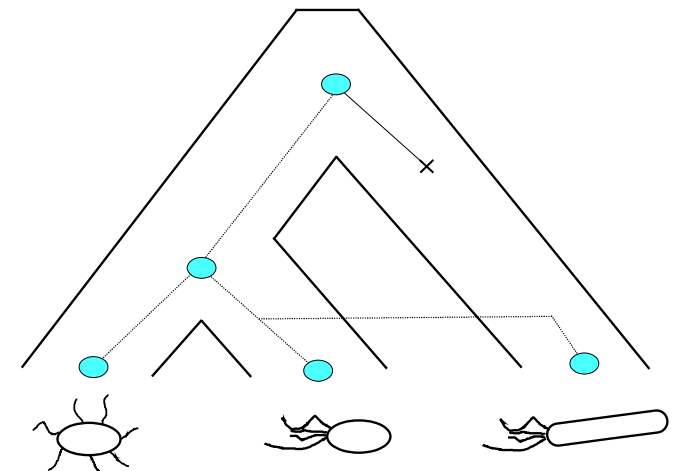
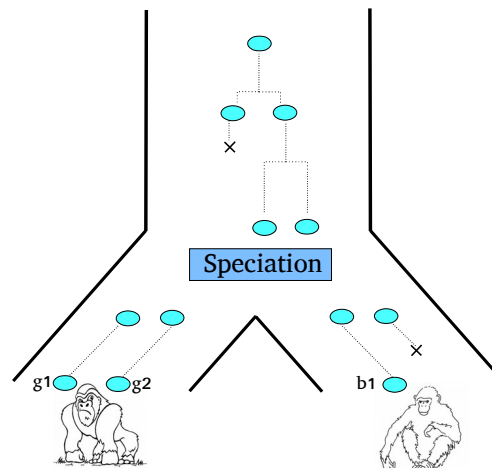
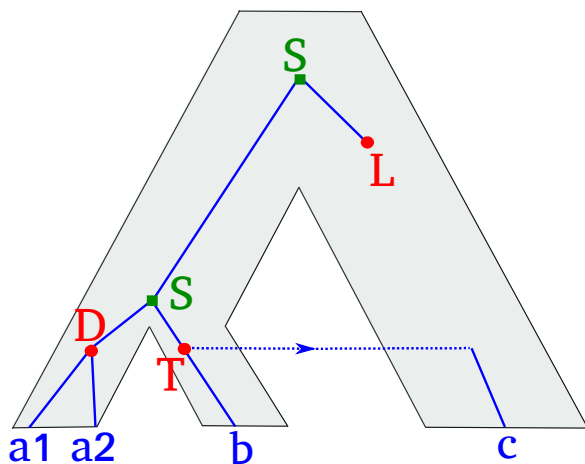
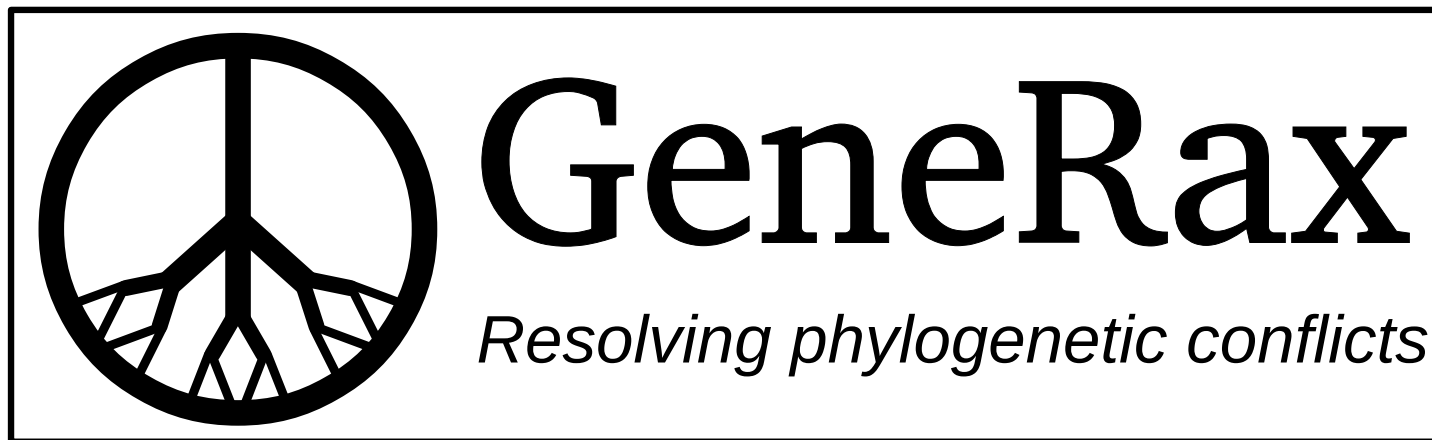
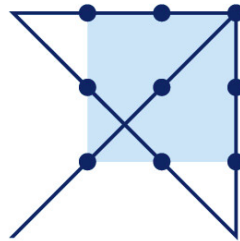


# Gene tree inference under gene duplication, transfer and loss with



# About me

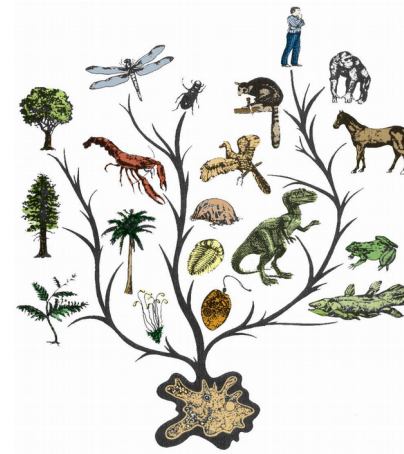
- Benoit Morel ([benoit.morel@h-its.org](mailto:benoit.morel@h-its.org))
- PhD student (advisor: Alexandros Stamatakis)
- Working on phylogenetics tools development



**HITS**

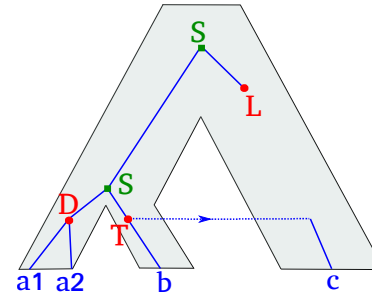
**Heidelberg Institute for  
Theoretical Studies**

# Outline



- Phylogenetics, species and gene trees

- Gene tree correction and reconciliation problem



- Our solution



# Definitions

In the scope of this talk:

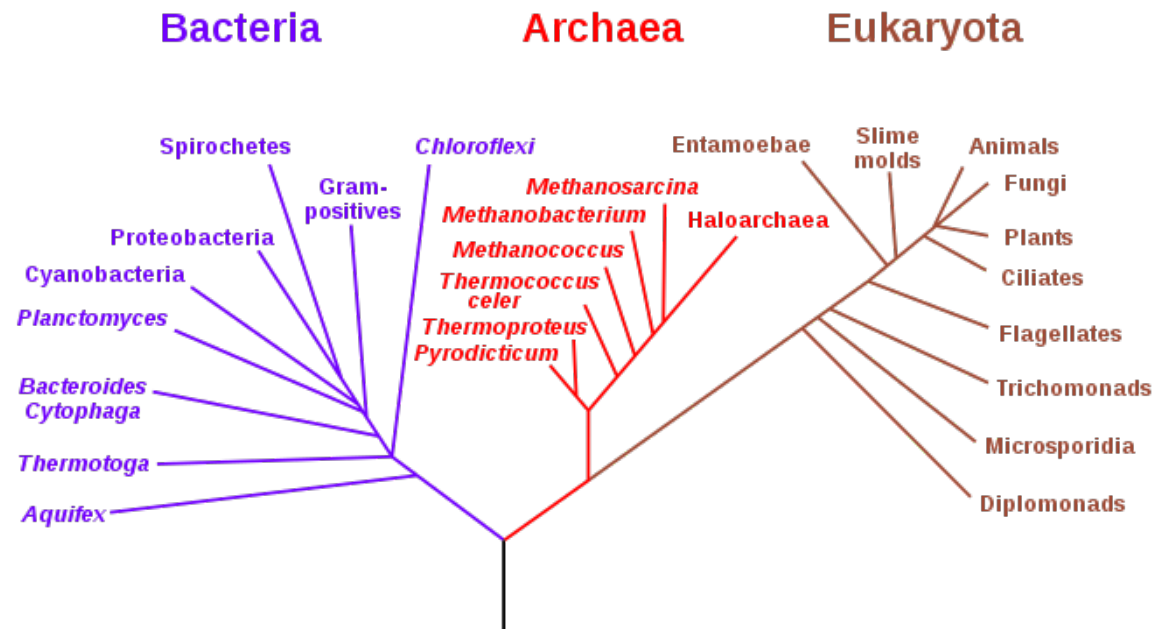
**Species**: group of organisms that share a pool of genes and that are able to interbreed

**Gene**: a DNA or protein sequence that belongs to a species

**Gene family (or homologous genes)**: set of genes that share a common ancestor.

# (Phylogenetic) species tree

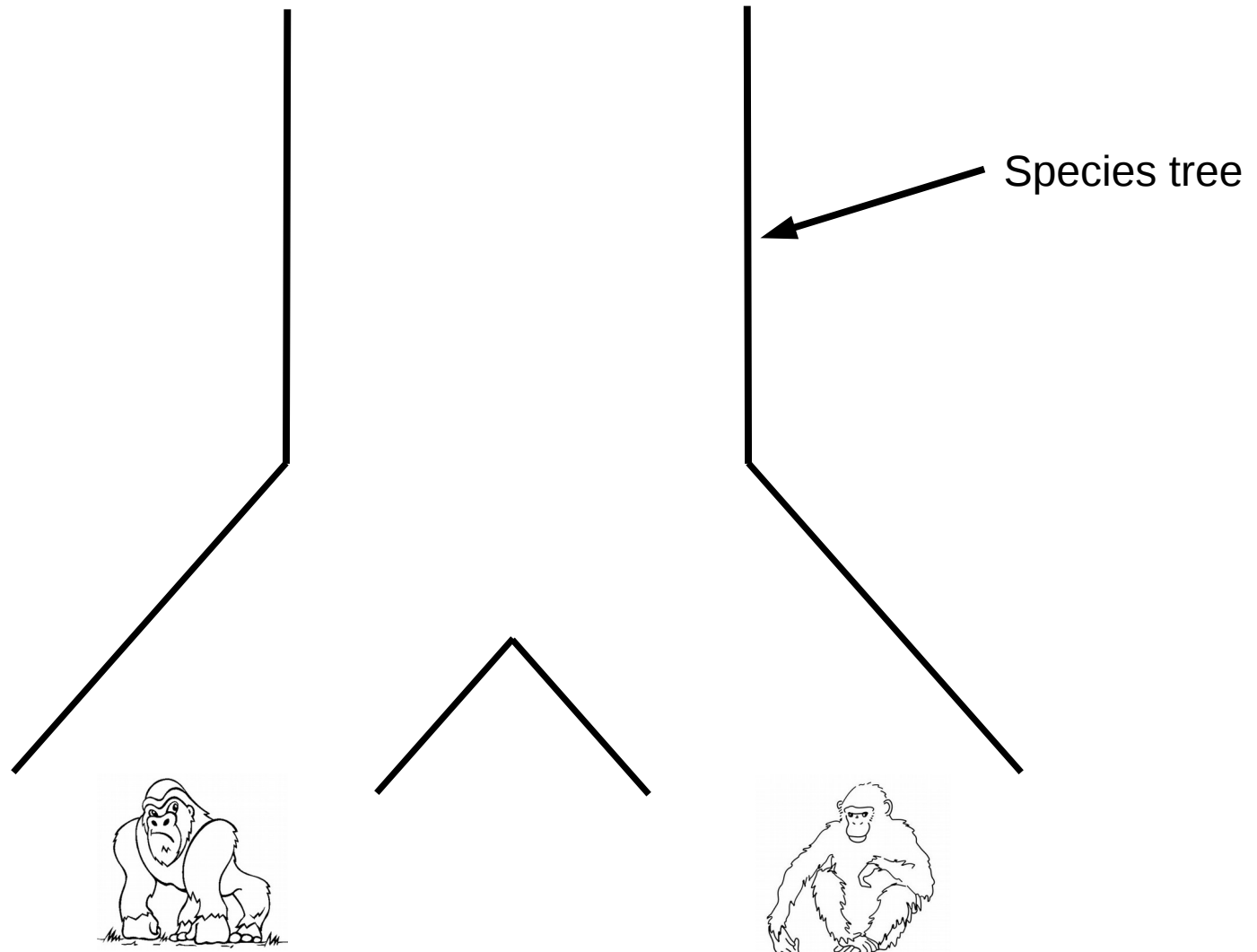
Represent the evolutionary history of species



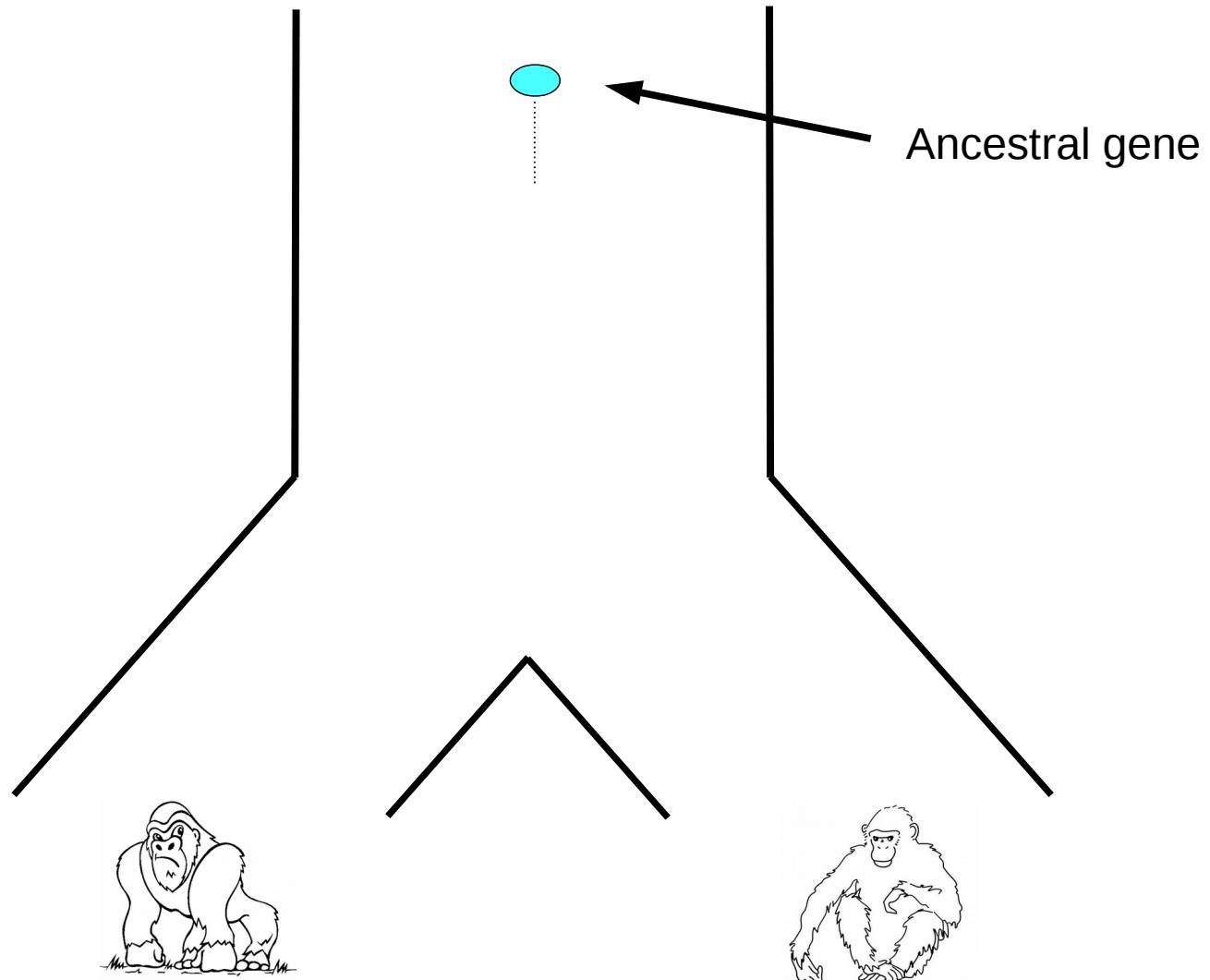
# Gene tree

- Represent the evolutionary history of homologous genes
- High correlation with the species tree

# Gene trees evolve along a species tree

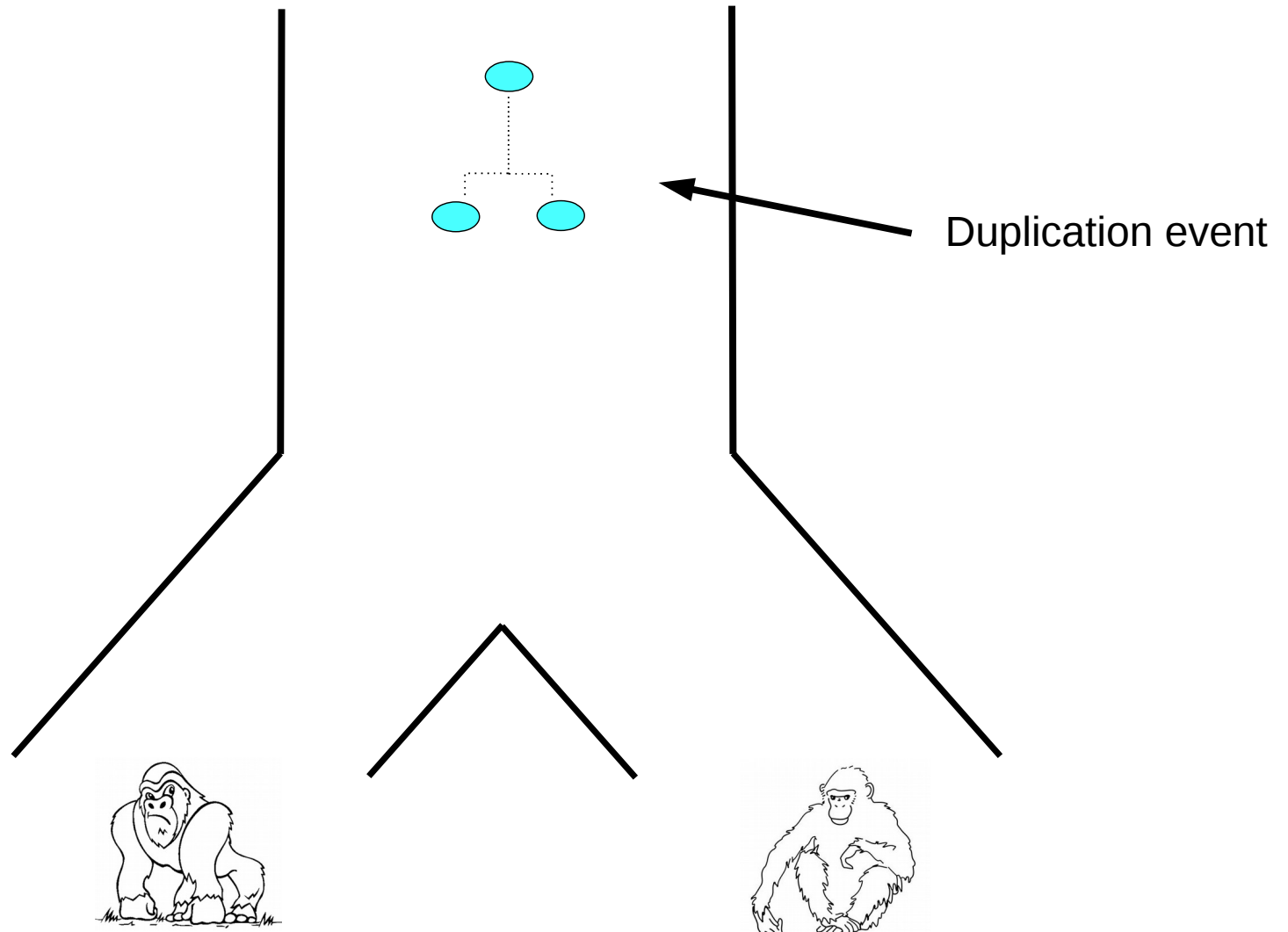


# Gene trees evolve along a species tree

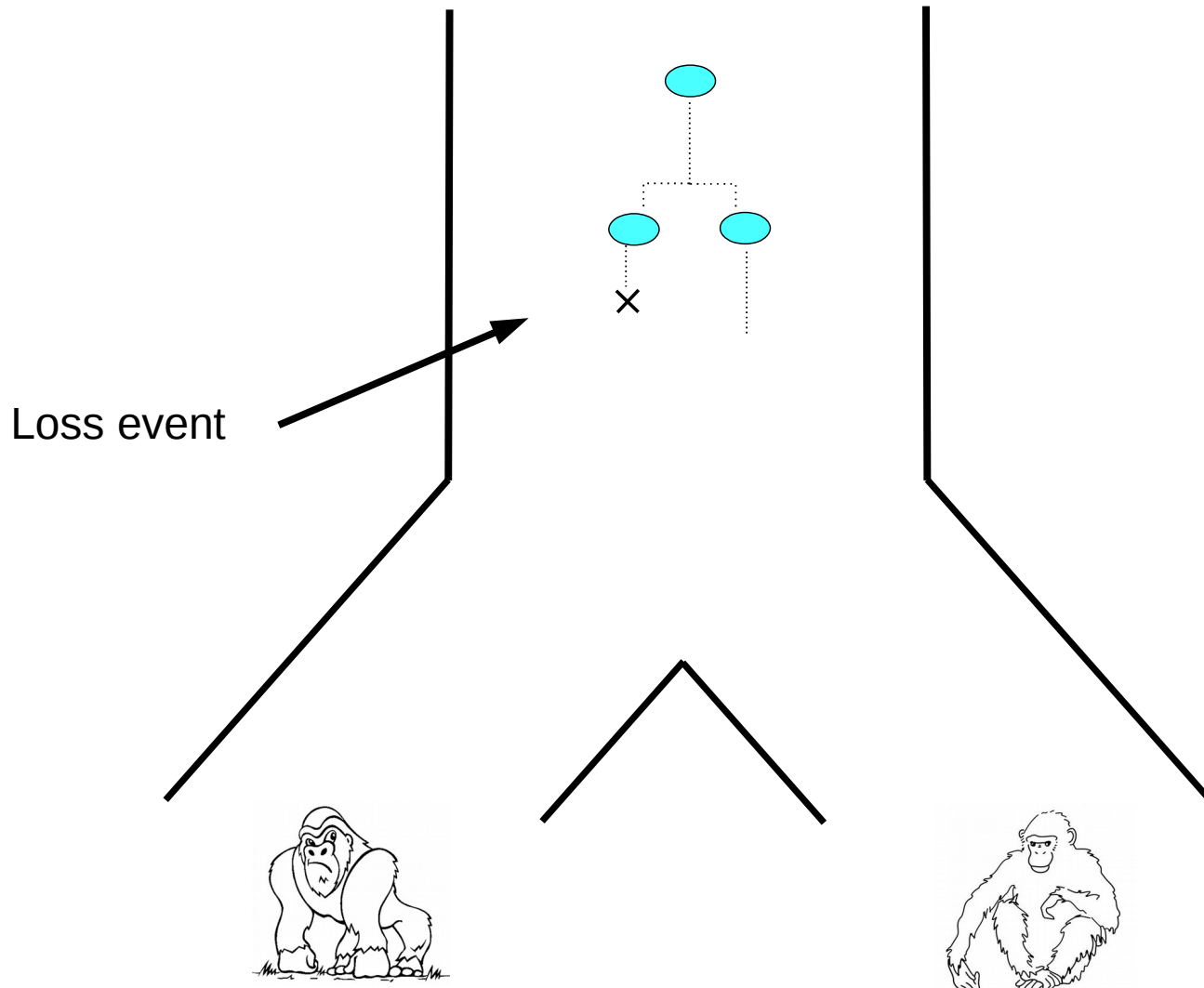




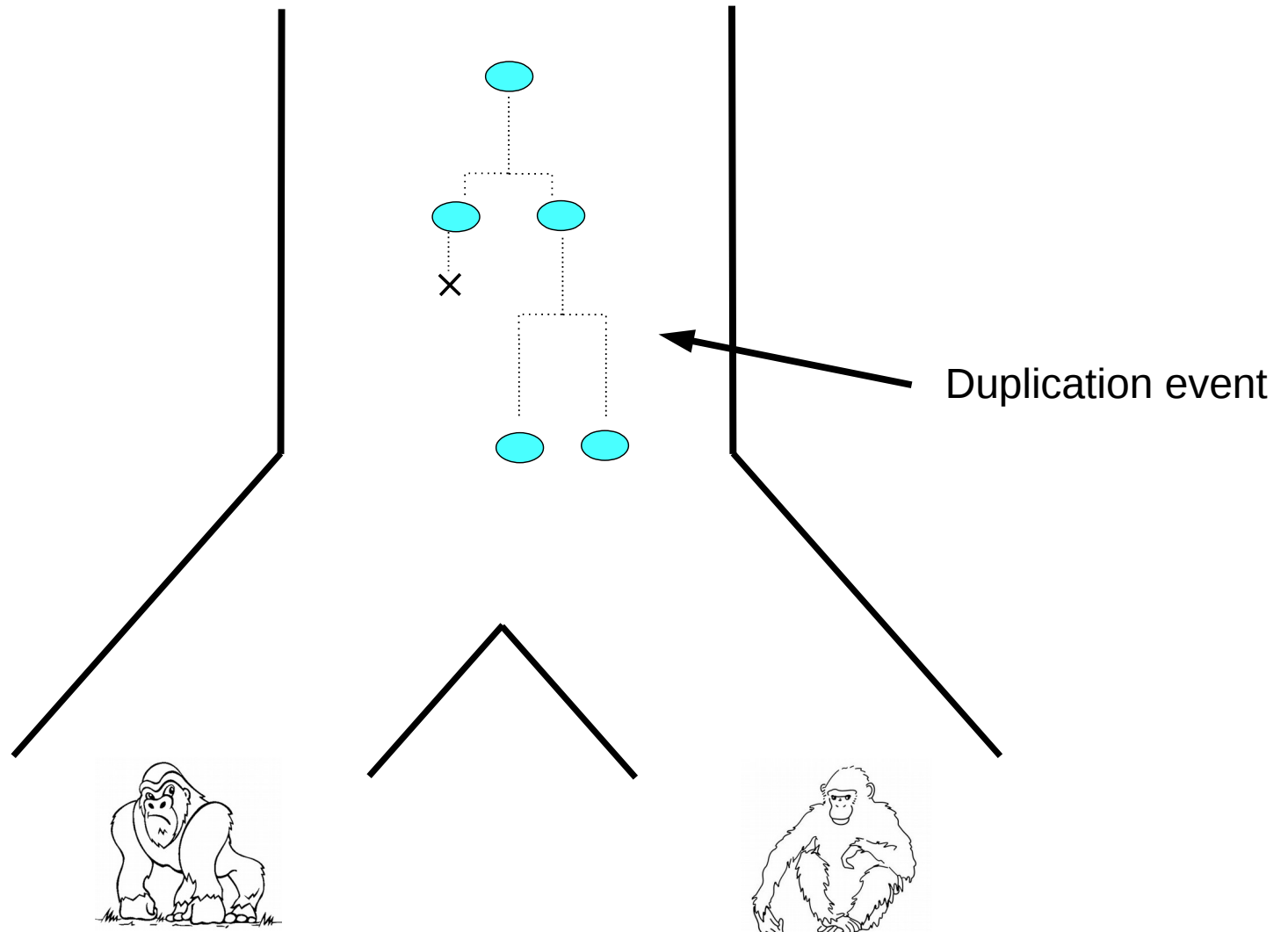
# Gene trees evolve along a species tree



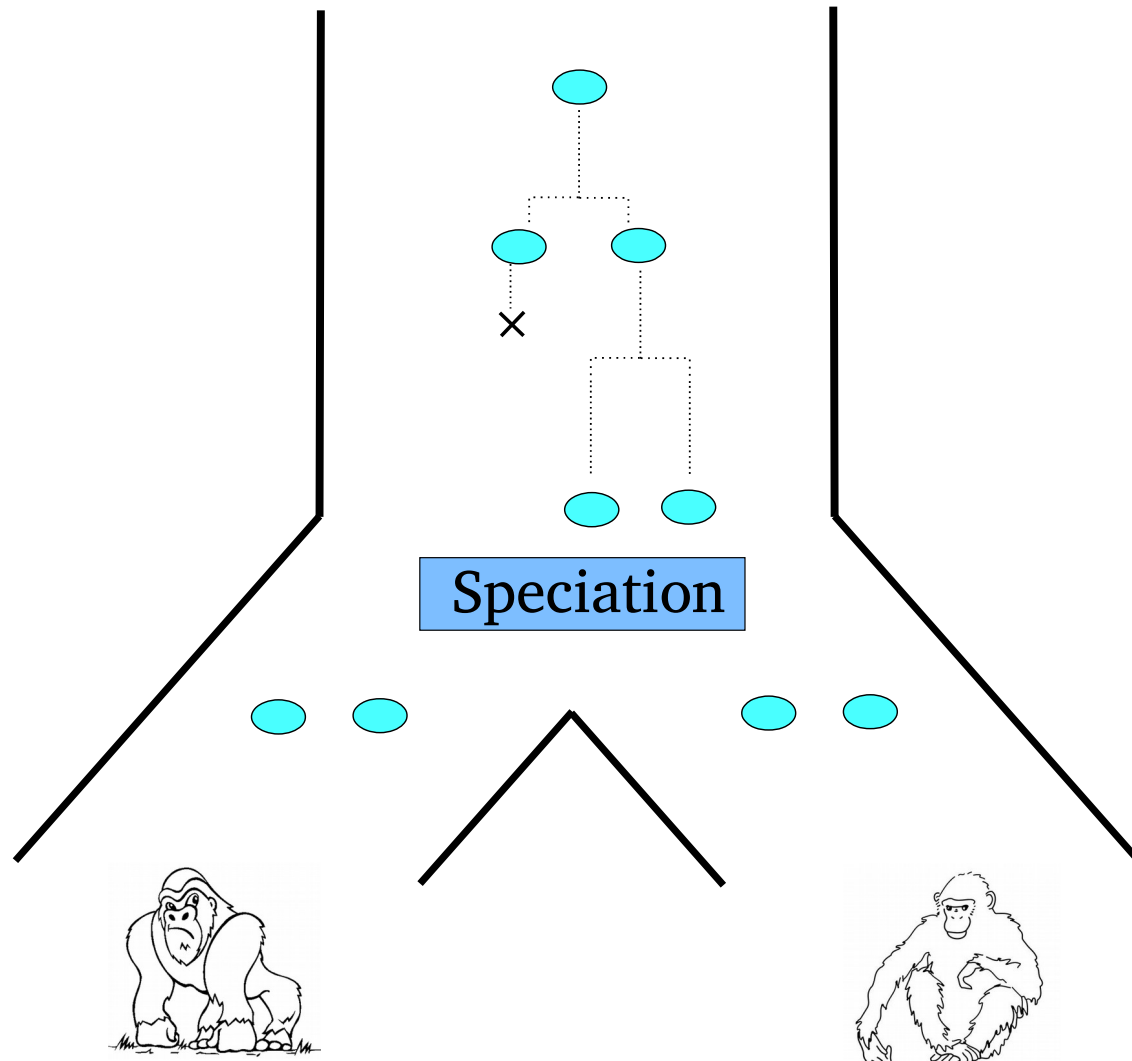
# Gene trees evolve along a species tree



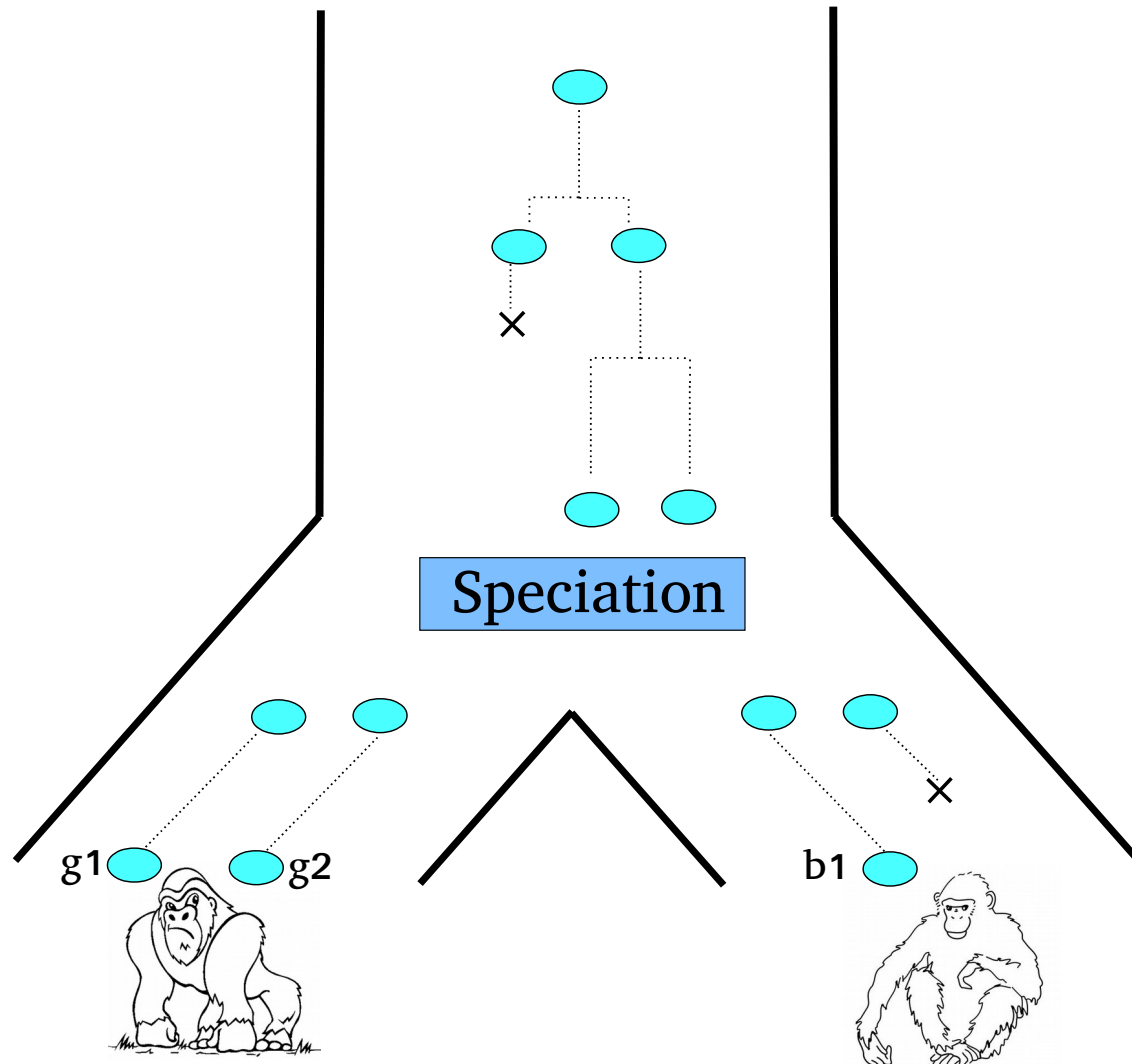
# Gene trees evolve along a species tree



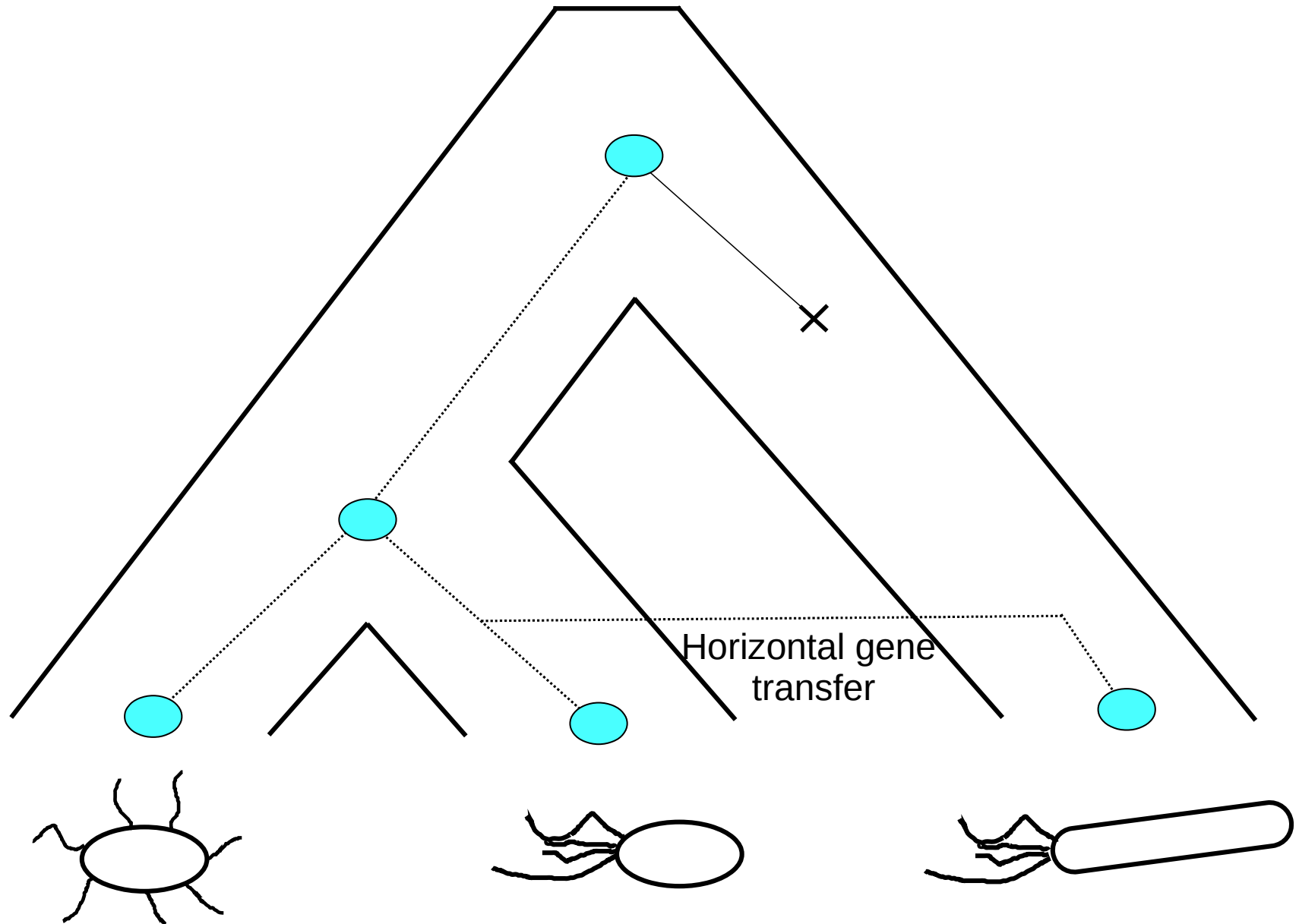
# Gene trees evolve along a species tree



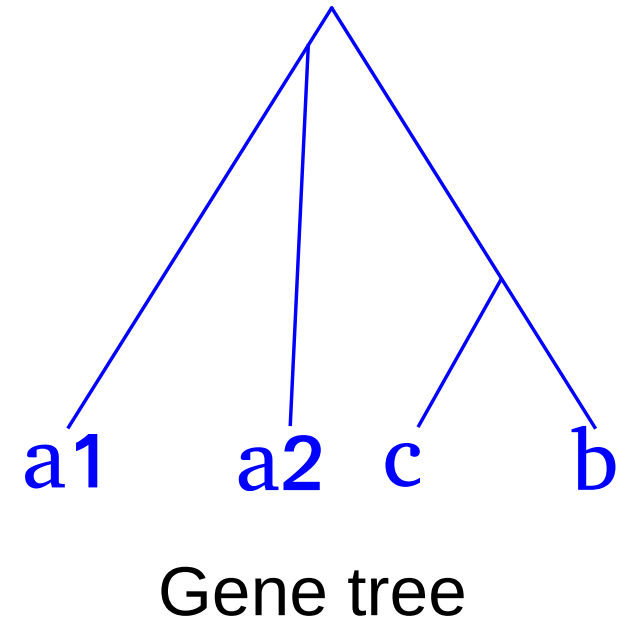
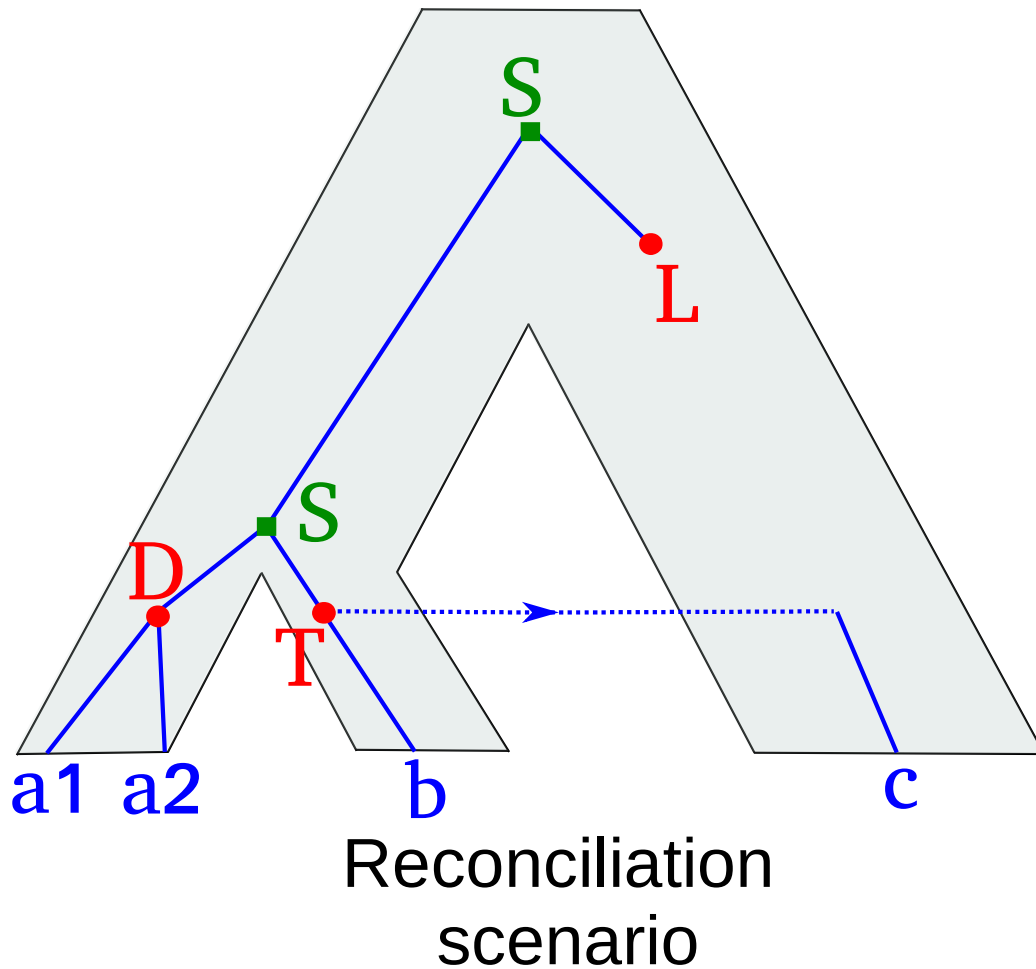
# Gene trees evolve along a species tree



# Gene trees evolve along a species tree



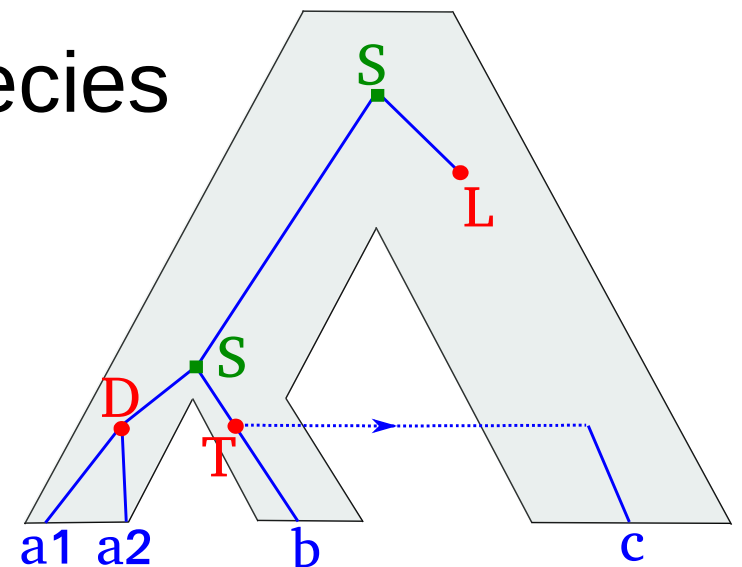
# The reconciliation scenario corresponds to a gene tree



# The UndatedDTL model

A gene lineage present in a given species lineage can either:

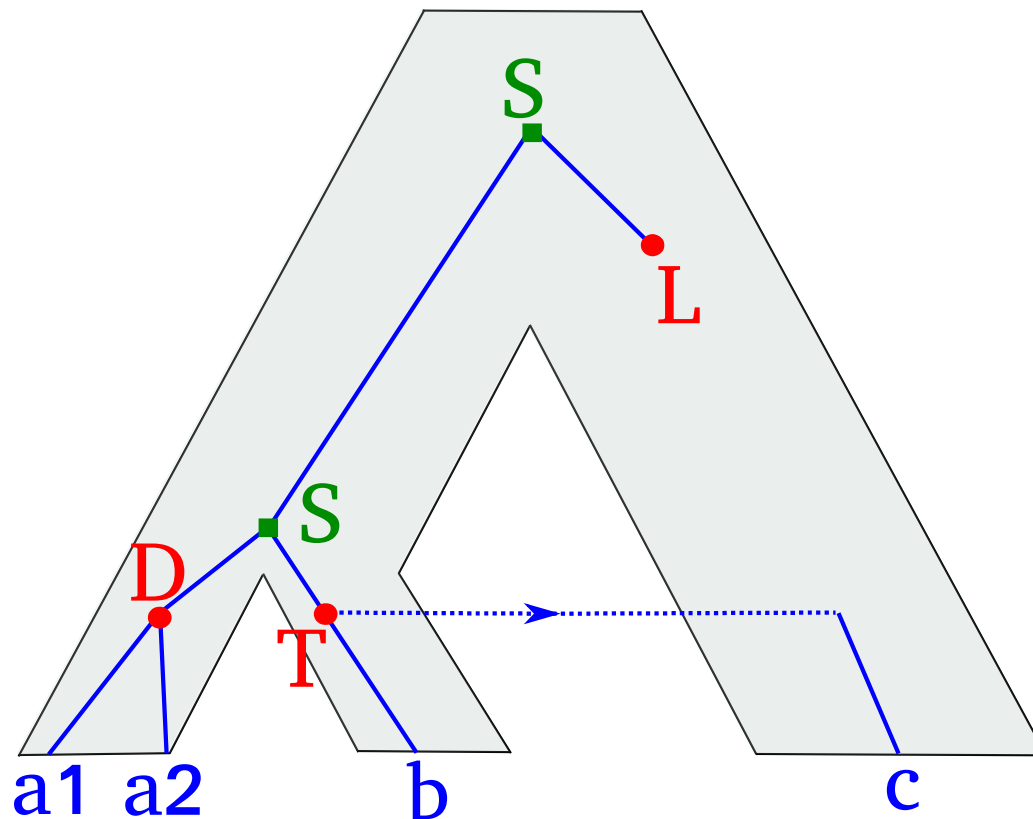
- Duplicate with probability  $p_D$
- Speciate with probability  $p_S$
- Get lost with probability  $p_L$
- Get transferred to another species with probability  $p_T$





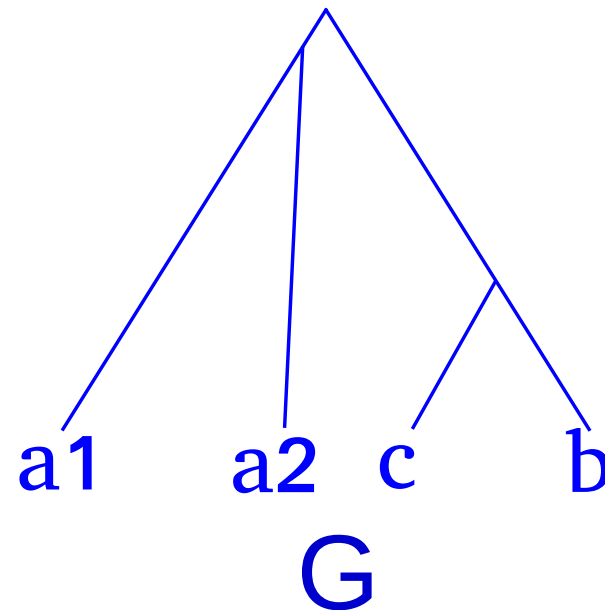
# Probability of a scenario

$$P(\text{scenario}) = P_s^6 * P_l * P_d * (P_t / \# \text{species})$$

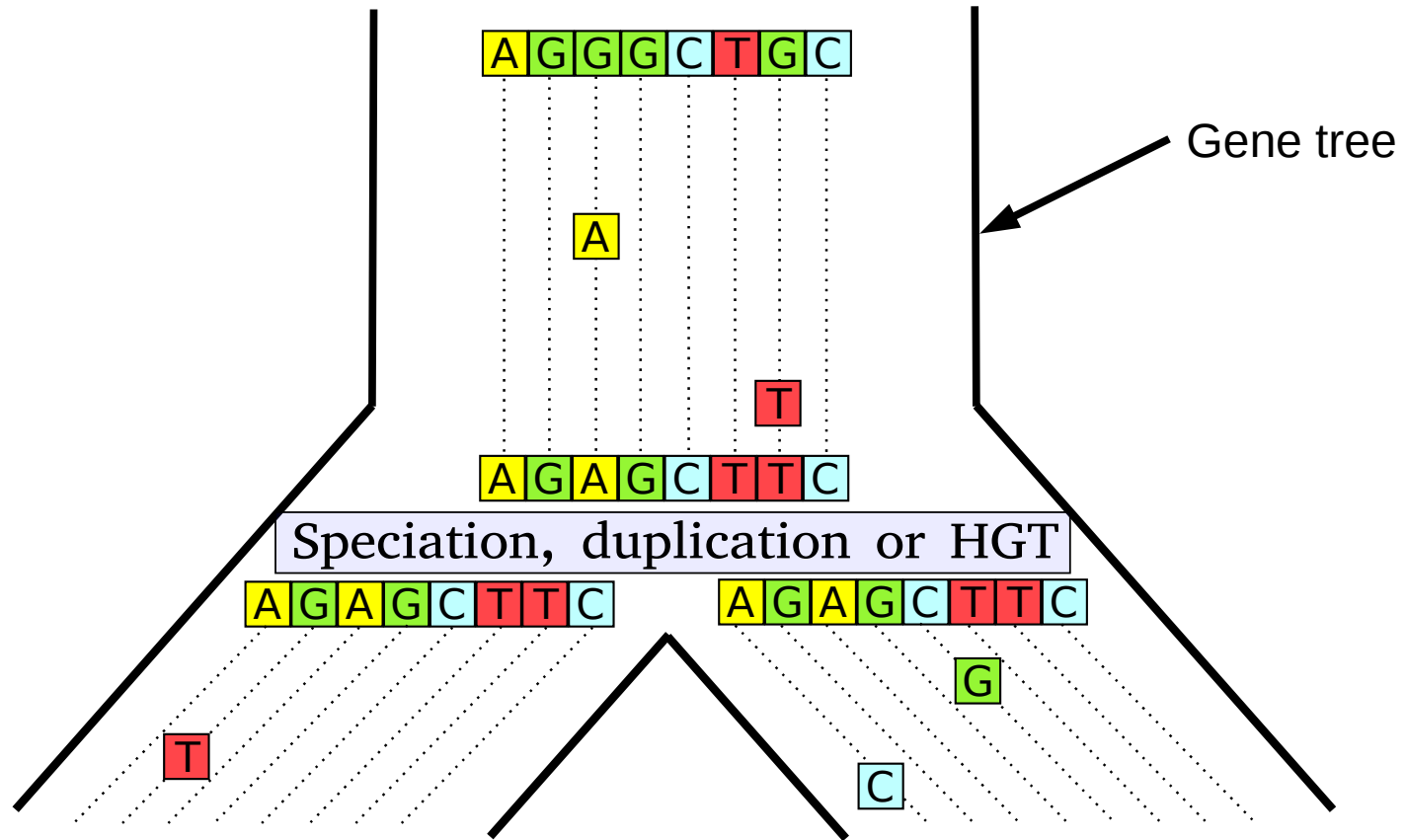


# The reconciliation likelihood

- $P(\mathbf{G}, \mathbf{S})$  = probability of observing a gene tree  $\mathbf{G}$  given a rooted species tree  $\mathbf{S}$
- Sum over all possible DTLs scenarios that would generate  $\mathbf{G}$



# Gene sequences evolve along a gene tree

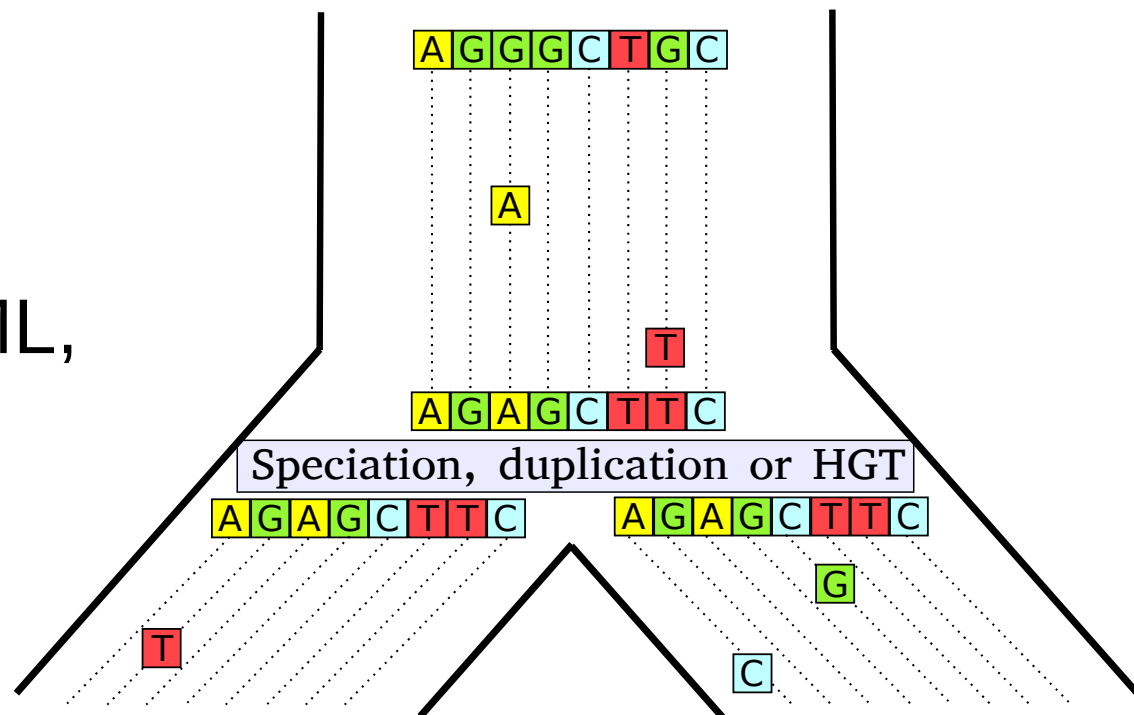


# The phylogenetic likelihood

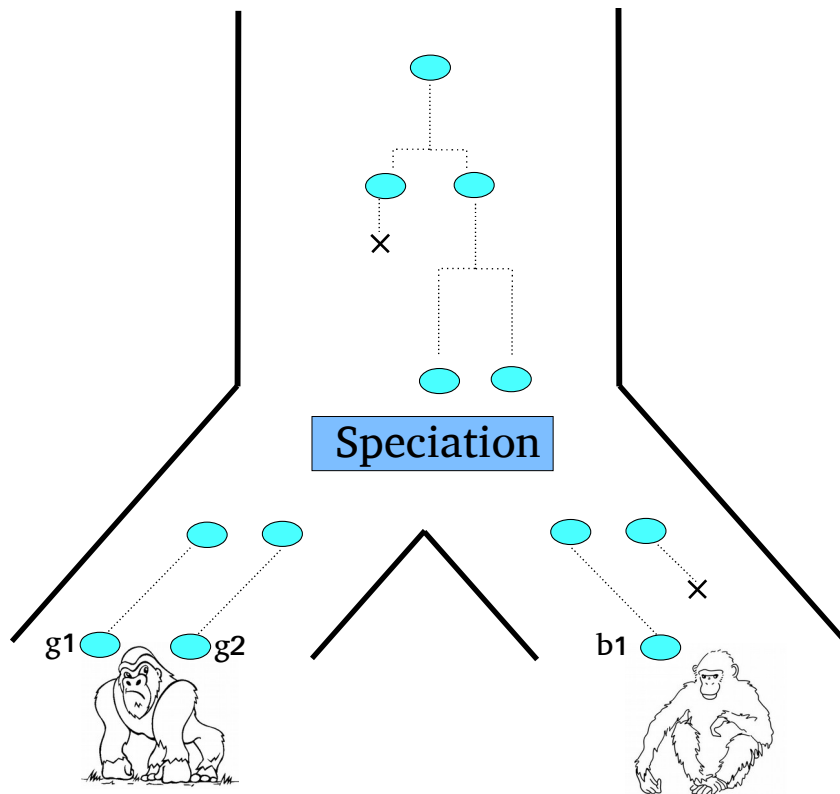
Sequence evolution models: JC, GTR, LG etc.

$P(\mathbf{A}, \mathbf{G})$  = probability of observing an alignment  $\mathbf{A}$  given a gene tree  $\mathbf{G}$

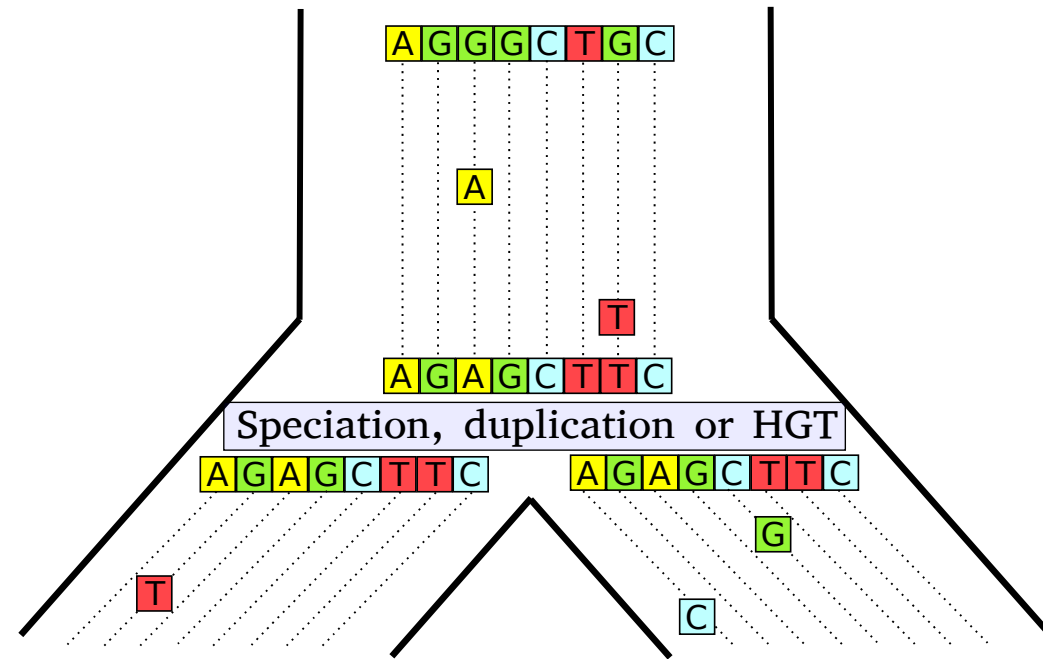
(Implemented in RaxML,  
IQTree, PhyML)



# The whole process

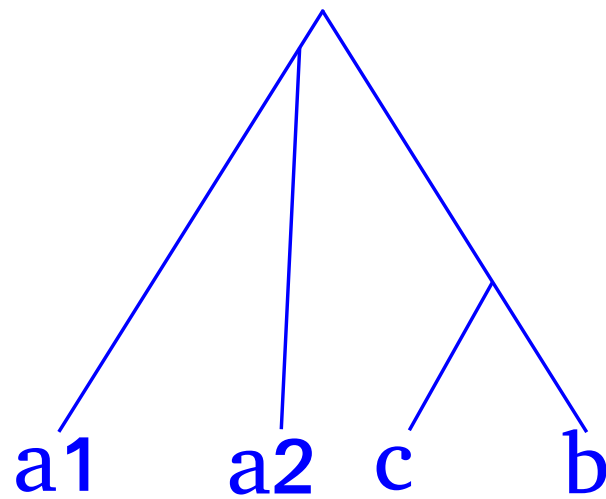


The gene tree evolves along the species tree

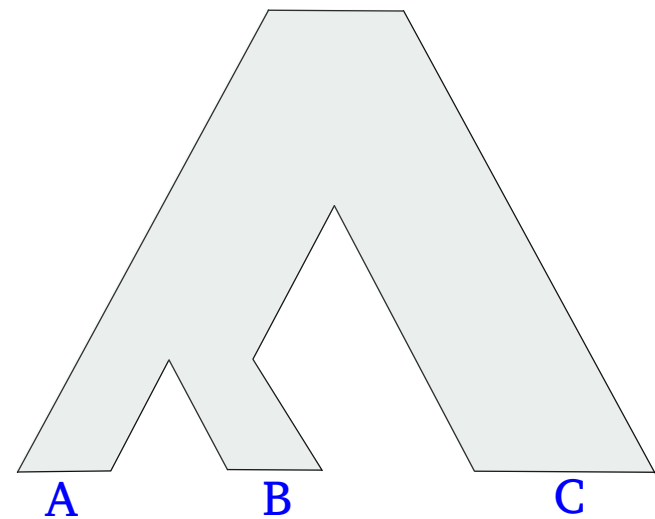


The sequences evolve along the gene tree

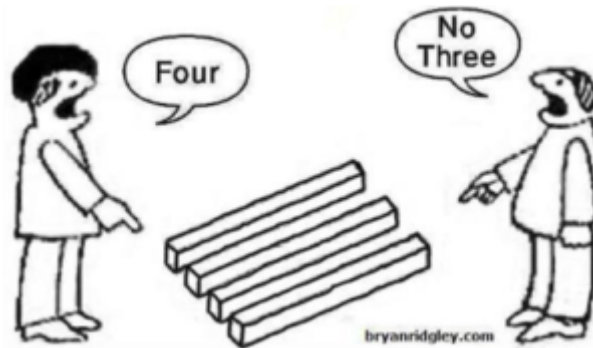
# Gene tree and species tree are different



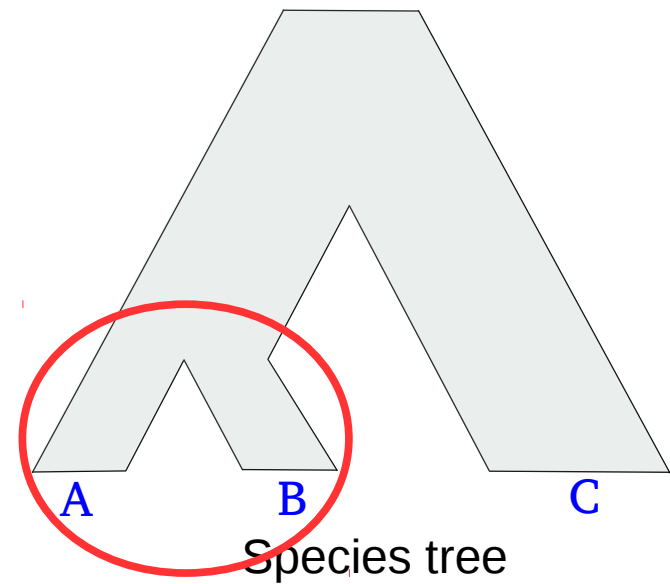
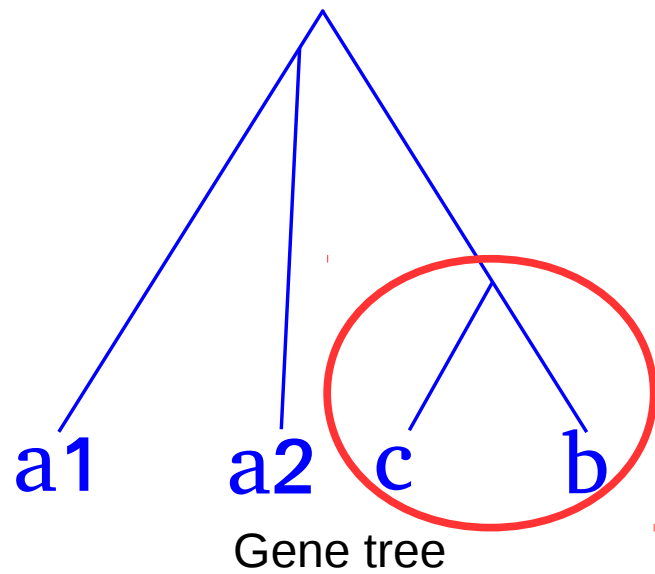
Gene tree



Species tree

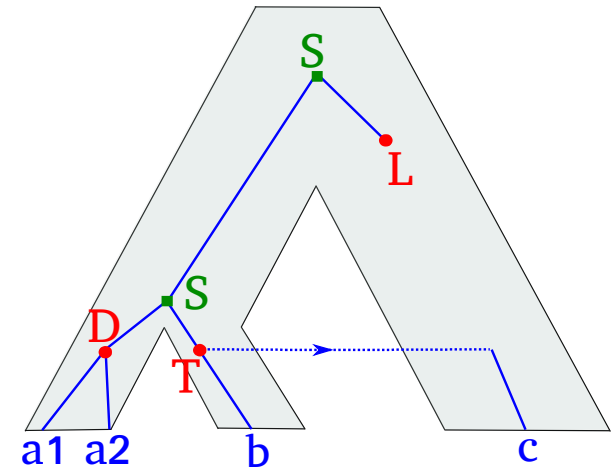


# Gene tree and species tree are different





# Gene tree and species tree are different

- Tree inference error
- DTL events
- (Others: Incomplete lineage sorting, hybridization etc.)



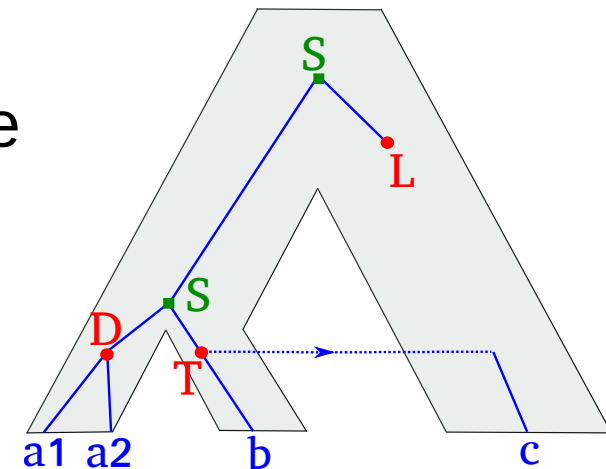
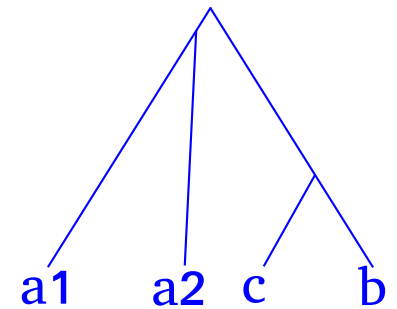
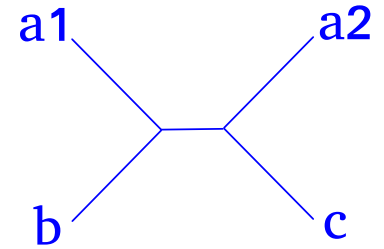


# Gene tree and species tree are different

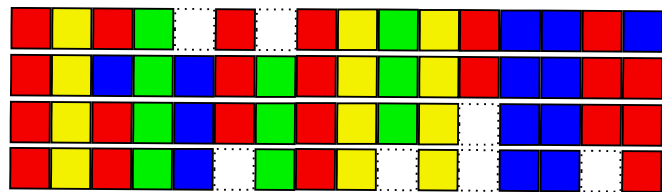
- Tree inference error  Need to be fixed!  
(Gene tree correction)
- DTL events  Need to be explained!  
(Gene tree reconciliation)

# Resolving the conflicts to recover the truth

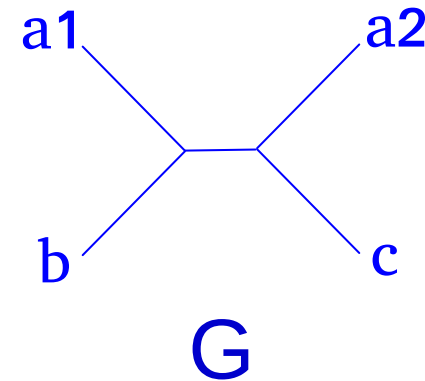
- **Gene tree inference:** infer a gene tree from the sequences
- **Gene tree correction:** correct the gene tree topology using the species tree
- **Gene tree – species tree reconciliation:** explain how the gene tree evolved within the species tree



# Gene tree inference (from the sequences)



A

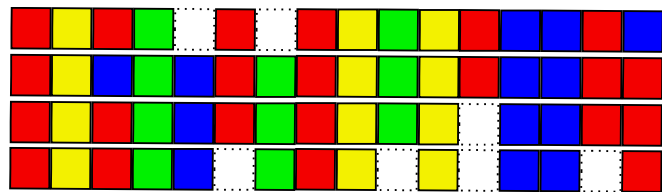


G

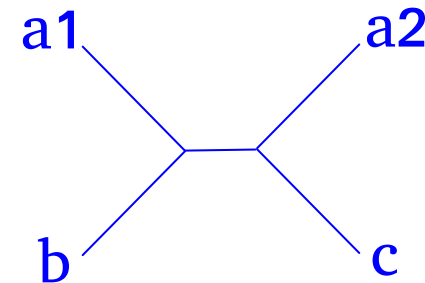
Maximize  $P(A|G)$

Standard software: RAxML, IQTree, PhyML, (FastTree)

# Gene tree inference (from the sequences)



A

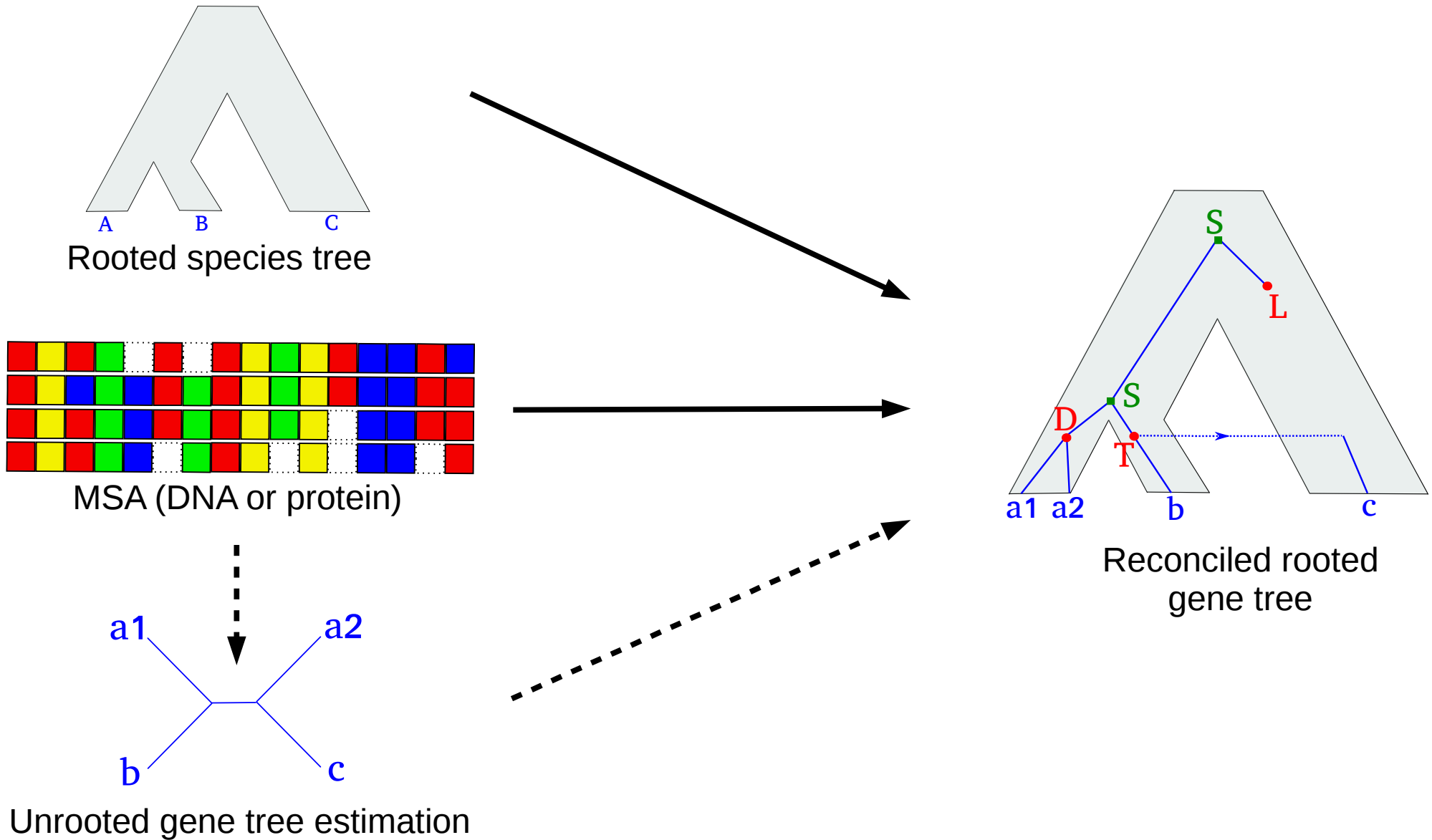


G

Limitations:

- Inaccurate (lack of signal)
- Does not infer DTL events

# Gene tree correction and reconciliation with a species tree



# Related work

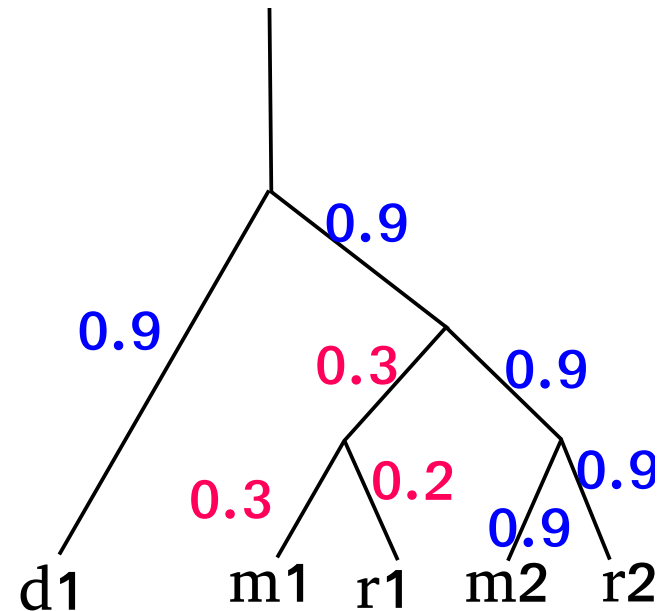
- Parsimony methods:
  - Notung, EcceTERA, Treerecs, Ranger-DTL etc.
- Bayesian methods:
  - ALE

# Related work: parsimony

- Infer a gene tree from the sequences
- Identify parts of the tree we do not trust
- Rearrange with parsimonuous reconciliation

A gene tree with support values

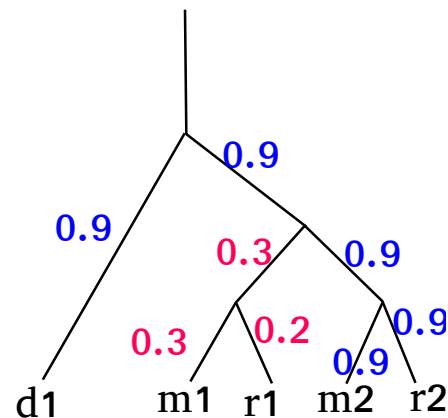
Low support values are in red



# Related work: parsimony

Tools: Notung, EcceTERA, Ranger-DTL etc.

- the correction step is very fast
- BUT inferring support values is slow
- how to pick the support value threshold??





# Related work: bayesian

- Infer a gene tree distribution from the sequences (MrBayes, ExaBayes etc.)
- Sample gene trees from this distribution under the UndatedDTL model

# Related work: bayesian

Tool: ALE

- More accurate than parsimony methods
- BUT inferring the gene tree distribution is very **slow**

# Maximum likelihood

**Parsimony** → not accurate

→ expansive precomputations

**Bayesian** → very expansive precomputations

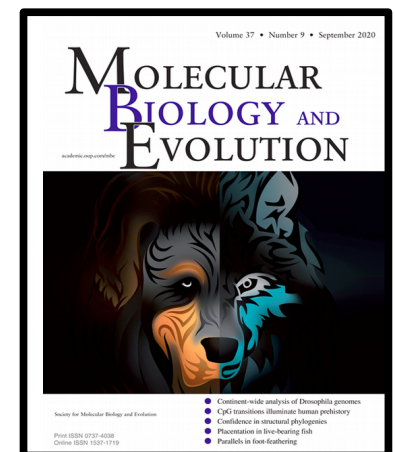
What about trying **maximum likelihood**?



Our solution: maximum likelihood method  
accounting for both sequences and species  
tree

- more accurate than parsimony
- does not need any slow precomputation step

Published in MBE

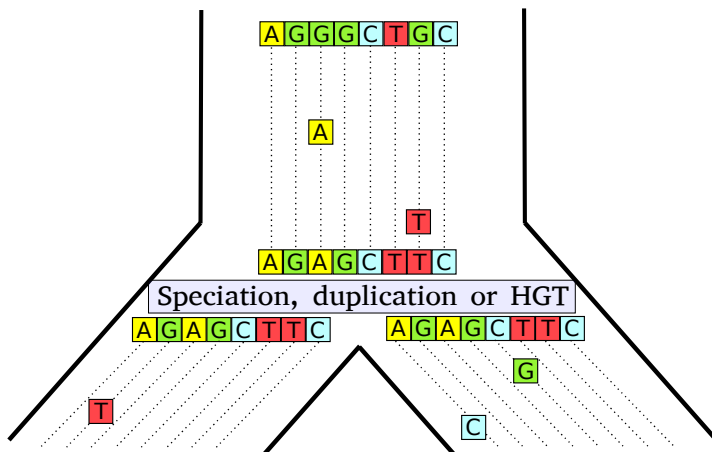


# GeneRax

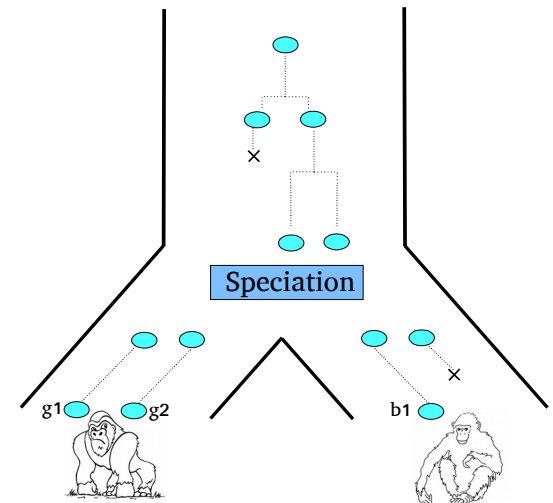
Maximizes the joint likelihood:

$$L(G|S) = P(A|G) P(G|S)$$

Phylogenetic likelihood



Reconciliation likelihood

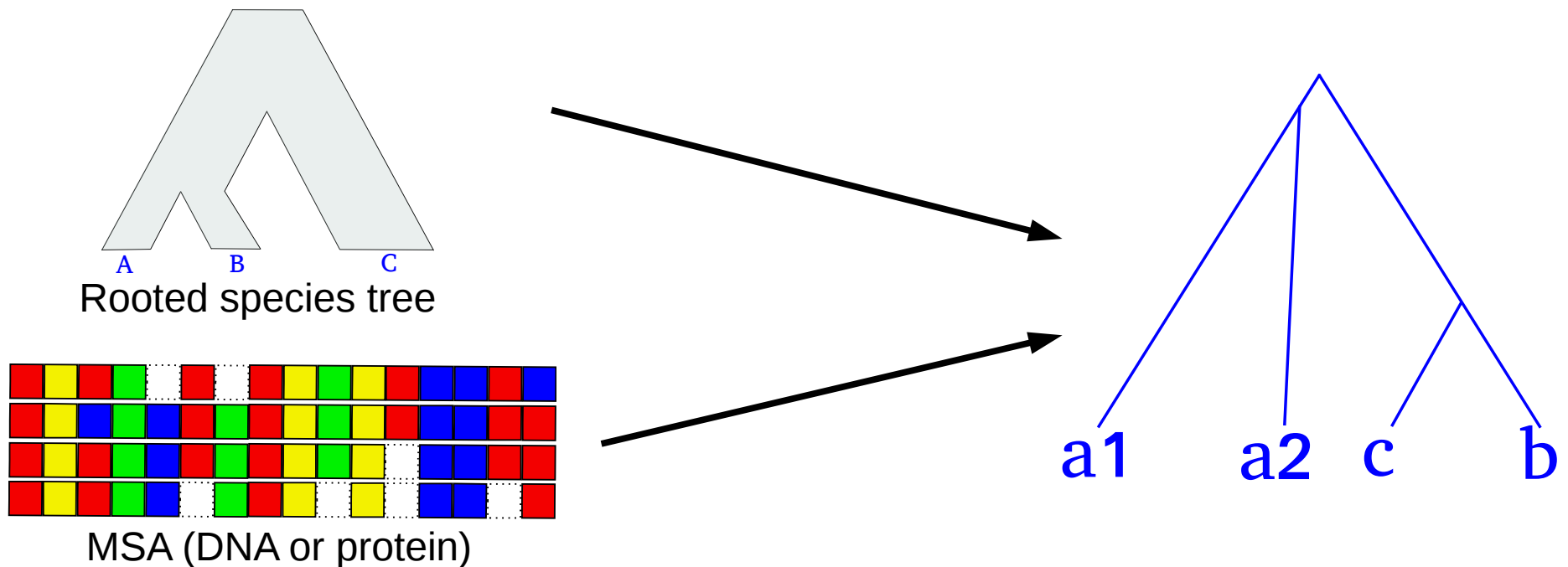


# Tree search heuristic

- Start from an initial gene tree
- Apply small changes to the tree and keep the ones that improve the likelihood
- Stop when we cannot find any better tree

# GeneRax

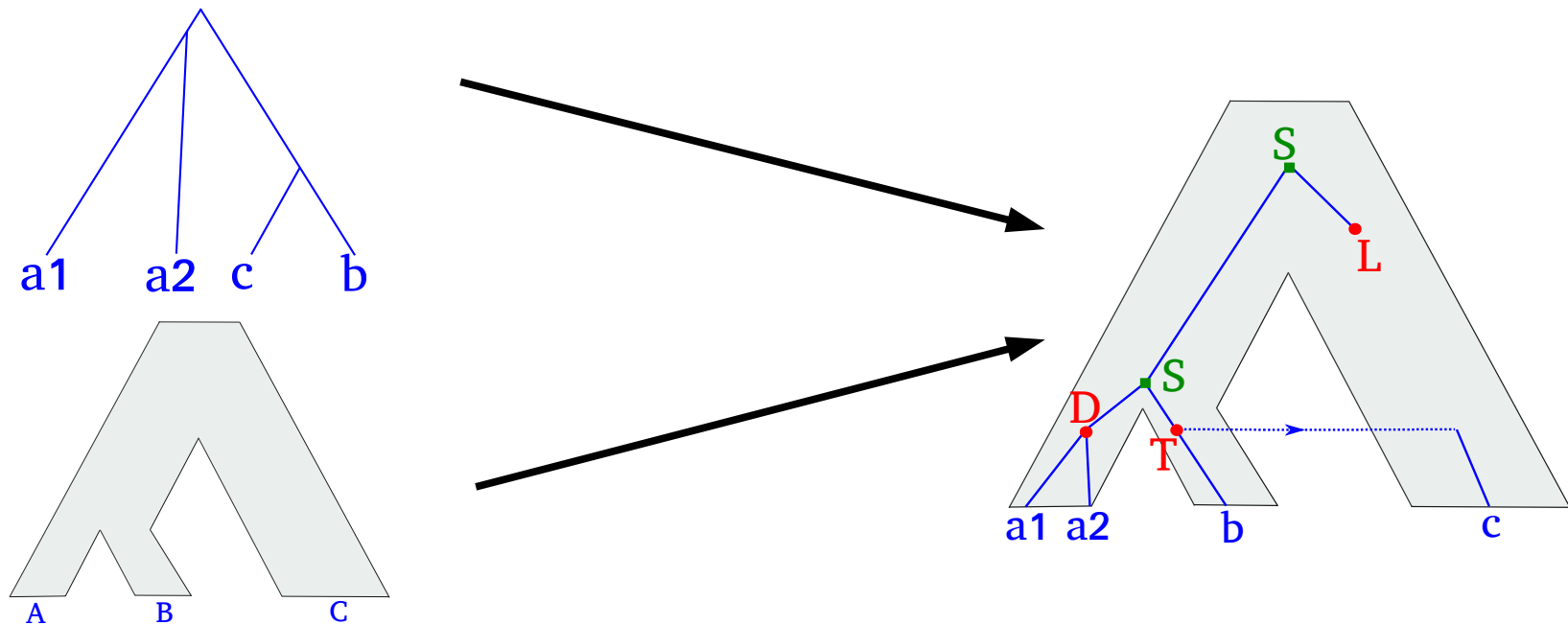
Result: the gene tree that is the “best” compromise between the signal from the sequences and the signal from the species tree



# Reconciliation with GeneRax

Once we inferred the best gene tree, we can either return:

- the maximum likelihood scenario
- a stochastic sample of plausible scenarios

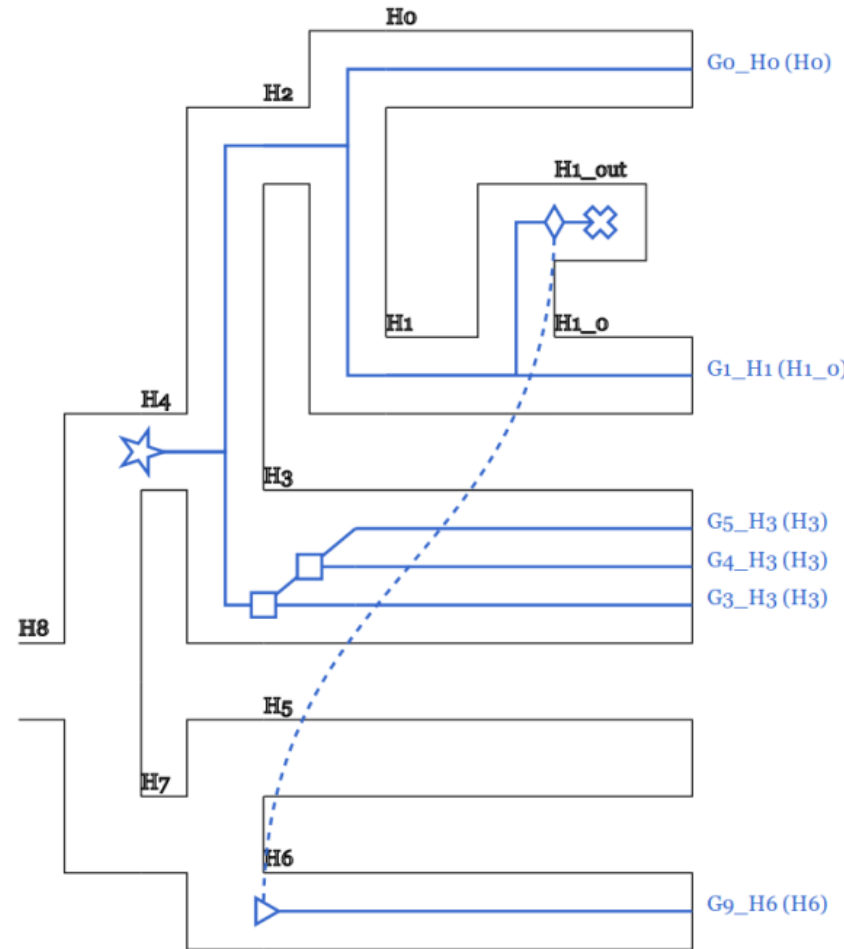




# Displaying reconciliations

Output in RecPhyloXML

Can be viewed with  
RecPhyloVisu



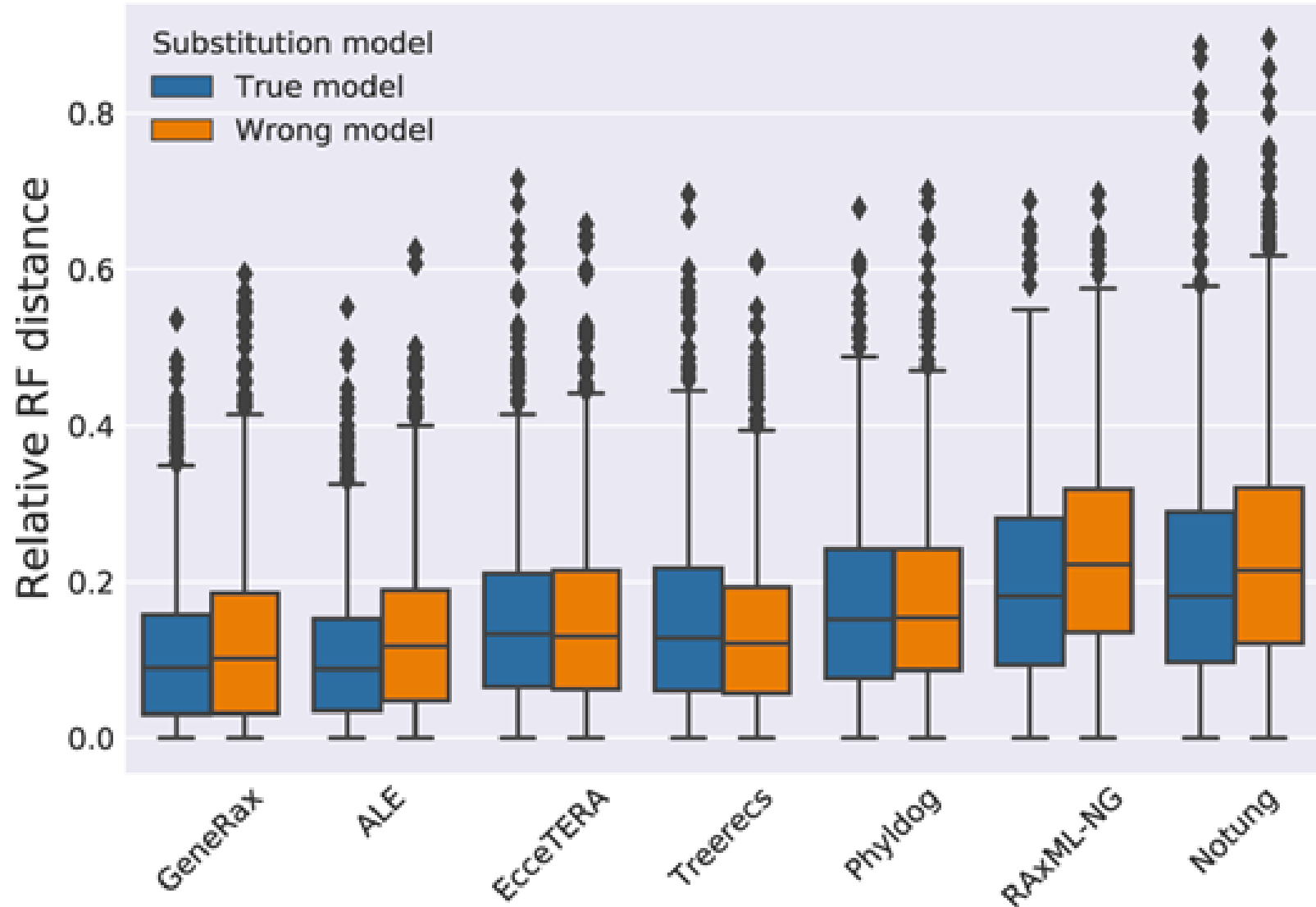
<http://phylariane.univ-lyon1.fr/recphyloxml/recphylovisu>

# GeneRax is suitable for large analyses

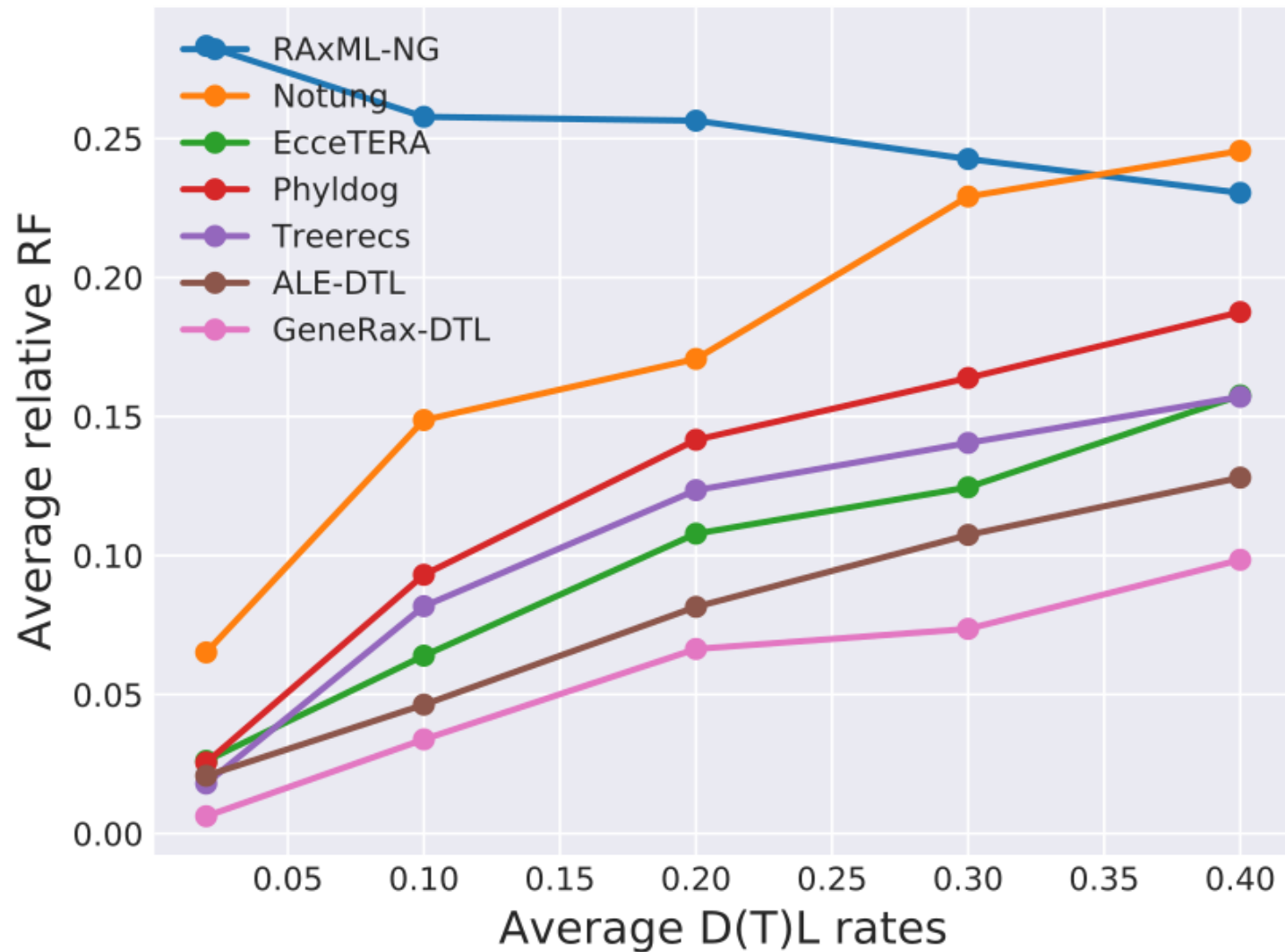
- Process single or multiple gene families in parallel with MPI
  - can also be deployed on a cluster
- Checkpoint system (a run can be restarted after an interruption)

# Results (accuracy)

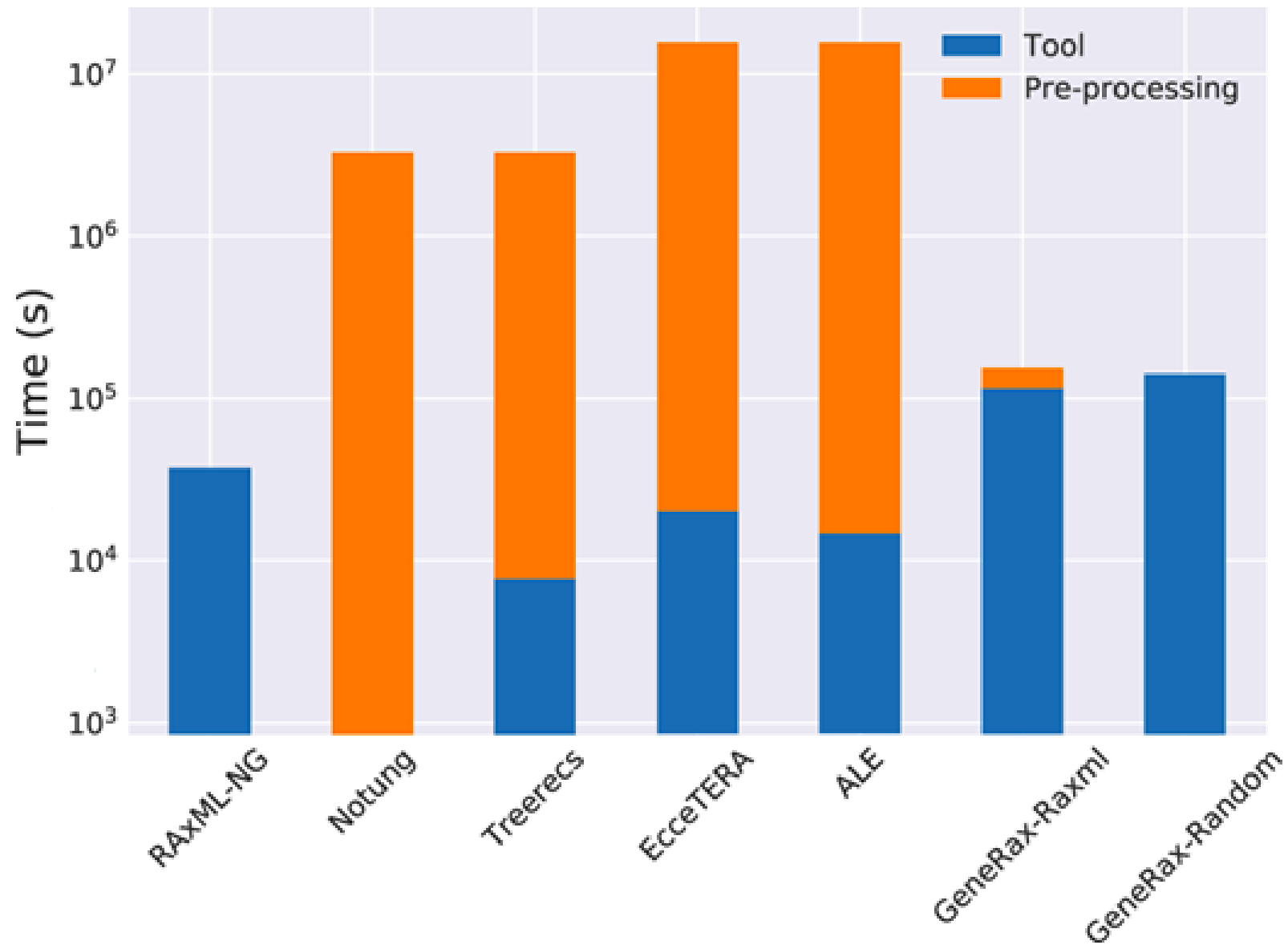
Simulated cyanobacteria (1099 families)



# Results (accuracy)



# Results (runtime)



# GeneRax strenghts

- fast and accurate
- mathematically satisfying
- suitable for large analyses
- simplifies pipelines
- actively maintained
- available on github and bioconda

# Future work on GeneRax

- Infer the species tree from gene trees
- Provide post-analysis tools
- Ask the users what they would like to have!

# Thanks!

Contact: [benoit.morel@h-its.org](mailto:benoit.morel@h-its.org)

GeneRax on Github:

<https://github.com/BenoitMorel/GeneRax/>

GeneRax on BioConda:

<https://anaconda.org/bioconda/generax>

