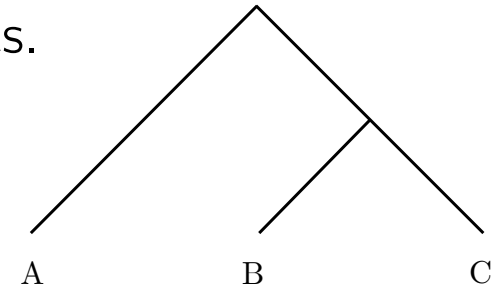
MLMSC
A flexible new model of gene family evolution

Celine Scornavacca
joint work with Nicolas Galtier, Yao-ban Chan and Qiuyi Li

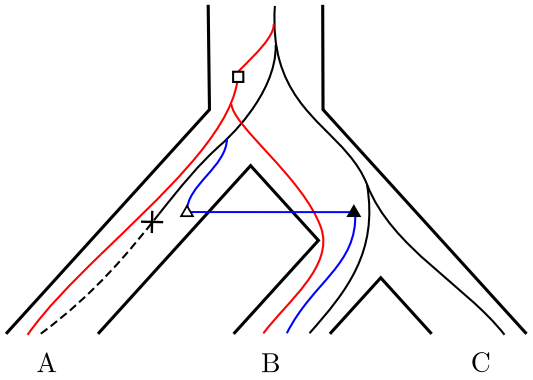
05/12/2023

Species/gene trees

Species **trees** : depict the evolutionary history of a set of organisms.
Internal nodes represent speciation events.
Branch lengths represent divergence times/amounts.

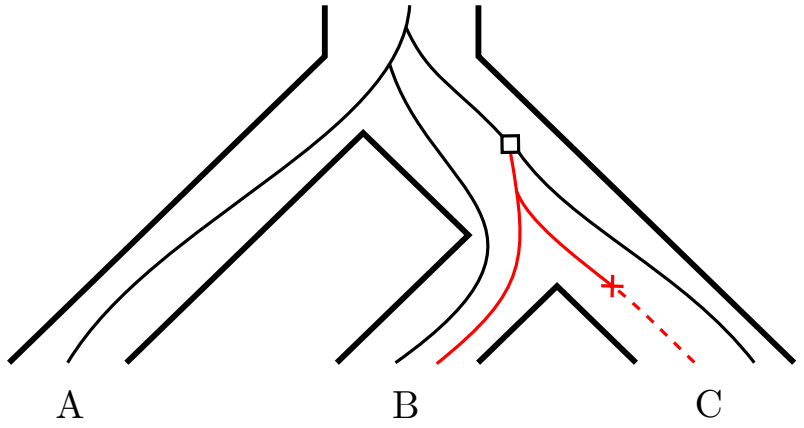


Gene **trees** : depict the evolutionary history of a gene family, i.e., homologous molecular sequences appearing in the genome of different organisms.
They reflect a complex evolution potentially involving many diverse processes, in addition to speciations.

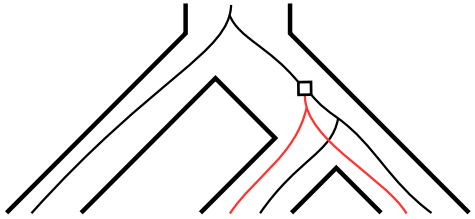


Species/gene trees can significantly differ

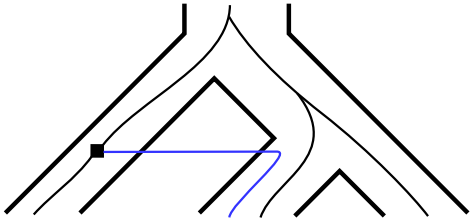
- methodological causes
  - sequencing errors
  - contamination
  - inexact species delimitation
  - inaccurate clustering into homologous groups
  - misalignement
  - model inadequacy
  - ...
- biological processes
  - D+L
  - HGT
  - ILS
  - ancient population structure
  - low/high mutation rate
  - ...



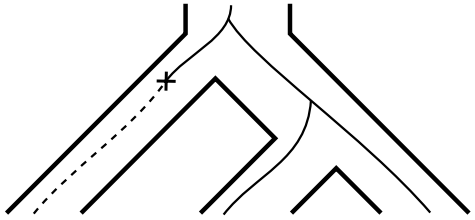
The DTL processes



Gene Duplication: a single gene copy gives rise to two copies at two distinct loci

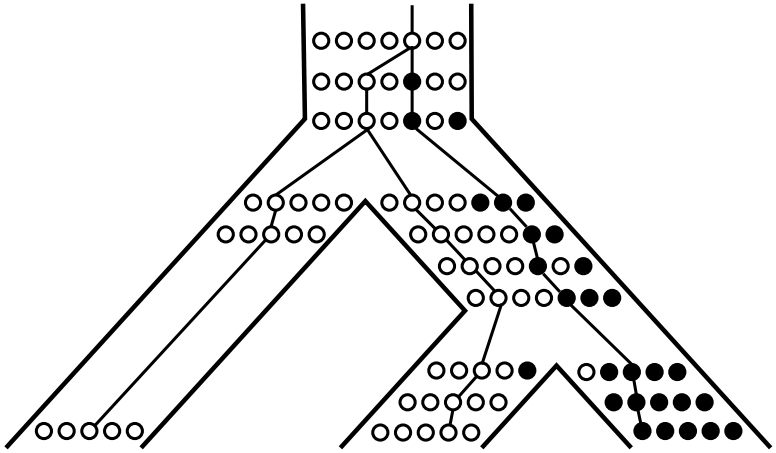


Gene Transfer: a gene from one species enters the genome of another contemporary species



Gene Loss: a gene is removed from the genome

Incomplete Lineage Sorting (ILS)

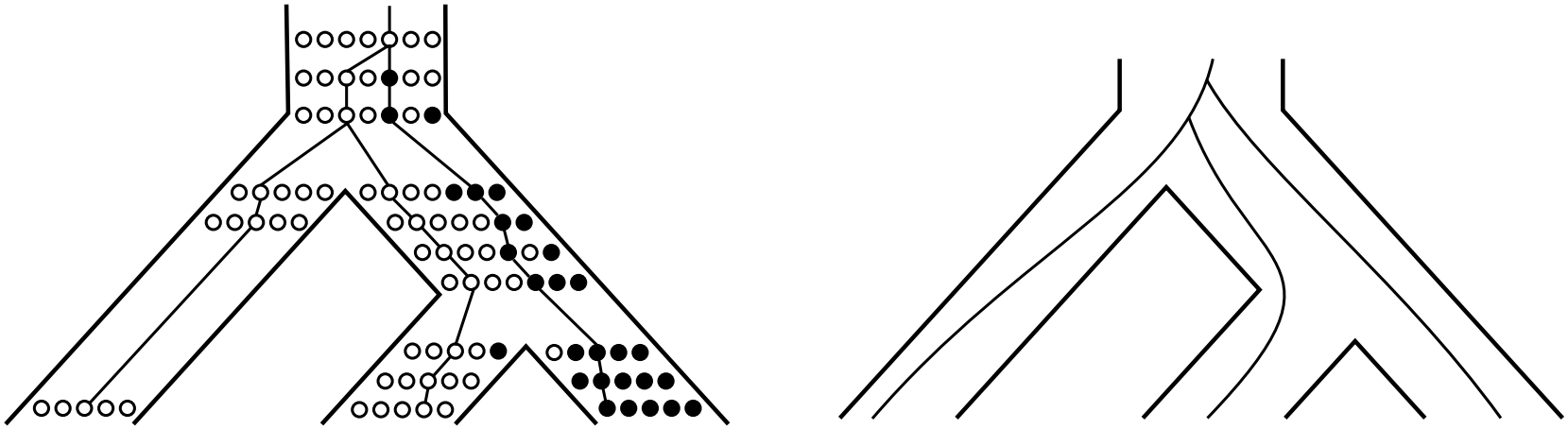


The originating population contains a single white allele.

1. a mutation leads to a new black allele at the locus,
2. the first speciation takes place,
3. rapidly followed by a second one.

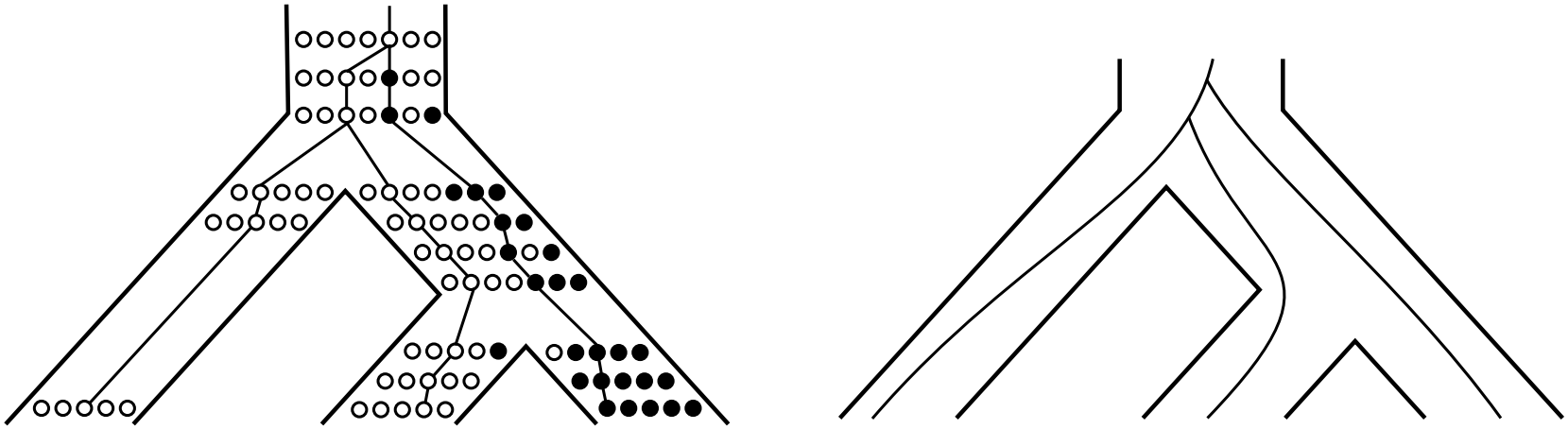
As the white and black alleles still coexist when the second speciation takes place, both alleles may be fixed in separate descendant species.

Incomplete Lineage Sorting (ILS)



This results in a gene history which **differs** from the species history.

Incomplete Lineage Sorting (ILS)



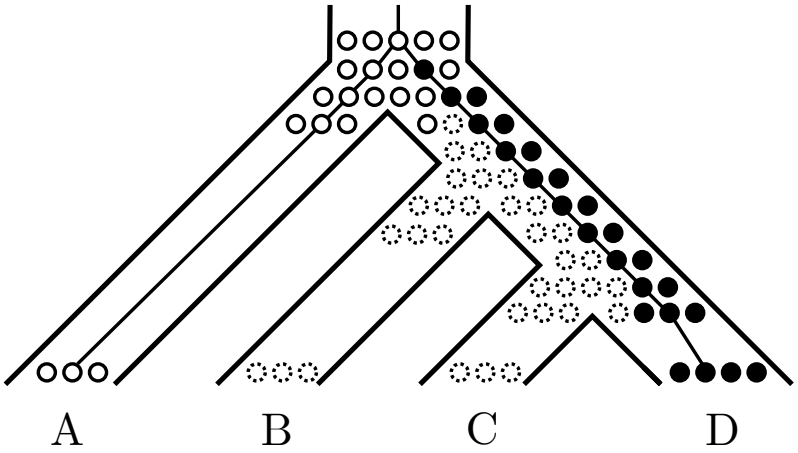
This results in a gene history which **differs** from the species history.

The multispecies coalescent (**MSC**) model predicts the effect of ILS on gene tree branch lengths and topology as a function of the effective population size and timing of speciations.
More likely for

→ fast successive speciations

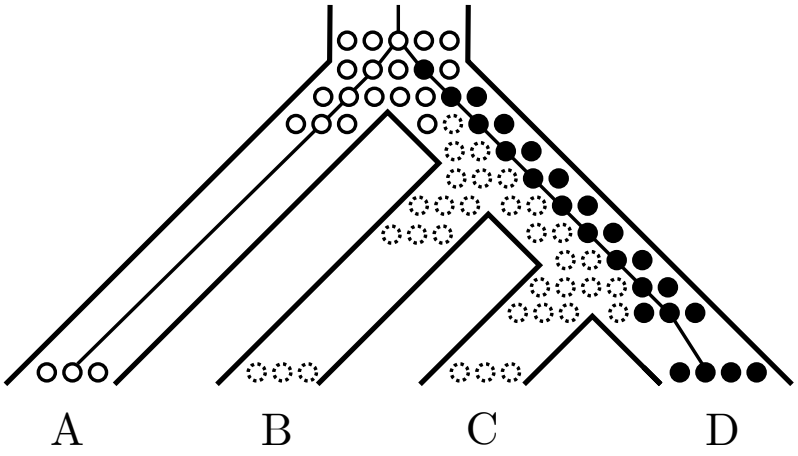
→ large effective sizes

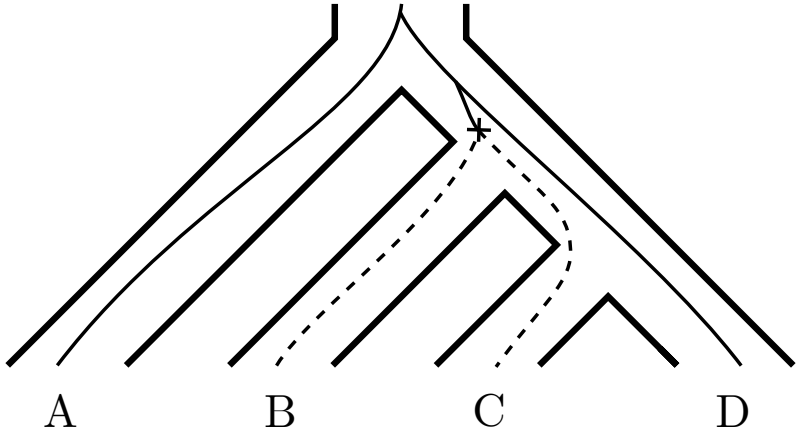
ILS and DTL interaction, examples



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

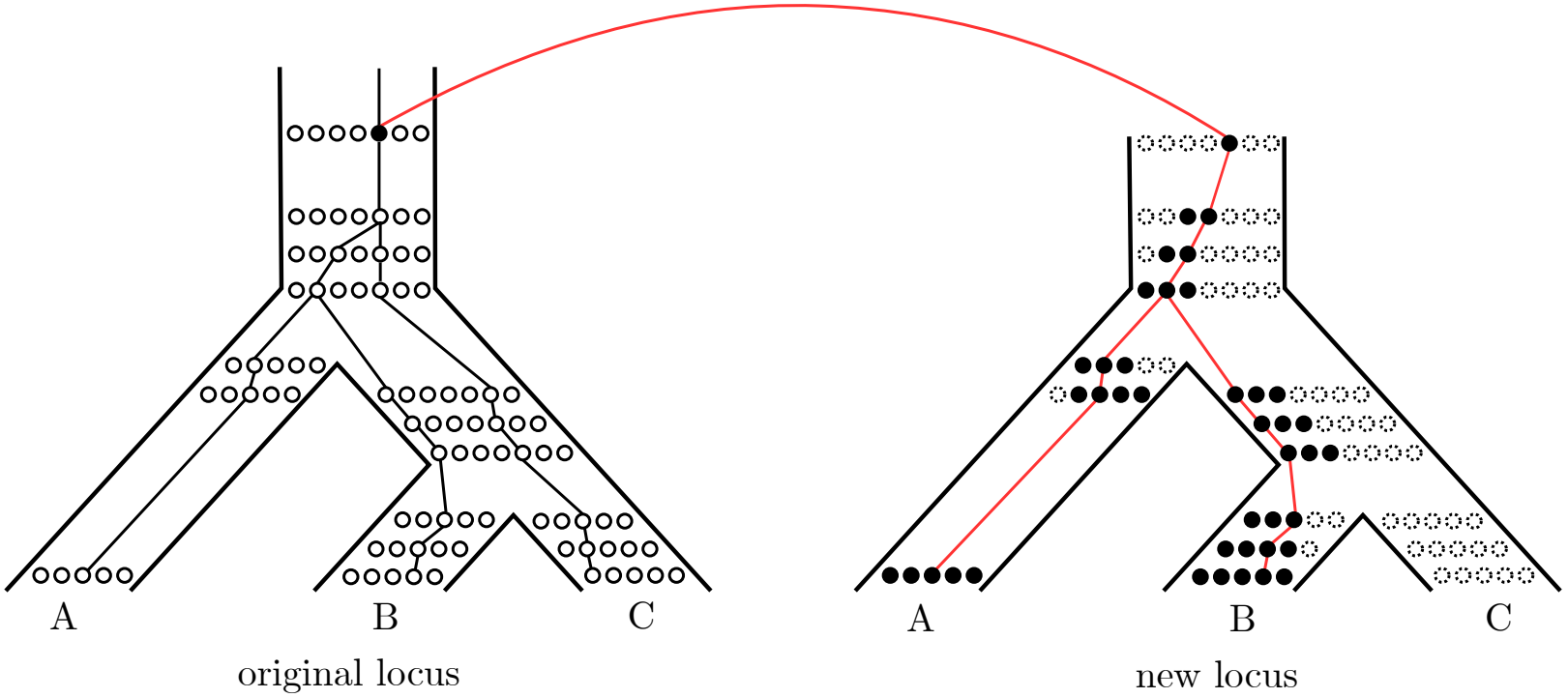
ILS and DTL interaction, examples





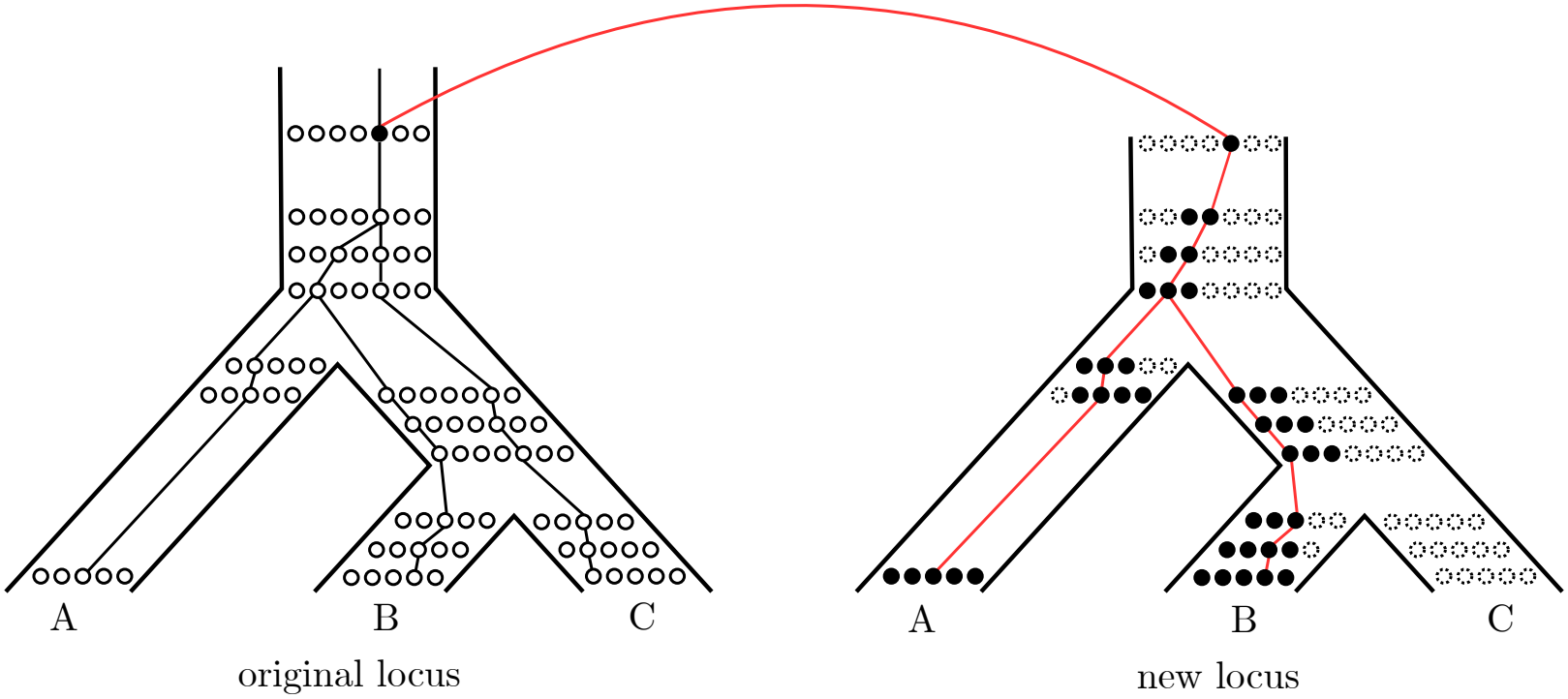
ILS interacting with loss:
with only one loss event, two descendant species may end up with an empty locus due to the presence of ILS

ILS and DTL interaction, examples



ILS interacting with duplication:
the duplicated gene only fixes in species A and B.
Species C has only one copy, without a gene loss event

ILS and DTL interaction, examples

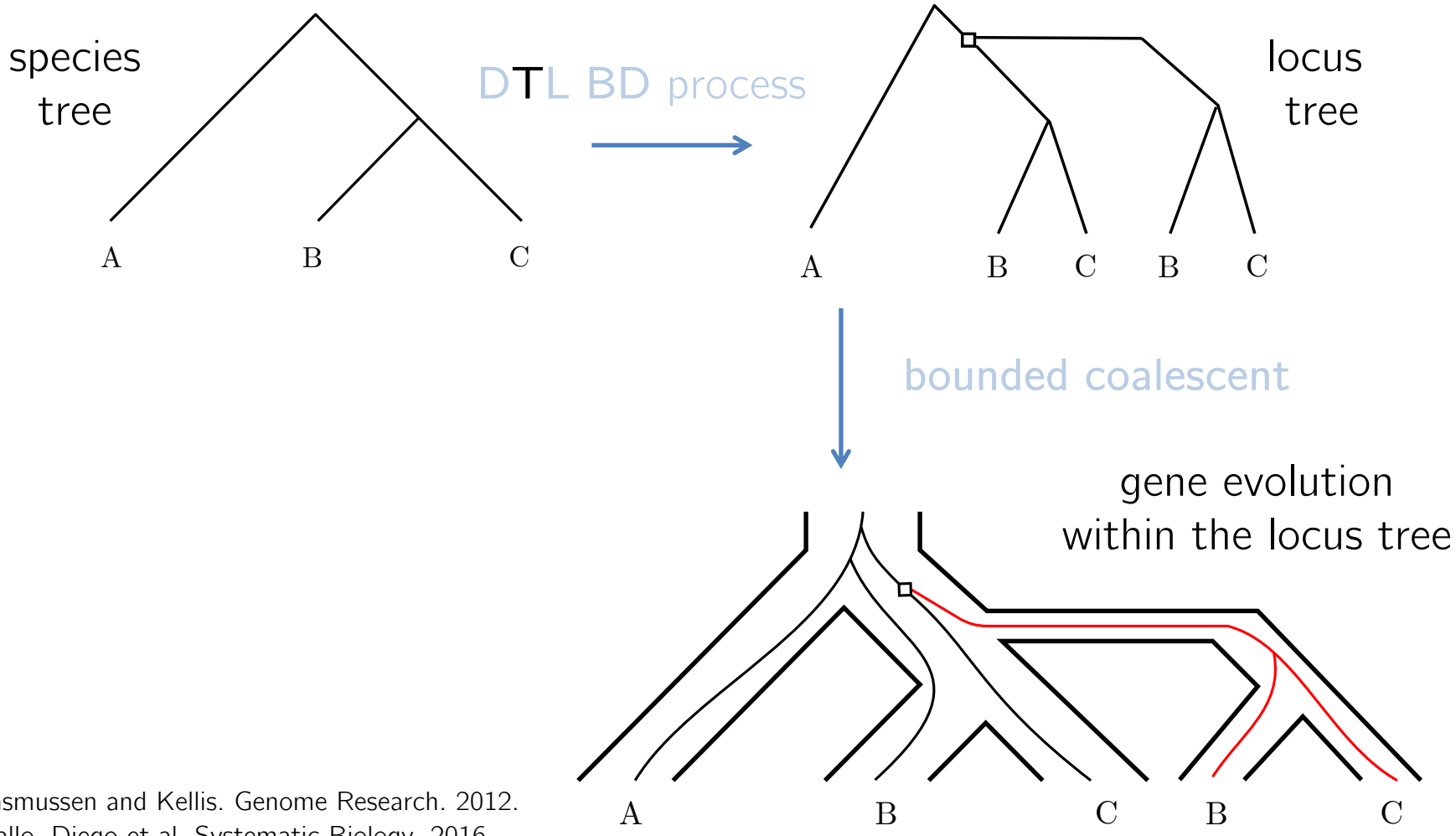


ILS interacting with duplication:
the duplicated gene only fixes in species A and B.
Species C has only one copy, without a gene loss event

Copy Number Hemiplasy (CNH)

Existing models accounting for ILS and DTL

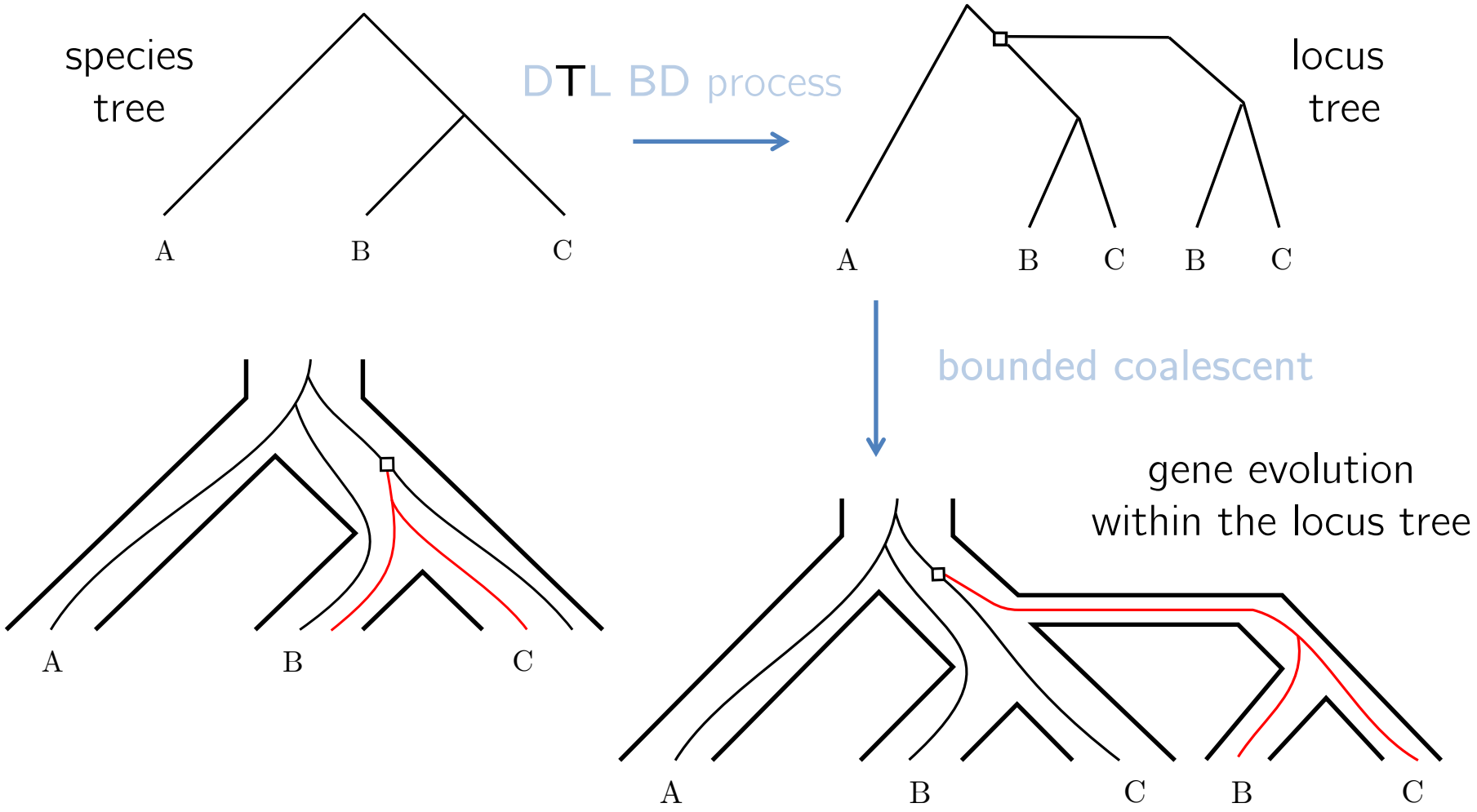
Locus tree model (DLCoal, DLCpar and SimPhy)



Rasmussen and Kellis. Genome Research. 2012.
Mallo, Diego et al. Systematic Biology. 2016

Existing models accounting for ILS and DTL

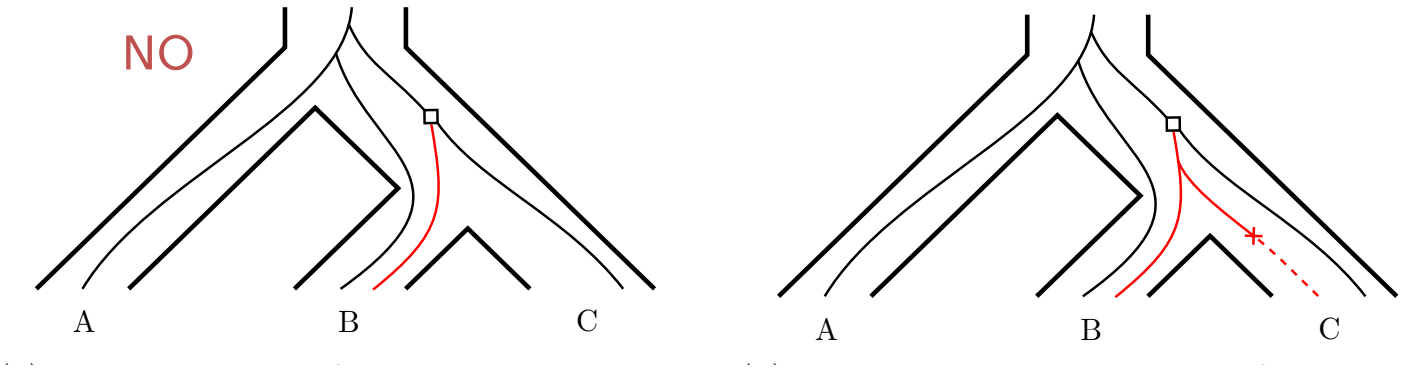
Locus tree model (DLCoal, DLCpar and SimPhy)



Existing models accounting for ILS and DTL

Locus tree model (DLCoal, DLCpar and SimPhy)

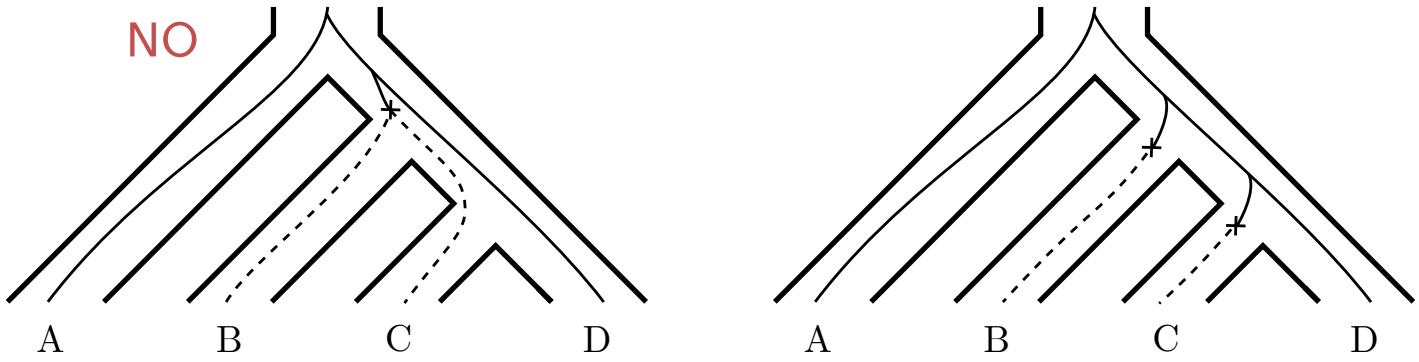
ILS does not interfere with gene **duplication** and loss in generating variation in gene copy number among species.



Existing models accounting for ILS and DTL

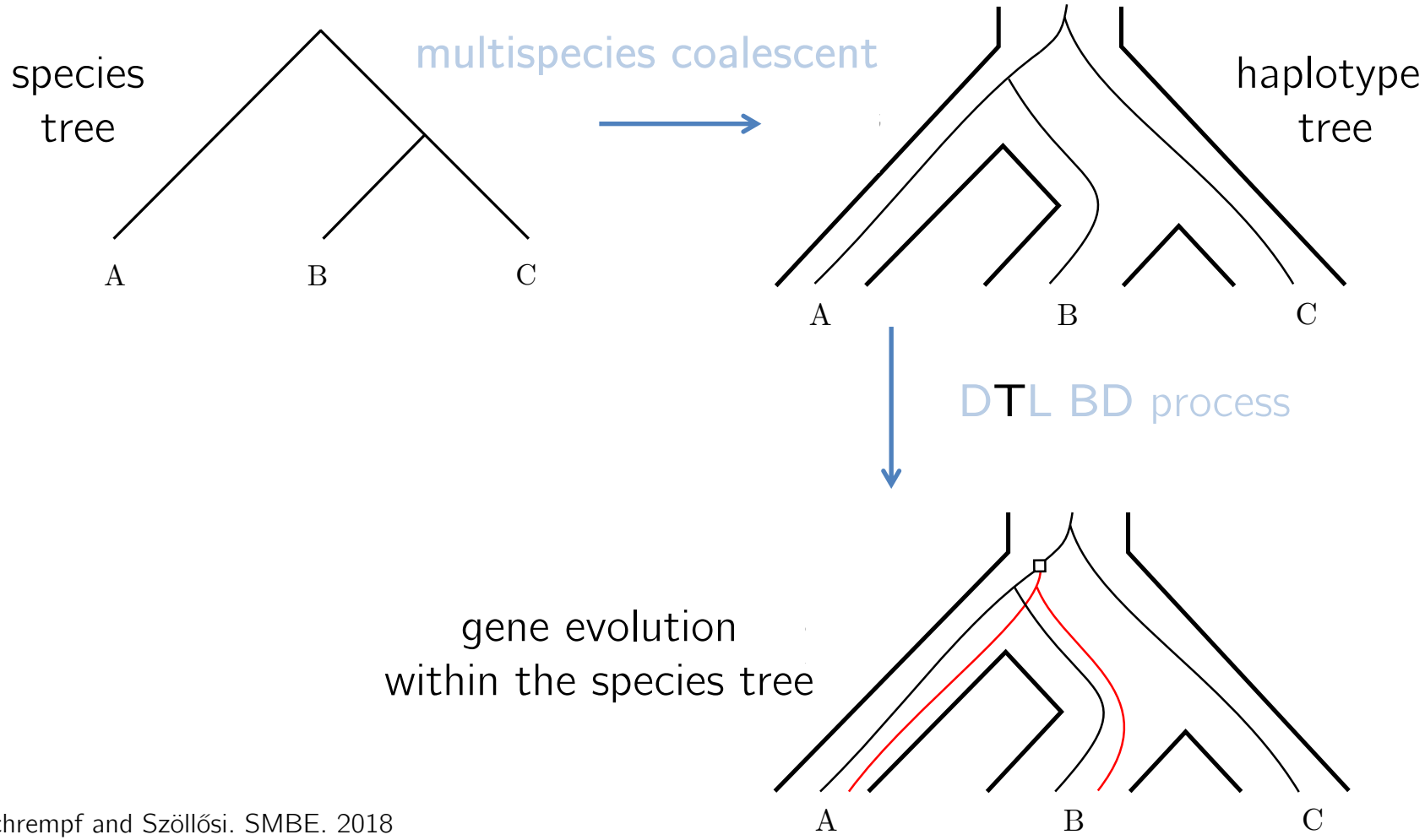
Locus tree model (DLCoal, DLCpar and SimPhy)

ILS does not interfere with gene **duplication** and loss in generating variation in gene copy number among species.



Existing models accounting for ILS and DTL

Haplotype tree model

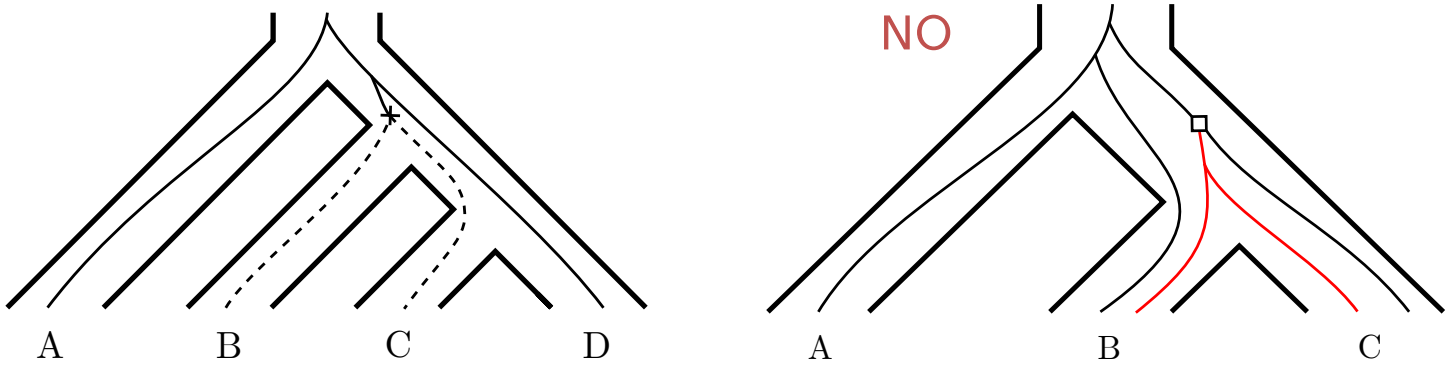


Schrempf and Szöllősi. SMBE. 2018

Existing models accounting for ILS and DTL

Haplotype tree model

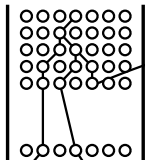
Does not fully allow for copy number hemiplasy since duplicated/transferred gene must be sorted into the same species than the original copy.



The Wright-Fisher Process with Dup and Loss

We start by focusing on a single randomly mating population of 2N haploid genomes with discrete, non-overlapping generations in the *original* locus.

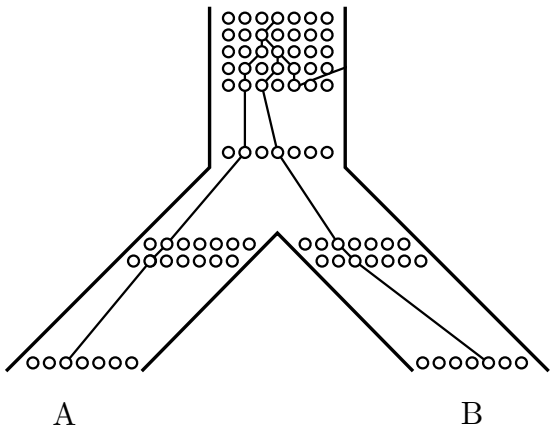
In this locus, the genealogical process is identical to the Wright-Fisher process: at every generation each individual descends from a random member of the previous generation.



The Wright-Fisher Process with Dup and Loss

In addition to the genealogical process, three events may occur:

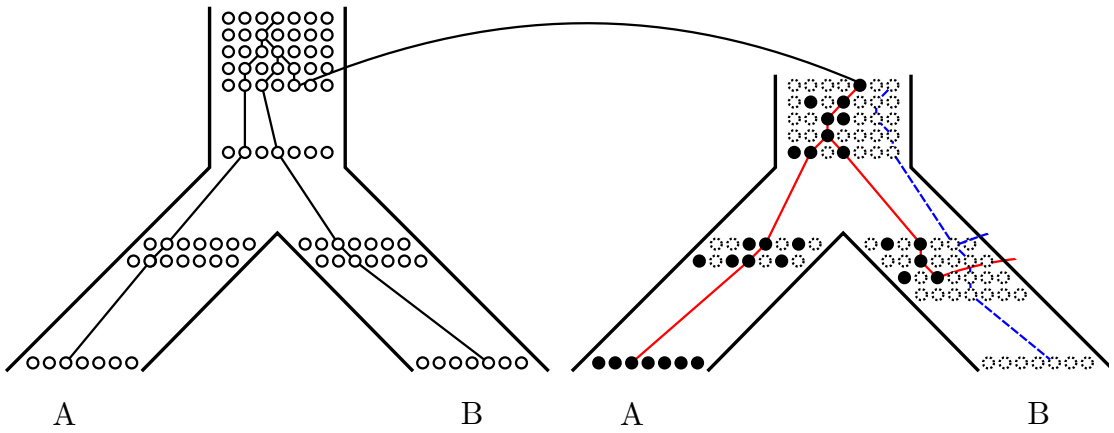
1. **Speciations** Each population in each locus splits into two separate populations. For each child population, each individual in the first generation descends from a random individual of the last generation of the parent population, as in the standard WF model. The WFDL process is then run recursively and independently in both child populations.



The Wright-Fisher Process with Dup and Loss

In addition to the genealogical process, three events may occur:

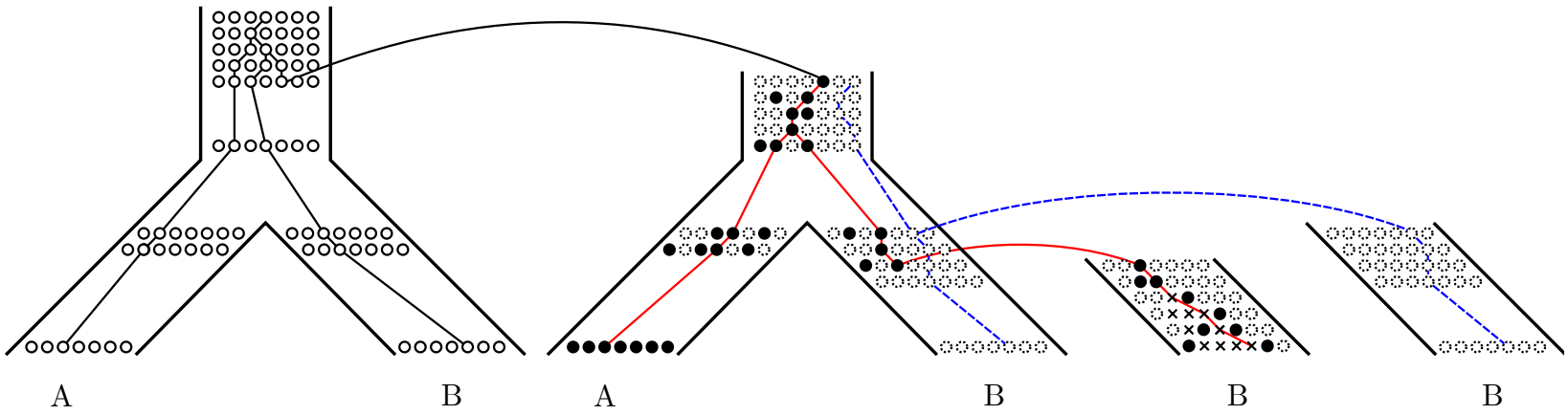
2. **Duplication** at rate r_d per individual per generation in each locus. An individual from the population is selected (uniformly at random) to duplicate. A new child locus is created with a population of size 2N. In the child population, a single individual descends from the duplicating individual in the parent locus. The WFDL process is then run recursively and independently in the new population.



The Wright-Fisher Process with Dup and Loss

In addition to the genealogical process, three events may occur:

3. **Loss** at rate r_l per individual per generation in each locus. An individual from the population is selected (uniformly at random) to lose the gene copy (if the locus for that individual is not empty) at that locus.

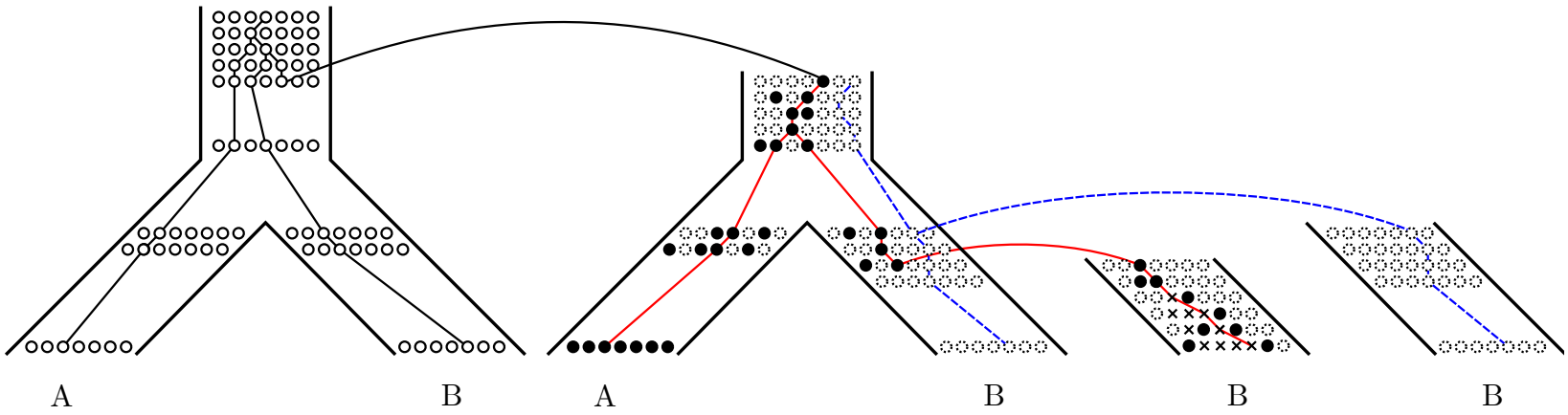


The Wright-Fisher Process with Dup and Loss

In addition to the genealogical process, three events may occur:

3. **Loss** at rate r_l per individual per generation in each locus. An individual from the population is selected (uniformly at random) to lose the gene copy (if the locus for that individual is not empty) at that locus.

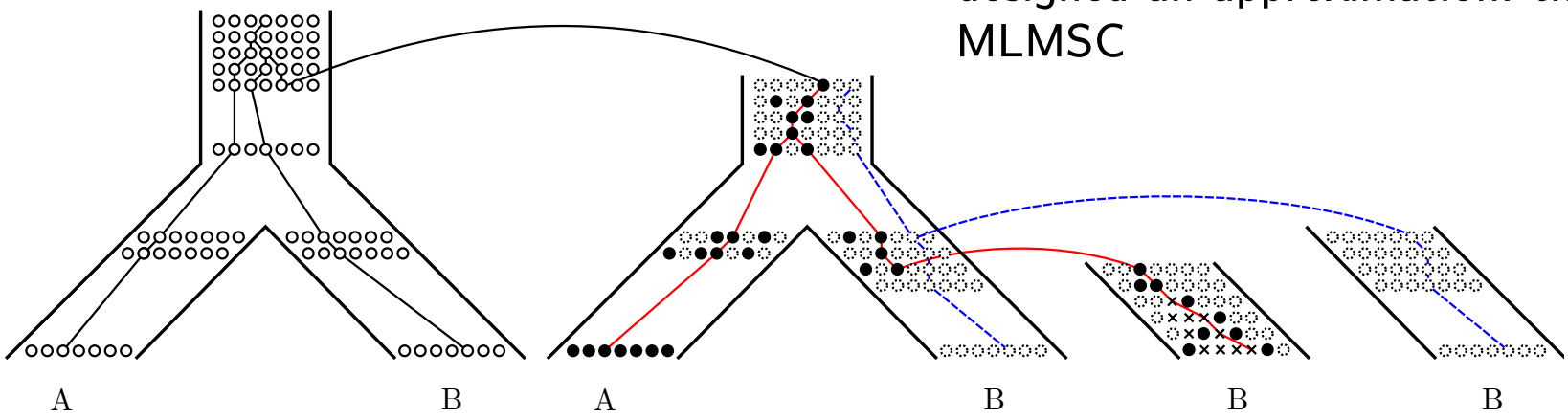
We define the rates of duplication and loss as $r'_d = 2N\, r_d$ and $r'_l = 2N\, r_l$



The Wright-Fisher Process with Dup and Loss

When the WFDL process reaches the present time, we sample one genome in each species → one individual per population in each locus.
We trace the lineages back to their most recent common ancestor.
If a sampled individual did not descend from the original locus, it is discarded to obtain the *joined haplotype tree*.
We truncate the tree at each loss to produce the final *gene tree.*

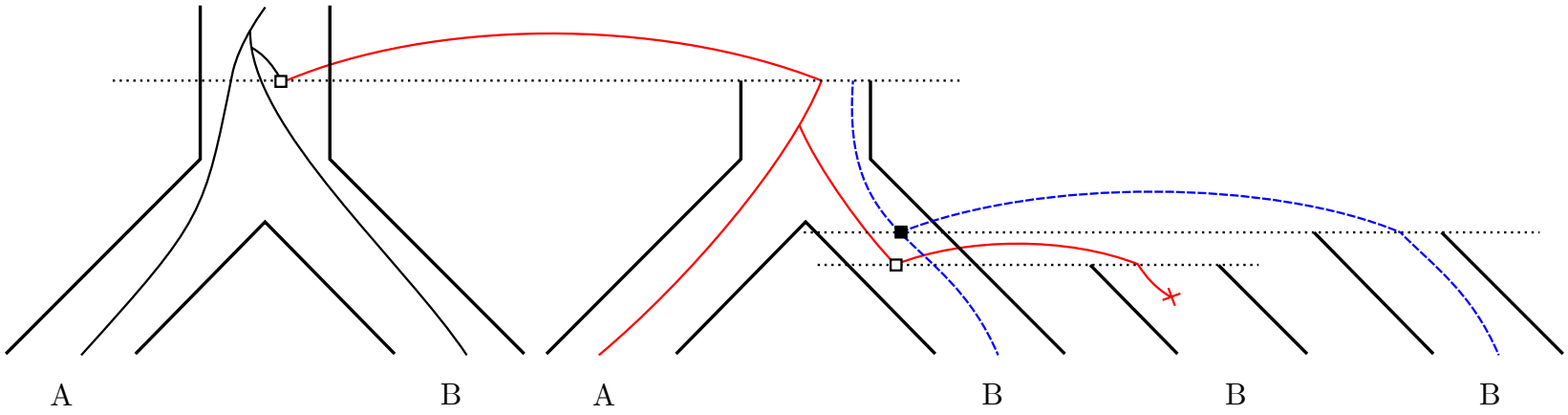
Intractable and pointless, we designed an approximation: the MLMSC



The MLMSC, a new gene family evolution model

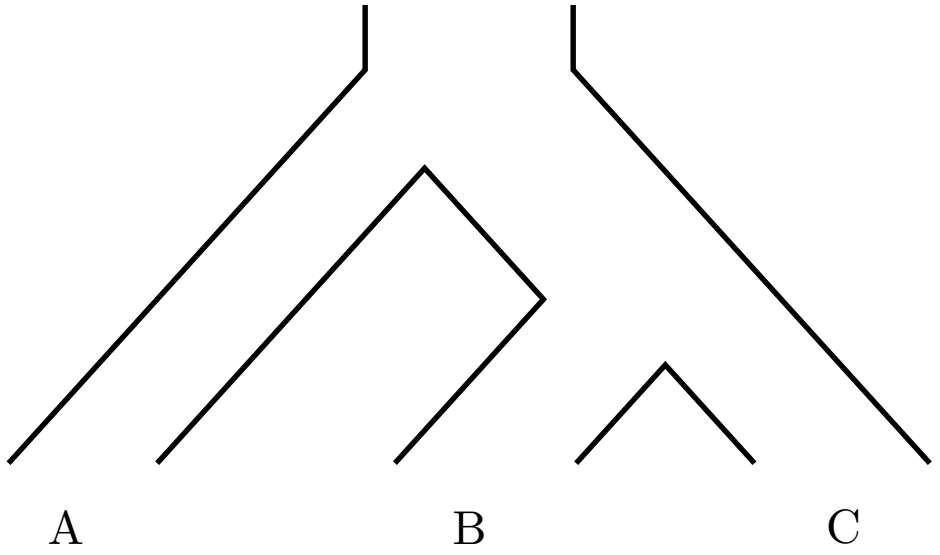
Given a species tree, we generate a gene tree in a recursive manner by defining a succession of *unilocus trees*, each representing the evolution of one locus.
Within each unilocus tree, we generate a *haplotype tree*, which represents the genealogy of the individuals in that locus.
We sample duplication events within the locus to create new unilocus and then haplotype trees. The haplotype trees are then joined, and losses simulated, to produce the full gene tree.

(Actually the MLMSC is more complex, it also models transfers and `linked' duplications)



The original locus

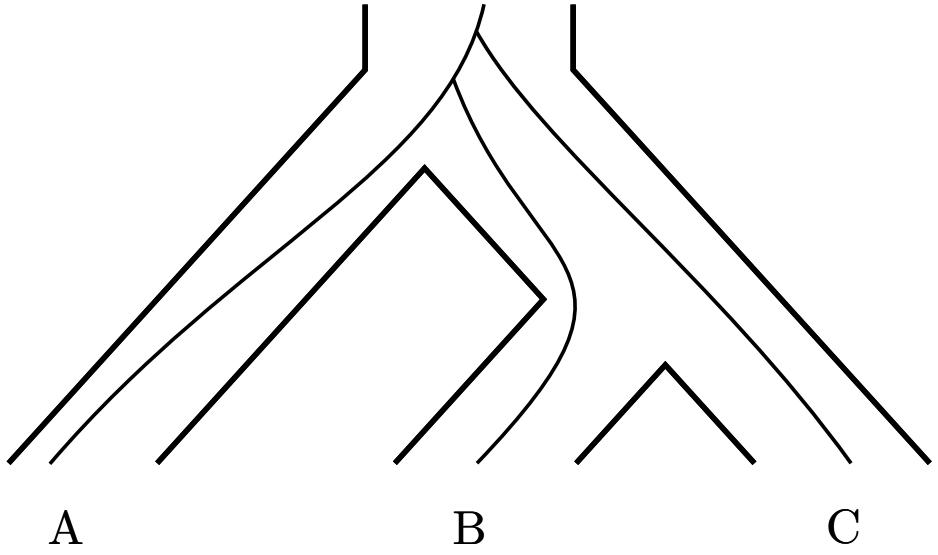
The unilocus tree for the original locus is the same as the species tree.



The original locus

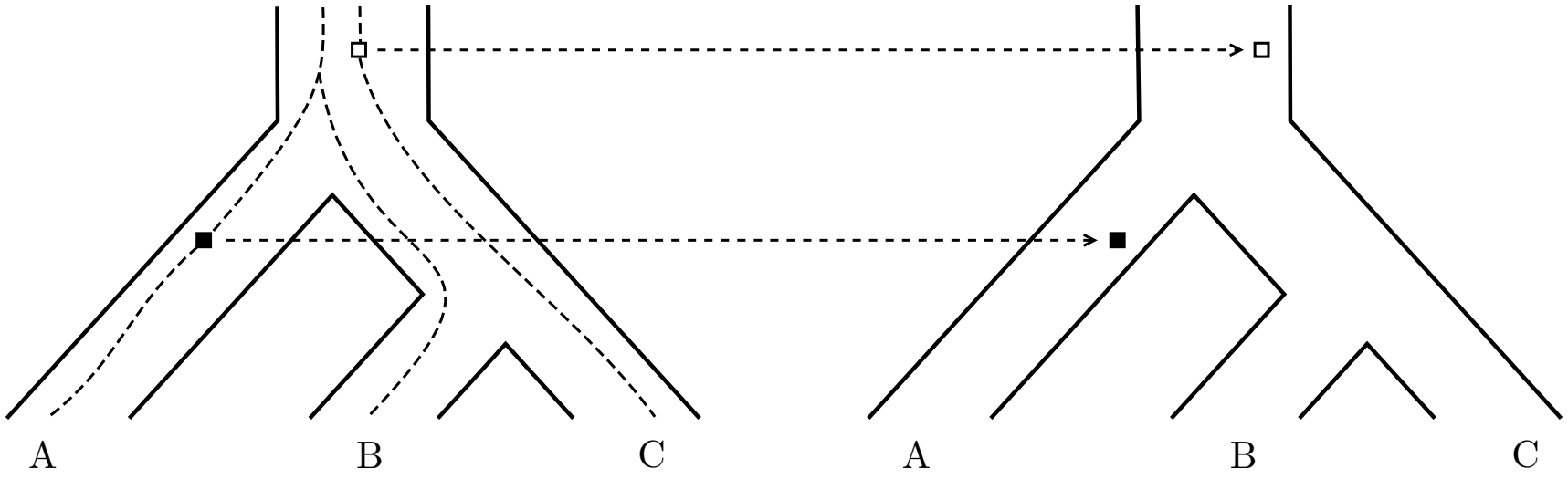
The unilocus tree for the original locus is the same as the species tree.

The haplotype tree for the original locus is produced via a standard multispecies coalescent process.



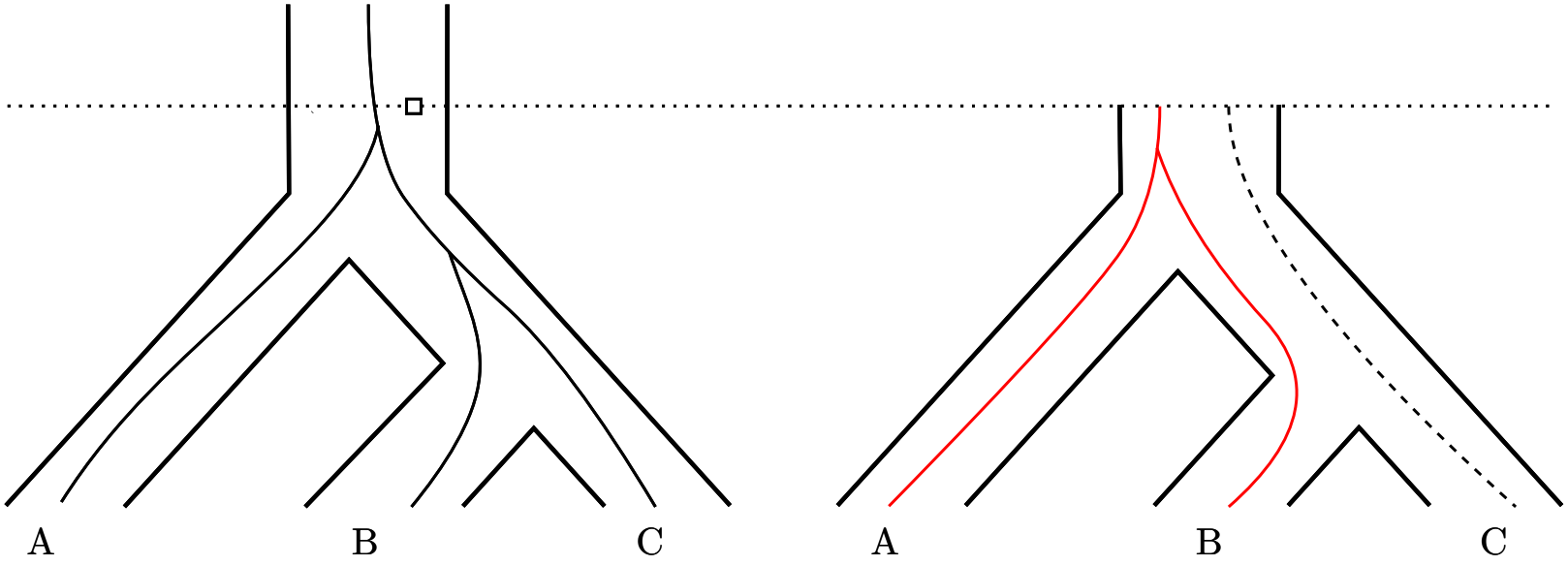
The duplication process (unilocus tree)

Duplications are simulated within the branches of the unilocus tree using a *coalescent-rate process*: a temporary multispecies coalescent is simulated within the unilocus tree, duplications are simulated at constant rate r'_d on the branches of the temporary tree, and the temporary tree is then discarded.



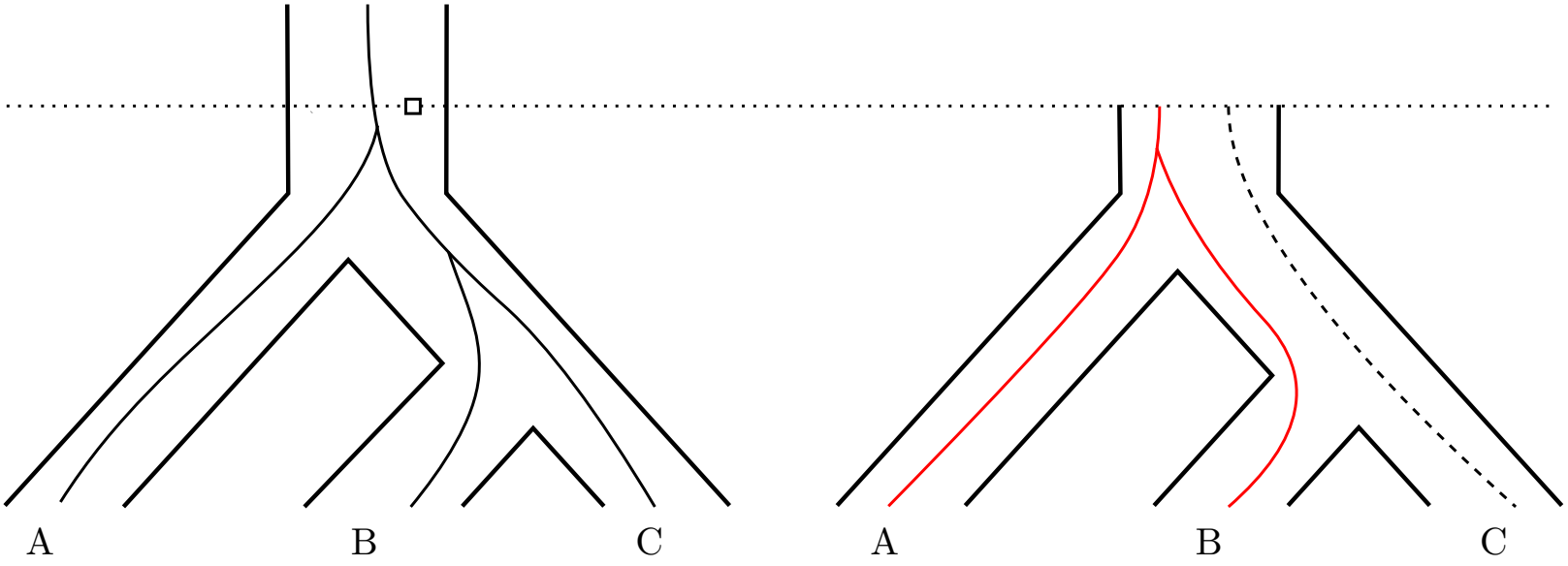
The duplication process (unilocus tree)

A new unilocus tree is created at each duplication point. This tree is the subtree of the species tree starting from the time and branch when the duplication happens.



The duplication process (haplotype tree)

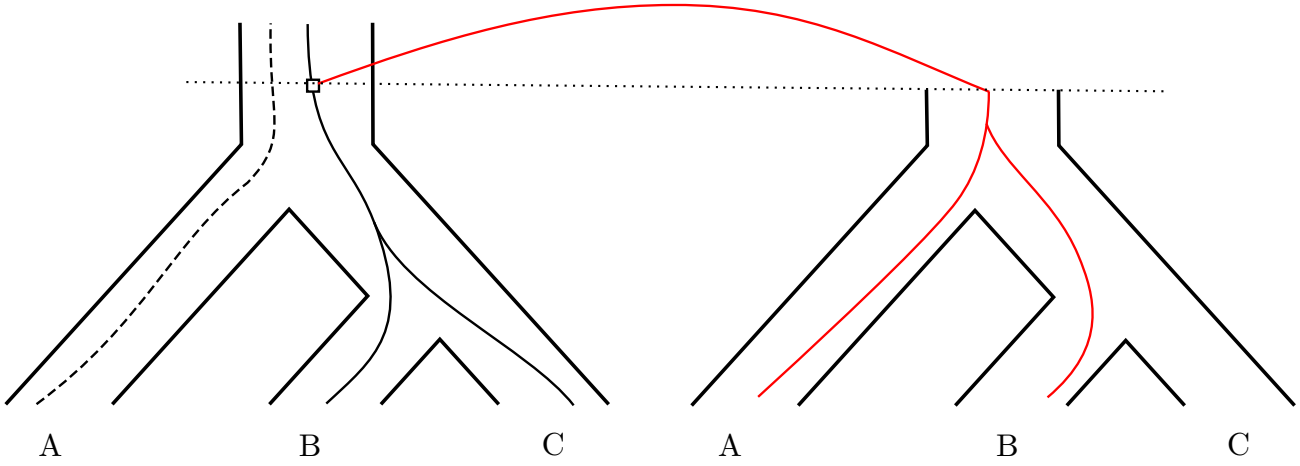
We run a multispecies coalescent process, this generates a *haplotype forest*. We keep the generated haplotype forest in proportion to the number of trees k it contains (probability k/n).
Otherwise, a new haplotype forest is generated (and kept or not, etc.), until a forest is kept.
We choose one of the trees at random to be the locus haplotype tree .



n is the total number of species

The duplication process (joining)

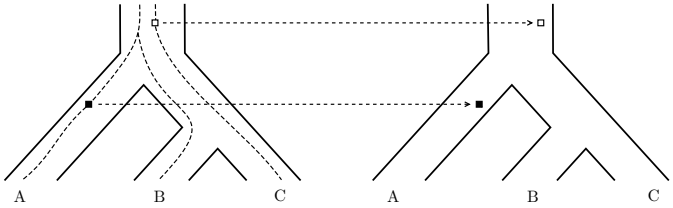
Duplications are joined back to the haplotype tree in the locus in which they occur, using the multispecies coalescent.

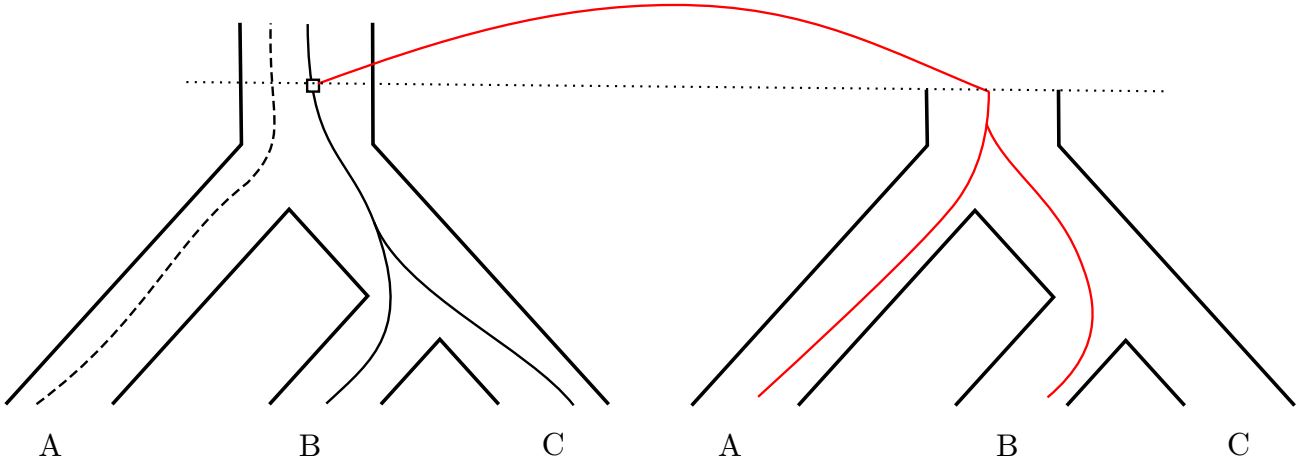


The duplication process (joining)

Duplications are joined back to the haplotype tree in the locus in which they occur, using the multispecies coalescent.

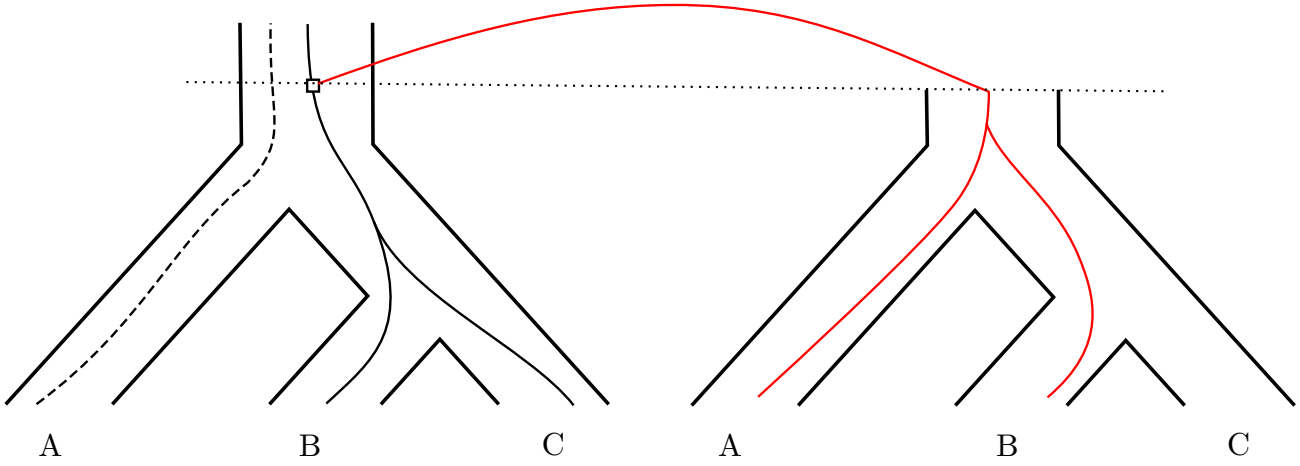
Continue to recursively generate new unilocus trees and haplotype trees until no more duplications are produced.





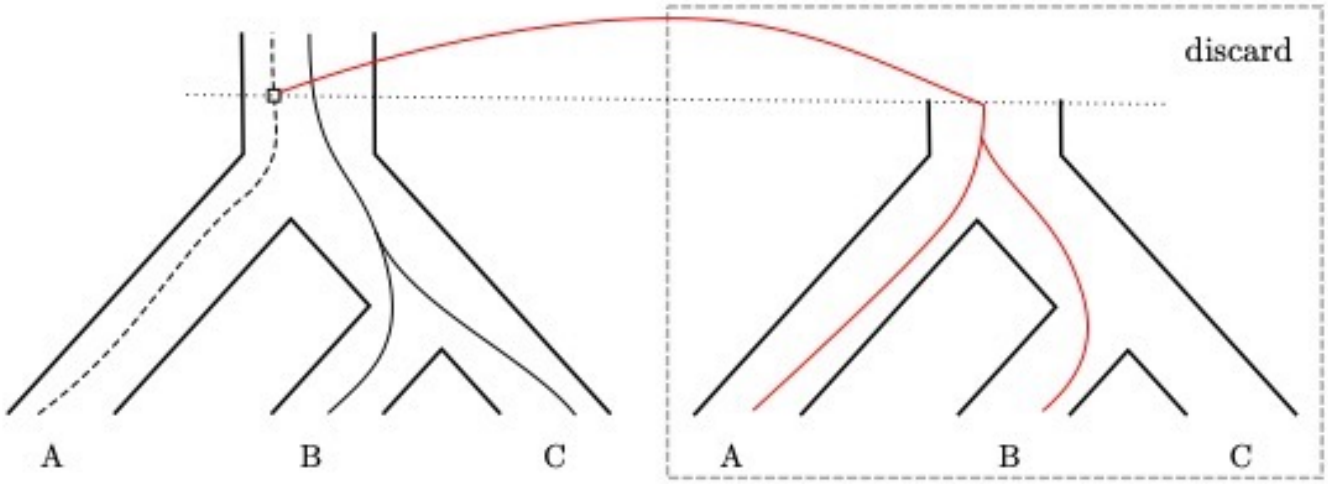
The duplication process (discarding)

In the joining step, duplications are joined to haplotype forests, not just the haplotype tree. If a duplication does not coalesce by the time of origination of the locus, or coalesces to an element of the haplotype forest that is not the haplotype tree (these are only possible in a duplicated locus, not the original locus), the duplication is discarded.



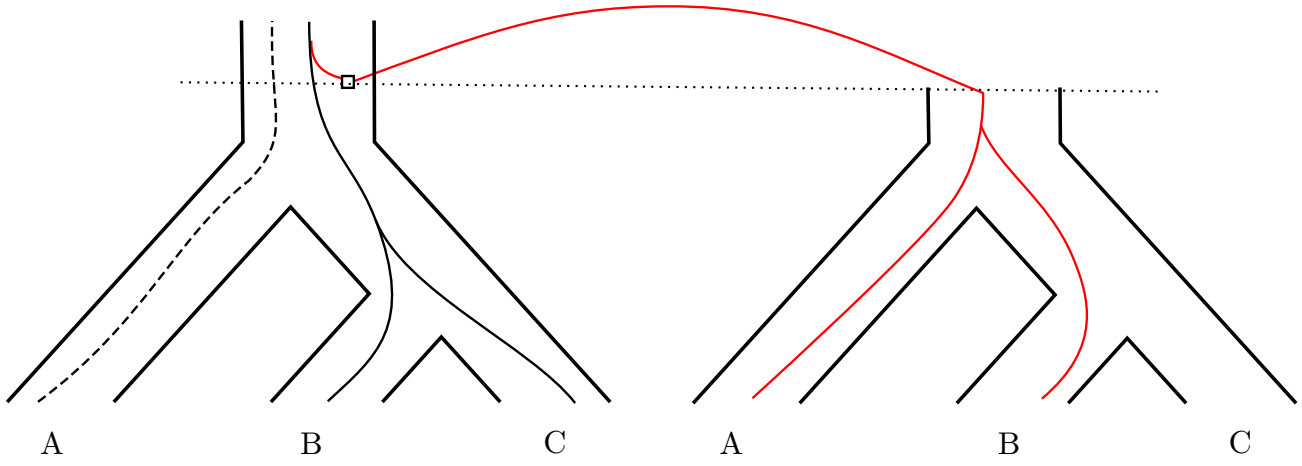
The duplication process (discarding)

In the joining step, duplications are joined to haplotype forests, not just the haplotype tree. If a duplication does not coalesce by the time of origination of the locus, or coalesces to an element of the haplotype forest that is not the haplotype tree (these are only possible in a duplicated locus, not the original locus), the duplication is discarded.



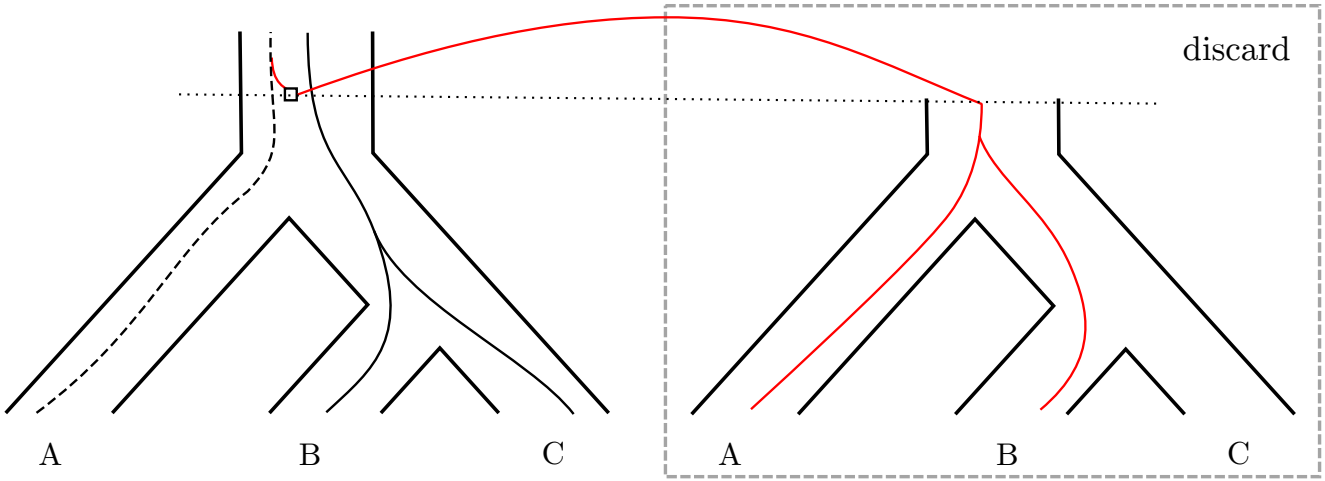
The duplication process (discarding)

In the joining step, duplications are joined to haplotype forests, not just the haplotype tree. If a duplication does not coalesce by the time of origination of the locus, or coalesces to an element of the haplotype forest that is not the haplotype tree (these are only possible in a duplicated locus, not the original locus), the duplication is discarded.



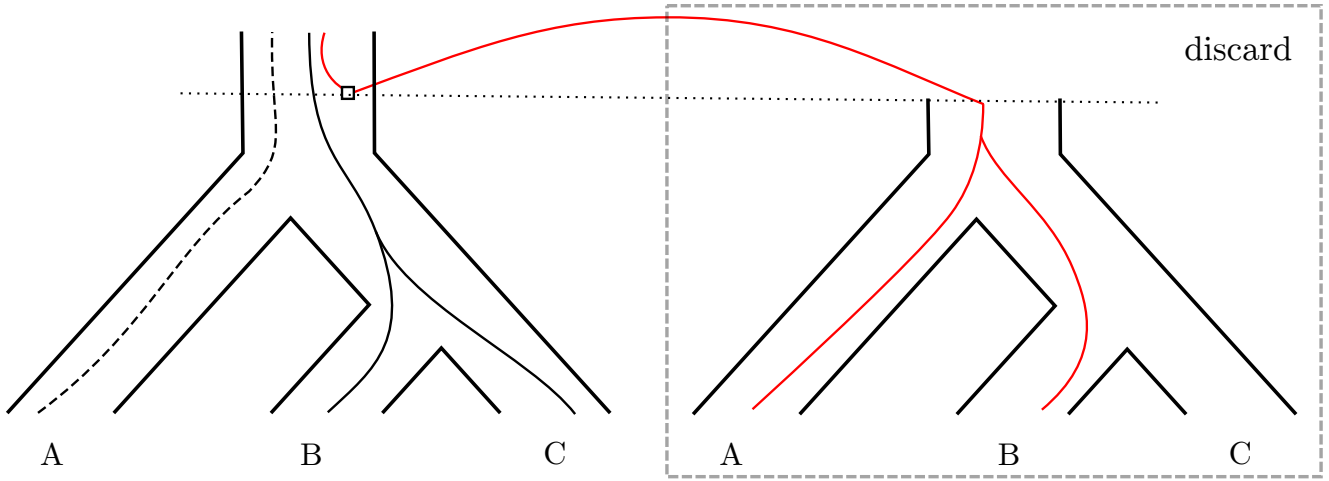
The duplication process (discarding)

In the joining step, duplications are joined to haplotype forests, not just the haplotype tree. If a duplication does not coalesce by the time of origination of the locus, or coalesces to an element of the haplotype forest that is not the haplotype tree (these are only possible in a duplicated locus, not the original locus), the duplication is discarded.



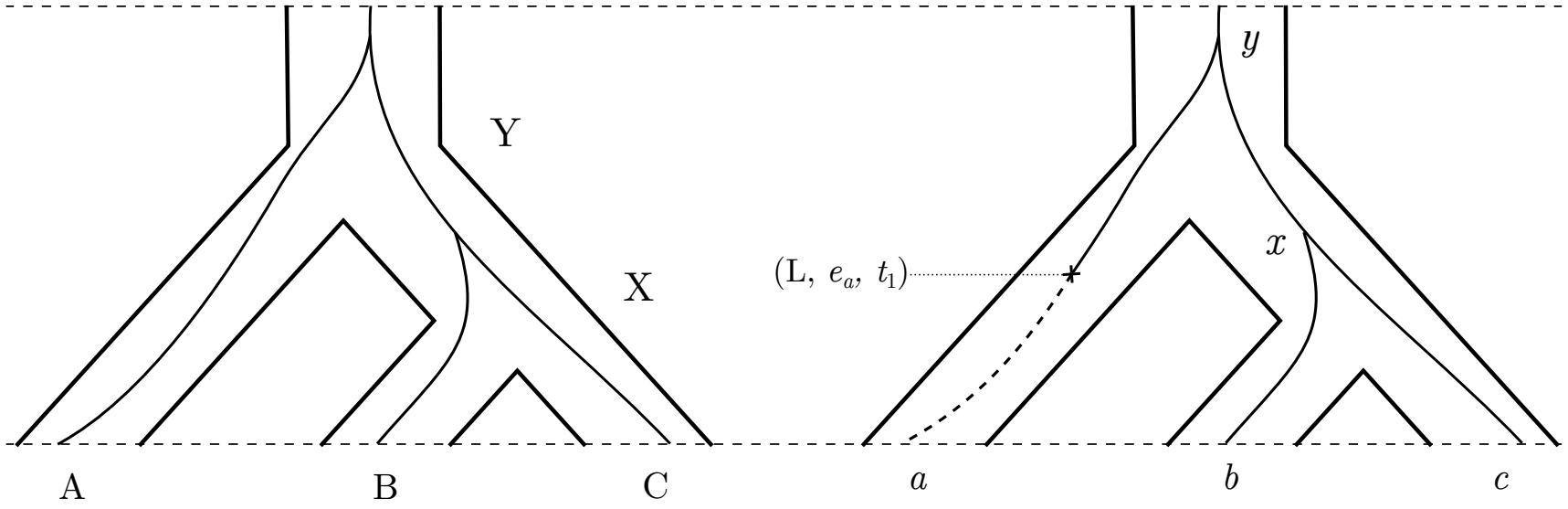
The duplication process (discarding)

In the joining step, duplications are joined to haplotype forests, not just the haplotype tree. If a duplication does not coalesce by the time of origination of the locus, or coalesces to an element of the haplotype forest that is not the haplotype tree (these are only possible in a duplicated locus, not the original locus), the duplication is discarded.



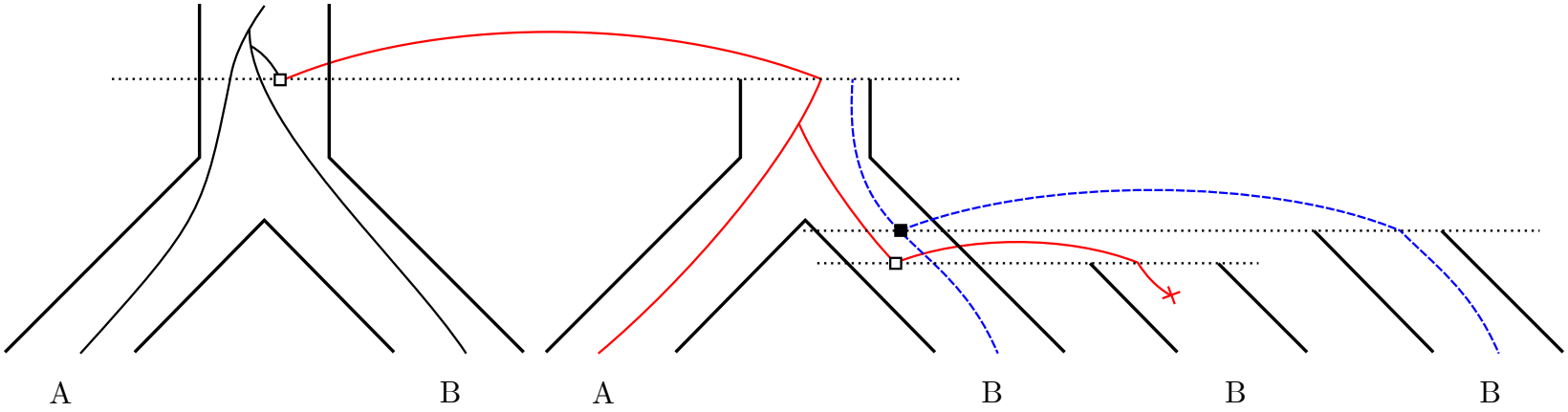
The duplication process (haplotype tree)

The haplotype trees are joined together to form the joined haplotype tree. Losses are simulated at constant rate r'_l on the branches of the joined haplotype tree, and the tree is truncated at these losses to form the final gene tree.



The duplication process (haplotype tree)

The haplotype trees are joined together to form the joined haplotype tree. Losses are simulated at constant rate r'_l on the branches of the joined haplotype tree, and the tree is truncated at these losses to form the final gene tree.

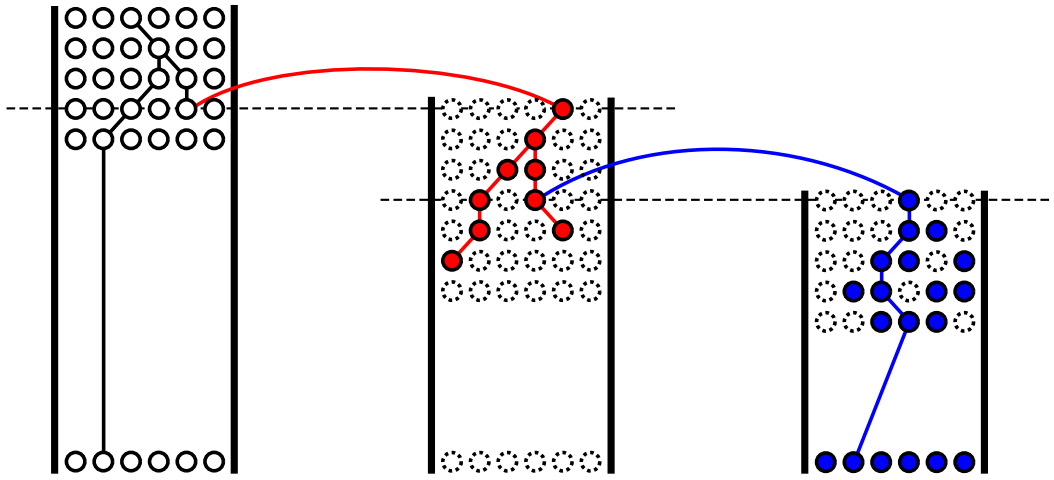


MLMSC is a good estimation of WFDL

If all duplications are *self-descending*, MLMSC = WFDL

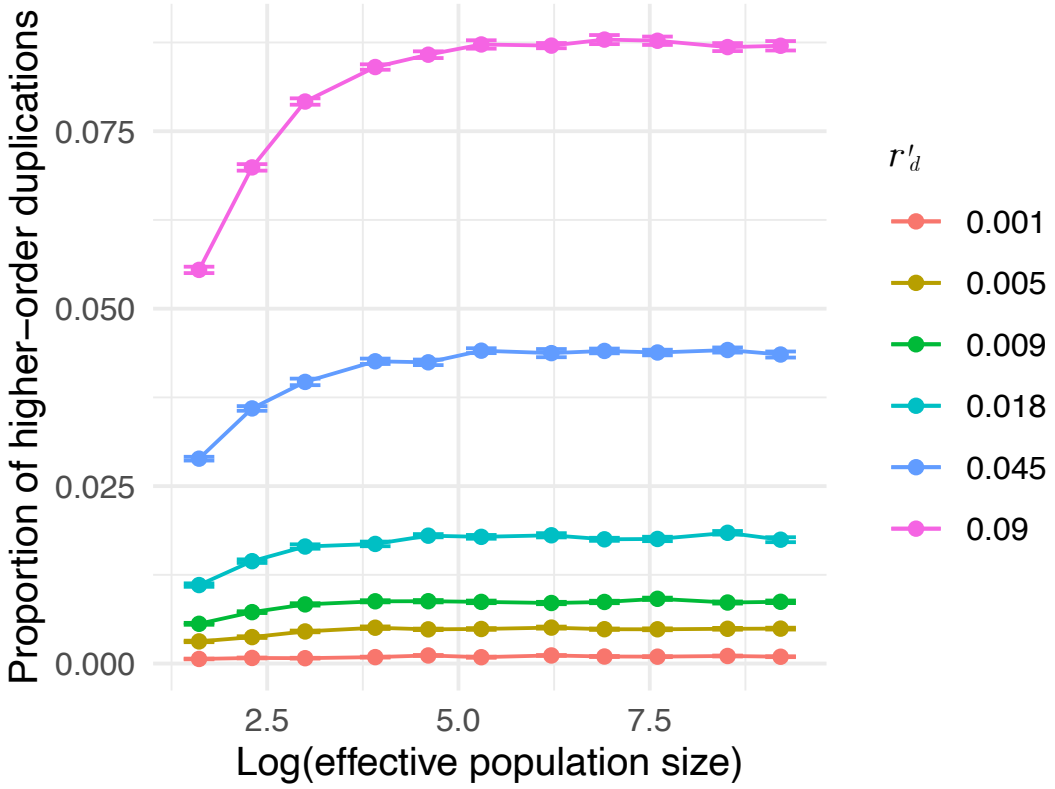
- incomplete coalescent and rejection sampling
- calculating rates of events with the coalescent-rate process

If not, the true rate of observed duplications is slightly higher because we must also consider duplications that are not self-descending, but are observed in descendant loci (*higher-order duplications*).



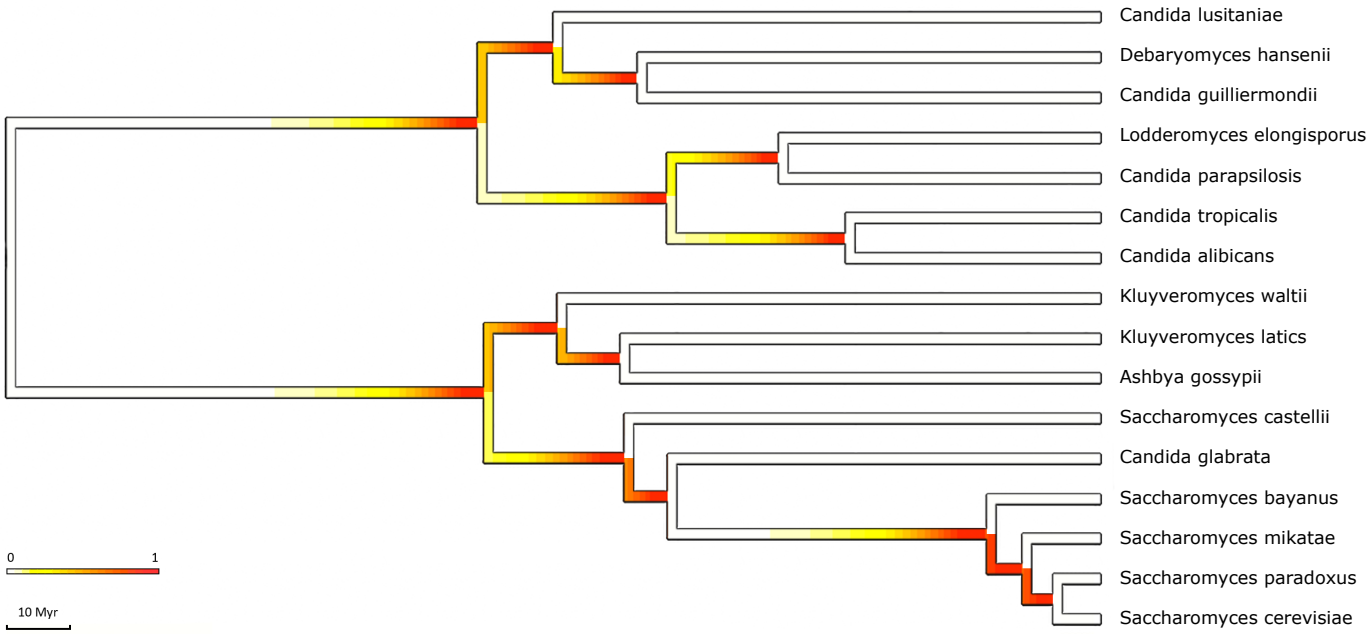
MLMSC is a good estimation of WFDL

We find that under realistic parameters, we can expect the frequency of higher-order duplications to be about 1% of all observed duplications.



Probability of copy number hemiplasy

Difficult to write down an closed formula for but relatively simple to calculate it for any given species tree, branch and time ($2N = 9 \times 10^7$).



Impact in reconstructing species trees

ASTRAL is the most used summary method, which takes a set of gene trees as input and *summarises* them in some way to reconstruct a species tree.

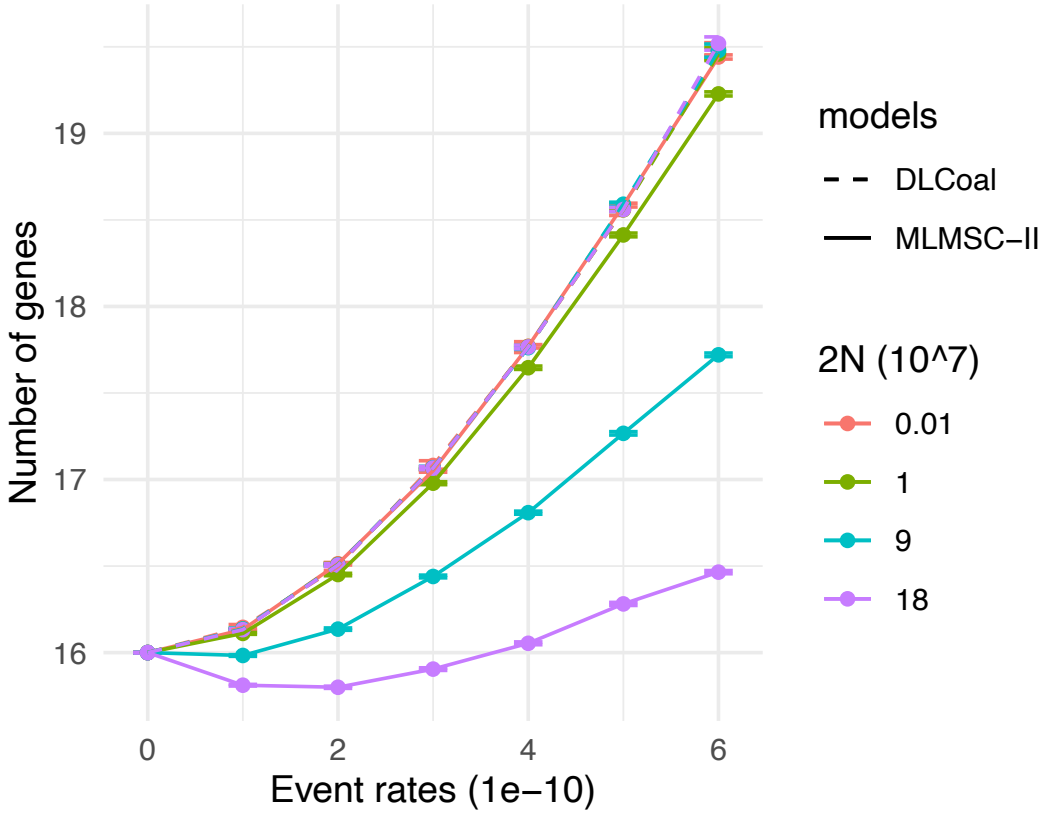
The accuracy of ASTRAL has been heavily studied, both theoretically and via **simulations:**

- Take a data set (species tree, gene trees, etc)
- Estimate the duplication rate by calculating the expected number of genes under a birth-and-death DL process, and matching this to the observed number of genes. ($r_{d=}r_l$).
- Simulate gene trees via DLCoal
- Test the accuracy

We studied the effect of copy number hemiplasy on the performance of ASTRAL and the practical estimation of duplication rate.

Impact in reconstructing species trees

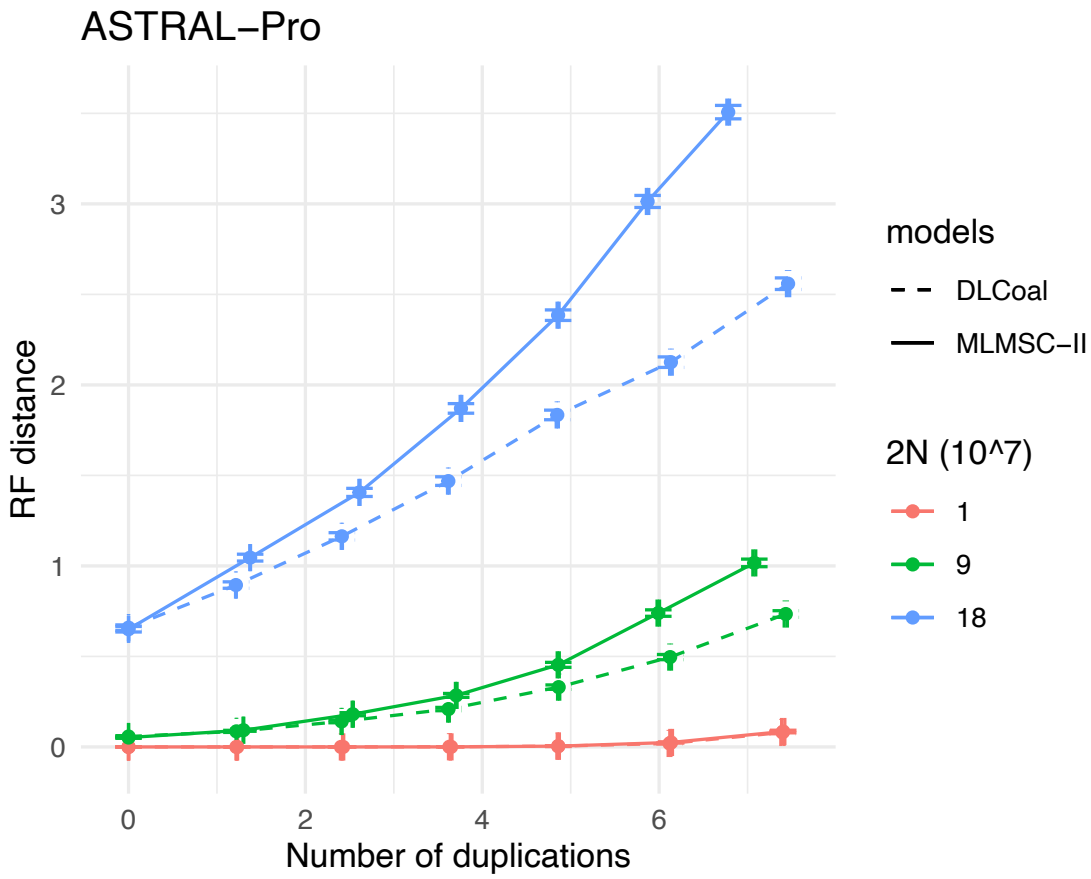
The duplication rates calibrated from real data have been potentially heavily underestimated.



The shapes of the trees generated by the two models differ in subtle but significant ways

Impact in reconstructing species trees

The performances of ASTRAL degrade, when CNH is modelled compared to when it is not.



Conclusions

We showed the MLMSC model is appropriate to study the effect of copy number hemiplasy.

We showed that the assumption that copy number hemiplasy does not have a noticeable impact on gene family evolution is often not satisfied, by proving that the frequency of CNH is higher than previously thought, and has a noticeable effect on both the shape of gene trees and their suitability for species tree inference.

It is essential to develop new methods that fully take into account copy number hemiplasy.