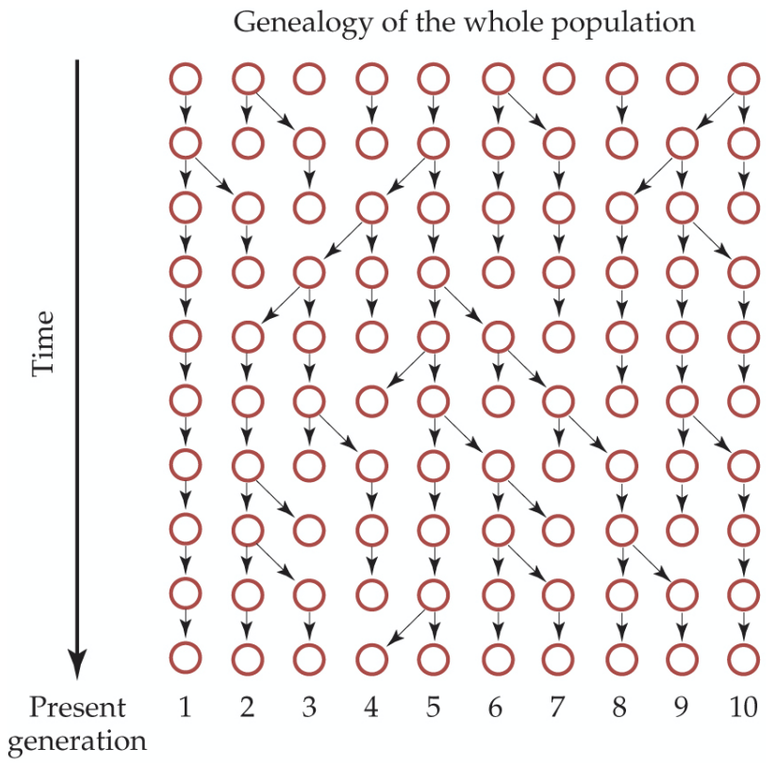
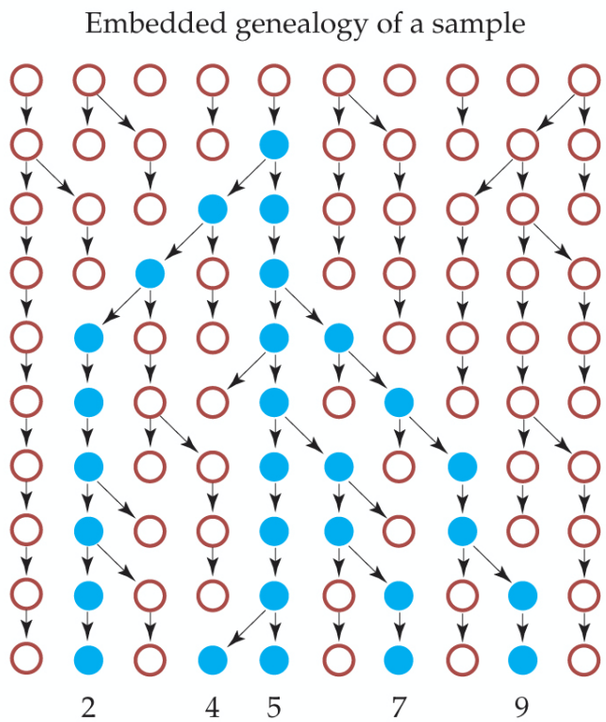
A unified model of duplication, loss, introgression, and coalescence

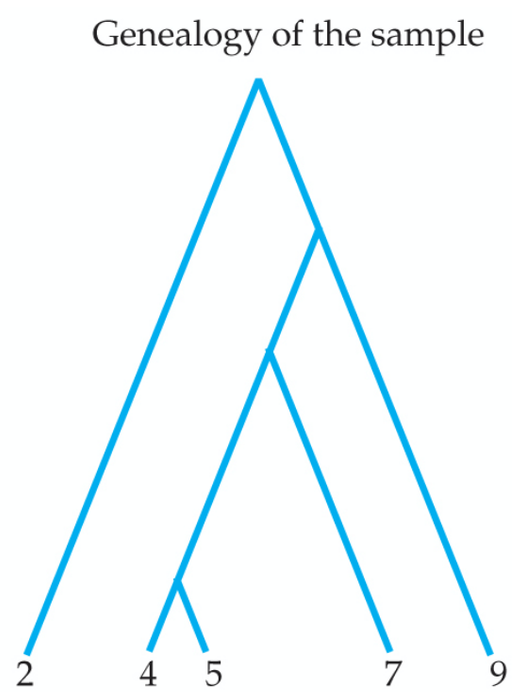
Matthew Hahn

Department of Biology
& Department of Computer Science
Indiana University

The coalescent model

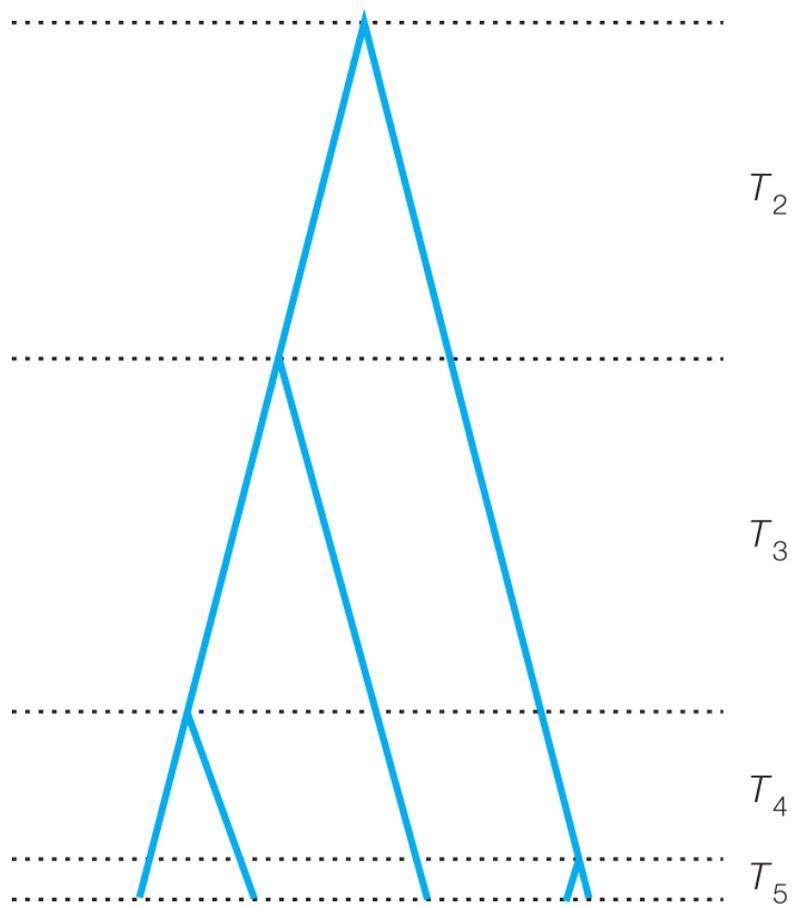






Kingman (1982)

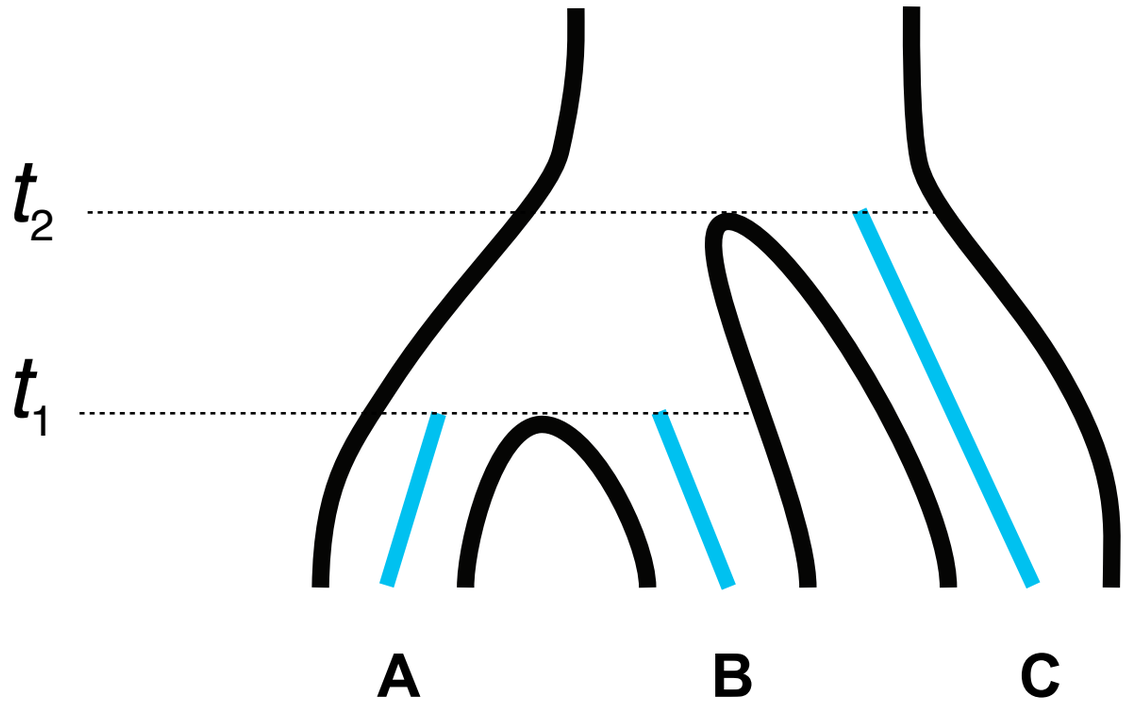
The coalescent model



The coalescent model

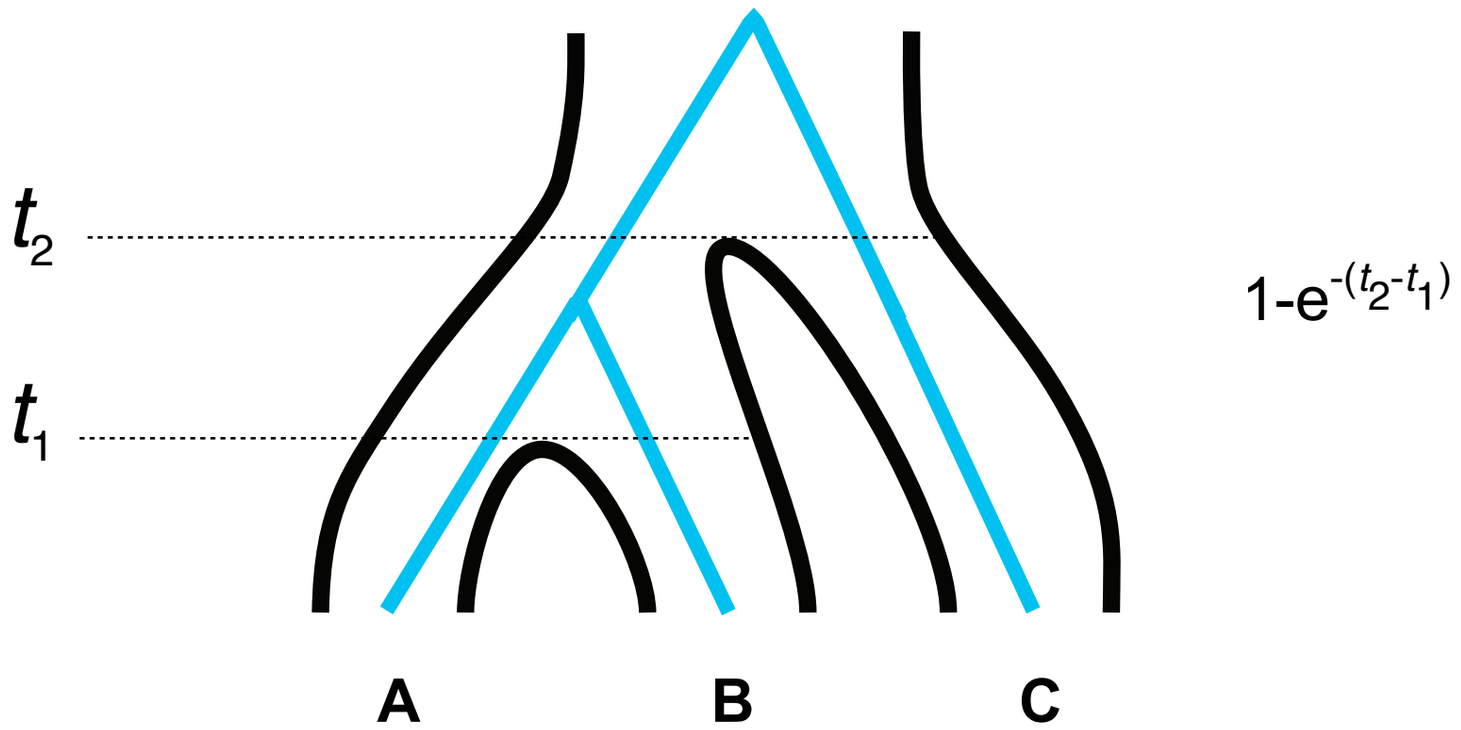
The coalescent model

The multispecies coalescent



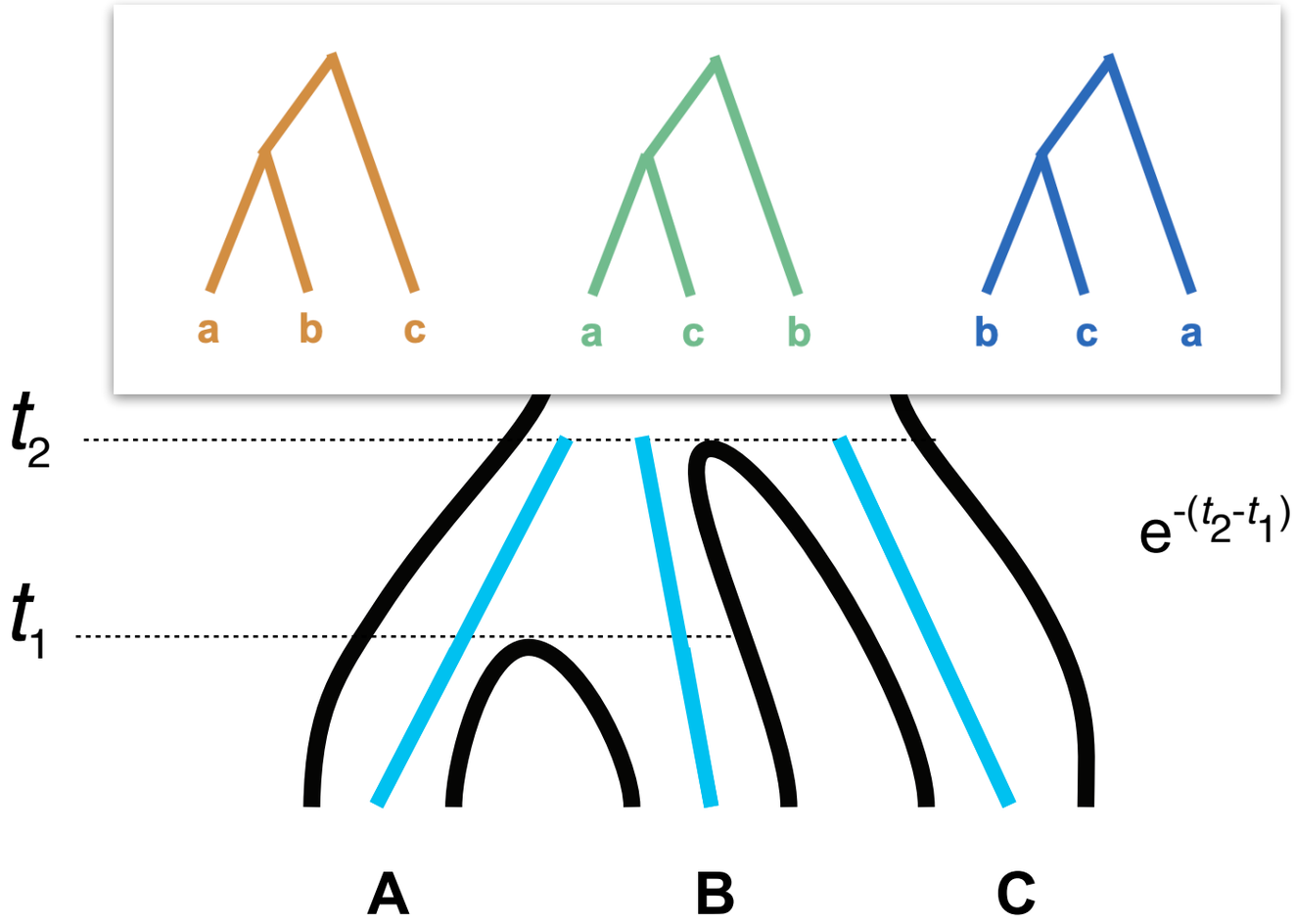
Hudson (1983); Tajima (1983)

The multispecies coalescent



Hudson (1983); Tajima (1983)

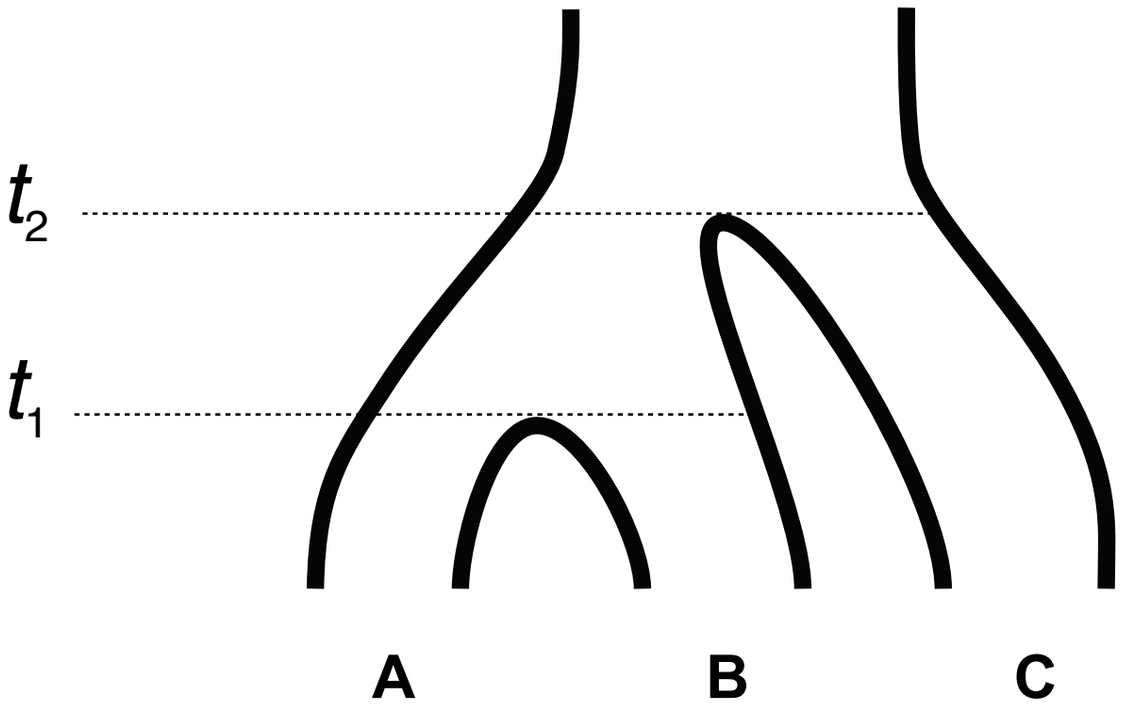
The multispecies coalescent



Incomplete lineage sorting

Hudson (1983); Tajima (1983)

Incomplete lineage sorting



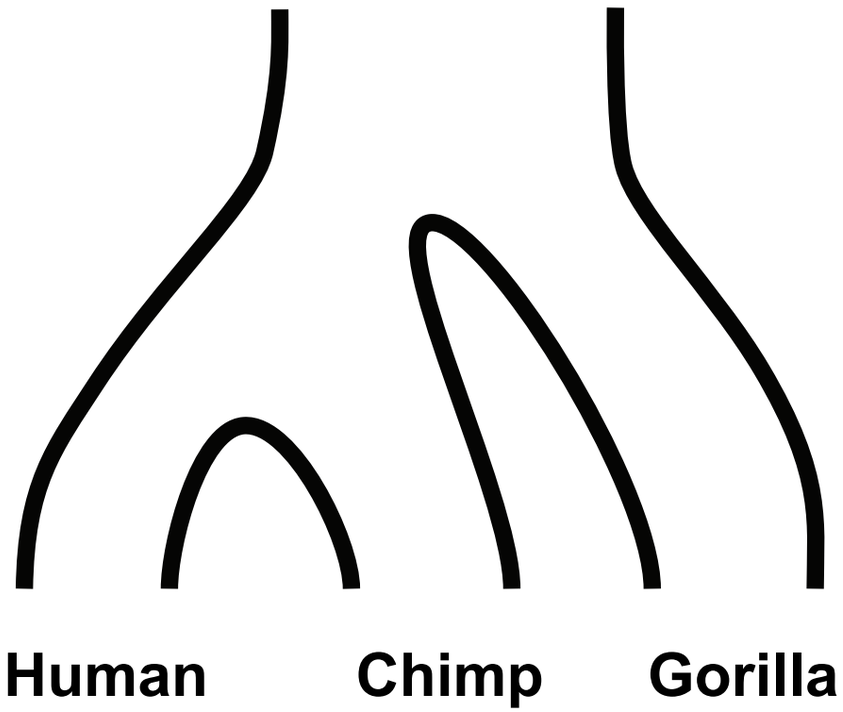
$1-(2/3)e^{-(t_2-t_1)}$

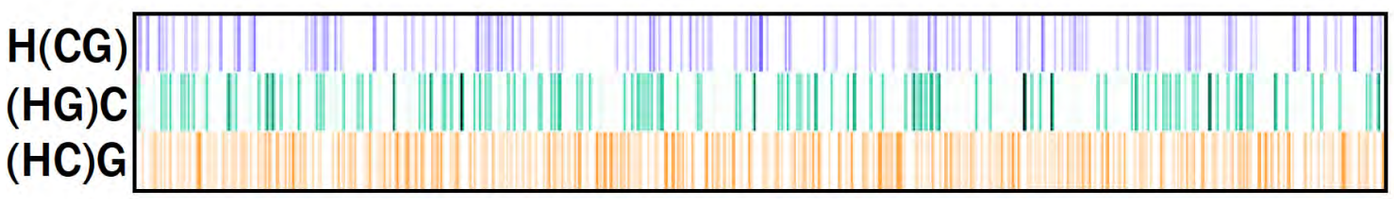
$(1/3)e^{-(t_2-t_1)}$

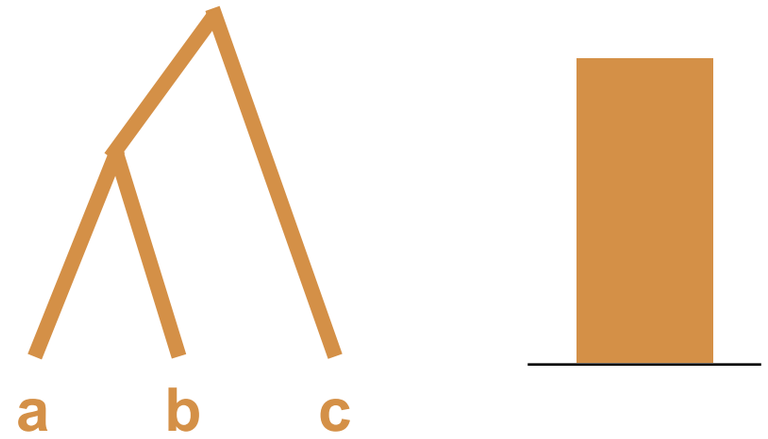
$(1/3)e^{-(t_2-t_1)}$

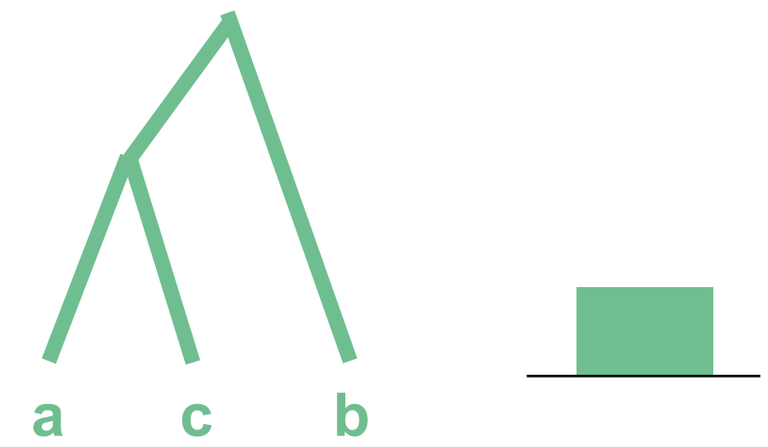
Hudson (1983)

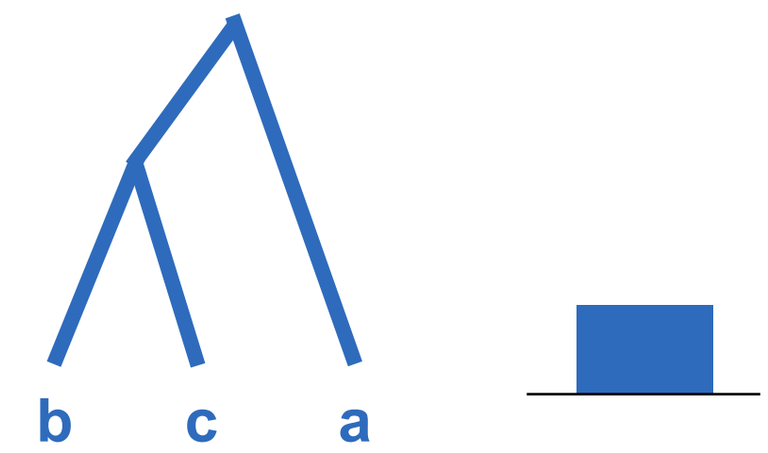
Incomplete lineage sorting in the great apes









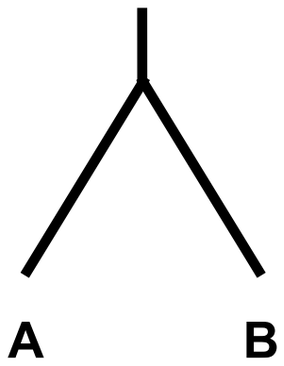


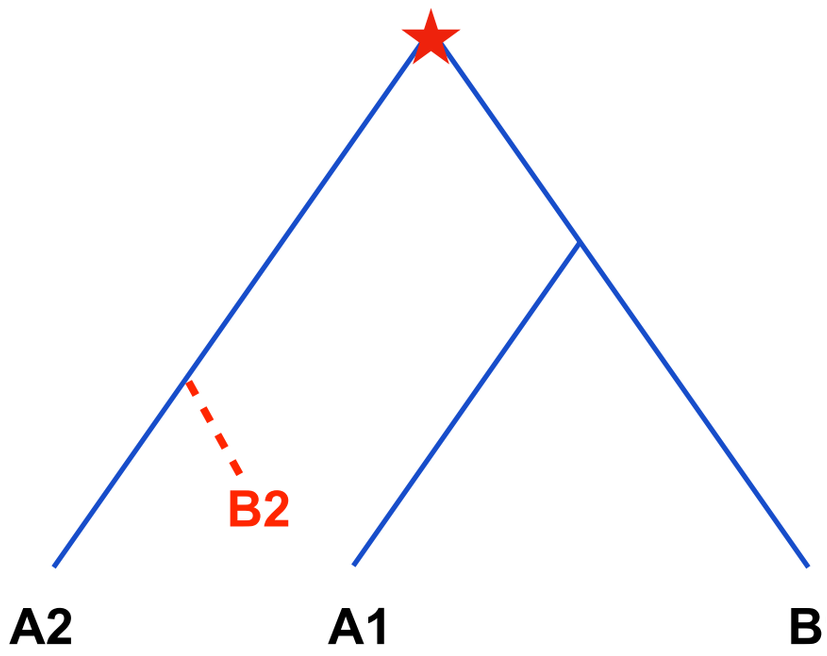
Adding duplications and losses

Sarthak Mishra

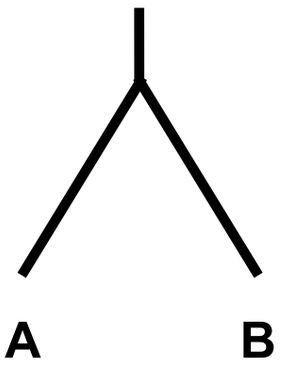
Megan Smith

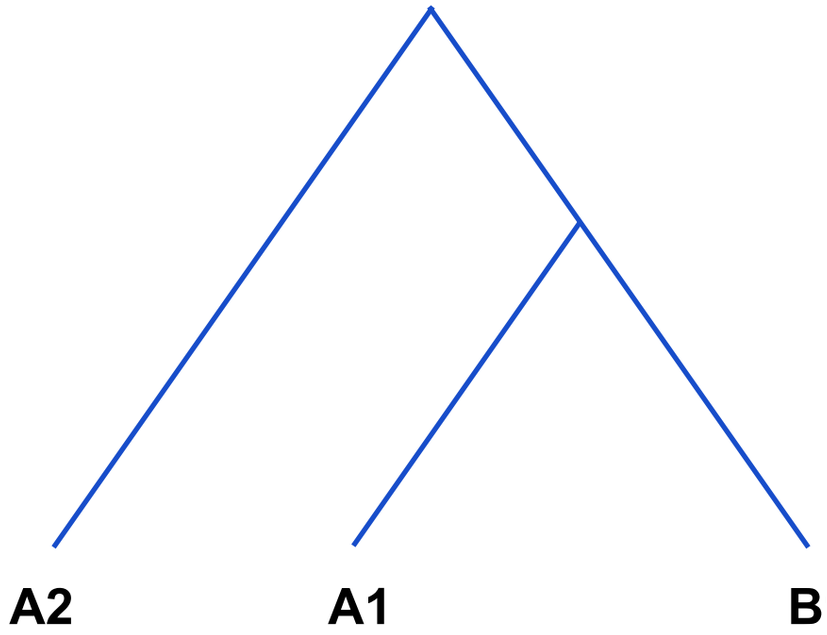
What is the history of this gene tree?





What is the history of this gene tree?





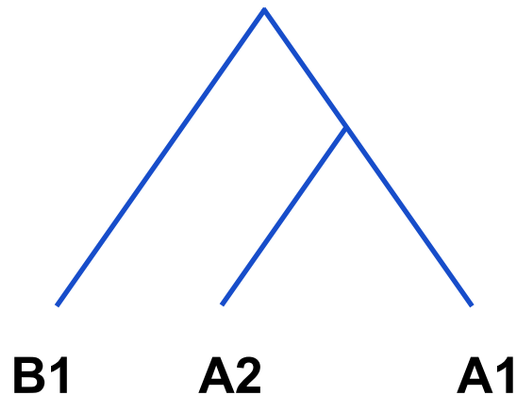
How many duplications?

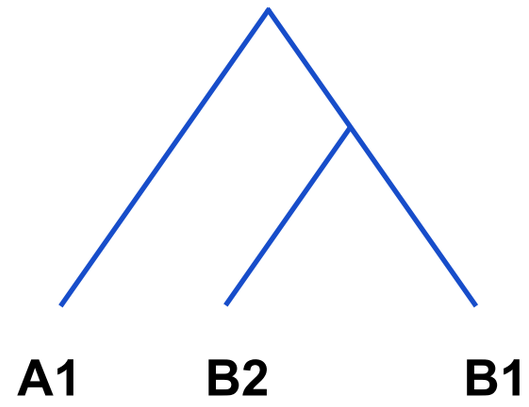
How many losses?

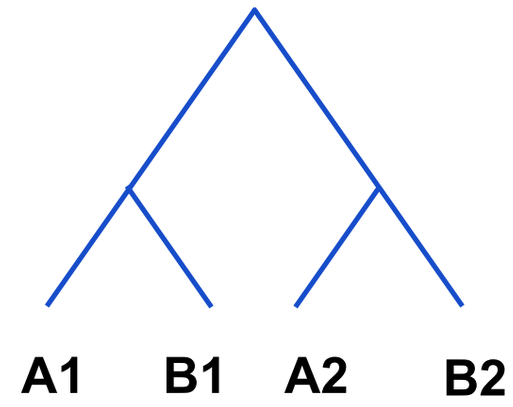
When did duplication/loss occur?

Which genes are orthologs/paralogs?

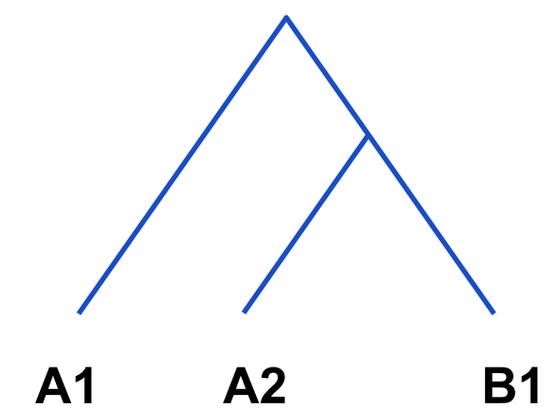
How many trees are possible with 1 duplication?

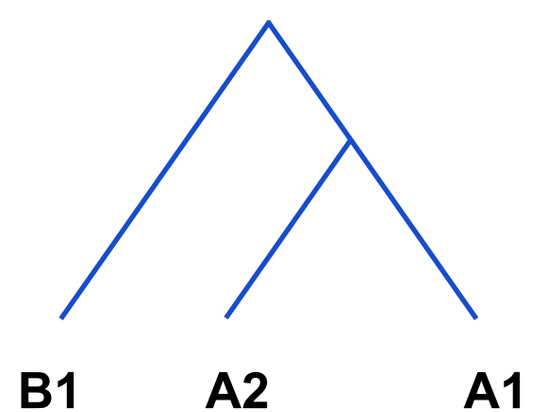


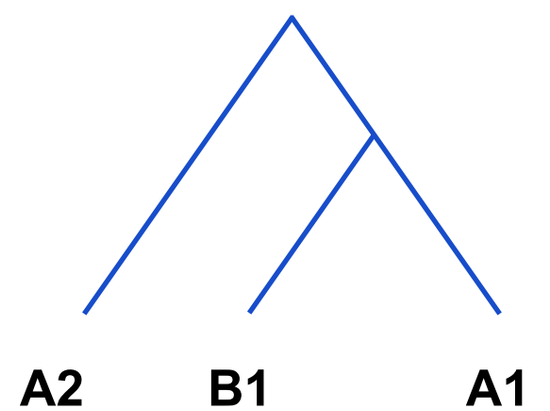


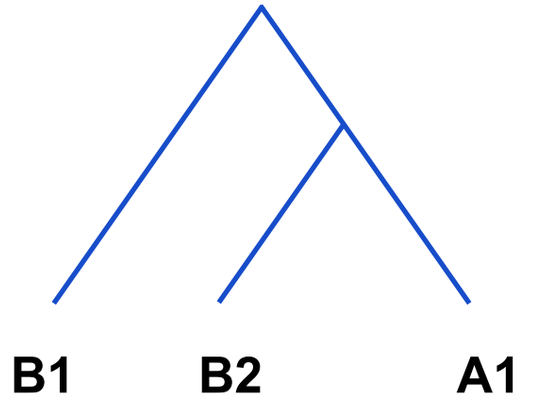


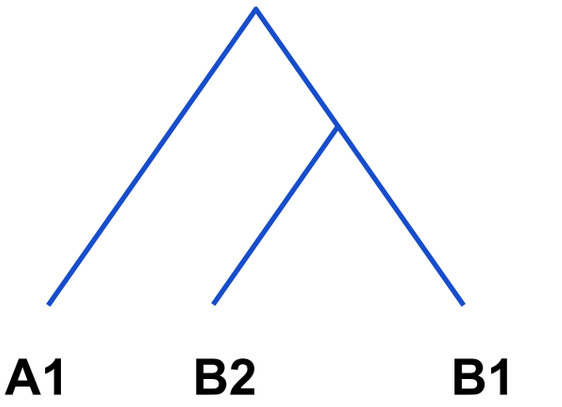
How many trees are possible with 1 duplication?

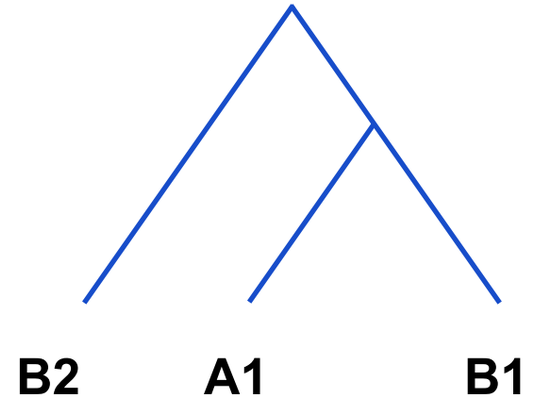


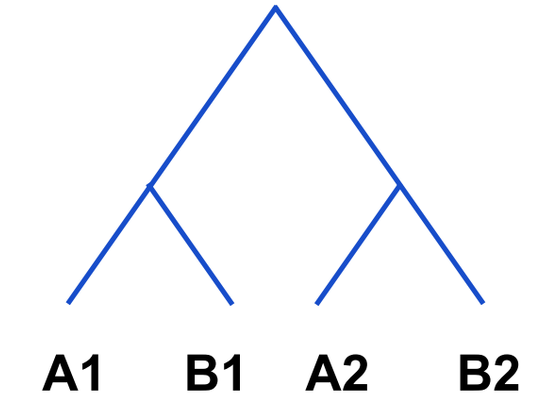


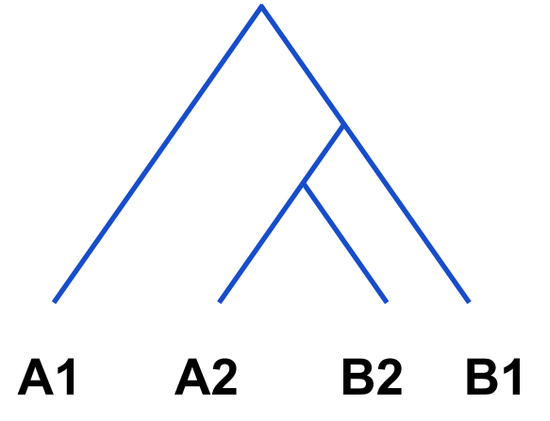


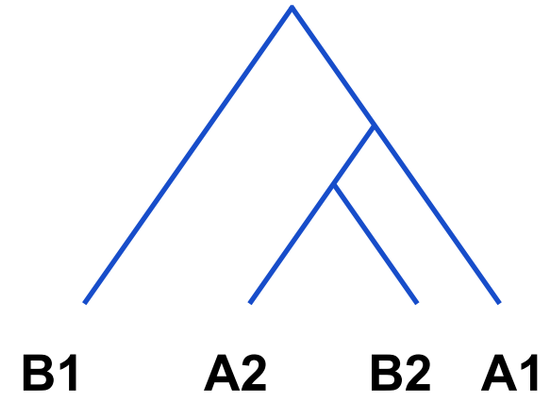




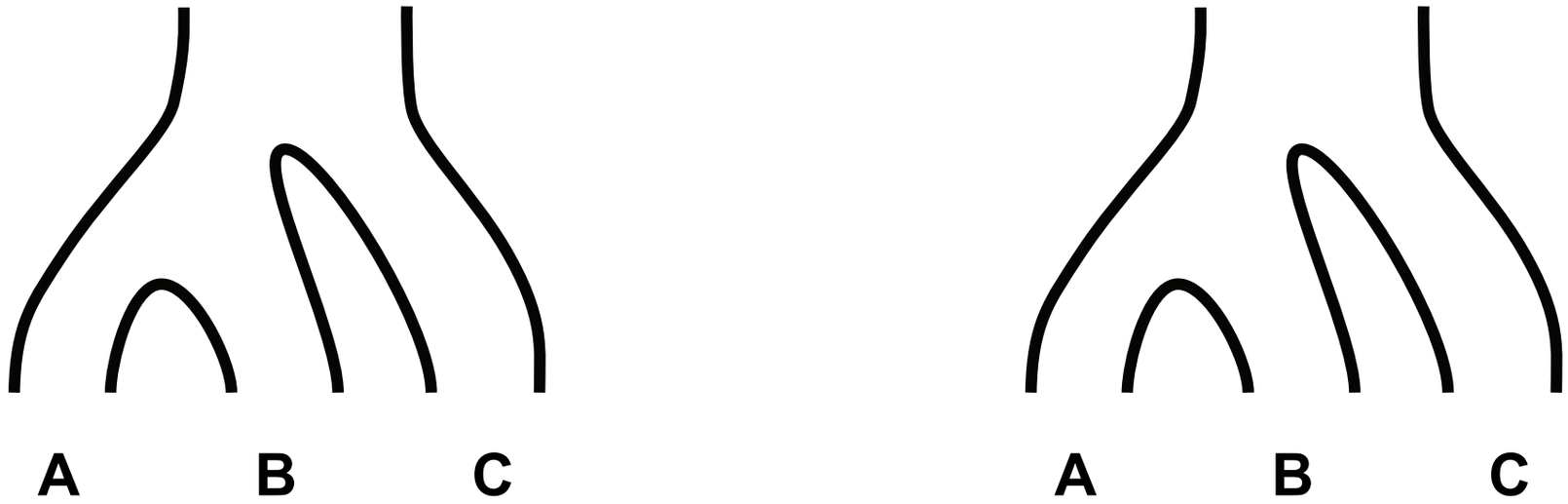






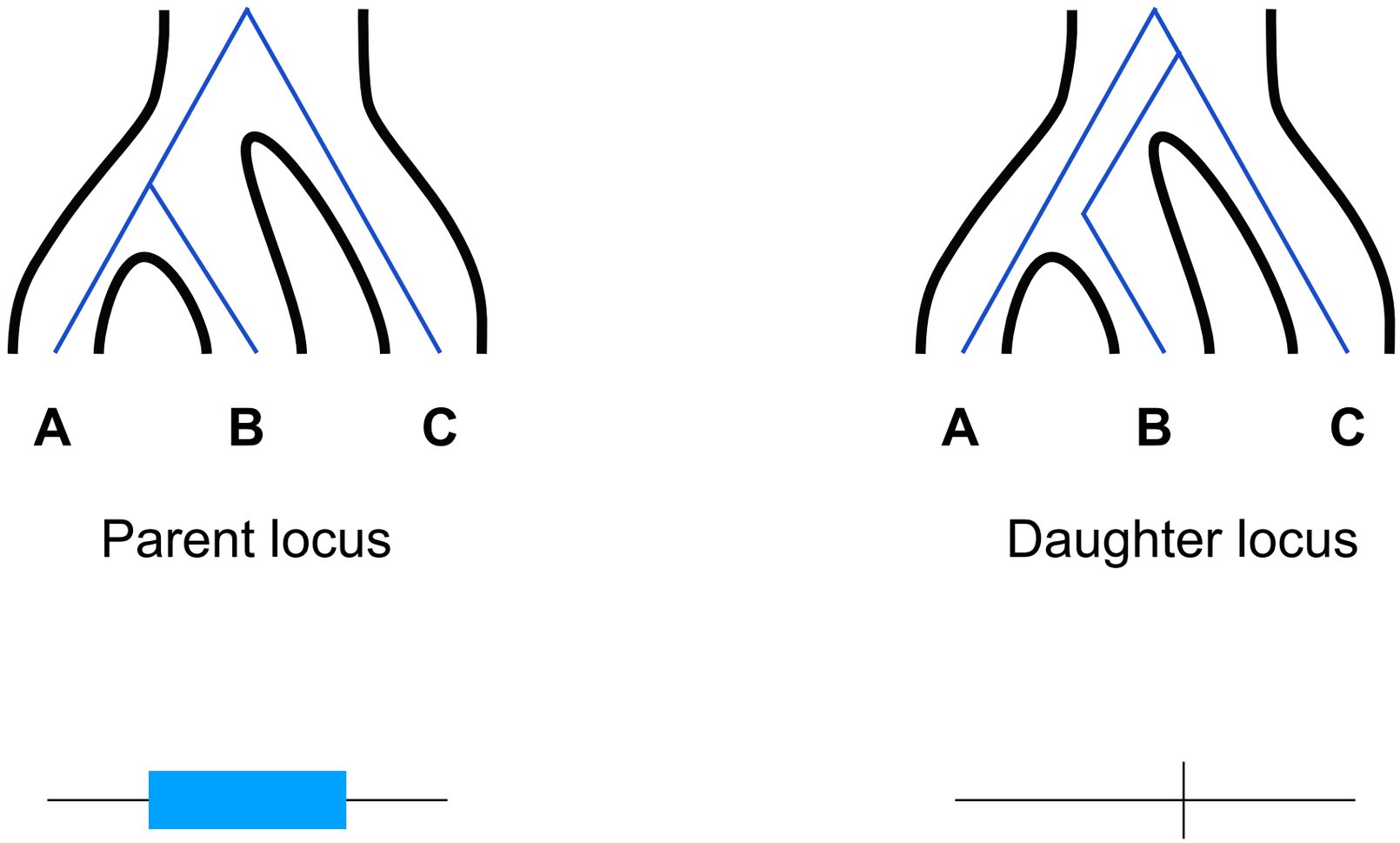


The multilocus multispecies coalescent (MLMSC)



Li et al. (2021)

The multilocus multispecies coalescent (MLMSC)

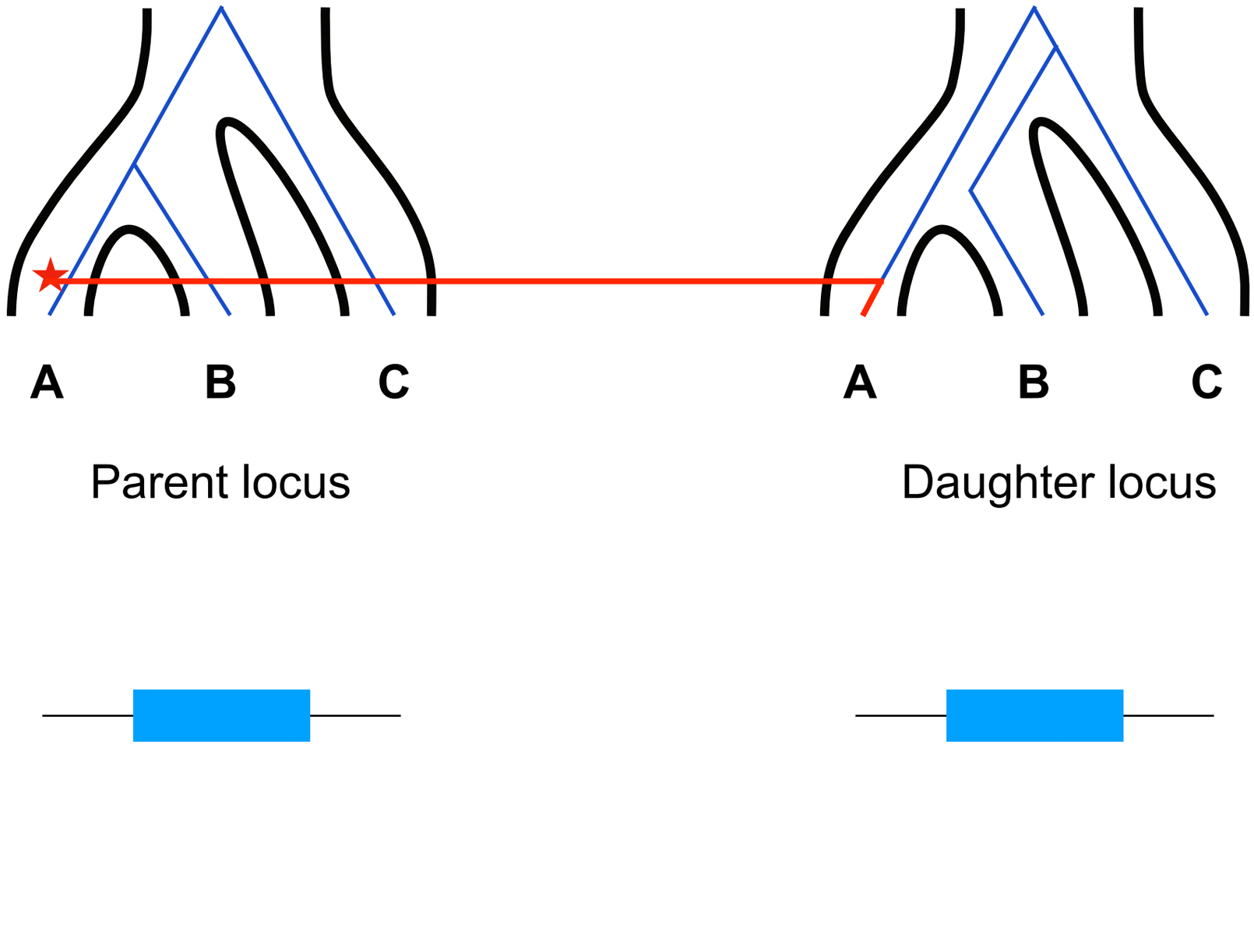


Parent locus

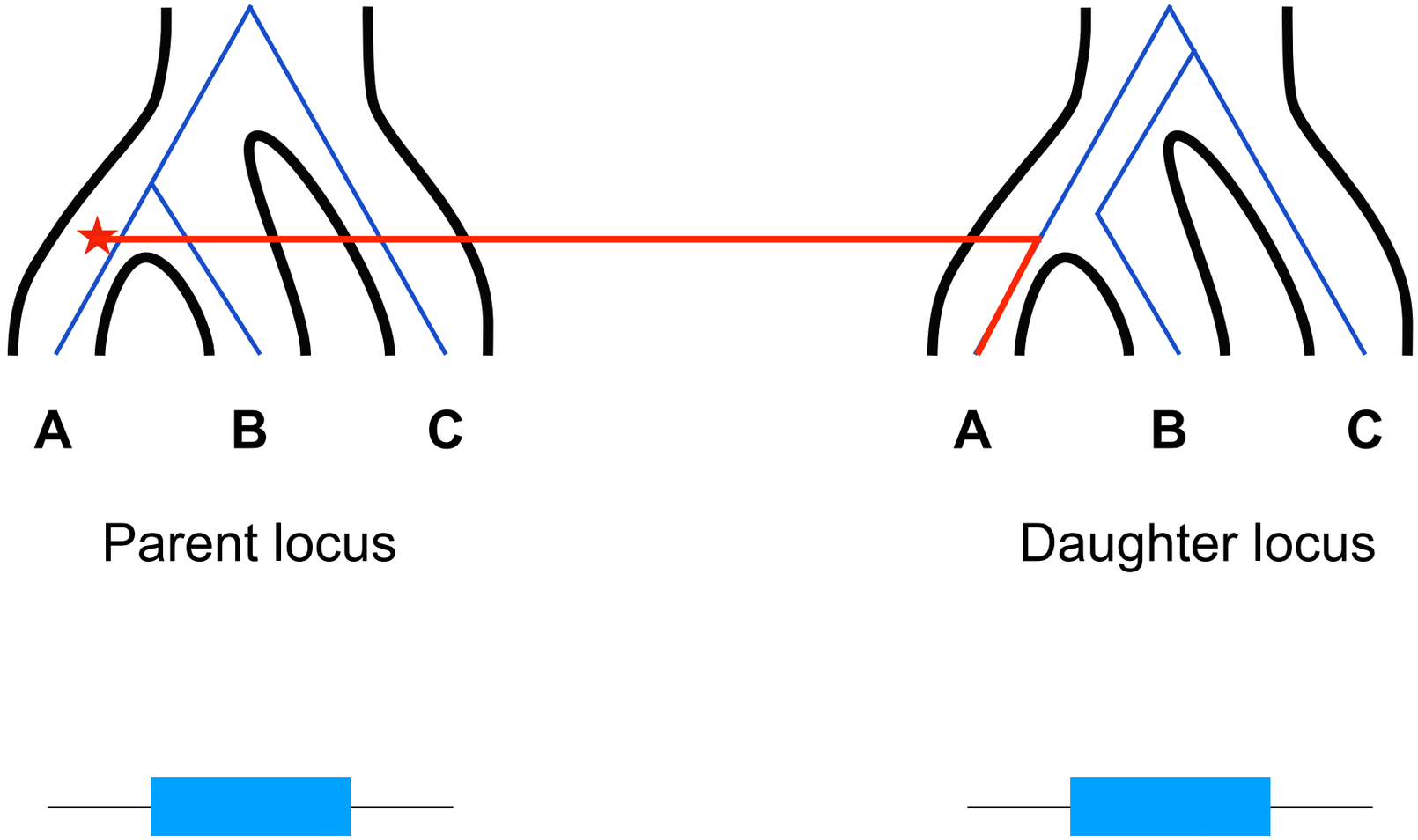
Daughter locus

1. Every locus has its own gene tree drawn from the MSC

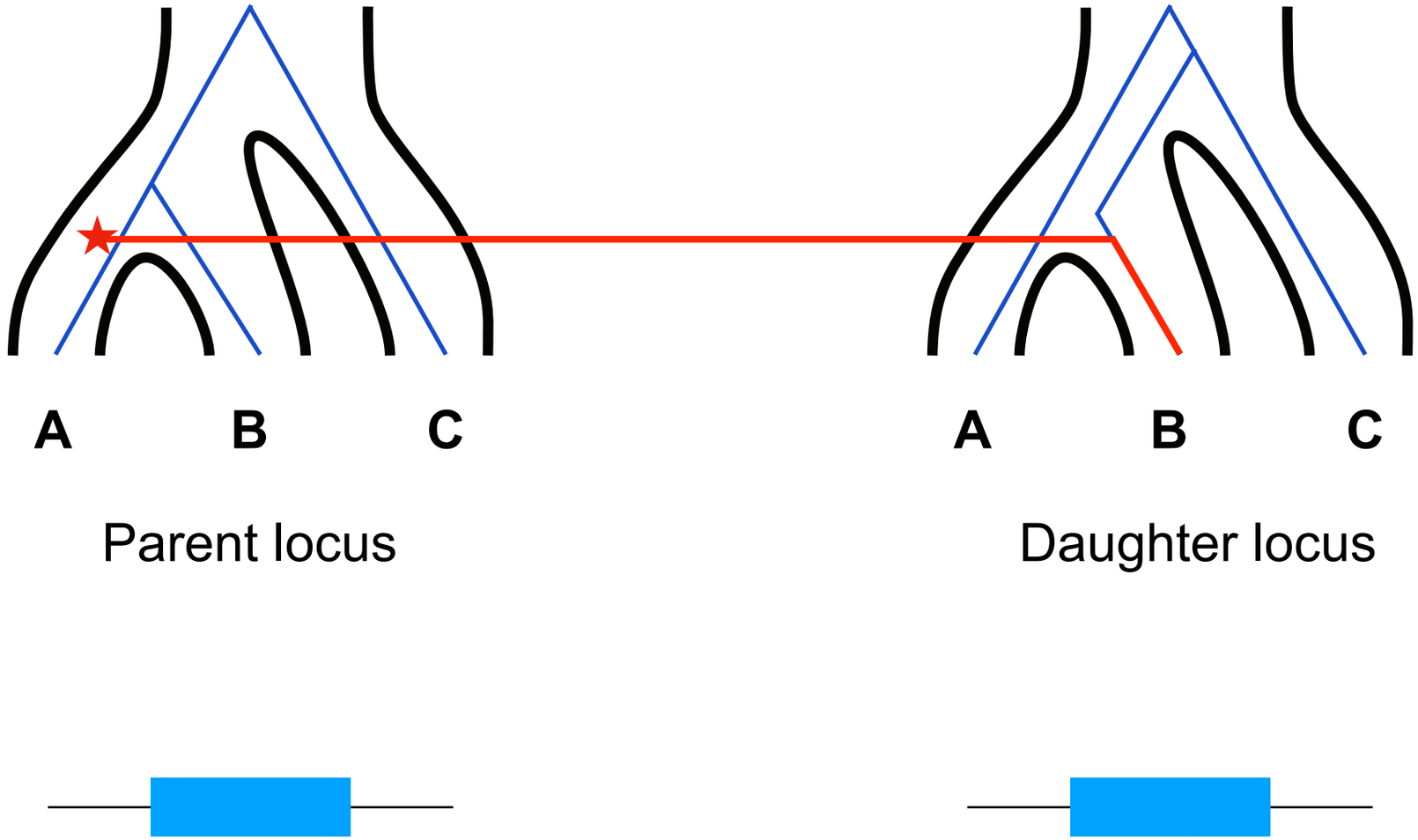
Duplications follow the local gene tree



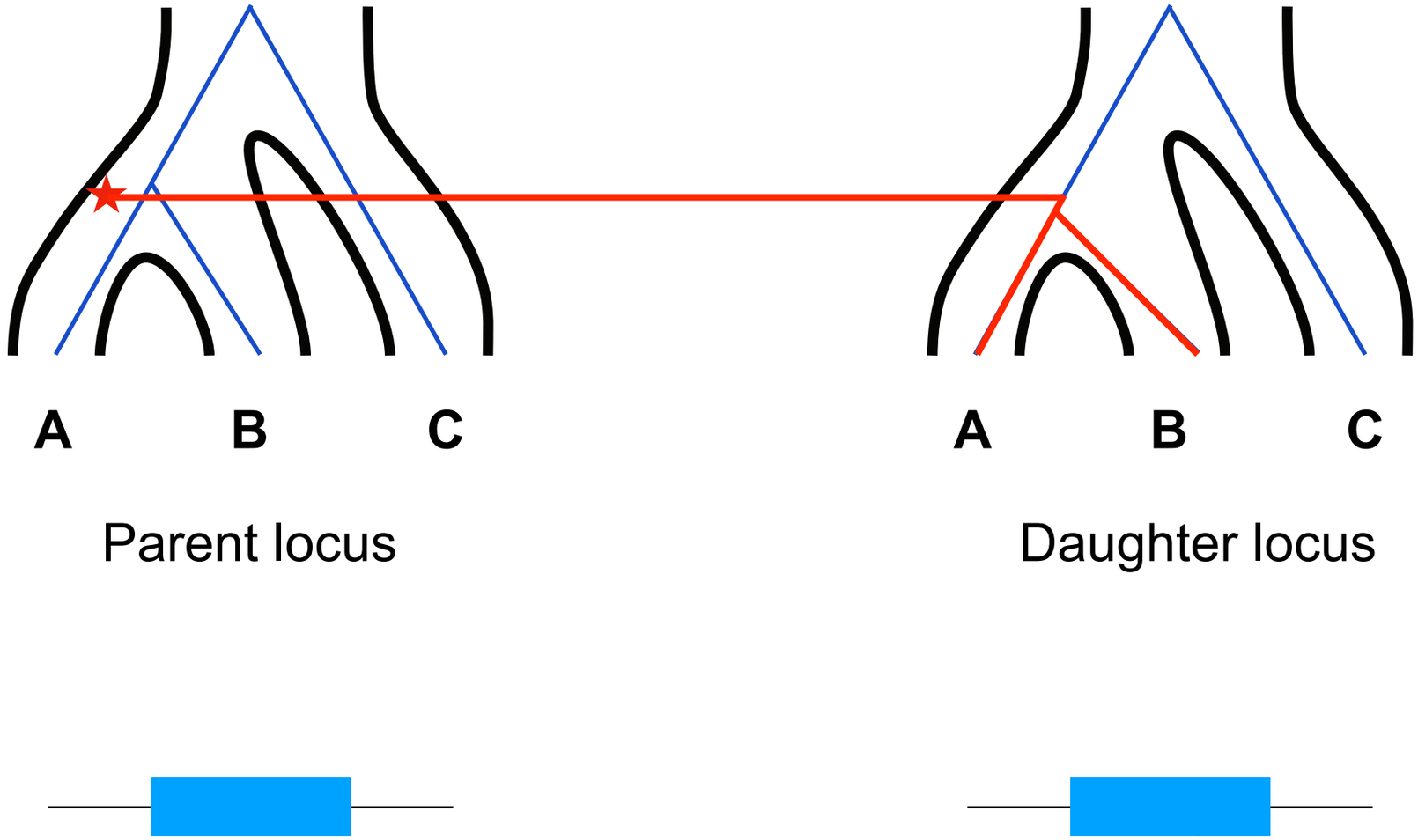
It gets more complicated



It gets more complicated

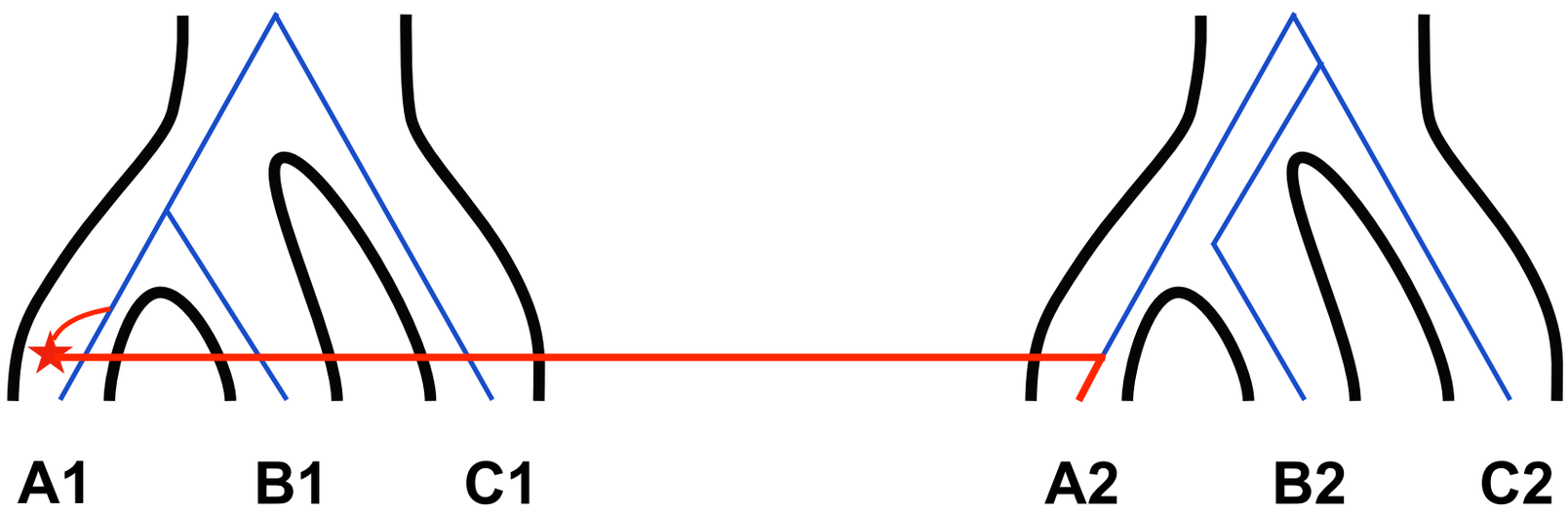


It gets more complicated



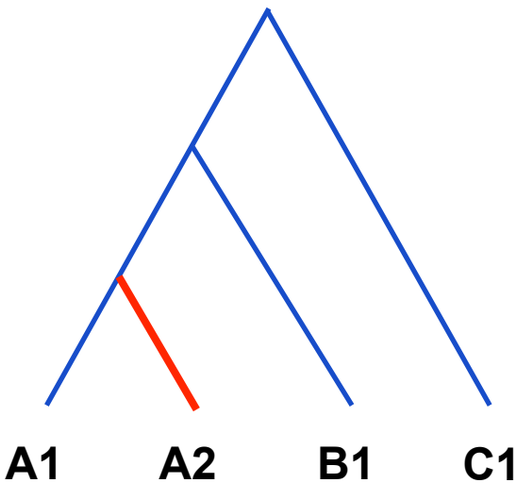
2. Duplications (and losses) follow the gene tree at the locus at which they occur

Coalescence joins parent and daughter trees

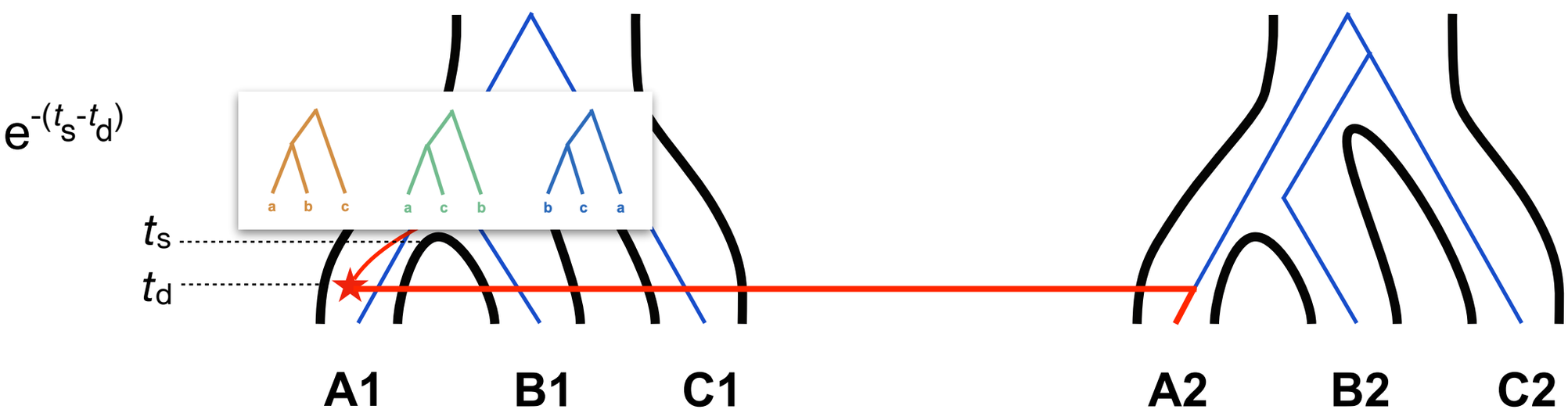


Parent locus

Daughter locus

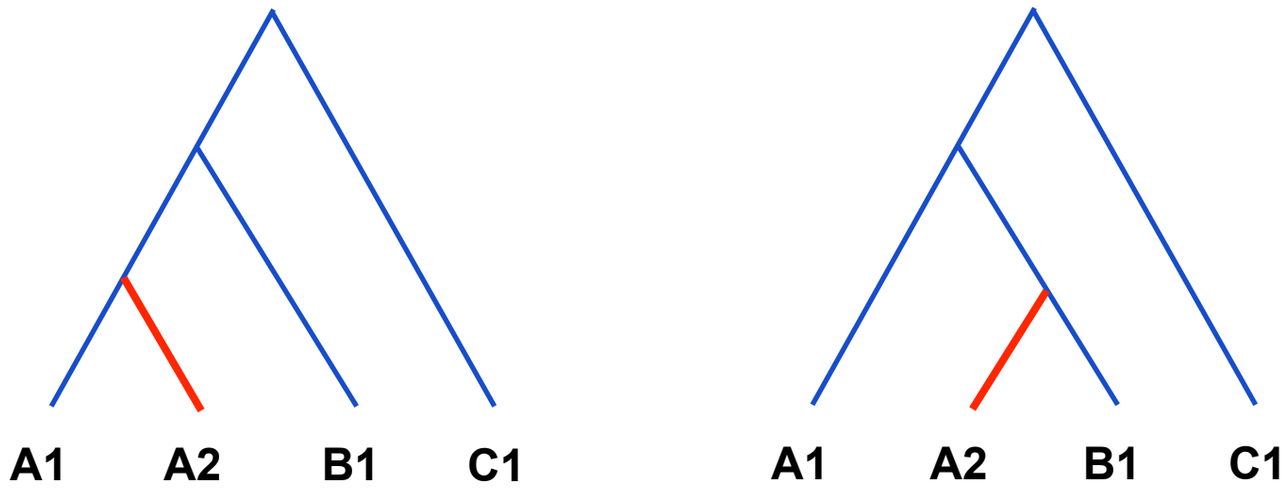


Coalescence joins parent and daughter trees

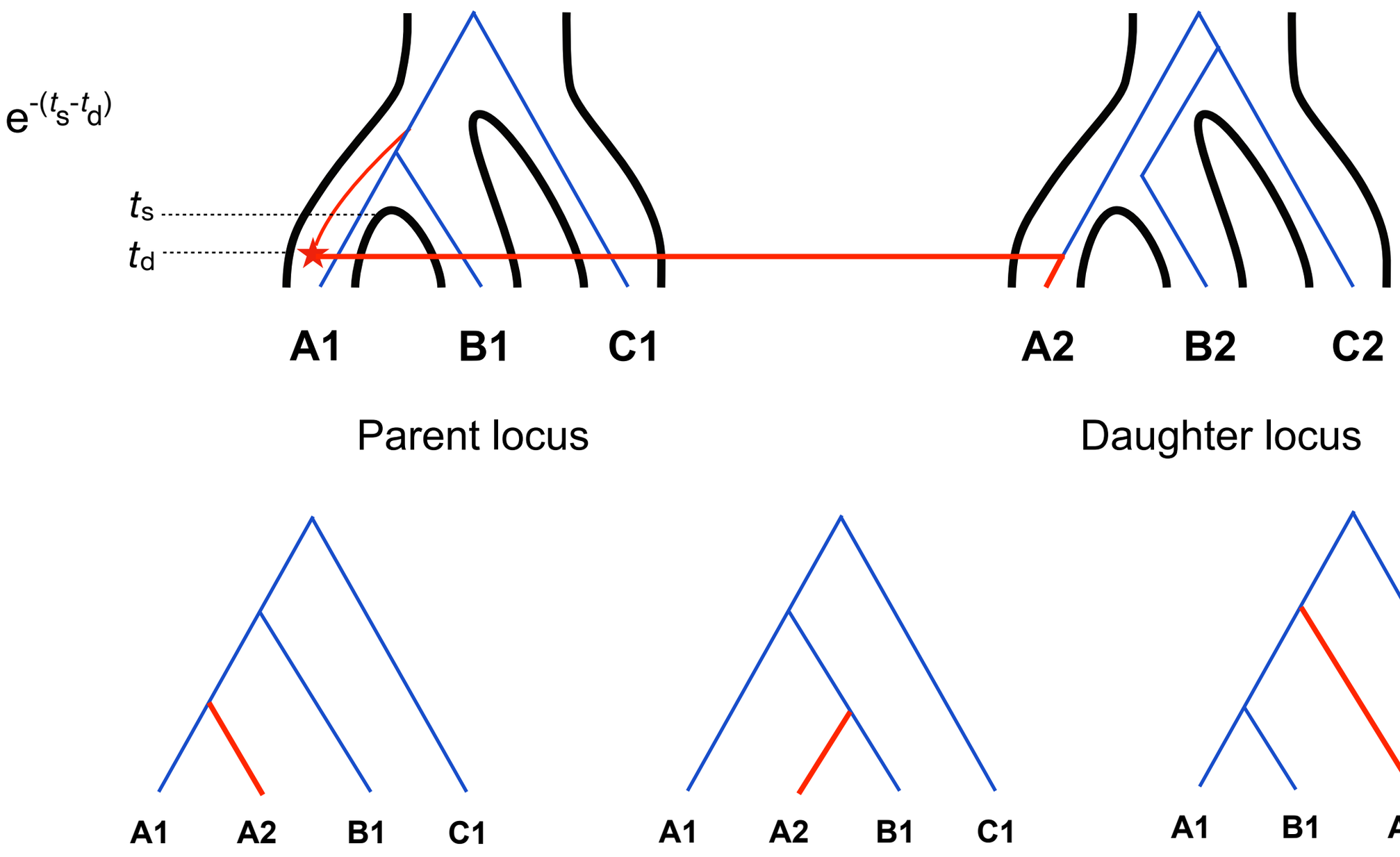


Parent locus

Daughter locus

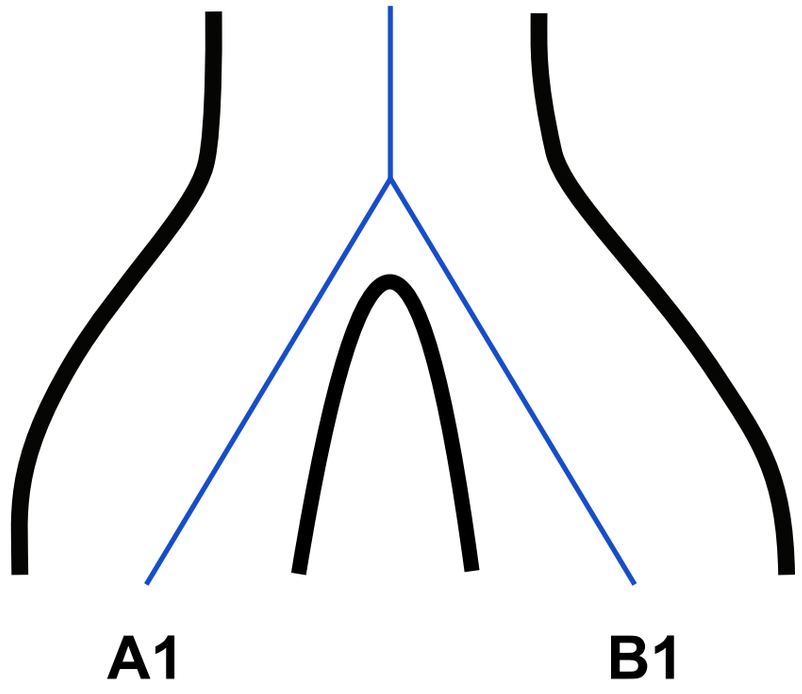


Coalescence joins parent and daughter trees

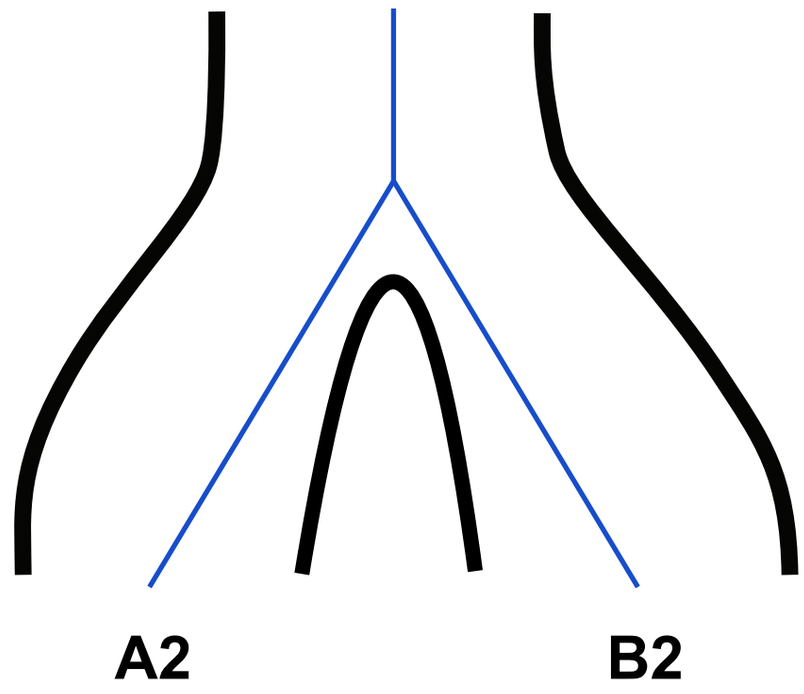


3. All parent and daughter loci must coalesce

How many trees are possible with 1 duplication?



Parent locus

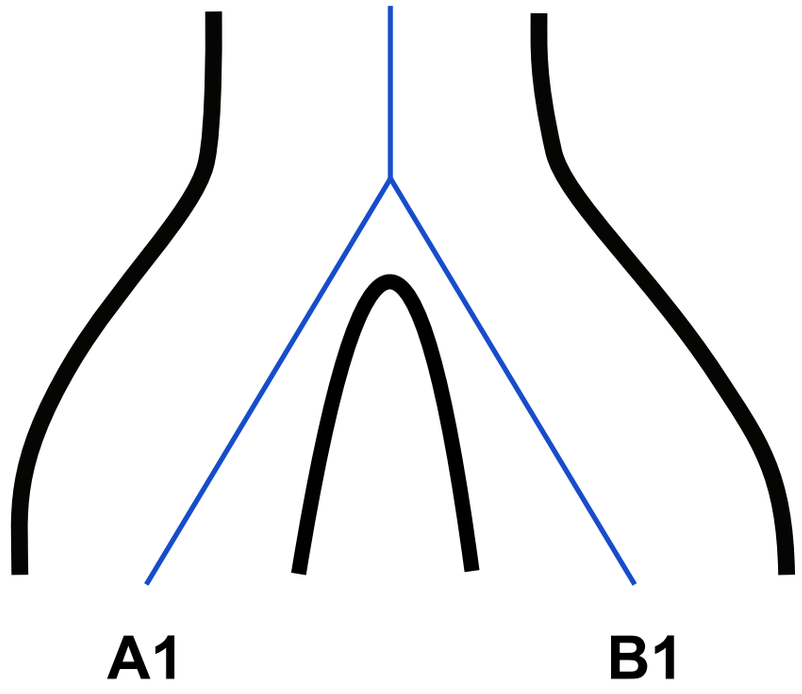


Daughter locus

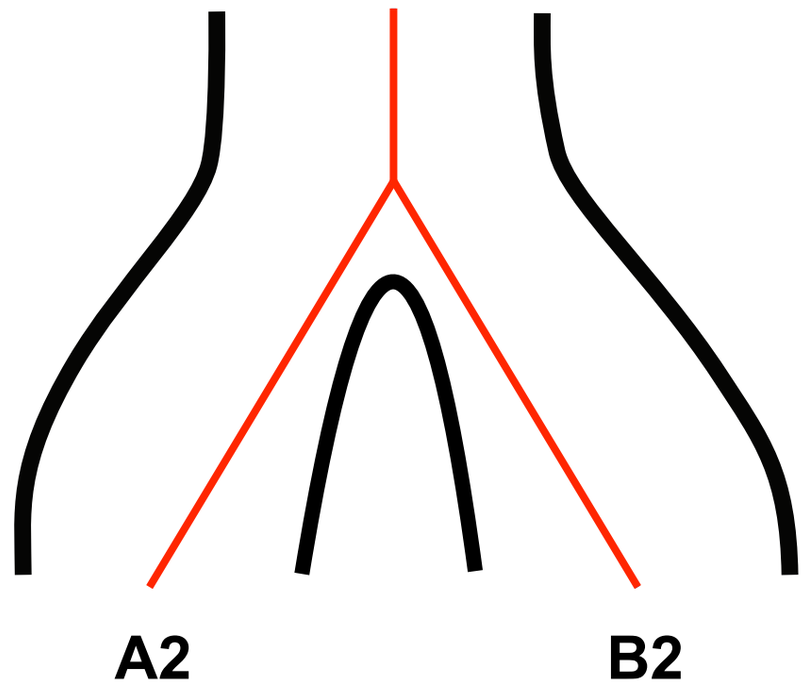
How many branches can the duplicate occur on?

How many branches can the duplicate coalesce on?

How many trees are possible with 1 duplication?



Parent locus

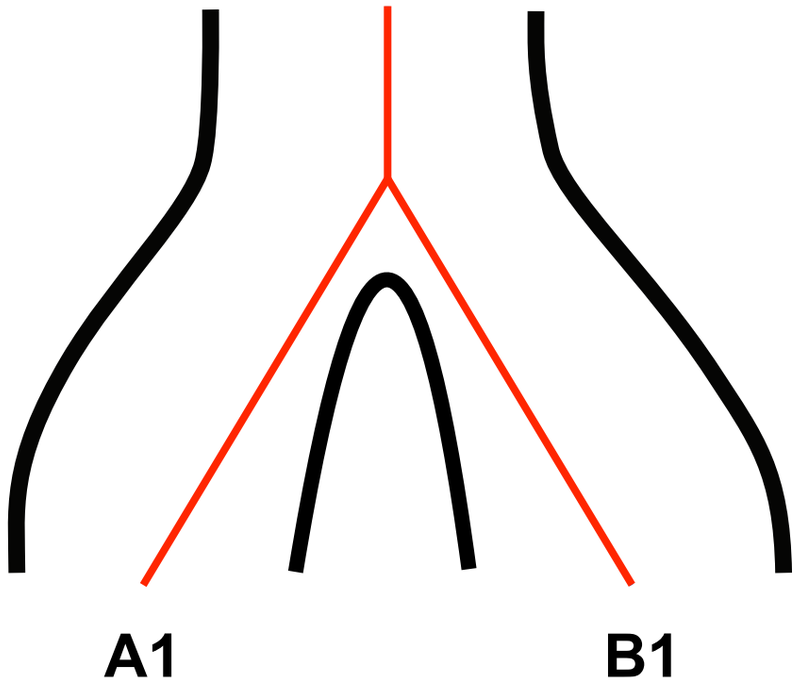


Daughter locus

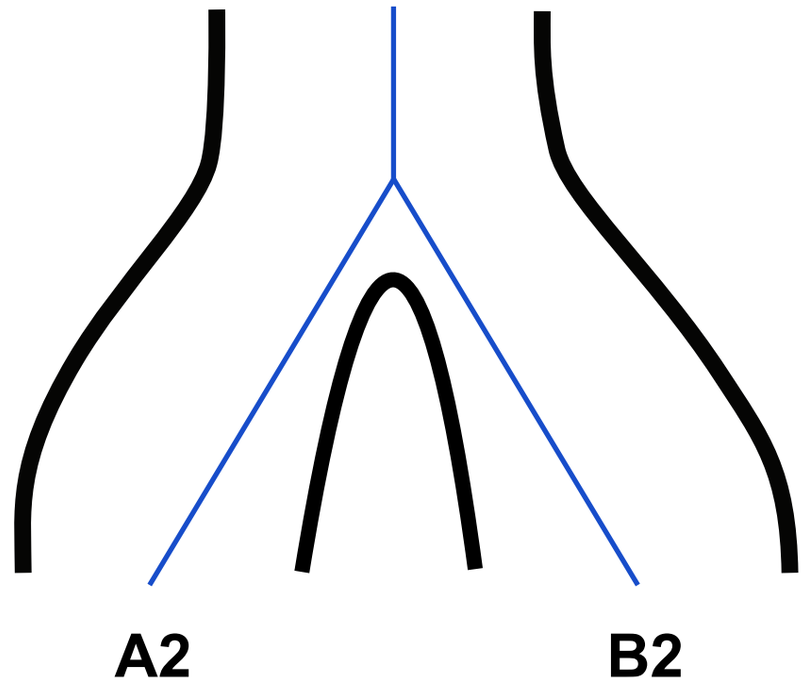
How many branches can the duplicate occur on? 3

How many branches can the duplicate coalesce on?

How many trees are possible with 1 duplication?



Parent locus

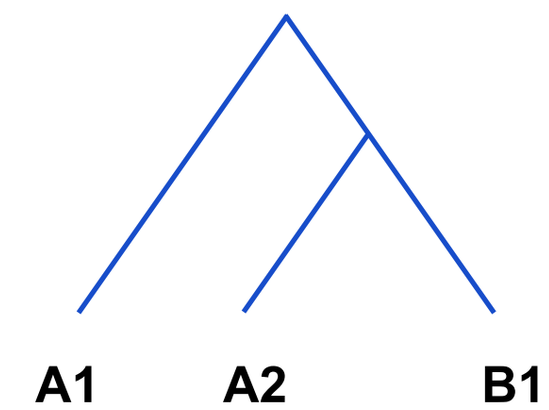


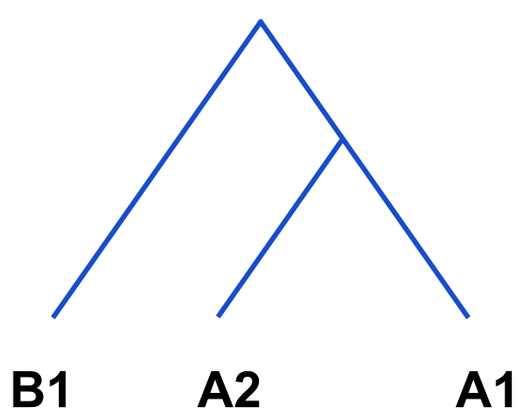
Daughter locus

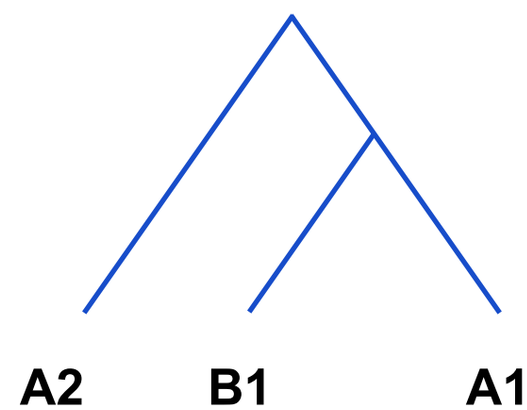
How many branches can the duplicate occur on? 3

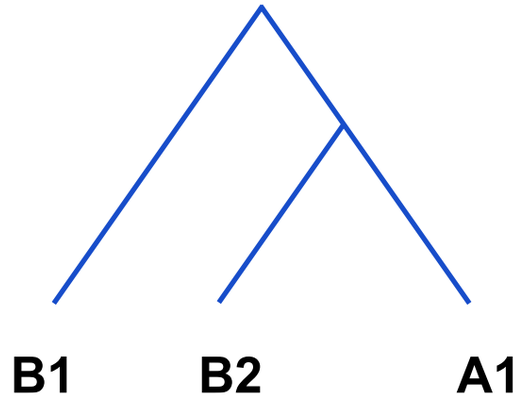
How many branches can the duplicate coalesce on? 3

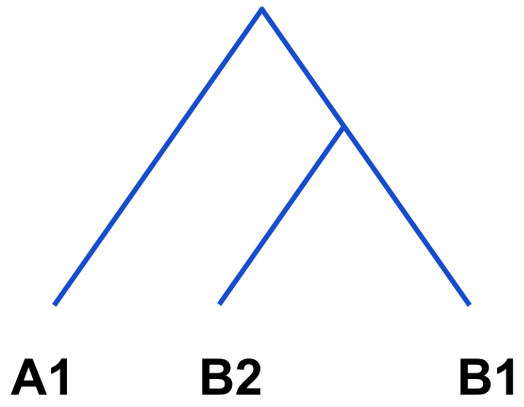
How many trees are possible with 1 duplication?

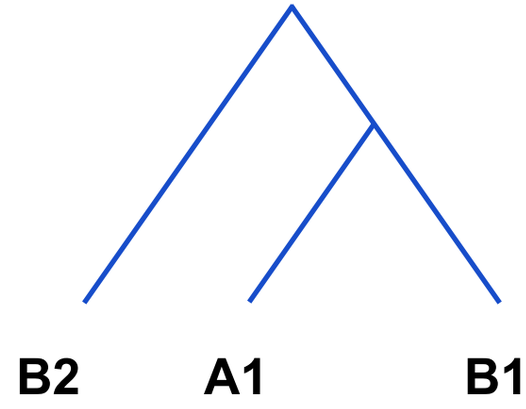


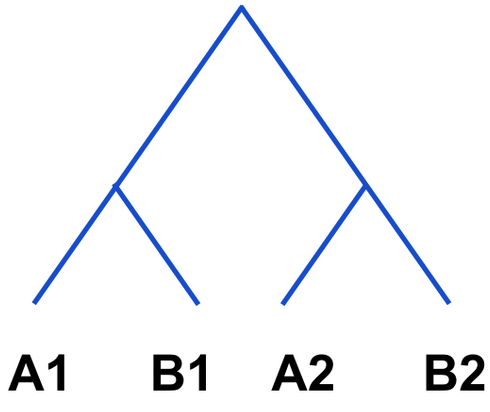


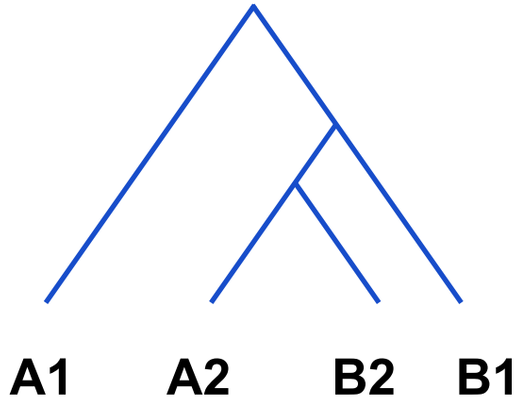


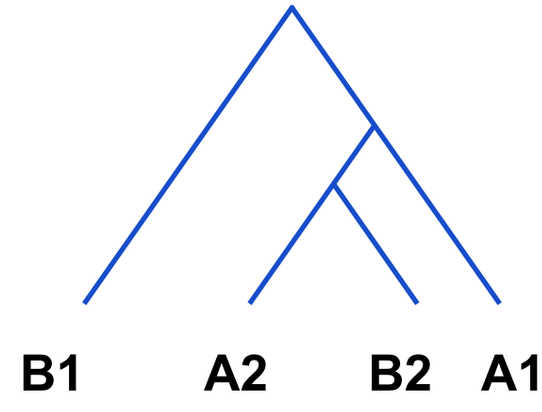










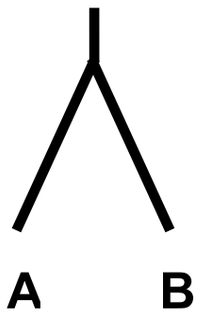


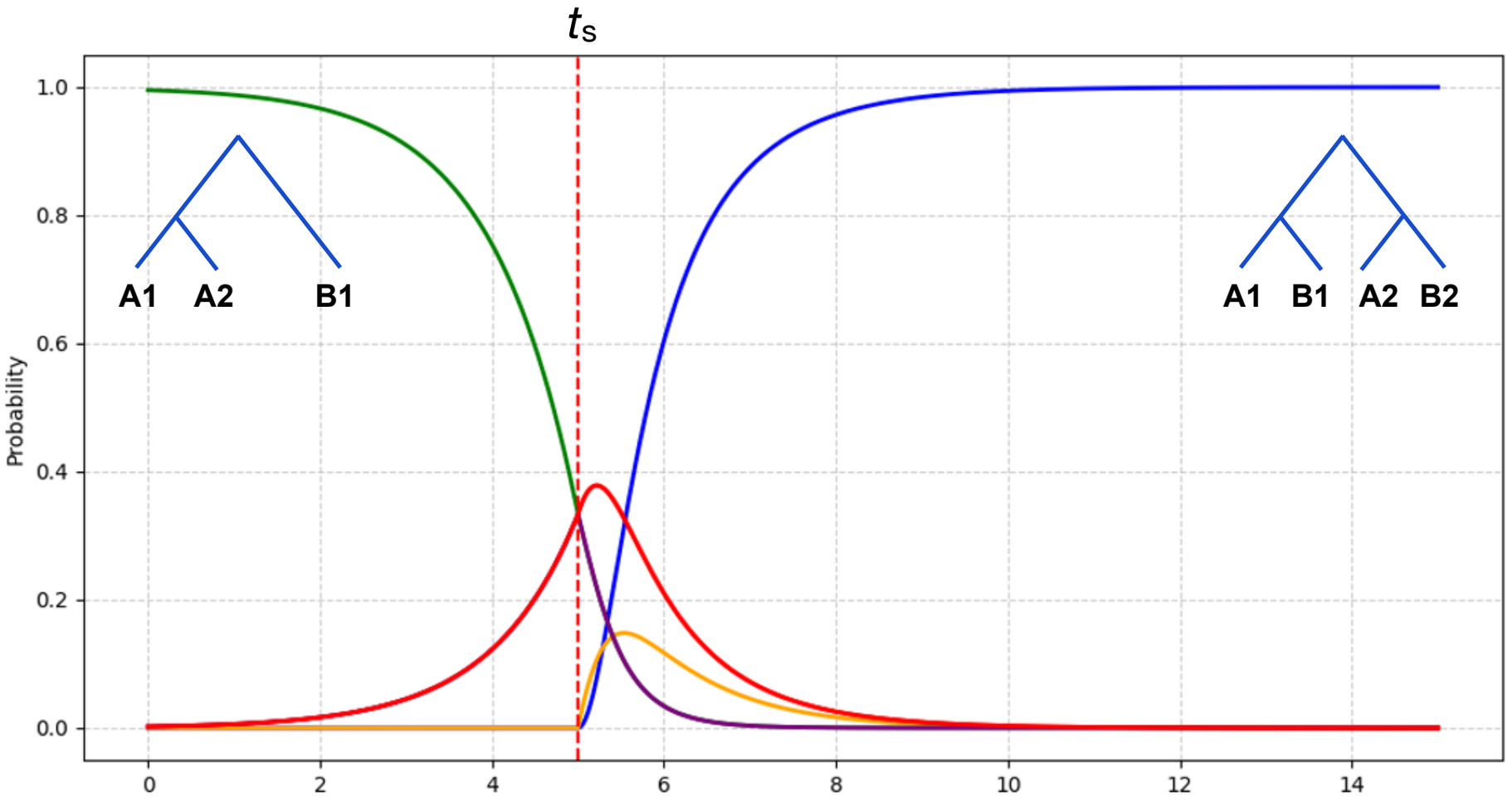
How many trees are possible with 1 duplication?

$$\underbrace{\sum_{k=1}^{N=\text{number of species}} \binom{N}{k} \times (2k-3)!!)}_{\text{Number of duplication branches}} \times \underbrace{(2N-3)!! \times (2N-1)}_{\text{Number of coalescence branches}}$$

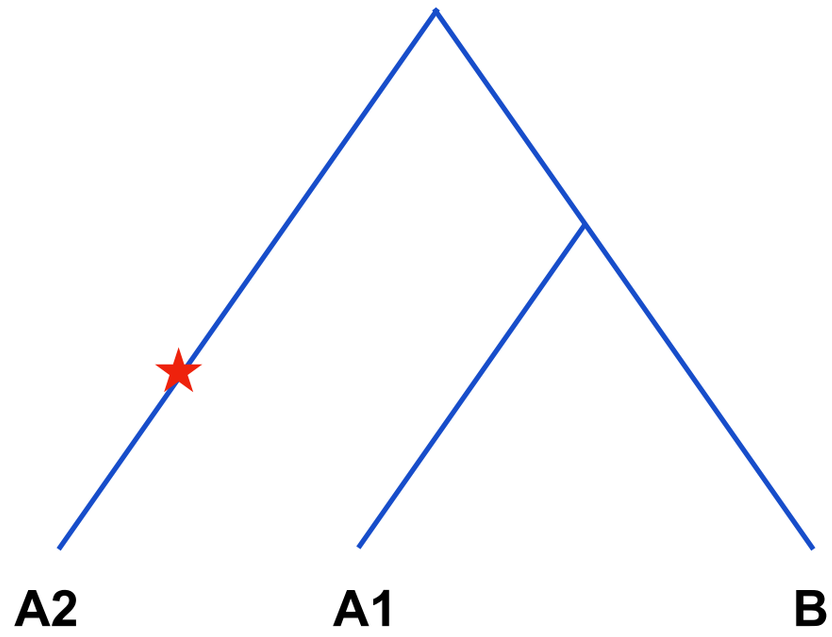
N	Possible Duplication Branches	Possible Coalescence Branches	Total Trees
2	3	3	9
3	9	15	135
4	37	105	3885
5	225	945	212625
6	1881	10395	19552995

Not all gene trees are equally probable





What is the history of this gene tree?



How many duplications?

How many losses?

When did duplication/loss occur?

Which genes are orthologs/paralogs?

reconcILS

reconcILS: A gene tree-species tree reconciliation algorithm that allows for incomplete lineage sorting

Sarthak Mishra, Megan L. Smith, Matthew W. Hahn

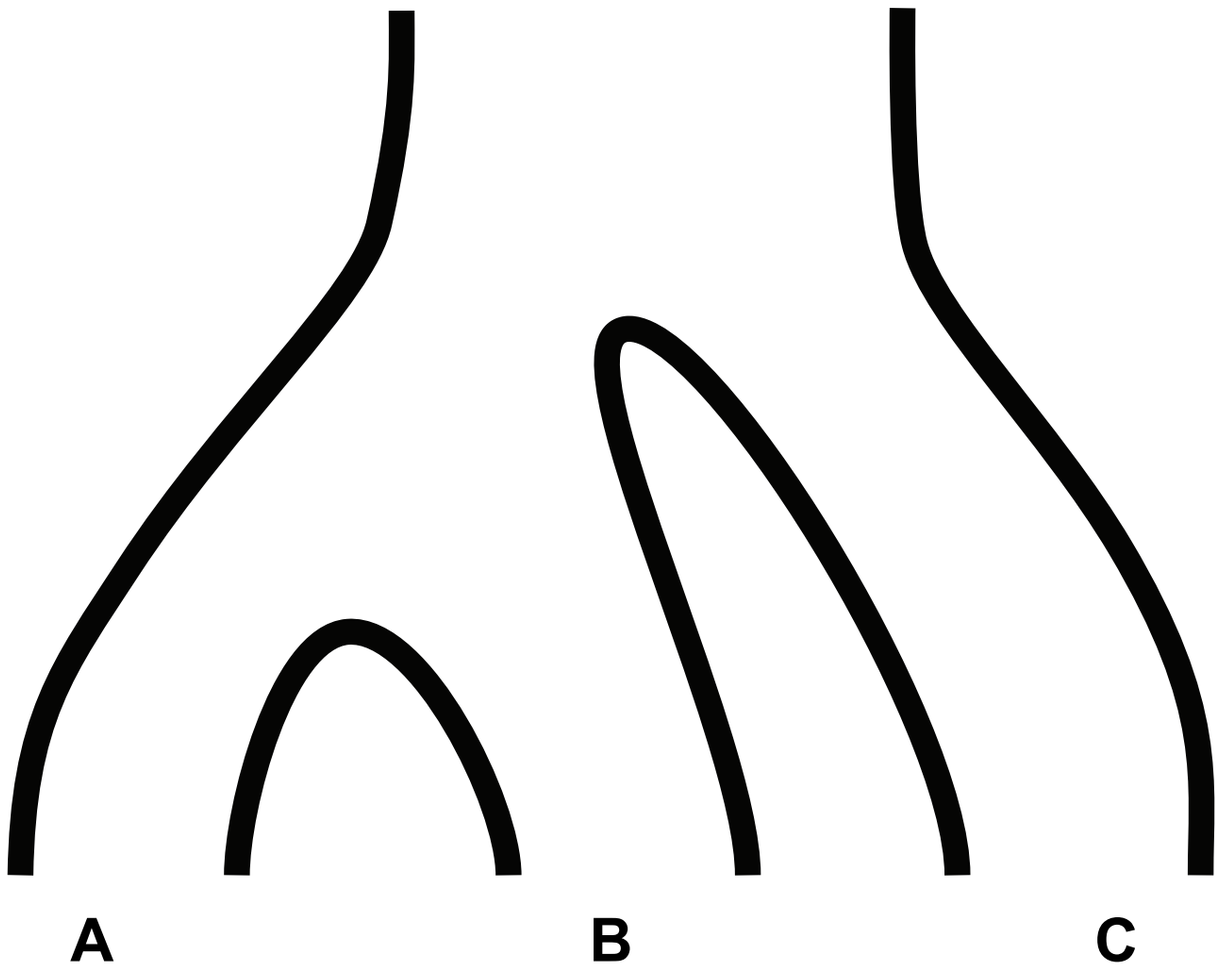
doi: https://doi.org/10.1101/2023.11.03.565544

Adding introgression

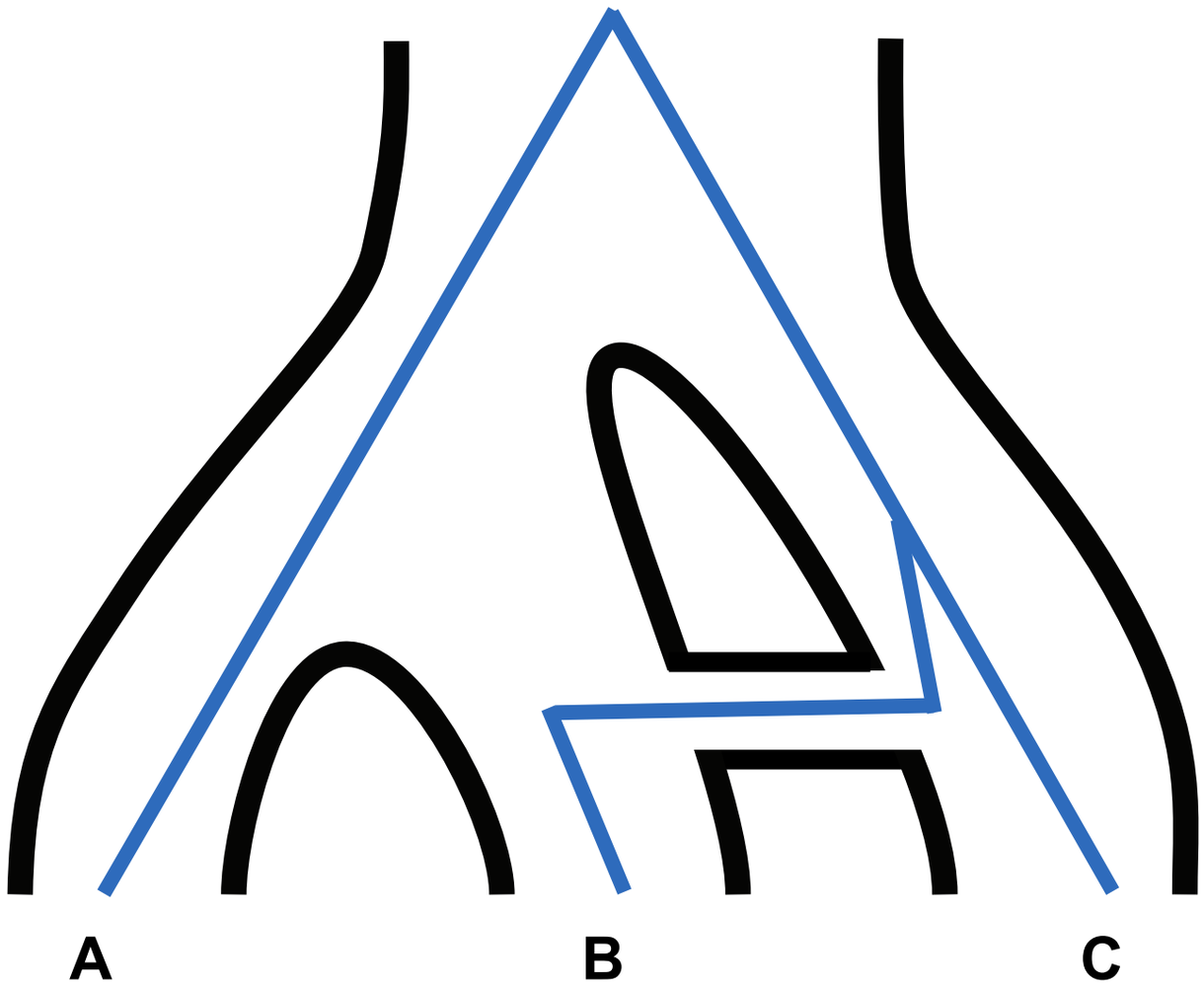
Sarthak Mishra

Rob Lanfear

Introgression

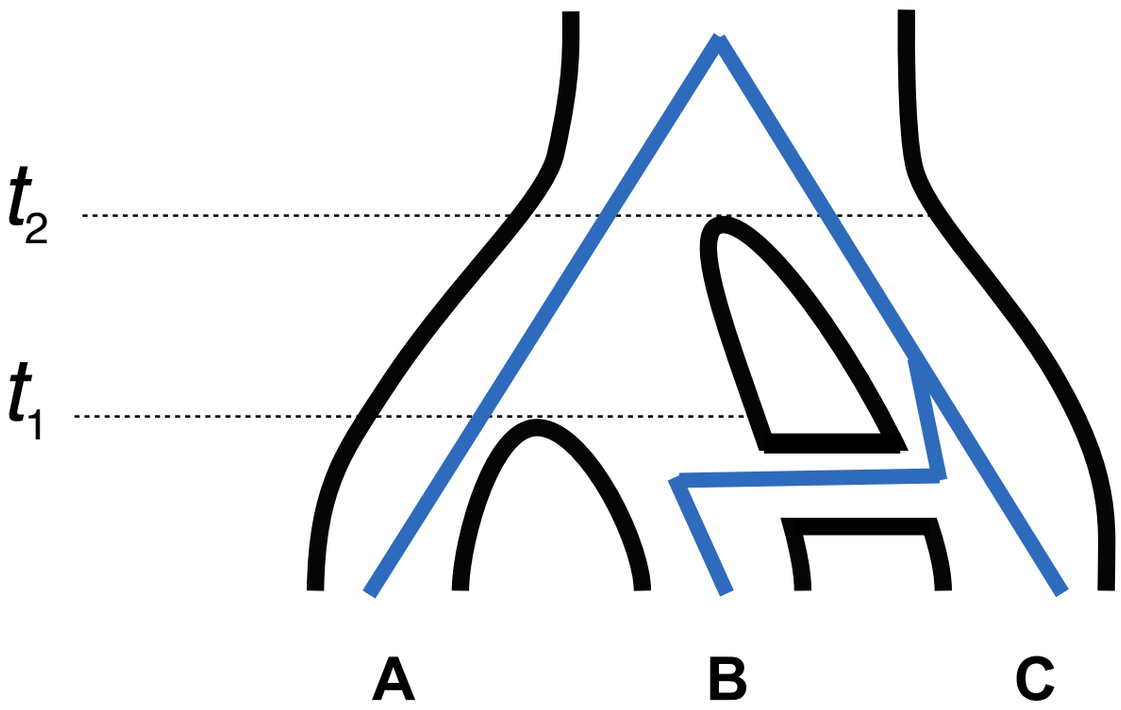


The multispecies network coalescent (MSNC)



Yu et al. (2011)

Triplet asymmetry to detect introgression



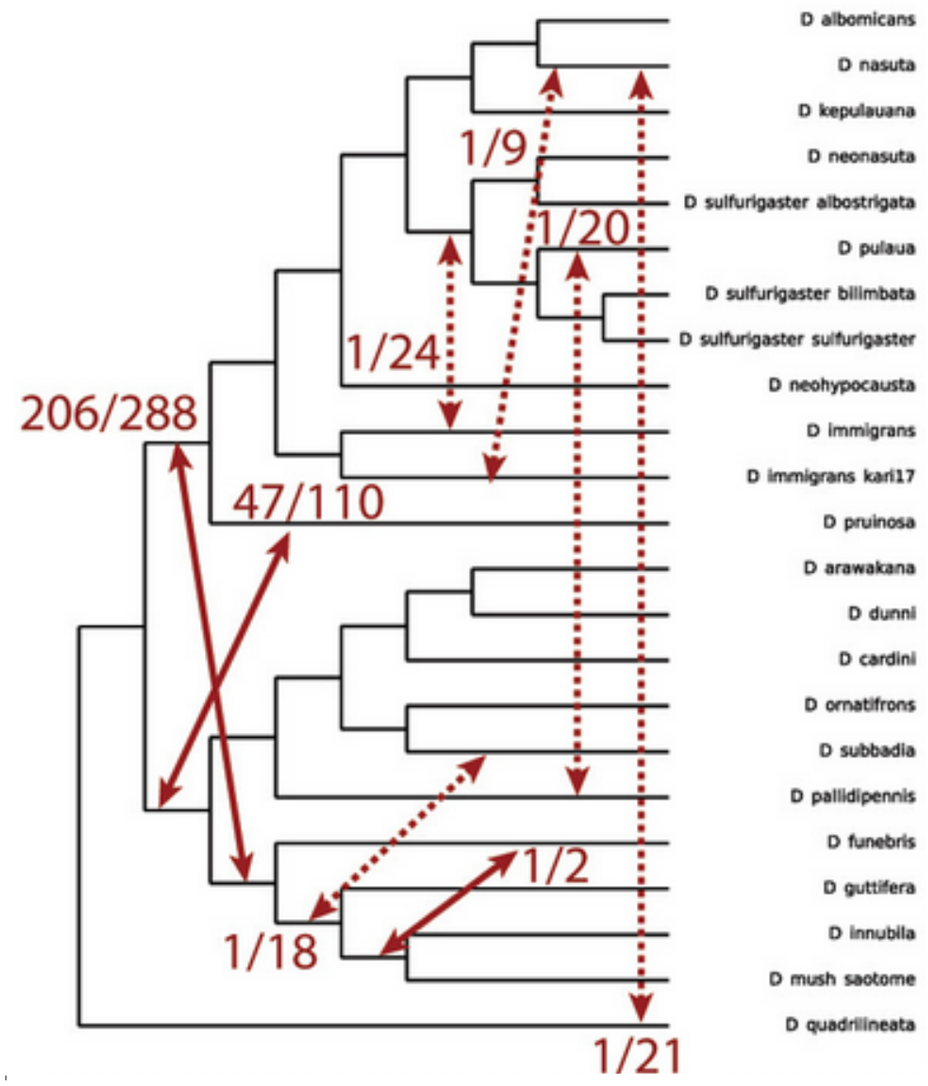
$1-(2/3)e^{-(t_2-t_1)}$

$(1/3)e^{-(t_2-t_1)}$

$(1/3)e^{-(t_2-t_1)}$

Huson et al. (2005); Green et al. (2010)

The problem with triplets



Suvorov et al. (2022)

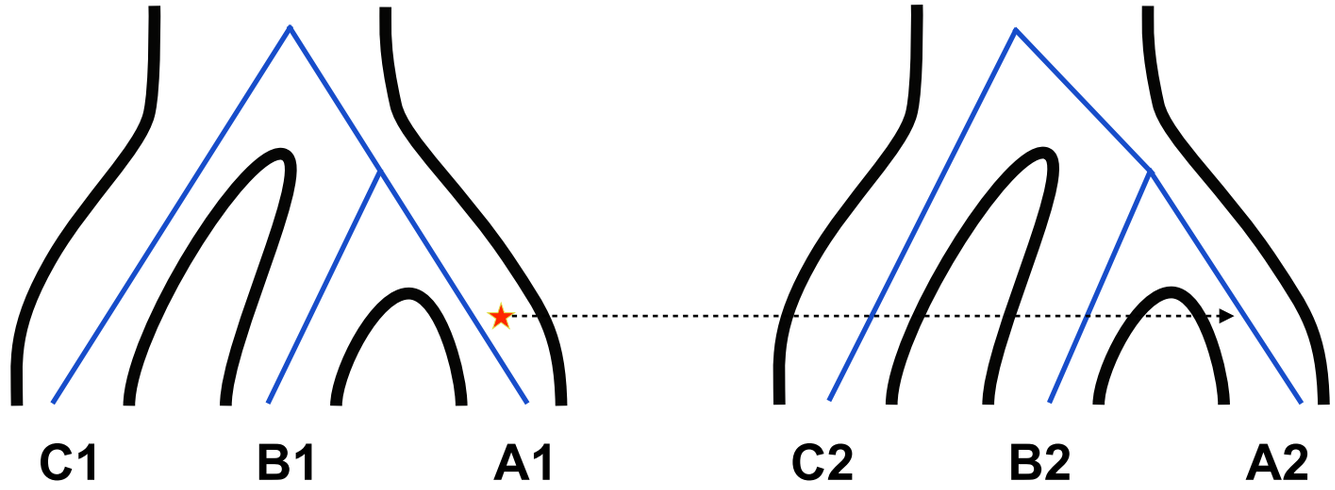
What we would like to know

Who is involved in introgression?

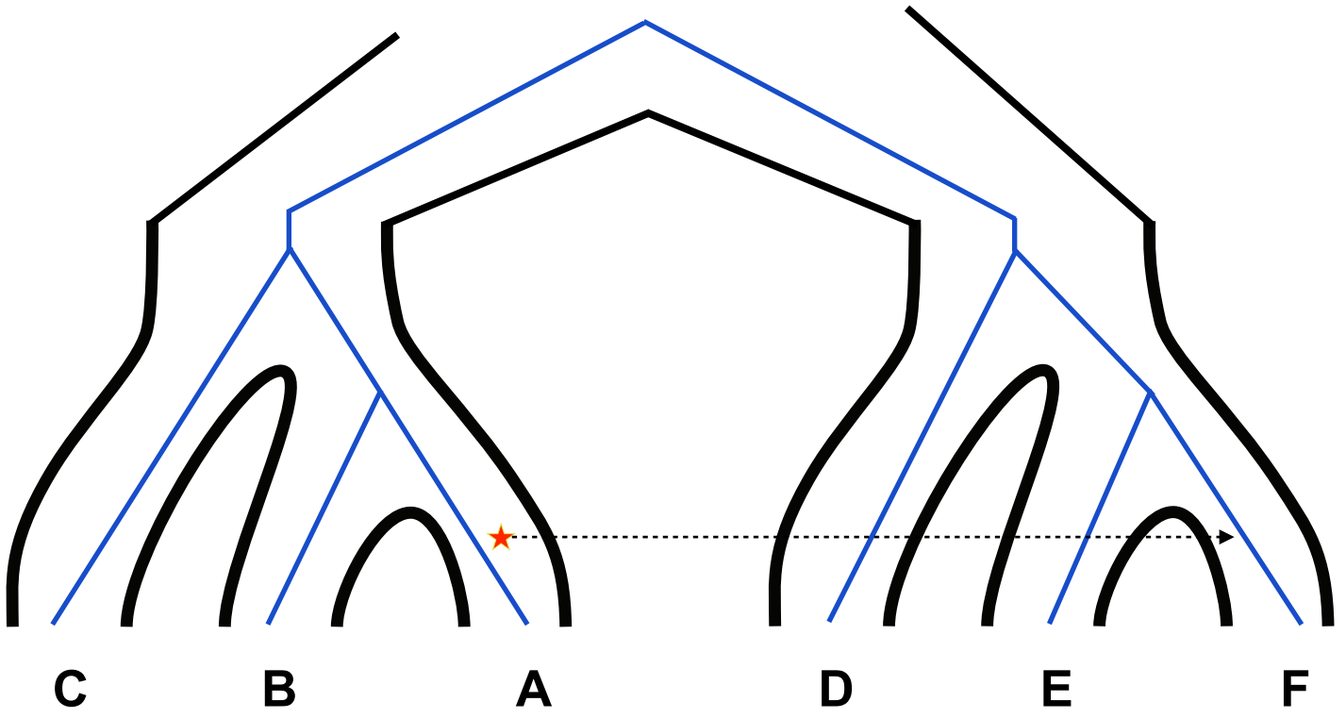
What is the direction of introgression?

How many introgression events?

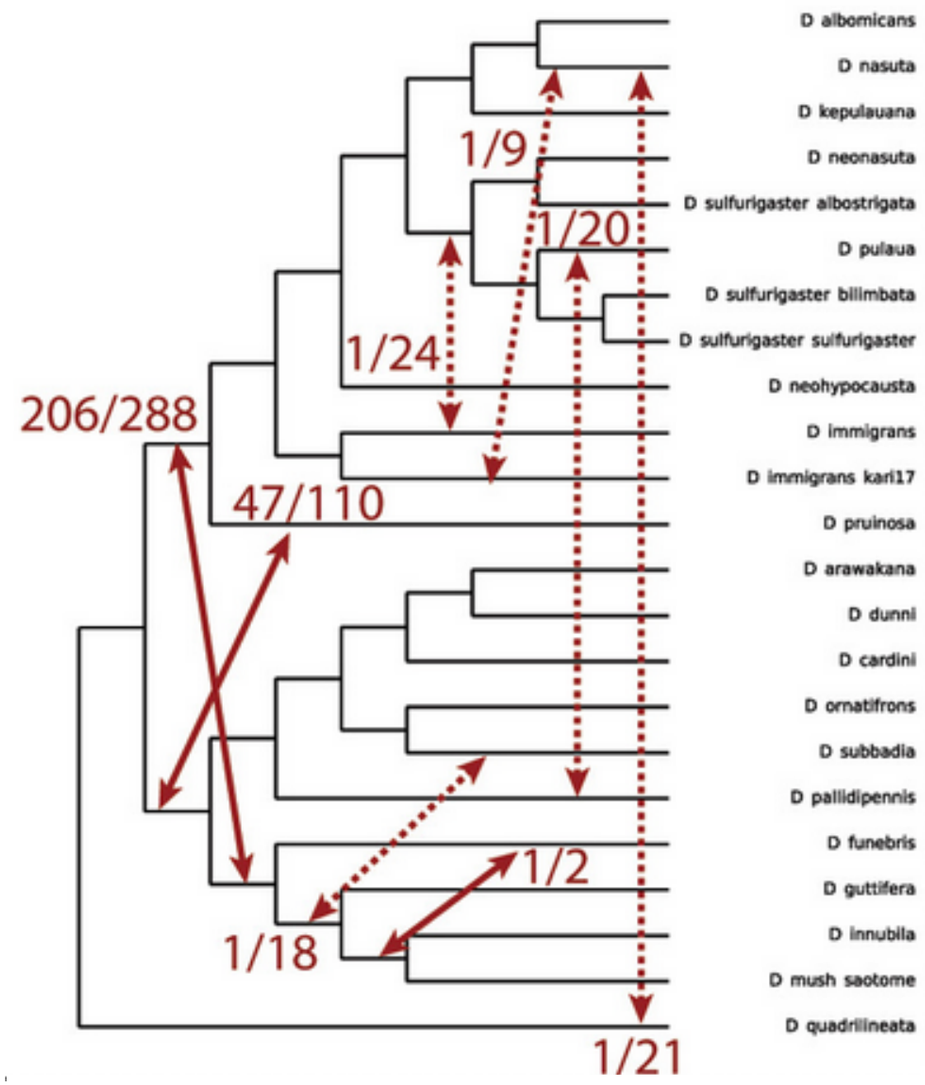
The effects of introgression on gene trees



The effects of introgression on gene trees

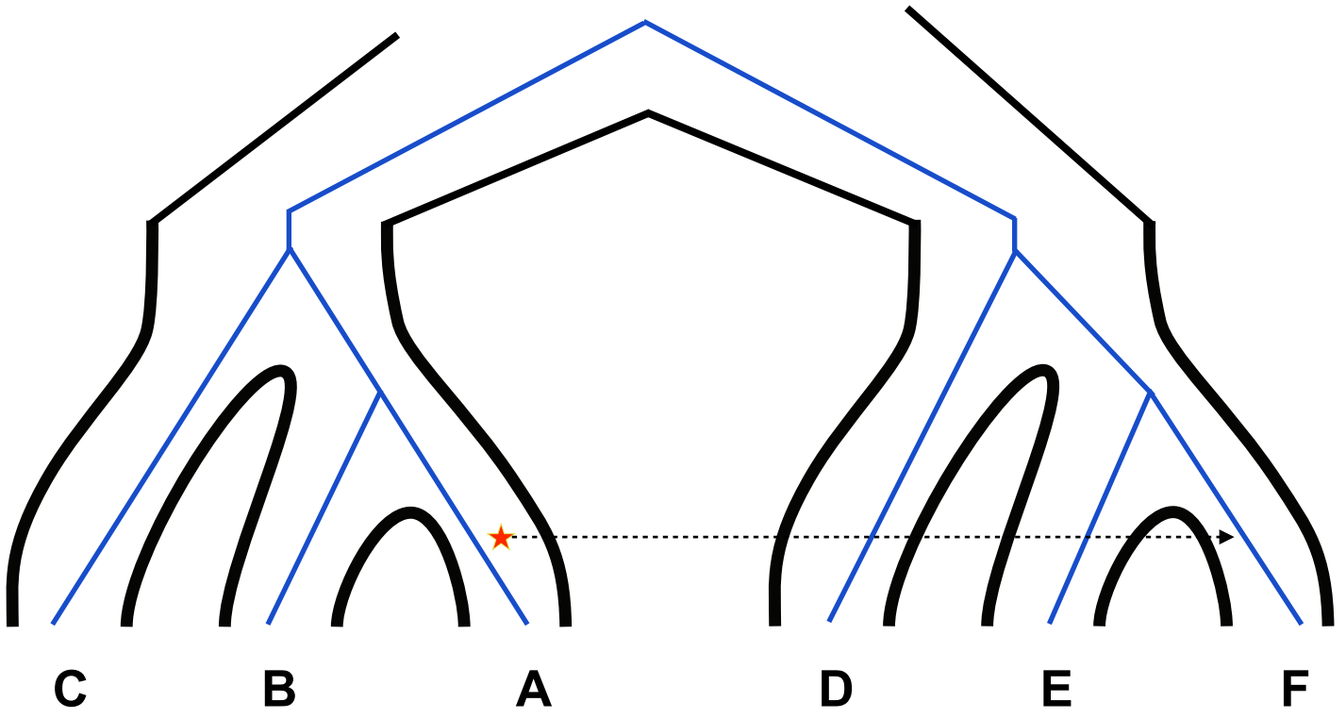


The effects of introgression on gene trees

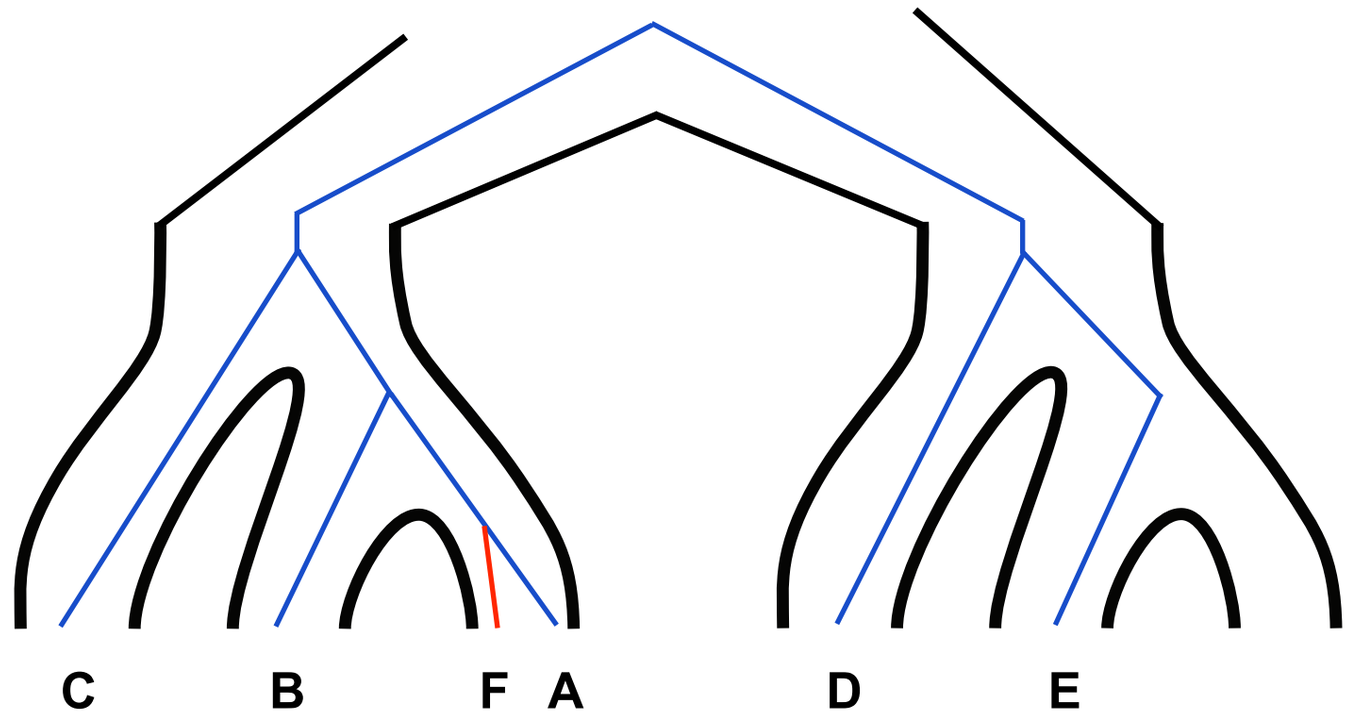


Suvorov et al. (2022)

The effects of introgression on gene trees

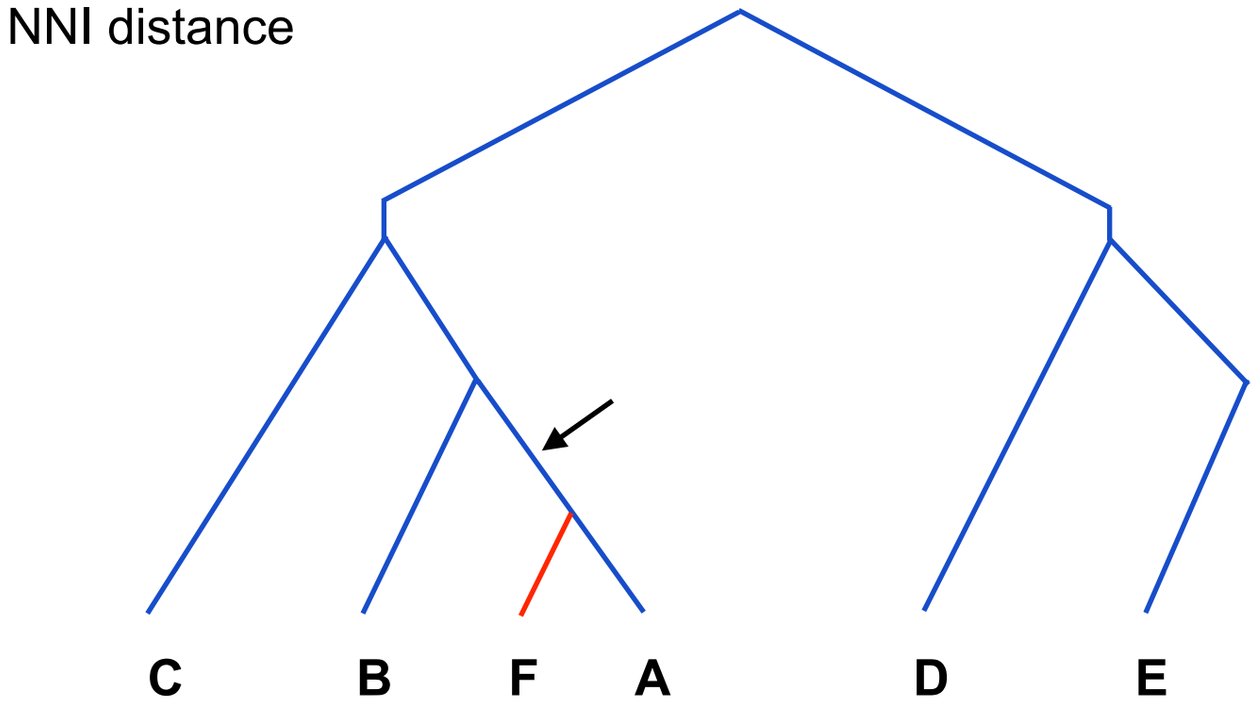


The effects of introgression on gene trees

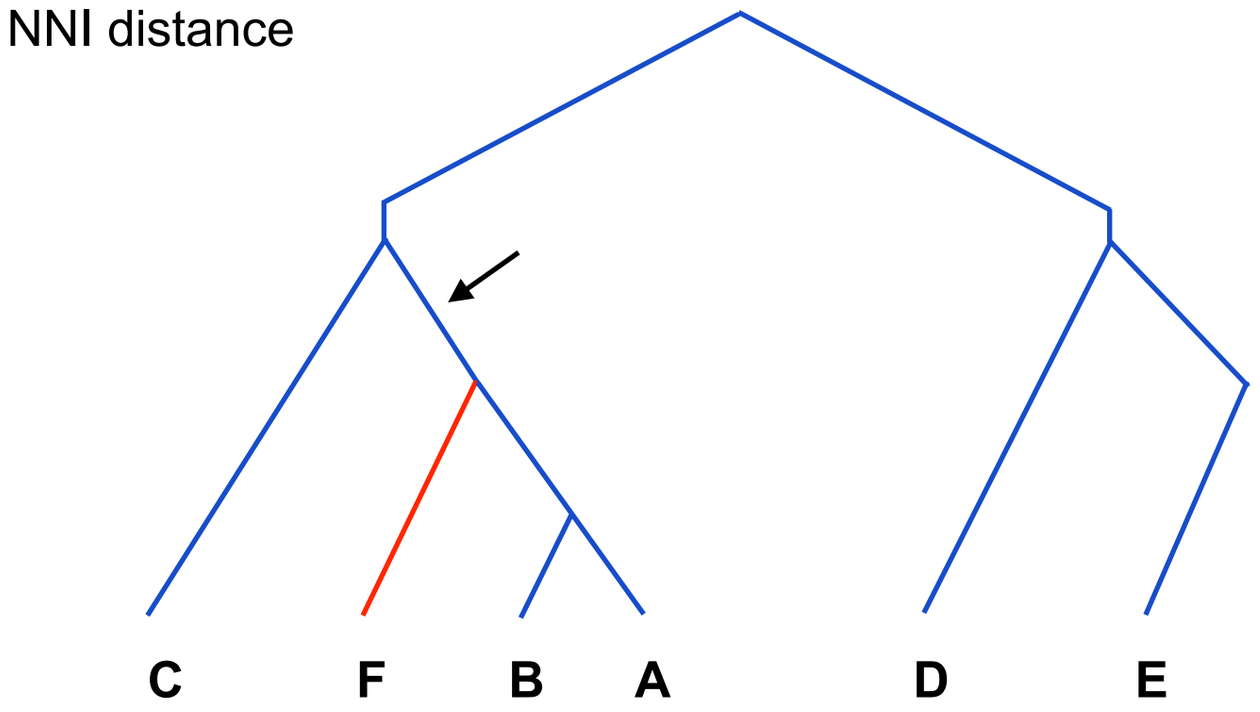


“attachment”

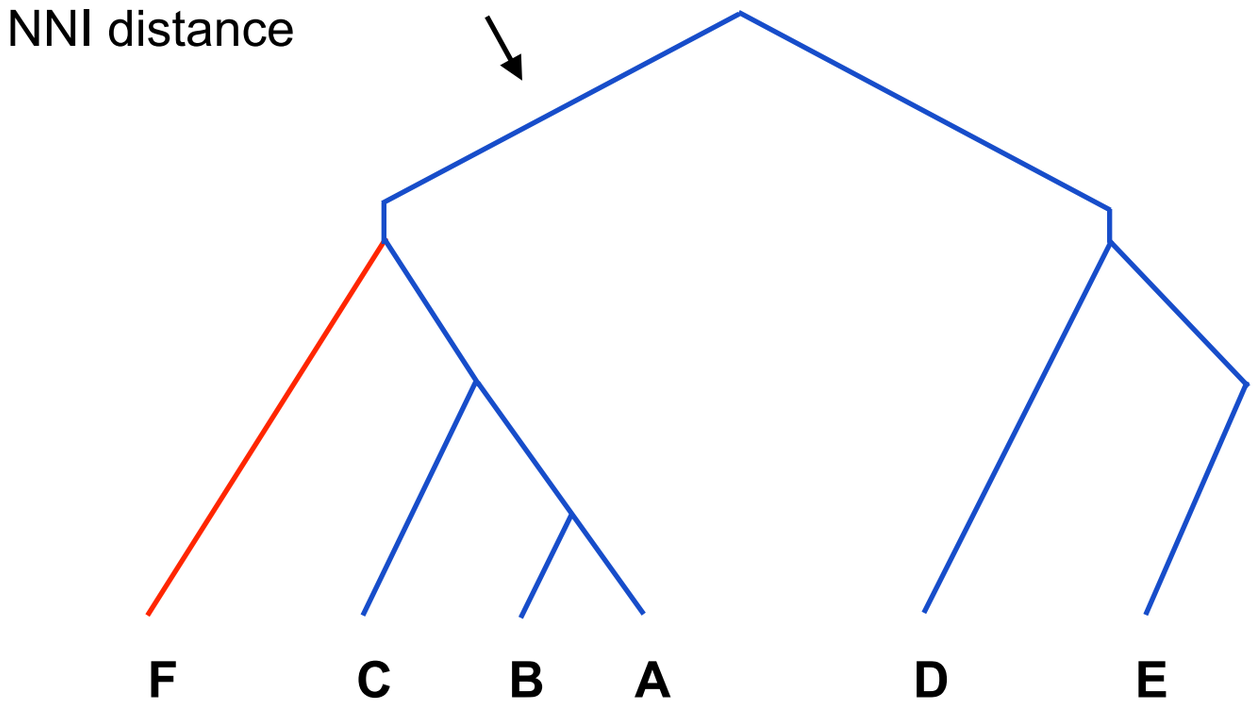
How far apart are the attached lineages?



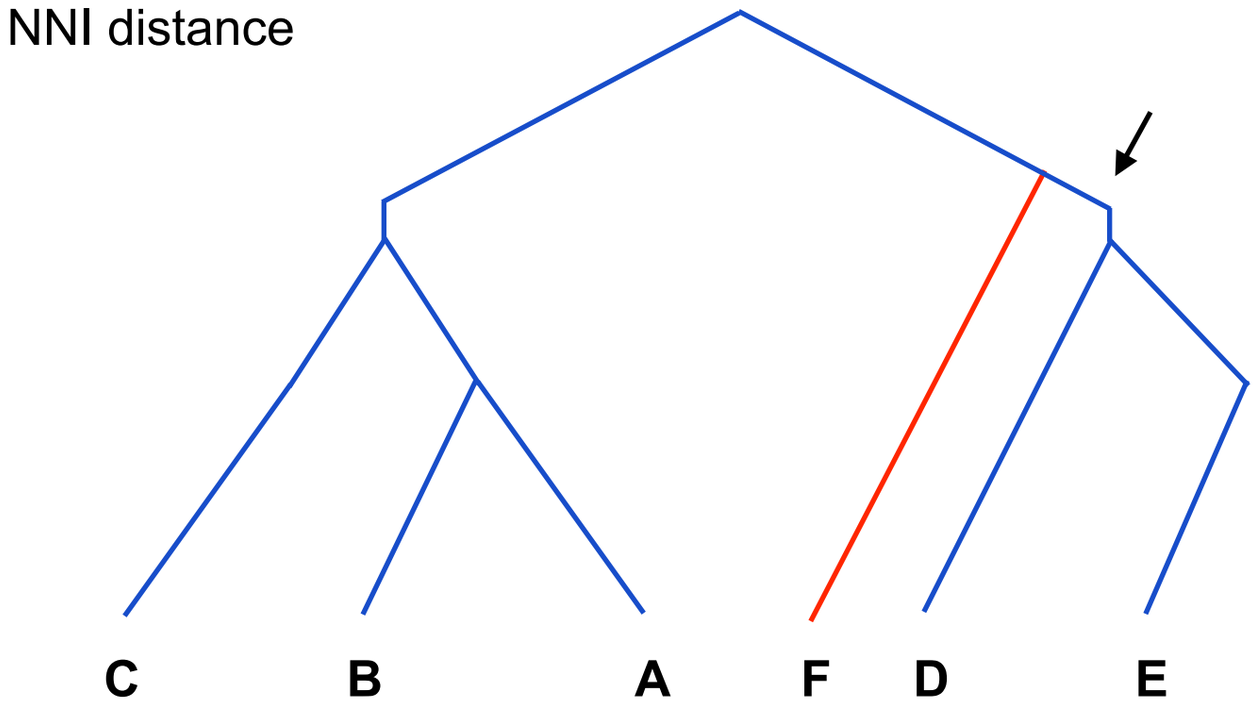
How far apart are the attached lineages?



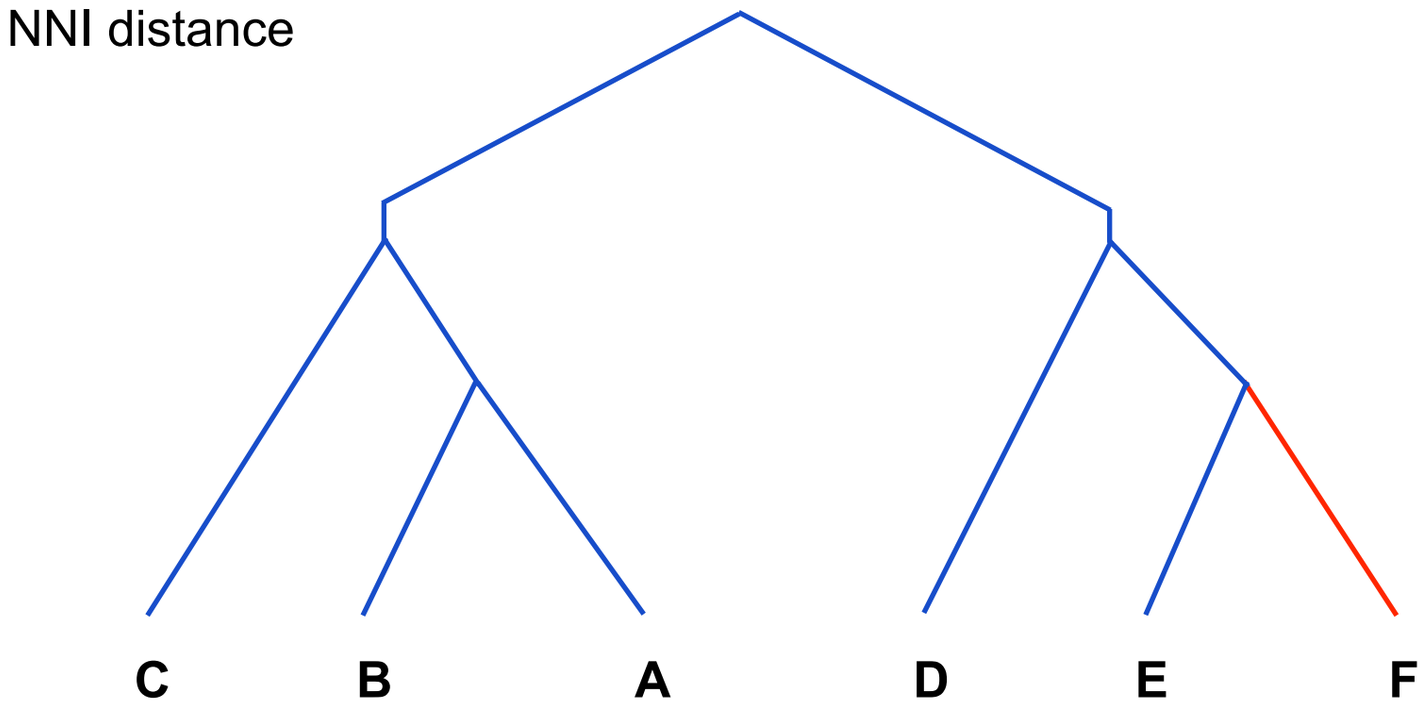
How far apart are the attached lineages?



How far apart are the attached lineages?



How far apart are the attached lineages?



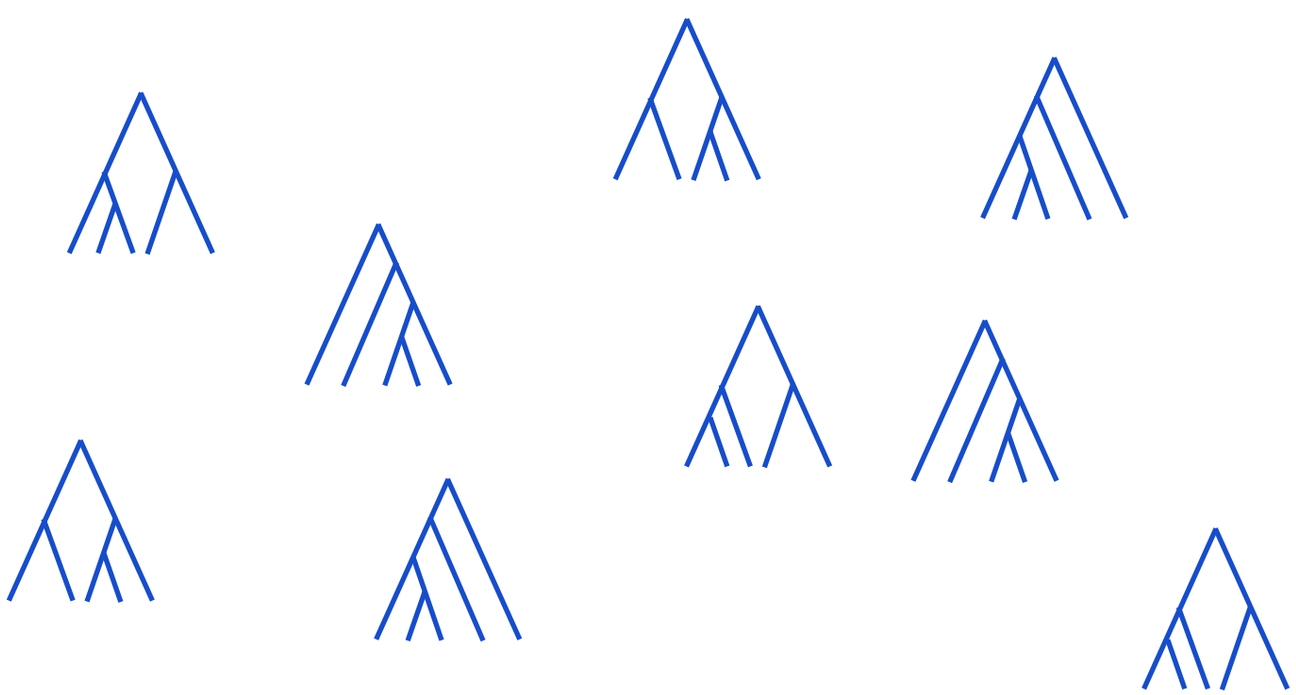
Attachment Distance = 4 NNIs

DAFT

(Discordant attachment frequency test)

Counting attachment frequencies

Input: all single-copy gene trees and a species tree

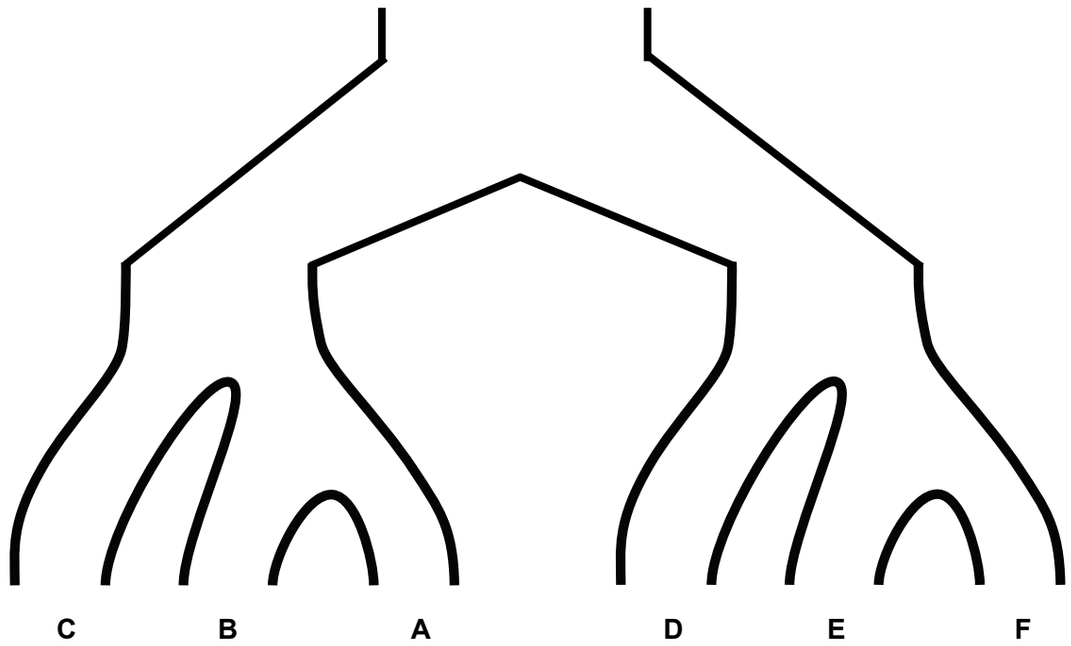


Counting attachment frequencies

Output: attachments at all branches of the species tree

Attached to branch E:

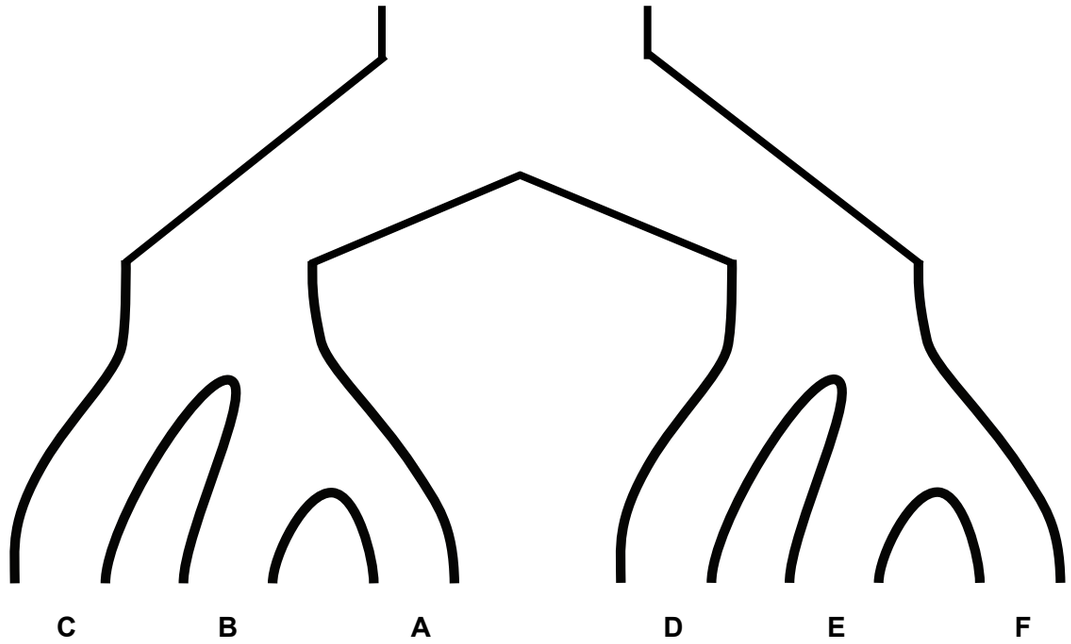
F
D
C(B,A)
C
(B,A)
B
A



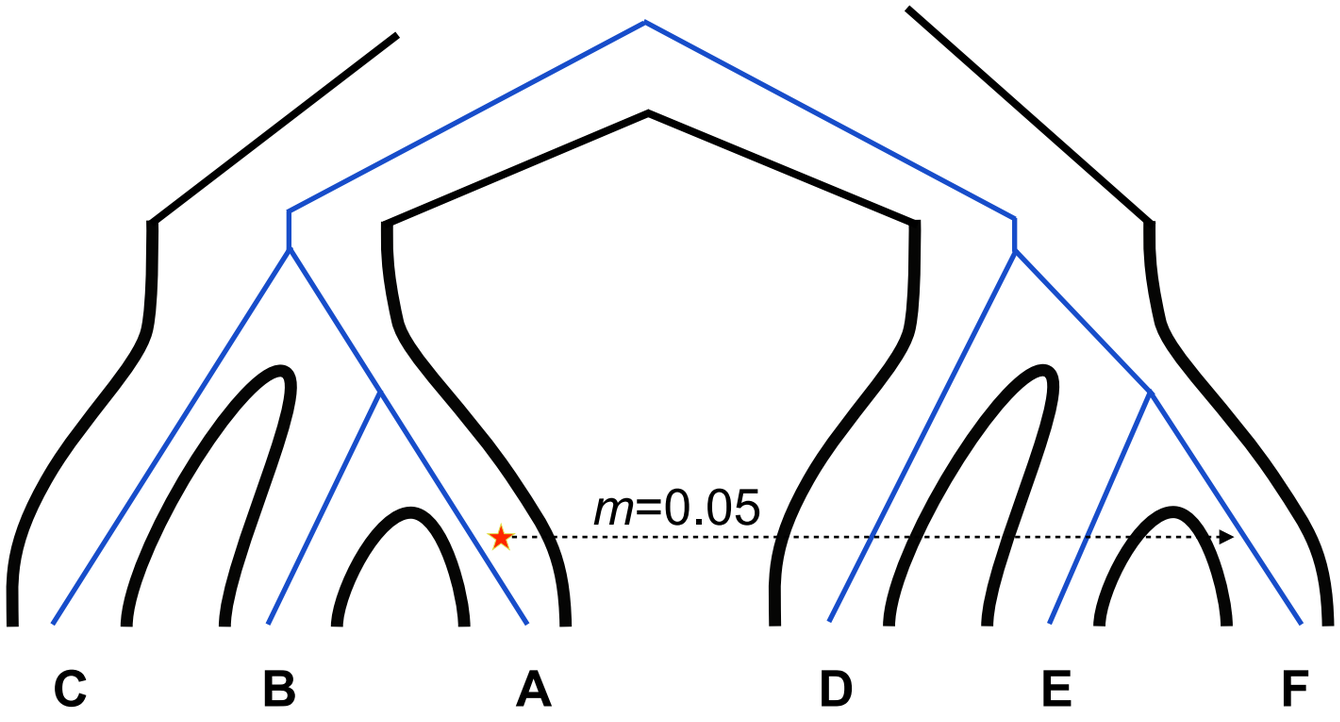
Counting attachment frequencies

Output: attachments at all branches of the species tree

NNI distance	**Attached to branch E:**	
0	**F**	565
1	**D**	311
2	**C(B,A)**	99
3	**C**	68
3	**(B,A)**	59
4	**B**	12
4	**A**	14



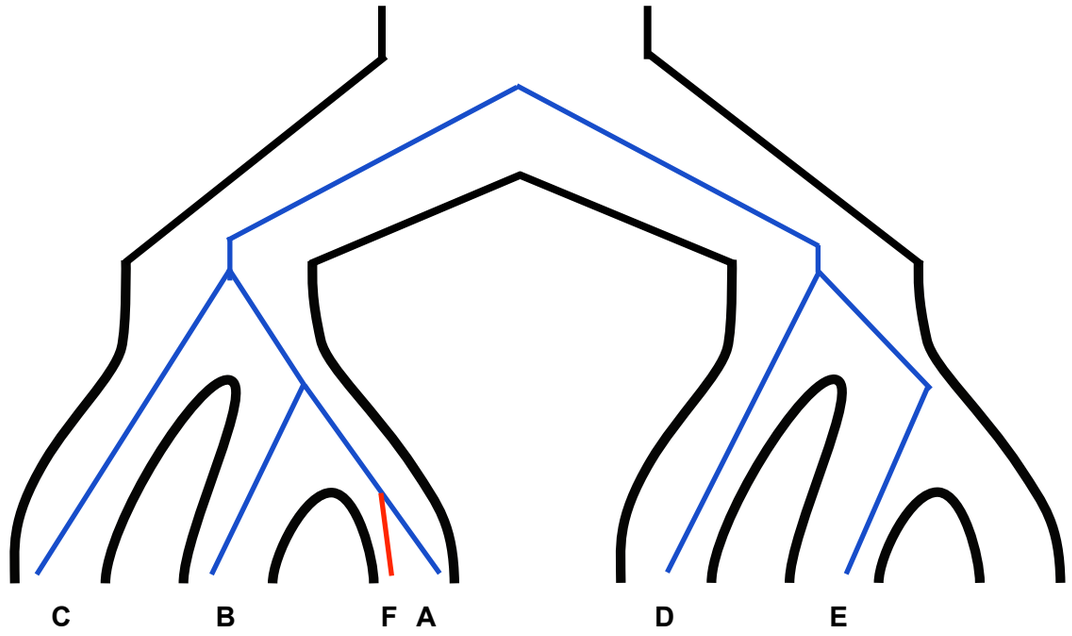
The effects of introgression on gene trees



Counting attachment frequencies

NNI distance **Attached to branch A:**

NNI distance		Attached to branch A:	
0		**B**	501
1		**C**	299
2		**D(E,F)**	45
3	|	**D**	16
3	|	**(E,F)**	13
4	|	**E**	2
4	|	**F**	50 ←

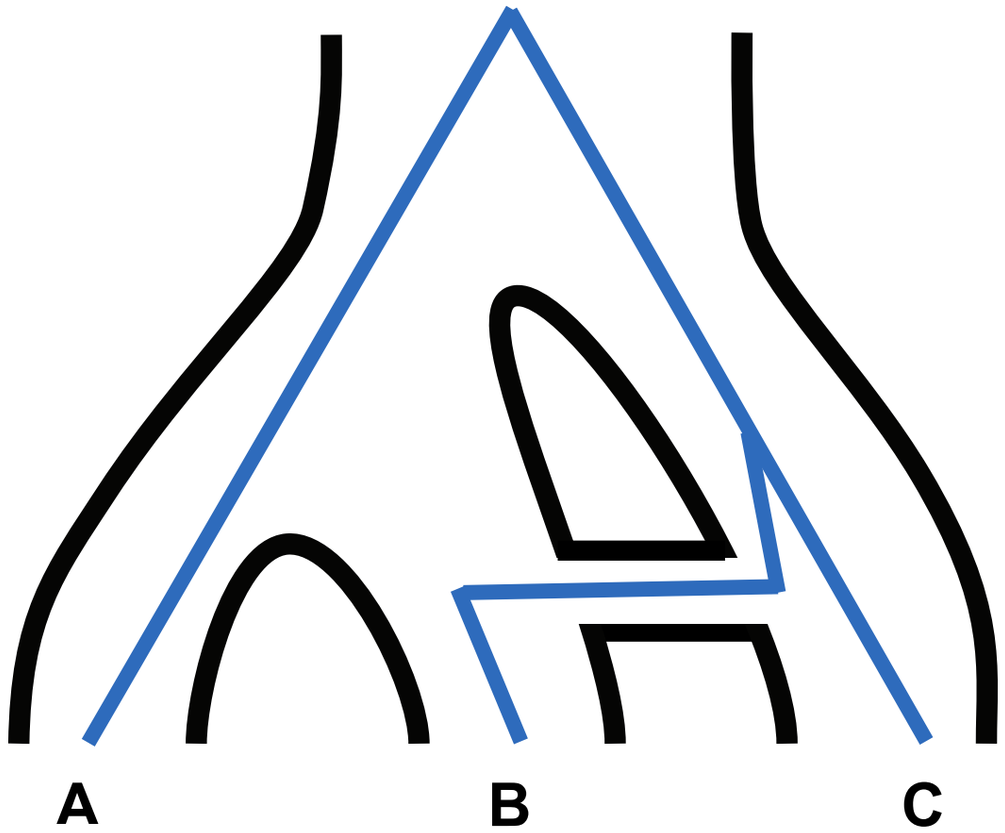


Calculate a *z*-test for significance

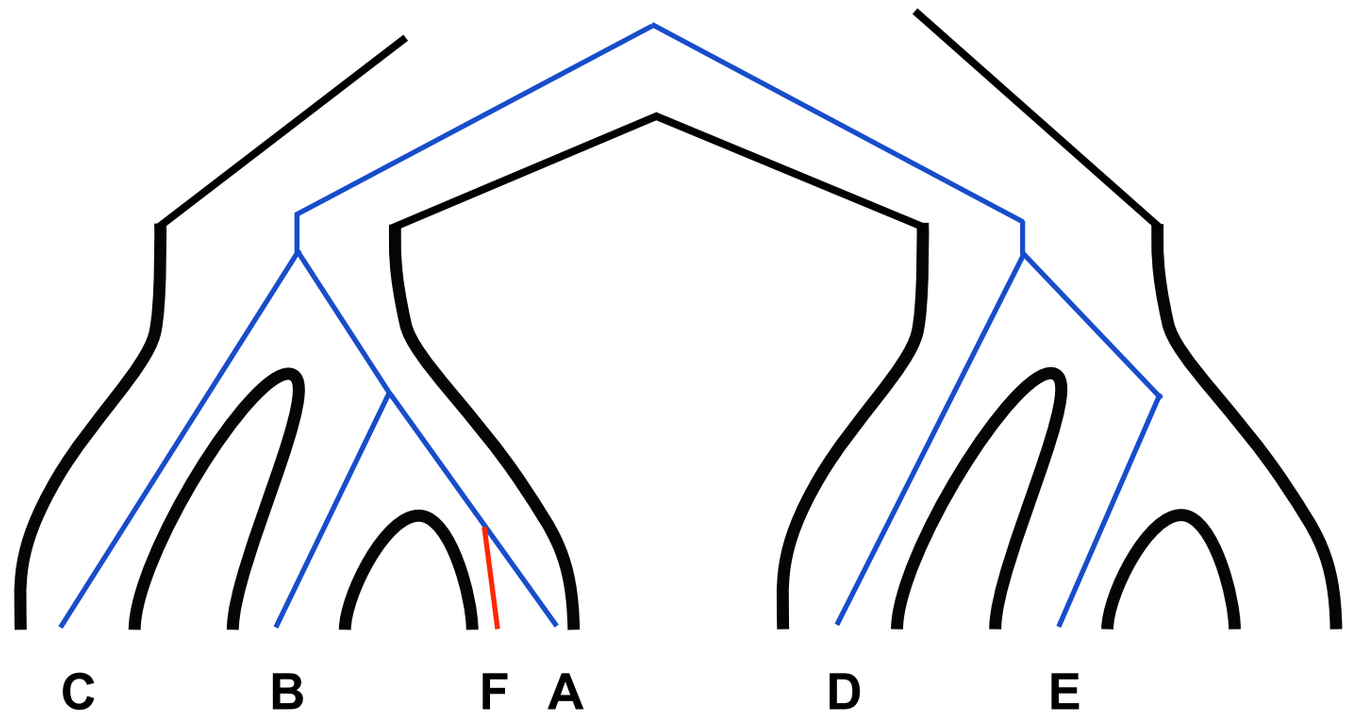
What we would like to know

What is the direction of introgression?

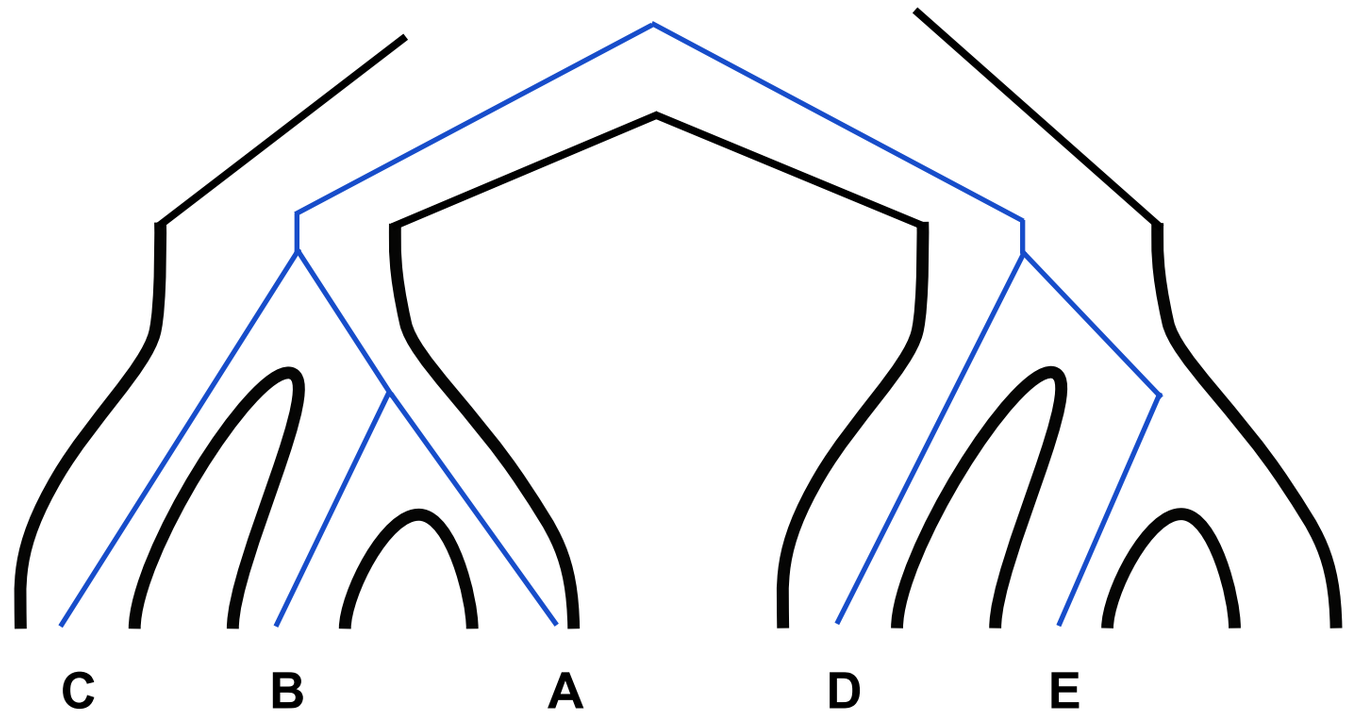
What is the direction of introgression?



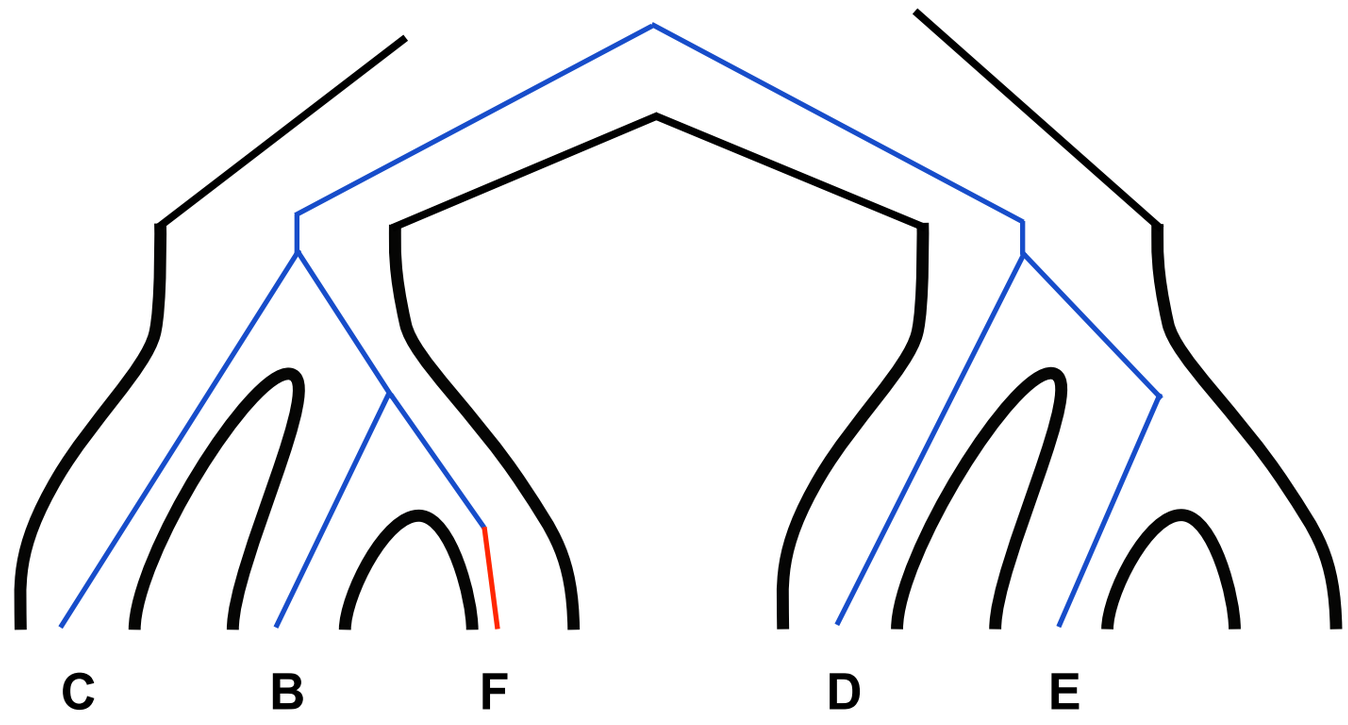
Who moved?



Who moved?

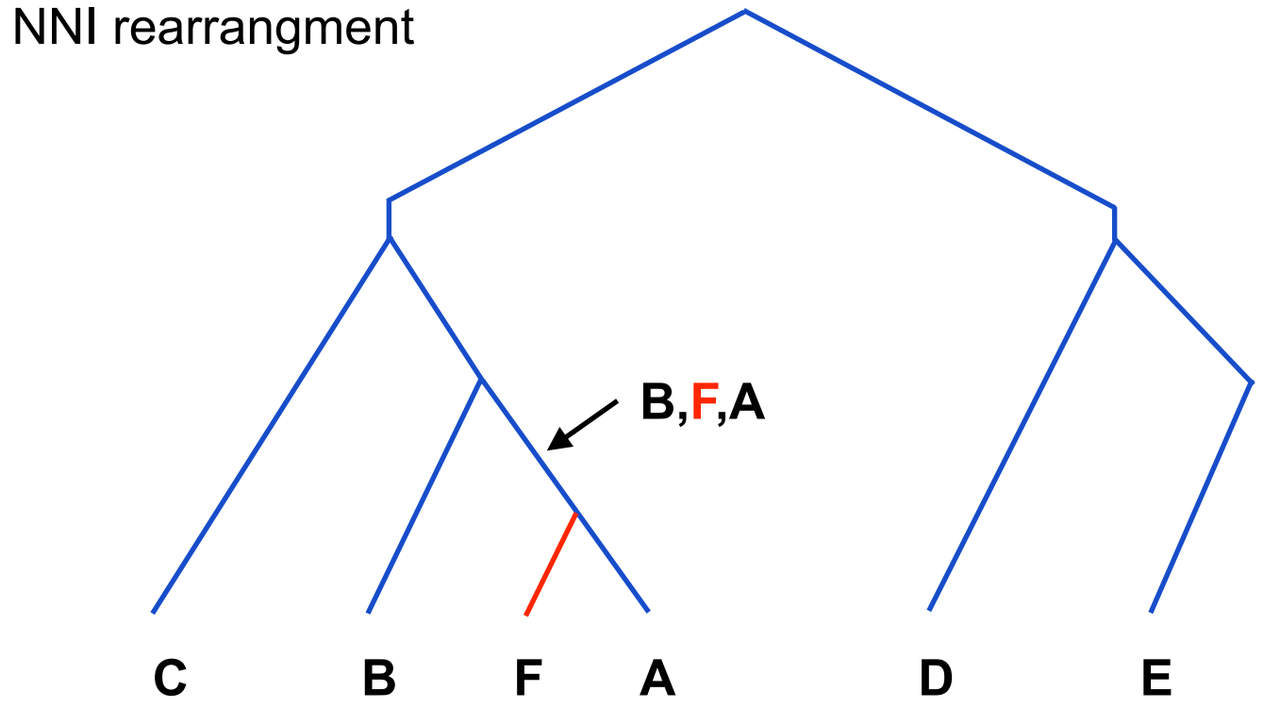


Who moved?

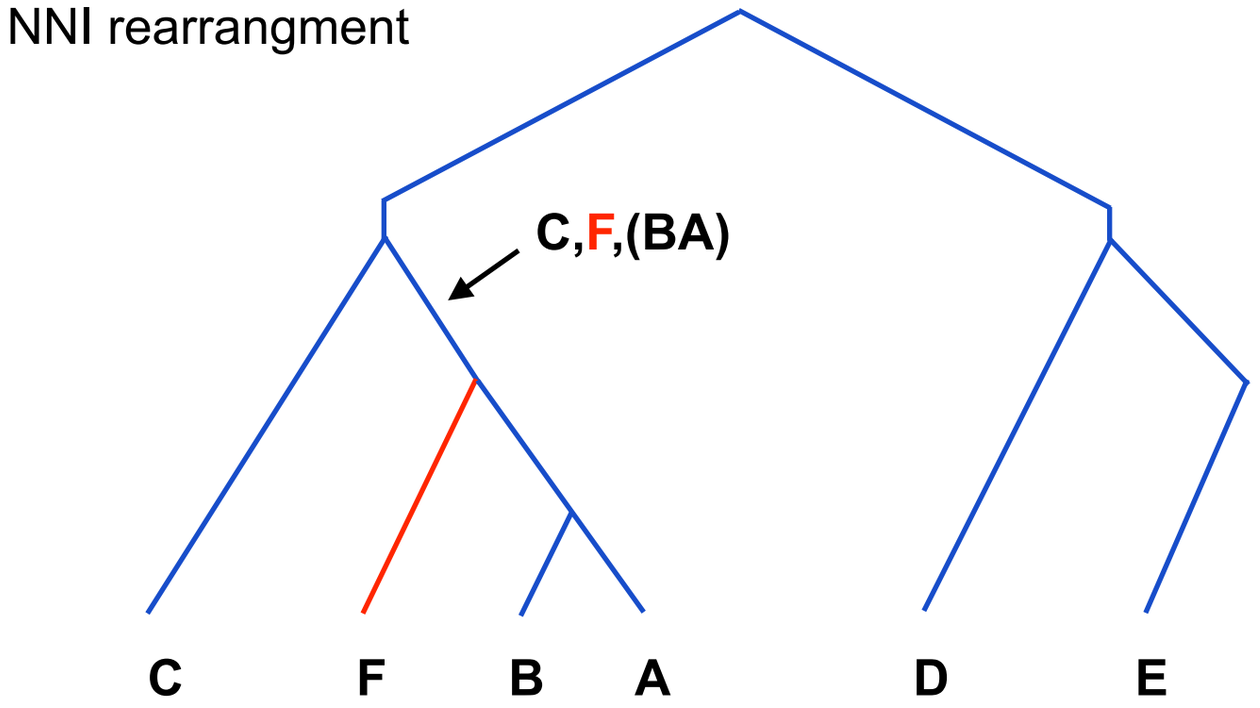


Heuristic: removing the “donor” lineage leads to an incorrect tree

Determining the direction of introgression

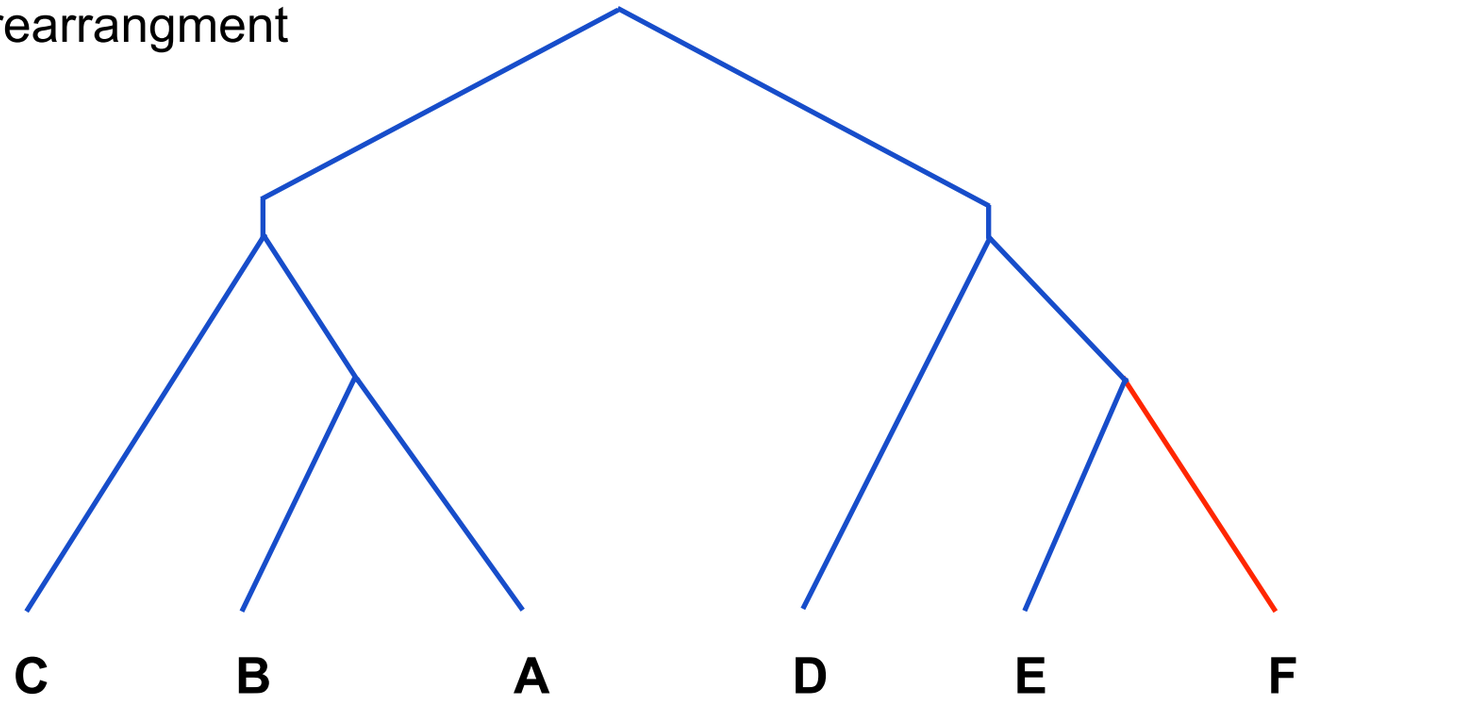


Determining the direction of introgression



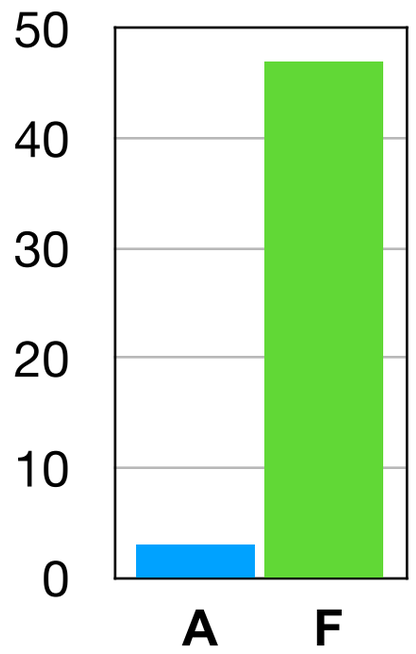
Determining the direction of introgression

NNI rearrangment

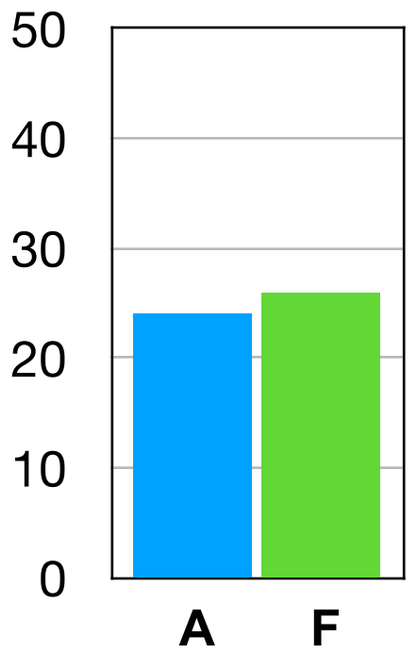


B,F,A
C,F,(BA)
(CBA),F,(DE)
D,F,E

Determining the direction of introgression

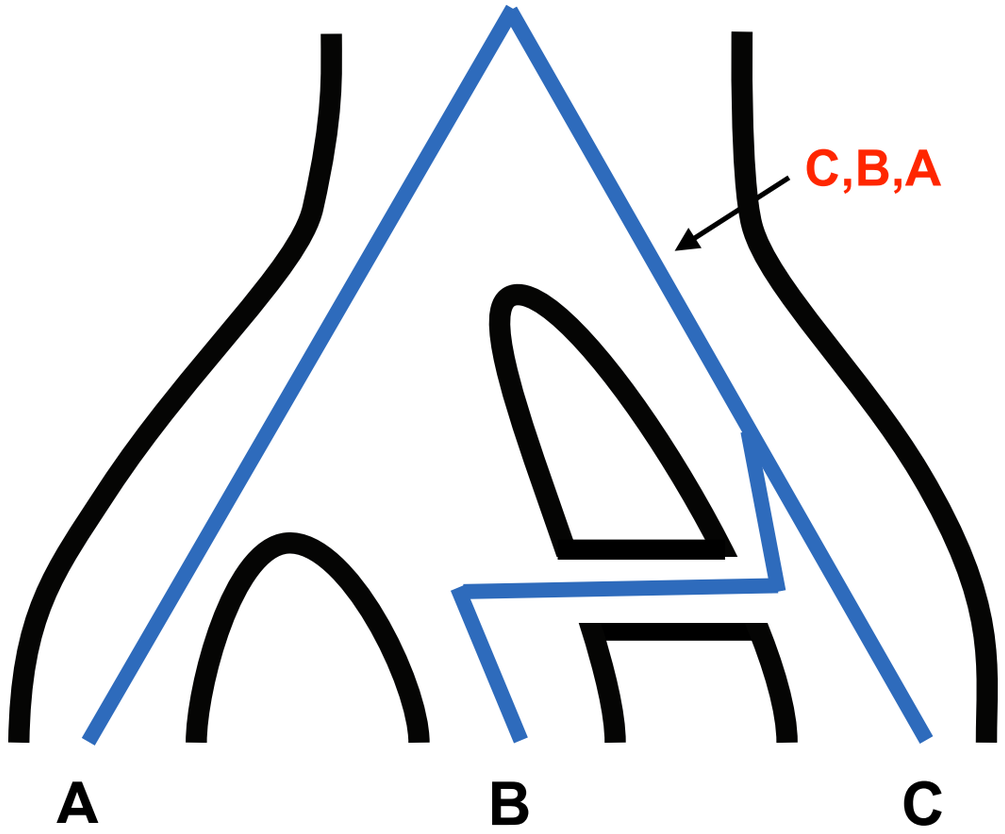


Determining the direction of introgression



We can also detect bidirectional introgression!

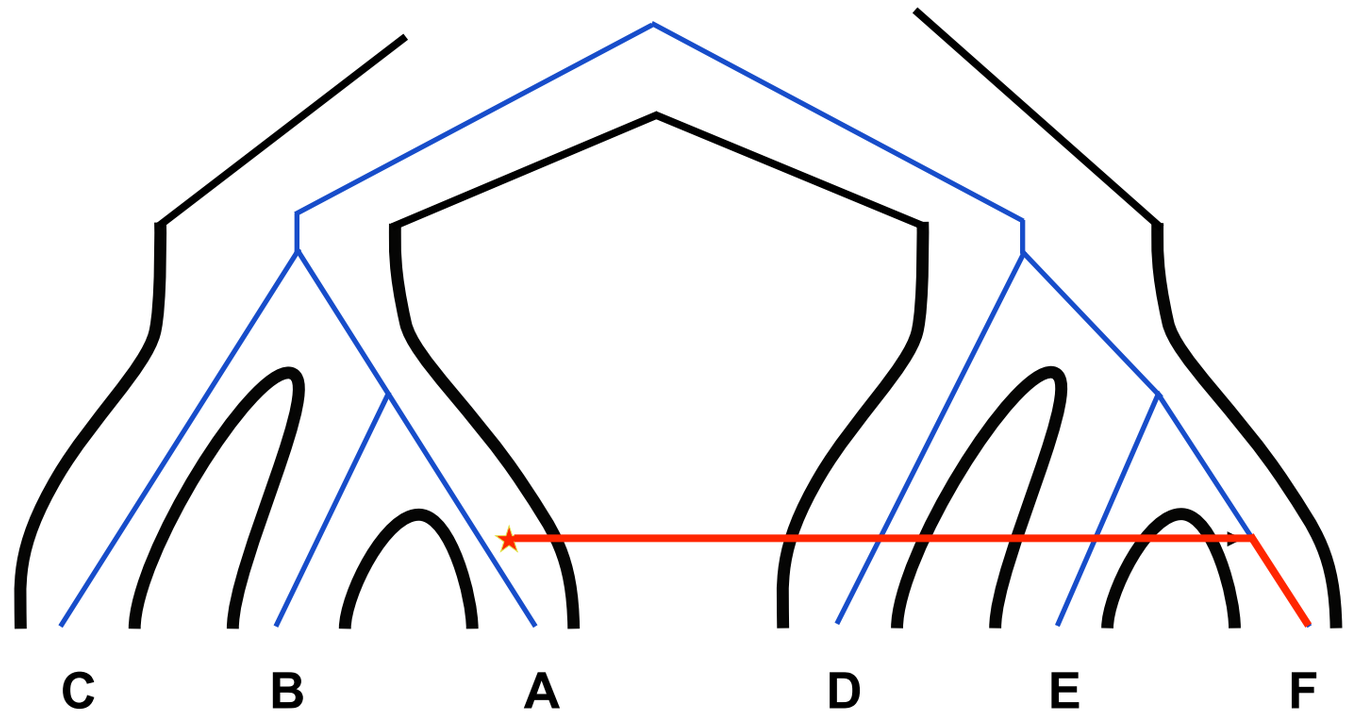
What is the direction of introgression?



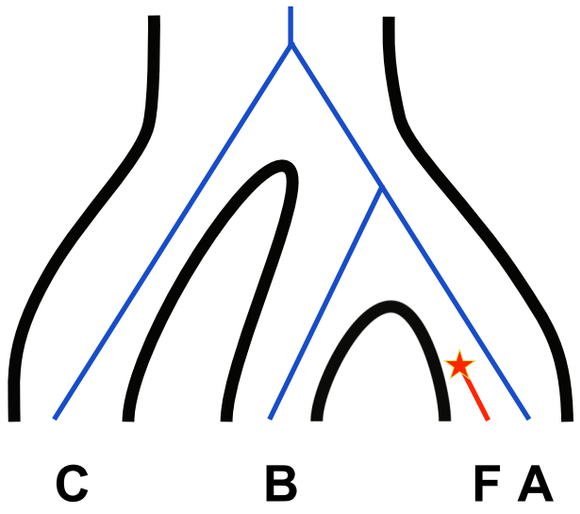
What we would like to know

How many introgression events?

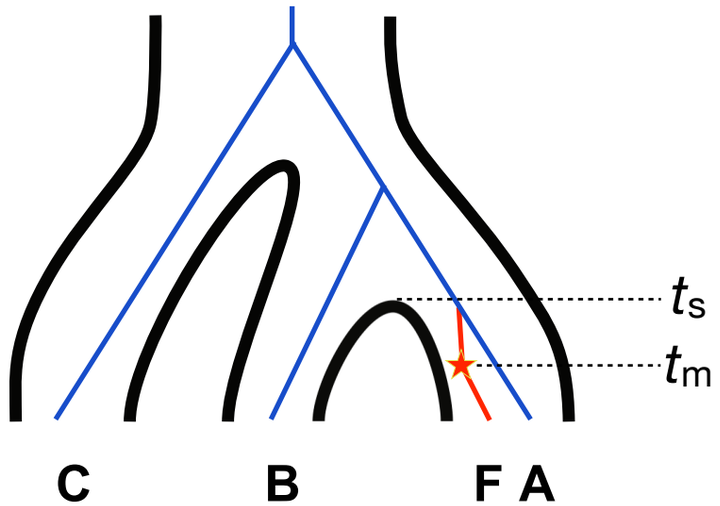
Which gene trees are produced by this introgression?

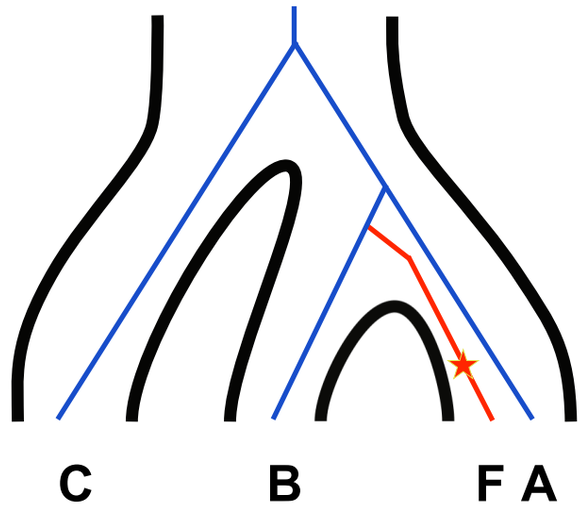


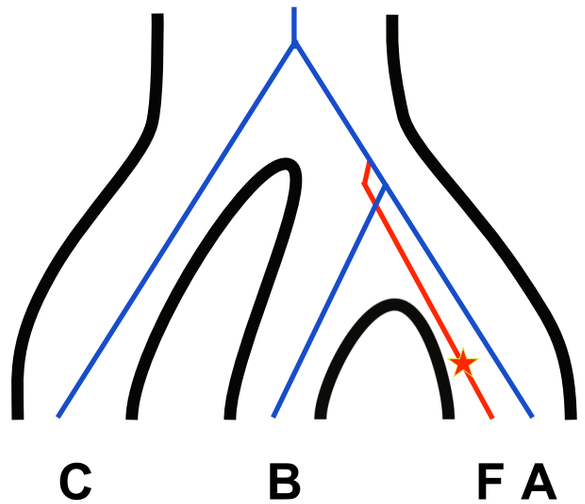
Which gene trees are produced by this introgression?

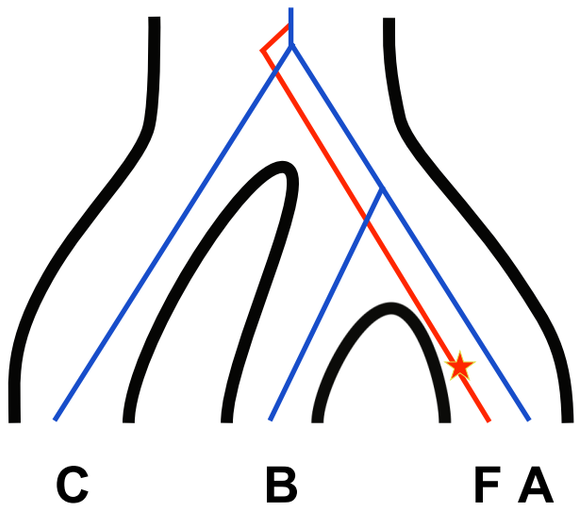


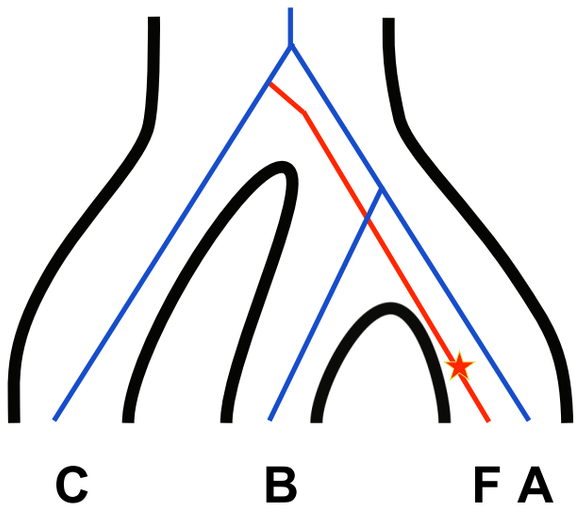
Which gene trees are produced by this introgression?



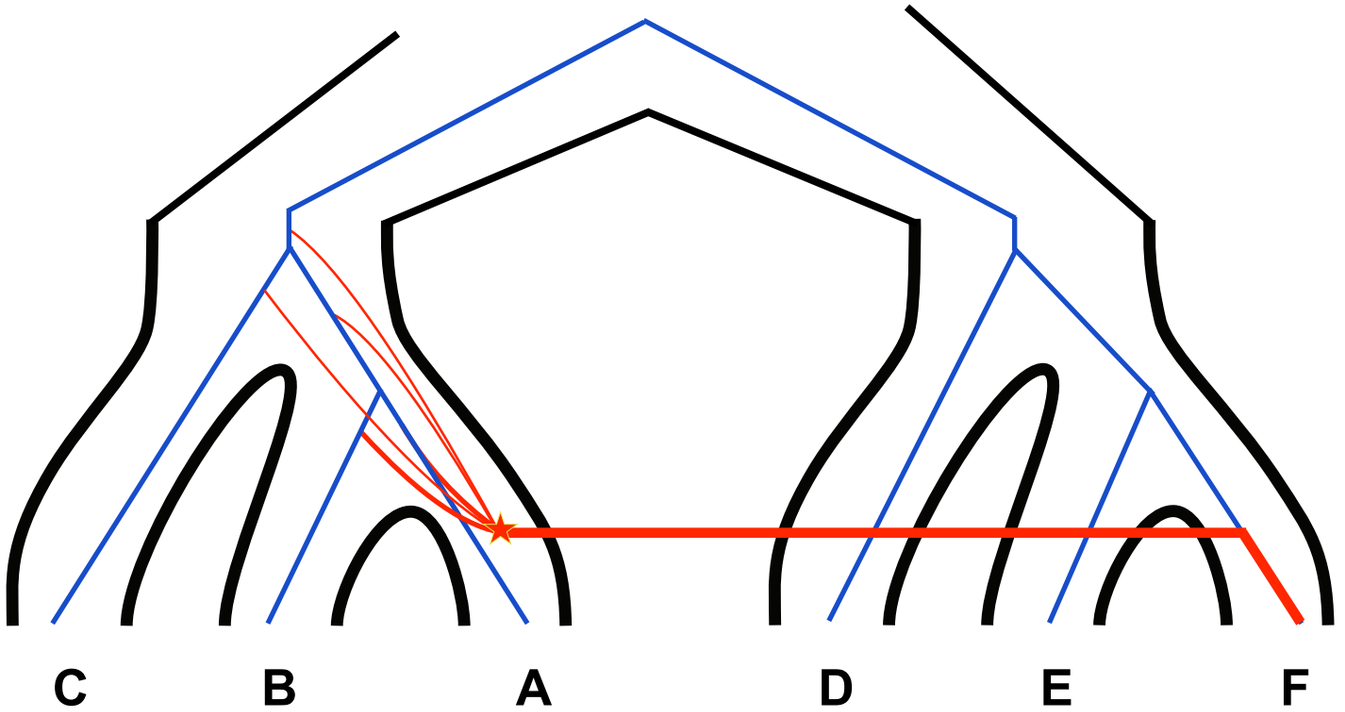




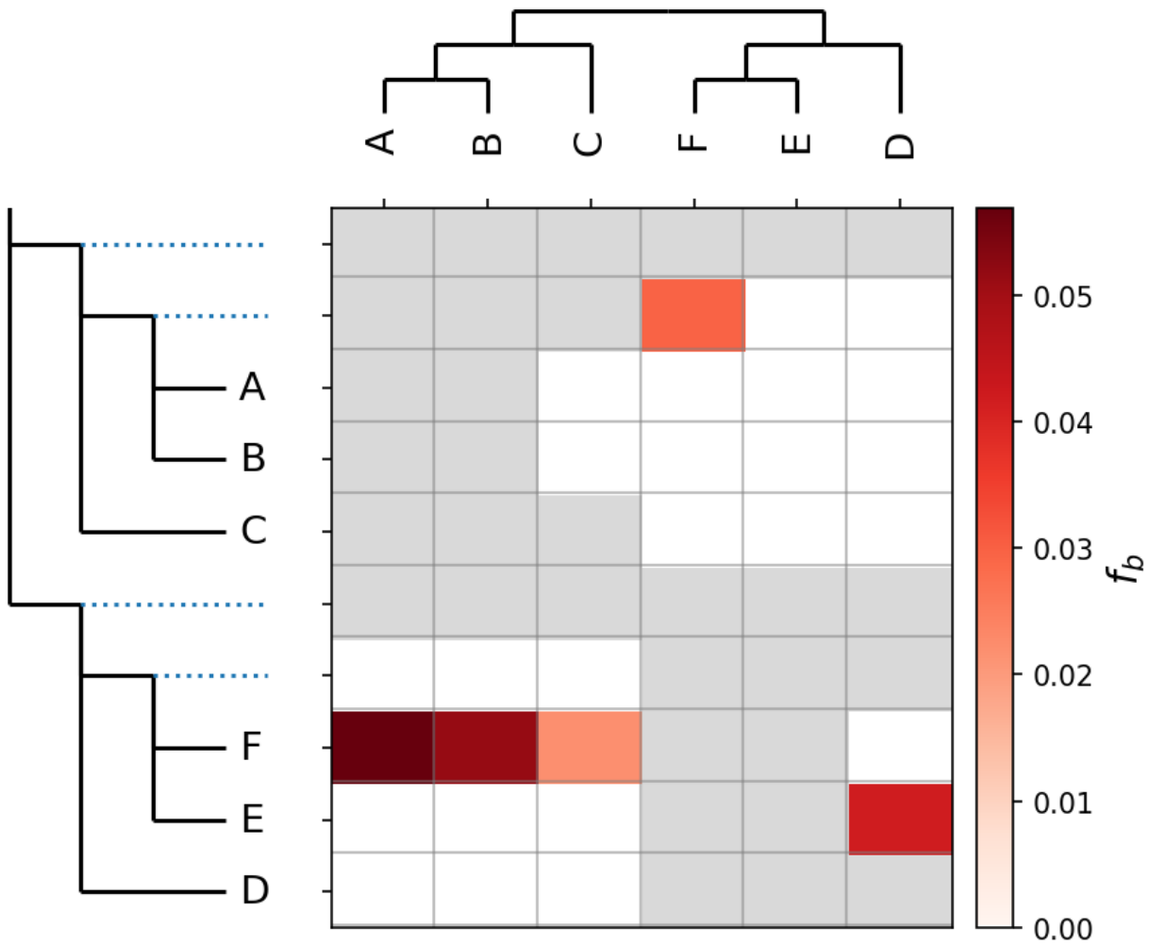




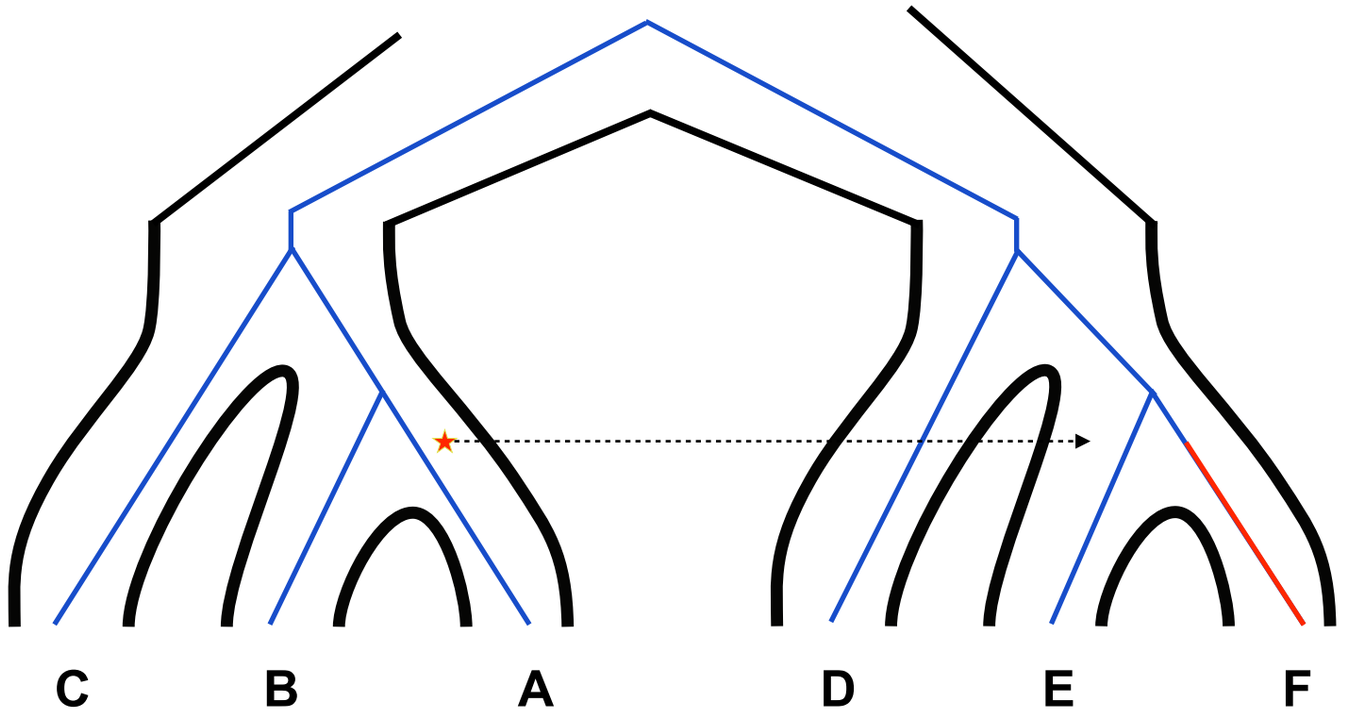
One introgression, many discordant attachments



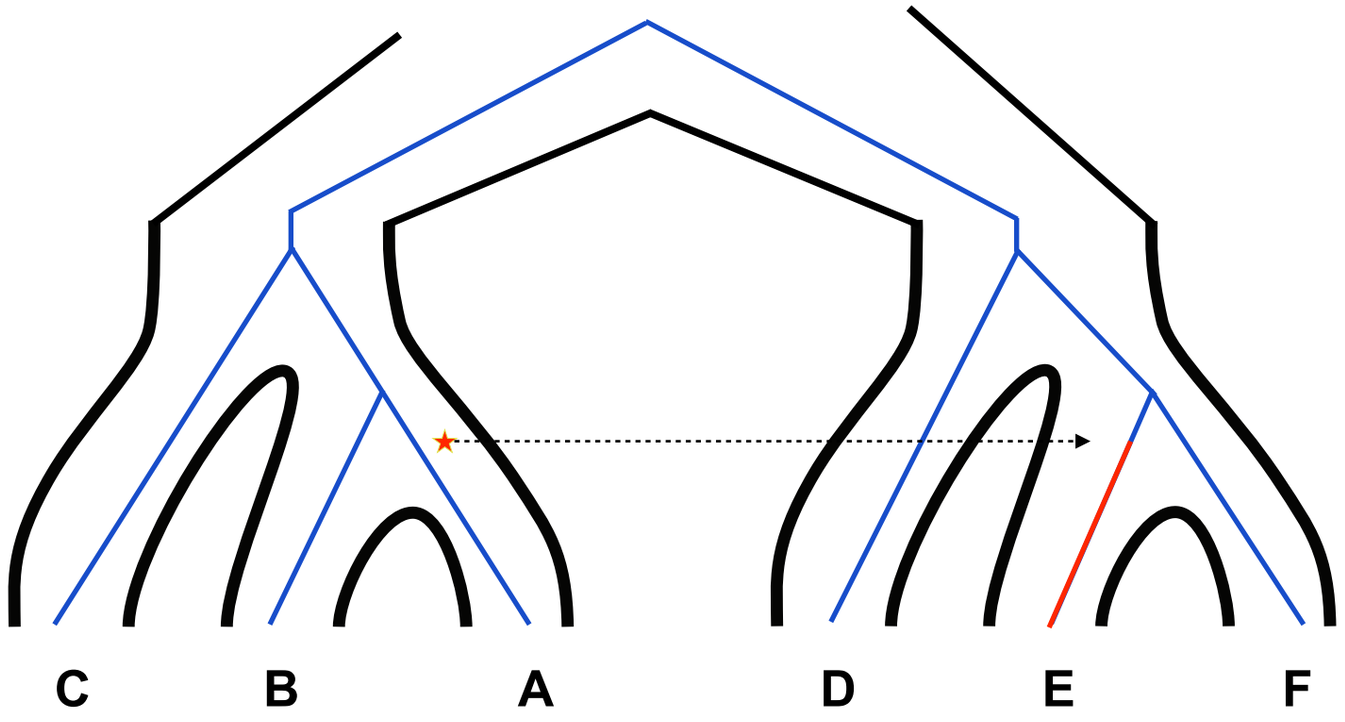
One introgression, many discordant attachments



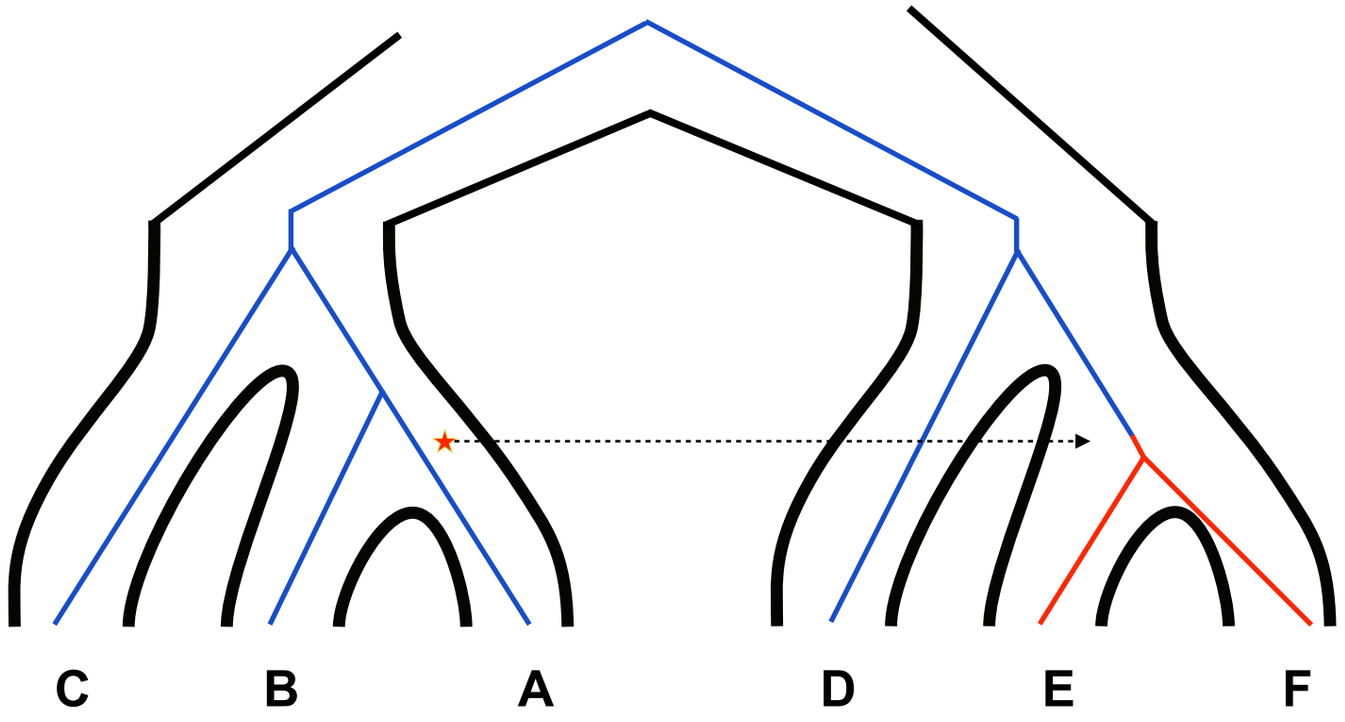
It gets more complicated



It gets more complicated



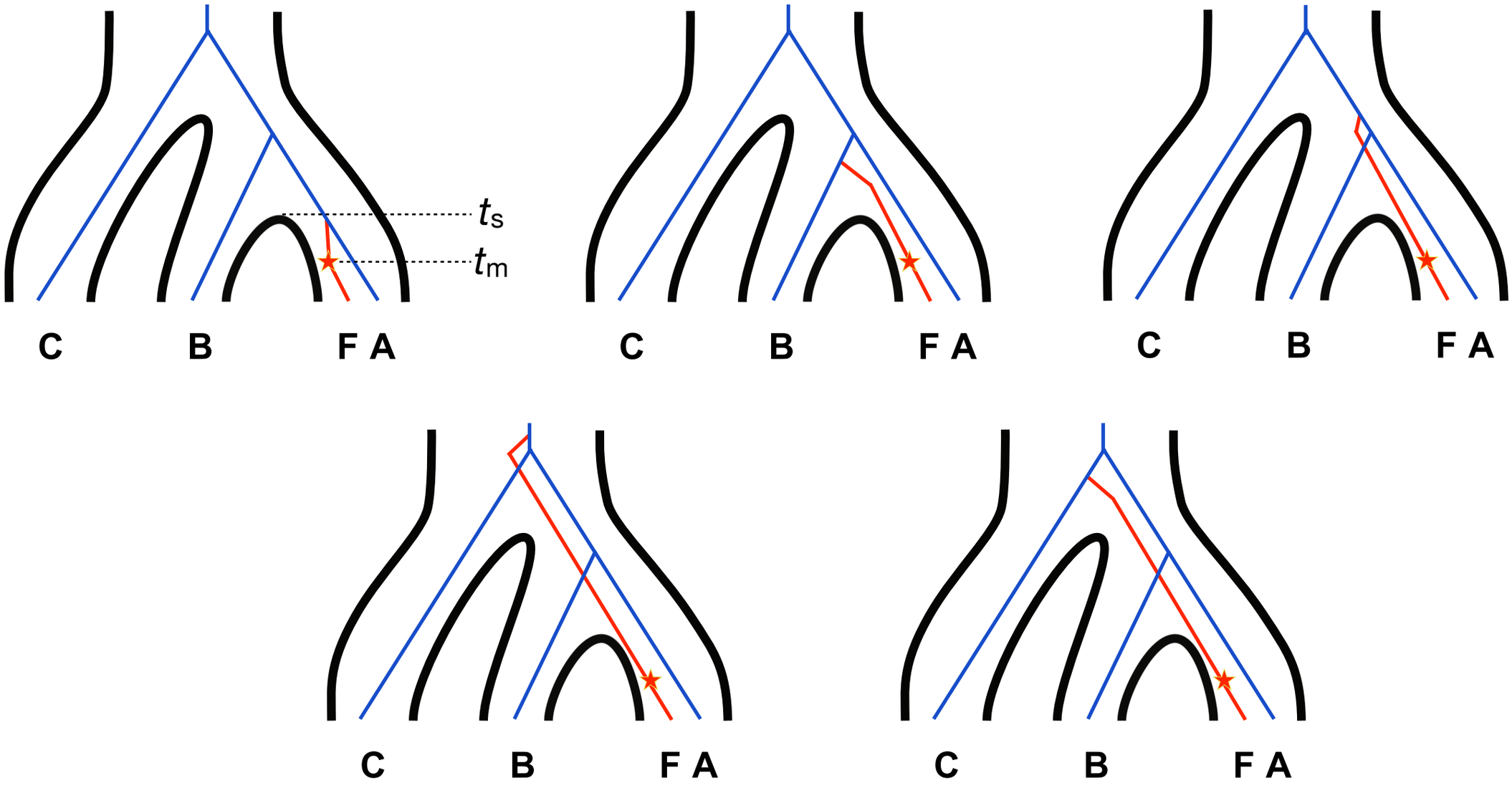
It gets more complicated



Some rules for combining introgression events

If **donor** is a tip lineage, donor attaches > than all others

If **donor** is a tip lineage, donor attaches > than all others

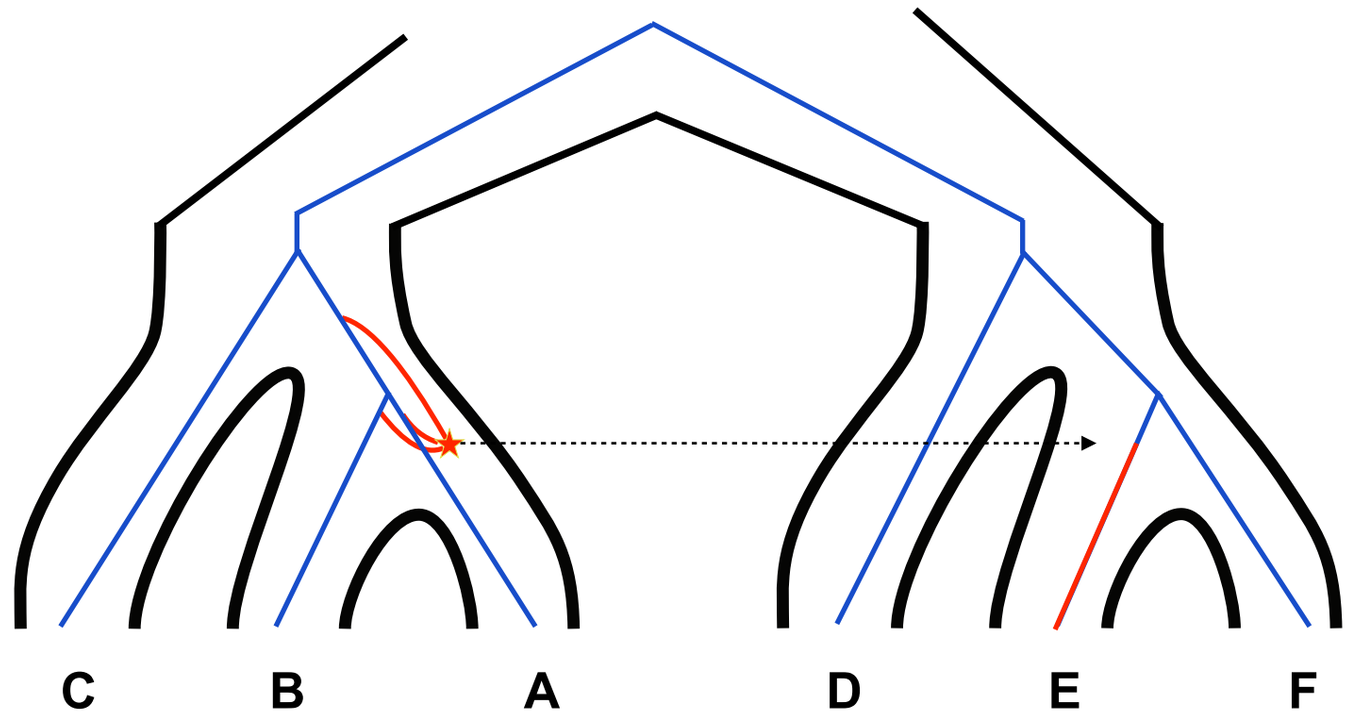


Some rules for combining introgression events

If **donor** is a tip lineage, donor attaches > than all others

If **donor** is an ancestral lineage, 2 daughter lineages are =

If **donor** is an ancestral lineage, 2 daughter lineages are =



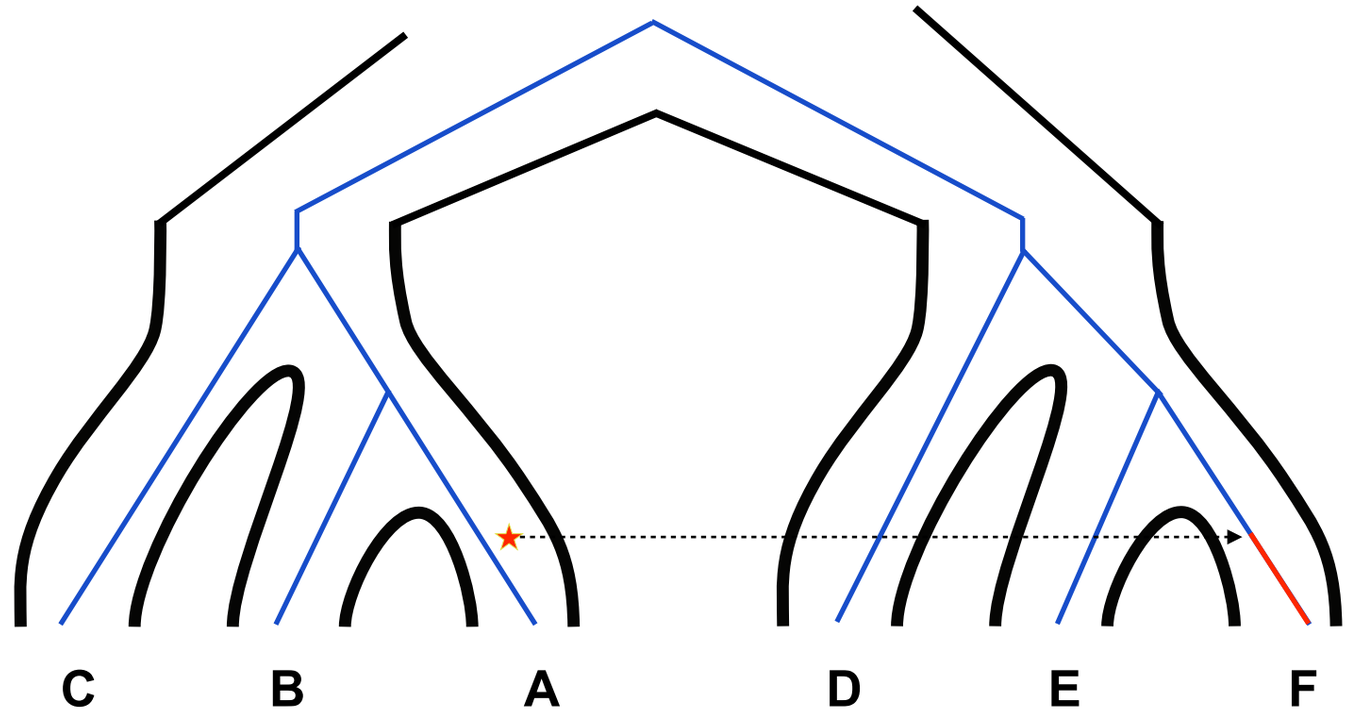
Some rules for combining introgression events

If **donor** is a tip lineage, donor attaches > than all others

If **donor** is an ancestral lineage, 2 daughter lineages are =

If **recipient** is a tip lineage, it is the only one moving

If **recipient** is a tip lineage, it is the only one moving



Some rules for combining introgression events

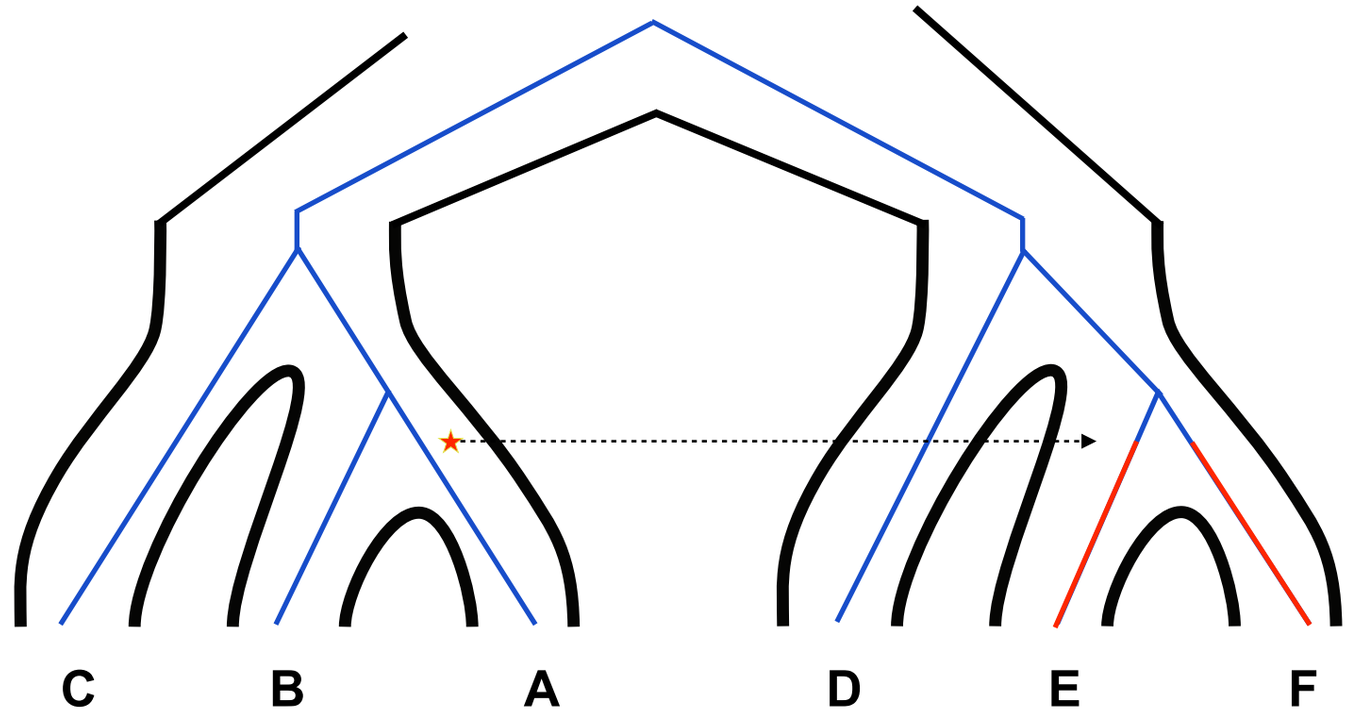
If **donor** is a tip lineage, donor attaches > than all others

If **donor** is an ancestral lineage, 2 daughter lineages are =

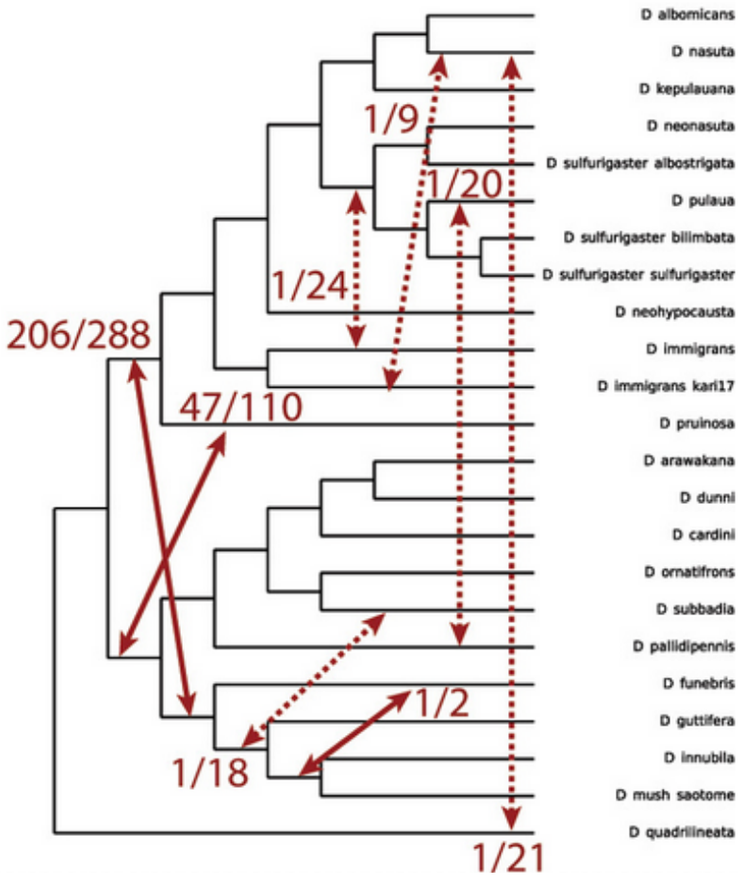
If **recipient** is a tip lineage, it is the only one moving

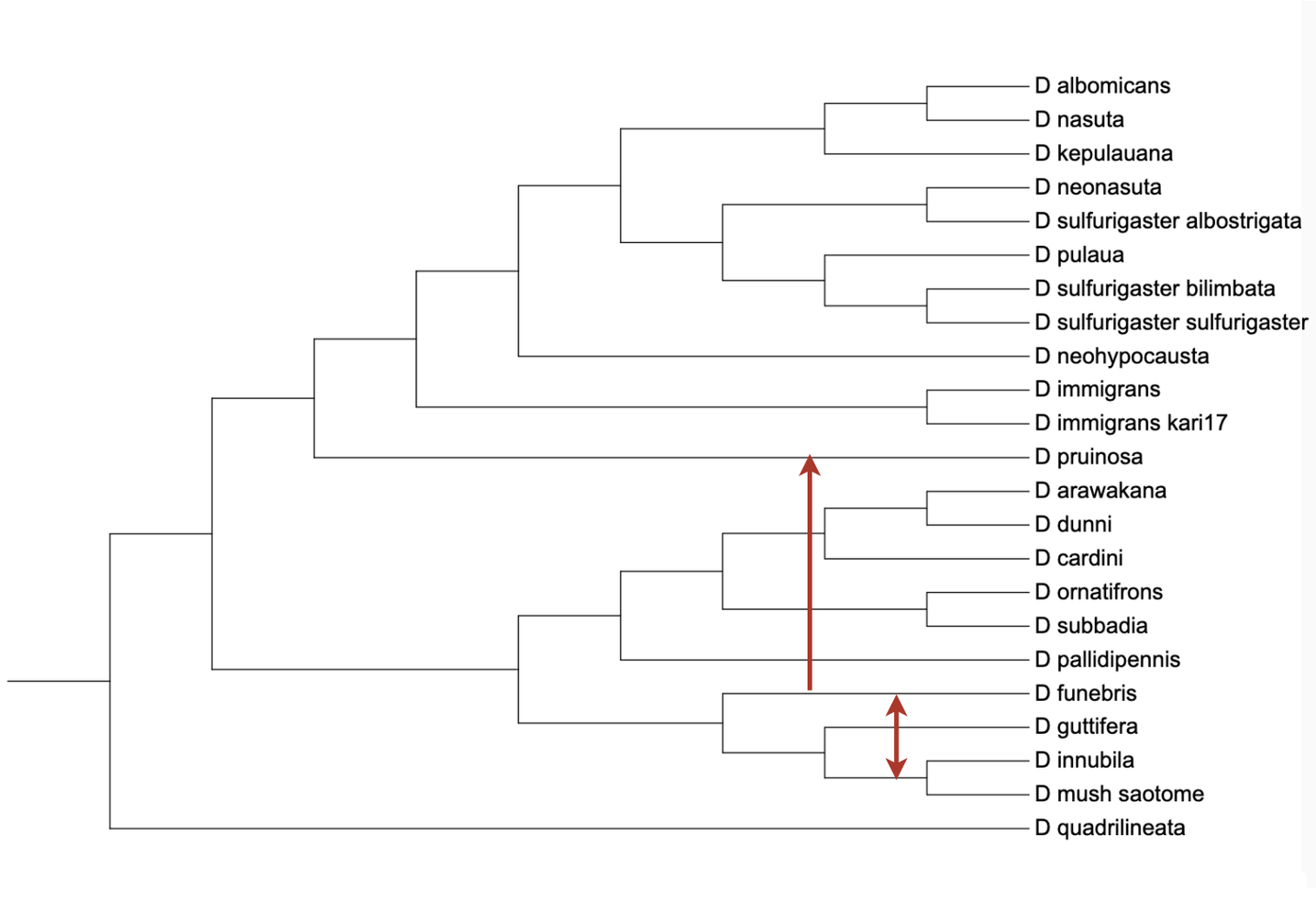
If **recipient** is an ancestral lineage, 2 daughter lineages are =

If **recipient** is an ancestral lineages, 2 daughter lineages are =



Results from *Drosophila*



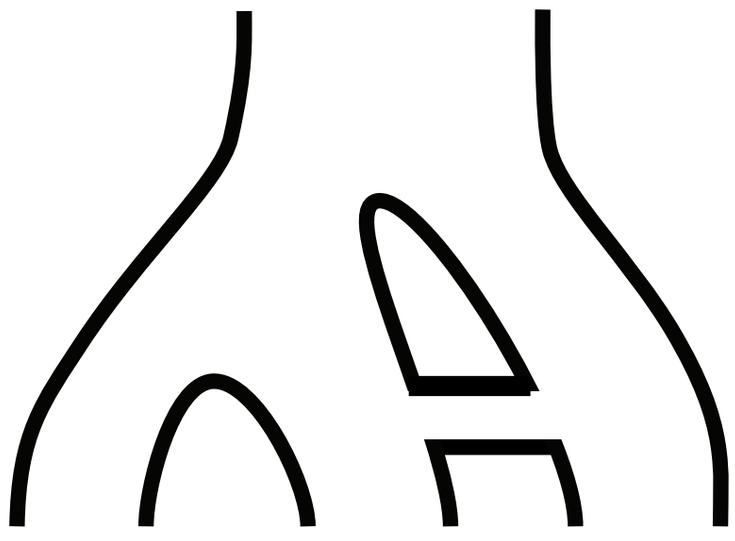


What we would like to know

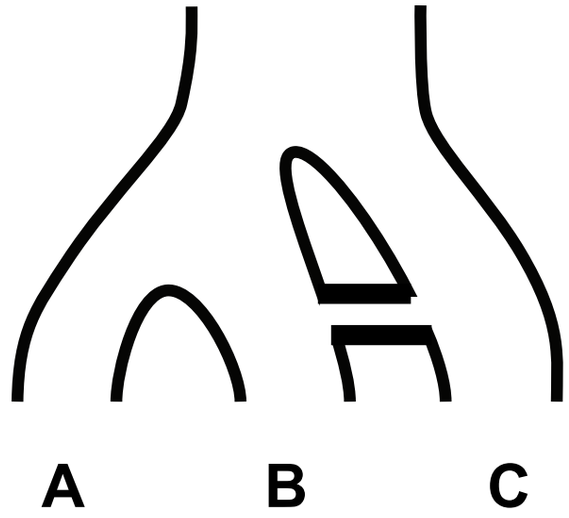
Who is involved in introgression?

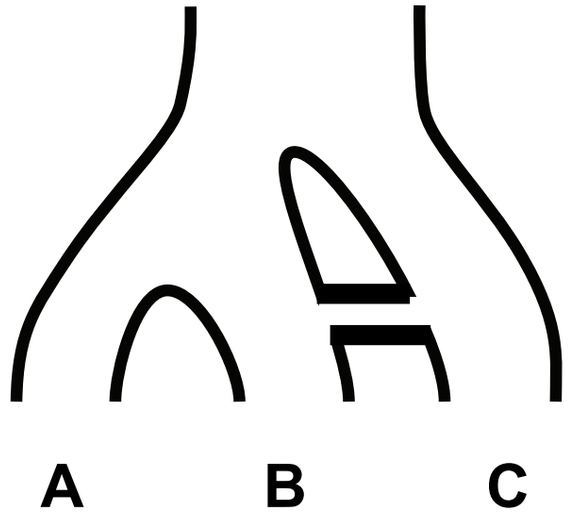
What is the direction of introgression?

How many introgression events?



A unified model





Thanks!

Sarthak Mishra

Megan Smith

Rob Lanfear

Anton Suvorov

Dan Schrider