Gene tree - species tree models, with a focus on gene duplications, transfers, and losses

Bastien Boussau

Laboratoire de Biométrie et Biologie Évolutive

Université de Lyon

CoME, 2022

Plan

I. The need for gene tree species tree models

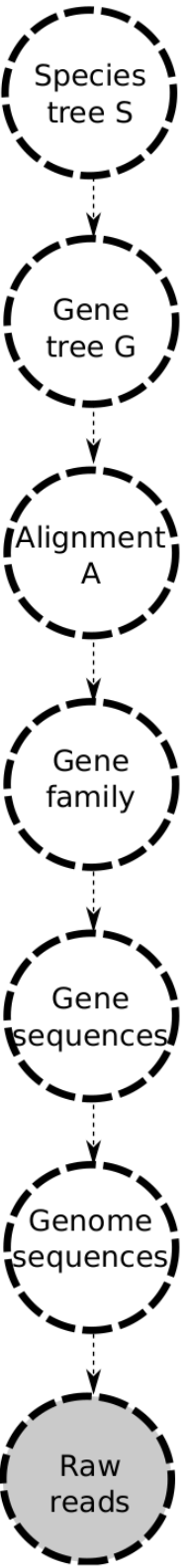
II. Using a model of gene transfers, duplications and losses (DTL) to reconstruct species and gene trees

III. Using a DTL model to compare genome evolution in Fungi vs Cyanobacteria

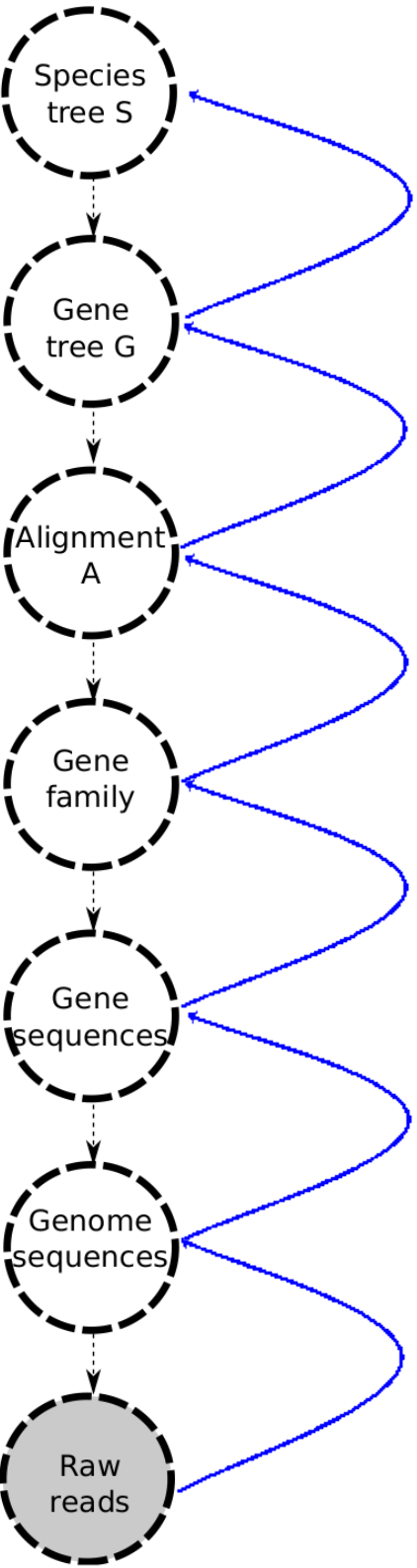
IV. Using a DTL model to reconstruct ancestral genomes in Archaea

V. Using a DTL model to date species phylogenies

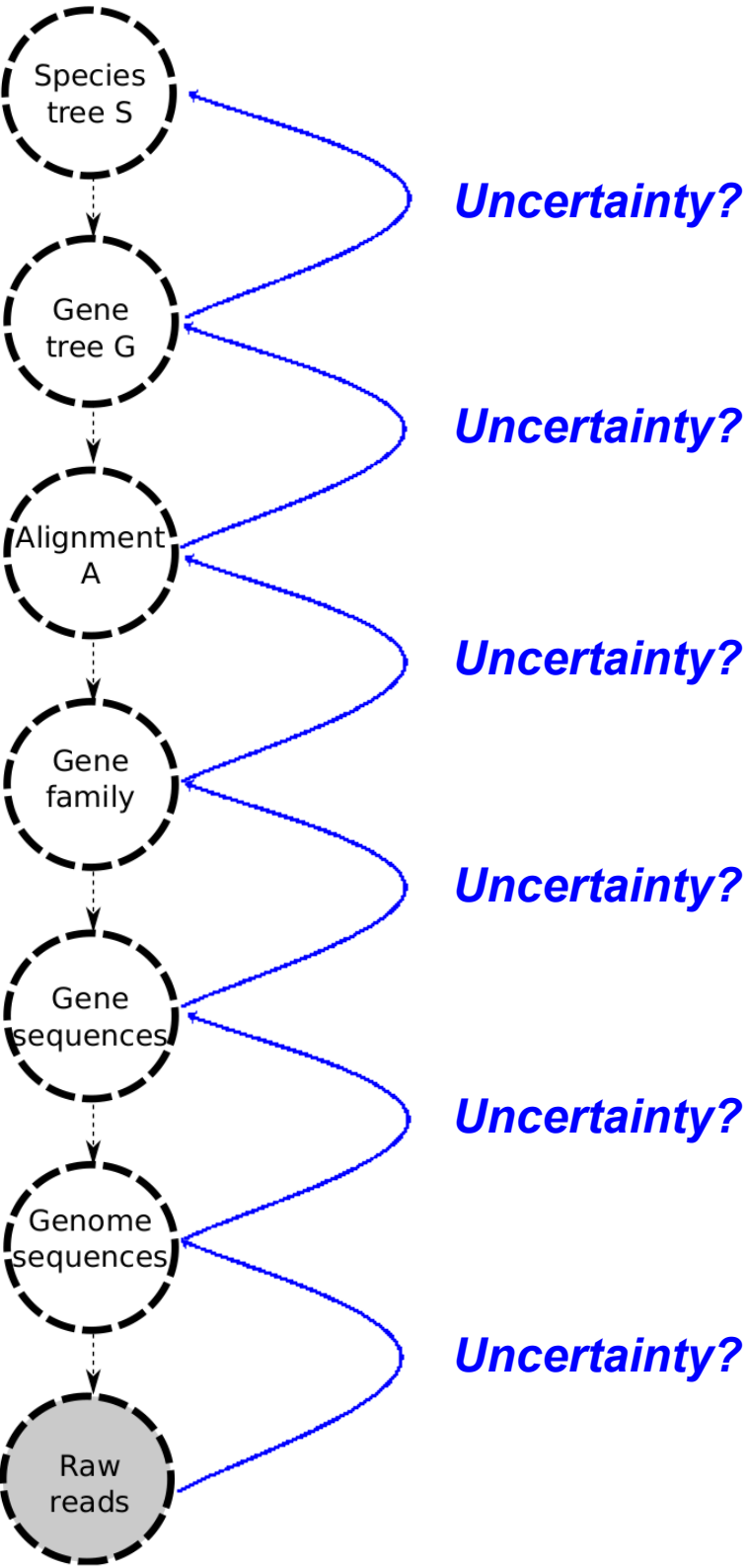
The phylogenomics pipeline

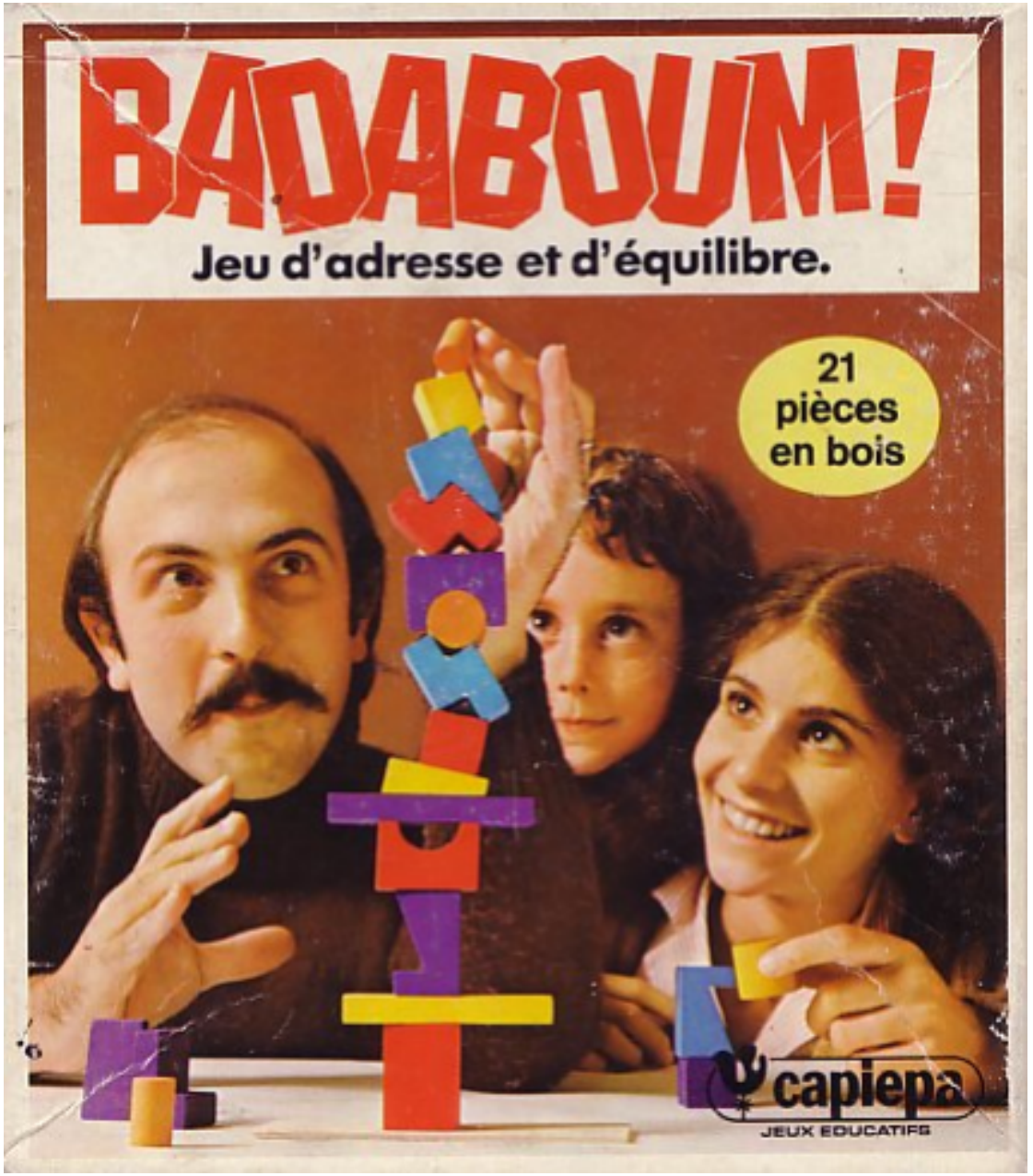


The phylogenomics pipeline

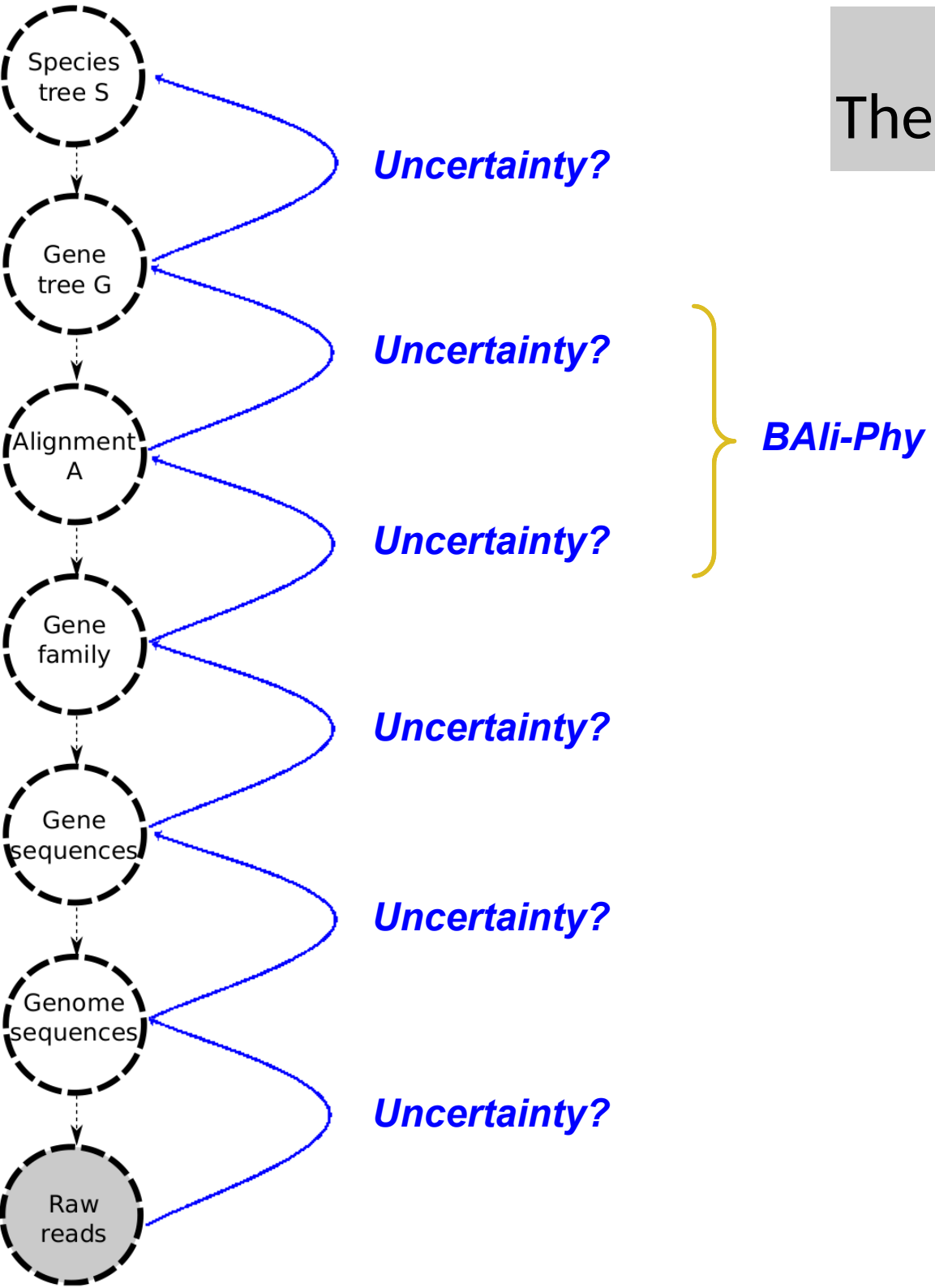


The phylogenomics pipeline

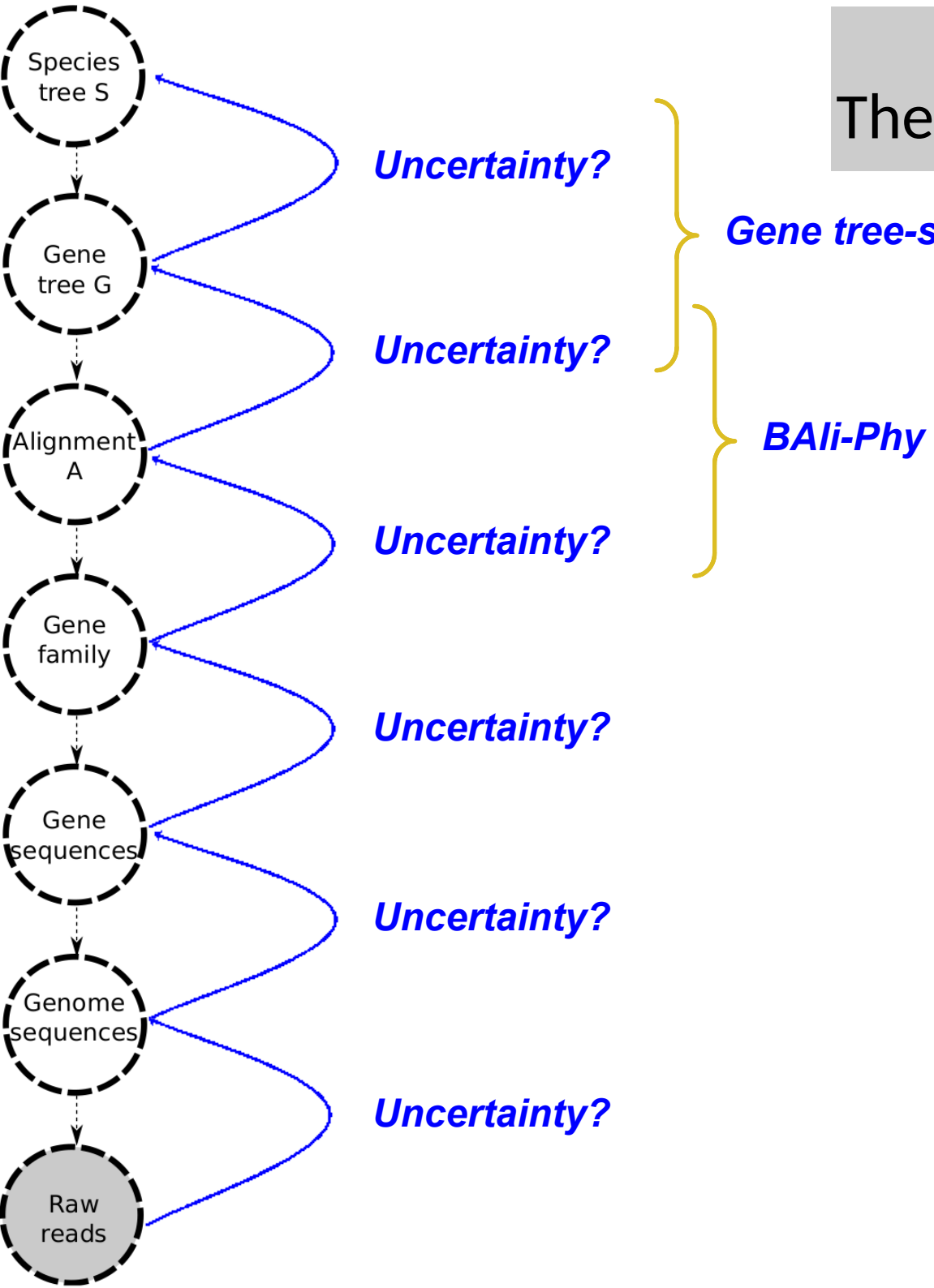




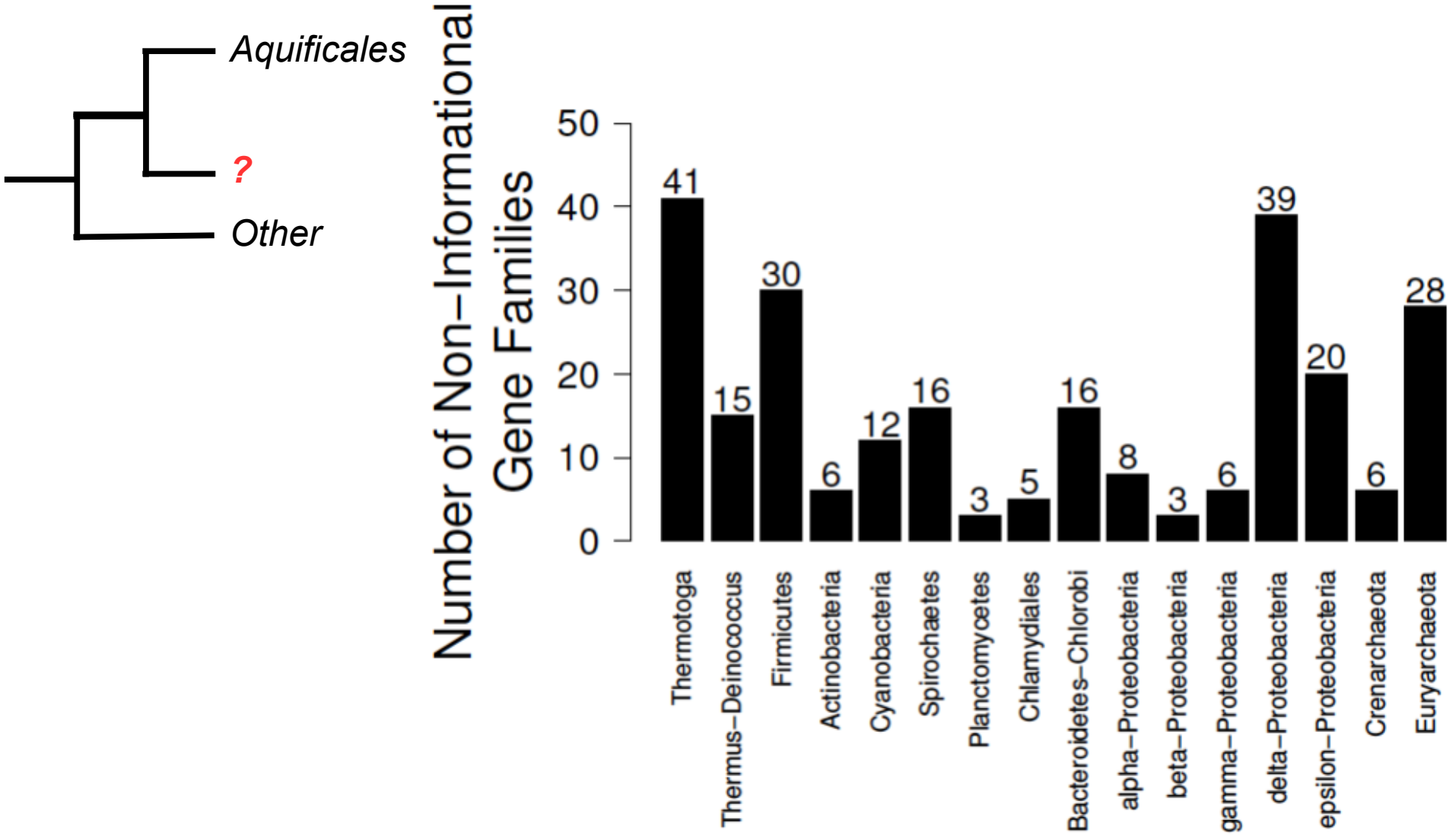
The phylogenomics pipeline



The phylogenomics pipeline



Gene trees provide confusing signal about species relationships



HOGENOM database

Boussau, Gueguen, Gouy, BMC Evol. Biol. 2008

The need for a gene tree-species tree model

- There is a lot of incongruence between gene trees, and therefore between gene trees and the species tree
- To know how much of it is biological and how much of it is methodological, we need to model biological processes, notably gene transfers

Incongruence ~ *Biological* **+** *Methodological*

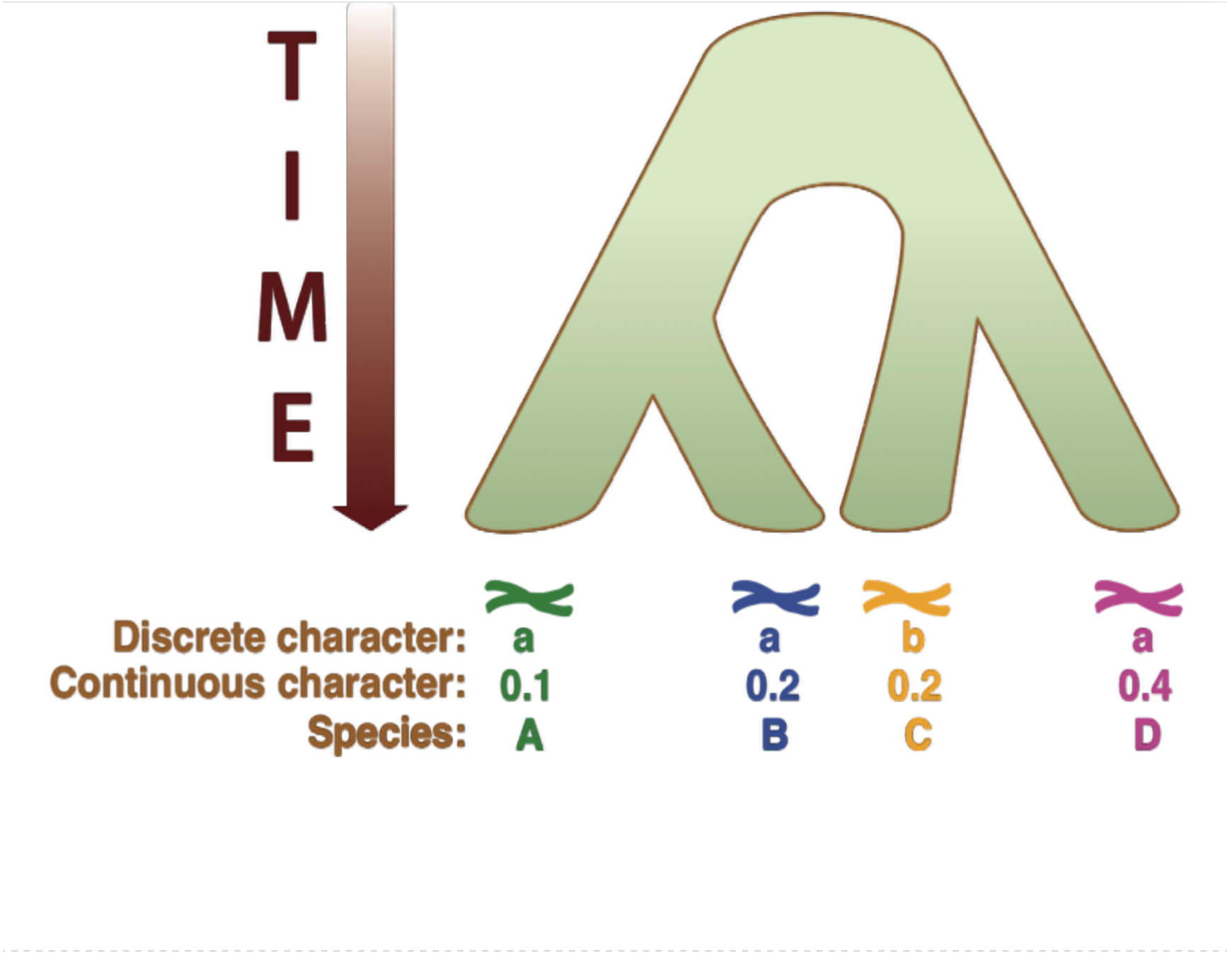
The need for a gene tree-species tree model

- There is a lot of incongruence between gene trees, and therefore between gene trees and the species tree
- To know how much of it is biological and how much of it is methodological, we need to model biological processes, notably gene transfers

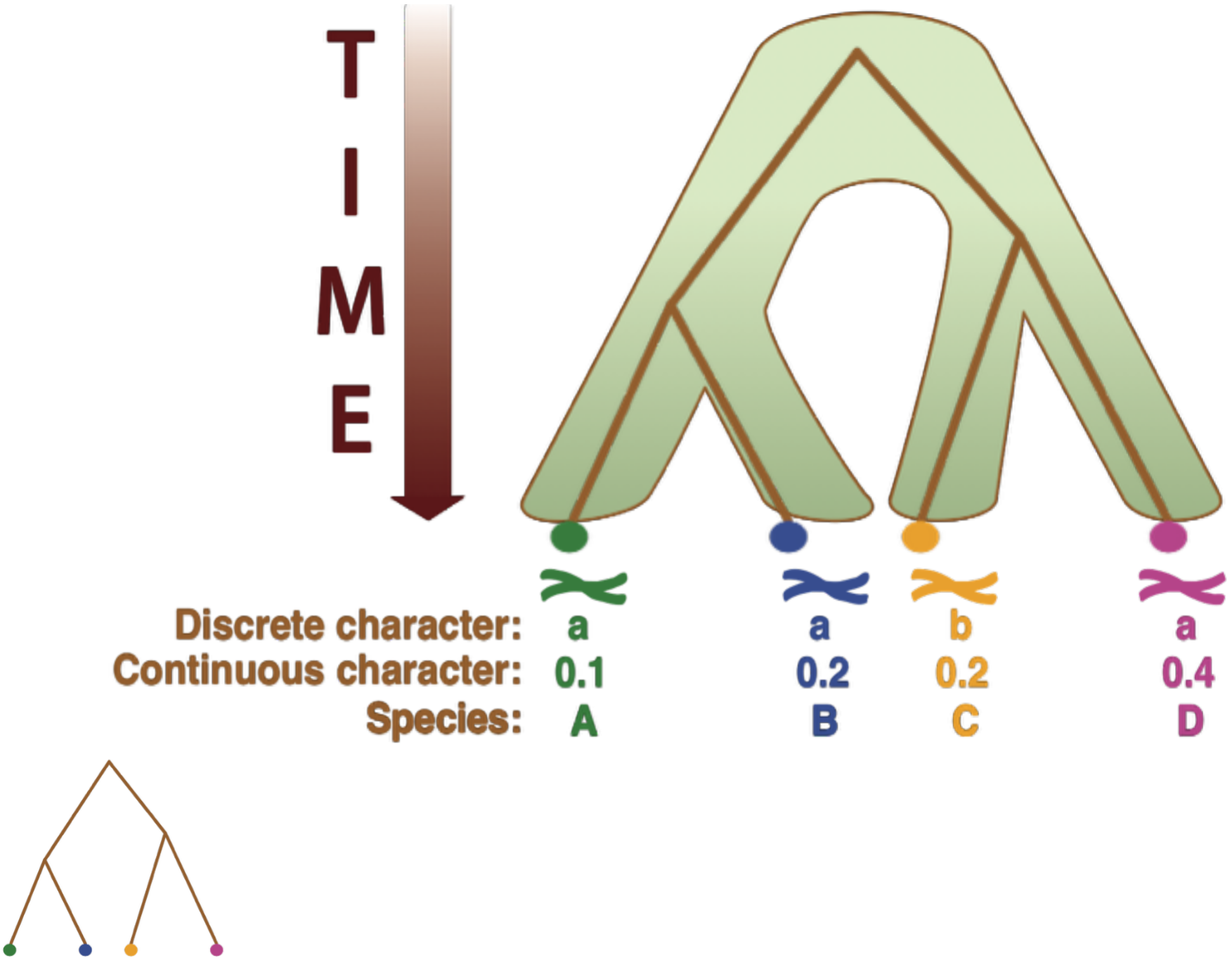
Incongruence ~ *Biological* **+** *Methodological*

Incongruence ~ Duplications + losses + Transfers + ILS + Error

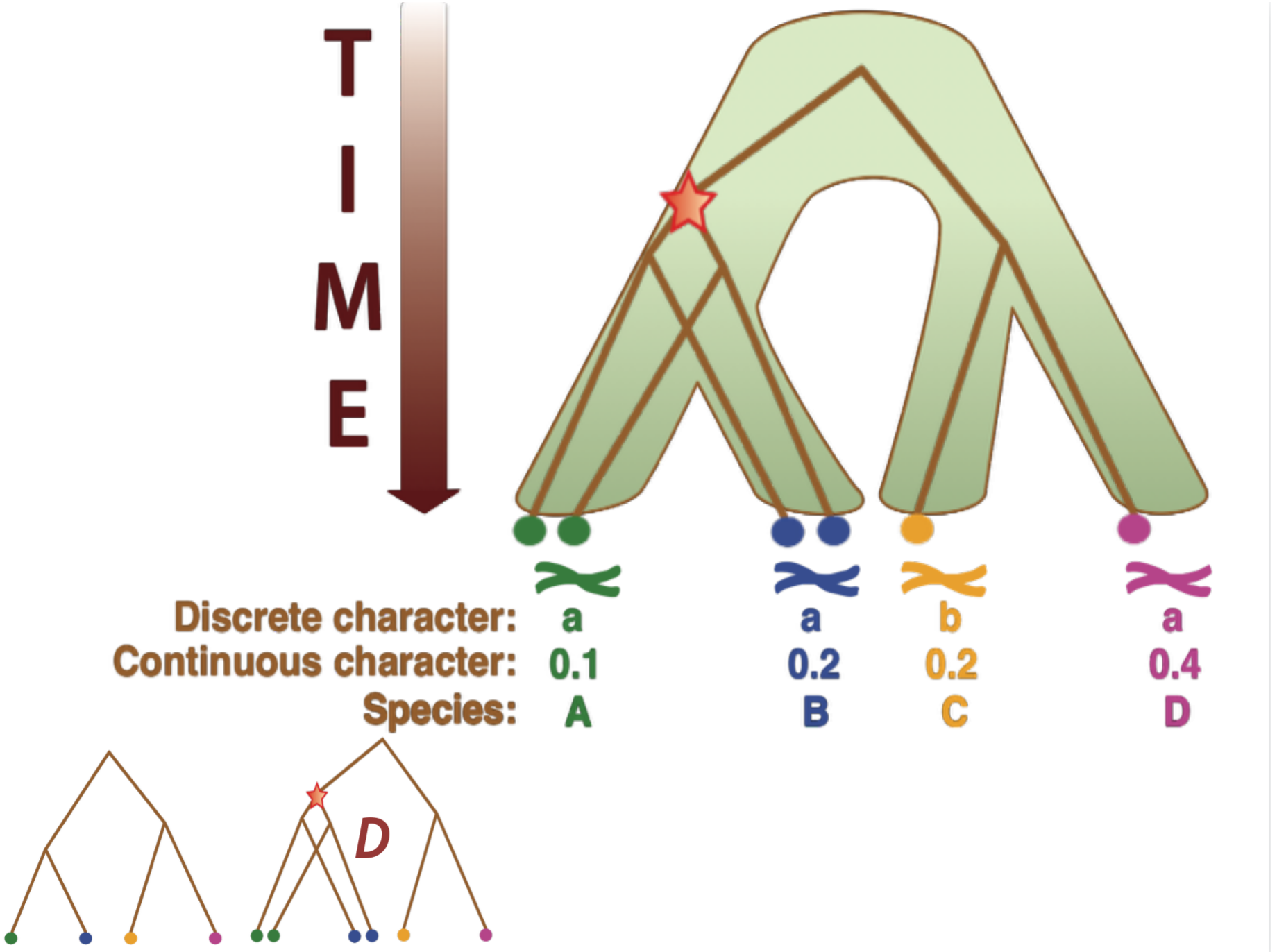
Biological processes creating incongruence



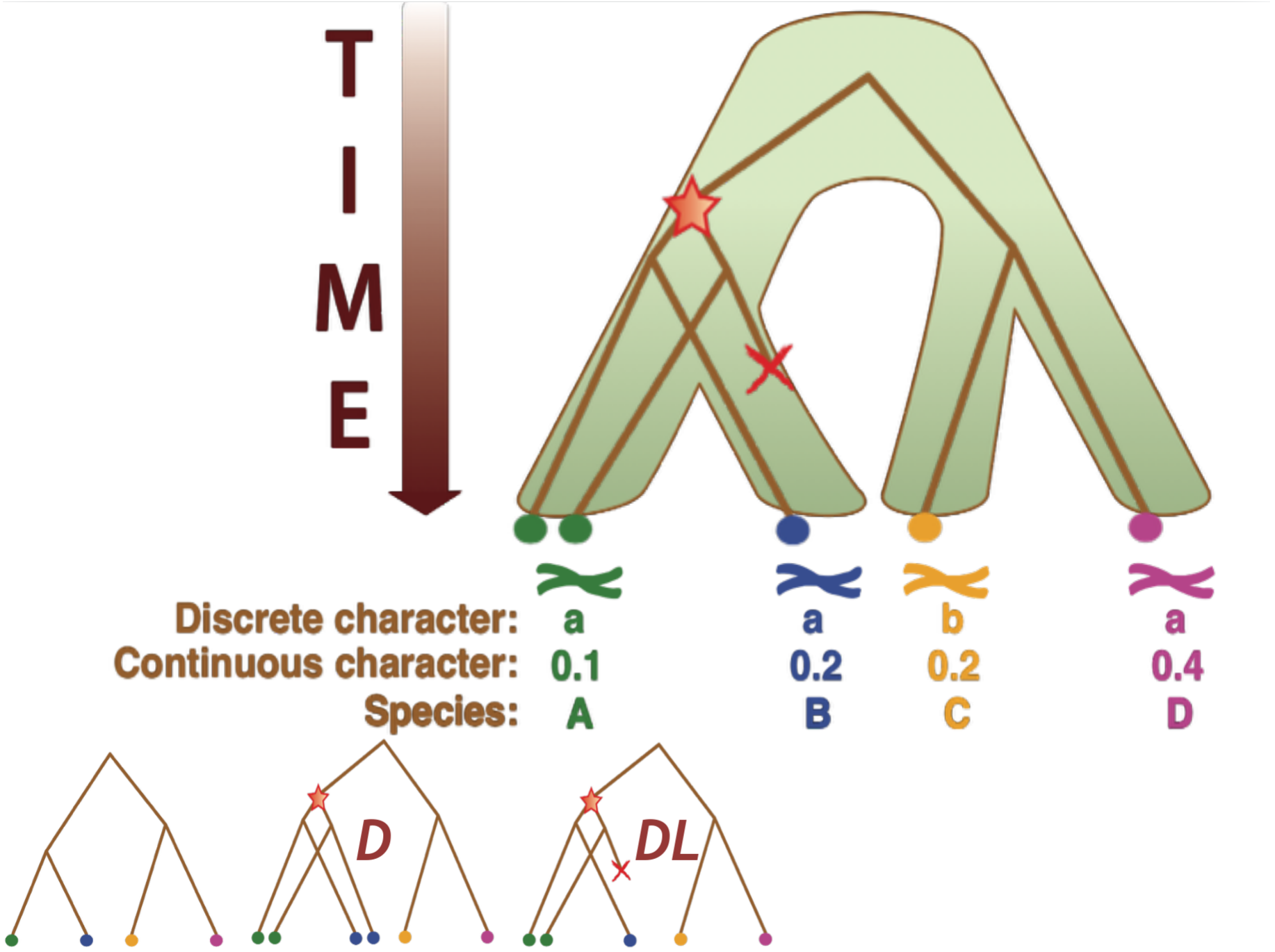
Biological processes creating incongruence



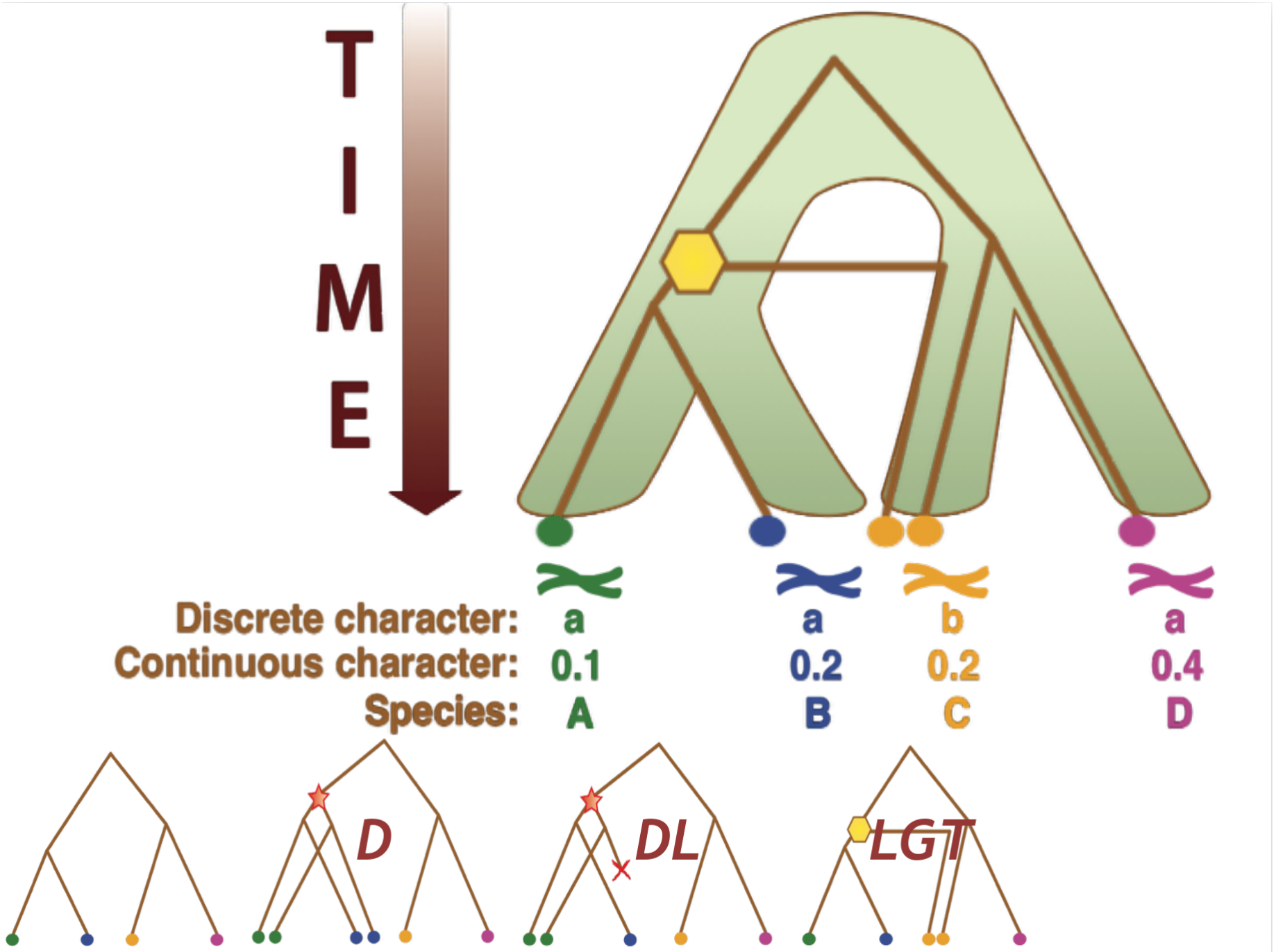
Biological processes creating incongruence



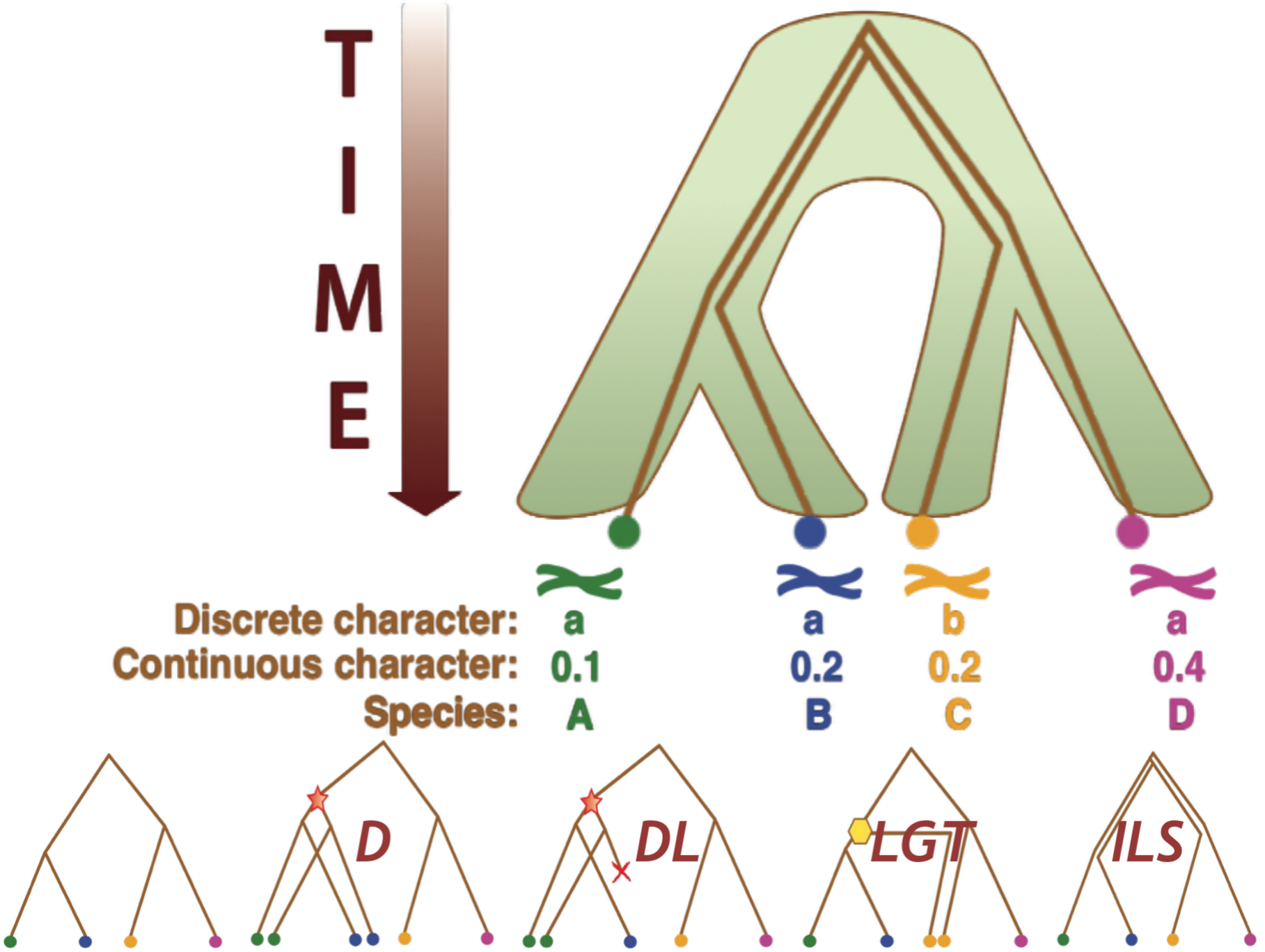
Biological processes creating incongruence



Biological processes creating incongruence



Biological processes creating incongruence



Plan

I. The need for gene tree species tree models

II. Using a model of gene transfers, duplications and losses (DTL) to reconstruct species and gene trees

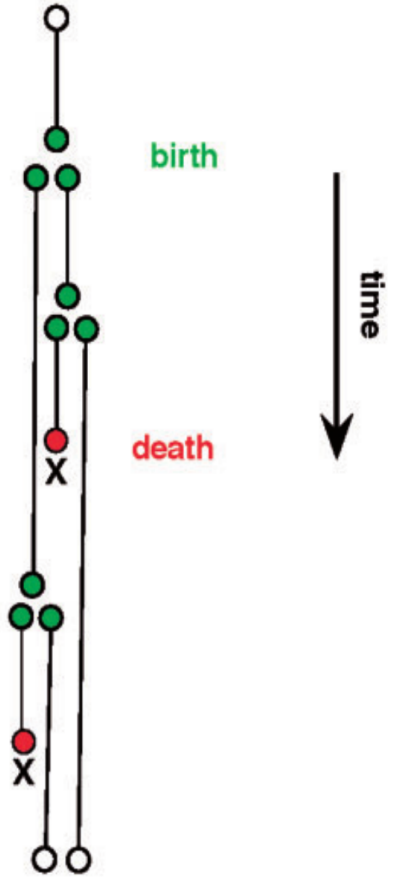
III. Using a DTL model to compare genome evolution in Fungi vs Cyanobacteria

IV. Using a DTL model to reconstruct ancestral genomes in Archaea

V. Using a DTL model to date species phylogenies

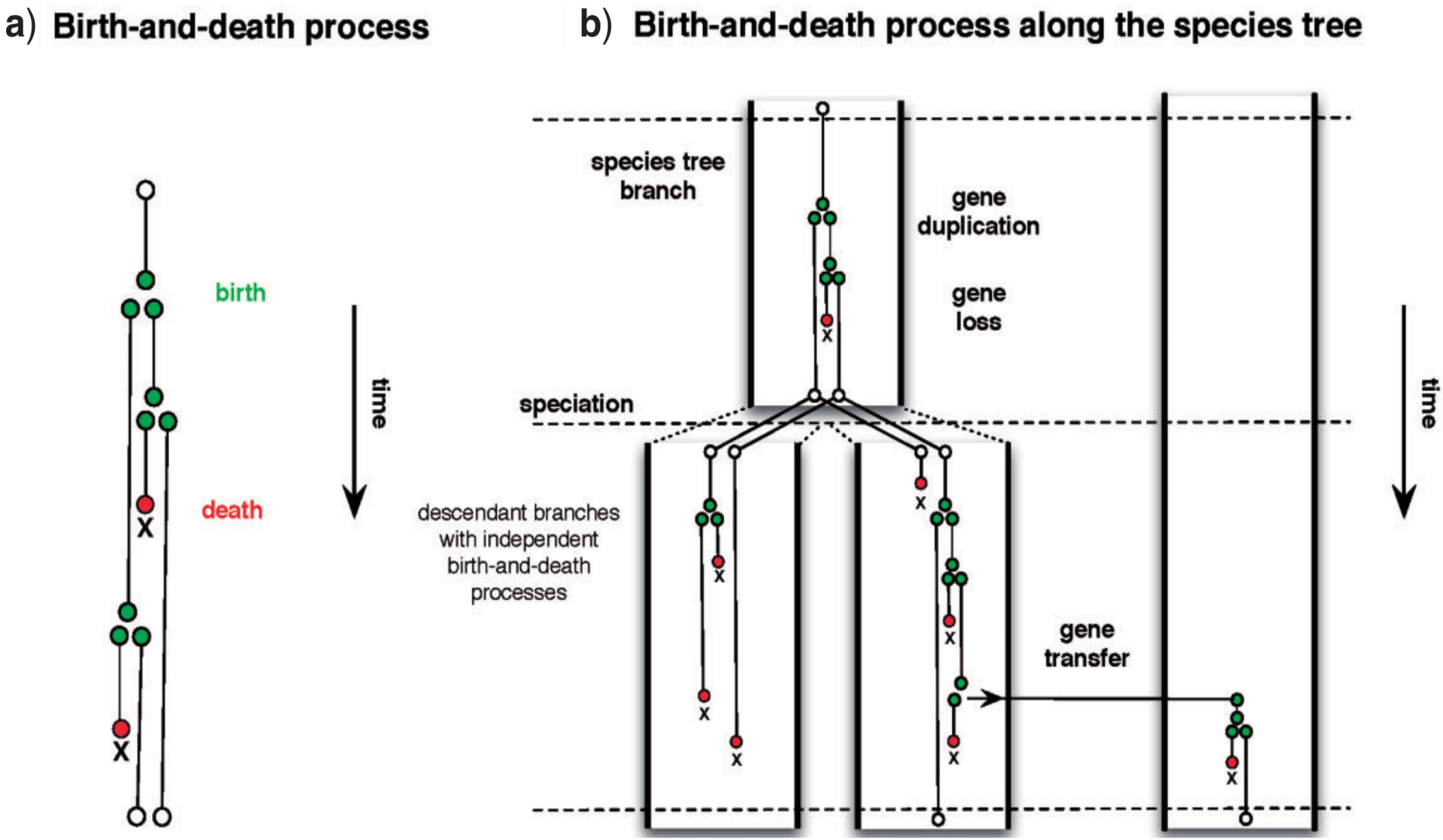
Modelling gene family evolution within a species tree

a) Birth-and-death process



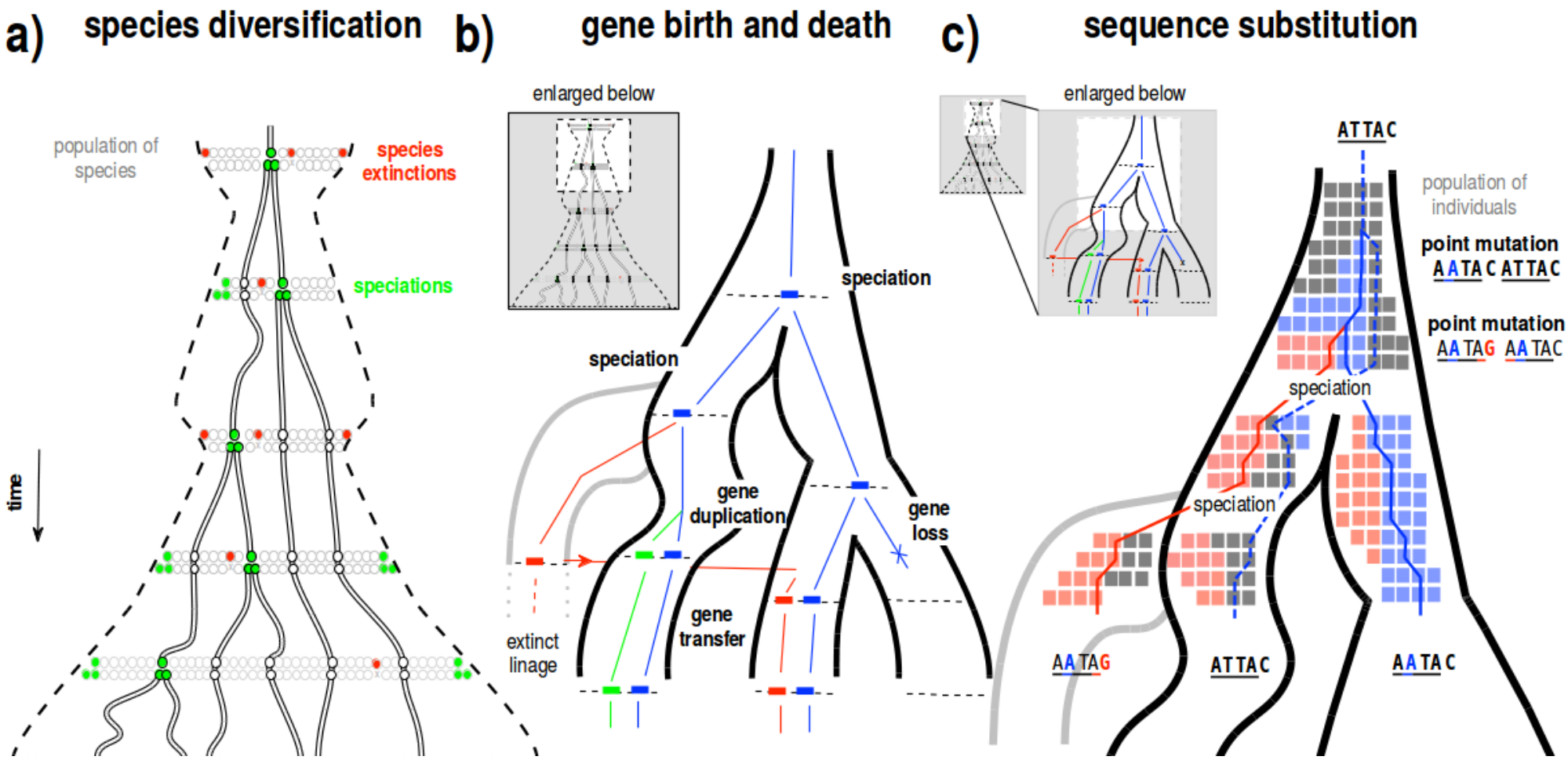
Szöllősi et al. Syst. Biol. 2015

Modelling gene family evolution within a species tree



Szöllősi et al. Syst. Biol. 2015

Hierarchical model of sequence evolution



Szöllősi et al. Syst. Biol. 2015

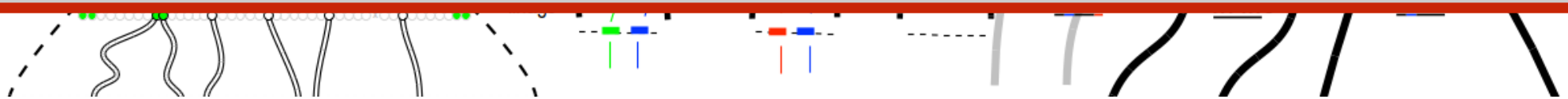
Hierarchical model of sequence evolution

a) species diversification b) gene birth and death c) sequence substitution

****Caveat****

Lots of reconstruction methods do not use dated trees :

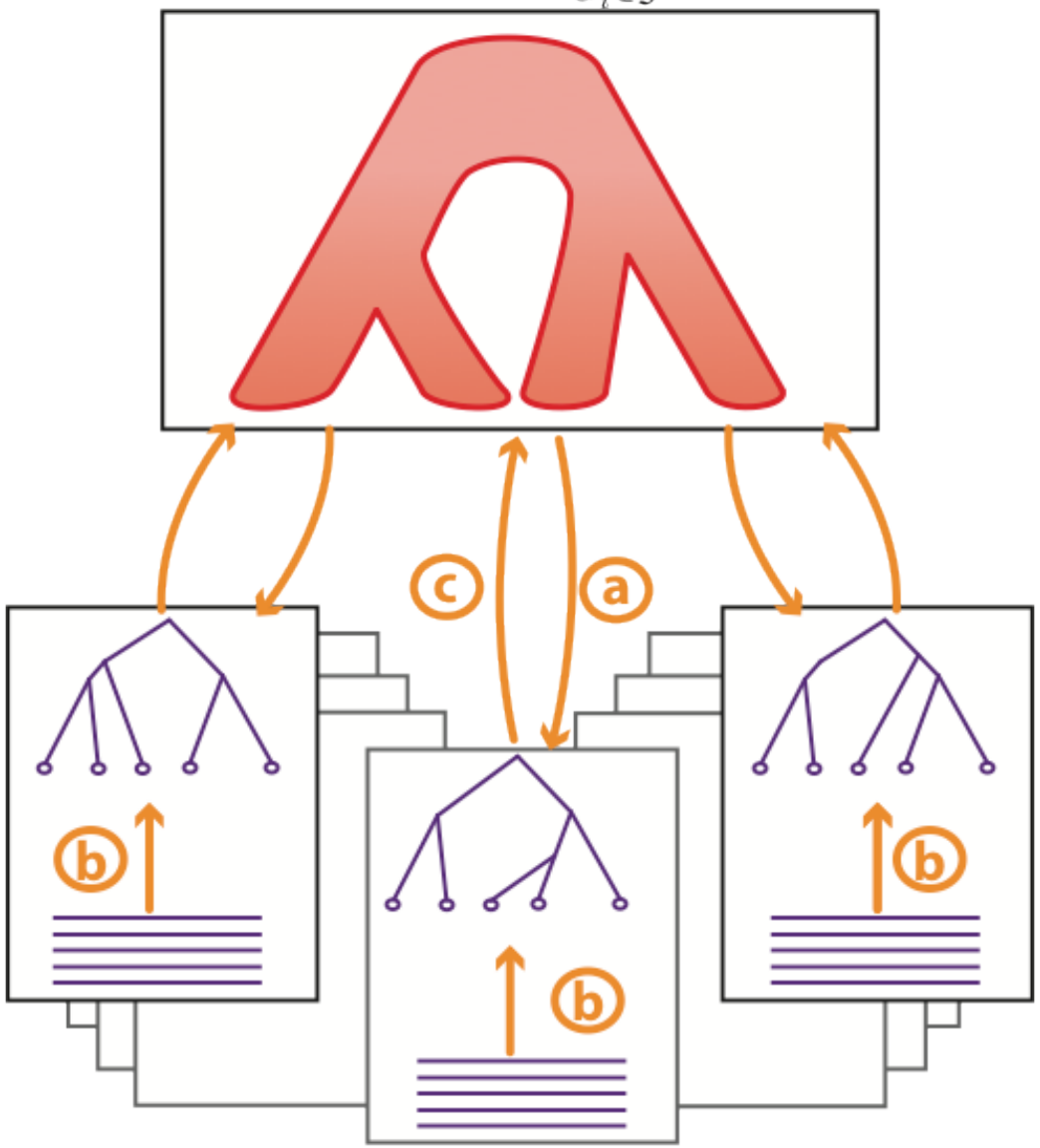
- *they allow placing DTL events on branches of the species tree*
- *they do not date them in time*



Szöllősi et al. Syst. Biol. 2015

Inference with gene tree-species tree models

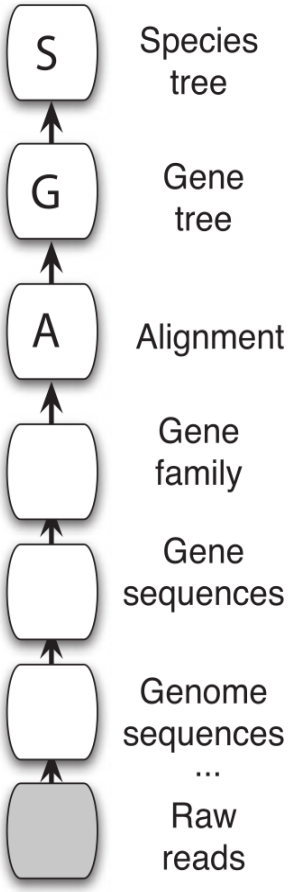
$$L(T, S, N|A) = \prod_{G_i \in \mathcal{G}} L(G_i)$$



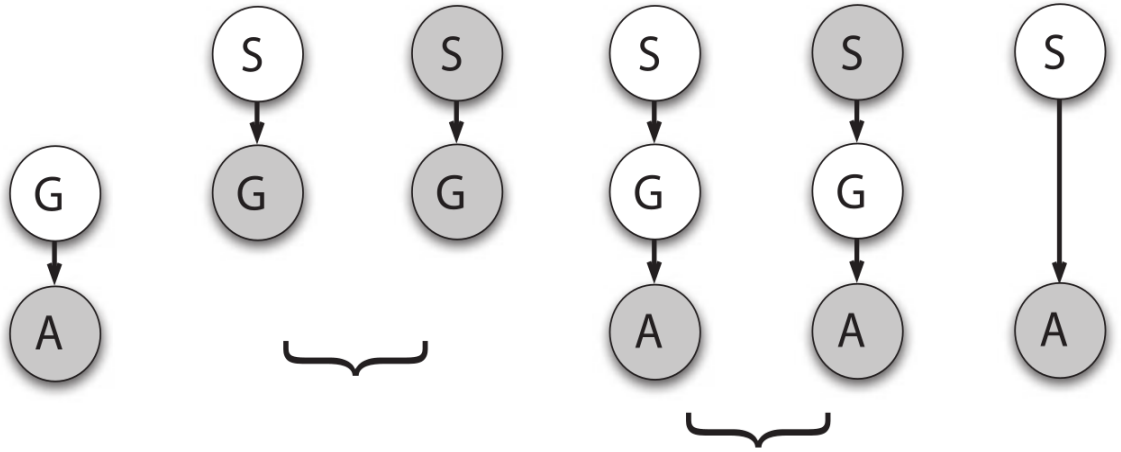
- Inference made by maximizing a score (e.g. ML) or sampling according to a posterior probability distribution (e.g. Bayesian MCMC)
- Often computationally intensive
- Some have been parallelized
- Some focus on inferring good gene trees, others on inferring a good species tree, others on both

Different methods target different problems

Phylogenomics inference pipeline



Gene tree-species tree models published in the literature



P(A|G)

PhyML [Guindon & Gascuel, 2003];
RAxML [Stamatakis et al., 2005];
MrBayes [Ronquist et al., 2012];
BEAST [Drummond et al., 2012];

P(G|S)

BEST [Edwards et al., 2007];
DLCoal [Rasmussen & Kellis, 2012];
MP-EST [Liu et al., 2010];
ODT [Szöllősi et al., 2012];
GeneRax (Morel et al. 2020)
Treerecs (Comte et al. 2020)

P(A|G,S)

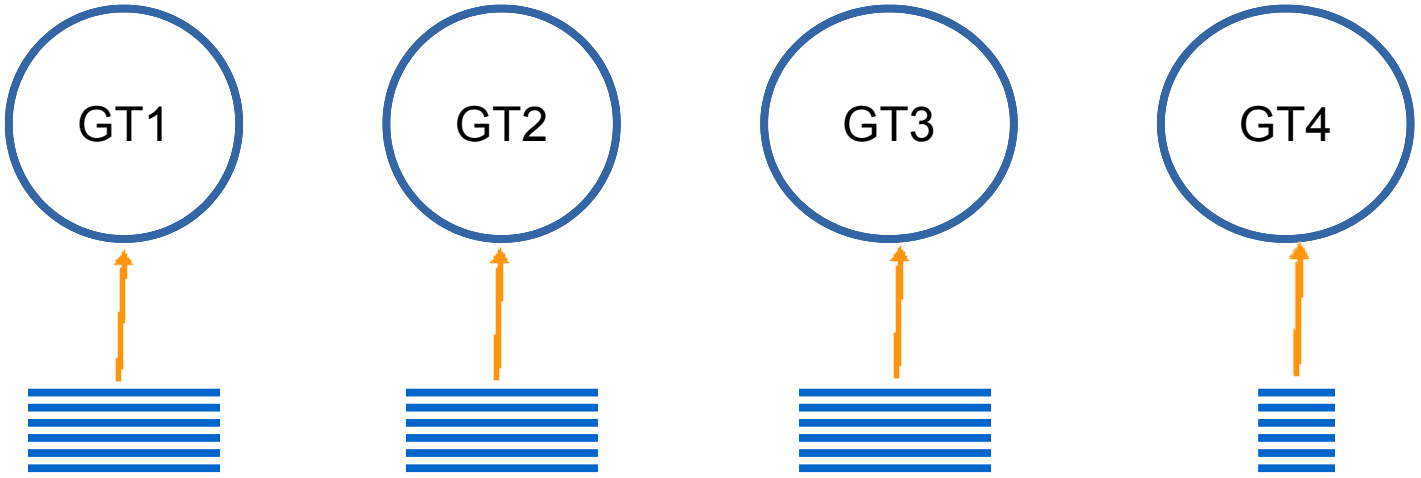
*BEAST [Heled & Drummond, 2008];
PHYLDOG [Boussau et al., 2013];
Prime-GSR [Akerborg et al., 2009];
IMa2 [Hey, 2011];
Prime-DLRS [Sjőstrand et al., 2012];
exODT [Szöllősi et al., 2013a];

P(A|S)

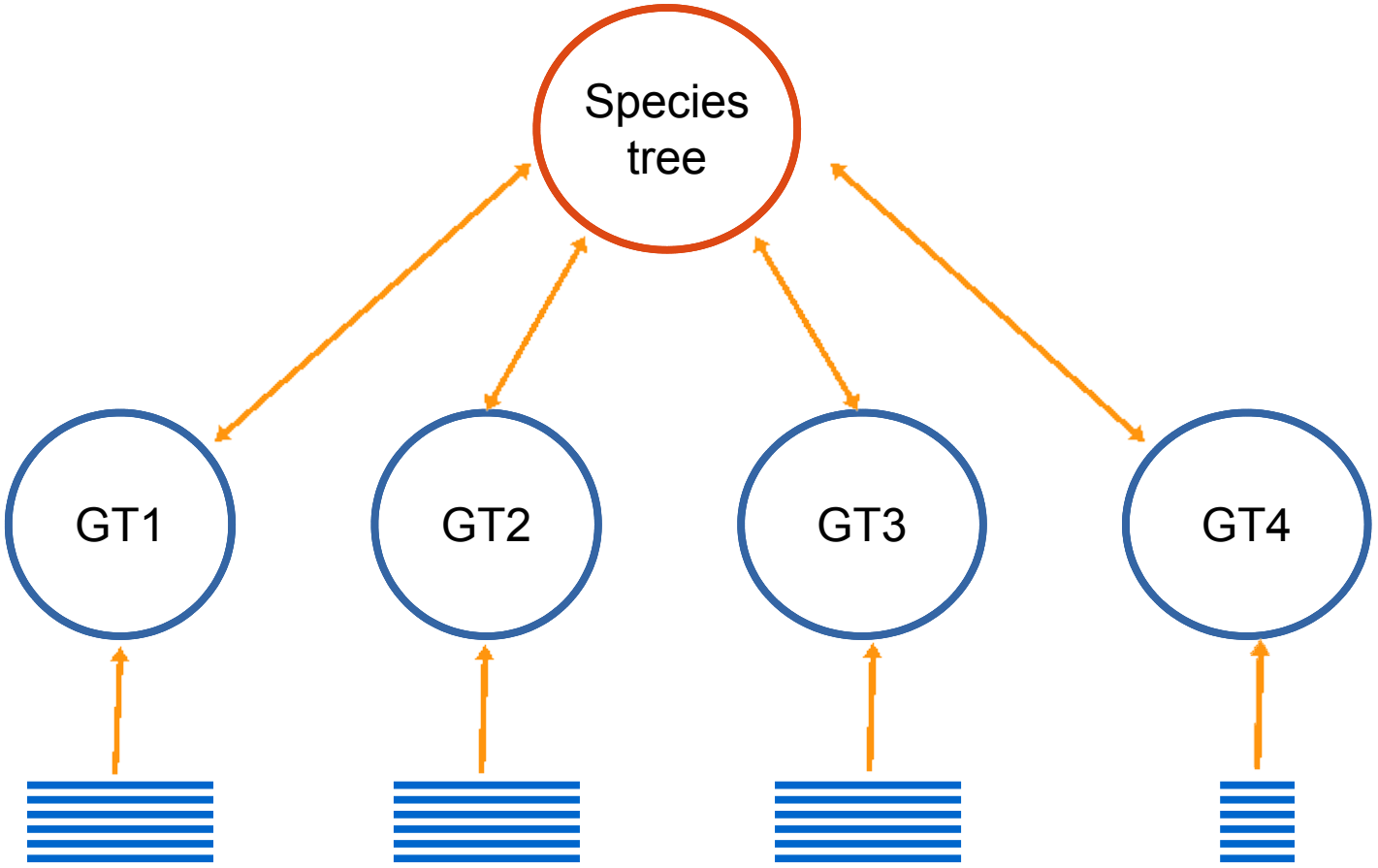
SNAPP [Bryant et al., 2012];
ALE [Szöllősi et al., 2013b];

Szöllősi et al. Syst. Biol. 2015

Without GTST models, there may be shortage of information

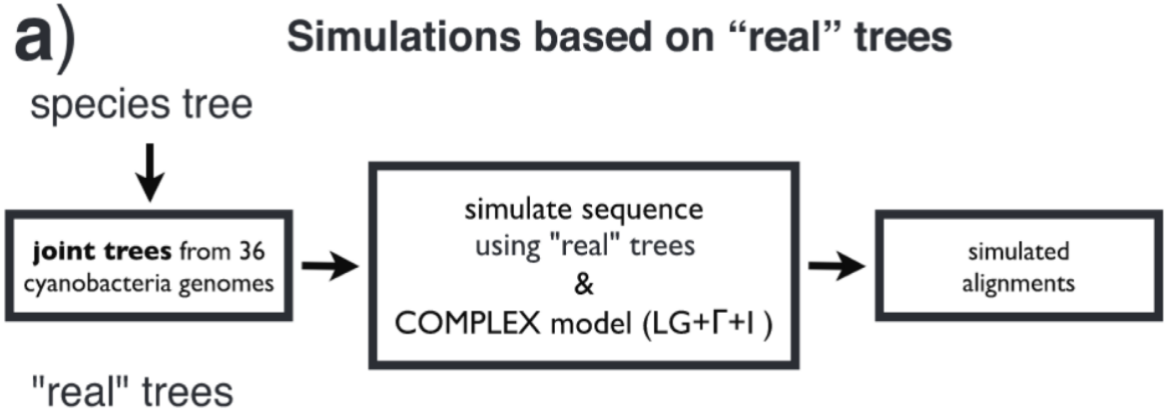


GTST models allow sharing of information thanks to a hierarchical model



Gene trees improved with a GTST model

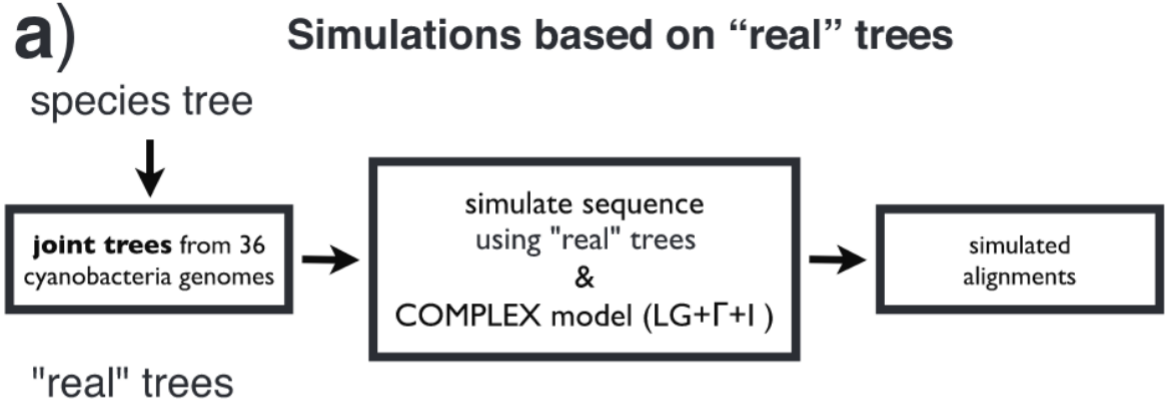
Protocol:
- use reconstructed gene trees
- simulate sequence evolution along those gene trees using a LG model + Discrete Gamma distribution to model rate heterogeneity across sites. Sequence sizes match true sequences.

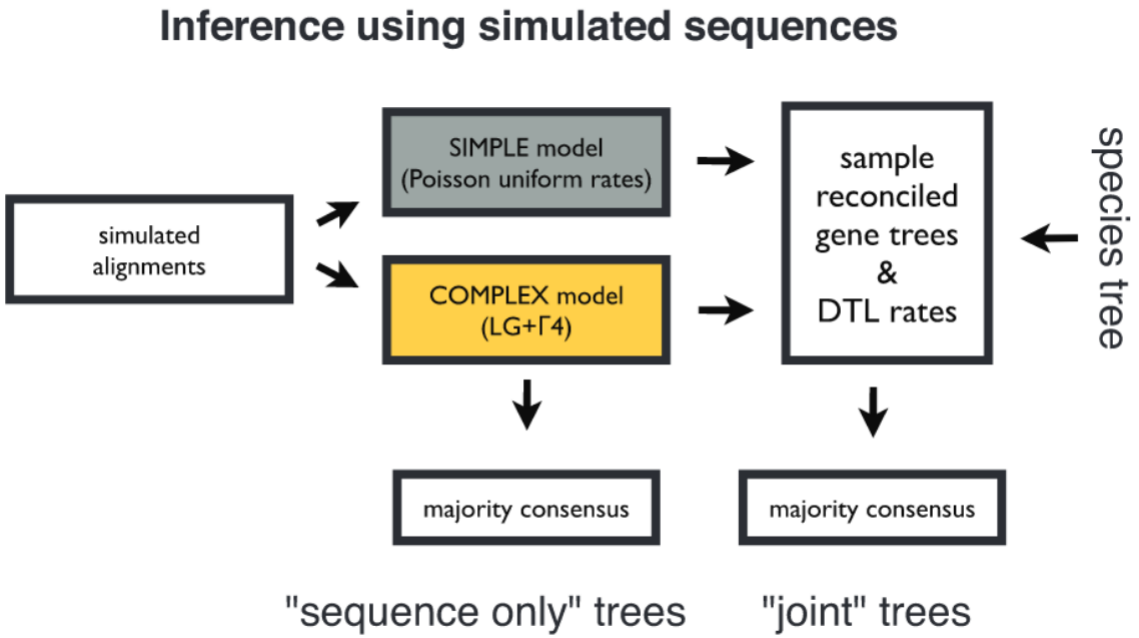


Szöllősi et al. Syst. Biol. 2013

Gene trees improved with a GTST model

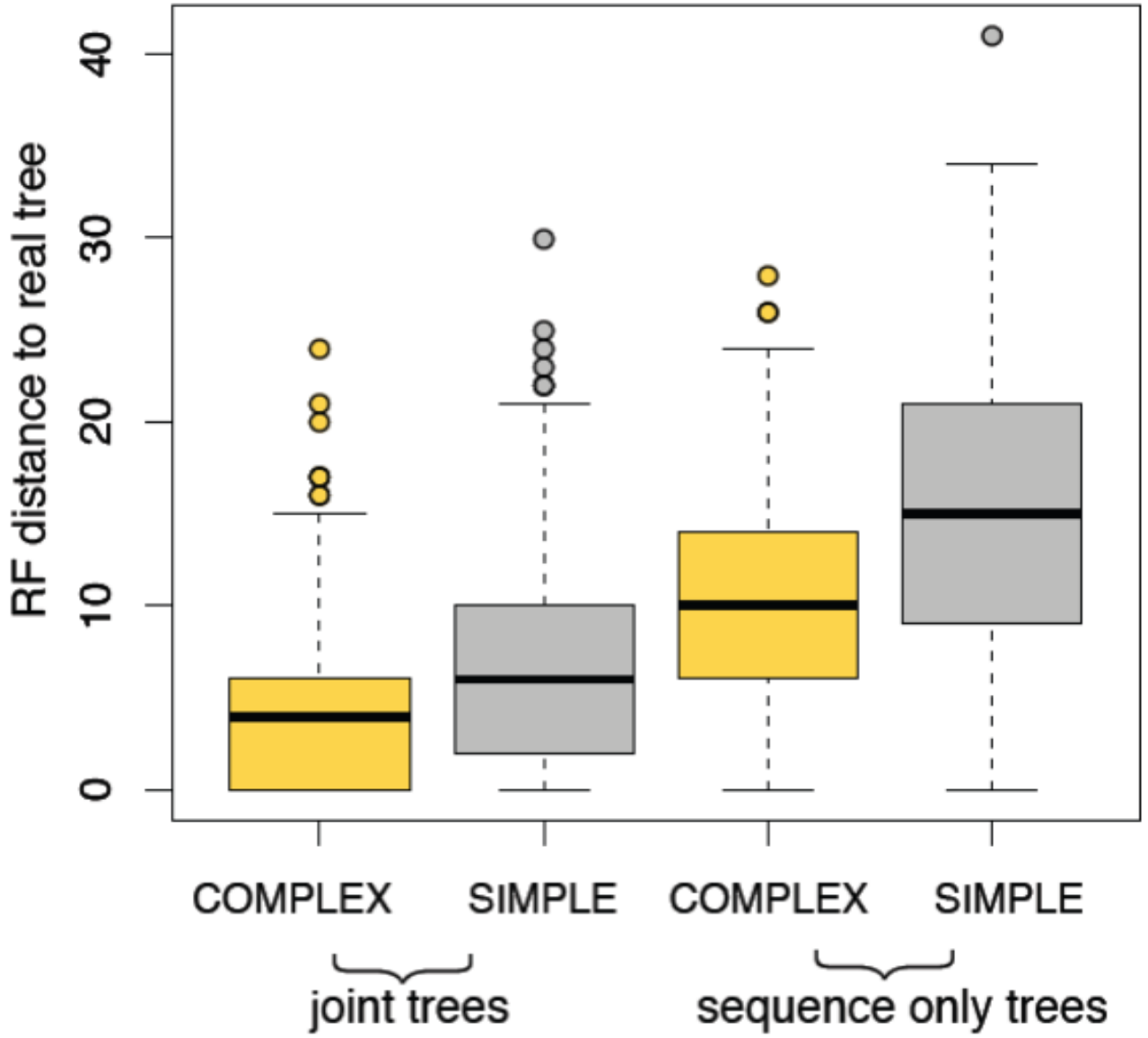
Protocol:
- use reconstructed gene trees
- simulate sequence evolution along those gene trees using a LG model + Discrete Gamma distribution to model rate heterogeneity across sites. Sequence sizes match true sequences.





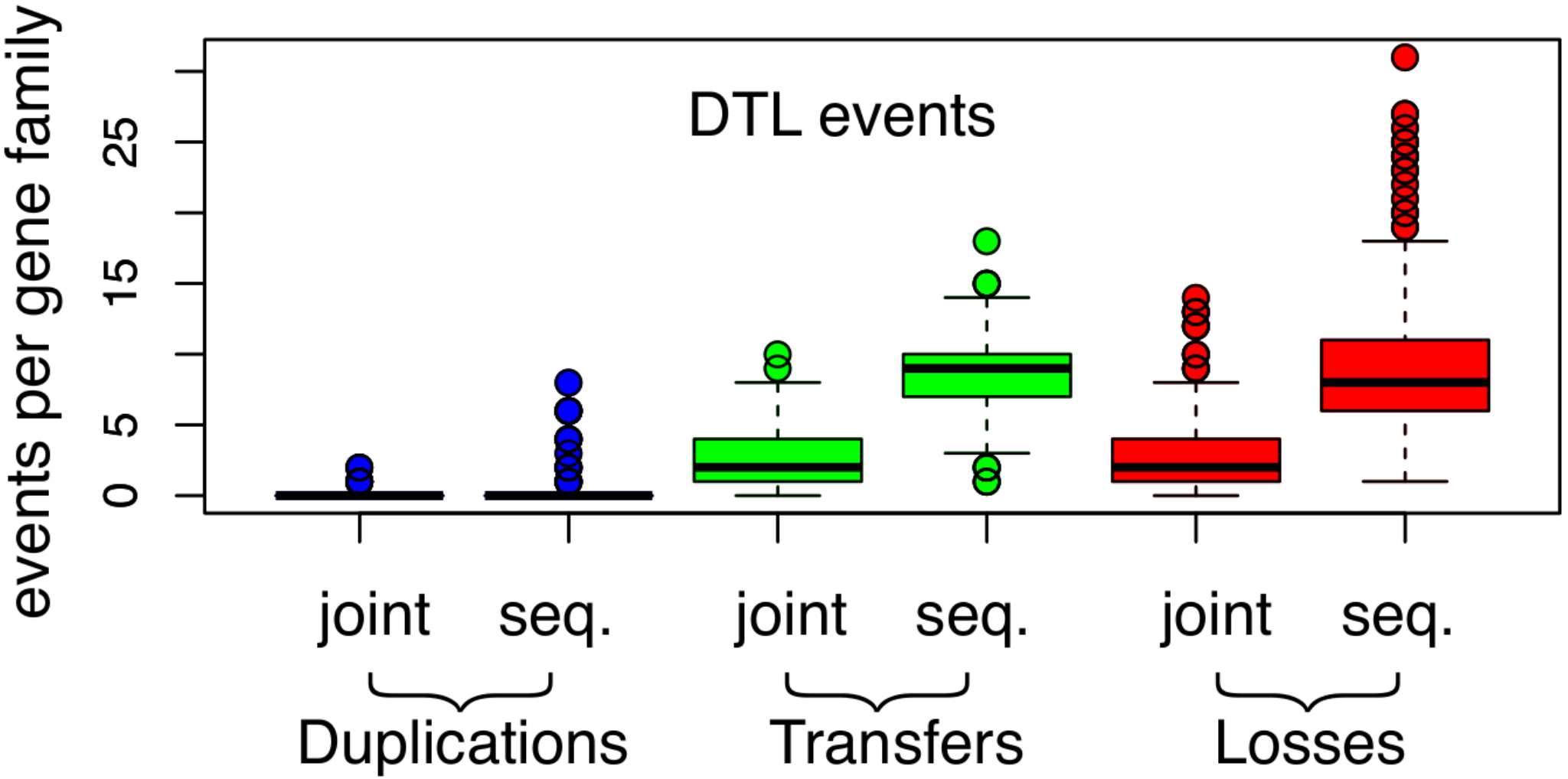
Szöllősi et al. Syst. Biol. 2013

Gene trees improved with a GTST model



Szöllősi et al. Syst. Biol. 2013

Gene trees improved with a GTST model



Szöllősi et al. Syst. Biol. 2013

Plan

I. The need for gene tree species tree models

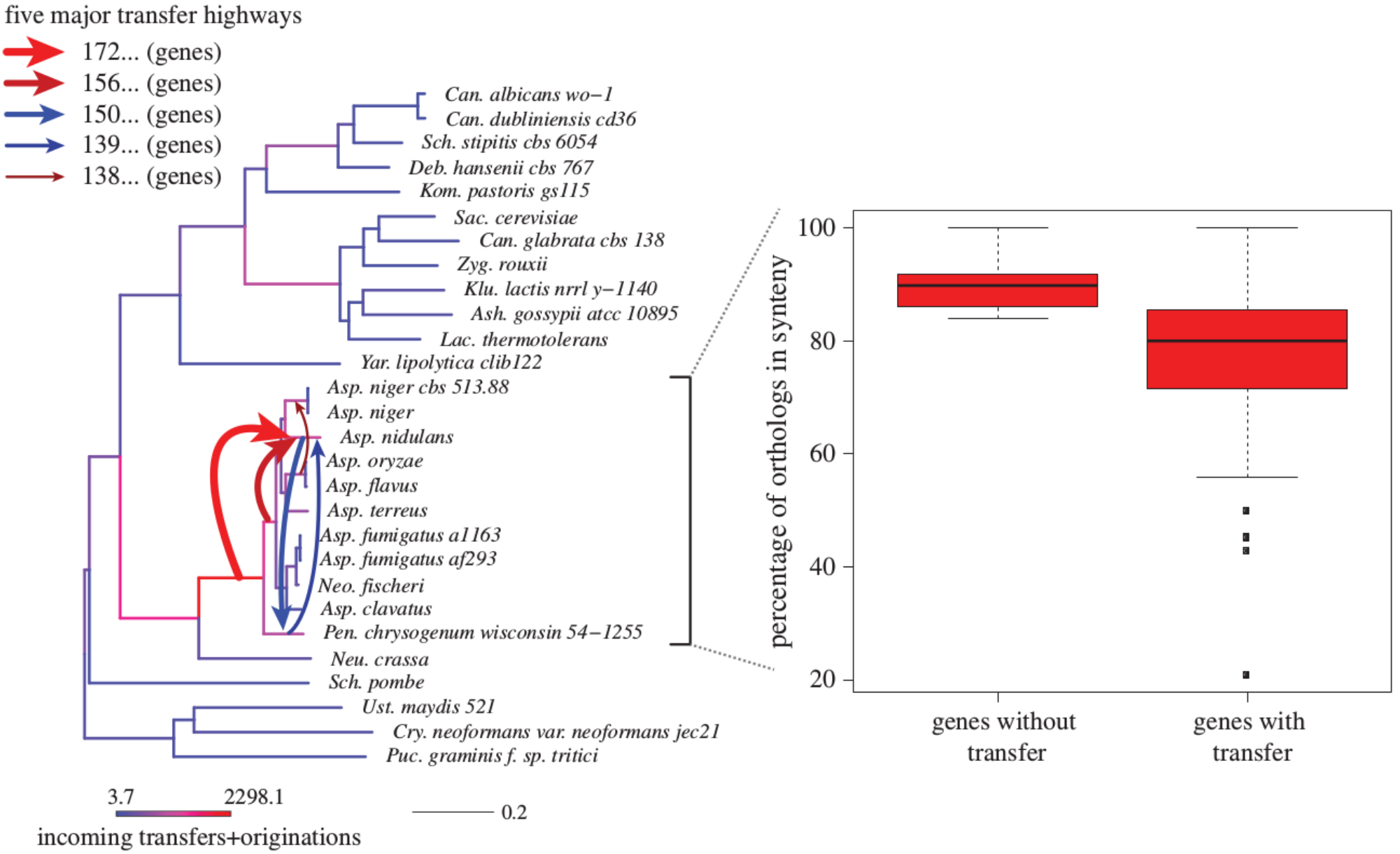
II. Using a model of gene transfers, duplications and losses (DTL) to reconstruct species and gene trees

III. Using a DTL model to compare genome evolution in Fungi vs Cyanobacteria

IV. Using a DTL model to reconstruct ancestral genomes in Archaea

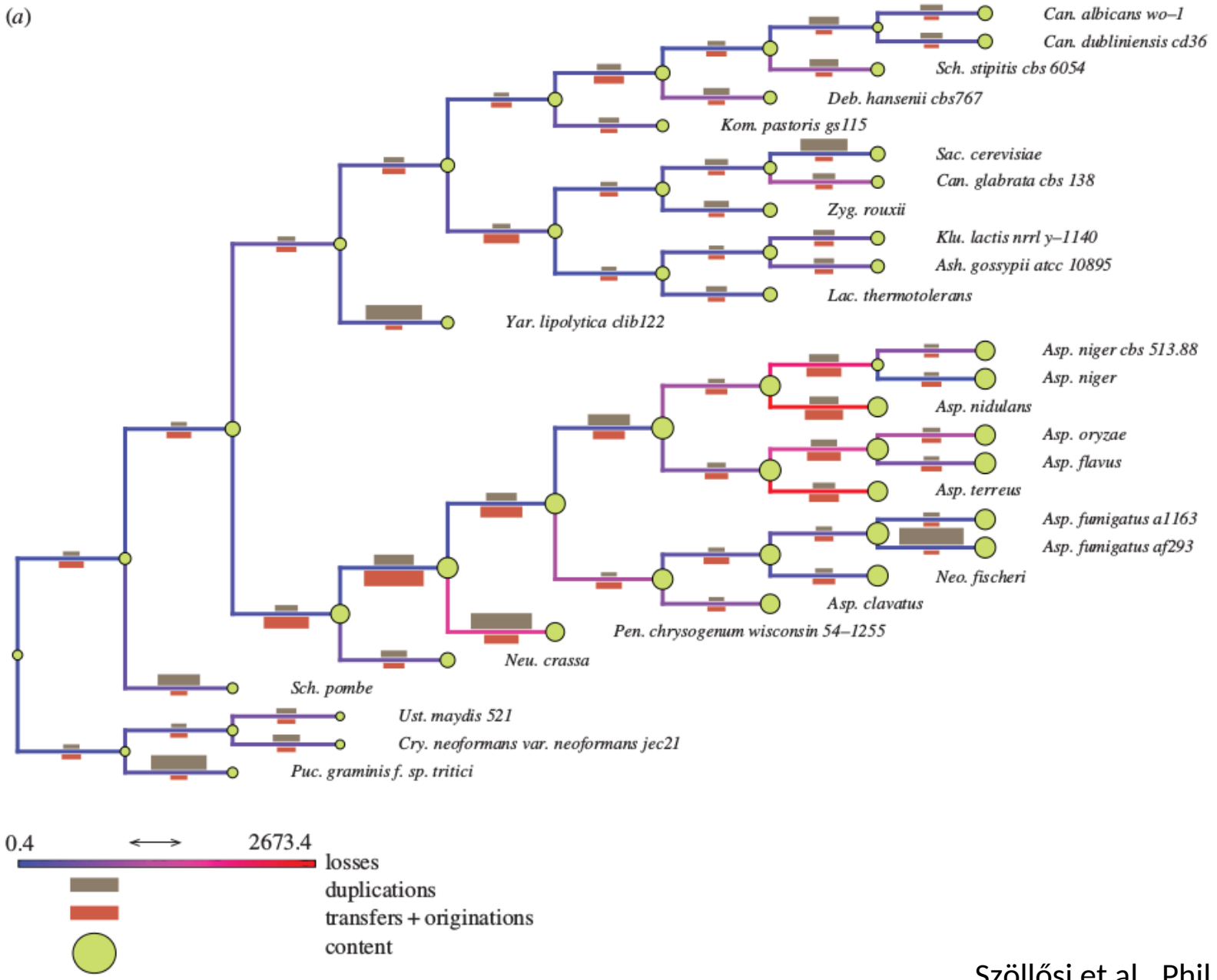
V. Using a DTL model to date species phylogenies

Highways of gene transfers in Fungi



Szöllősi et al., Phil Trans 2015

Genome evolution in Fungi



Szöllősi et al., Phil Trans 2015

Plan

I. The need for gene tree species tree models

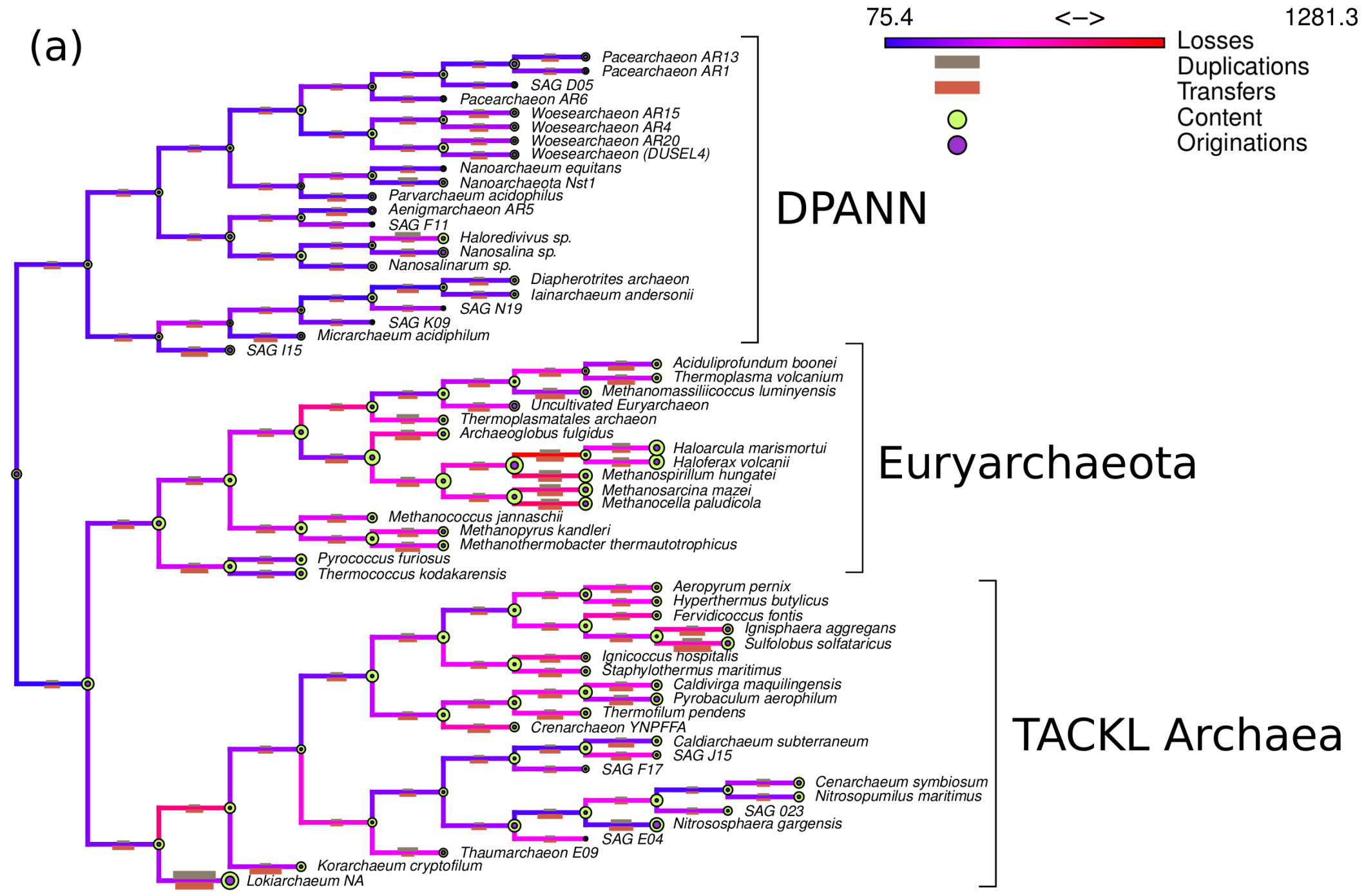
II. Using a model of gene transfers, duplications and losses (DTL) to reconstruct species and gene trees

III. Using a DTL model to compare genome evolution in Fungi vs Cyanobacteria

IV. Using a DTL model to reconstruct ancestral genomes in Archaea

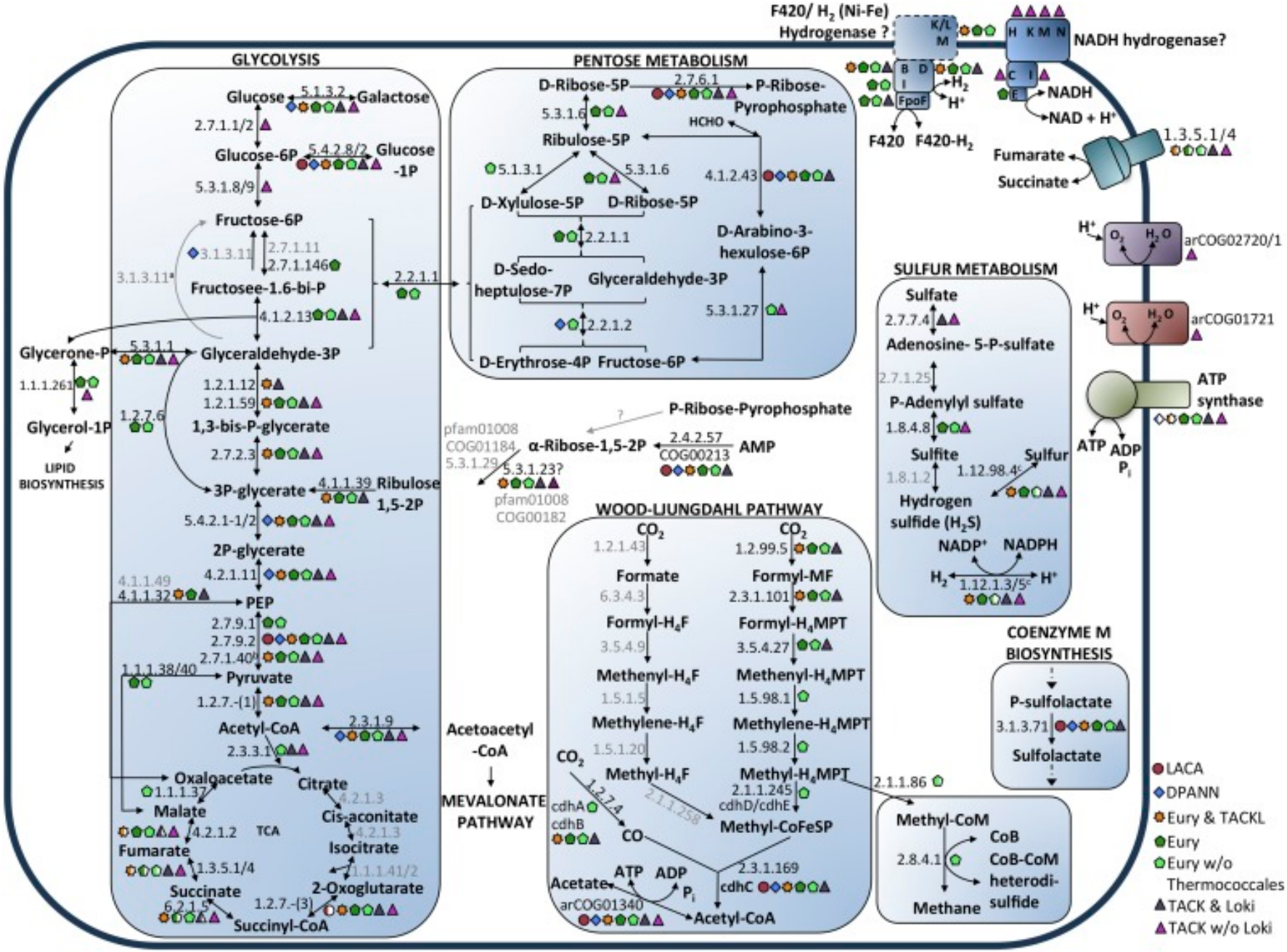
V. Using a DTL model to date species phylogenies

Inferring ancient lifestyles from ancestral gene contents



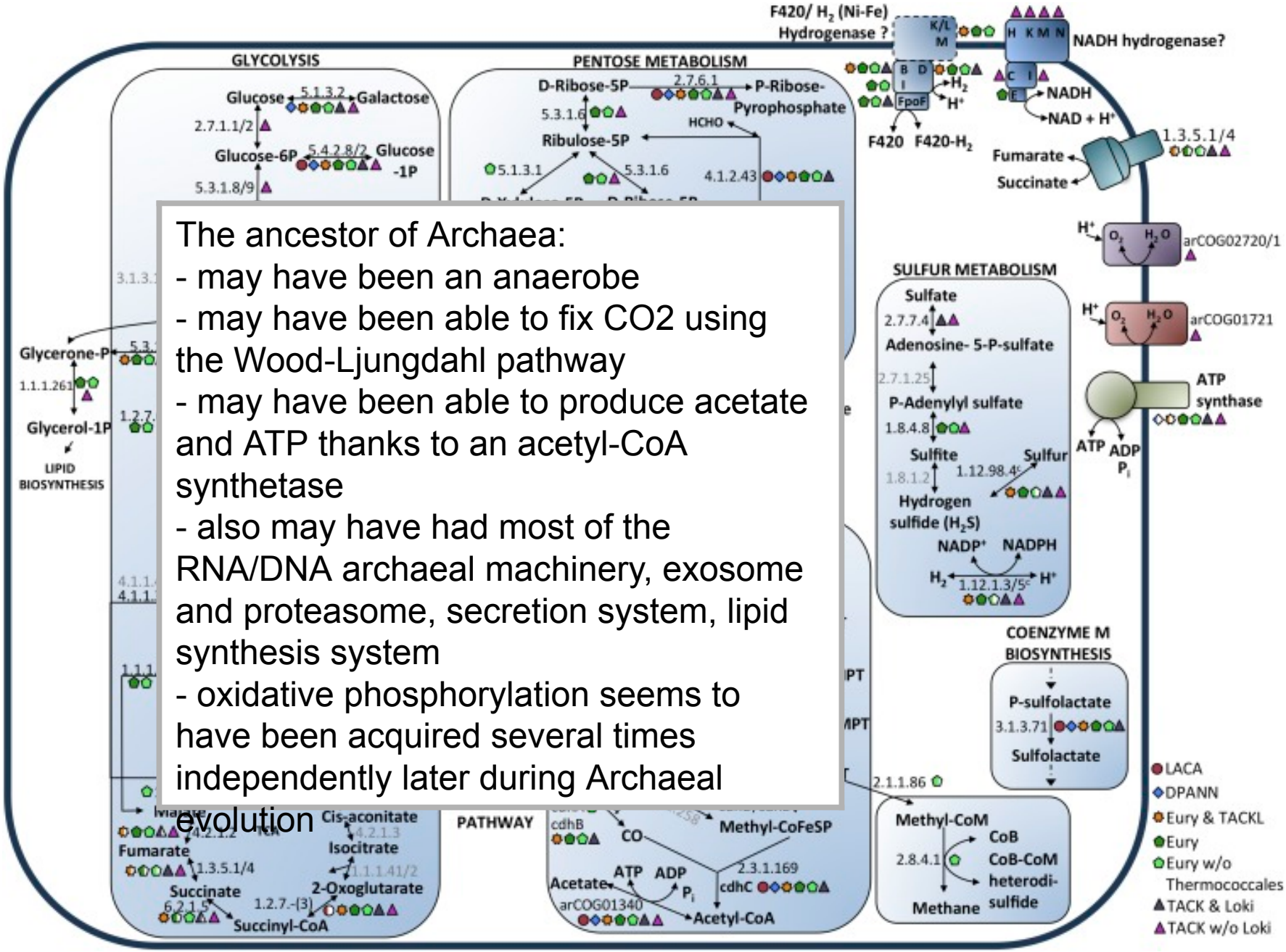
Williams et al. PNAS 2017

Inferring ancient lifestyles from ancestral gene contents



Williams et al. PNAS 2017

Inferring ancient lifestyles from ancestral gene contents



The ancestor of Archaea:
- may have been an anaerobe
- may have been able to fix CO2 using the Wood-Ljungdahl pathway
- may have been able to produce acetate and ATP thanks to an acetyl-CoA synthetase
- also may have had most of the RNA/DNA archaeal machinery, exosome and proteasome, secretion system, lipid synthesis system
- oxidative phosphorylation seems to have been acquired several times independently later during Archaeal evolution

Williams et al. PNAS 2017

Plan

I. The need for gene tree species tree models

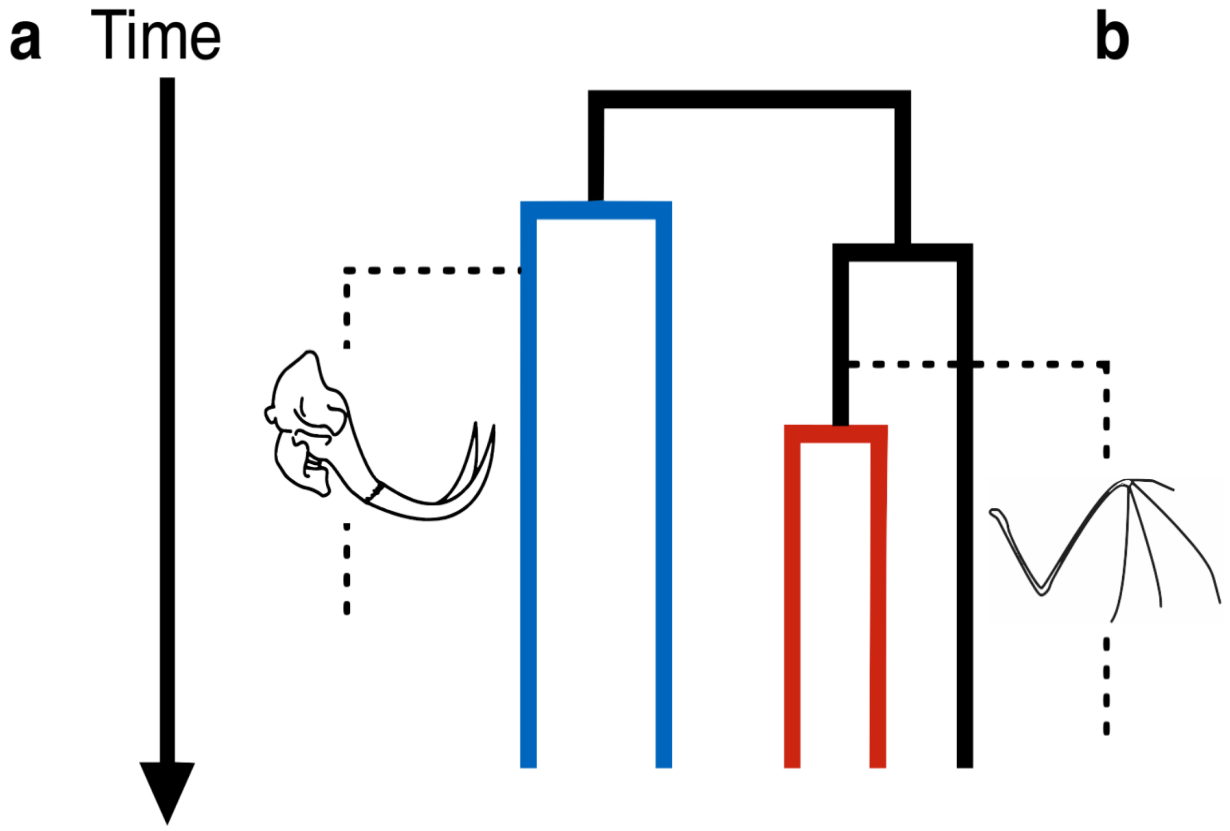
II. Using a model of gene transfers, duplications and losses (DTL) to reconstruct species and gene trees

III. Using a DTL model to compare genome evolution in Fungi vs Cyanobacteria

IV. Using a DTL model to reconstruct ancestral genomes in Archaea

V. Using a DTL model to date species phylogenies

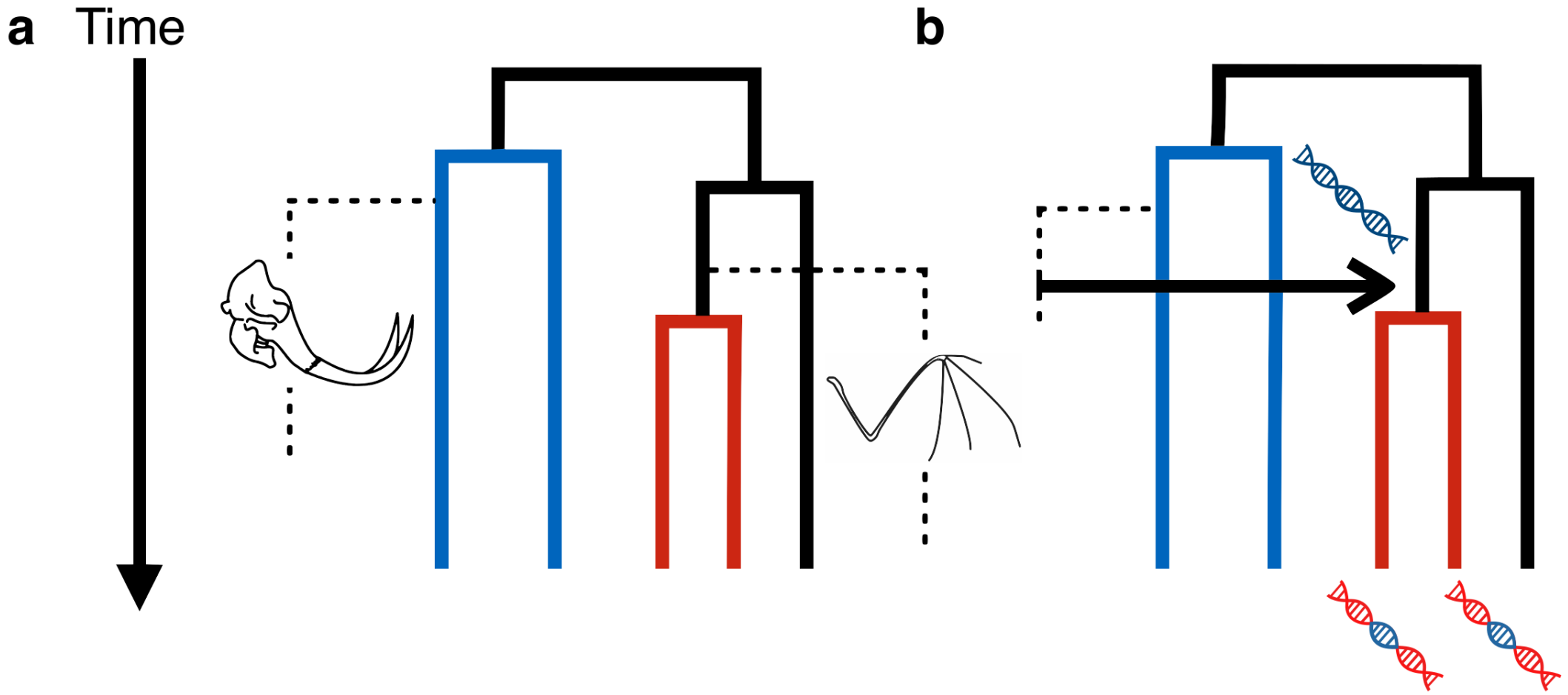
How transfers carry information about relative time



Fossils carry **absolute** timing information

Davín et al., Nature EE 2018

How transfers carry information about relative time



Fossils carry **absolute** timing information

Transfers carry **relative** timing information

Davín et al., Nature EE 2018

Dating with gene transfers

- There is a very large number of transfers that may provide a huge amount of unexplored information for dating
- This information is most abundant in clades that do not fossilize well
- Beyond gene transfers, other transfers may be useful, for instance transfers of microbiota or parasites between host species

Combining clocks and transfers

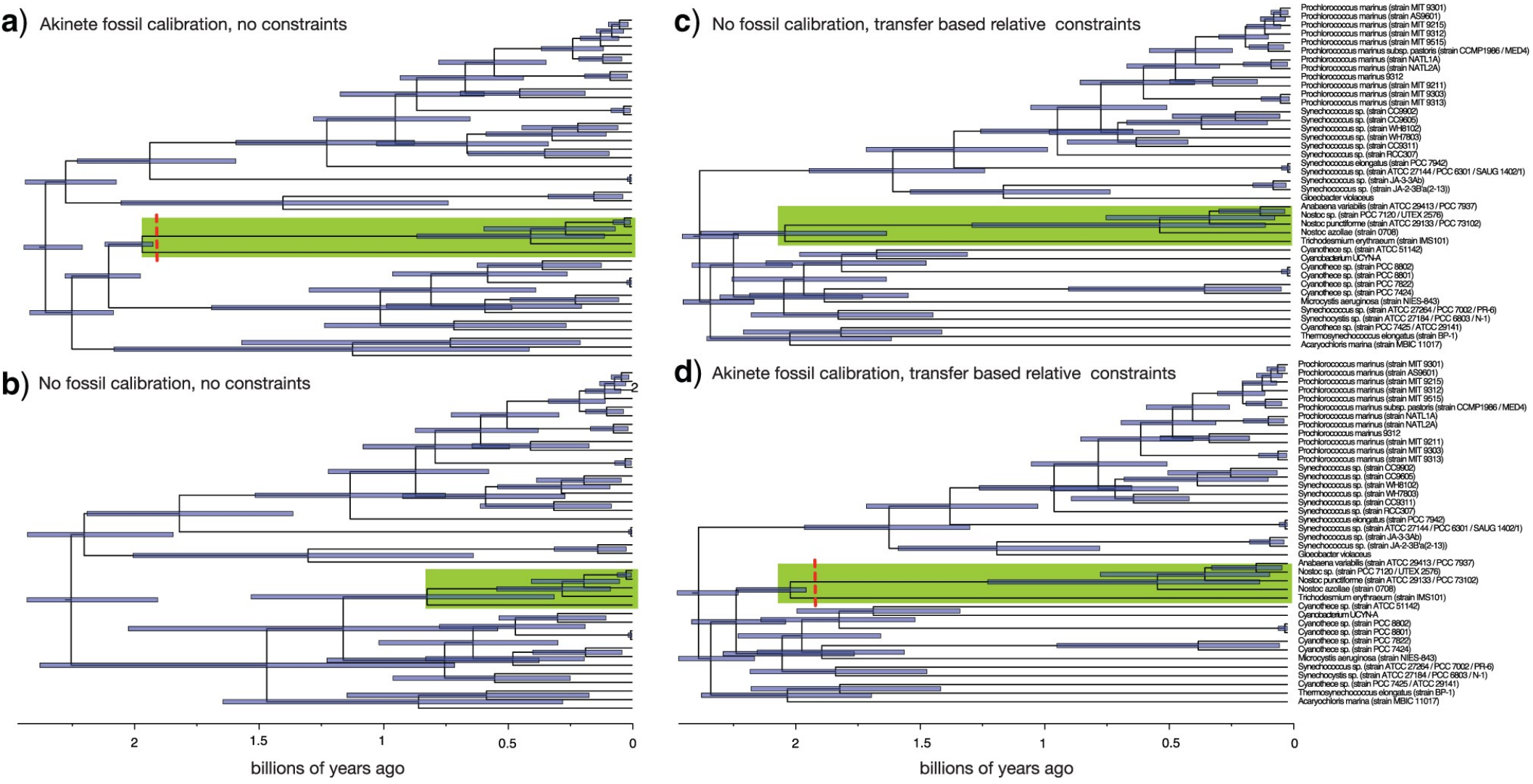
- Chronograms that disagree with a relative constraint get a 0 probability
- Compatible with existing tree priors
- Compatible with all relaxed clocks and other models
- New move to ease MCMC mixing

Method implemented in RevBayes (*https://github.com/revbayes*)

Protocol :

- estimate gene trees
- detect gene transfers using ALE (https://github.com/ssolo/ALE)
- use highly supported gene transfers as constraints
- compare dated Cyanobacteria phylogeny with constraints to phylogeny dated with fossil calibration

Constraints improve dating



Szöllősi et al., Syst Biol 2021

Conclusion

- Species tree-gene tree models allow propagating uncertainty from gene tree reconstruction into species tree reconstruction
- They enable more accurate gene tree reconstruction
- We can reconstruct ancestral genomes and infer ancient metabolisms and lifestyles
- ALE detects transfers that carry *bona fide* dating information
- This information in found across the tree of life
- We can combine transfer-based constraints with fossil calibrations and relaxed molecular clock