Genomes as documents of evolutionary history

Vincent Daubin

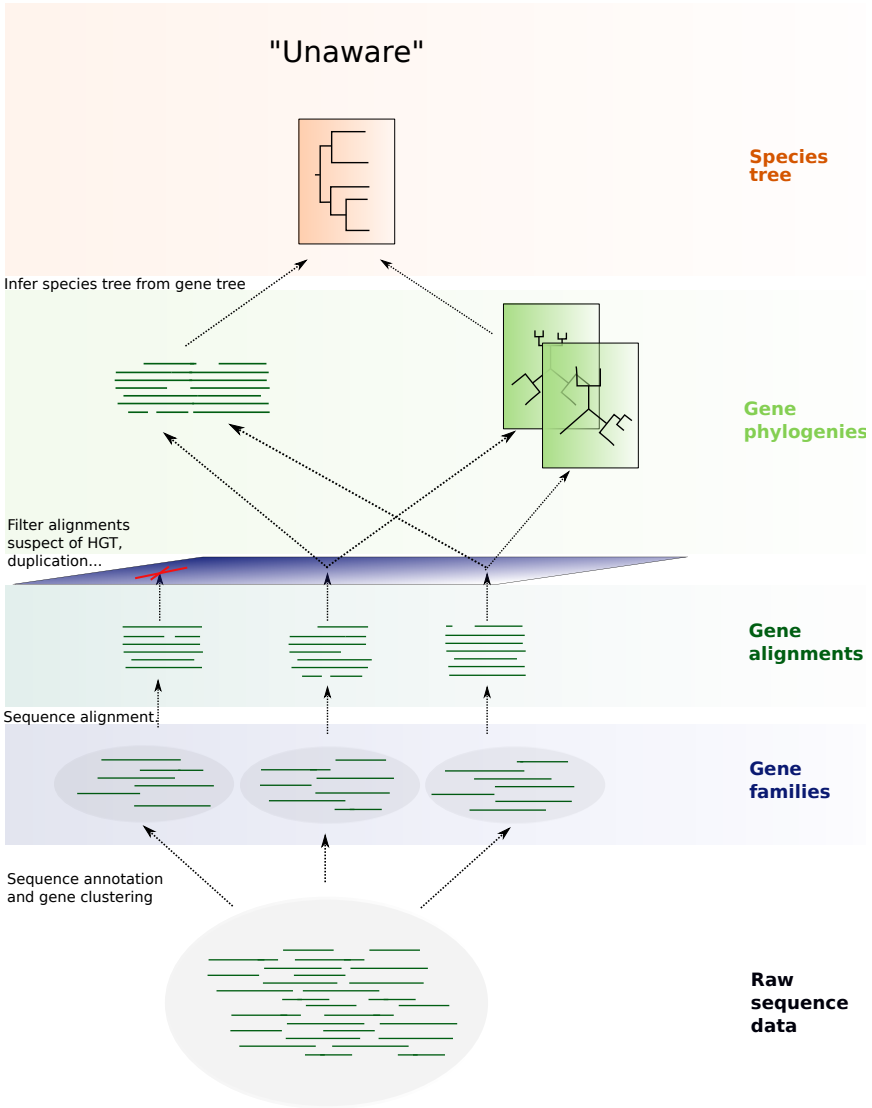
Laboratoire de Biométrie et Biologie Evolutive, CNRS, Université de Lyon, FRANCE

September 6, 2010

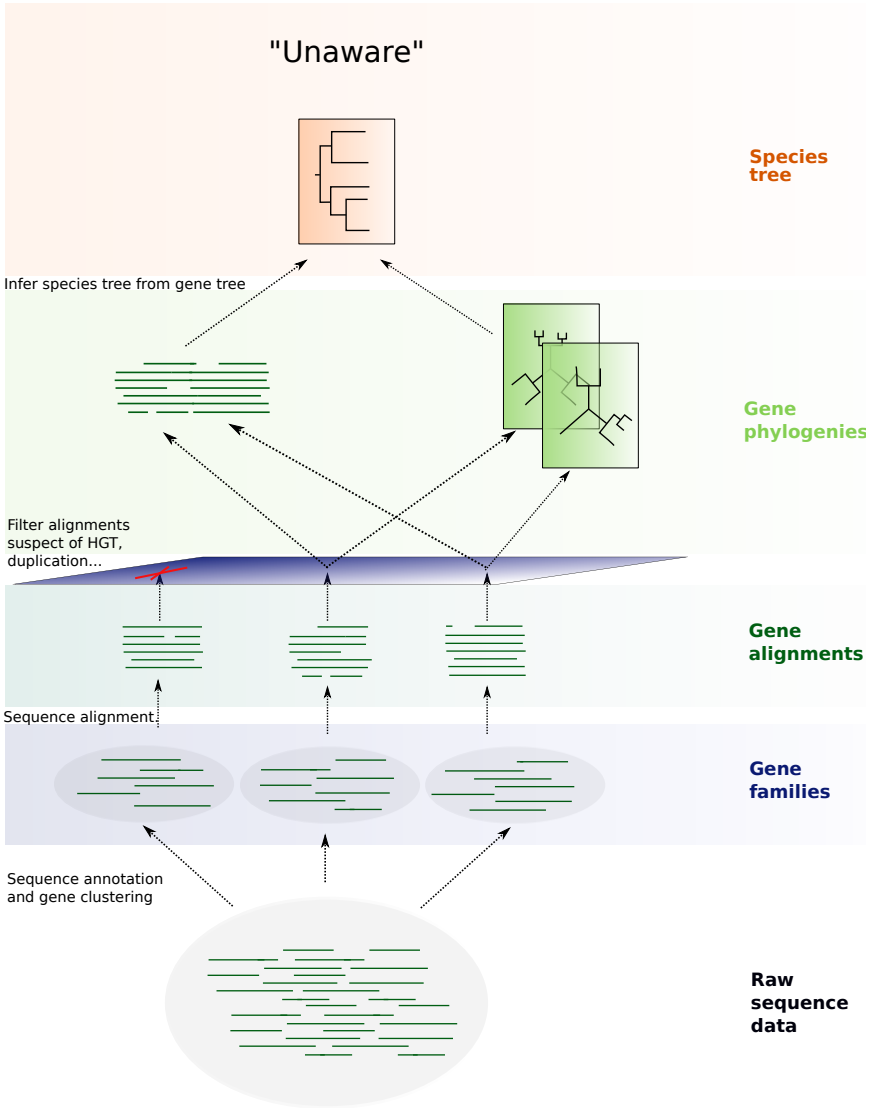
Réunion Bioinformatique et Biodiversité 2010

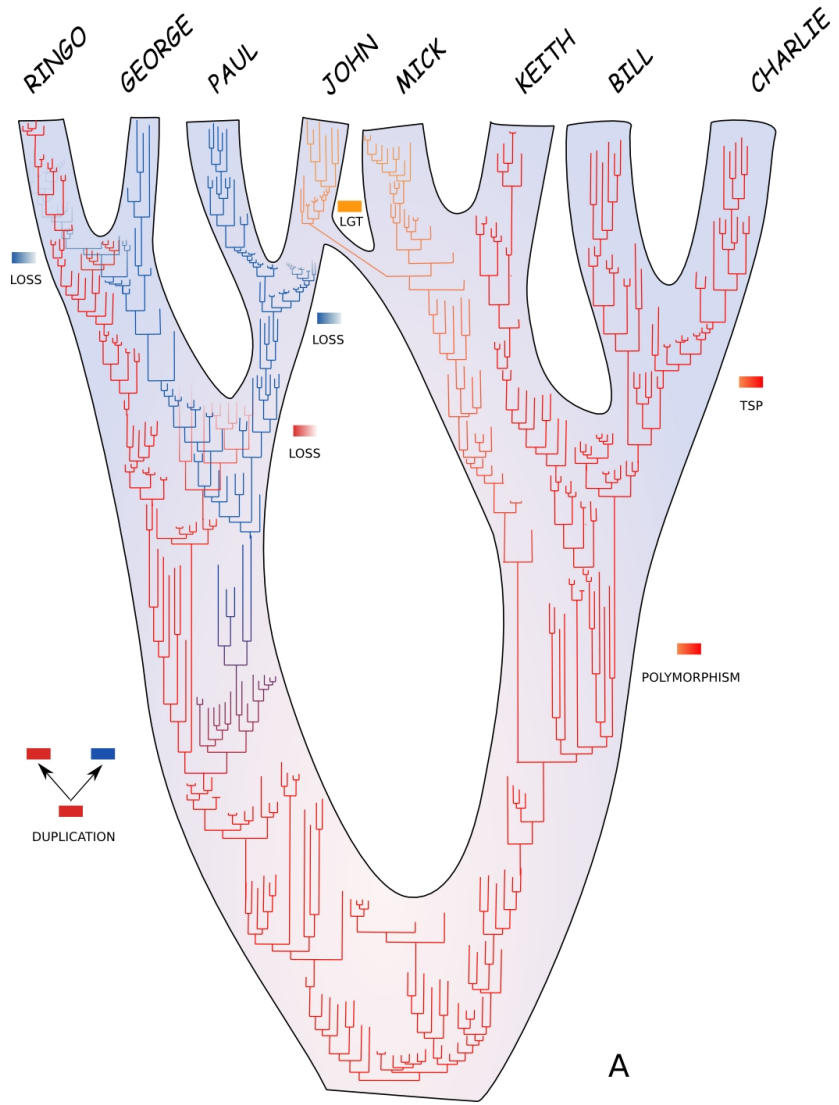
Drawings by Lenke Sifko

Using genetic information to reconstruct the history of life



Using genetic information to reconstruct the history of life





Distribution of gene families in Ensembl v.57

- **17976** gene families with more than 2 genes

Distribution of gene families in Ensembl v.57

- **17976** gene families with more than 2 genes

No obvious duplications

- **3525** families with at most one gene per species

Obvious duplications

- **14451** families

Distribution of gene families in Ensembl v.57

- **17976** gene families with more than 2 genes

No obvious duplications

- **3525** families with at most one gene per species
- Median size: **6** species

Obvious duplications

- **14451** families
- Median size: **41** species

Distribution of gene families in Ensembl v.57

- **17976** gene families with more than 2 genes

No obvious duplications

- **3525** families with at most one gene per species
- Median size: **6** species
- **13** families with 51 species

Obvious duplications

- **14451** families
- Median size: **41** species
- **713** families with 51 species

Distribution of gene families in Ensembl v.57

- **17976** gene families with more than 2 genes

No obvious duplications

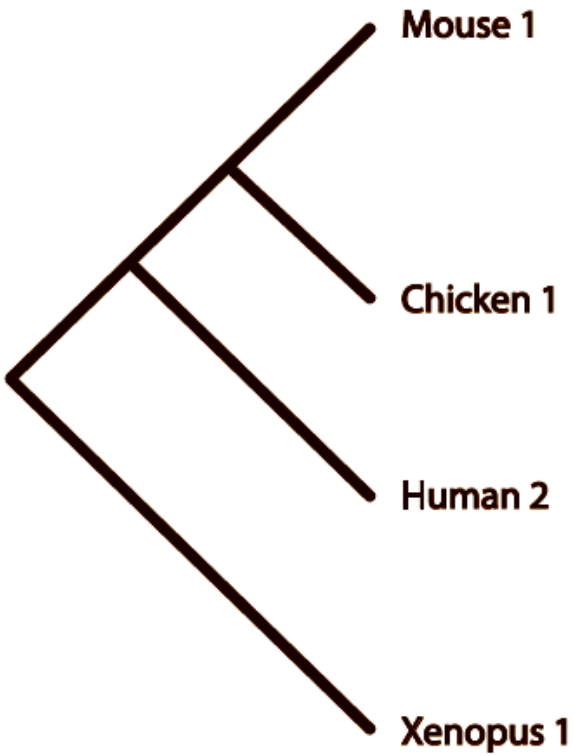
- **3525** families with at most one gene per species
- Median size: **6** species
- **13** families with 51 species

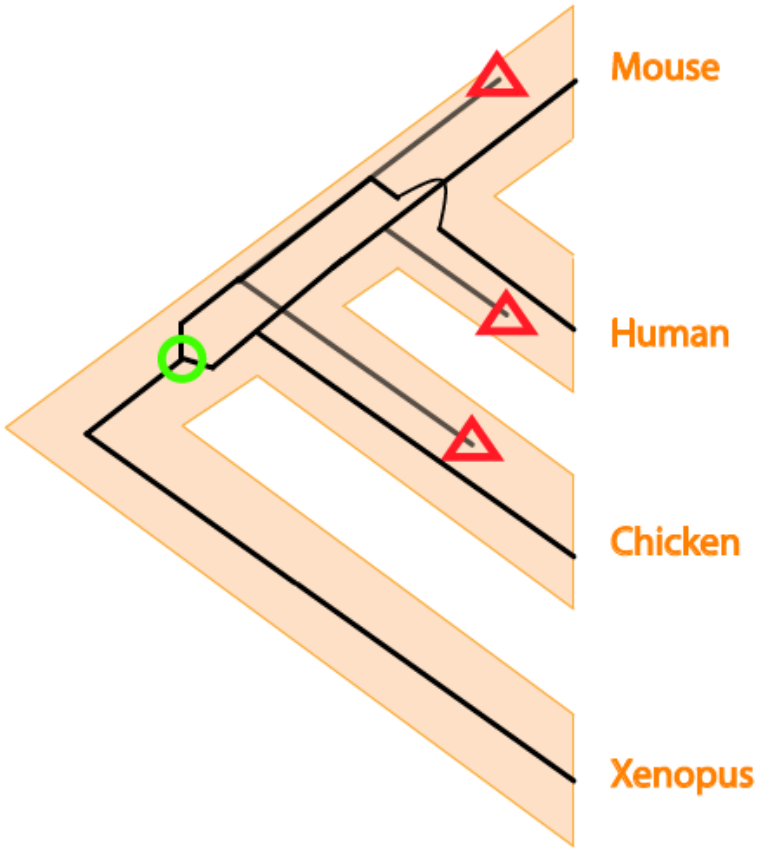
Obvious duplications

- **14451** families
- Median size: **41** species
- **713** families with 51 species

Most commonly used methods waste a lot of potentially valuable information

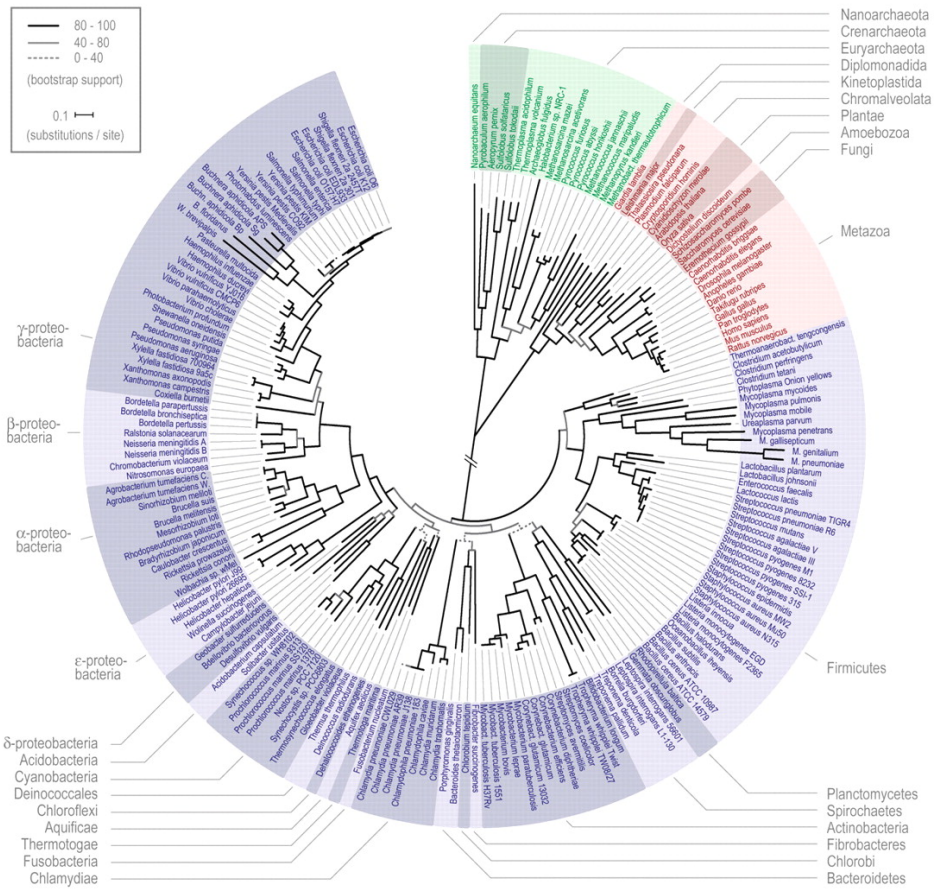
Hidden paralogy





A "highly resolved" tree of life

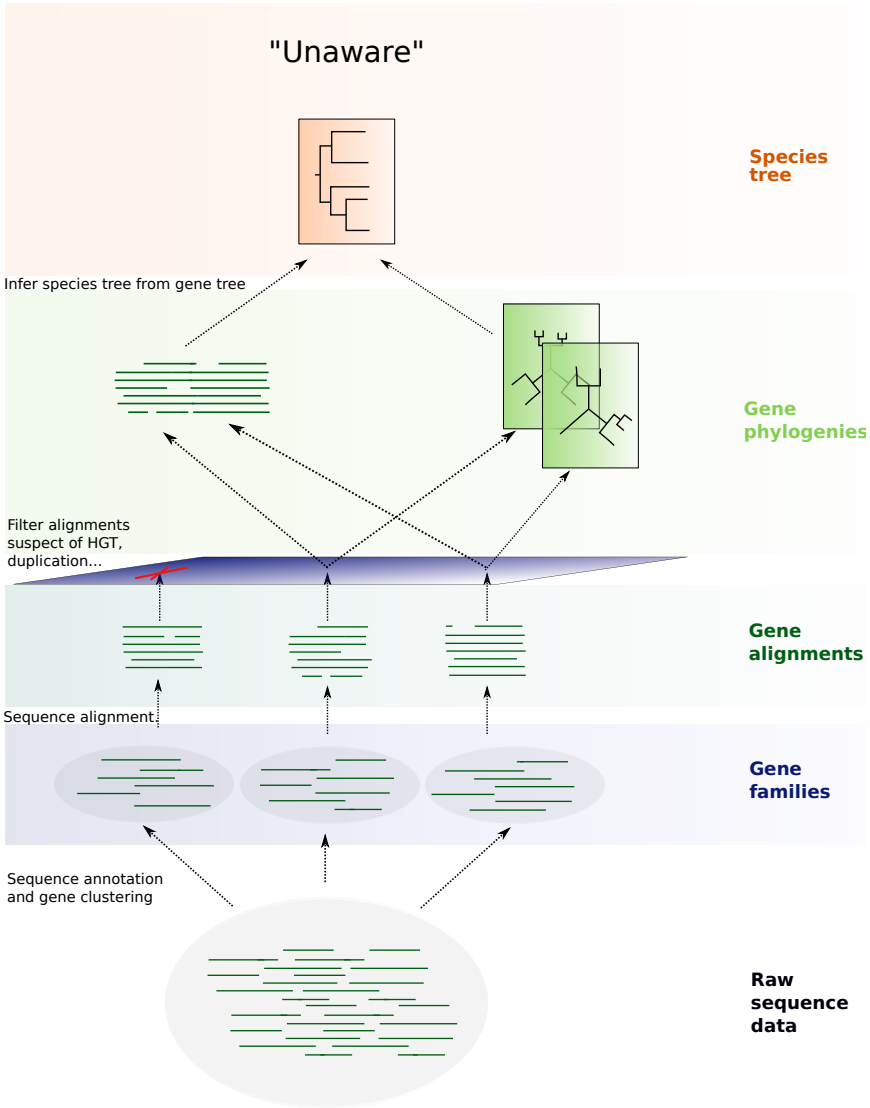
Ciccarelli et al. 2006, *Science*

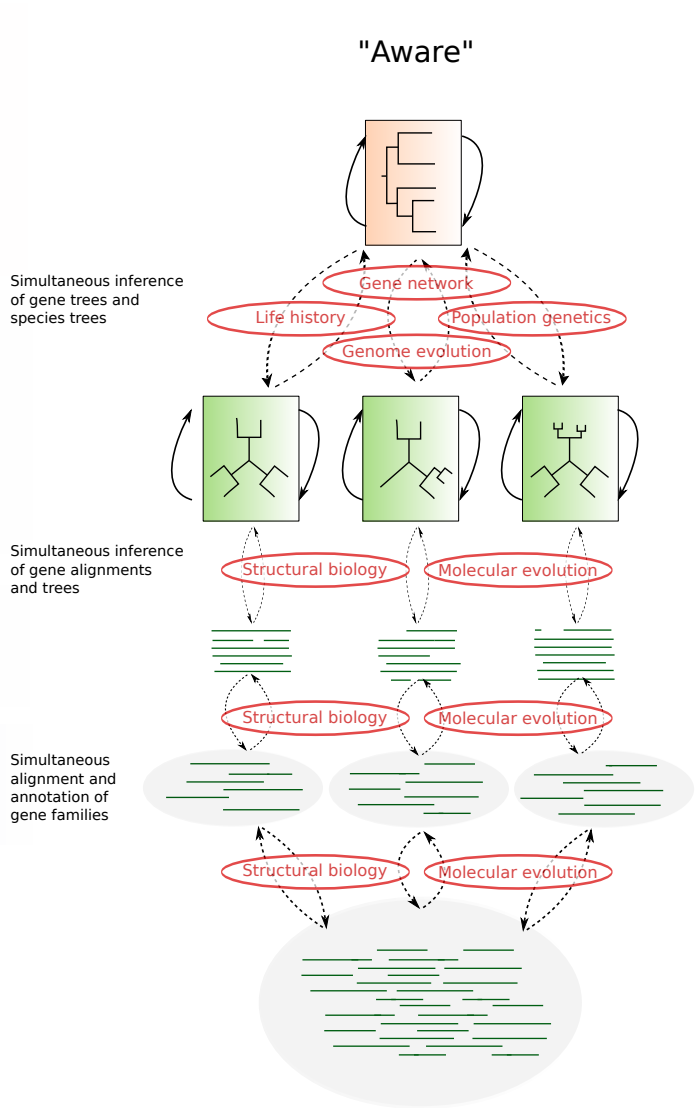


- most deep nodes have only weak support
- based on 31 universal single copy genes ("The tree of 1%"[a])

[a] Dagan and Martin. 2006, *Genome Biol*

Using genetic information to reconstruct the history of life





Reconciling organisms and gene histories

General idea

- There is a strong phylogenetic signal in gene histories

Reconciling organisms and gene histories

General idea

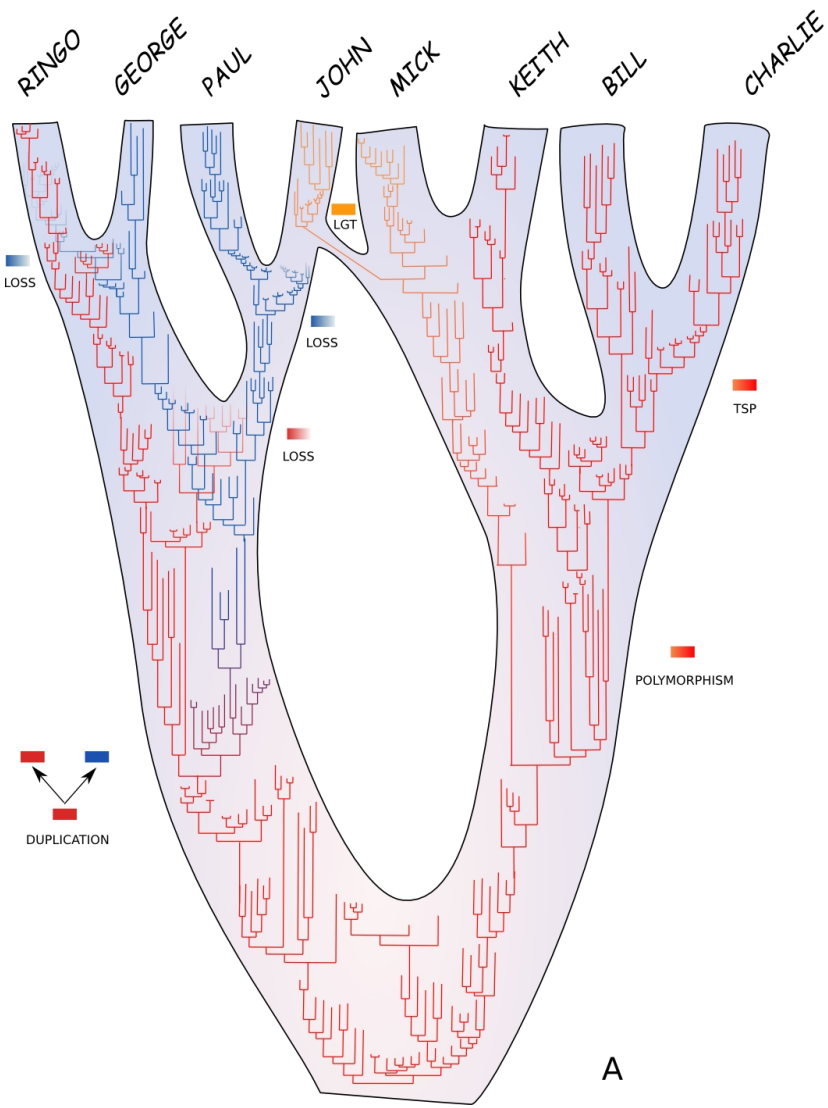
- There is a strong phylogenetic signal in gene histories
- This signal is a mix of horizontal and vertical transfer, gene duplication and loss

Reconciling organisms and gene histories

General idea

- There is a strong phylogenetic signal in gene histories
- This signal is a mix of horizontal and vertical transfer, gene duplication and loss
- In order to reconstruct the history of life, we would like to sort out (and take advantage of) these signals

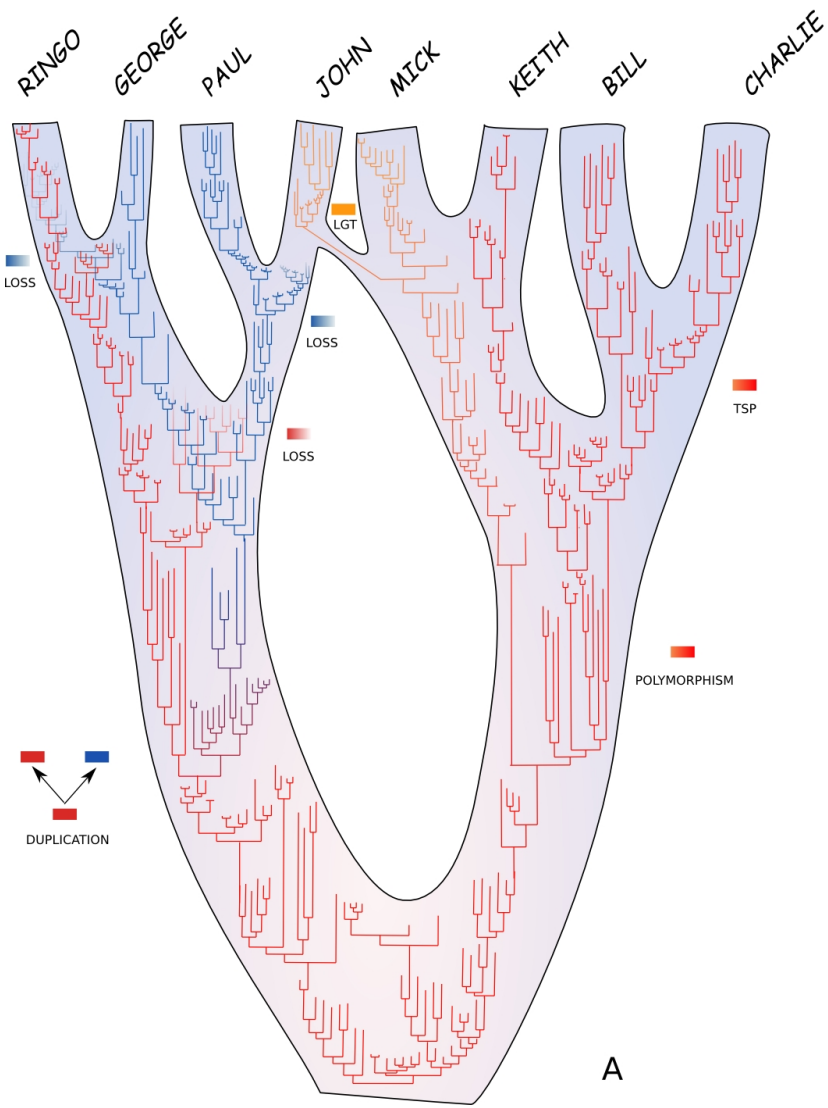
Using genetic information to reconstruct the history of life



In this talk:

- Phyldog: Taking duplication and loss into account (B. Boussau)

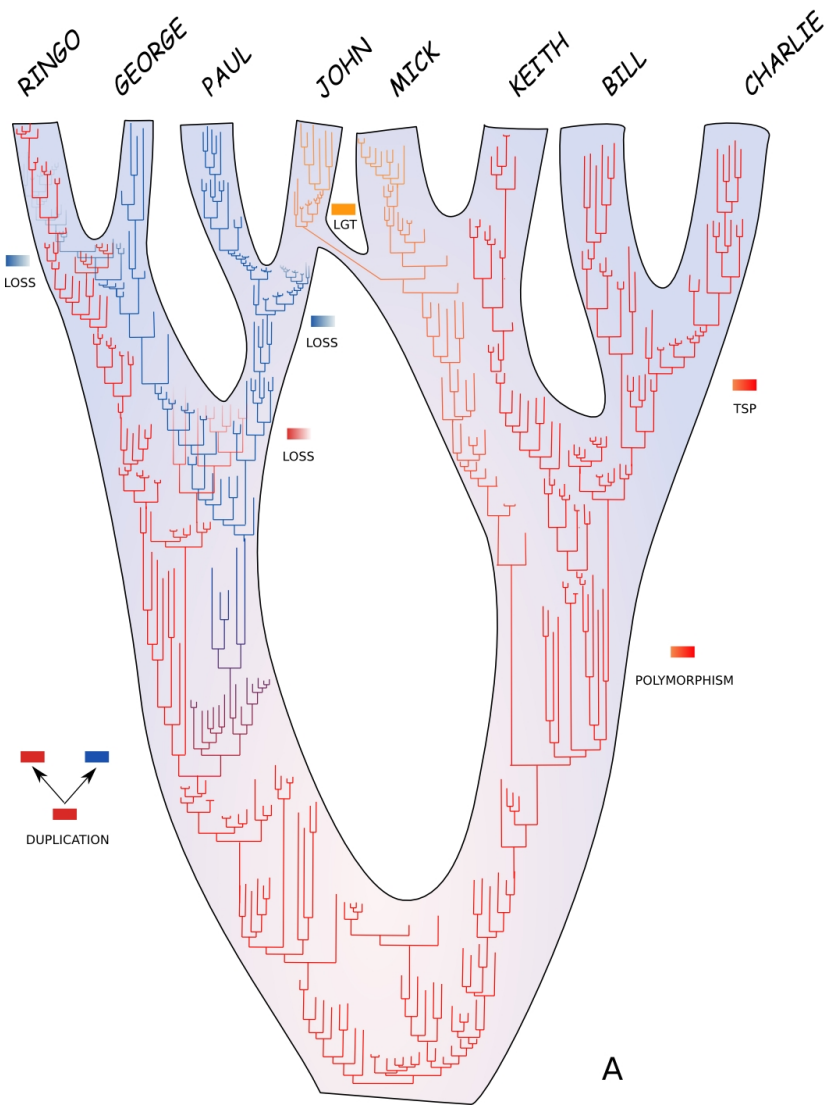
Using genetic information to reconstruct the history of life



In this talk:

- Phyldog: Taking duplication and loss into account (B. Boussau)
- HGT as a molecular clock: a model with duplication, transfer and loss (DTL) (G. Szollosi)

Using genetic information to reconstruct the history of life



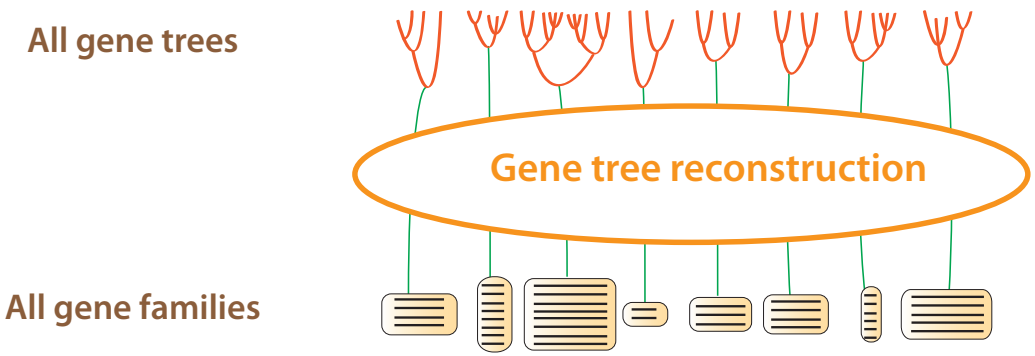
In this talk:

- Phyldog: Taking duplication and loss into account (B. Boussau)
- HGT as a molecular clock: a model with duplication, transfer and loss (DTL) (G. Szollosi)

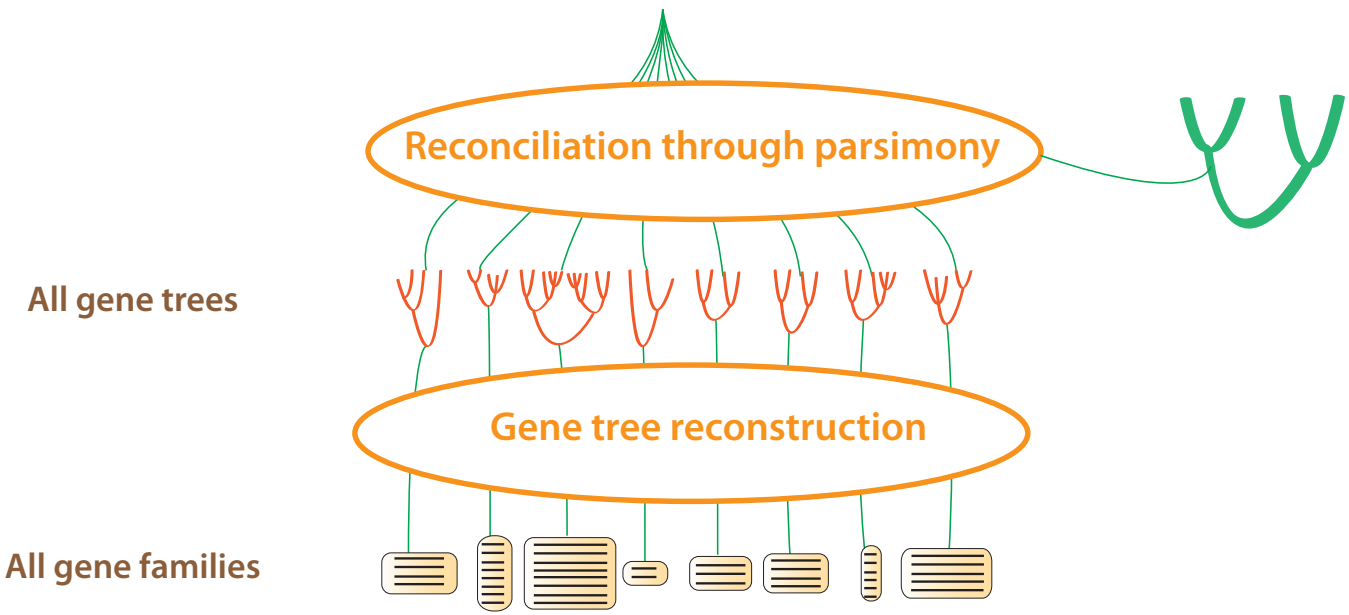
How to consider duplication and loss events

All gene families

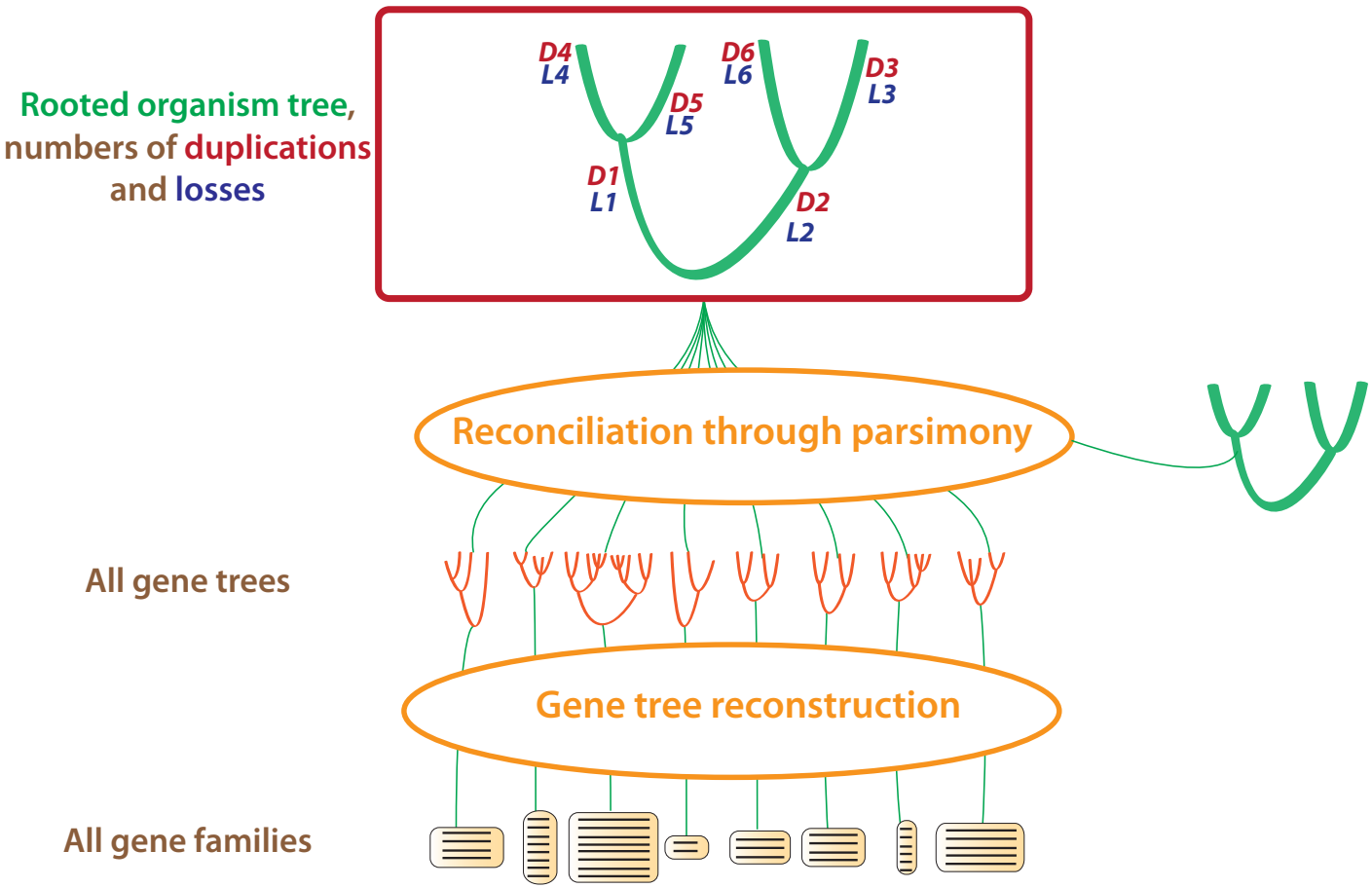
How to consider duplication and loss events



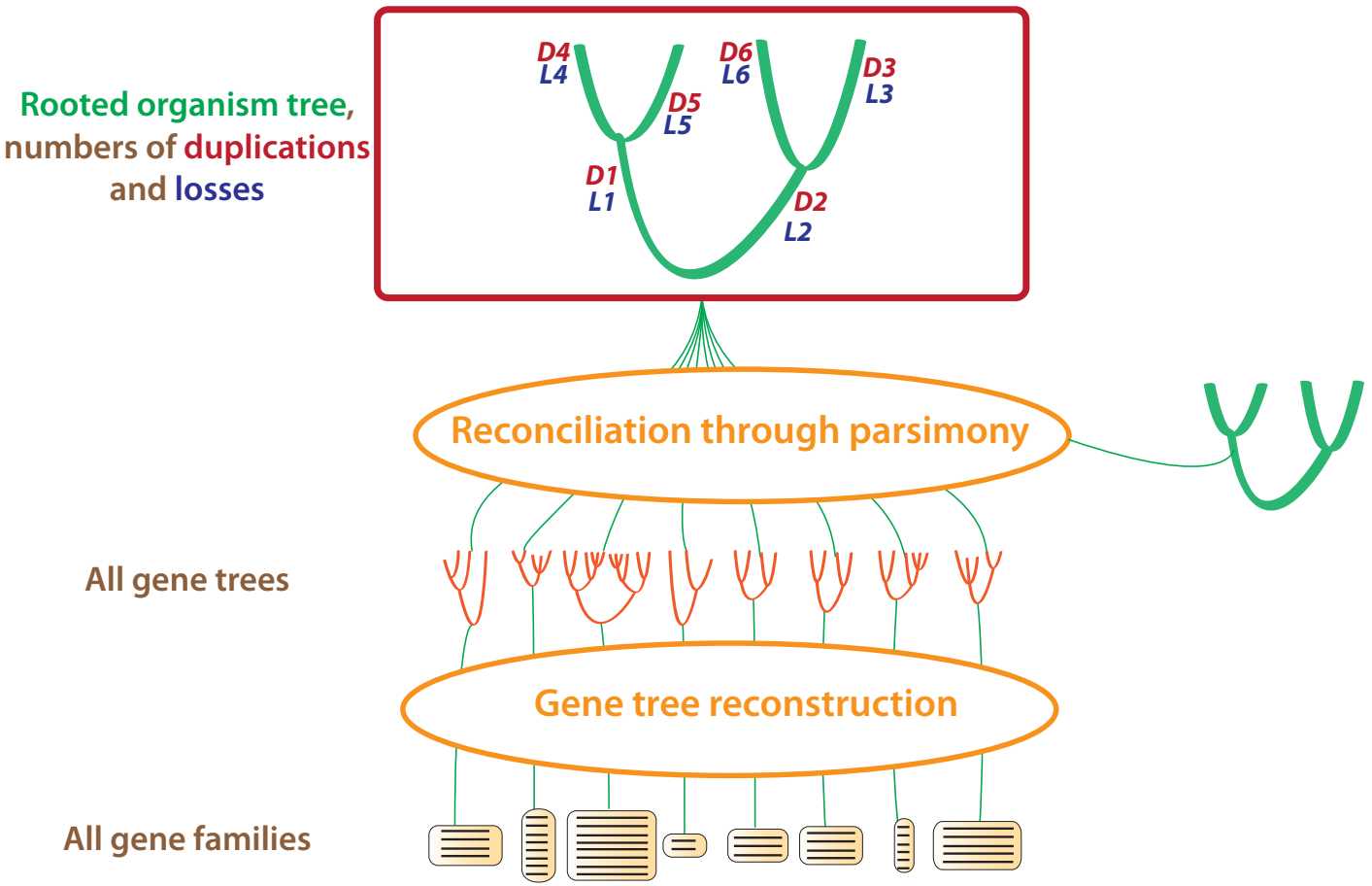
How to consider duplication and loss events



How to consider duplication and loss events



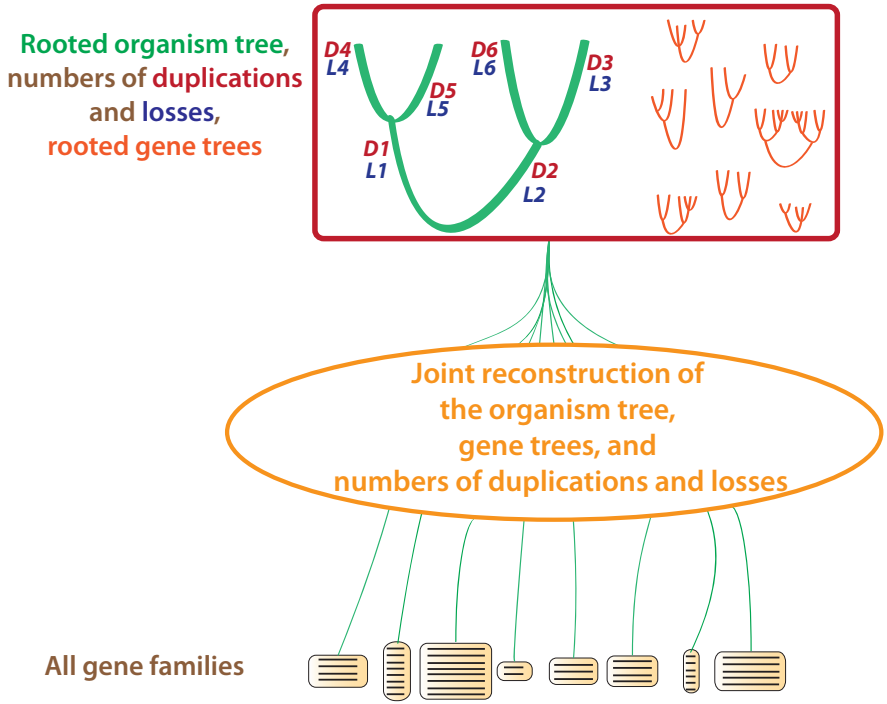
How to consider duplication and loss events



Numbers of duplications and losses obtained by this type of approach are **highly biased because of errors in gene trees**.

(Hahn MW, Genome Biol. 2007).

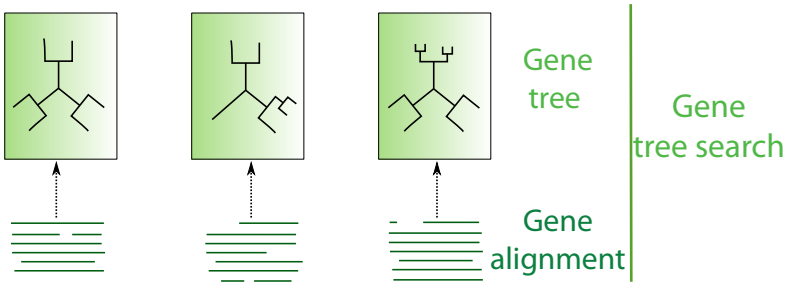
Our method: *PHYLDOG*



- Can use all gene families in a probabilistic framework (ML)
- Simultaneously:
  - reconstructs an organism tree
  - reconstructs gene trees
  - maps events of duplications and losses
  - estimates branch-wise expected numbers of duplications and losses per gene family

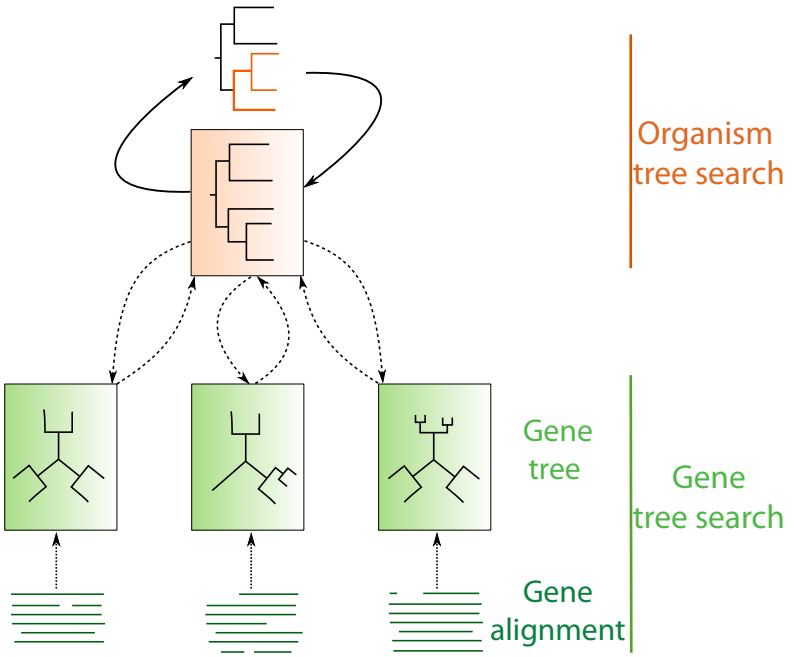
Computing the likelihood of organism and gene trees

Hierarchical structure



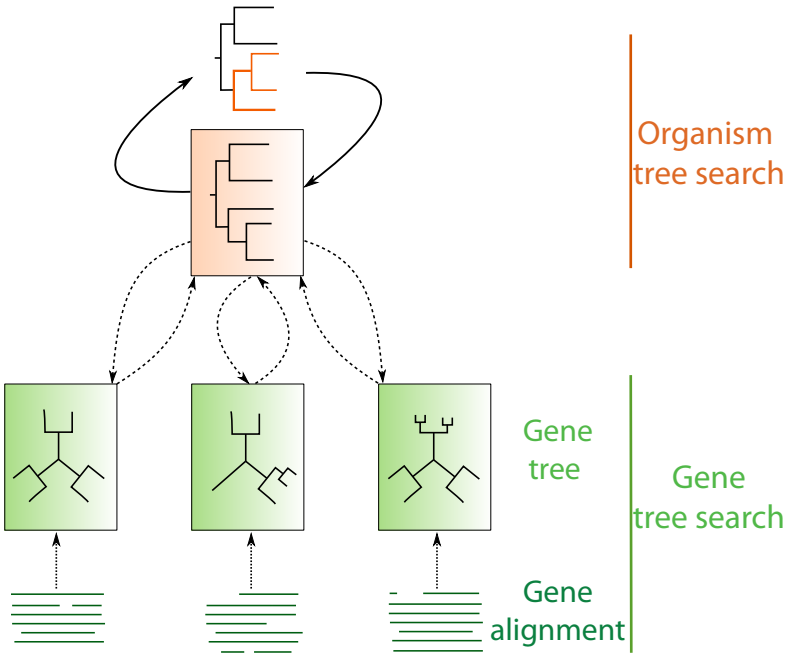
Computing the likelihood of organism and gene trees

Hierarchical structure



Computing the likelihood of organism and gene trees

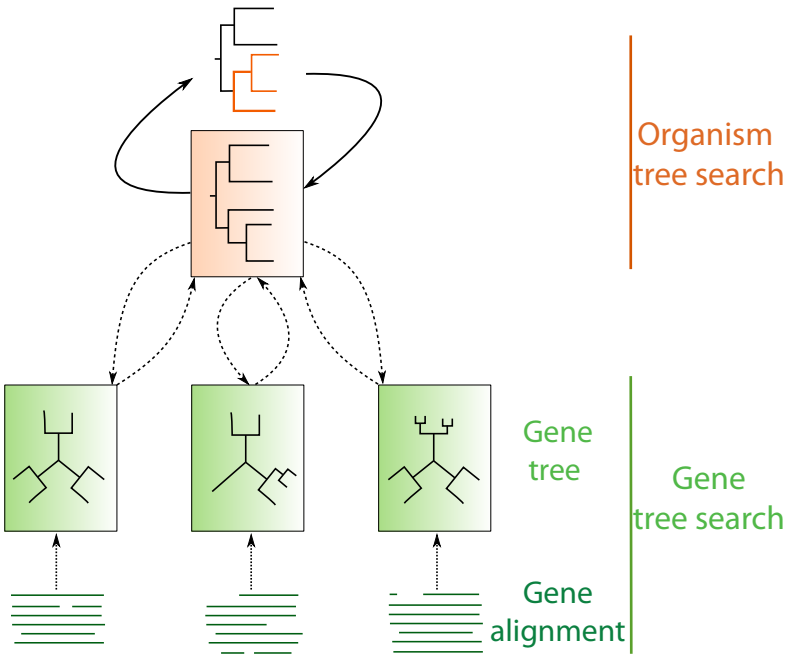
Hierarchical structure



$$L_G(family) = L_G(reconciliation) \times L_{G,Felsenstein}(gene\ tree)$$

Computing the likelihood of organism and gene trees

Hierarchical structure

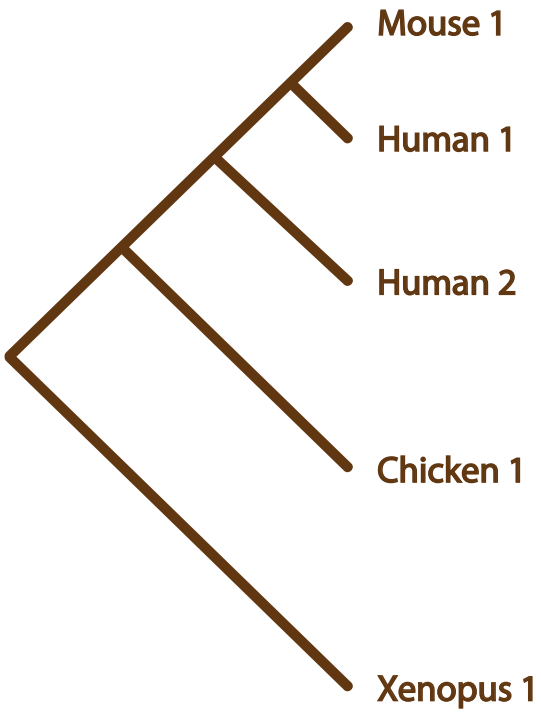


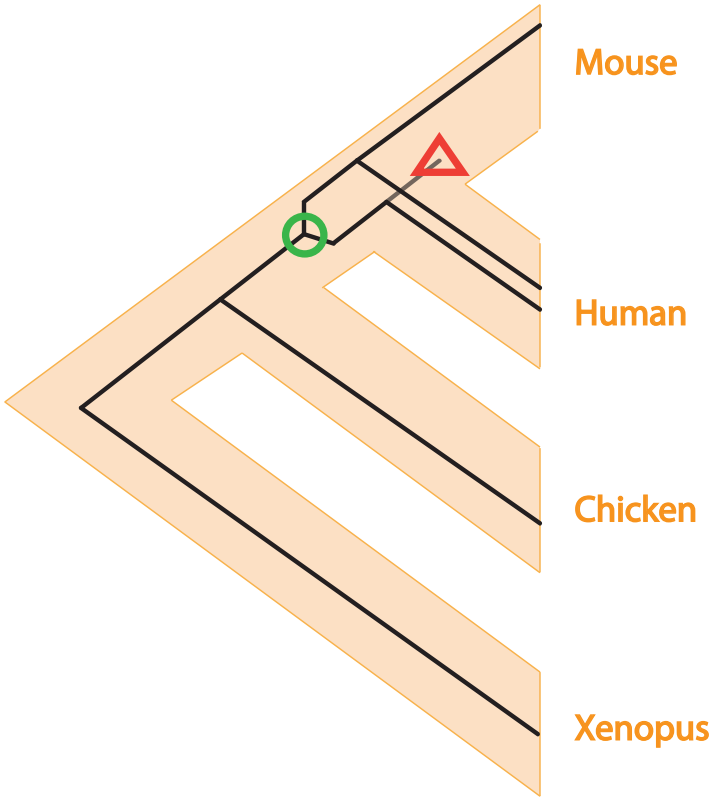
- **Likelihood of an organism tree** is a function of gene family likelihoods
- **Likelihood of a gene family**:
  1. Likelihood of a gene tree *sensu* Felsenstein (1981)
  2. Likelihood of an evolutionary scenario made of speciations, duplications and losses = **reconciliation**

$$L(\textit{Organism \& gene trees}) = \prod_{G \in \{\textit{All gene families}\}} L_G(\textit{family})$$

$$L_G(\textit{family}) = L_G(\textit{reconciliation}) \times L_{G,\textit{Felsenstein}}(\textit{gene tree})$$

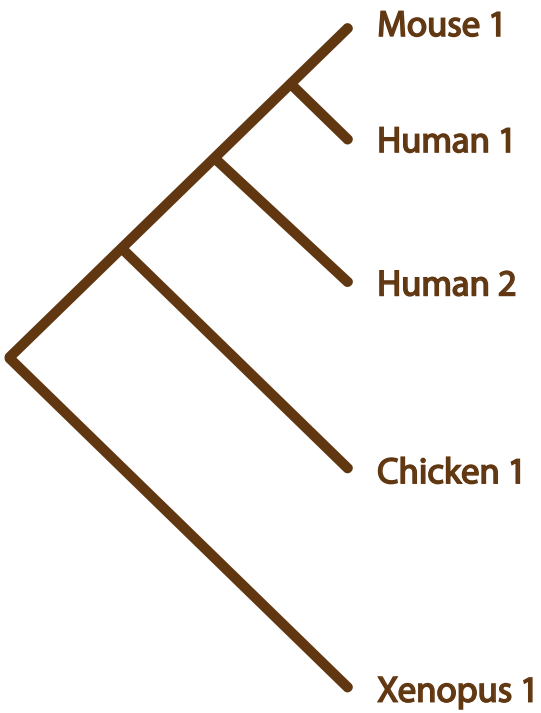
Scenarios considered by $L_G(reconciliation)$

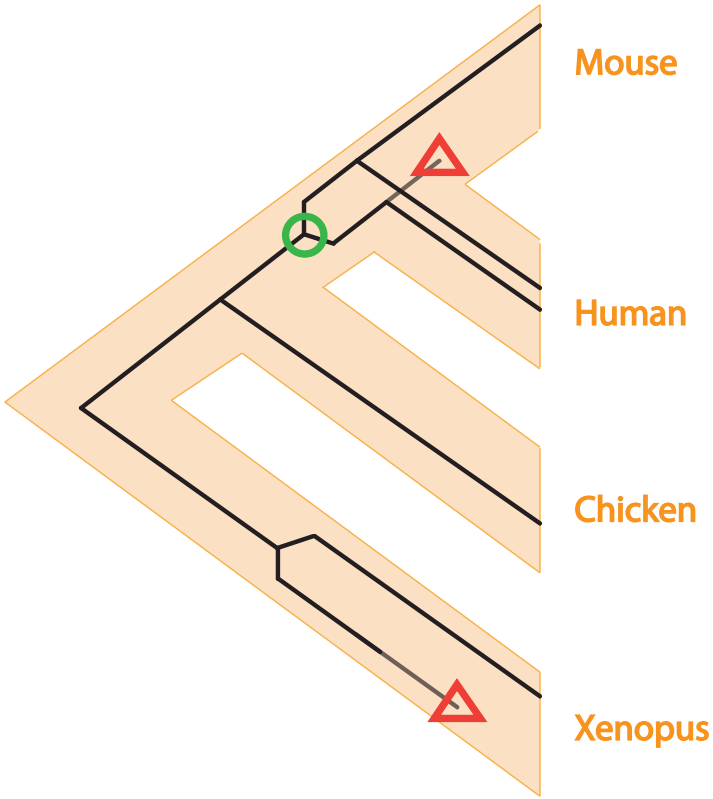




Most parsimonious scenario: **taken into account**

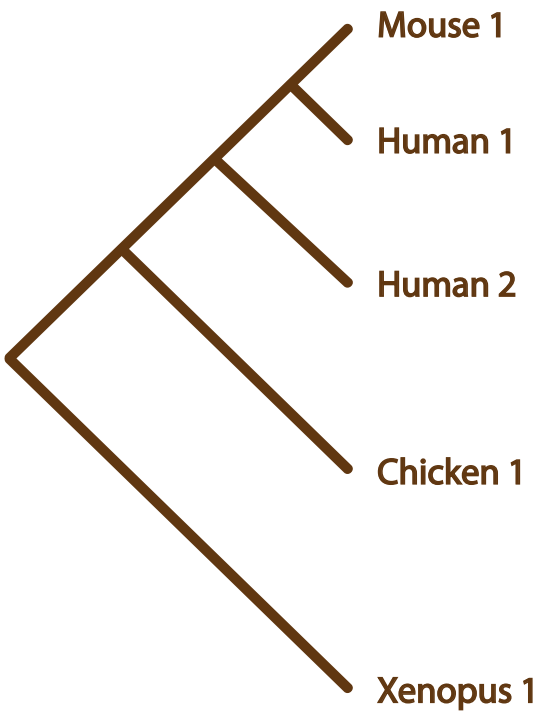
Scenarios considered by $L_G(reconciliation)$

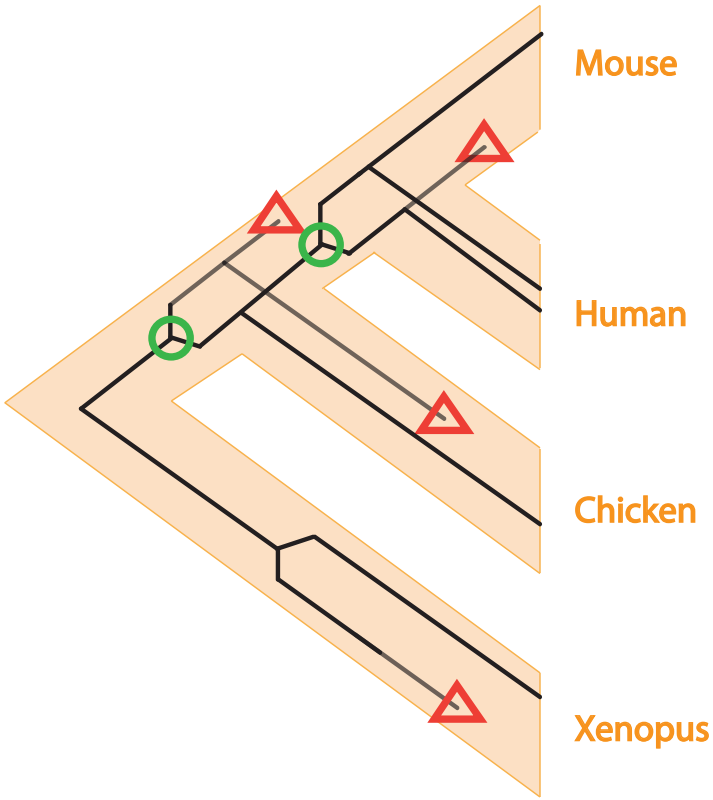




Most parsimonious scenario plus consecutive "hidden" duplications and losses in a single branch: **taken into account**

Scenarios considered by $L_G(reconciliation)$





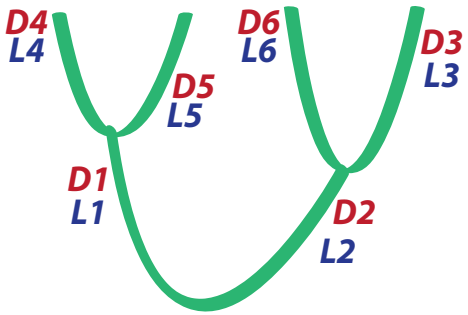
Scenario including a "hidden" duplication in one branch followed by losses in child branches: **NOT taken into account**

Tests on simulations

rooted organism tree

Tests on simulations

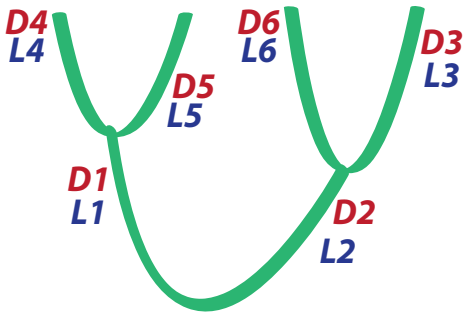
rooted
organism tree



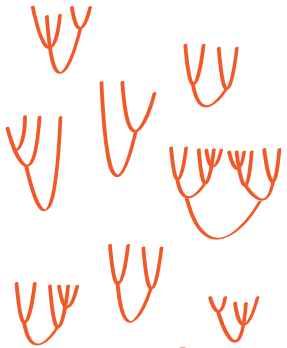
numbers of
duplications
and losses

Tests on simulations

rooted
organism tree



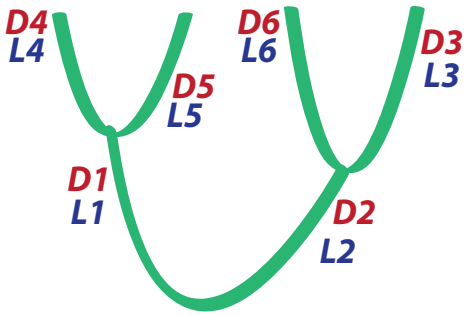
numbers of
duplications
and losses



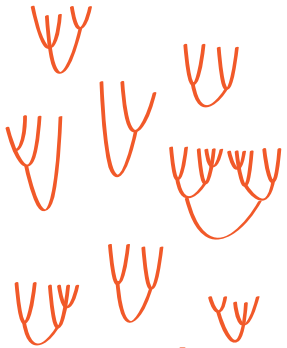
rooted
gene trees

Tests on simulations

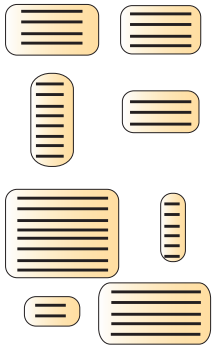
rooted
organism tree



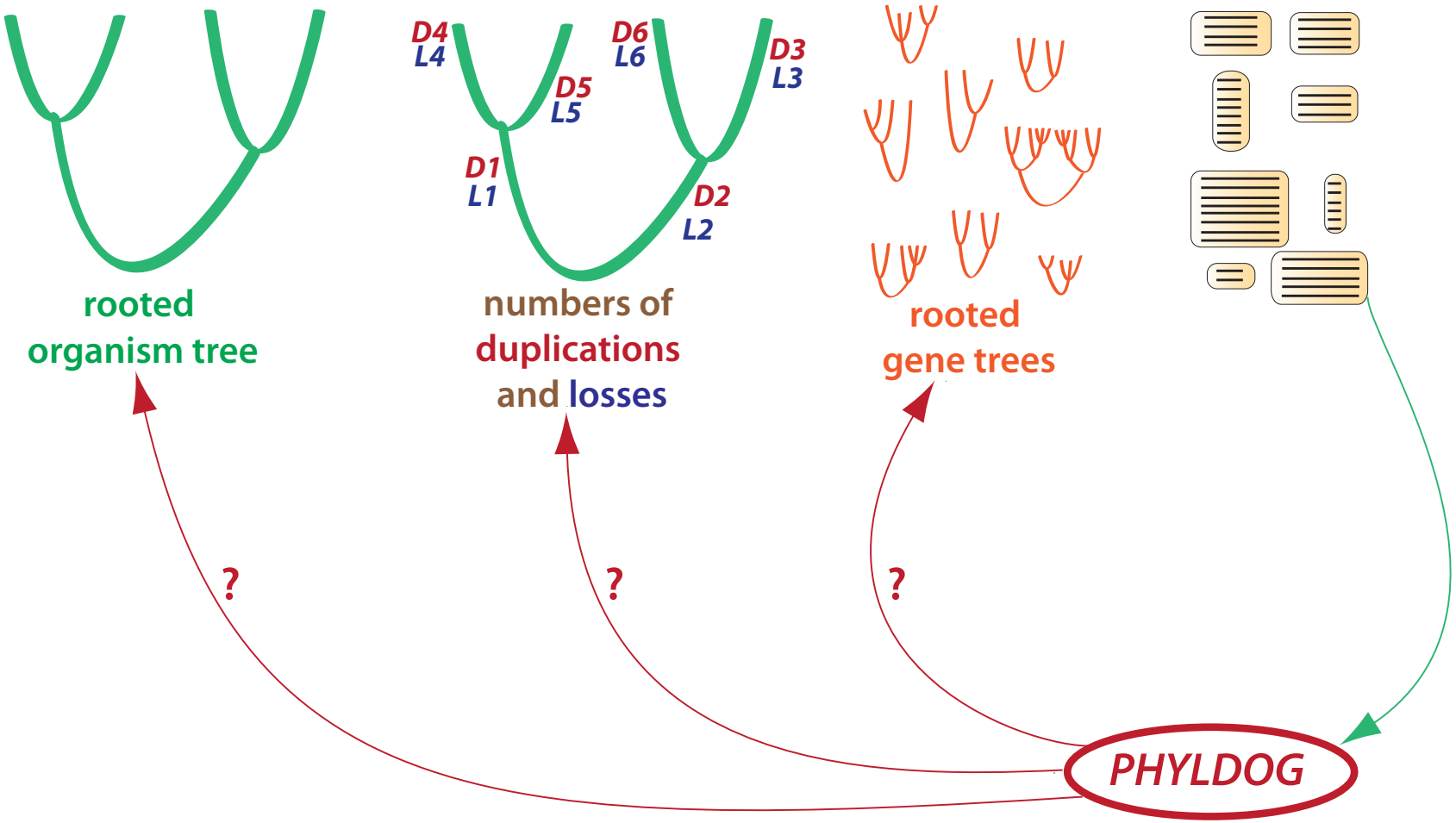
numbers of
duplications
and losses



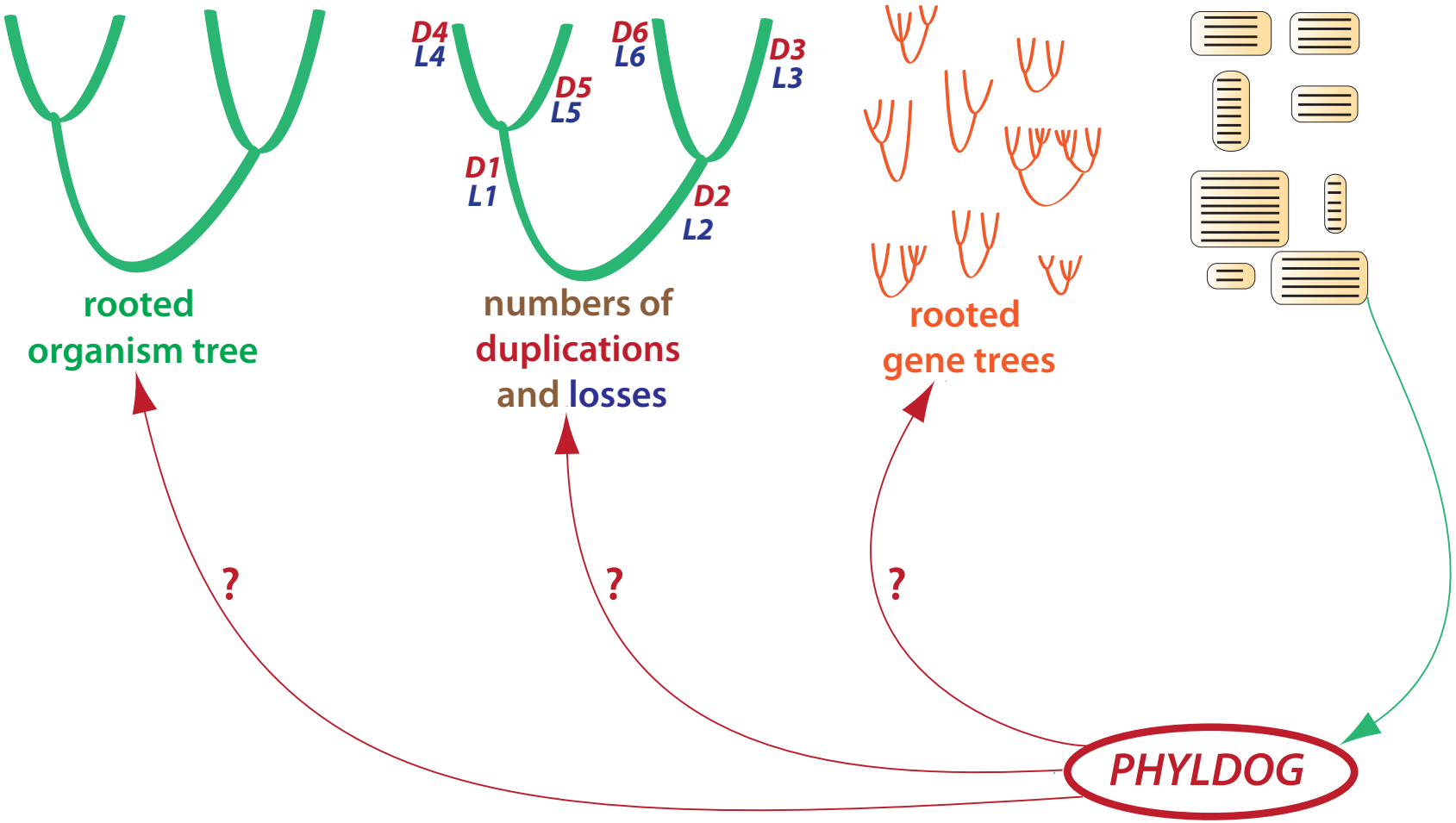
rooted
gene trees



Tests on simulations



Tests on simulations

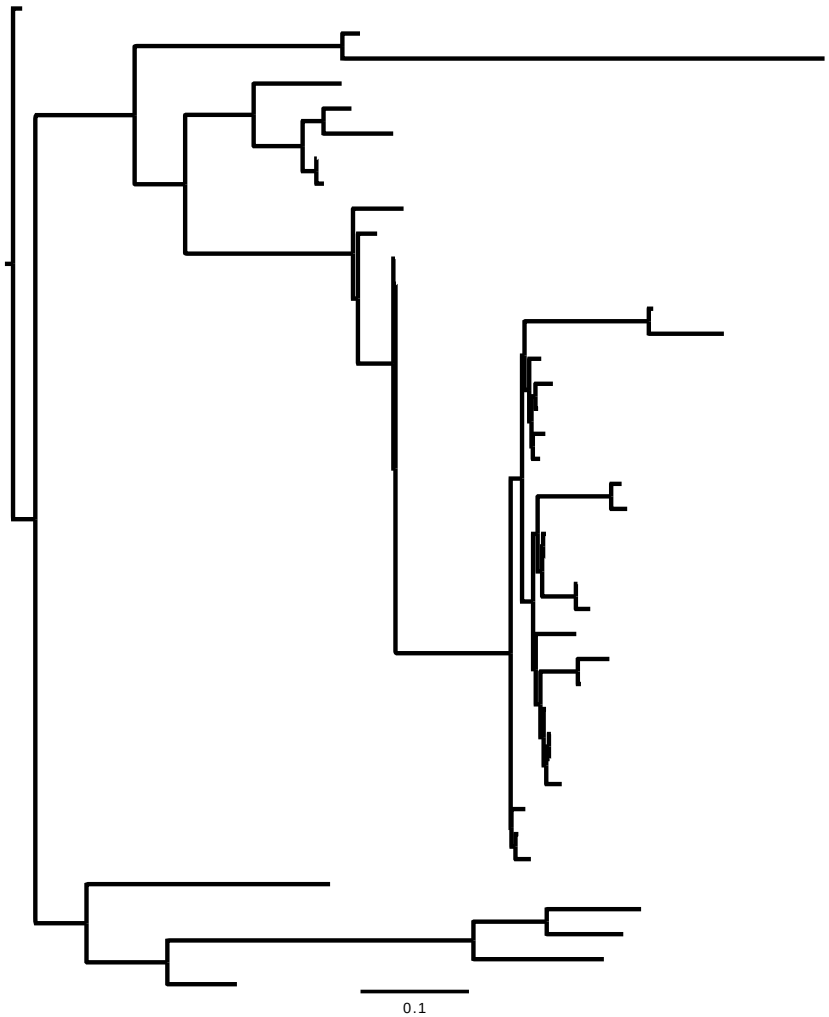


Simulations

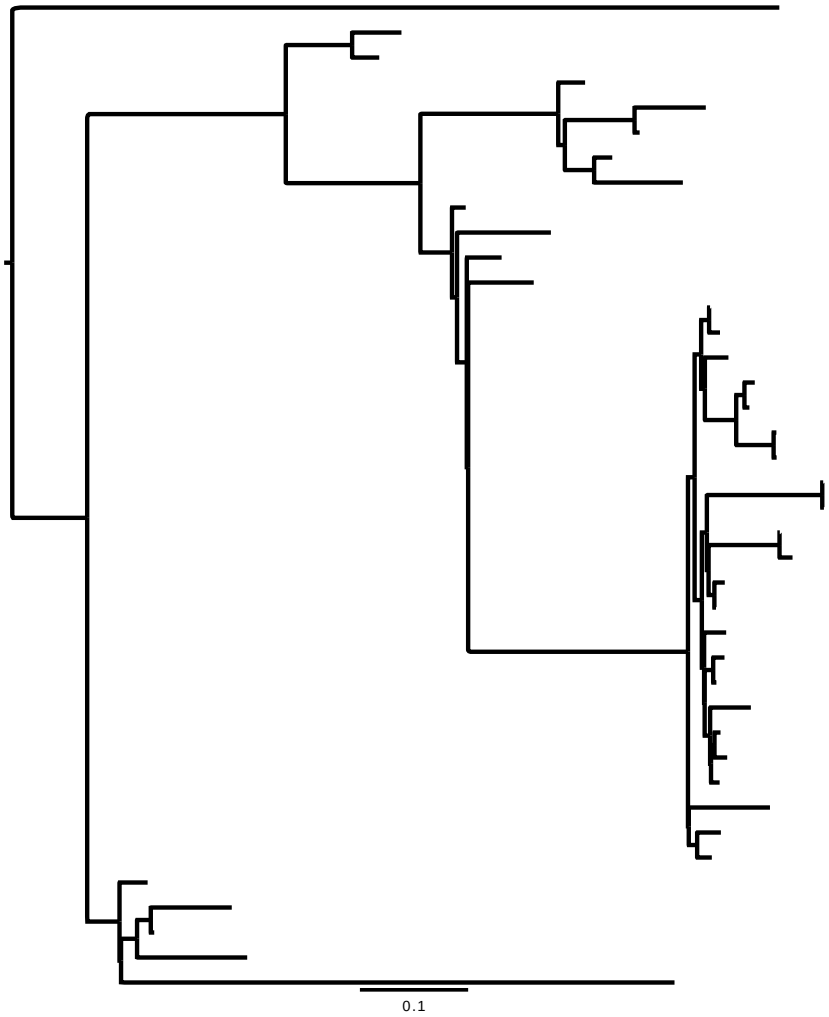
- 40 species
- Randomly pick duplication and loss rates per branch

Expected numbers

Duplications

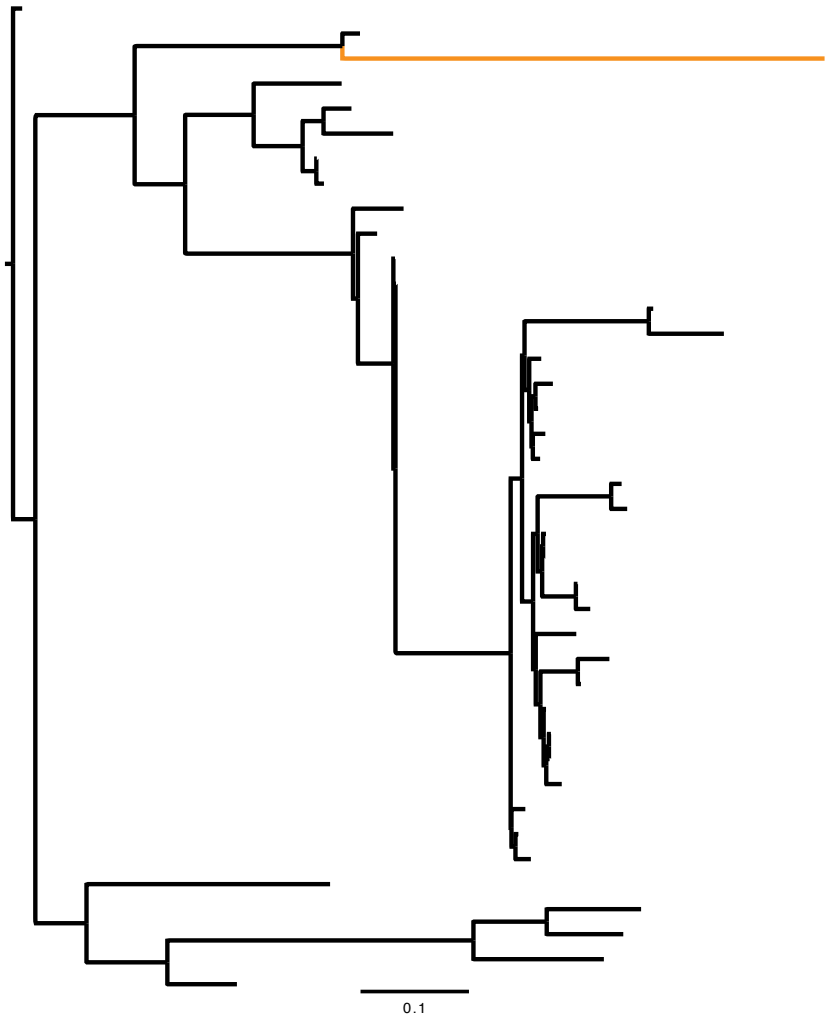


Losses

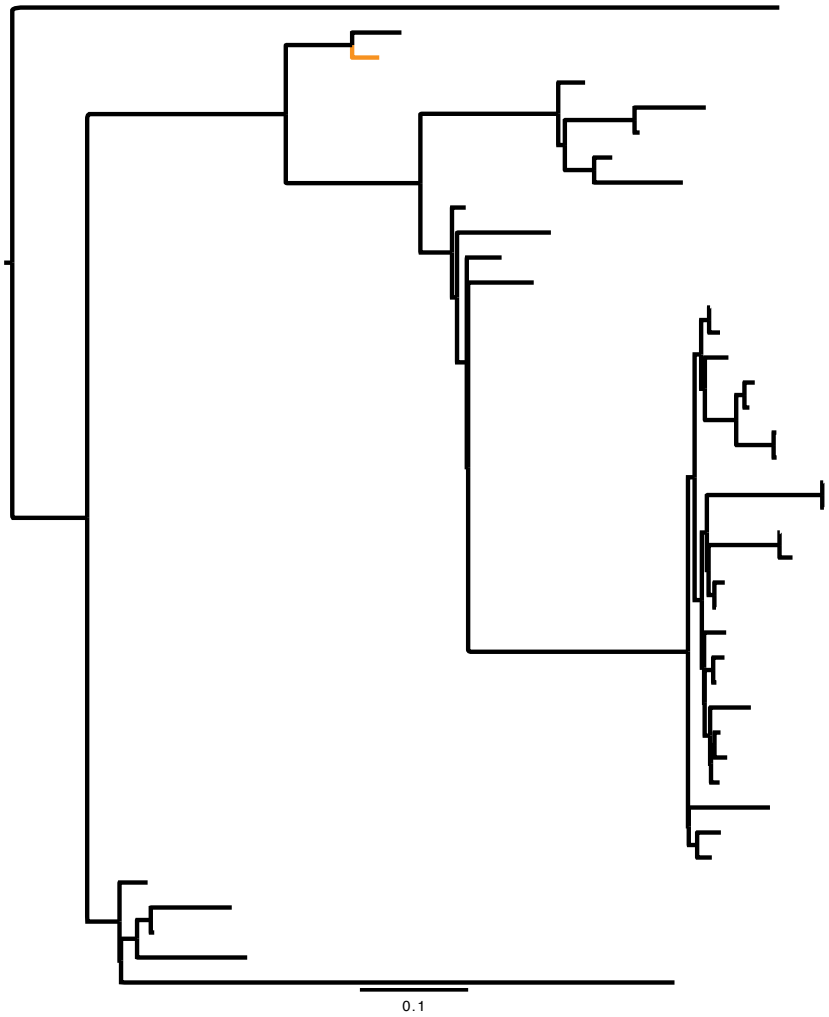


Expected numbers

Duplications



Losses



Tests on simulations

Simulations

- 40 species
- Randomly pick duplication and loss rates per branch
- Simulate 500 gene family trees using Sennblad and Lagergren's model (Syst. Biol., 2009), with randomly drawn branch lengths
- For each gene family tree, simulate 1000 base sequences using the HKY85+GI model

Tests on simulations

Simulations

- 40 species
- Randomly pick duplication and loss rates per branch
- Simulate 500 gene family trees using Sennblad and Lagergren's model (Syst. Biol., 2009), with randomly drawn branch lengths
- For each gene family tree, simulate 1000 base sequences using the HKY85+GI model

Reconstruction through *PHYLDOG*

- Start from a random organism tree
- Use a (very simple) JC69 model
- Use 80 processors in parallel

Tests on simulations

Simulations

- 40 species
- Randomly pick duplication and loss rates per branch
- Simulate 500 gene family trees using Sennblad and Lagergren's model (Syst. Biol., 2009), with randomly drawn branch lengths
- For each gene family tree, simulate 1000 base sequences using the HKY85+GI model

Reconstruction through *PHYLDOG*

- Start from a random organism tree
- Use a (very simple) JC69 model
- Use 80 processors in parallel

Questions

1. Can the true rooted organism tree be recovered?
2. Can the true gene trees be recovered?

Tests on simulations

Simulations

- 40 species
- Randomly pick duplication and loss rates per branch
- Simulate 500 gene family trees using Sennblad and Lagergren's model (Syst. Biol., 2009), with randomly drawn branch lengths
- For each gene family tree, simulate 1000 base sequences using the HKY85+GI model

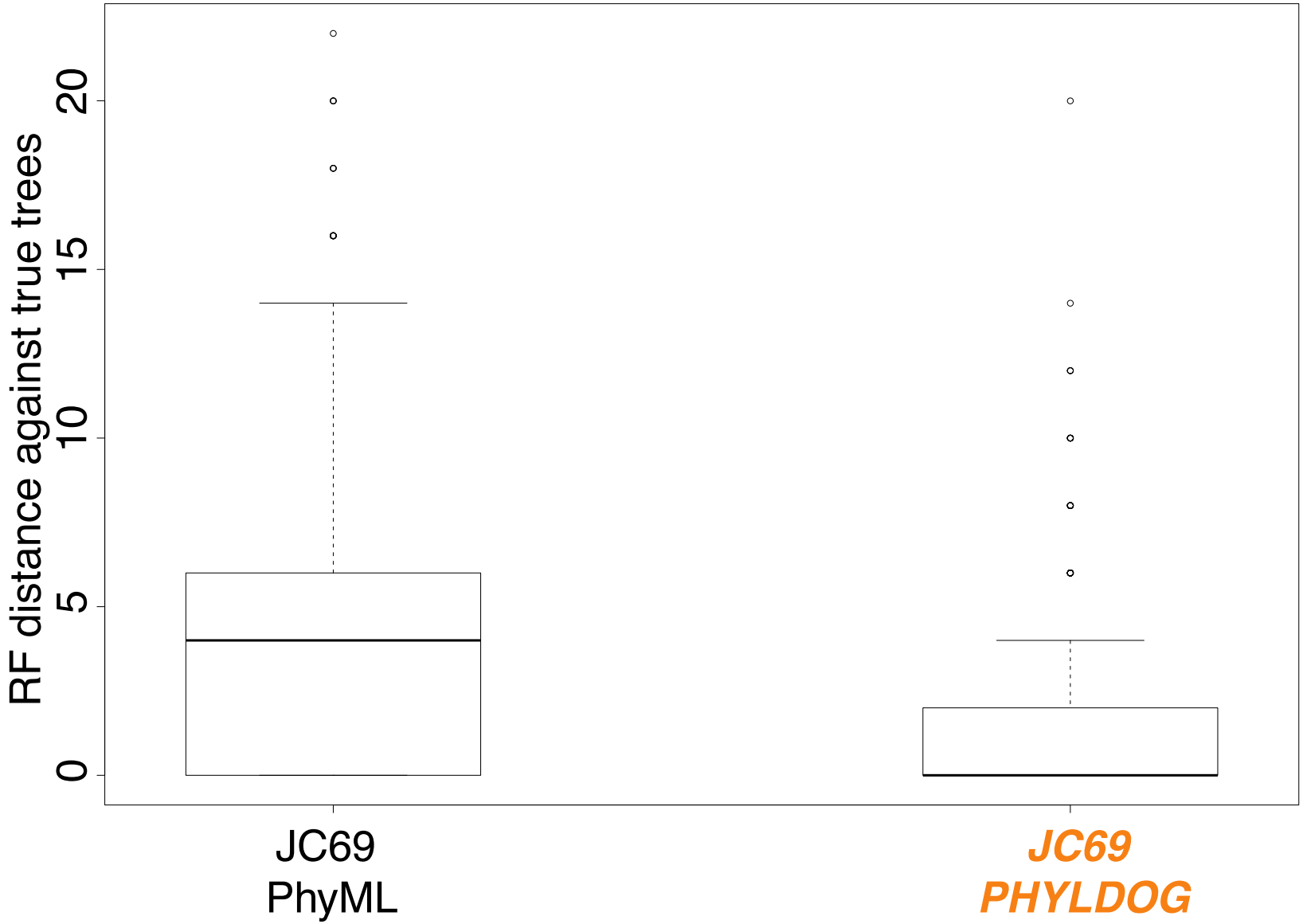
Reconstruction through *PHYLDOG*

- Start from a random organism tree
- Use a (very simple) JC69 model
- Use 80 processors in parallel

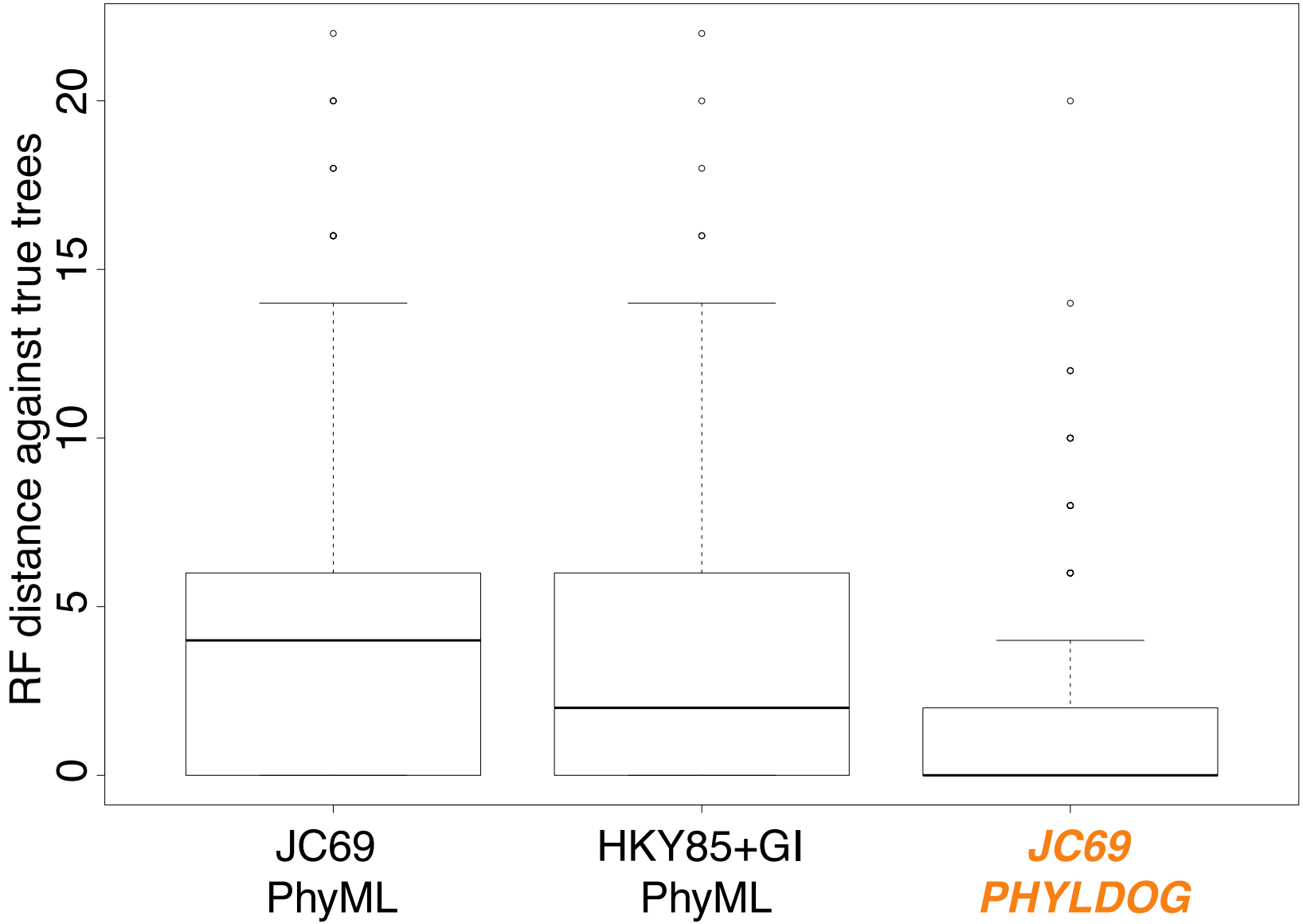
Questions

1. Can the true rooted organism tree be recovered? **YES**
2. Can the true gene trees be recovered?

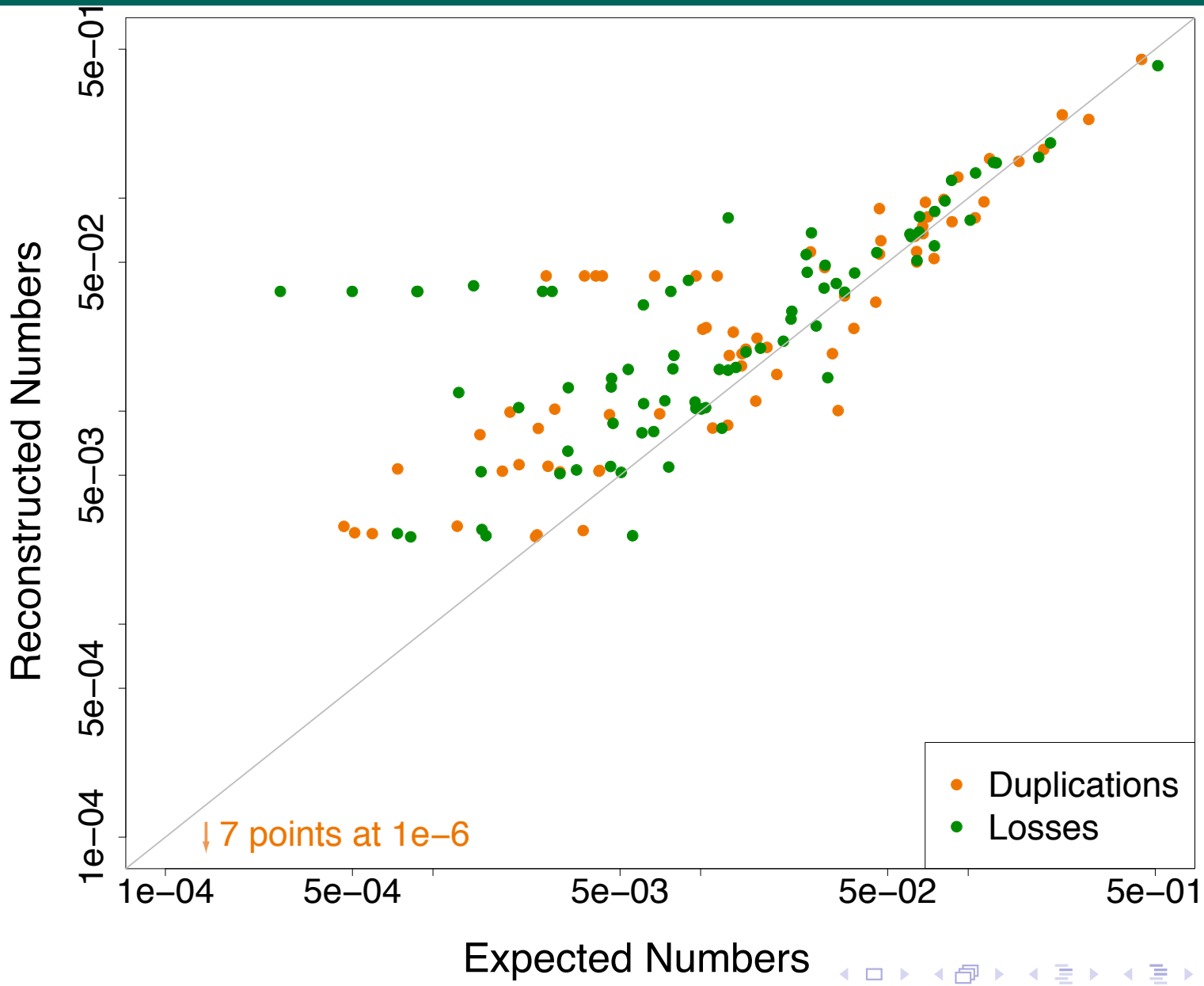
Good gene trees are obtained



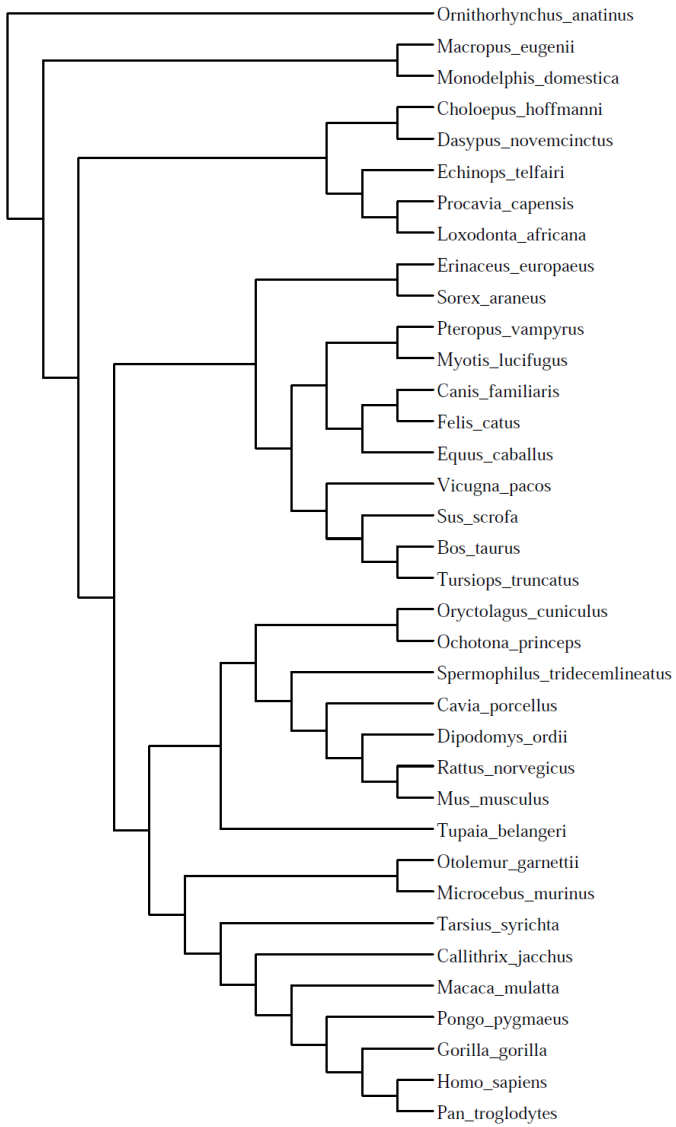
Good gene trees are obtained



The expected numbers of events can be recovered

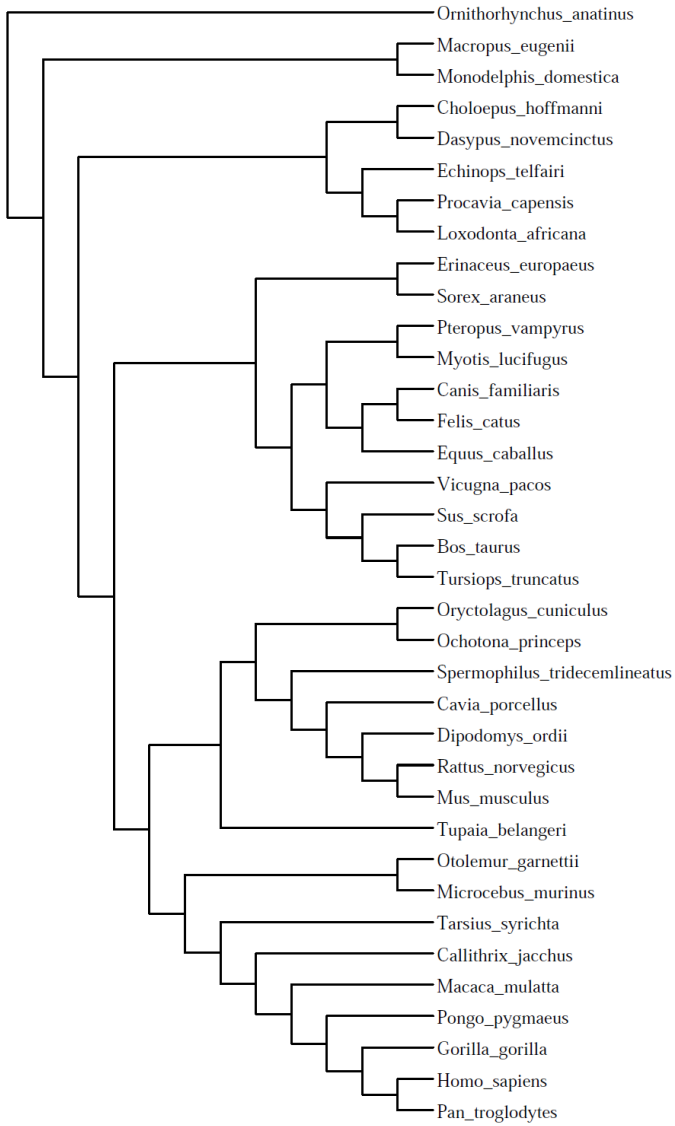


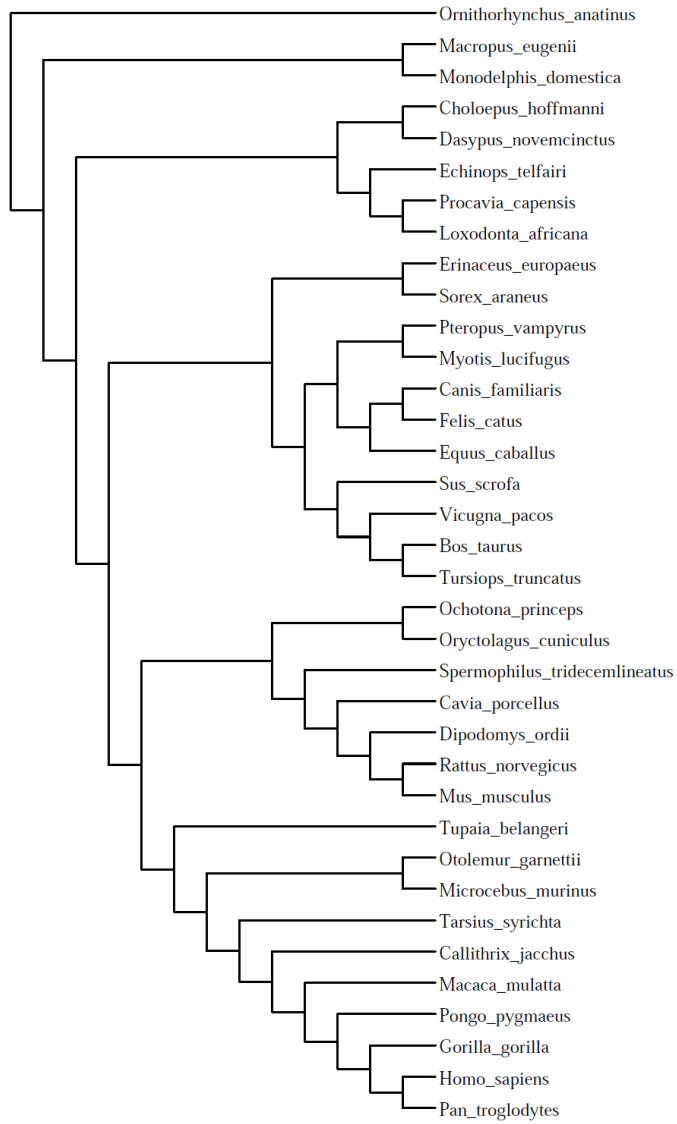
Application to the tree of Mammals



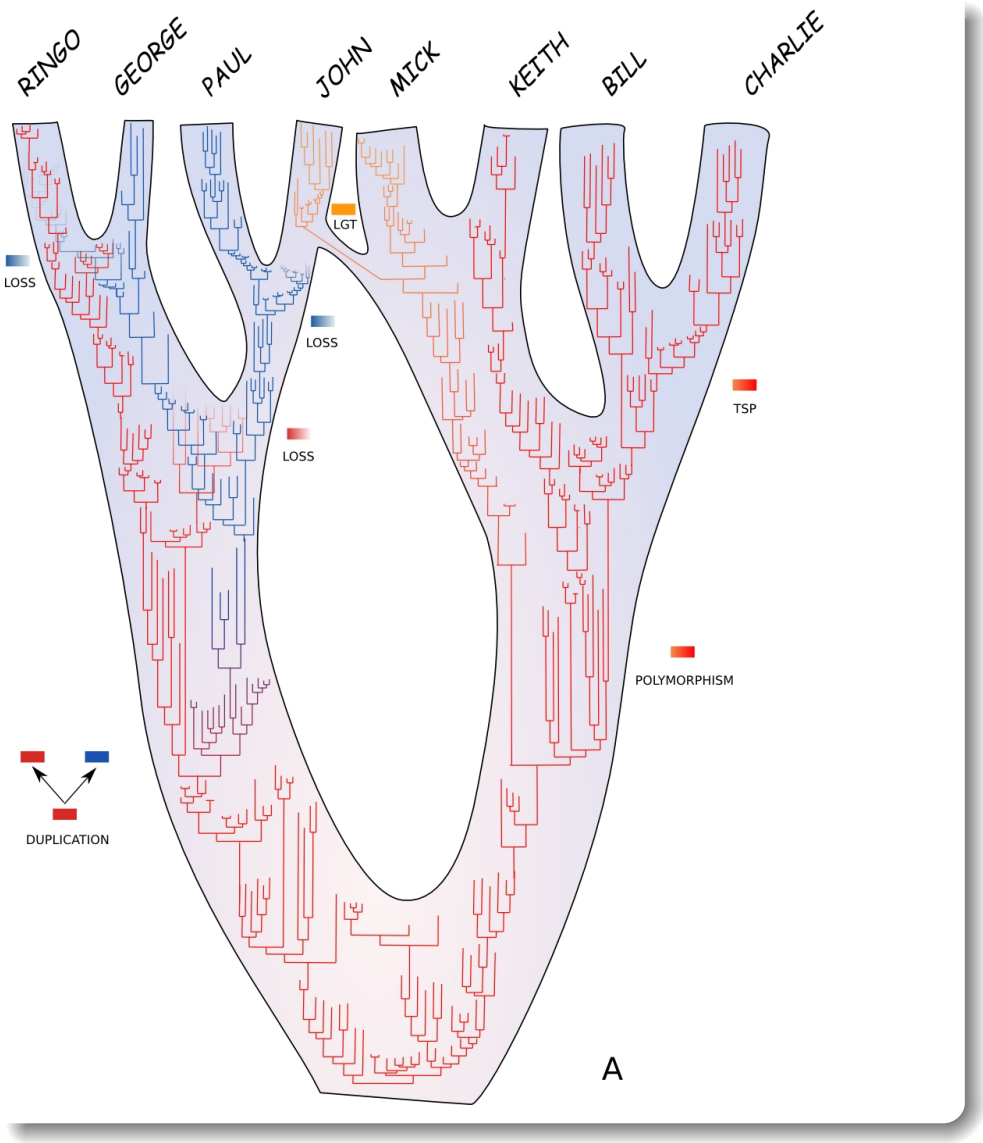
- Ensembl v57
- families with more than 2 genes and less than 100 genes
- alignments with more than 500 sites
- $\Longrightarrow$ 9771 gene families

Application to the tree of Mammals





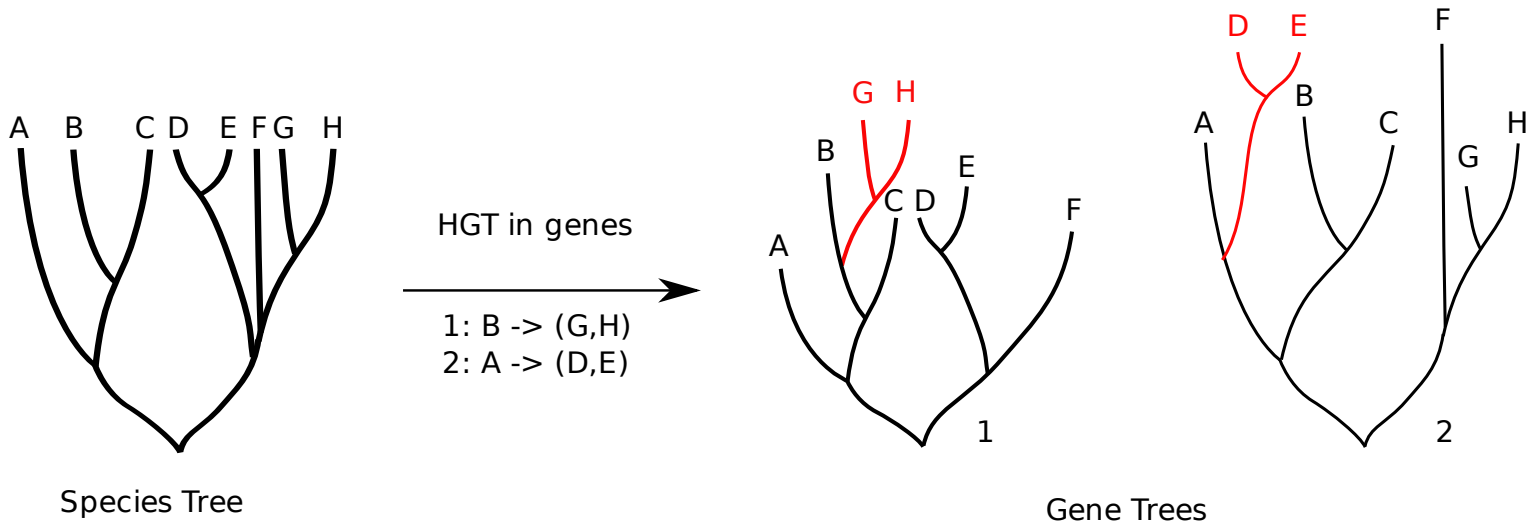
Using genetic information to reconstruct the history of life



In this talk:

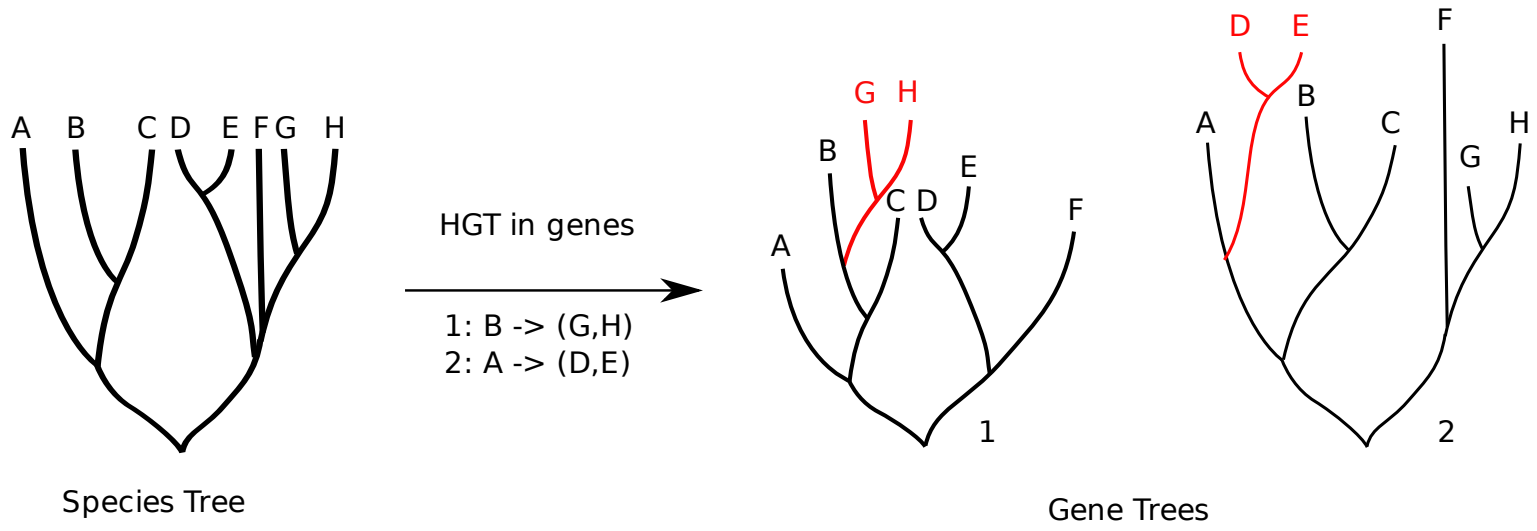
- Phyldog: Taking duplication and loss into account (B. Boussau)
- HGT as a molecular clock: a model with duplication, transfer and loss (DTL) (G. Szollosi)

HGT contains information on the relative order of speciations



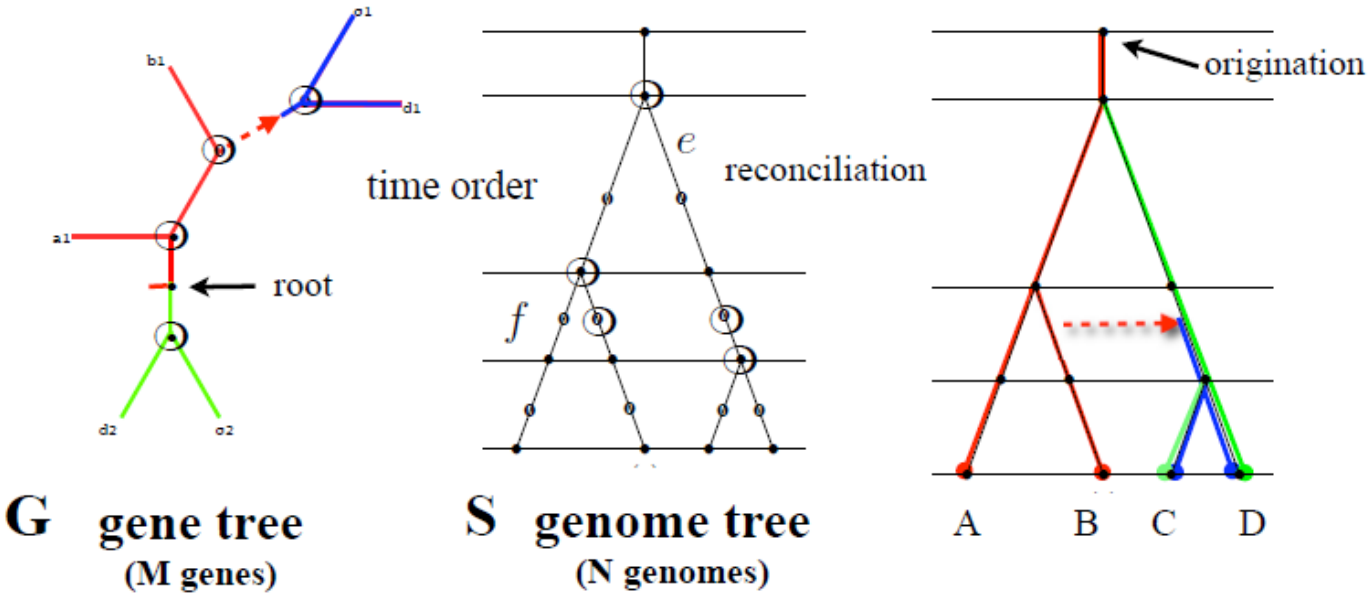
- The ancestor of (G,H) is more recent than the ancestor of (B,C)

HGT contains information on the relative order of speciations



- The ancestor of (G,H) is more recent than the ancestor of (B,C)
- The ancestor of (D,E) is more recent than the ancestor of (A,B,C)

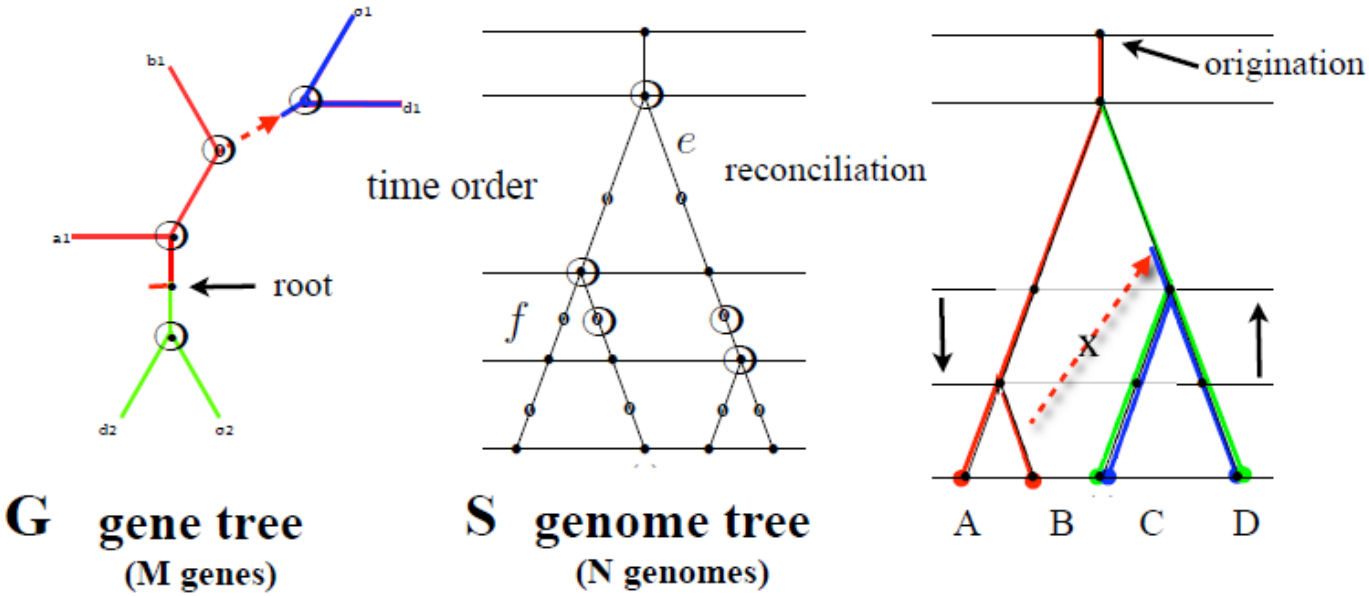
A model for DTL reconciliation with "time slices"



- Constructing a probabilistic model of reconciliation that includes transfer requires time order
- Tofigh et al., 2009 $\Rightarrow$ $\mathbf{L}($ $\mathbf{|\ DTL\ rates,\ S)}$

$$\mathbf{L(\ S\ ) = L(}\ \text{ }\ \mathbf{) \times L(}\ \text{ }\ \mathbf{) \times L(}\ \text{ }\ \mathbf{) \times \ldots}$$

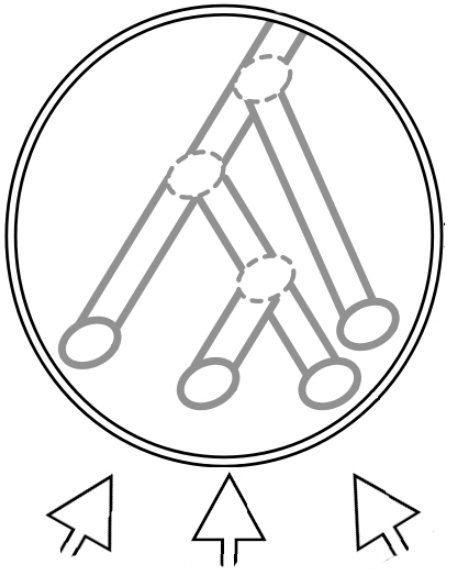
A model for DTL reconciliation with "time slices"

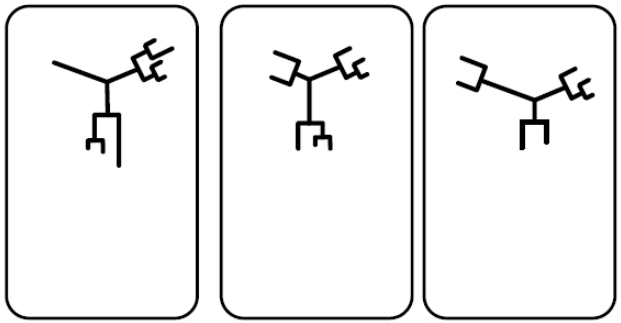


- Constructing a probabilistic model of reconciliation that includes transfer requires time order
- Tofigh et al., 2009 $\Rightarrow$ $\mathbf{L}($ $\mathbf{|\ DTL\ rates,\ S)}$

$$\mathbf{L(\ S\ ) = L(}\ \text{ }\ \mathbf{) \times L(}\ \text{ }\ \mathbf{) \times L(}\ \text{ }\ \mathbf{) \times \ldots}$$

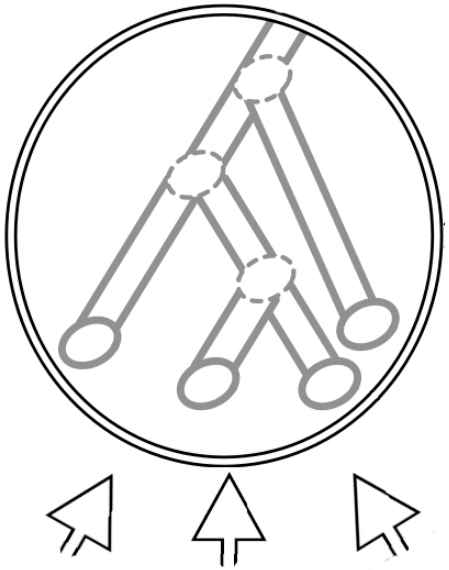
Reconstructing "time order" in a species tree

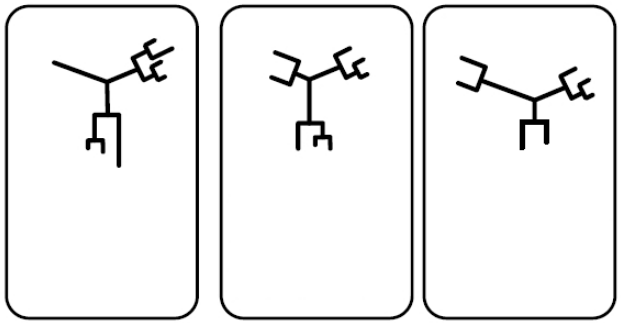




- Fixed gene trees

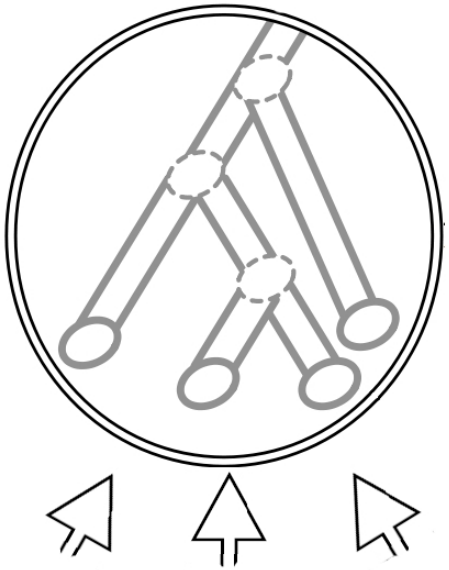
Reconstructing "time order" in a species tree

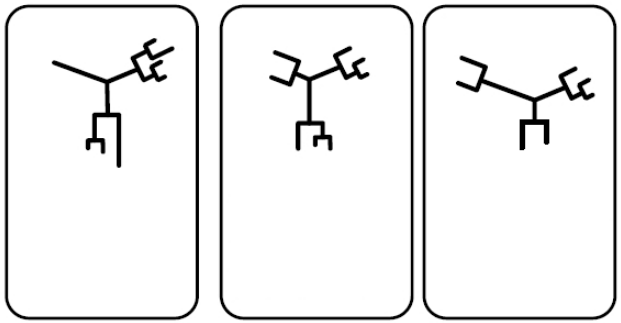




- Fixed gene trees
- "Reasonable" starting species tree (or fixed species tree)

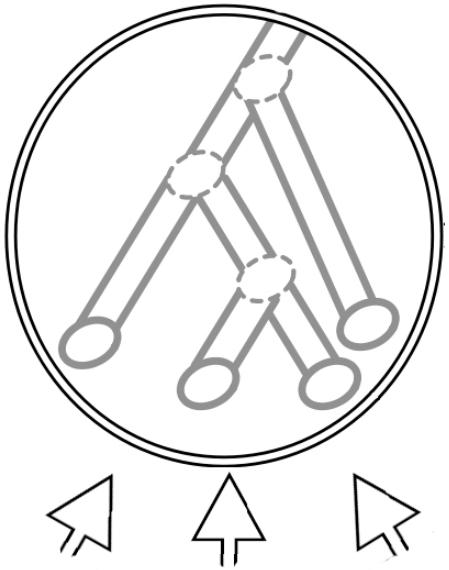
Reconstructing "time order" in a species tree

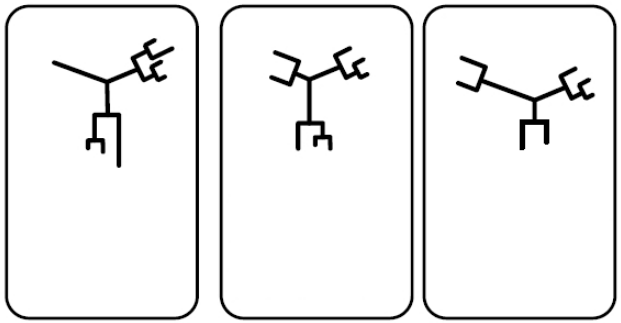




- Fixed gene trees
- "Reasonable" starting species tree (or fixed species tree)
- Exploration of "time orders"

Reconstructing "time order" in a species tree

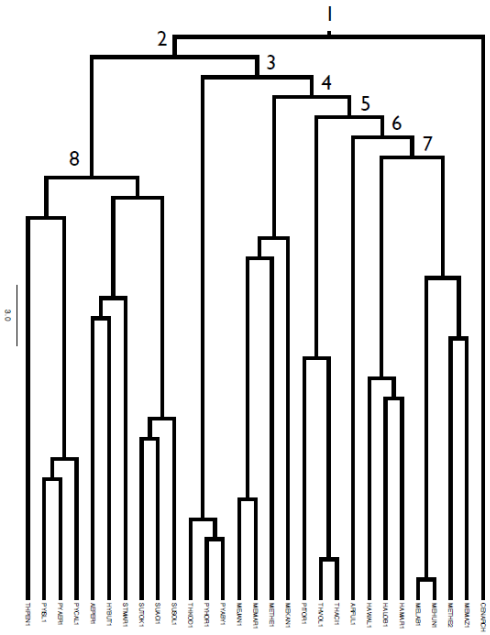




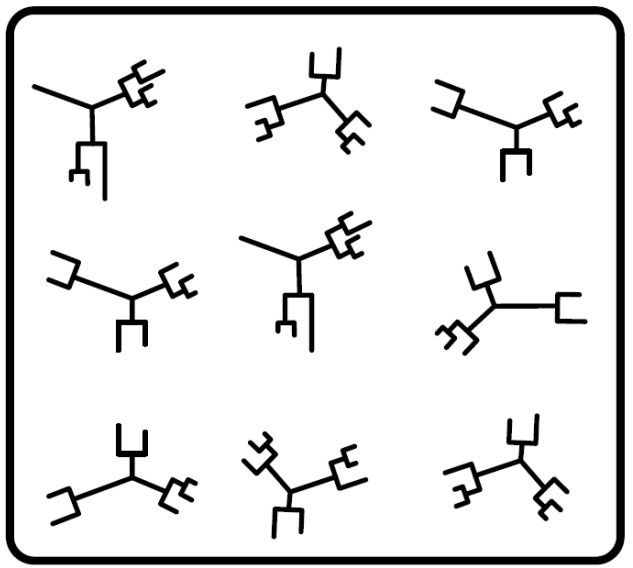
- Fixed gene trees
- "Reasonable" starting species tree (or fixed species tree)
- Exploration of "time orders"
- Estimation of all DTL parameters

Simulation on the DTL model

S with time order

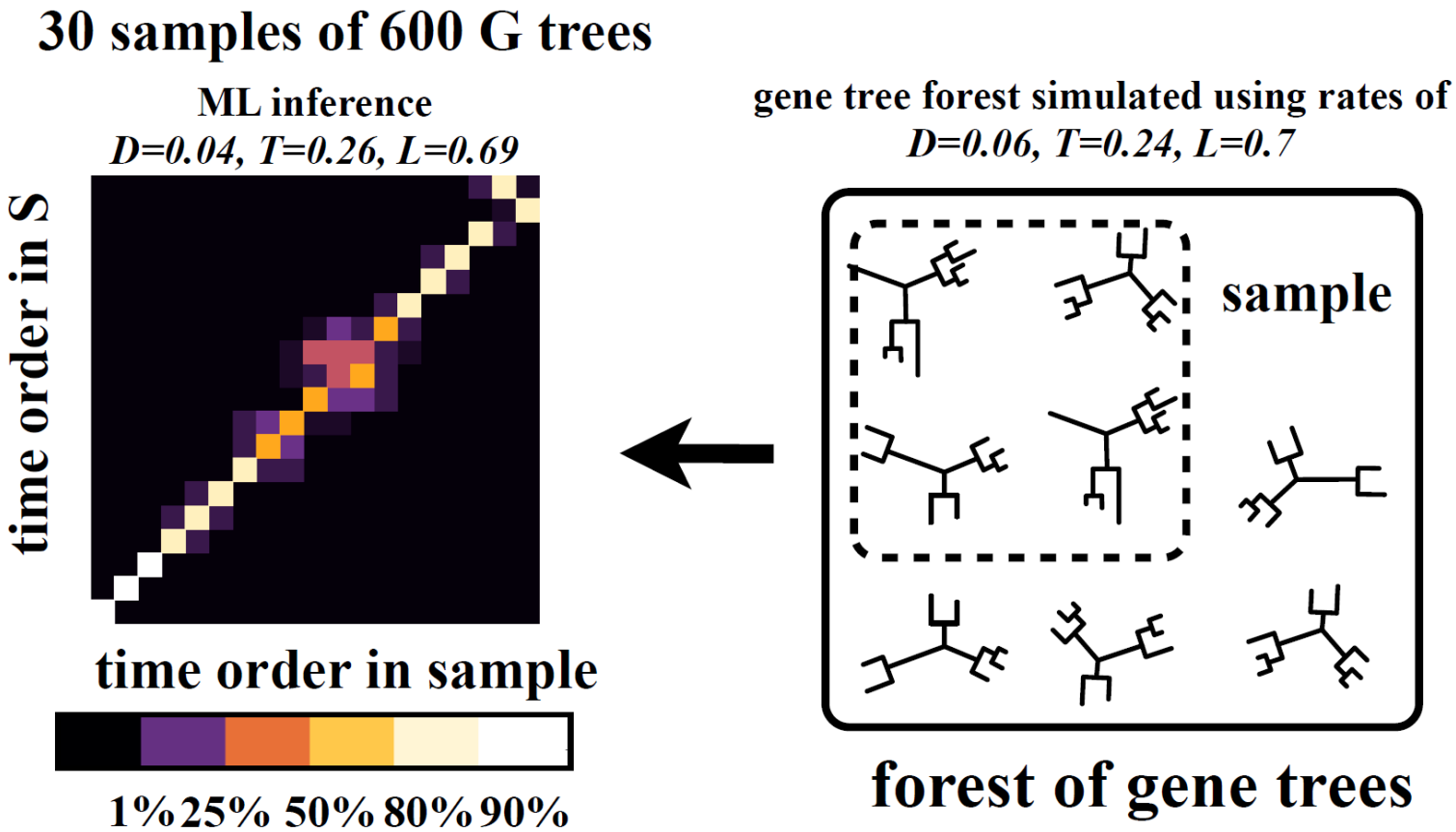


gene tree forest simulated using rates of
D=0.06, T=0.24, L=0.7

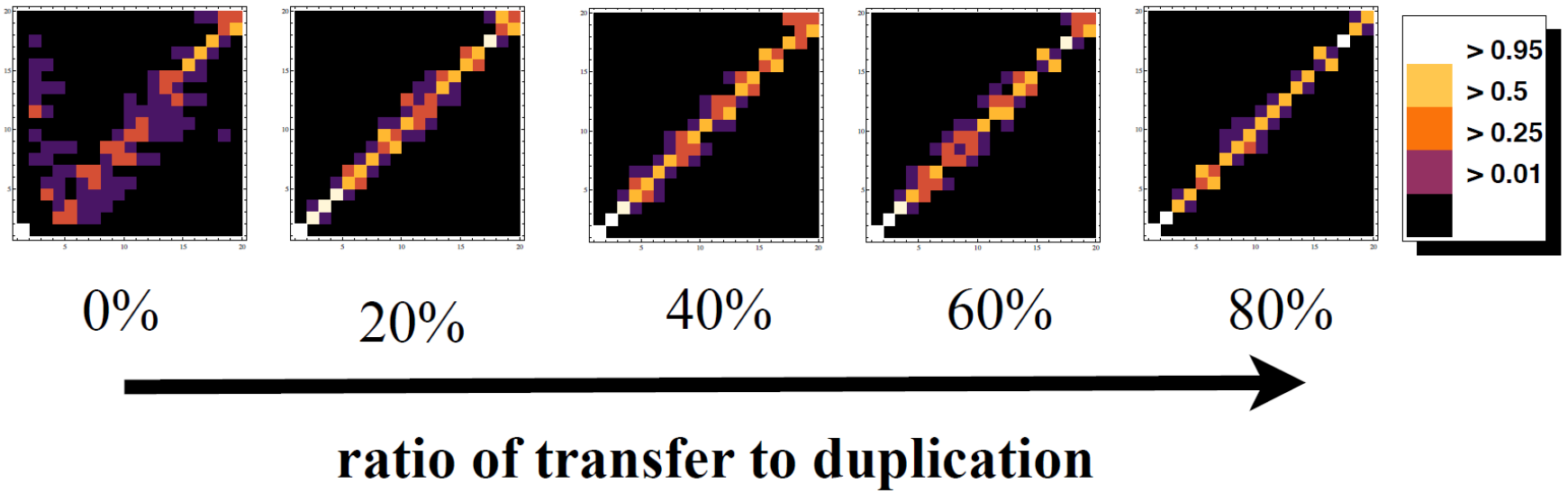


forest of gene trees

Simulation on the DTL model



Sensitivity to TH

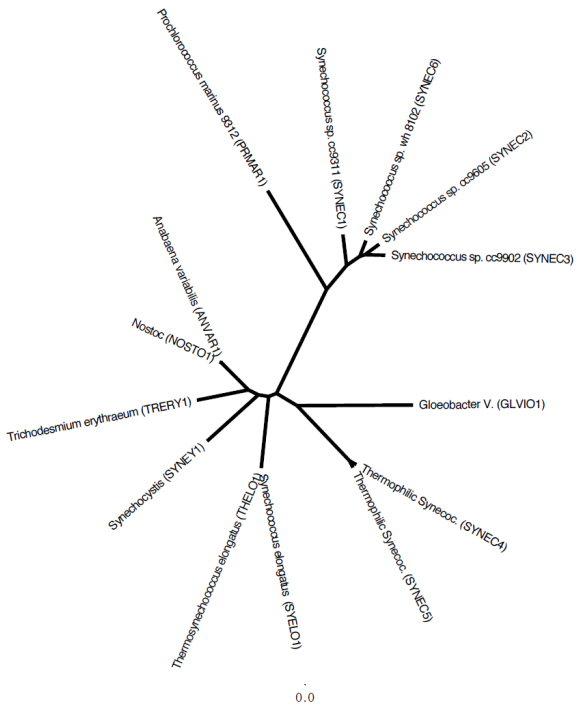


Inference is robust to several caveats:

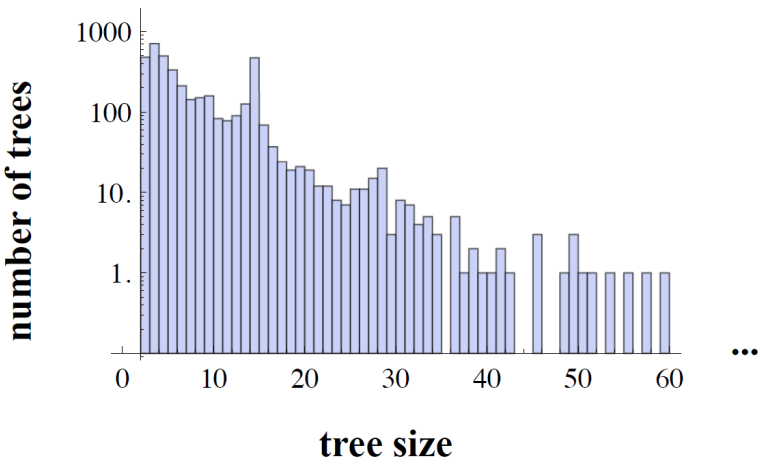
- Different degree of transfer
- Origination inside the tree
- Transfer from outside
- Rate heterogeneity

Application to Cyanobacteria

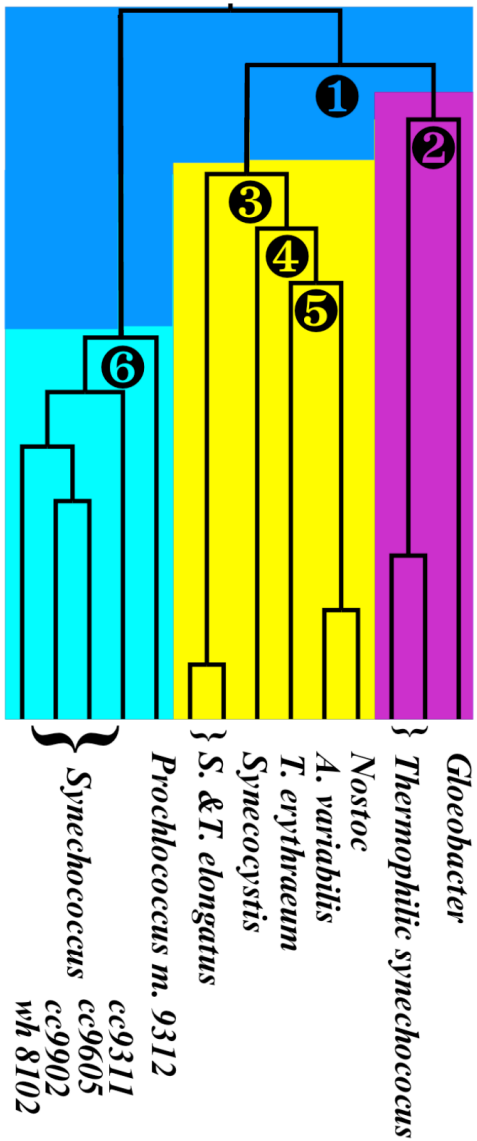
unrooted reference tree
14 species
concatenate of universal genes

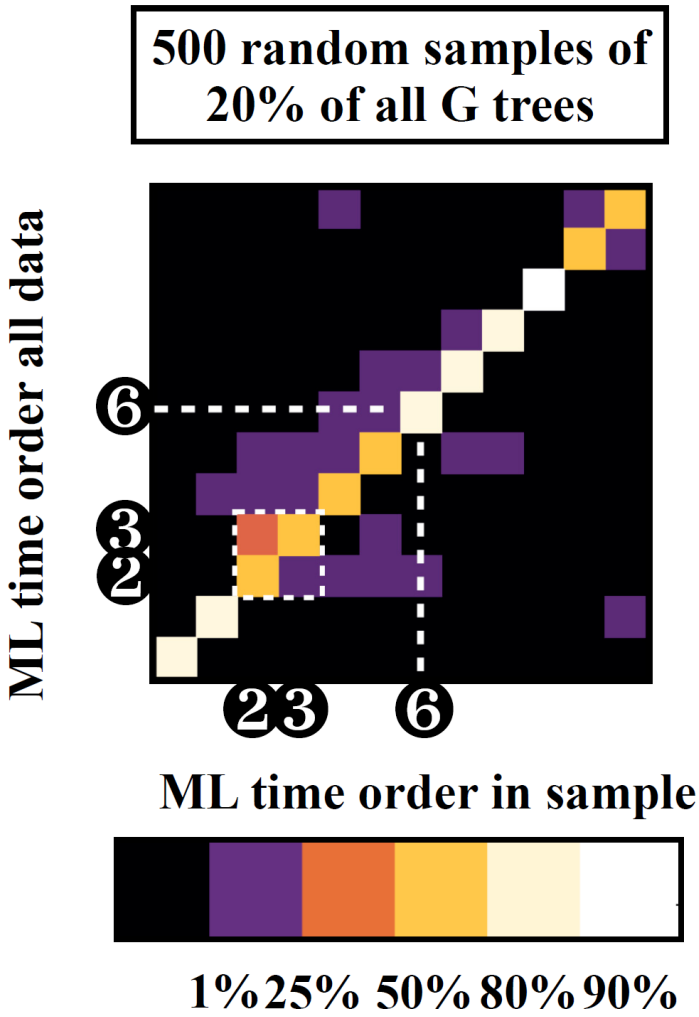


3886 G trees
31 263 genes
All families with trees in HOGENOM 4

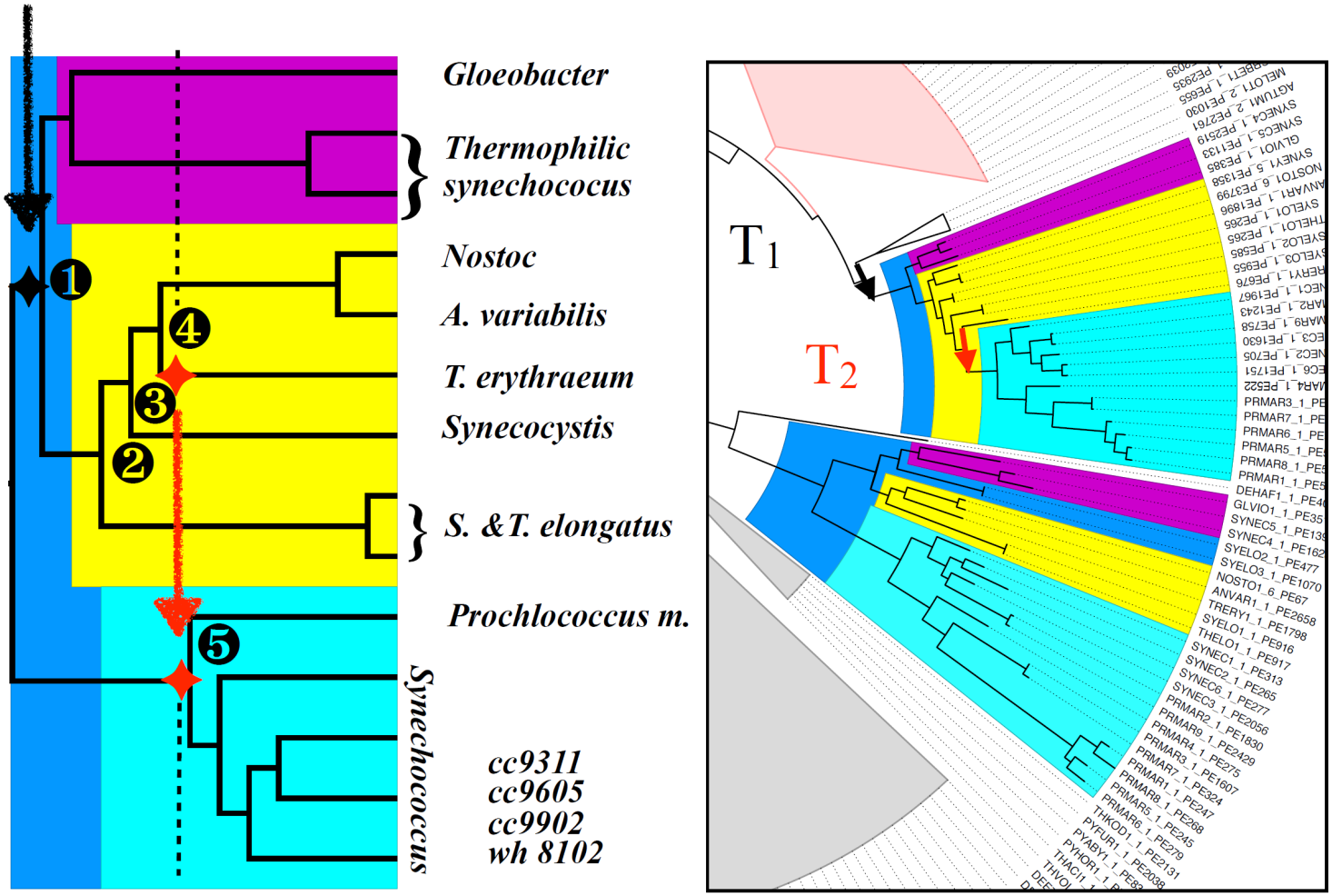


Time order in the tree of Cyanobacteria



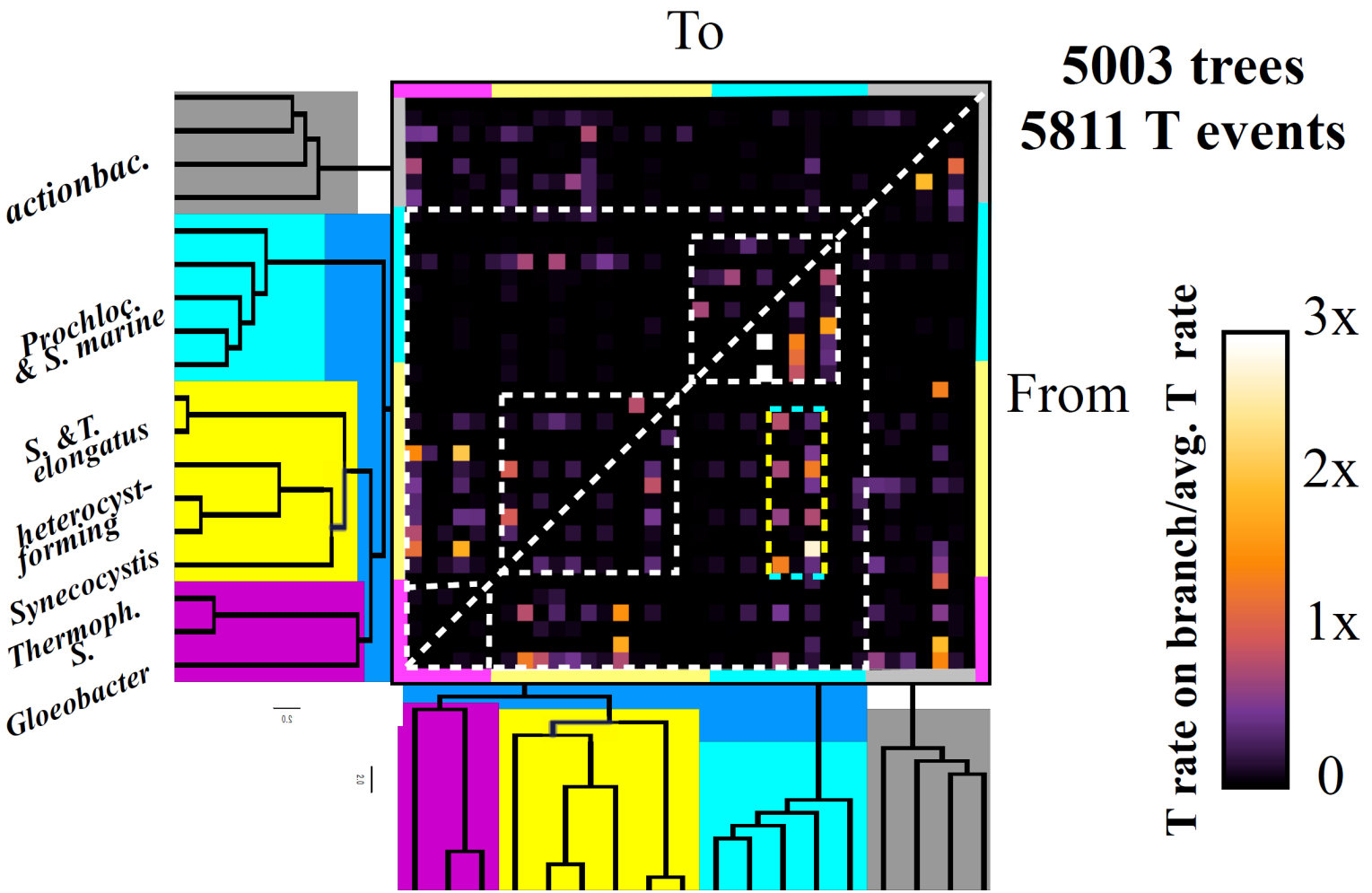


Transfer events in the history of Cyanobacteria



Cob(I)alamin Adenosyltransferase related protein

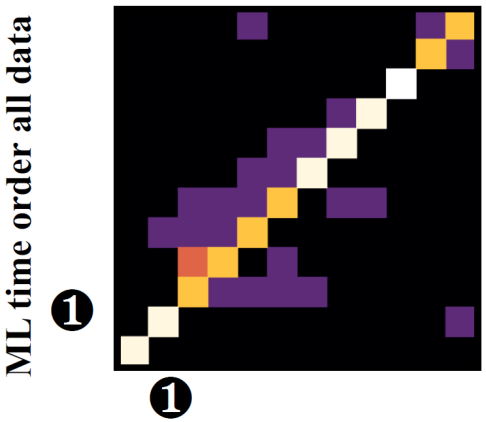
Transfer events in the history of Cyanobacteria



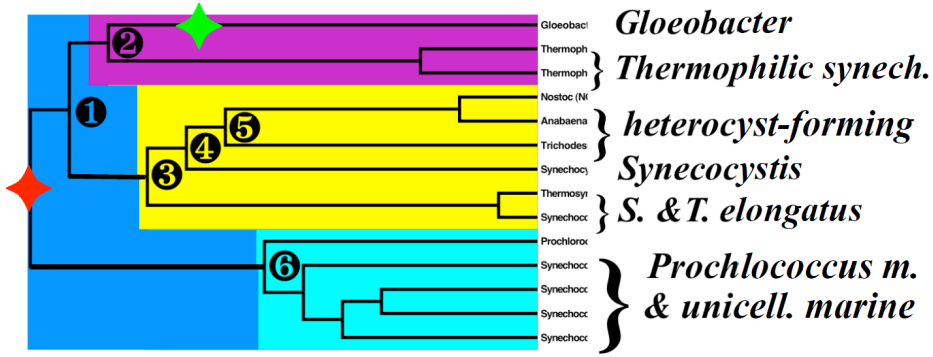
A new root for the Cyanobacterial tree?

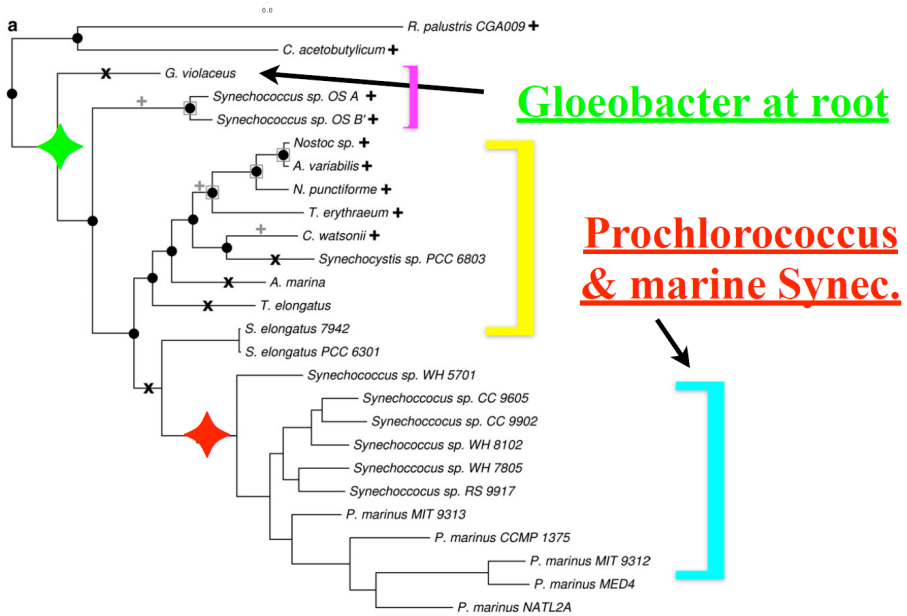
strong resampling support for root

500 random samples of 20% of all G trees



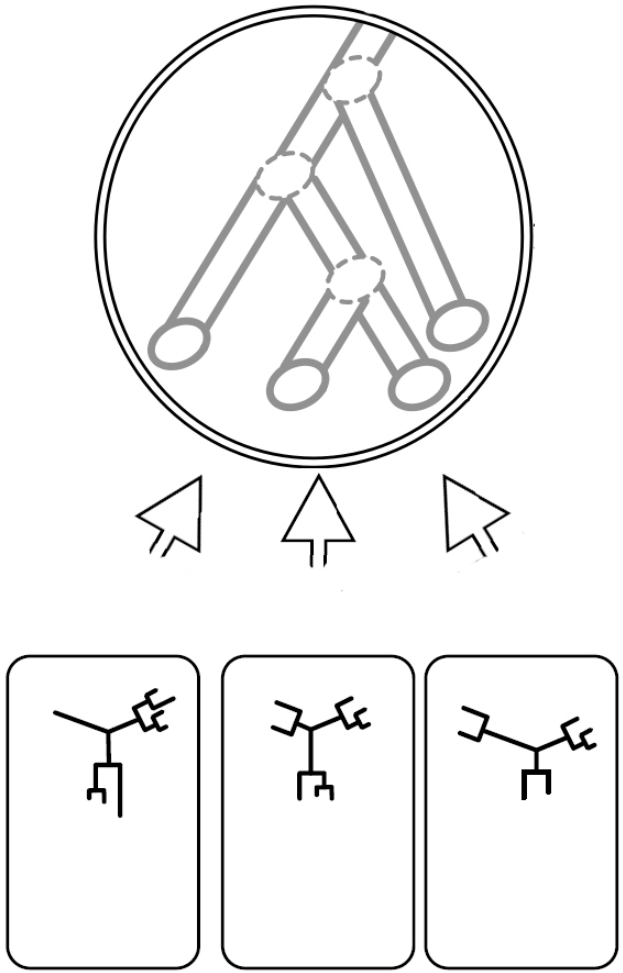
same root found in 96% of samples



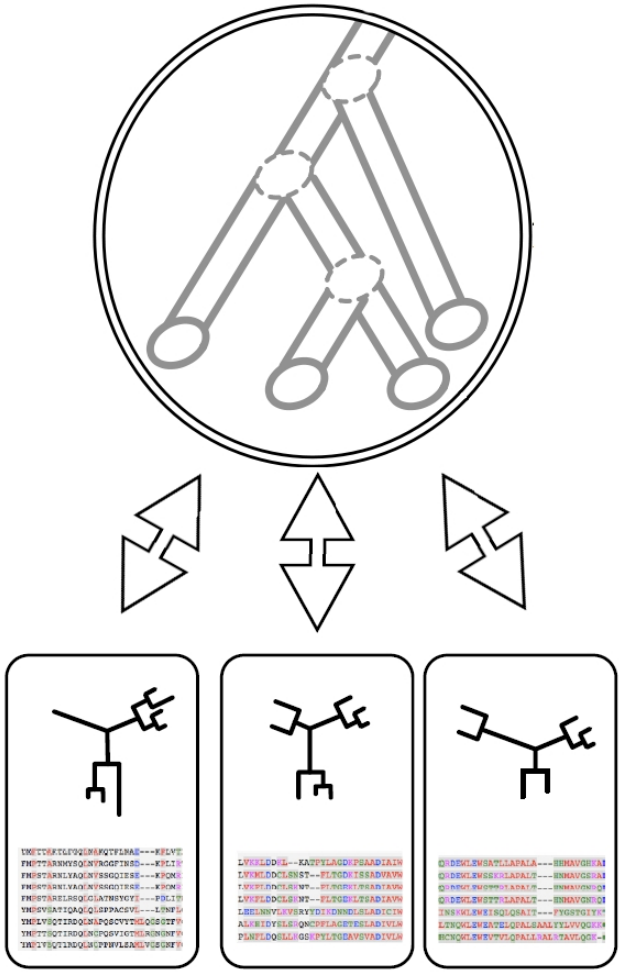


Swingley, Blankenship, Raymond MBE 2008

Simultaneous reconstruction of gene trees and species trees



Simultaneous reconstruction of gene trees and species trees



Take home message

Modeling the processes of genome evolution during phylogenetic reconstruction allows

- Reconstructing better species trees

Take home message

Modeling the processes of genome evolution during phylogenetic reconstruction allows

- Reconstructing better species trees
- Reconstructing better gene histories

Take home message

Modeling the processes of genome evolution during phylogenetic reconstruction allows

- Reconstructing better species trees
- Reconstructing better gene histories
- Documenting the tempo and rates of genome evolution

Take home message

Modeling the processes of genome evolution during phylogenetic reconstruction allows

- Reconstructing better species trees
- Reconstructing better gene histories
- Documenting the tempo and rates of genome evolution
- Documenting particular evolutionary events

Take home message

Modeling the processes of genome evolution during phylogenetic reconstruction allows

- Reconstructing better species trees
- Reconstructing better gene histories
- Documenting the tempo and rates of genome evolution
- Documenting particular evolutionary events
- HGT can be used to date divergence events (a clock without rock)

Acknowledgements

- Bastien Boussau, Gergely Szollosi, Eric Tannier
- Sophie Abby, Simon Penel and Manolo Gouy
- ANR Phylariane (LIRMM, ISEM, LBBE)
- Computing power from LBBE, IN_2P_3 and CINES (Jade)
- Thank *you*!