this locus. The root of the red tree within the new unilocus tree represents the individual carrying the new copy.

Not requiring an MRCA to be reached has a further implication, which is that a locus can be entirely lost via sorting effects only. Consider a locus 1 originating at time t_1. Now

assume that there is a duplication in locus 1 at time t_2, resulting in the origination of locus 2. Under the incomplete coalescent, it might be that the duplicating lineage fails to coalesce with any lineages in locus 1 by time t_1. This can be interpreted as a duplication arising in an individual who does not actually contain any gene copy at that locus — i.e., an impossibility. In this case, the duplication is discarded and no new locus is created (see Figure 9). A similar situation can occur with transfers, or a combination of the two types of event.

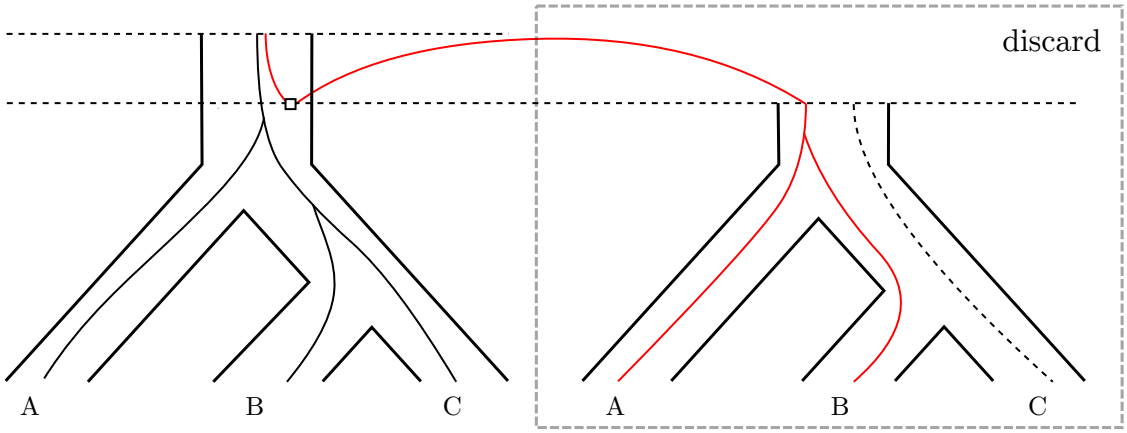


Fig. 9: The new haplotype tree does not coalesce with the haplotype forest in the parent locus by the time of its creation, so the duplication is discarded.

Additionally, to model copy number hemiplasy, we must be able to lose an allele which is not completely fixed in a population. To achieve this, losses are simulated on gene lineages (branches of the haplotype trees), rather than on loci (as is done in the locus tree model).

Modelling Linked Loci

Some duplications, and all transfers, appear in *unlinked* loci; that is, the genealogy of the population (and thus the evolution of the gene) in the new locus is independent from that at the parent locus. This is a reasonable assumption which occurs in many cases, for example with large-scale duplications. However, some duplications may produce a new locus which is physically close to the parent locus on the chromosome, in which case their evolution will be dependent. We call these *linked* loci.

To model linked loci, we introduce a new process called the *linked coalescent* to generate the haplotype forest and tree in a linked locus created by duplication. In this process, the haplotype forest in the parent locus is mirrored into the new unilocus tree and used as a pre-existing genealogy which can be followed by the lineages in the new locus. The incomplete coalescent is then run, during which gene lineages can coalesce with each other, as usual, or with the pre-existing genealogy. In this way, the genealogy of the new locus depends on the genealogy of the existing locus; in the absence of recombination, the two will be identical.

To model the effects of recombination, lineages in the new locus which are coalesced with the pre-existing genealogy can 'uncoalesce' from it, representing a recombination event between the two loci. In contrast, lineages in the new locus which coalesce with each other cannot uncoalesce, because they represent actual lineages. By varying the rate at which recombination events occur, we can model any level of dependence between loci. An example of this process is given in Figure 10.

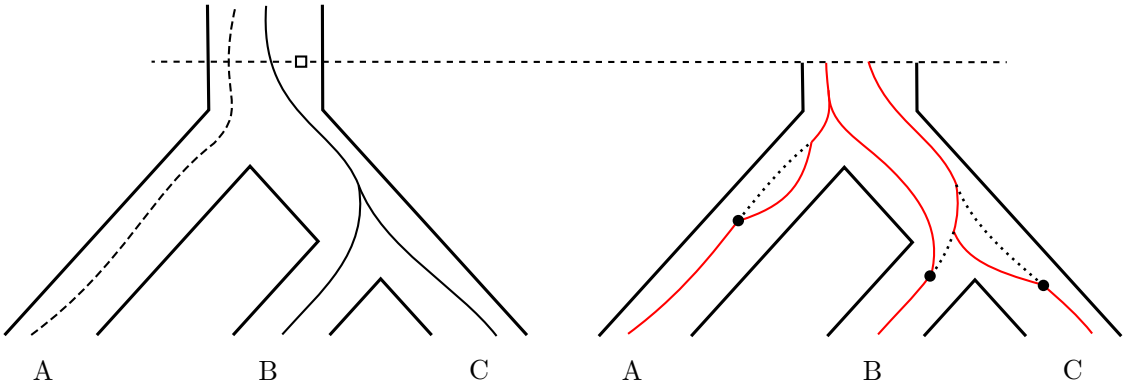


Fig. 10: The linked coalescent. The lineages in the created locus can coalesce with each other, and coalesce and 'uncoalesce' (due to recombination) with the genealogy copied from the parent locus.

A duplication may either be *ancestral* (the duplicating individual is a direct ancestor of a sampled lineage in the genealogy of the parent locus), or *non-ancestral* (the lineage of the duplicating individual does not survive, or is not sampled, in the parent locus). For unlinked loci, it can be assumed that only non-ancestral duplications occur, as the

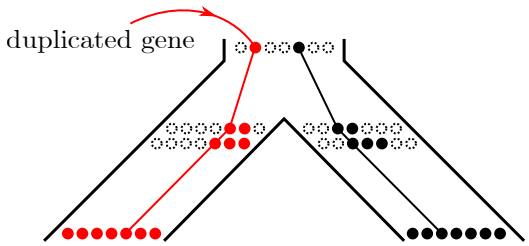
probability of survival in the parent locus, which is independent of the existence of the duplicated locus, is extremely small and can be safely ignored. However, when modelling linked loci, it is important to distinguish between these two cases. By associating lineages in the new locus with lineages in the parent locus, the linked coalescent provides a natural way to determine if a linked duplication is ancestral or non-ancestral.

Simulating Surviving Events

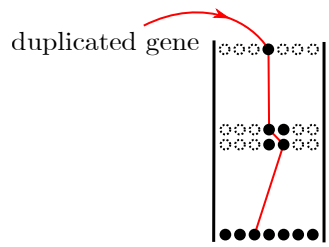
It is important to realise that, here, we do not want to simulate all DTL events that occur, because the vast majority of these events will simply fail to fix in the population, and thus be unobserved. Instead, we only wish to simulate events which survive to the present day and are observed in at least one sampled individual. (We use the terms 'surviving' and 'observed' interchangeably, as a lineage which survives in an unsampled individual is undetectable.) In order to do this, we must consider the probability of survival of each event.

In simpler models (e.g., the locus tree model), the survival probability is constant in time and across lineages, but unrealistic assumptions are needed for this to occur. In the MLMSC model, the survival probability is not constant. This can be seen by considering the following simple example. A duplication occurs 'just before' a speciation into two species leaves. In this case, the duplication can survive because it is observed in either of the descendant species (and possibly in both). Since these probabilities are independent, the total survival probability is roughly twice that of a duplication which occurs in a terminal branch of the species tree (see Figure 11).

In our model, we develop the *coalescent-rate process*, which simulates the exact rate of surviving events, under the Wright-Fisher model, anywhere in the species tree. The simplest version of this process — applied to unlinked duplications — runs an incomplete coalescent in a unilocus tree, and then simulates events at constant rate on the branches of the coalescent trees. The events are then considered to occur in the corresponding branches of the unilocus tree (see Figure 12). This allows us to simulate surviving duplications at the



(a) Duplication 'just before' a speciation. A duplication in either of the filled individuals will survive.



(b) Duplication in a species leaf. Only a duplication in one of the individuals can survive.

Fig. 11: Although duplications occur at a constant rate, they do not survive with constant probability.

correct rate. Some modifications must be made for transfers and linked duplications, and we give further details in Section "Simulating Events".

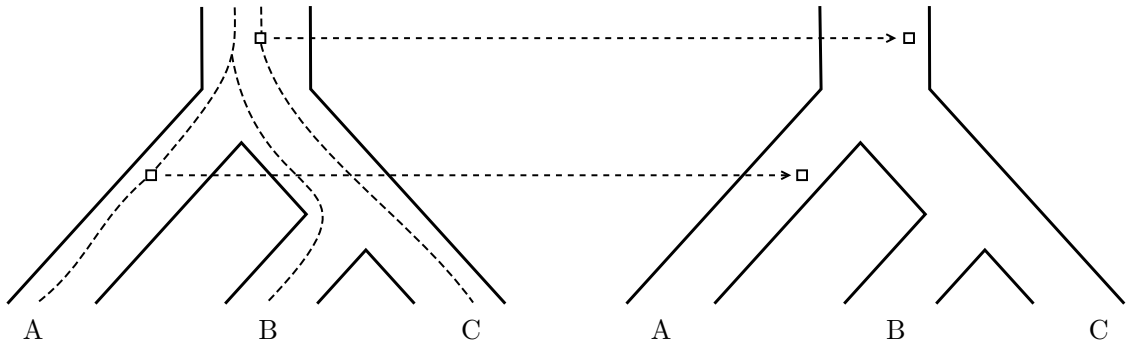


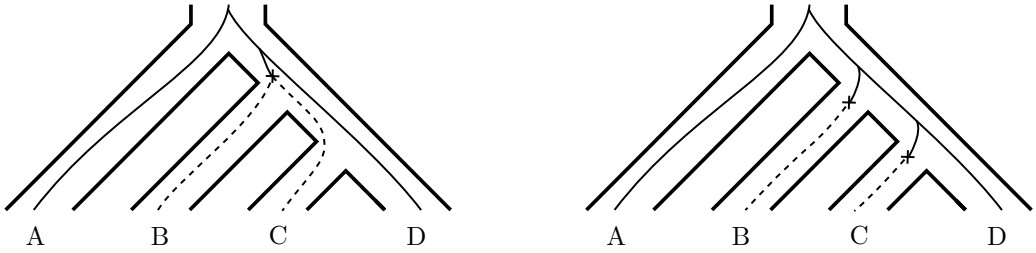
Fig. 12: An example of the coalescent-rate process. Firstly, coalescent trees are sampled from the multispecies coalescent, and then events are sampled at constant rate on the branches of the coalescent trees. Finally, the coalescent trees are removed and the sampled events are considered to occur in the unilocus tree.

Model Comparison

In this section, we discuss the limitations of the current models in the literature, and show how the MLMSC model is subject to none of these limitations.

MLMSC vs Locus Tree Model

The locus tree model simulates losses on the locus tree, instead of the haplotype tree. Assigning a loss to a locus, instead of a branch of the haplotype tree, means that the loss of an allele due to lineage sorting cannot be modelled. Consider Figure 13a: here, two alleles were present in the ancestral population of B, C, and D, one of which was subsequently lost, i.e., replaced by the null allele. The null allele is then sorted into species B and C. The locus tree model can only produce this gene tree by invoking at least two losses, as in Figure 13b. In order to properly capture this scenario, losses must be placed on the haplotype tree rather than the locus (or unilocus) tree, which is exactly what is done in the MLMSC model.



(a) An allele is lost, with the null allele being sorted into species B and C.

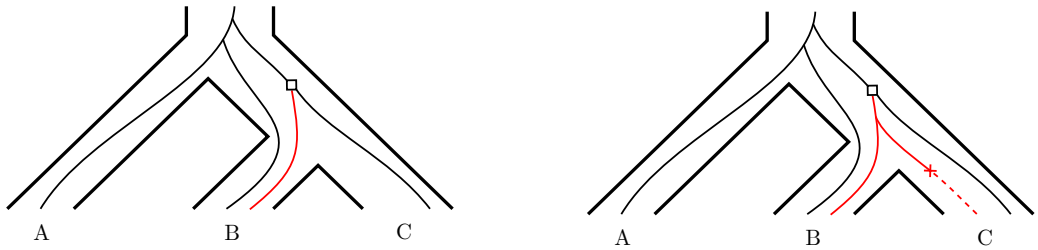
(b) The locus tree model must invoke at least two losses to reproduce this gene tree.

Fig. 13: The locus tree model cannot model lost alleles.

In a similar vein, the locus tree model also assigns duplications to the locus tree. These duplications must then rejoin the haplotype tree via the multispecies coalescent; in other words, they are assumed to be non-ancestral. As discussed above, this is a reasonable consequence of the model assumption that all loci evolve independently (i.e., are unlinked), since the probability of an ancestral unlinked duplication is vanishingly small. However, in the more realistic case that some duplicated loci may be linked to their parent loci, the survival of a duplicated gene is correlated to the survival of its parent gene, and the probability of an ancestral linked duplication, knowing that the new duplicate exists, is non-negligible. Thus it is important to model both ancestral and non-ancestral duplications, as the MLMSC model does.

Finally, in the locus tree model, a duplicated gene is either lost, or fixed in all

possible descendant species. This means there can be no copy number hemiplasy. For example, the locus tree model cannot model the scenario in Figure 14a: an additional loss is needed, as in Figure 14b. By use of the incomplete coalescent, the MLMSC model can model both scenarios.



(a) The duplication fixes only in species B without a loss.

(b) The locus tree model must infer a loss in species C.

Fig. 14: The locus tree model does not model copy number hemiplasy.

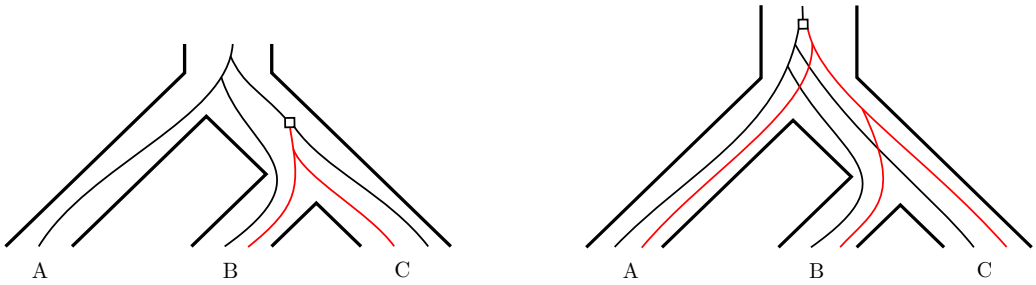
MLMSC vs Haplotype Tree Model

In the haplotype tree model, a duplicated gene is assumed to have exactly the same genealogy as its parent gene; the model cannot model duplicated genes with different coalescent histories. For this to occur, we must assume that there is no recombination between loci, which is too restrictive. It is more realistic and flexible to allow the loci to be linked, where the evolution in the loci are dependent but not necessarily identical, or unlinked, where the evolution are completely independent, as is done in the MLMSC model.

Because the haplotype tree model applies the coalescent first, duplications and losses are assigned to gene lineages, rather than loci as in the locus tree model. It is reasonable to apply duplications to gene lineages when there is no recombination (and thus only ancestral duplications can be observed). But, as discussed above, a more realistic model is to allow recombination, and model both ancestral and non-ancestral duplications, as done by the MLMSC model. On the other hand, the disadvantages of the locus tree model do not apply to the haplotype tree model; for example, it can model both scenarios in Figure 13.

The haplotype tree model also does not fully allow for copy number hemiplasy;

instead, keeping with the assumption of fully dependent loci, it enforces a restricted version wherein a duplicate must undergo the same coalescent process, and therefore be sorted into the same species as the parent gene. For example, the haplotype tree model cannot model the scenario in Figure 15a, where the duplicated gene is sorted into different species than the parent gene. It also cannot model the scenario in Figure 15b, where the duplicated gene undergoes a different genealogy from the parent gene. By allowing recombination between linked loci, the MLMSC model can model both these scenarios even if the duplicated loci are considered to be linked.



(a) The haplotype tree model cannot have a duplication which is sorted into different species from the parent lineage.

(b) The haplotype tree model cannot have a duplication which is sorted in a different way from the parent lineage.

Fig. 15: Limits of the haplotype tree model. Duplicated lineages are in red, while the parent lineages are in black.

MLMSC vs IDTL Model

The IDTL model offers a kind of 'halfway house' between the locus tree and haplotype tree models, where duplications, transfers, and losses are applied to gene lineages rather than loci, and duplicates can be sorted in different ways than their parent genes. However, some of the assumptions made in the model are computationally convenient but biologically questionable — for example, a duplicated gene must be sorted into the same species as its parent gene (i.e, no recombination allowed), but it may be sorted in a different way (i.e, recombination is allowed). For example, Figure 14a is also not allowed by the IDTL model. The MLMSC model is based explicitly on a model of the coalescent with

recombination, and therefore incorporates recombination events in each species separately, allowing a greater range of scenarios.

MLMSC vs Notung

As discussed above, Notung only allows ILS at a polytomy, and each possible sorting of genes at the polytomy is equivalent and equally likely; there is no 'correct' sorting which agrees with a specified species tree. In particular, this means that ILS is not penalised, and so any gene tree-species tree discrepancy which can be attributed to ILS is attributed to it, with the remaining discrepancies then explained by DTL. In contrast, the MLMSC allows ILS everywhere in the tree, with probabilities based on the branch lengths of the species tree, and balances that with the DTL processes. It also specifies an underlying 'true' binary species tree which is always the most likely outcome for the gene sorting. Internal branch lengths in the MLMSC can be made arbitrarily short, which effectively covers what Notung represents by polytomies.

For computational convenience, the Notung model assumes that whenever a transfer occurs, the parent lineage must survive until the present day. This means that transfers must be ancestral. (In reconciliation terminology, this means that there are no 'transfer-loss' events.) As discussed above, this is not a realistic assumption, as only the transferred lineage needs to be observed; in fact, assuming selective neutrality, the chance of the parent lineage also surviving in its unilocus tree is $O(\frac{1}{2N})$, where $2N$ is the effective population size. The MLMSC model has no such restriction and can model transfer events accurately.

Discussion

The MLMSC model generalises the multispecies coalescent to the case of duplications, transfers, and losses. By using the incomplete coalescent, haplotype forests, and disjoint unilocus trees, the MLMSC accounts for scenarios in which ancestral gene copy number polymorphisms are incompletely sorted among the descendant species, in contrast

with existing models. By allowing for both linked and unlinked loci, the MLMSC model can also model more complex evolutionary scenarios than the existing models, while taking recombination into account in a natural way. Finally, the MLMSC recognizes the fact, so far ignored, that the realised rate of duplication, transfer and loss becomes inhomogeneous if ILS is at work. Hence, the MLMSC model is a better alternative for the evolutionary process of genetic evolution within species than all previous models.

An implicit assumption in our model is that we can determine the linked or unlinked status of each locus solely with respect to the locus from which it originated. In this respect, it is not fully cognisant of the linear structure of the chromosome. For example, if a locus (say locus 1) has two linked duplications in rapid succession (say loci 2 and 3), then one of the three loci must lie between the other two. If the order is (say) 1–2–3, then a recombination event between loci 1 and 2 must imply a recombination event between loci 1 and 3; this kind of dependency is not modelled by the MLMSC, which effectively represents the relations between linked loci as a tree rather than a linear structure. Moreover, in a scenario such as this, it may not be reasonable to assume that the recombination rate between loci 1 and 2 is equal to the recombination rate between 1 and 3. A full model which incorporates the linear structure of the chromosome would have to explicitly model the position of the gene copies, which we have elected not to do here for reasons of computational convenience.

The development of the MLMSC model opens up several avenues for further research. Firstly, because the model produces a more realistic and powerful framework for simulating gene family evolution, it can be used to compare the accuracy of current reconciliation algorithms in a ‘neutral’ setting. Secondly, the simulations themselves can be explored under biologically realistic parameter values to ascertain the relative importance of population-level processes to macro-evolutionary ones. Finally, reconciliation methods based on the MLMSC model can now be developed; as we have a statistical implementation of the model, these could either be parsimonious or likelihood-based.

We have implemented a simulator based on this model, available at `https://github.com/QiuyiLi/MLMSC`.

Full Model Specification

In this section, we detail a full statistical implementation of the MLMSC model. The model starts with the species tree as the unilocus tree for the original locus. Then we alternately generate a haplotype tree and forest for that locus, or new loci and unilocus trees, until all simulated duplication and transfer events are accounted for. The haplotype trees are then concatenated together to form the full gene tree. The formal statistical details of these operations follow.

Assumptions

We first clearly set out the assumptions in the model, which helps to distinguish the possible scenarios that can arise. One of the primary conceptual difficulties in unifying ILS with a DTL model has been that ILS is traditionally implemented by a backwards-in-time coalescent model, whereas the birth-death process used to implement DTL events runs forwards in time. The MLMSC model is based on a Wright-Fisher model for the population-level genealogical process, which also runs forwards in time and is therefore easier to merge with a birth-death process. To calculate the genealogy of alleles, the coalescent model is still used in practice, as it is far more efficient than the Wright-Fisher; however, the Wright-Fisher is used to specify the possible scenarios.

In the MLMSC model:

1. The evolution of the gene inside the species is considered forward in time from one common ancestral gene towards the contemporary genes.

2. Duplication (D), horizontal gene transfer (T), and loss (L) are the possible events which can shape gene histories. (ILS is not considered to be an event but rather a

name for the discrepancies arising from population-level processes.)

3. Each species evolves as a panmictic population. Within a population, in each generation each member may pass any (or all) of its genes to any member of the next generation. Several alleles may be maintained in a population in any generation.

4. When a speciation occurs, two independent new populations are created with the same loci as the parent species. Each member of these populations is descended from any member of the parent species. If several alleles are present in the parental population, a sequence of rapid speciations may thus result in topological discrepancies between the gene and species trees, or ILS. (Items 3 and 4 describe the ordinary multispecies coalescent.)

5. When a duplication occurs:

   (a) An individual in a population is chosen to duplicate. If this individual carries a copy of the gene, a new locus is created in the genome at the given species. For this locus, the population contains one individual that exhibits the duplicated gene, while the remainder have no gene copy at this locus.

   (b) The duplication can be of two types:

      i. *unlinked duplication*, where the new locus is unlinked to the parent locus, and thus evolves independently of the parent locus. In particular, the genealogy of the population at the new locus is independent from that at the parent locus;

      ii. *linked duplication*, where the new locus is linked to the parent locus. In this case, the genealogy of the population at the new locus is identical to that at the parent locus, except for when there is a recombination event: when a recombination event occurs, a single individual in the population has different parents in the two loci, breaking the dependency between the loci.

   (c) The duplicated gene may consequently fix in any subset (including none or all) of the descendants of the species containing the duplication (CNH).

6. When a transfer occurs, the same process is followed as for a duplication, except that the transferred gene appears in a new locus of a contemporary species of the origin species, and this locus is always unlinked to any existing locus.

7. When a loss occurs, an individual in a population is chosen to lose a copy of the gene, if it has one. This gene may or may not be an allele. The loss may consequently fix in any subset (including none or all) of the descendants of the species containing the loss (CNH). If a population loses its last copy of a gene in a given locus in a species, that locus is lost in that species.

Generating a Haplotype Tree for the Original Locus

We start with a species tree S. In the original locus (which we denote as locus 0), the unilocus tree S_0 is the original species tree S. For this locus only, the haplotype tree G_0 is generated according to the standard multispecies coalescent, starting from a single copy of the gene in each leaf of the tree. We also set the haplotype forest $\mathcal{G}_0$ to be the set $\{G_0\}$.

Note. For simplicity, we assume (here and in all further coalescents) that the effective population size $2N$ is constant over time and across species. The process can easily be generalised to more variable cases, with the caveat that the population size is a property of the species and must therefore remain the same across loci.

For loci created by duplication or transfer, a more complex process is required. We first describe how to generate new loci and unilocus trees, then return to generating haplotype trees within those unilocus trees.

Simulating Events

Suppose that we are simulating events from a locus l, with unilocus tree S_l, haplotype tree G_l, and haplotype forest $\mathcal{G}_l$. Unilocus and haplotype trees have branch lengths given in coalescent units, and by convention the present day is time $t = 0$, and t increases as we go backwards in time. We are given a set of parameters $\boldsymbol{r} = (r_d, r_t, r_l, r_r, p_u)$,

where r_d, r_t, and r_l are the rates of duplications, transfers, and losses respectively, r_r is the rate of recombination for linked duplications (see Section "Generating a Haplotype Forest"), and p_u is the (fixed) probability that a duplication will be unlinked. The rates r_d, r_t, r_l, and r_r are given in units of events per individual per coalescent unit.

Recall that, as discussed in Section "Simulating Surviving Events", we only wish to model surviving DTL events, not all of them. A duplication may either be ancestral (the duplicating individual is a direct ancestor of a sampled lineage in the parent locus), or non-ancestral. Either of these kinds of duplication can survive, and so we can consider a duplicating individual to be any one of the population, and thus sample duplications from the unilocus tree. Likewise, (surviving) transfers can also be sampled from the unilocus tree.

On the other hand, a loss in a lineage which does not survive to the present day will not be observed; in other words, losses are always ancestral. Hence losses should be sampled from the haplotype tree at constant rate r_l. The constant rate is justified because a population which contains multiple alleles will lose each of them at the same rate as a completely fixed gene, and thus the overall loss rate will be proportional to the number of existing lineages.

In order to sample surviving duplications and transfers, we must consider both the rate at which the events occur, and their probability of being observed. We assume that the events (all events, not surviving events) occur at a constant rate per individual per coalescent unit, as given by the parameters. Thus the overall rate of each event across a population is $2N$ multiplied by the per-individual rate. We then need to scale this rate by the probability of survival of the event. This probability differs for linked duplications, transfers, and unlinked duplications. We consider each in turn.

Simulating unlinked duplications — The locus tree model only considers unlinked duplications. Because no CNH is allowed, the survival probability of a duplication is constant, and since unlinked duplications occur at a constant rate, surviving unlinked duplications can also be simulated at a constant rate. In the MLMSC model, we use the

incomplete coalescent to model CNH, so this property no longer holds.

In the MLMSC model, unlinked duplications occur at a constant rate of $2Np_ur_d$. To calculate the survival probability of an unlinked duplication at a locus l, consider a duplication which occurs at a certain time and species in S_l, and consider the sampled individuals in the extant descendants of this species. Each of these individuals will have a single direct ancestor at the time of the duplication, which may or may not be distinct from each other. The duplication will be observed if and only if one of these ancestors is the duplicating individual in a population of size $2N$.

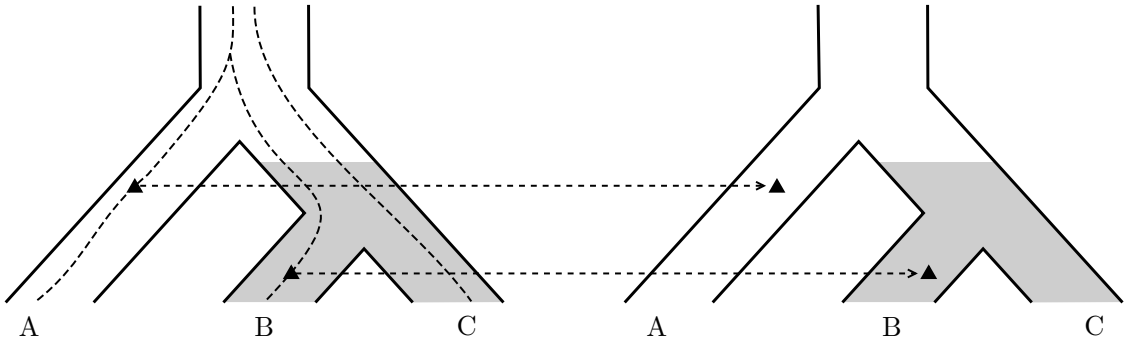
The probability that such a duplication is observed is therefore $\frac{1}{2N}$ multiplied by the expected number of lineages at the time of the duplication, under the multispecies coalescent. It is difficult to obtain a closed formula for this number, but there is a simple way to simulate at the correct rate. To do this, we run a (incomplete) multispecies coalescent within S_l, and then sample events with a constant-rate Poisson process at rate p_ur_d from the branches of the resulting coalescent trees. These events will be considered as duplications at the corresponding branch and time in S_l. An example of this process is given in Figure 12.

At any time, the coalescent trees will, on average, trivially have the expected number of lineages under the coalescent, and so the resulting rate is correct. We call this a *coalescent-rate process*. We must also specify a rate for the coalescent-rate process, which is the rate at which events are sampled from the branches of the coalescent trees (in this case p_ur_d).

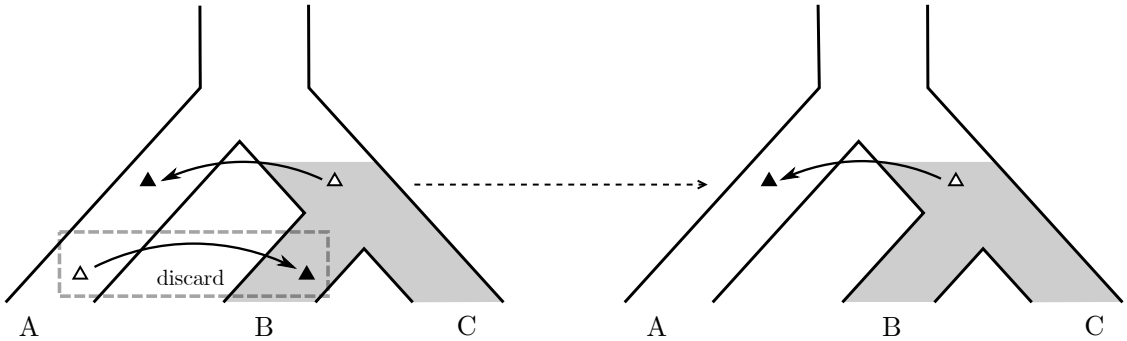
Simulating transfers — Transfers (which are always unlinked) can be simulated using a similar process to that of duplications, but with some modifications. It is important to note that the coalescent-rate process reflects the probability of a gene lineage surviving if it *appears* (by duplication or transfer) in a particular species. A duplication always appears in the same species that it originates from, but for a transfer, this is not the case. Thus the coalescent-rate process must be used to select the target species of the transfer, not its origin species. Additionally, while a transfer from locus l must originate from within S_l, it need not

appear in S_l, but could appear in any species.

In order to simulate surviving transfers from locus l, we thus perform a coalescent-rate process with rate r_t over the *entire* species tree S. Each resulting event marks the target species of a transfer, and we subsequently choose an origin species uniformly at random from *all* other contemporaneous species. If this origin species does not lie in S_l, the entire transfer is discarded. An example of this process is given in Figure 16.



(a) Transfer targets (represented by black triangles) are generated using a coalescent-rate process over the entire species tree.



(b) Transfer origins are simulated from all possible origins. Transfers originating from outside the unilocus tree are discarded.

Fig. 16: Simulating transfers. Here, the unilocus tree is shaded.

In practice, it is less efficient to simulate many transfers and discard some, so we take advantage of the fact that transfers originate from each species at an equal rate, since we assume that all population sizes are equal. Hence, we scale the overall transfer rate to each species by the proportion of possible origin species which are also contained in S_l, and then select an origin species uniformly at random from the possible origin species in S_l. It is easy

to see that this is equivalent to the process described above, but without having to discard some simulated transfers.

Simulating linked duplications — In a linked locus created by duplication, the probability of the duplication surviving is dependent on the genealogy of the parent locus. However, the principle is the same: a duplication will survive if and only if it is the direct ancestor of a sampled lineage in the new locus. The linked coalescent (described in Sections "Modelling Linked Loci" and "Generating a Haplotype Forest") directly models the genealogy in such a locus. Thus, to simulate surviving linked duplications originating from locus l, we run a linked coalescent in the unilocus tree S_l, allowing the lineages to coalesce and uncoalesce (through recombination) with the existing haplotype forest $\mathcal{G}_l$. We then sample linked duplications as events at constant rate $(1 - p_u)r_d$ on the branches of the resulting coalescent forest. These events are then interpreted as duplications at the corresponding branch and time inside S_l. For brevity, we refer to this as a *coalescent-rate process in the presence of* $\mathcal{G}_l$.

Generating new loci and unilocus trees

When a new locus is created, we create a unilocus tree for it, which shows the evolution of all species which could possibly contain the locus (i.e., all descendants of the species where the locus is created). This is the subtree of the species tree starting from the time (and branch) of the creation of the locus.

In order to formally describe this process, we first introduce some notation. Given a tree T, $V(T)$ and $E(T)$ are the sets of their nodes and edges (branches) respectively. For a node $v \in V(T)$, e_v represents the unique edge in $E(T)$ which has v as its target node.

Now, given a unilocus tree S_l and corresponding haplotype tree G_l, we simulate losses as described above, producing a sequence of events $\boldsymbol{e^l} = \{(b_i, t_i)\}$, where $b_i \in E(G_l)$, and $t_i \geqslant 0$ is the time of occurrence of the loss. (We must have $t_2 < t_i < t_1$, where t_1 and t_2

are the times of the top and bottom nodes of branch b_i respectively.) At each event, we decorate G_l at that branch and time. An example of this process is shown in Figure 17.

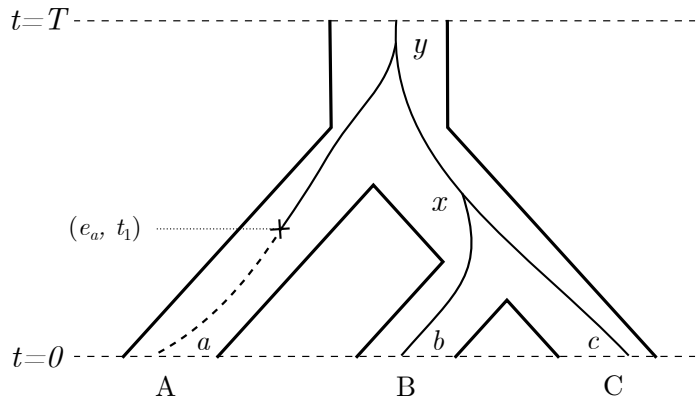


Fig. 17: Losses are sampled from the haplotype tree with constant rate. Here, we have $\boldsymbol{e^l} = \{(e_a, t_1)\}$.

We then simulate (as described above):

1. unlinked duplications in $\boldsymbol{e^{ud}} = \{(b_i, t_i)\}$, where $b_i \in E(S_l)$;
2. transfers in $\boldsymbol{e^t} = \{(b_i, t_i)\}$, where $b_i \in E(S)$;
3. linked duplications in $\boldsymbol{e^{ld}} = \{(b_i, t_i)\}$, where $b_i \in E(S_l)$.

In all cases, the same constraints on t_i given for losses apply. At each duplication, we decorate S_l at that branch and time, and at each transfer, we choose and decorate the origin branch of S_l at that time. An example of the generation of unlinked duplications is given in Figure 18.

Now, the effect of each event is applied in a forwards-in-time order.

At each loss event, the haplotype tree is truncated on branch b_i at time t_i, with further events on the same or descendant branches of G_l (but not S_l) having no effect. (Note that the haplotype forest is *not* truncated.)

At each duplication event $(b_i, t_i) \in \boldsymbol{e^{ud}} \cup \boldsymbol{e^{ld}}$, a new locus m is created, with a unilocus tree S_m which is a copy of the subtree of S_l which starts at time t_i on the species branch b_i (see Figure 19).

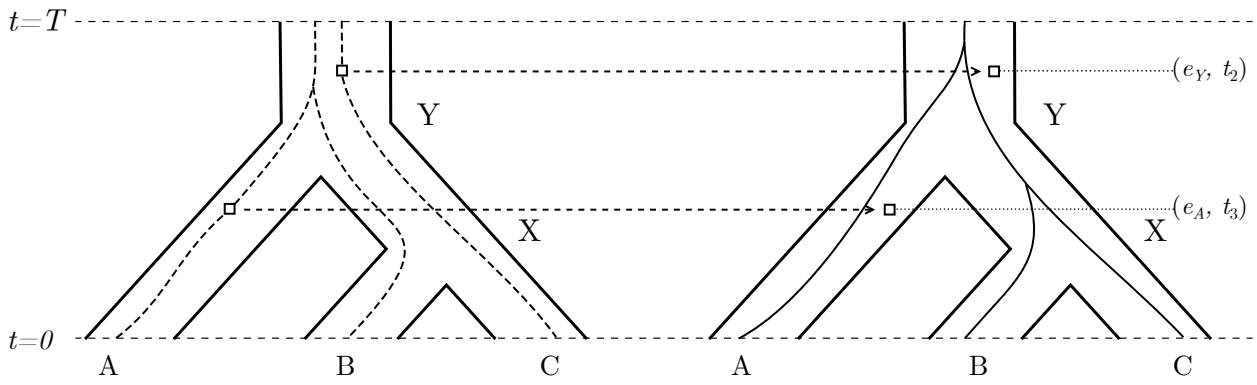
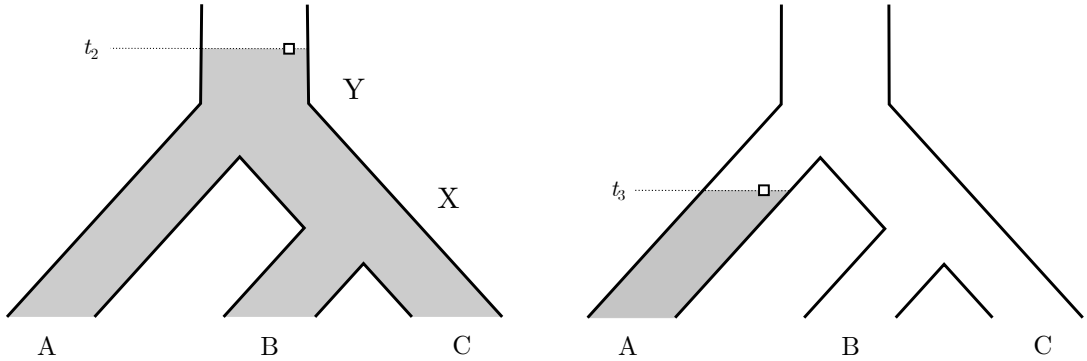


Fig. 18: Duplications are sampled with a coalescent-rate process on the unilocus tree. Here, we have $\boldsymbol{e^{ud}} = \{(e_Y, t_2), (e_A, t_3)\}$.



(a) For the duplication at time t_2, the new unilocus tree is the parent unilocus tree with the branch e_Y slightly shorter, i.e., starting at time t_2 instead of T.

(b) For the duplication at time t_3, the new unilocus tree is the leaf branch e_A of the parent unilocus tree, starting at time t_3.

Fig. 19: The new unilocus trees generated by the events in Figure 18.

Likewise, at each transfer event $(b_i, t_i) \in \boldsymbol{e^t}$, b_i represents the target of the transfer, not the source (as discussed above). The new locus will then have a unilocus tree created, which is a copy of the subtree of S starting from branch b_i at the transfer time t_i.

Generating a Haplotype Forest

Once we have generated the unilocus tree for a new locus, we must then simulate the haplotype tree and forest. If the new locus is unlinked to the parent locus (i.e., it is created by transfer or unlinked duplication), then the two loci evolve completely independently. We

can then use the incomplete coalescent process as specified in Section "Modelling Copy Number Hemiplasy". For the new locus m, the haplotype forest $\mathcal{G}_m$ is the collection of trees produced by the incomplete coalescent, while the haplotype tree G_m is the single tree selected by the incomplete coalescent.

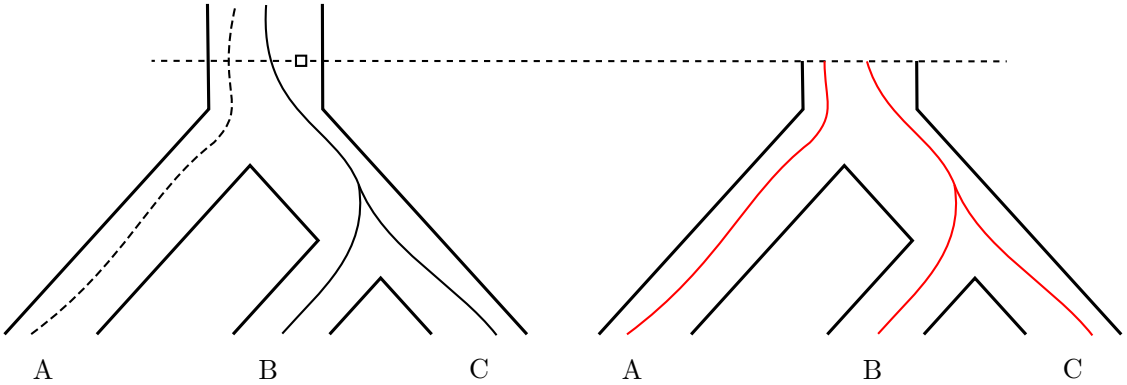
Suppose a locus l has a linked duplication, creating a new locus m. We need to consider that the genealogies of the two loci are dependent, and take this into account when constructing the haplotype tree and forest. We have some genealogical information from the parent locus l, represented by the haplotype forest $\mathcal{G}_l$. Since S_m is a copy of a subtree of S_l, we can identify all subtrees of the haplotype forest $\mathcal{G}_l$ which lie in the subtree of S_l corresponding to S_m. Formally, if locus m is created from an event (b_i, t_i), then we take all subtrees of $\mathcal{G}_l$ which lie in the species branch b_i and its descendants, truncated so that they start at time t_i. These subtrees are then copied to the new locus m (see Figure 20b for an example).

We now run the 'linked multispecies coalescent' in the new unilocus tree S_m, which we now detail. We start from a single lineage in each extant species, as with the ordinary multispecies coalescent. Because there is only one sampled individual per population, the lineage in each species begins coalesced to the copied subtree which has that species as a leaf. By construction, each extant species has one leaf in $\mathcal{G}_l$, so there is always exactly one possibility for this.

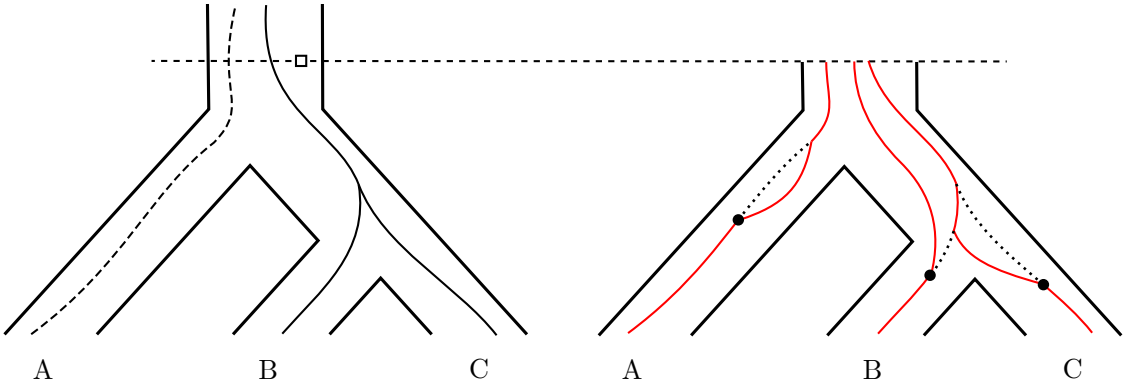
The coalescent then proceeds backwards in time, where the lineages which are coalesced with a copied subtree follow the same genealogy as that subtree. In particular, lineages which are coalesced with copied subtrees will coalesce with each other when the corresponding subtrees coalesce in the parent locus. However, with a constant rate r_r, each lineage which is coalesced with a copied subtree may have a recombination event (i.e., a recombination occurs between the two loci). If there are no recombination events, the haplotype forest in the new locus will be identical to that of the parent locus (Figure 20a).

When a recombination event occurs, the lineage in the new locus becomes

'uncoalesced' from the copied subtree, representing the fact that the individual in the population had different ancestors for the two loci. This lineage now becomes 'free' (not coalesced with a copied subtree) and is also followed backwards in time, where it may coalesce with a copied subtree of $\mathcal{G}_l$, or with another lineage in the new locus (which may itself be coalesced with a copied subtree, or not). Note that lineages in the new locus m which coalesce with each other cannot uncoalesce, since they exist in the same locus; a lineage can only uncoalesce with a copied subtree from the parent locus l, because a copied subtree represents a 'backbone' for the process rather than a lineage of the new locus. An example of this process is given in Figure 20b.



(a) Each lineage begins coalesced with a copied subtree. Here no recombination events occur, so the new haplotype forest is identical to the parent one.



(b) When a recombination event occurs, the lineage 'uncoalesces' from the copied subtree. It may then subsequently coalesce with another lineage or copied subtree.

Fig. 20: The linked multispecies coalescent. The subtrees of the haplotype forest in the parent locus are copied to the new locus (dotted black lines). Then, a new multispecies coalescent (in red) is run, with possible recombination events. Finally, one of the red lineages is chosen to be the root of the new haplotype tree.

This process continues backwards in time, until we reach the time of creation of locus m. This produces the haplotype forest $\mathcal{G}_m$ for the new locus. At this time, we will have a certain number $|\mathcal{G}_m|$ of lineages, some of which may be coalesced with copied subtrees. We choose one uniformly at random, and the descendant tree becomes the new haplotype tree G_m.

This process only produces the haplotype tree and forest in the new locus m. The haplotype tree must then be joined back to the haplotype tree in the parent locus, which we detail in Section "Assembling the Full Gene Tree" below.

Observe how this process produces the appropriate results for either extreme of the recombination rate: if $r_r = 0$ (no recombination), then no recombination events occur, and the haplotype tree G_m in the new locus will be a subtree of $\mathcal{G}_l$ (it is possible however that it will not be a subtree of G_l, in which case the duplication does not carry a copy of the gene and must be discarded; see the next section). On the other hand, as $r_r \to \infty$ ('infinite' recombination), then any coalescence with $\mathcal{G}_l$ will immediately become uncoalesced, so the haplotype tree G_m will be completely independent from $\mathcal{G}_l$ as desired.

The linked coalescent is very similar to the model of Slatkin and Pollack (2006), with a few subtle differences. In our model, we have fixed haplotype information for one locus, and simulate the second locus conditioned on that information, whereas the model of Slatkin and Pollack considers both loci simultaneously. They are also used in different contexts, as Slatkin and Pollack (2006) are concerned only with the probability that the two haplotype trees will be identical. Finally, their model is limited to only 3 species; certainly there is no theoretical barrier to extending it to more species, but it quickly becomes too complicated for theoretical analysis, rather than simulation.

Assembling the Full Gene Tree

If a locus generates no duplications or transfers, it will not produce any new loci. Thus almost surely, the process of generating new unilocus and haplotype trees/forests will

stop. When this happens, we assemble the full gene tree by concatenating each haplotype tree in a created locus to the haplotype tree in its parent locus.

Recall that linked duplications may be ancestral (the duplicating lineage survives in the parent locus) or non-ancestral (otherwise). Suppose that we wish to re-attach the haplotype tree from a locus m created by duplication, to its parent locus l. Once the haplotype tree and forest is generated in locus m, we can recognise three cases:

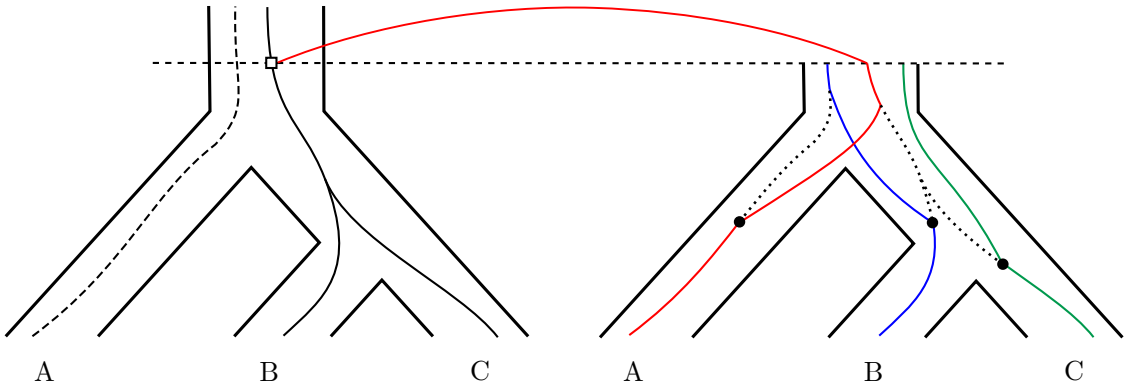
1. The lineage which is chosen to be the haplotype tree G_m in locus m is coalesced (at the time of creation of locus m) with (a copy of) the haplotype tree G_l in locus l. Then the duplication is an ancestral duplication, and we attach G_m to G_l at the lineage and time where the duplication occurred (Figure 21a).

2. The lineage which is chosen to be the haplotype tree G_m is coalesced with a copied subtree of $\mathcal{G}_l$ which is not the haplotype tree G_l. This corresponds to an ancestral duplication of a lineage in locus l which does not carry a copy of the gene. Thus locus m will not carry a copy of the gene, and the locus (along with the duplication, and all loci derived from m) is discarded (Figure 21b).

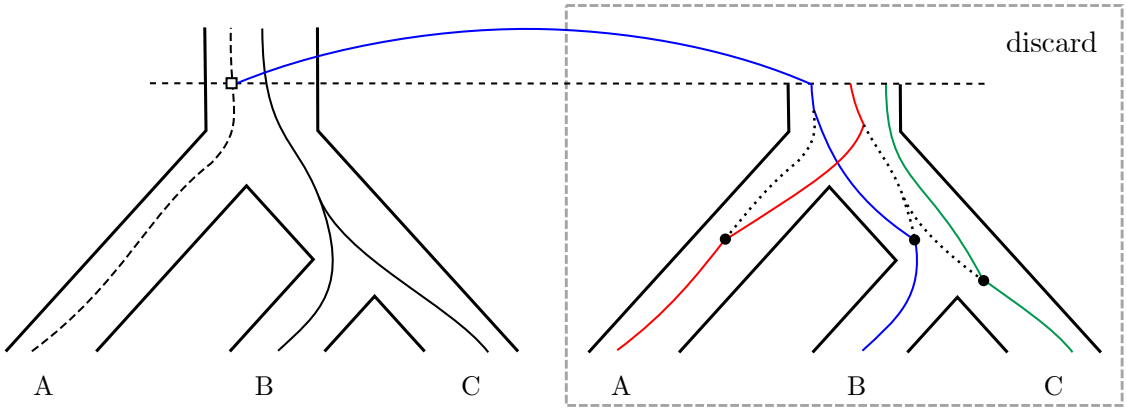
3. The lineage which is chosen to be the haplotype tree G_m is not coalesced with a copied subtree of $\mathcal{G}_l$. Then the duplication is a non-ancestral duplication, and is joined as specified below.

Events which create an unlinked locus (transfers and unlinked duplications) can always be considered to be non-ancestral; since the new locus is independent of the parent locus, the fact that the gene survives in the new locus gives no information about whether it survives in the parent locus (i.e., is ancestral). Therefore the probability of survival in the parent locus is extremely small ($O(\frac{1}{2N})$), and can be safely ignored.

Now, if locus m is created by a non-ancestral event (without loss of generality we will say it is a duplication), the duplicated gene does not survive in the parent locus l. Therefore we can treat the duplicating individual as simply another member of the population in the



(a) Here the red subtree is chosen to be the new haplotype tree. As it has coalesced with the haplotype tree copied from the parent locus, it is joined directly to that tree at the creation time of the new locus.



(b) Here the blue subtree is chosen to be the new haplotype tree. It has coalesced with a copied subtree which is not the haplotype tree in the parent locus, so it is discarded.

Fig. 21: If a duplication is ancestral, the lineage is joined directly to the subtree which it has coalesced with.

originating branch of S_l at the duplication time t_m. We follow this lineage backwards in time (in locus l) via the multispecies coalescent to where it coalesces with $\mathcal{G}_l$. Again, there are three cases:

1. The lineage coalesces with the haplotype tree G_l. We then attach G_m to G_l at the point where the coalescence occurs (Figure 22a).
2. The lineage coalesces with an element of $\mathcal{G}_l$ which is not the haplotype tree G_l. This corresponds to a lineage with no copy of the gene, so locus m is discarded (Figure 22b).
3. The lineage does not coalesce with $\mathcal{G}_l$ by the time of creation of locus l. Again, this lineage does not carry a copy of the gene, so locus m is discarded (Figure 22c).

If we do attach G_m to G_l, we must finally check that the lineage from the root of G_m to the point where it coalesces with G_l has no losses on it, which we do by running a constant-rate loss process (with rate r_l) on that branch. If a loss exists on that branch, the new locus m (and all derived loci) must be discarded.

This process is done backwards in time order by first attaching (or discarding) each haplotype tree with no child loci to the haplotype trees of their parent loci, and then those to the haplotype trees of their parent loci, and so on until the full gene tree is assembled.

The full pseudocode for this algorithm is given in Algorithms 1 and 2. For clarity, we have chosen to be brief in the specification of the multispecies coalescent processes in lines 14, 18, and 25; for full details, refer to the above discussion. The process can also generate a reconciliation that maps nodes of the gene tree G to branches in the species tree S.

Algorithm 1 CONSTRUCT_GENE_TREE

Input: Species tree S with branch lengths in coalescent units, event parameters $\mathbf{r} = (r_d, r_t, r_l, r_r, p_u)$

Output: Gene tree G

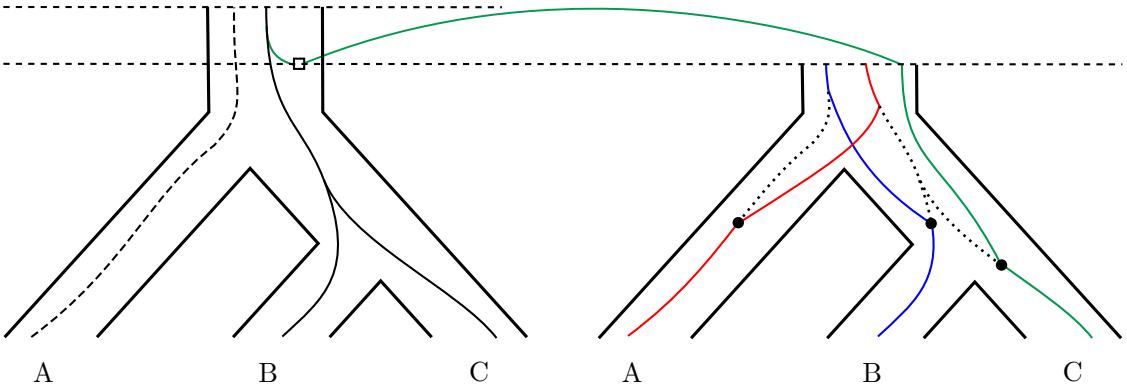
1: $S_0 \leftarrow S$
2: $G_0 \leftarrow$ multispecies coalescent within S_0
3: $\mathcal{G}_0 \leftarrow \{G_0\}$
4: $G_0^l \leftarrow$ ADD_NEW_LOCI$(S_0, G_0, \mathcal{G}_0, G_0, \mathbf{r})$
5: **return** G_0^l

Algorithm 2 ADD_NEW_LOCI

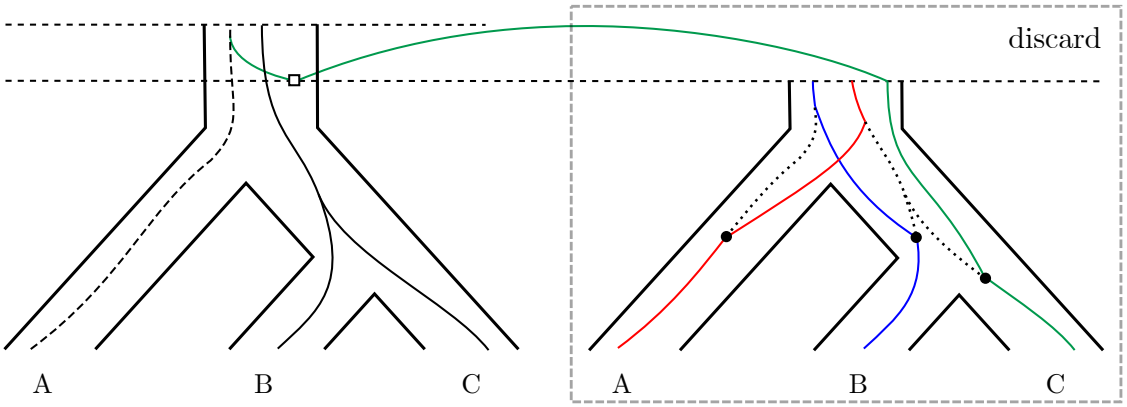
Input: Unilocus tree S_i, haplotype tree G_i, haplotype forest $\mathcal{G}_i$, partial gene tree G_i^l, event parameters $\mathbf{r}$

Output: Gene tree G_i^l

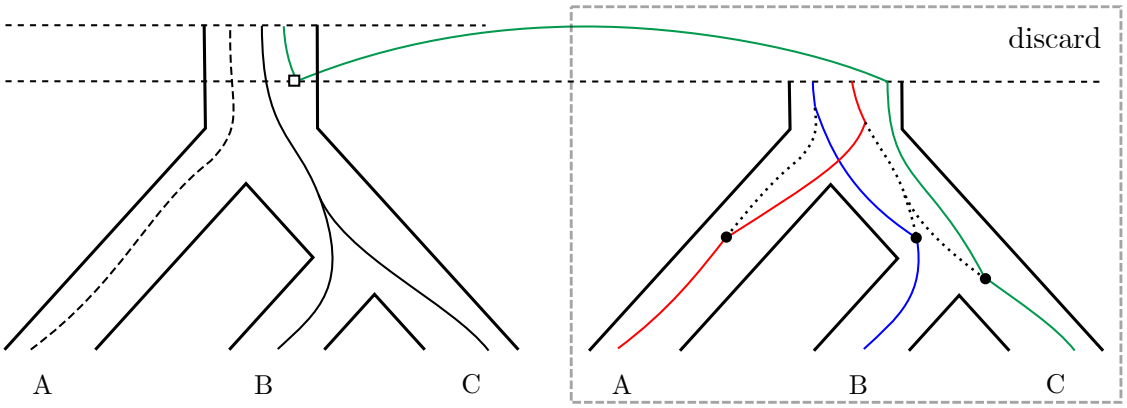
$\boldsymbol{e^{ud}} \leftarrow$ D coalescent-rate process on S_i with rate parameter $p_u r_d$
$\boldsymbol{e^{t}} \leftarrow$ T coalescent-rate process on S with rate parameter r_t, scaled by the proportion of origin species in S_i
$\boldsymbol{e^{ld}} \leftarrow$ D coalescent-rate process in the presence of $\mathcal{G}_i$ on S_i with rate parameter $(1-p_u)r_d$ and recombination rate r_r
$\boldsymbol{e^{l}} \leftarrow$ L constant-rate process on G_i with rate parameter r_l
for $(b,t) \in \boldsymbol{e^{ud}} \cup \boldsymbol{e^{ld}} \cup \boldsymbol{e^{t}} \cup \boldsymbol{e^{l}}$ in time order **do**
  if $(b,t) \in \boldsymbol{e^{l}}$ **then** ▷ here $b \in E(G_i)$
    $g \leftarrow$ a unary node on G_i^l at time t along the branch b
    $G_i^l \leftarrow$ cut G_i^l from g
    remove all events from $\boldsymbol{e^{l}}$ below b which occur at a later time
  else if $(b,t) \in \boldsymbol{e^{ud}} \cup \boldsymbol{e^{ld}} \cup \boldsymbol{e^{t}}$ **then** ▷ here $b \in E(S)$
    create a new locus j
    $S_j \leftarrow$ subtree of S_i rooted at b at time t
    if $(b,t) \in \boldsymbol{e^{ud}} \cup \boldsymbol{e^{t}}$ **then**
      $\mathcal{G}_j \leftarrow$ incomplete coalescent within S_j to time t
      $G_j \leftarrow$ an element of $\mathcal{G}_j$
    else if $(b,t) \in \boldsymbol{e^{ld}}$ **then**
      $\mathcal{G}'_j \leftarrow$ all subtrees of $\mathcal{G}_i$ within S_j
      $\mathcal{G}_j \leftarrow$ linked coalescent in the presence of $\mathcal{G}'_j$ within S_j to time t, with recombination rate r_r
      $G_j \leftarrow$ an element of $\mathcal{G}_j$
    if $(b,t) \in \boldsymbol{e^{ld}}$ and G_j has coalesced with a subtree of G_i at time t **then**
      attach G_j to G_i^l at time t
    else if $(b,t) \in \boldsymbol{e^{ld}}$ and G_j has coalesced with a subtree of $\mathcal{G}_i$ at time t **then**
      discard locus j
    else
      $G_i^l \leftarrow$ incomplete coalescent "joining" of G_j to G_i^l in the presence of $\mathcal{G}_i$ within S_i (keep G_i^l)
      $\boldsymbol{e^{l}} \leftarrow$ L constant-rate process from root of G_j to G_i^l with rate parameter r_l
      if $|\boldsymbol{e^{l}}| > 0$ **then**
        discard locus j
    $G_i^l \leftarrow$ ADD_NEW_LOCI$(S_j, G_j, \mathcal{G}_j, G_i^l, \mathbf{r})$
return G_i^l



(a) Here the green subtree is chosen to be the new haplotype tree. It coalesces with the haplotype tree in the parent locus, and is joined at the coalescent point.



(b) Here the green subtree is chosen to be the new haplotype tree. It coalesces with a lineage which is not in the haplotype tree in the parent locus, so it is discarded.



(c) Here the green subtree is chosen to be the new haplotype tree. It does not coalesce by the time of creation of the parent locus, so it is discarded.

Fig. 22: If a duplication is non-ancestral, we follow the lineage backwards in time until it coalesces with the haplotype forest in the parent locus.

Acknowledgements

YBC and CS acknowledge the Partenariat Hubert Curien (Hubert Curien Partnership) for providing grants for collaborative travel. QL would like to acknowledge

Geoffrey Law for providing assistance with programming.

References

Arvestad, L., Berglund, A.-C., Lagergren, J., and Sennblad, B. (2004). Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. In *Proceedings of the eighth annual international conference on Research in computational molecular biology*, pages 326–335. ACM.

Arvestad, L., Lagergren, J., and Sennblad, B. (2009). The gene evolution model and computing its associated probabilities. *J. ACM*, 56(2):1–44.

Avise, J. C. and Robinson, T. J. (2008). Hemiplasy: a new term in the lexicon of phylogenetics. *Syst. Biol*, 57(3):503–507.

Chan, Y., Ranwez, V., and Scornavacca, C. (2017). Inferring incomplete lineage sorting, duplications, transfers and losses with reconciliations. *J. Theor. Biol*, 432:1–13.

David, L. A. and Alm, E. J. (2011). Rapid evolutionary innovation during an archaean genetic expansion. *Nature*, 469(7328):93.

Degnan, J. H. and Rosenberg, N. A. (2006). Discordance of species trees with their most likely gene trees. *PLoS Genet.*, 2(5):e68.

Doyon, J.-P., Ranwez, V., Daubin, V., and Berry, V. (2011). Models, algorithms and programs for phylogeny reconciliation. *Brief. Bioinform*, 12(5):392–400.

Doyon, J.-P., Scornavacca, C., Gorbunov, K. Y., Szöllősi, G. J., Ranwez, V., and Berry, V. (2010). An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. In *RECOMB International Workshop on Comparative Genomics*, pages 93–108. Springer.

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. *IEEE/ACM Trans. Comput. Biol. Bioinform.*

Durand, D., Halldórsson, B. V., and Vernot, B. (2006). A hybrid micro–macroevolutionary approach to gene tree reconstruction. *J. Comput. Biol*, 13(2):320–335.

Goodman, M., Czelusniak, J., Moore, G. W., Romero-Herrera, A. E., and Matsuda, G. (1979). Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst. Biol*, 28(2):132–163.

Guerrero, R. F. and Hahn, M. W. (2018). Quantifying the risk of hemiplasy in phylogenetic inference. *Proc. Natl. Acad. Sci. U.S.A*, 115(50):12787–12792.

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol*, 23(2):183–201.

Kingman, J. F. (1982). On the genealogy of large populations. *J. Appl. Probab.*, 19(A):27–43.

Maddison, W. P. (1997). Gene trees in species trees. *Syst. Biol*, 46(3):523–536.

Mallo, D., de Oliveira Martins, L., and Posada, D. (2015). Simphy: phylogenomic simulation of gene, locus, and species trees. *Syst. Biol*, 65(2):334–344.

Mawhorter, R., Liu, N., Libeskind-Hadas, R., and Wu, Y.-C. (2019). Inferring pareto-optimal reconciliations across multiple event costs under the duplication-loss-coalescence model. *BMC Bioinform.*, 20(20):639.

Pamilo, P. and Nei, M. (1988). Relationships between gene trees and species trees. *Mol. Biol. Evol*, 5(5):568–583.

Rannala, B. and Yang, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, 164(4):1645–1656.

Rasmussen, M. D. and Kellis, M. (2010). A bayesian approach for fast and accurate gene tree reconstruction. *Mol. Biol. Evol*, 28(1):273–290.

Rasmussen, M. D. and Kellis, M. (2012). Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Res.*, 22(4):755–765.

Schrempf, D. and Szöllősi, G. J. (2018). Phylogenetic incongruences - opportunities to improve the reconstruction of a dated tree of life.

Sjöstrand, J., Arvestad, L., Lagergren, J., and Sennblad, B. (2013). Genphylodata: realistic simulation of gene family evolution. *BMC Bioinform.*, 14(1):209.

Slatkin, M. and Pollack, J. L. (2006). The concordance of gene trees and species trees at two linked loci. *Genetics*, 172(3):1979–1984.

Stolzer, M., Lai, H., Xu, M., Sathaye, D., Vernot, B., and Durand, D. (2012). Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics*, 28(18):i409–i415.

Tofigh, A., Hallett, M., and Lagergren, J. (2011). Simultaneous identification of duplications and lateral gene transfers. *IEEE/ACM Trans. Comput. Biol. Bioinform. (TCBB)*, 8(2):517–535.

Vernot, B., Stolzer, M., Goldman, A., and Durand, D. (2008). Reconciliation with non-binary species trees. *J. Comput. Biol*, 15(8):981–1006.

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 Res.*, 24(3):475–486.

Zhang, L. (1997). On a Mirkin-Muchnik-Smith conjecture for comparing molecular phylogenies. *J. Comput. Biol*, 4(2):177–187.