Multi-gene phylogenies (phylogenomics)

Large evolutionary-scale (deep) phylogenetic analyses

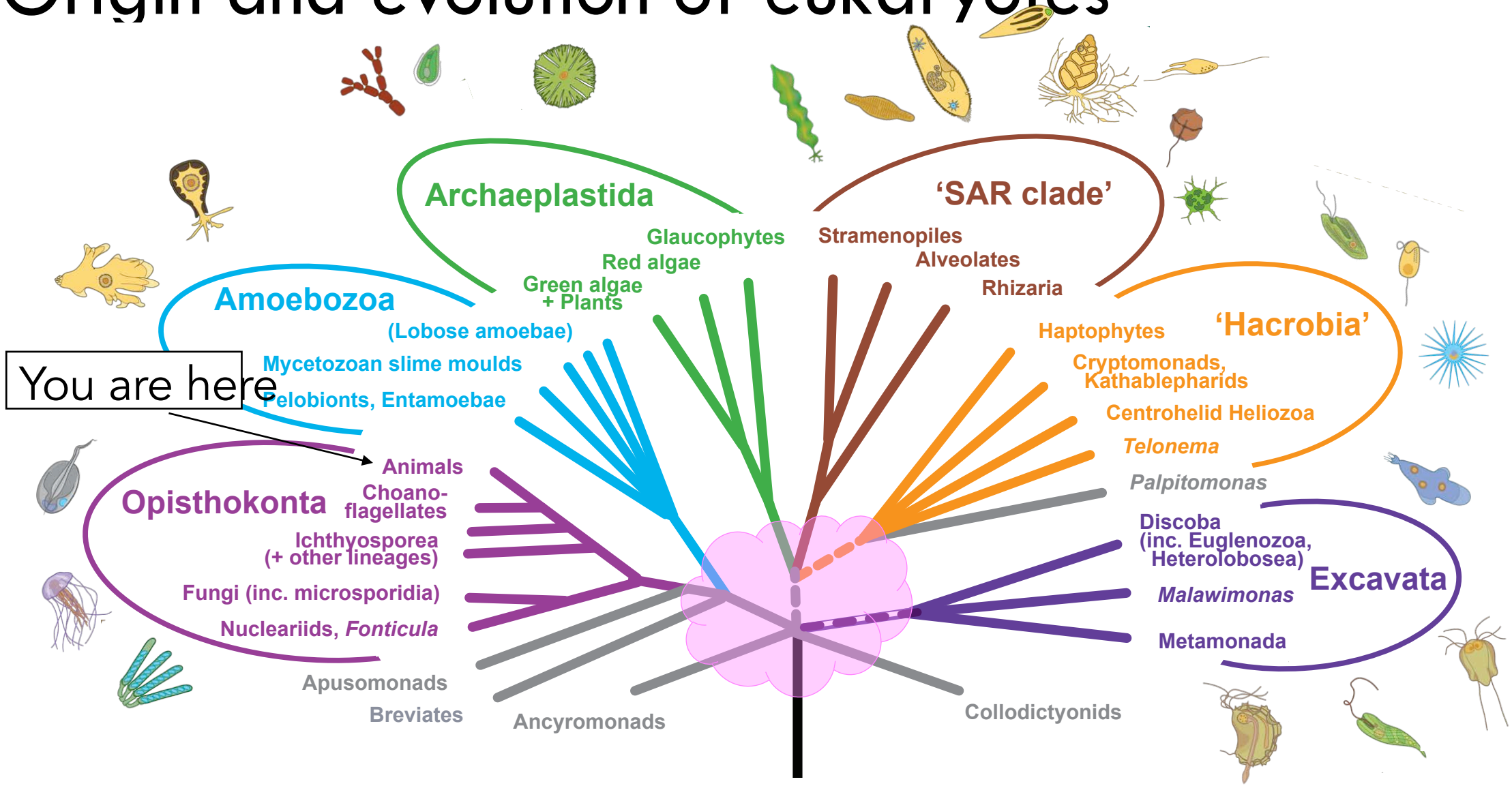
Laura Eme
Assistant Prof.
Paris-Sud University

http://emelaura.com

Disclaimer

What I do and how, and how that can hopefully apply to your questions

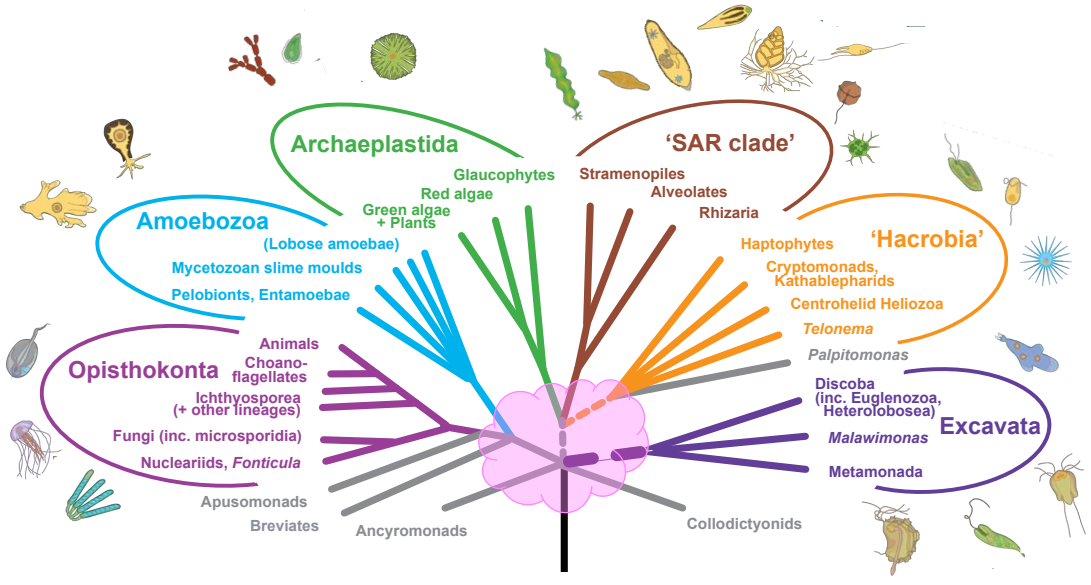
Origin and evolution of eukaryotes



Origin and evolution of eukaryotes

- Tree of eukaryotes (deep relationships between major groups)
- Origin of eukaryotes (i.e., their placement within the tree of life)
- Origin and evolution of specific eukaryotic genes/systems (e.g., endomembrane system)
- Evolution of gene content along the species tree (origin of major clades)
- Impact of LGT on genome evolution

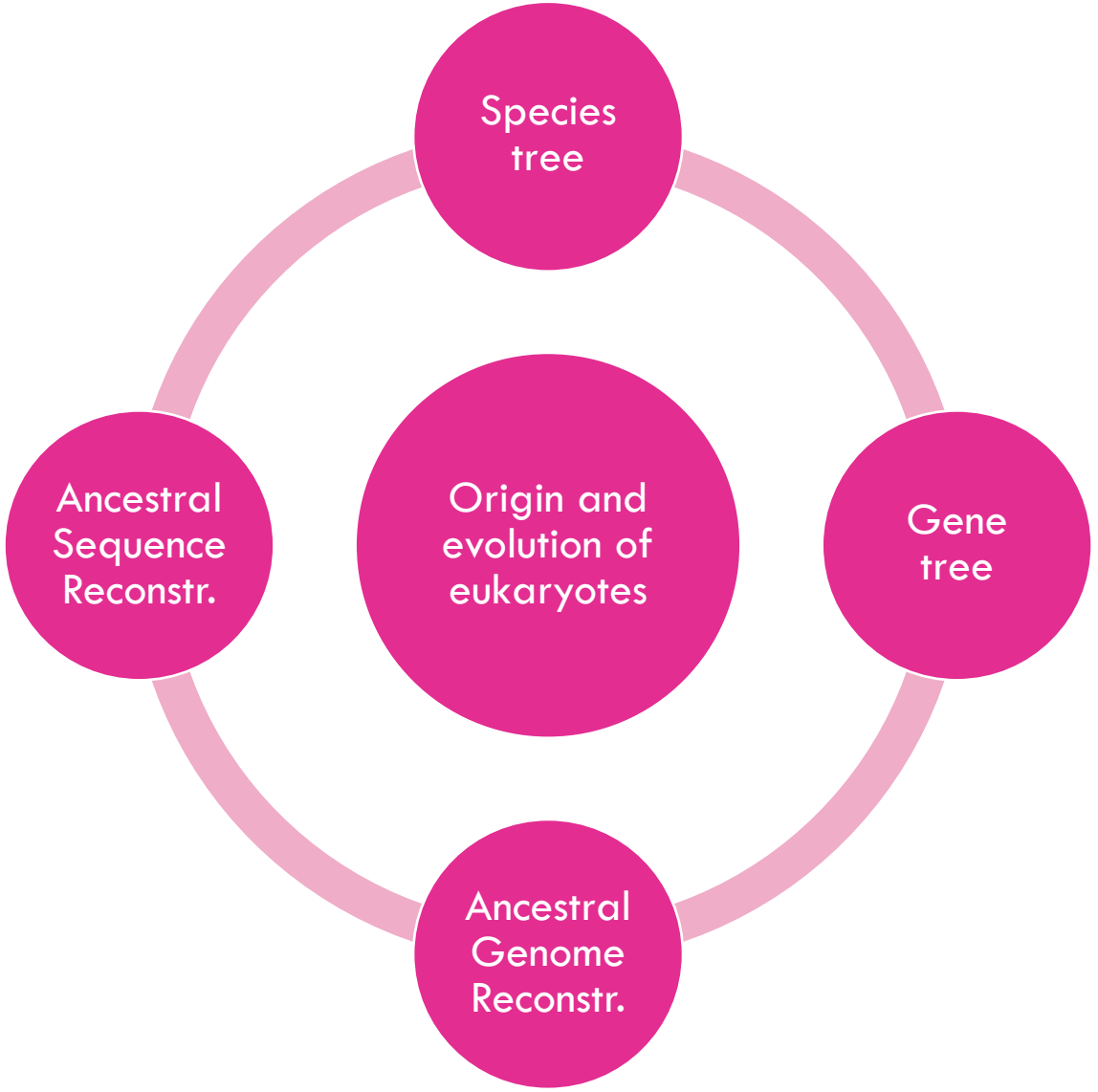
➔ Large evolutionary scale questions



Origin and evolution of eukaryotes

- Tree of eukaryotes (deep relationships between major groups)
- Origin of eukaryotes (i.e., their placement within the tree of life)
- Origin and evolution of specific eukaryotic genes/systems (e.g., endomembrane system)
- Evolution of gene content along the species tree (origin of major clades)
- Impact of LGT on genome evolution

➔ Large evolutionary scale questions



1 - Protein models of evolution

Empirical models, GTR model, Mixture models

1.1 Empirical models

Code degeneracy

Glu-Gly-Ser-Ser-Trp-Leu-Leu-Leu-Gly-Ser

Glu-Gly-Ser-Ser-Tyr-Leu-Leu-Ile-Gly-Ser

Asp-Gly-Ser-Ala-Trp-Leu-Leu-Leu-Gly-Ser

Asp-Gly-Ser-Ala-Tyr-Leu-Leu-Ala-Gly-Ser

GAA-GGA-AGC-TCC-TGG-TTA-CTC-CTG-GGA-TCC

GAG-GGT-TCC-AGC-TAT-CTA-TTA-ATT-GGT-AGC

GAC-GGC-AGT-GCA-TGG-TTG-CTT-TTG-GGC-AGT

GAT-GGG-TCA-GCT-TAC-CTC-CTG-GCC-GGG-TCA

Protein sequence evolves slower than nucleotide

Code degeneracy

- Base composition bias can lead to large difference in codon usage
- Comparing protein sequences can reduce the compositional bias problem

Evolutionary models for amino acid changes

Typically

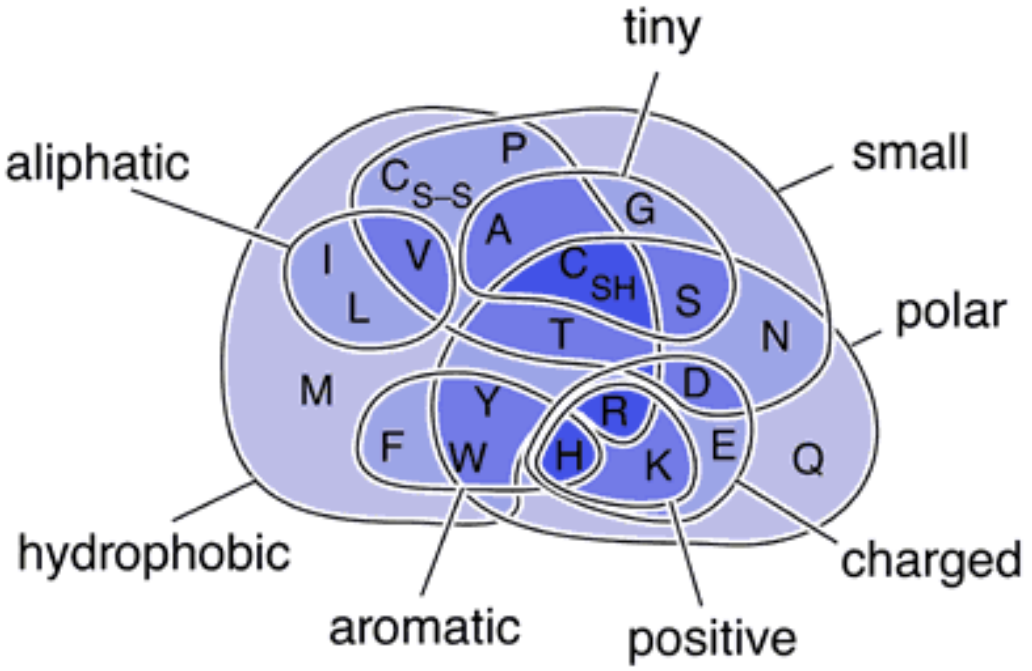
- A 20x20 rate matrix
- Assumes stationarity and reversibility

Amino acid physico-chemical properties

- AA can be categorized according to their physicochemical properties
- Major factor in protein folding (secondary, tertiary, quaternary structure)
- Key to protein functions (e.g., catalytic sites)

➔ Major influence in pattern of amino acid mutations

Some amino acid changes are more commonly fixed than others



Empirical models: amino acid substitution matrices based on observed substitutions

Summarise the substitution patterns from a large number of existing alignments (‘average’ models)

Empirical models: amino acid substitution matrices based on observed substitutions

Summarise the substitution patterns from a large number of existing alignments ('average' models)

Raw data: observed changes in pairwise comparisons

```
seq.1 AIDESLIIASIATATI
      |*||*||*||*||*||
seq.2 AGDEALILASAATSTI
```

```
seq.1 AIDESLIIASIATATI
      |*||*||*||*||*||
seq.2 AGEEALILASAATSTI
```

Raw matrix

Symmetrical

	A	S	T	G	I	L	E	D
A	3							
S	2	1						
T	0	0	1					
G	0	0	0	0				
I	1	0	0	1	2			
L	0	0	0	0	1	1		
E	0	0	0	0	0	0	1	
D	0	0	0	0	0	0	1	0

```
seq.1 AIDESLIIASIATATI
      |*||*||*||*||*||
seq.2 AGEEALILASAATSTI
```

Raw matrix

Symmetrical

	A	S	T	G	I	L	E	D
A	3							
S	2	1						
T	0	0	1					
G	0	0	0	0				
I	1	0	0	1	2			
L	0	0	0	0	1	1		
E	0	0	0	0	0	0	1	
D	0	0	0	0	0	0	1	0

```
seq.1 AIDESLIIASIATATI
      |*||*||*||*||*||
seq.2 AGEEALILASAATSTI
```

Raw matrix

Symmetrical

	A	S	T	G	I	L	E	D
A	3							
S	2	1						
T	0	0	1					
G	0	0	0	0				
I	1	0	0	1	2			
L	0	0	0	0	1	1		
E	0	0	0	0	0	0	1	
D	0	0	0	0	0	0	1	0

➔ The larger the dataset, the better the estimates

Amino acid exchange matrices

$$\begin{pmatrix} - & s1,2 & s1,3 & \dots & s1,20 \\ s1,2 & - & s2,3 & \dots & s2,20 \\ s1,3 & s2,3 & - & \dots & s3,20 \\ \dots & \dots & \dots & \dots & \dots \\ s1,20 & s2,20 & s3,20 & \dots & - \end{pmatrix}$$

X diag($\pi 1, \ldots, \pi 20$) = Q matrix

Q	Rate matrix
sij	Exchangeabilities of amino acid pairs ij
sij = sji	Time reversibility (usually)
πi	Stationarity of amino acid frequencies (typically the observed proportion of residues in the dataset)

Empirical models

- Summarise the substitution patterns from a large number of existing alignments ('average' models)
- Different substitution matrices come from:
  - Selection of specific proteins
    - Globular proteins, membrane proteins?
    - Mitochondrial proteins?
  - Range of sequence similarities used
  - Counting methods
    - On a tree
    - Pairwise comparison from an alignment

Empirical models

Dayhoff (Dayhoff et al., 1978): Nuclear encoded genes (~100 proteins) ➔ PAM matrices

JTT (Jones et al., 1992): 59,190 point mutations from 16,300 proteins from membrane spanning segments

Closely related sequence pairs (>85% identity):

1) count the number of amino acid changes of each type per pair
2) rescale these by the sequence divergence for the analyzed pair
3) Average over all sequence pairs

Limitation: for less similar sequences, no linearity between observed and real substitution rate (hidden substitutions)

Empirical models

Dayhoff (Dayhoff et al., 1978): Nuclear encoded genes, ~100 proteins ➔ PAM matrices

JTT (Jones et al., 1992): 59,190 point mutations from 16,300 proteins from membrane spanning segments

WAG (Whelan and Goldman, 2001): General matrix

LG (Le and Gascuel, 2008): General matrix

The WAG matrix

- Globular protein sequences
  - 3,905 sequences from 182 protein families

- Produced a phylogenetic trees for every family and used maximum likelihood to estimate the relative rate values in the rate matrix (i.e., maximizes the overall lnL over 182 different trees)

- Better fit of the model with most data (significant improvement of the tree lnL when compared to PAM or JTT matrices)

- Can be used for (more) distant homologues

Further improvements: the LG matrix

- Used the same phylogenetic approach as WAG
- Further refine the method by adding the variability of evolutionary rates across sites when estimating the matrix and increase the number of sequences used
- Better fit of the model with most data (significant improvement of the tree lnL when compared to WAG and other matrices)

Empirical models

Dayhoff (Dayhoff et al., 1978): Nuclear encoded genes, ~100 proteins ➔ PAM matrices

JTT (Jones et al., 1992): 59,190 point mutations from 16,300 proteins from membrane spanning segments

WAG (Whelan and Goldman, 2001): General matrix

LG (Le and Gascuel, 2008): General matrix

Mtrev24 (Adachi and Hasegawa, 1996) : Mitochondrial (vertebrates)

Mtmam (Yang et al., 1998): Mitochondrial (mammals)

mtART (Abascal et al., 2007): Mitochondrial (Arthropoda)

CpRev (Adachi et al., 2000): Chloroplast

VT (Müller and Vingron, 2000): General matrix

RtRev (Dimmic et al., 2002): Retrovirus

DayhoffDCMUT (Kosiol and Goldman, 2005): Revised Dayhoff matrix

(and more…)

Summary

- Many amino acid rate matrices exist
- One should make a rational choice (as much as possible):
  - How was the rate matrix produced?
  - What are the structural features of the sequences that you are analyzing? Globular/membrane protein? Overall level of sequence identity of the compared sequences? Specific compositional bias (mitochondrial proteins matrix: mtREV24; Transmembrane domains: PHAT)?
  - ModelTest, ModelFinder (IQtree), ProtTest… to compare models

Correcting for equilibrium frequencies

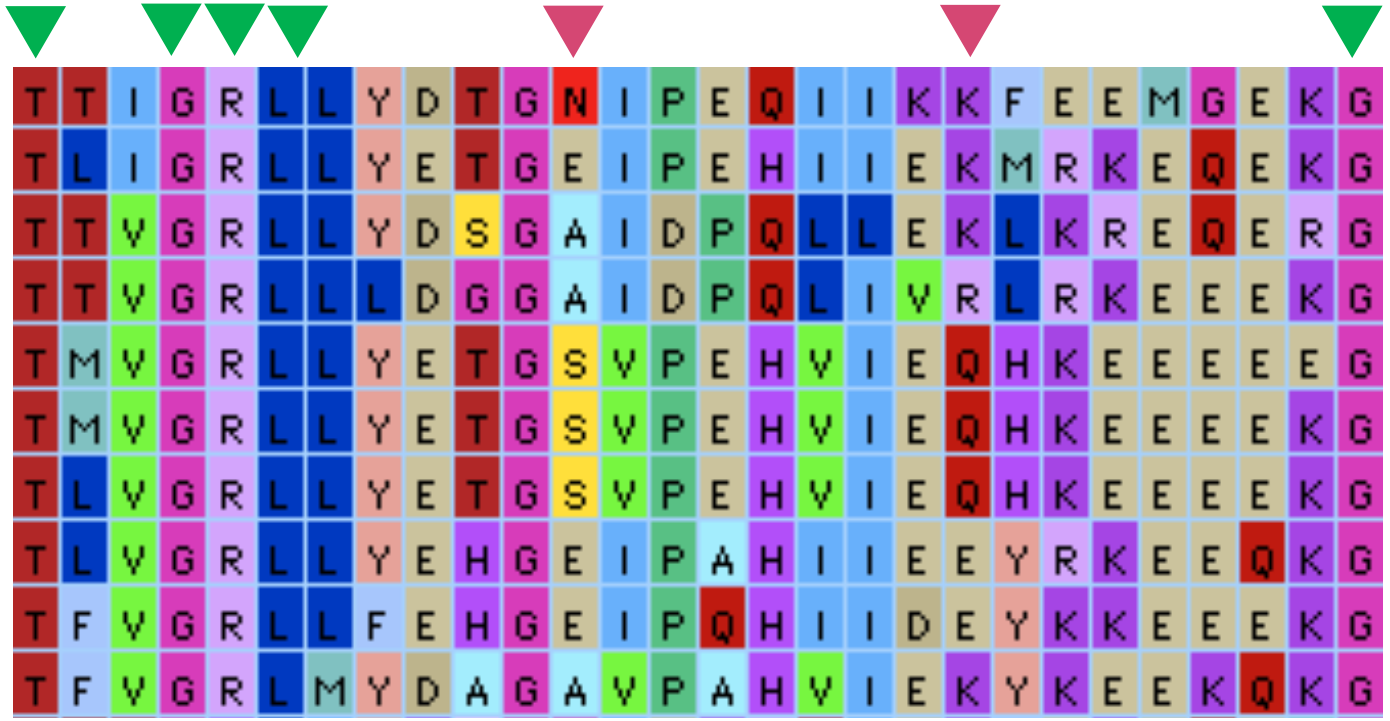
- Empirical matrices are obtained by averaging the observed changes and amino acid frequencies between numerous proteins and are used for your specific dataset

- With recent software, you can correct the πi values based on the observed frequencies in your data (“+F” option). E.g. LG+G+F

Rate heterogeneity parameter

- Not all sites “evolve” at the same speed depending on how it impacts function

Ef1-a



Rate heterogeneity parameter

- Discreticized Gamma distribution (+G)
  - Default is usually 4 categories but can be set to be more (but more computationally intensive)

Rate heterogeneity parameter

- Discreticized Gamma distribution (+G)
  - Default is usually 4 categories but can be set to be more (but more computationally intensive)

- FreeRate model (+R)
  - Does not follow a parametric distribution
  - Not all categories will have the same number of sites
  - More realistic but more computationally intensive
  - Typically fits data better than the +G model and is recommended for analysis of large data sets

1.2 Fully parameterized time-reversible model

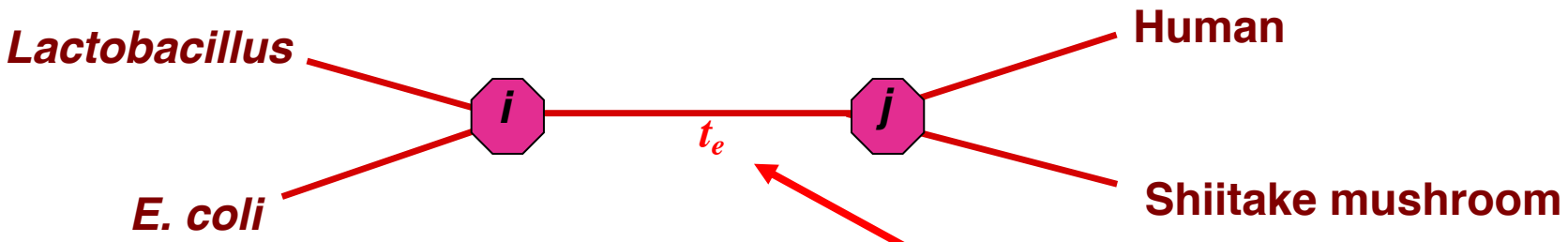
GTR (General time reversible)

- One can generate a dataset-specific model
- All parameters of the Q matrix are estimated from your data (exchangeabilities and equilibrium frequencies)
- GTR20: General time reversible model for amino-acids: 189 rate parameters!

WARNING Parameter-rich: parameter estimates might not be reliable if made on short alignments (not enough phylogenetic information)

1.3 Mixture models

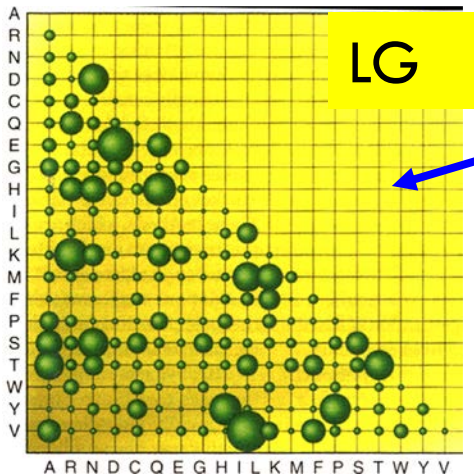
Your model is giving you the probability of going from amino acid *i* to *j* at site *x*, evolving at rate r_v on branch t_e



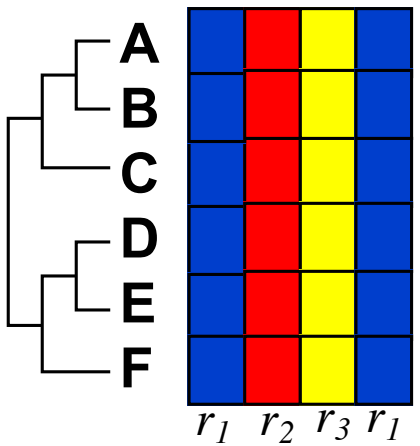
$$P(j \mid i;t) = \left[\exp(Q \times t_e \times r_v)\right]_{ij}$$

For $i \neq j$:

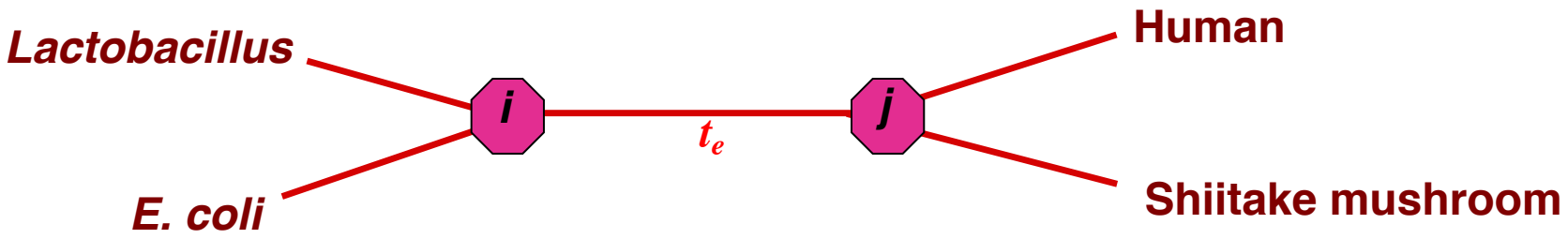
$$q_{ij} = r_{ij} \times \pi_j$$



$$\Pi = \begin{bmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_R & 0 & 0 \\ 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \pi_v \end{bmatrix}$$



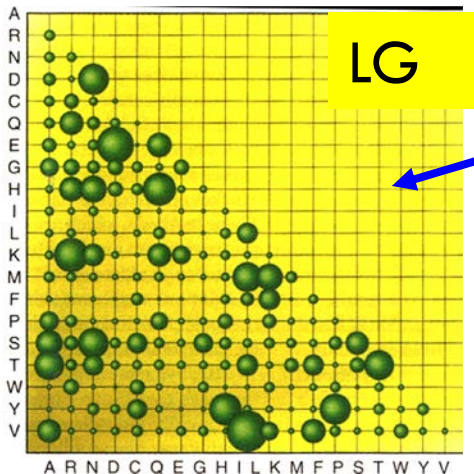
Your model is giving you the probability of going from amino acid *i* to *j* at site *x*, evolving at rate r_v on branch t_e



$$P(j \mid i;t) = \left[\exp(Q \times t_e \times r_v)\right]_{ij}$$

For $i \neq j$:

$$q_{ij} = r_{ij} \times \pi_j$$



$$\Pi = \begin{bmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_R & 0 & 0 \\ 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \pi_v \end{bmatrix}$$

Assumptions

-different sites in protein and organisms all evolve according to the same general 'rules'

- i.e. rate matrices (*R's*) and frequencies (*Π's*) are the 'same' for all sites and branches

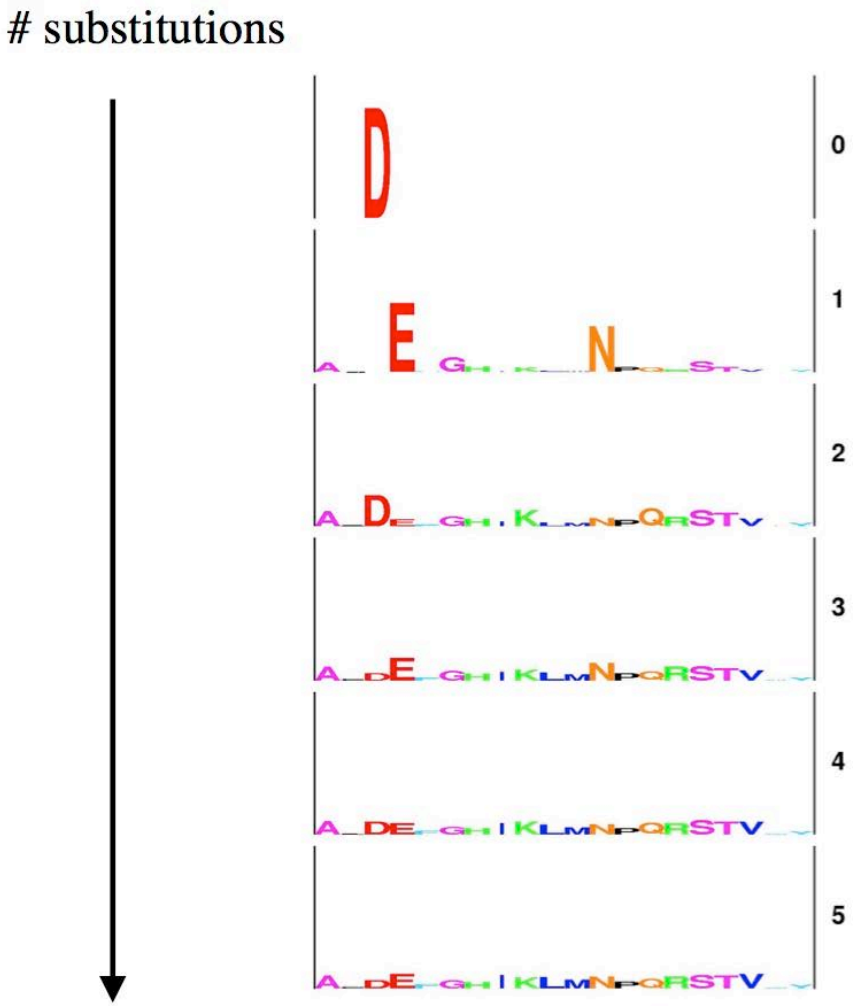
The problem…

- Such models are a dramatic over-simplification of what is really going on
  - Average over sites, average over different organisms, average across protein families

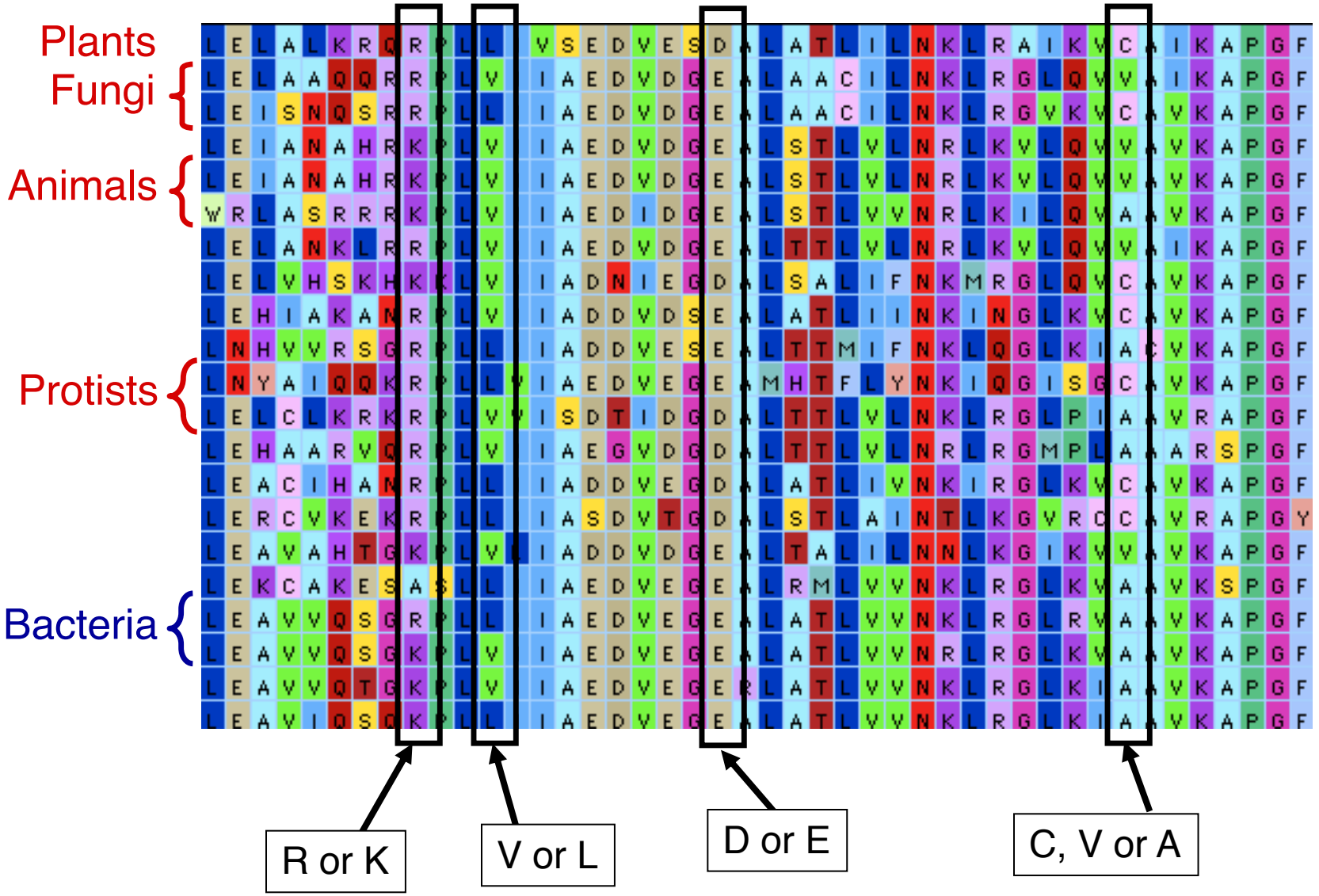
- Sites in proteins can change function over time
  - sites under negative selection ⟷ neutral ⟷ positive selection

- Every amino acid site in a protein has a unique structural/functional context
  - Hydrophobicity, polarity, charge, size, functional group, etc.
  - Different sites have different exchangeabilities
  - Different frequencies of AAs occur at different sites

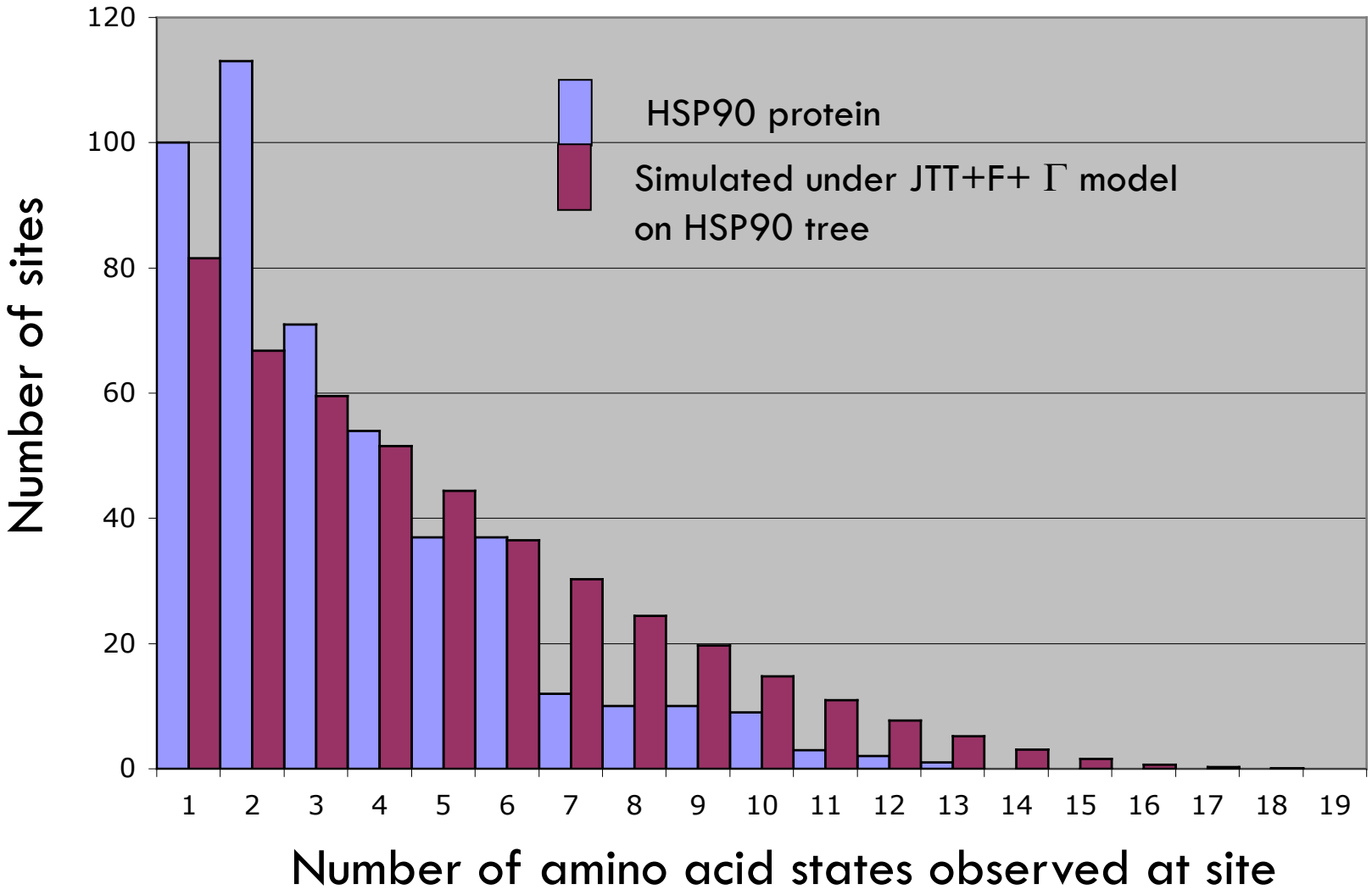
Starting at a D with a site homogeneous matrix (LG+F)



Evolution of chaperonin 60 over ~1.5 billion years

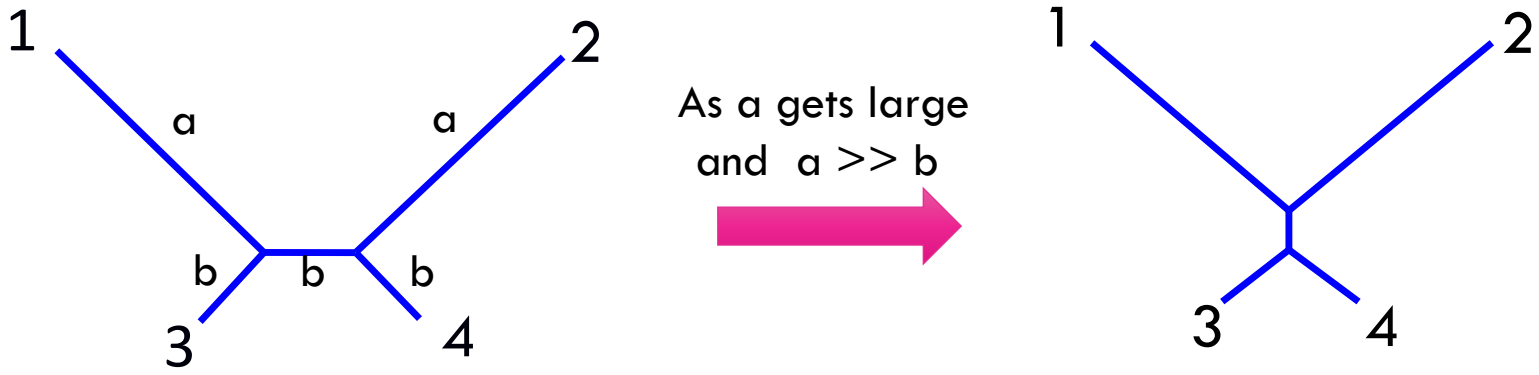


Distribution of the number of different amino acids at aligned sites



Wang *et al.* (2008) *BMC Evolutionary Biology* 8: 331

So what happens to phylogenetic estimation when you ignore site-heterogeneity?



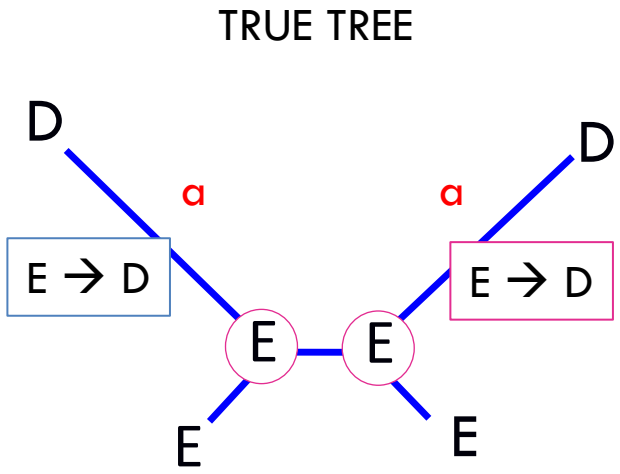
Long branch attraction

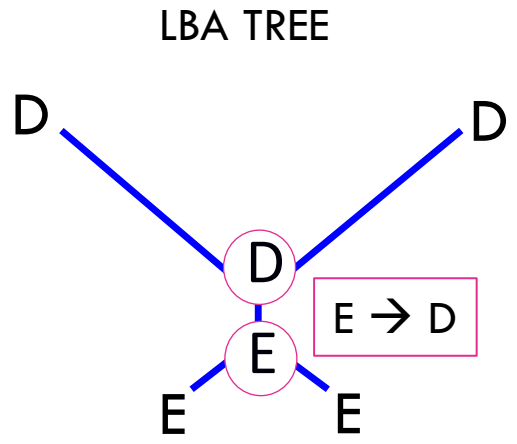
Susko et al. (2004) *Mol. Biol. Evol.* 21:1629

Lartillot and Philippe (2007) BMC Evol. Biol Suppl 1, S4.

Wang et al. (2008) *BMC Evol. Biol.* 8:331

Why long branch attraction (LBA)?





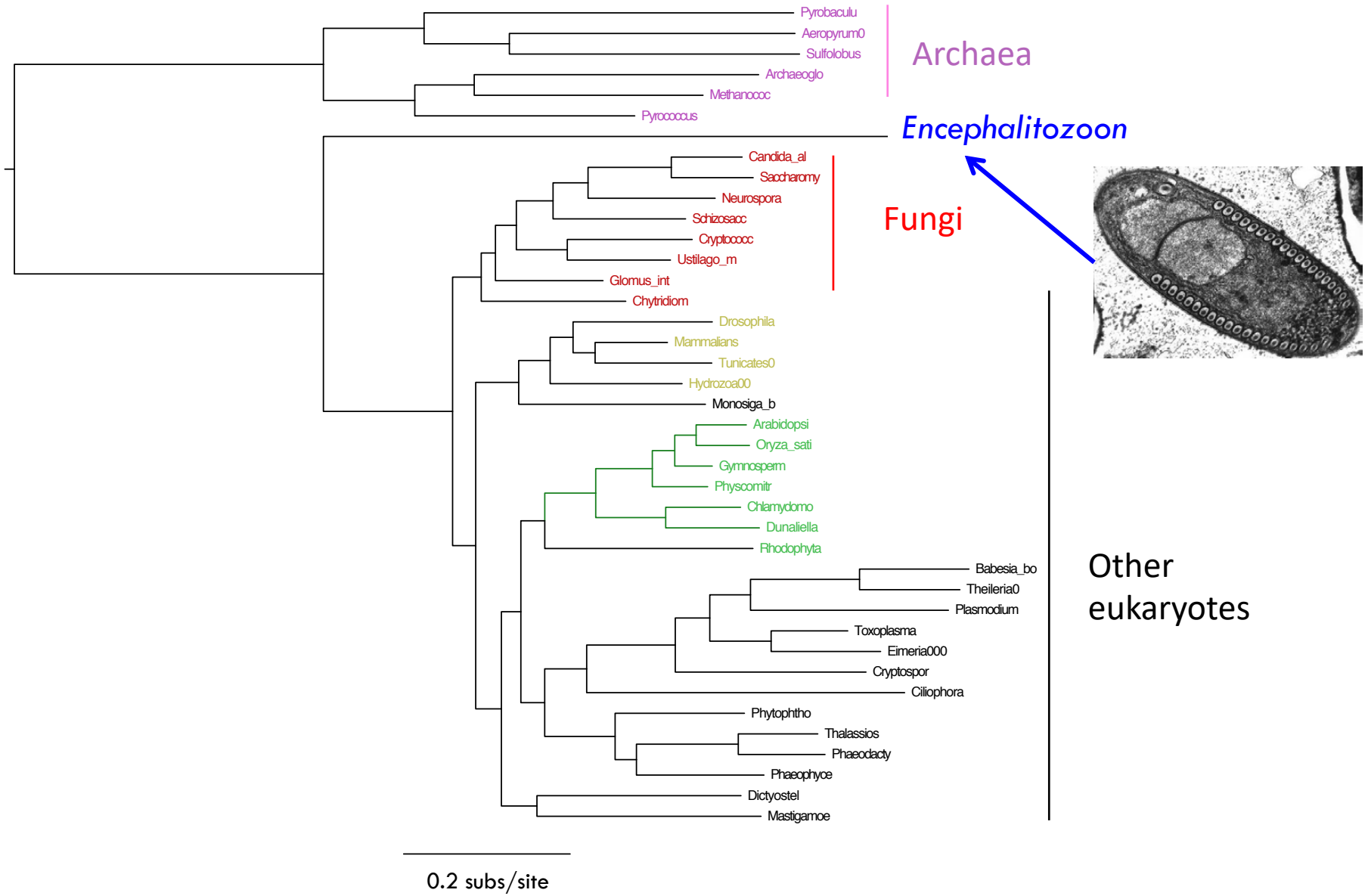
Under site homogeneous model (LG+F+G), the probability of converging on the same state: i.e. E → D twice is pretty low:

- if branch-length a is really long, then $P(\text{convergence})_{LG} \approx \pi_D^2 = (0.057)^2 = 0.0032$

Under a site-specific model where you can only be D or E (with equal frequency of 0.5):

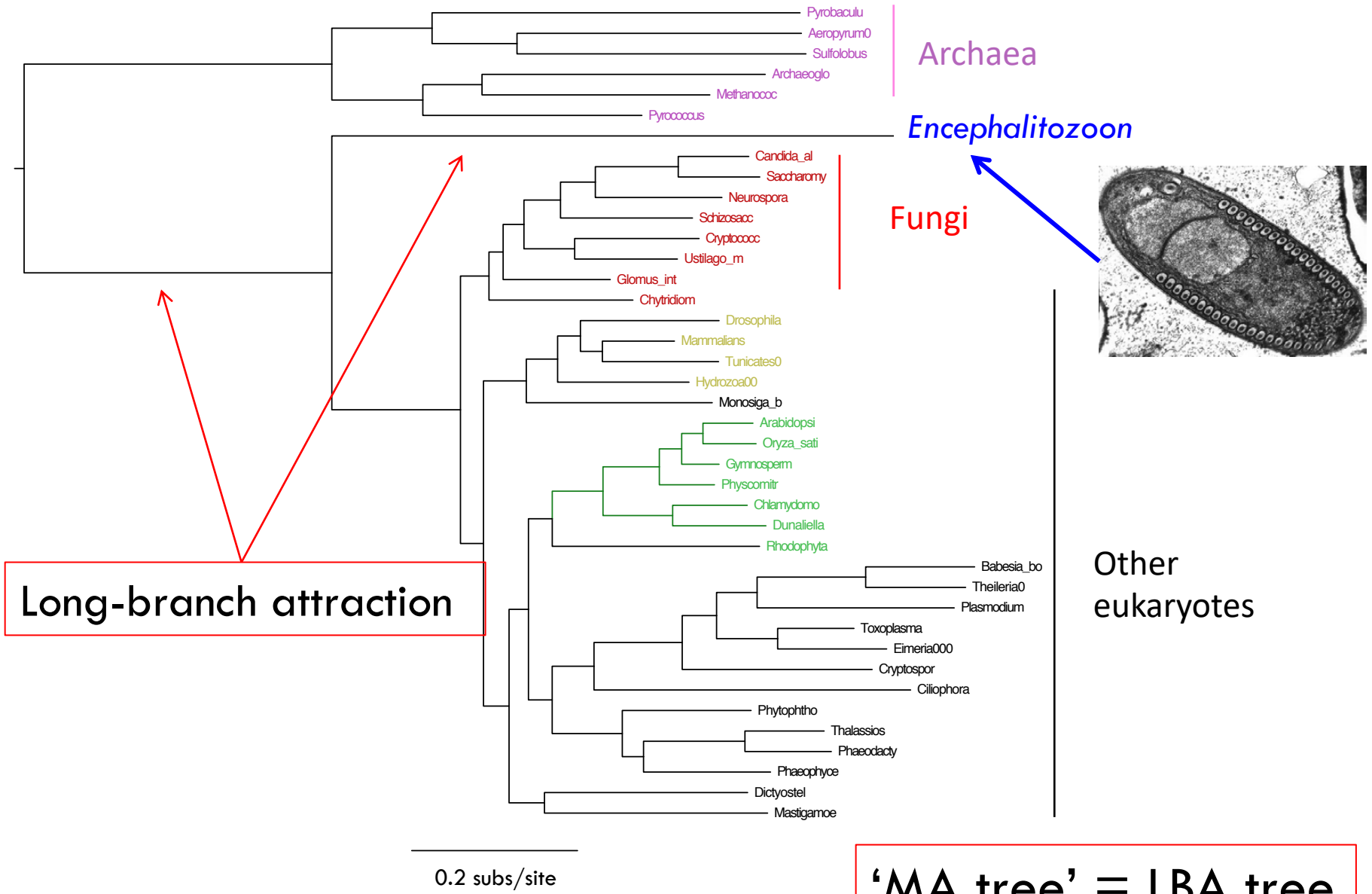
- $P(\text{convergence})_{ss} \approx \pi_D^2 = (0.5)^2 = 0.25$

ML tree based on site-homogeneous LG+F+Γ model



40 sequences, 133 genes, 24291 aa

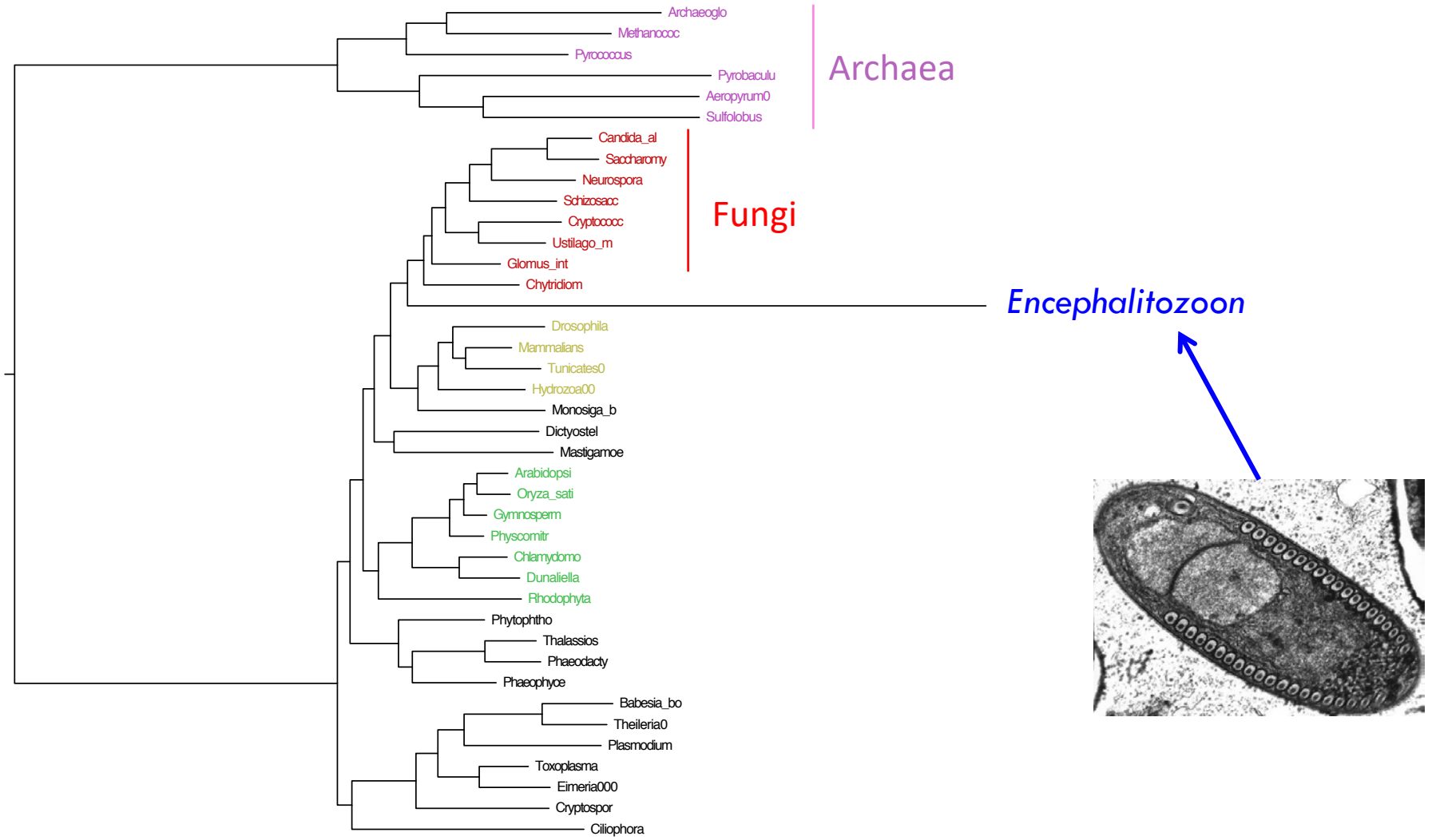
ML tree based on site-homogeneous LG+F+Γ model



40 sequences, 133 genes, 24291 aa

'MA tree' = LBA tree

ML tree based on site-heterogeneous LG+C20+F+Γ model



'MF tree' = true tree

Mixture models

- Standard protein substitution models: single Q matrix
- Mixture models: combine several amino-acid replacement matrices
- Same principle as rate heterogeneity gamma distribution
  - For each site, its likelihood is the **sum of its weighted likelihood under each Q matrix** that is are part of the mixture model

Mixture models: terminology warning

- Different kinds of mixture models!
- Rate-category gamma distribution is a mixture model
- **Usually people refer to mixture of amino-acid replacement matrices**
- Mixtures can be apply to any part of the model (e.g., branch lengths)

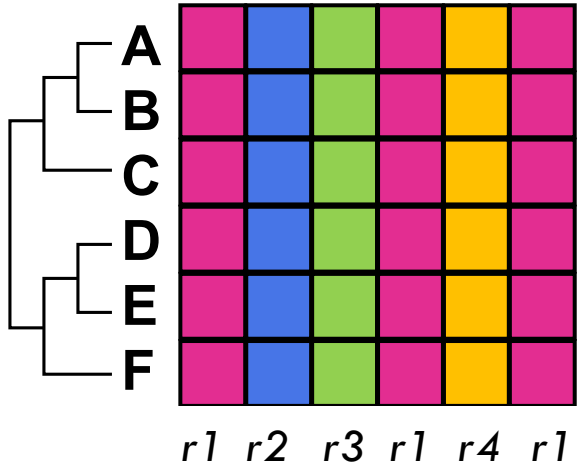
LG4M and LG4X mixture models

"the variability of evolutionary rates corresponds to one of the most apparent heterogeneity factors among sites, and **there is no reason to assume that the substitution patterns remain identical regardless of the evolutionary rate**" Le, Dang, Gascuel 2012

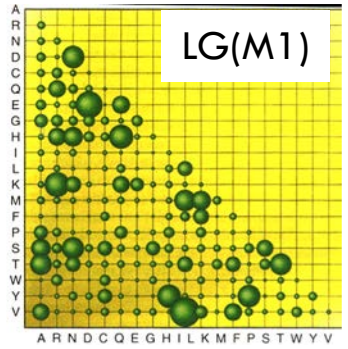
Standard LG+gamma model: only the global rate differs from one category to another

LG4M and LG4X mixture models

LG4M: each gamma rate category gets its own Q matrix (i.e., each of the 4 gamma-distributed rate category gets it own amino acid equilibrium distributions and exchangeabilities)

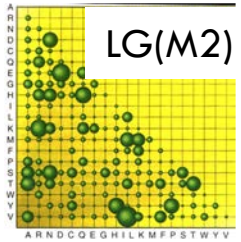


25%



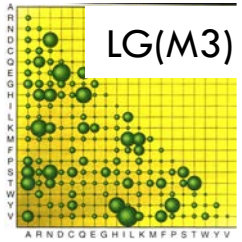
$$\Pi = \begin{bmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_R & 0 & 0 \\ 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \pi_v \end{bmatrix}$$

25%



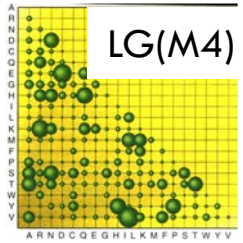
$$\Pi = \begin{bmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_R & 0 & 0 \\ 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \pi_v \end{bmatrix}$$

25%



$$\Pi = \begin{bmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_R & 0 & 0 \\ 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \pi_v \end{bmatrix}$$

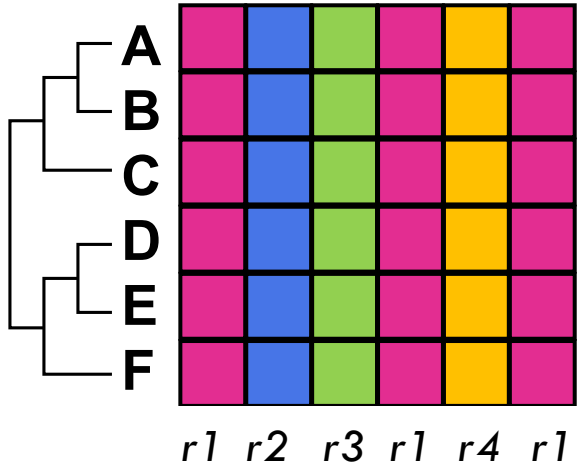
25%



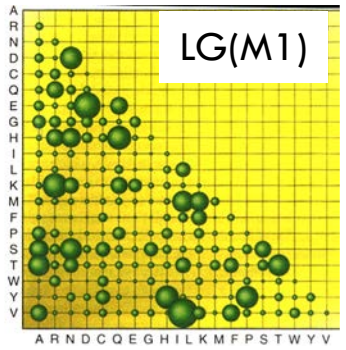
$$\Pi = \begin{bmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_R & 0 & 0 \\ 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \pi_v \end{bmatrix}$$

LG4M and LG4X

LG4X: each rate category gets its own Q matrix BUT rates and weights are left out of the gamma distribution assumption

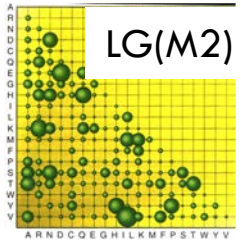


50%



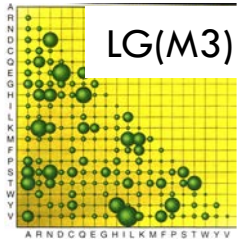
$$\Pi = \begin{bmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_R & 0 & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \pi_v \end{bmatrix}$$

15%



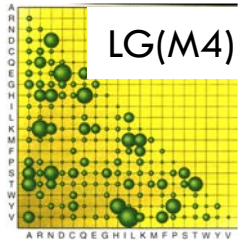
$$\Pi = \begin{bmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_R & 0 & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \pi_v \end{bmatrix}$$

30%



$$\Pi = \begin{bmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_R & 0 & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \pi_v \end{bmatrix}$$

5%



$$\Pi = \begin{bmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_R & 0 & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \pi_v \end{bmatrix}$$

Mixture models based on site exposure (buried or exposed) or secondary structure

Site evolution is highly heterogeneous and depends on many factors: genetic code; solvent exposure; secondary and tertiary structure; protein function; etc.

Mixture models based on site exposure (buried or exposed) or secondary structure

Site evolution is highly heterogeneous and depends on many factors: genetic code; solvent exposure; secondary and tertiary structure; protein function; etc.

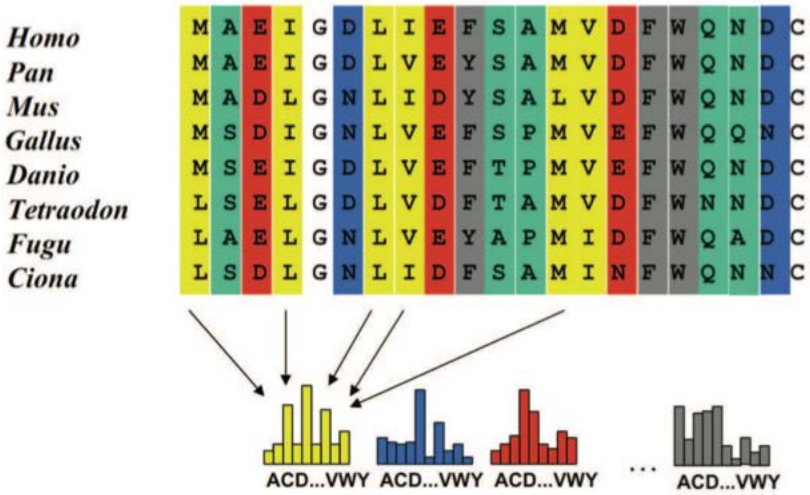
EX2 : Two-matrix model for exposed/buried AA sites

based on their relative accessibility to solvent

EX3 : Three-matrix model for highly exposed/intermediate/buried AA sites

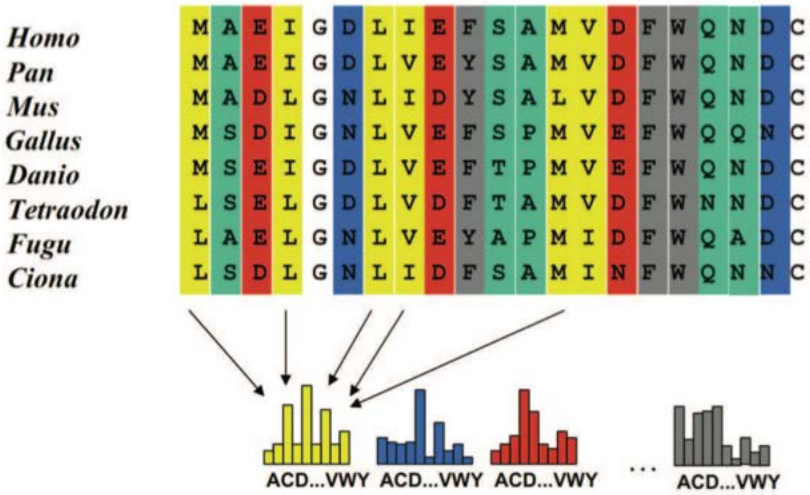
EHO : Three-matrix model for extended/helix/other sites

The CAT model



- Bayesian framework only
- Free number of profiles in the mixture model (estimated during the Bayesian procedure). “Infinite mixture model”
- Each profile corresponds in practice to a *biochemical profile*: only a small number of AA are highly probably, while the frequency of all others will be ~0.

The CAT model



- CAT-Poisson: very simple amino-acid replacement process (R matrix). Each time a substitution event occurs, a new amino-acid is chosen at random, according to the probabilities defined by the profile (*Poisson* or *proportional* amino-acid replacement process).
  Eg., any AA has the same probability to mutate to a Valine.

- CAT-GTR: GTR exchangeability matrix with 189 parameters!

C10, C20, …, C60 mixture models

- 10, 20, 30, 40, 50, 60-profile mixture models are approximations of the CAT model for ML

- 10 (20, 30…) different pre-computed (empirical) Q matrices that correspond to 10 (20, 30…) most-common types of biochemical profiles in proteins

- By default, assume Poisson AA replacement but can be combined with empirically estimated exchangeabilities, such as from the LG matrix. For example: LG+C10

$$q_{ij} = r_{ij} \times \pi_j$$

Problem with mixture models

- As the number of sites and proteins increases the computational cost becomes prohibitive
  - For an ML analysis of 104 taxa and ~90,000 sites (350 proteins concatenated) LG+C60+F+G model takes >350 GB of RAM and ~3 weeks on 12 cores to estimate **the ML tree** using IQTREE v. 1.5
  - **5.5 years to do true bootstrap analysis**

- Phylobayes-MPI using CAT+GTR takes weeks to get only 'thousands' of MCMC generations in the same time
  - Multiple chains almost never converge on the same posterior distribution of trees

➔ PMSF (Posterior Mean Site Frequency) approximation

PMSF (Posterior Mean Site Frequency) model

Implemented in IQtree

1) Reconstruct an ML tree under a 'reasonably good' model = guide tree

2) Using the guide tree, estimate, **for each site x**, the posterior probability of each amino-acid class c (e.g.: C1, C2, …, C60)

Posterior probability of 'class c' at site x

$$P(c|x) = \frac{w_c \times P(x|c)}{\sum_c w_c \times P(x|c)}$$

3) For each site x, estimate the **posterior mean frequency of each amino acid j**

Posterior mean frequency of amino acid j at site x over all c classes ($f_{j,x}$)

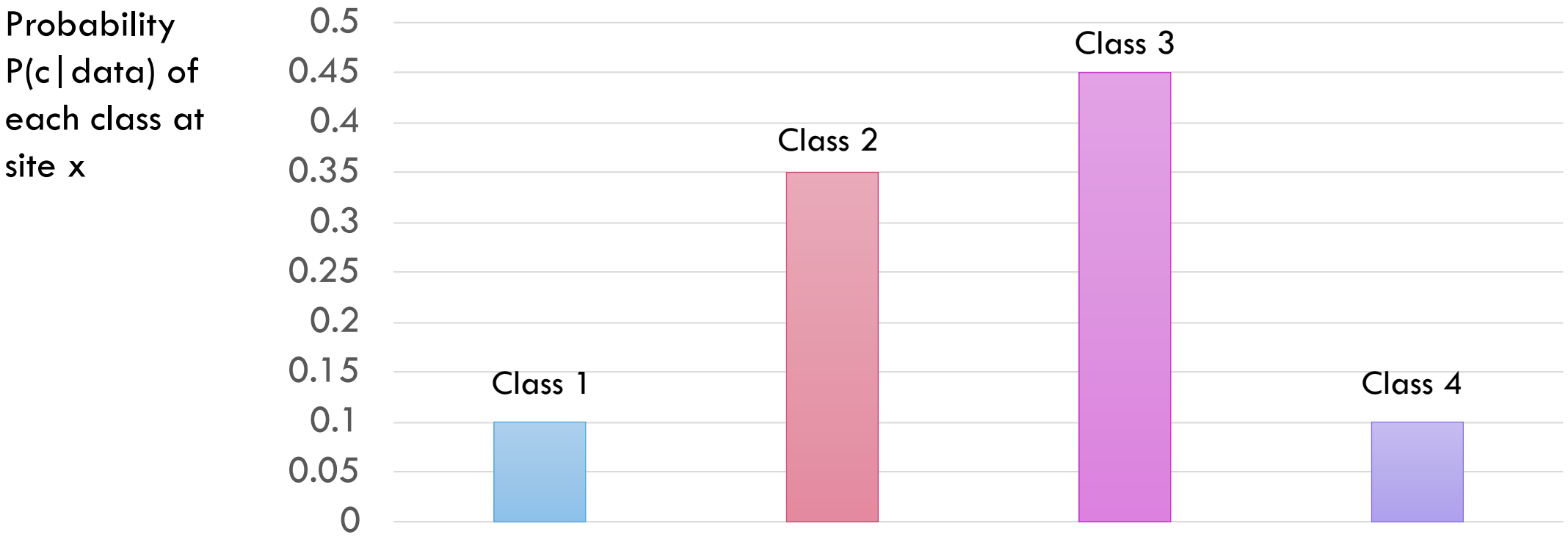
$$f_{j,x} = \sum_c f_{j,c} \times P(c|x)$$

Sum over all classes

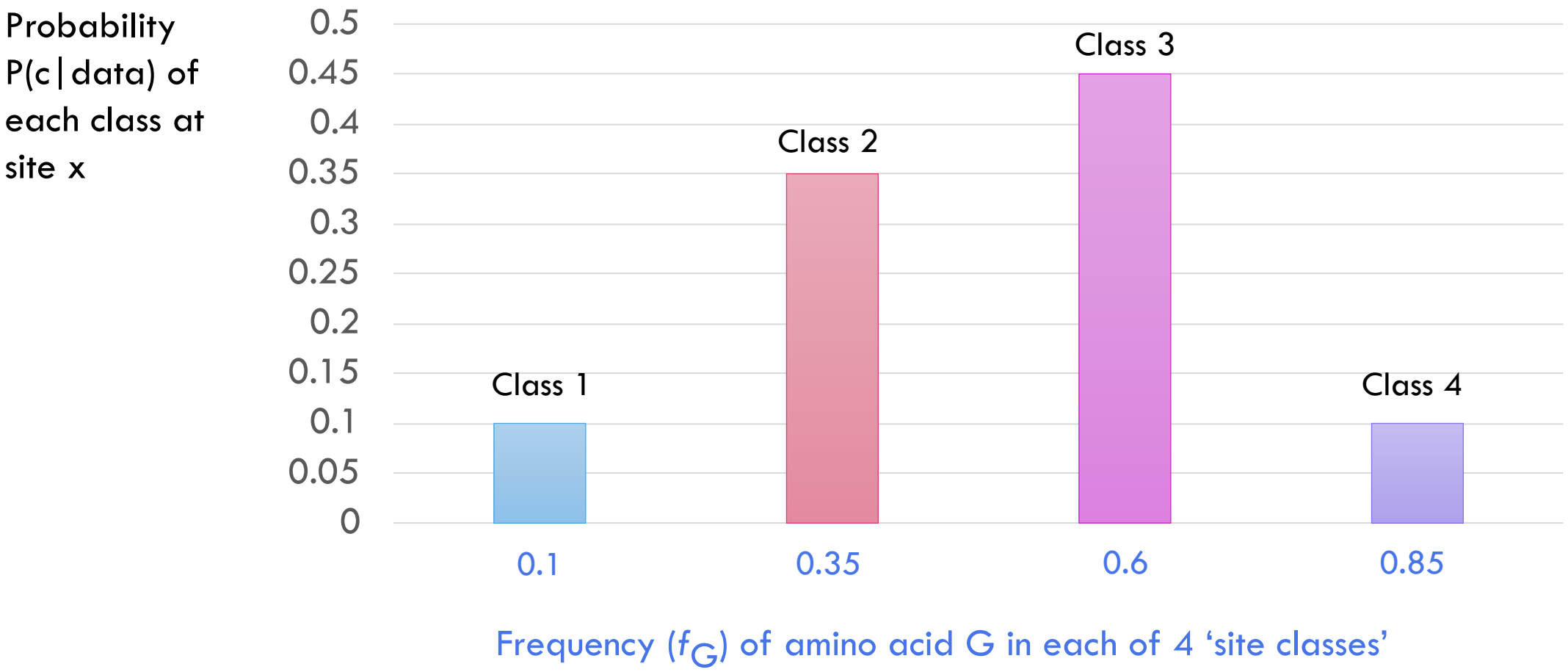
Freq of AA j for class c

Prob of class c at site x

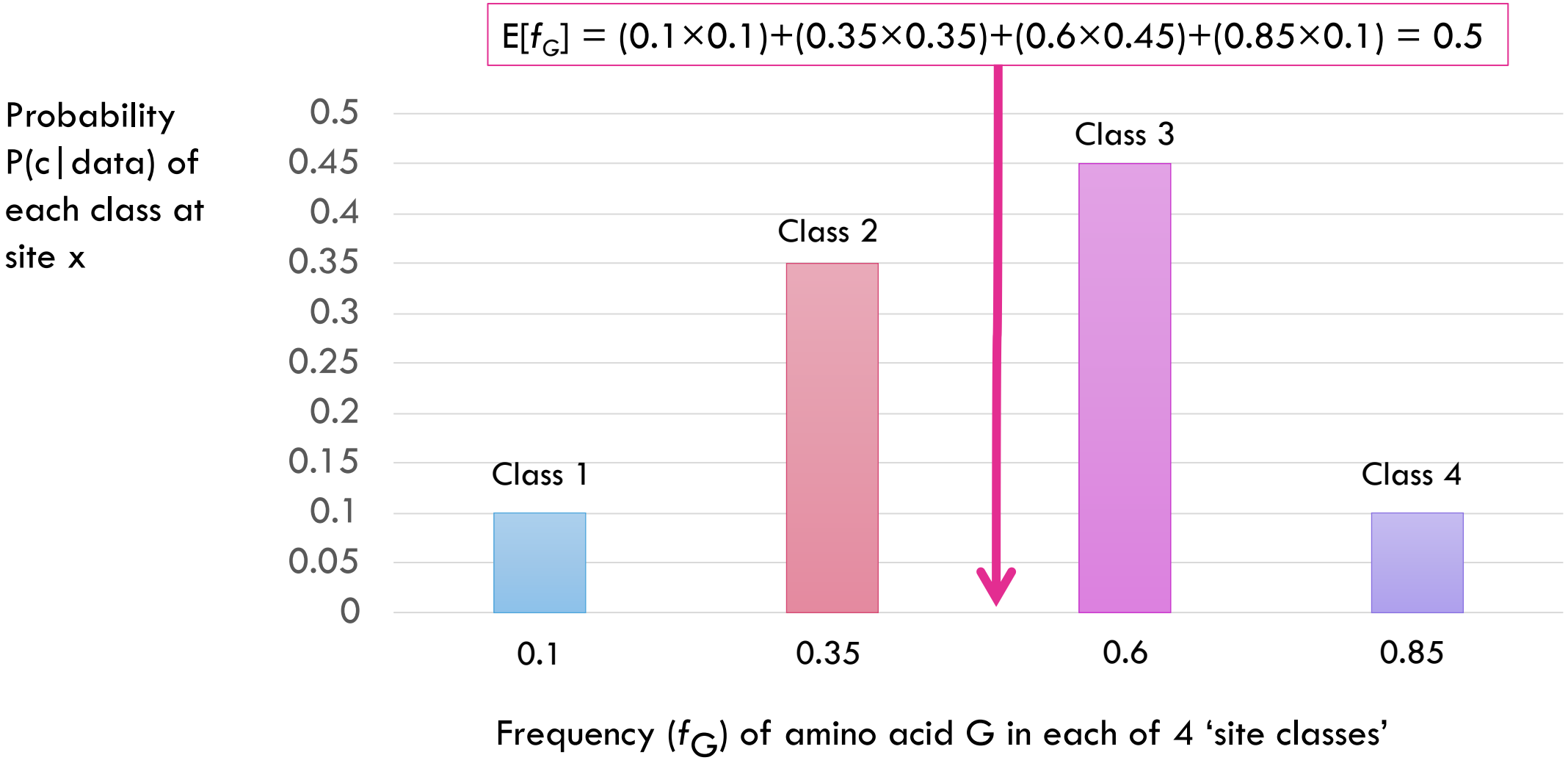
E.g.: Posterior mean site frequency for ‘G’ at a given site x, with a 4 class mixture model



E.g.: Posterior mean site frequency for ‘G’ at a given site x, with a 4 class mixture model



E.g.: Posterior mean site frequency for 'G' at a given site x, with a 4 class mixture model



PMSF (Posterior Mean Site Frequency) model

1) Reconstruct an ML tree under a 'reasonably good' model

2) Using the ML tree, estimate, for each site x, the posterior probability of each amino-acid class c of your preferred mixture model (e.g.: C60)

Posterior probability of 'class c' at site x

$$P(c|x) = \frac{w_c \times P(x|c)}{\sum_c w_c \times P(x|c)}$$

3) For each site x, estimate the posterior mean frequency of each amino acid j

Posterior mean frequency of amino acid j at site x over all c classes (f_j,x)

$$f_{j,x} = \sum_c f_{j,c} \times P(c|x)$$

4) Now, every site x has its own $\Pi = \begin{bmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_R & 0 & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \pi_v \end{bmatrix}$

PMSF (Posterior Mean Site Frequency) model

5) You estimate the ML tree using these pre-computed site-specific Q matrices: LG exchangeabilities + custom frequencies

➔ **Equivalent to LG+F, where F would be different for every site**

➔ Barely more computationally intensive than using the 'native' LG matrix

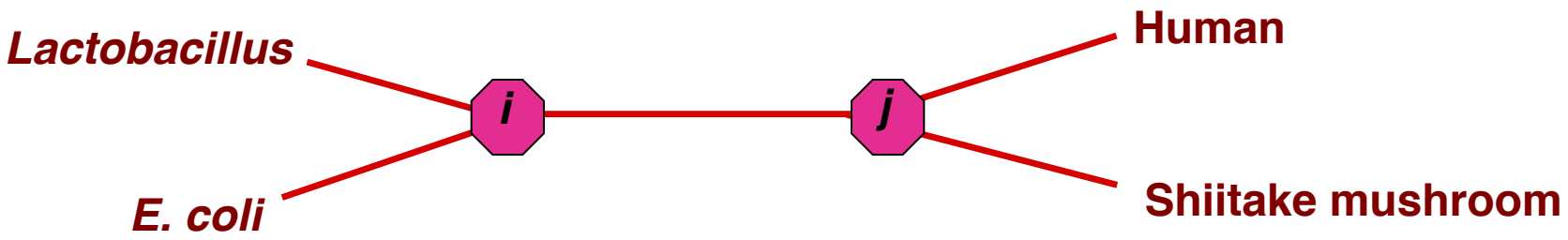
➔ Bootstrapping is dramatically faster

Take home

- Models are idealizations of the actual process of protein evolution
- Model misspecification (single-matrix models) often means systematic error (LBA)
- Mixture models deal with site-specific heterogeneity but are computationally expensive
- PMSF models provide a viable alternative for bootstrap analyses

Other types of mixture models

Probability of going from amino acid *i* to *j* at site *x*, evolving at rate r_v on branch t_e

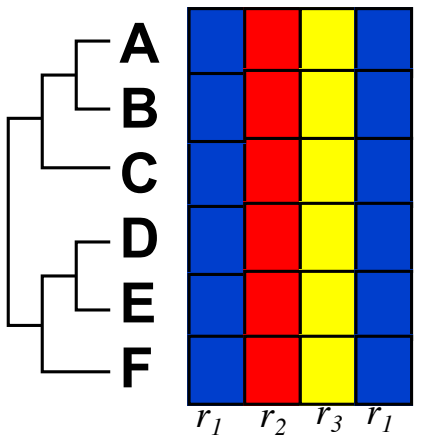


$$P(j \mid i;t) = \left[\exp(R \times \Pi \times t_e \times r_v)\right]_{ij}$$

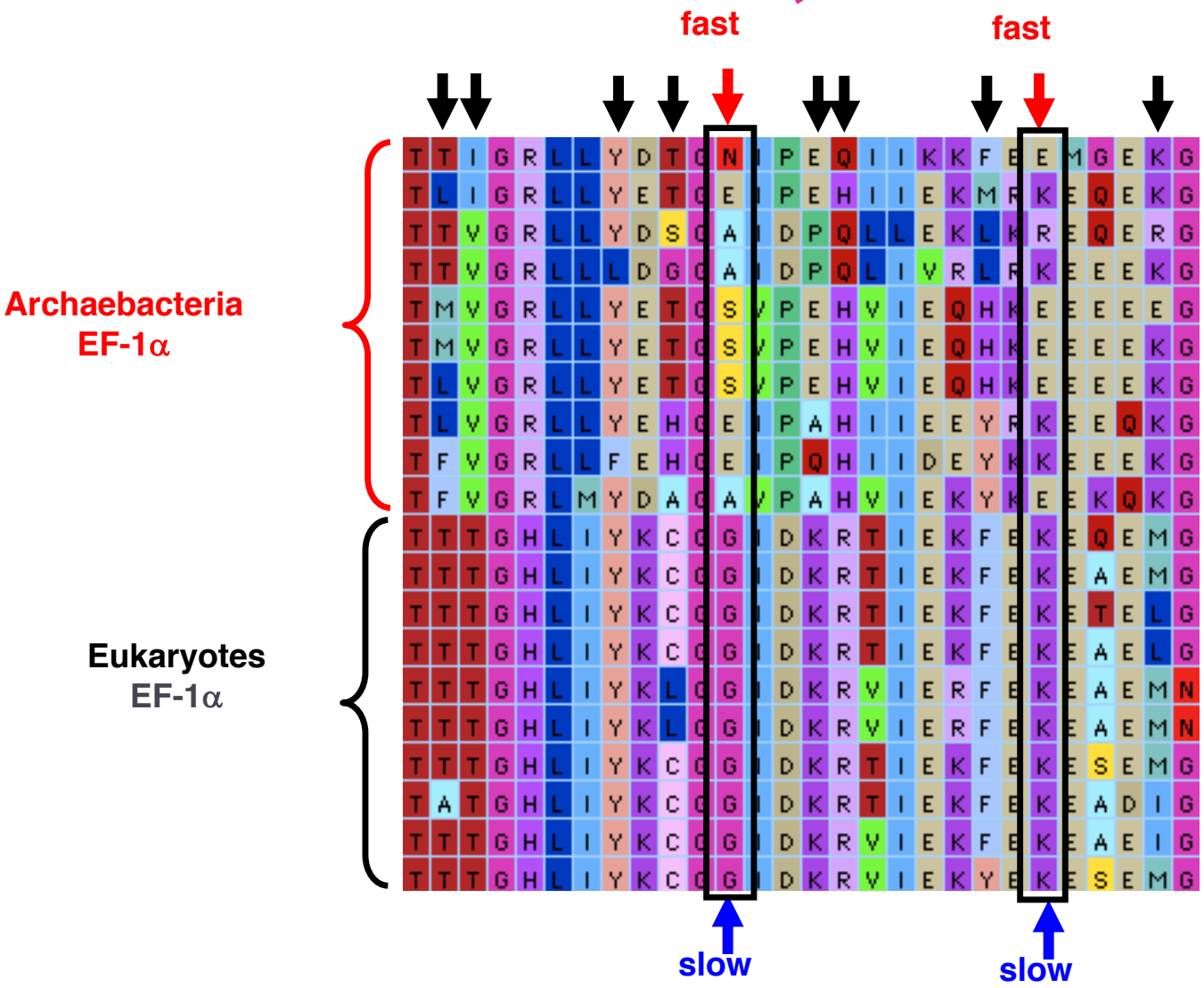
Assumptions

- 'fast-evolving' positions are always fast and slow-evolving positions are always slow

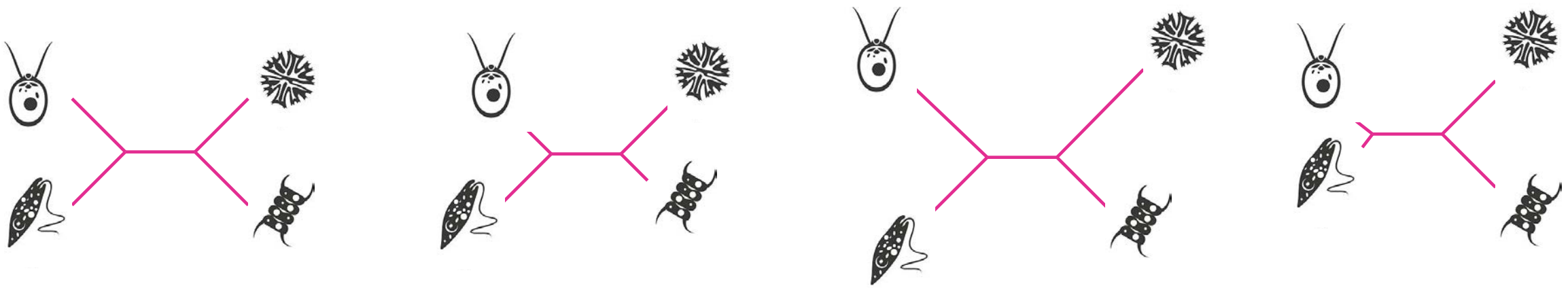
-Sites have the same rate of evolution (r_v) on different branches of tree



Changing rates of evolution at sites in different parts of the tree of life (=heterotachy)



Changing rates of evolution at sites in different parts of the tree of life (=heterotachy)



Models that deal with heterotachy (changing site rates across the tree)

- Covarion models (cf Joe's lecture)
  - Allow the sites "switch" between high rates and low rates over the tree
  - Computationally intensive

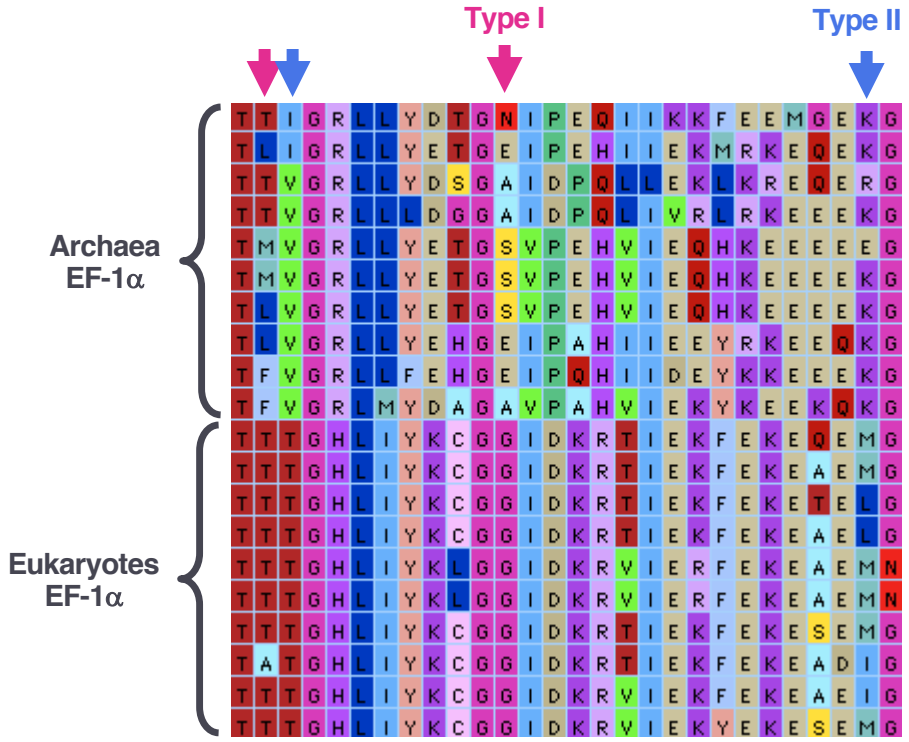
- Rate-shift models
  - Allows rates at many different sites to change abruptly on one branch

- Mixture of branch-length models
  - Allows different branch-lengths for different sites

Functionally divergent sites generate heterotachy

Functional shifts (functional divergence)

- Type I: 'rate-shifting' sites (sites that are conserved in one phylogenetic sub-group but not another).
- Type II: conserved-but-different' (conservation within both sub-groups of a phylogenetic tree but for amino acids with differing physico-chemical properties).



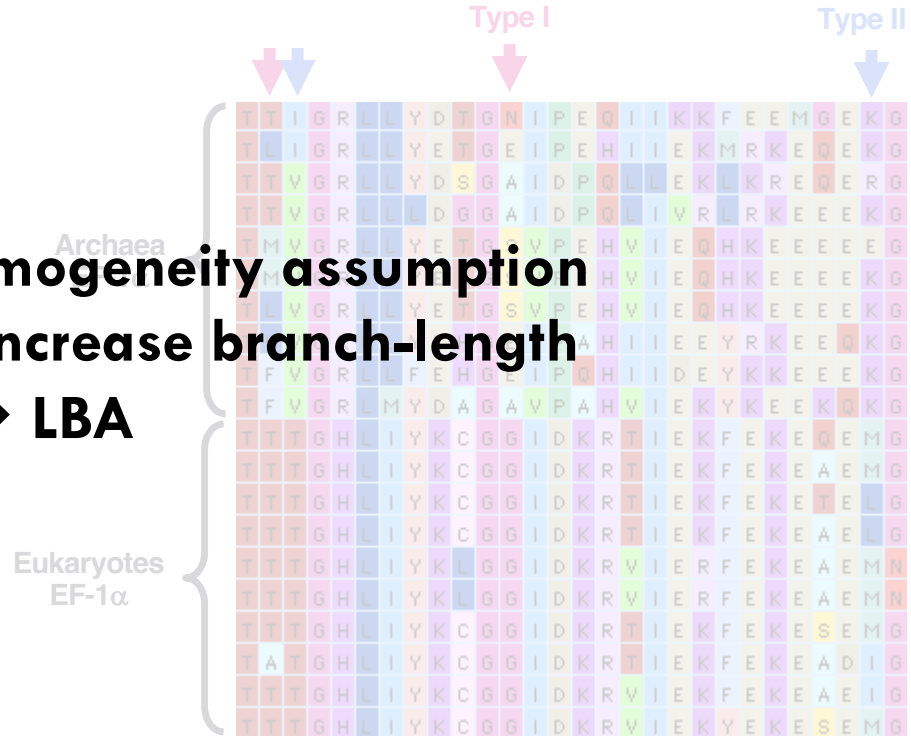
Functionally divergent sites generate heterotachy

Functional shifts (functional divergence)

- Type I: 'rate-shifting' sites (sites that are conserved in one phylogenetic sub-group but not another).
- Type II: conserved-but-different' (conservation within both sub-groups of a phylogenetic tree but for amino acids with differing physico-chemical properties).

FD sites violate homogeneity assumption and artefactually increase branch-length

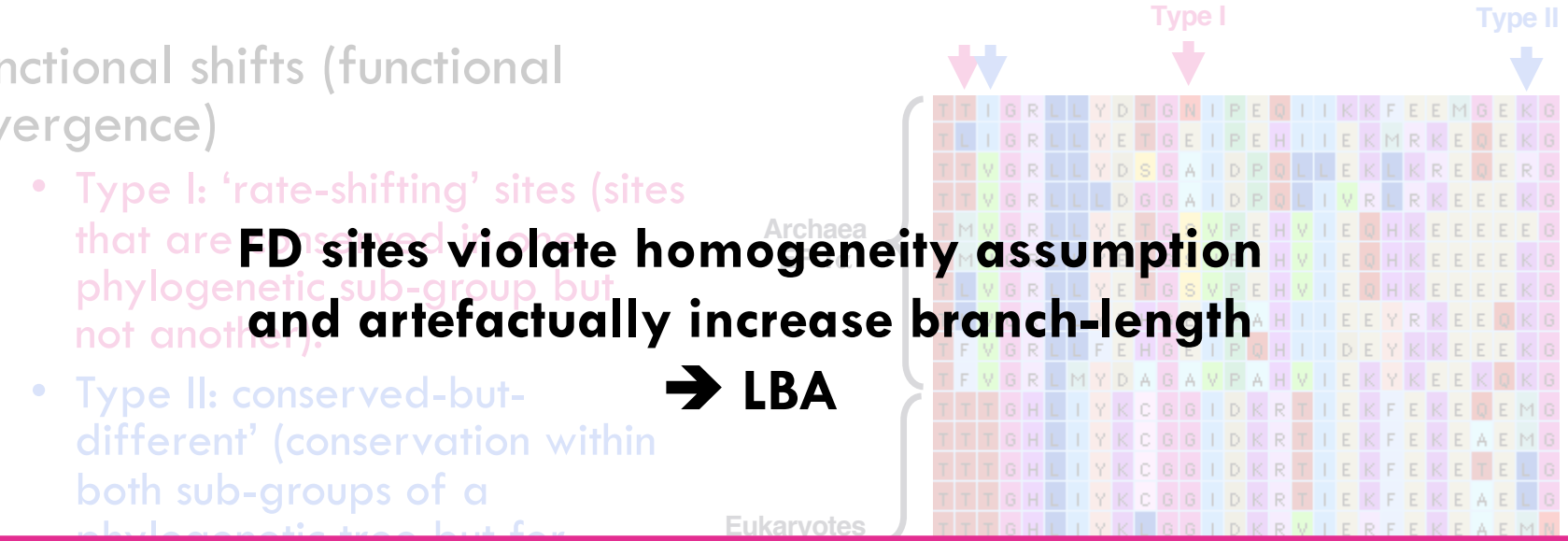
➔ LBA



Functionally divergent sites generate heterotachy

Functional shifts (functional divergence)

- Type I: 'rate-shifting' sites (sites that are conserved in one phylogenetic sub-group but not another).
- Type II: conserved-but-different' (conservation within both sub-groups of a



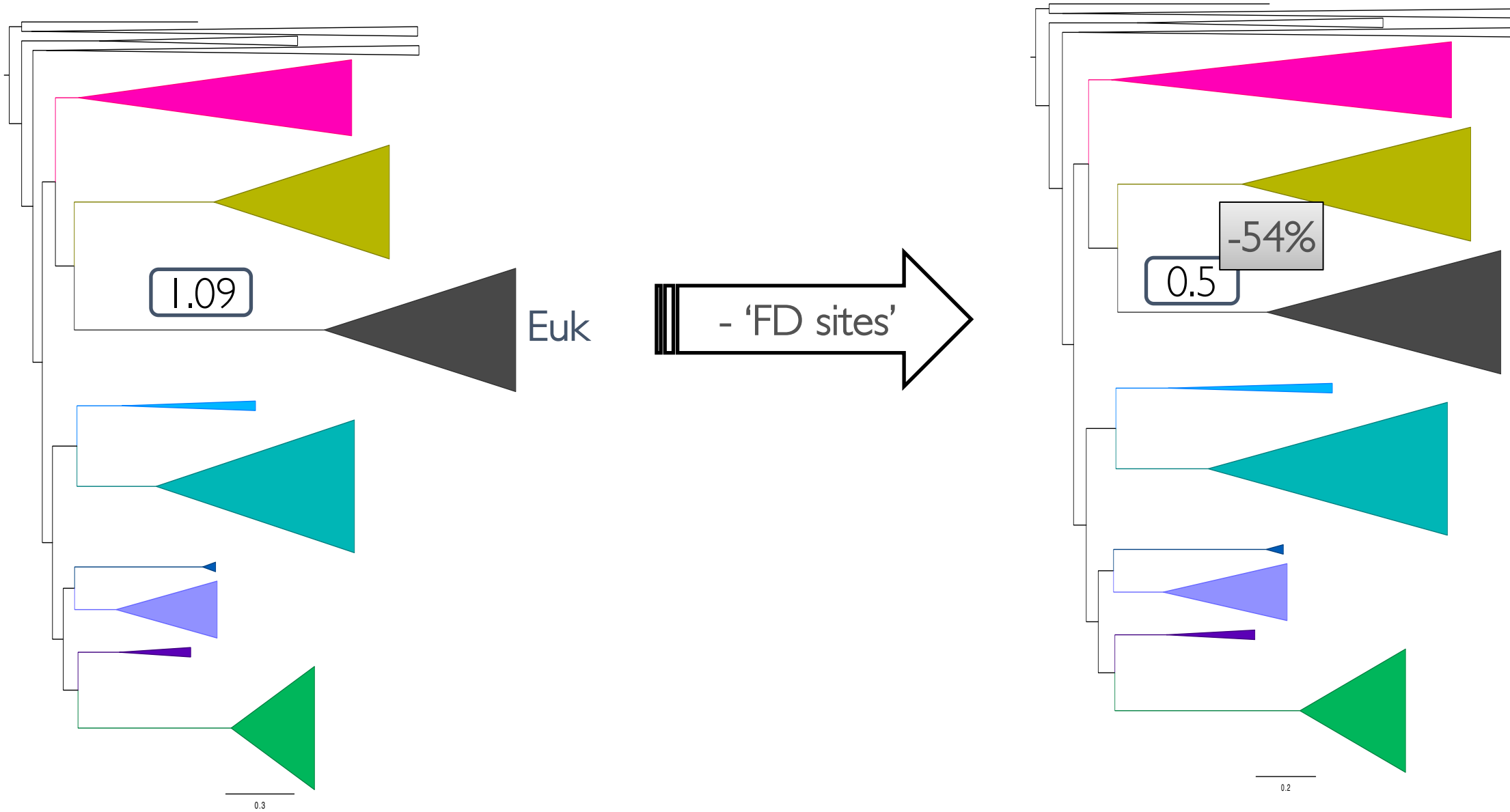
FD sites violate homogeneity assumption and artefactually increase branch-length

➔ LBA

FunDi : identifies FD sites along a specific branch taking into account the phylogeny (ML framework)

(Gaston, Susko, Roger, *Bioinformatics*)

Downside: you have to decide *a priori* which branch to analyze

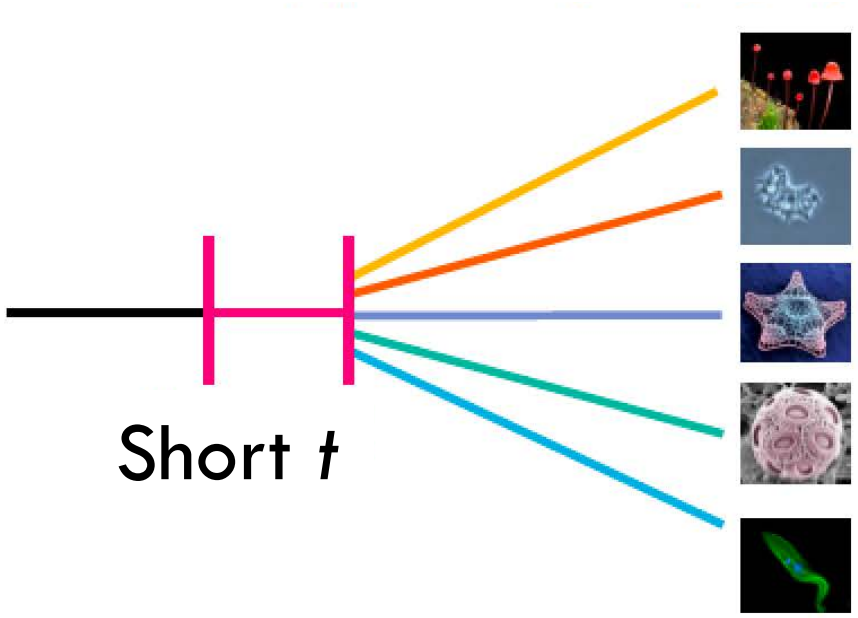

1.09
Euk
0.3
- ‘FD sites’
-54%
0.5
0.2

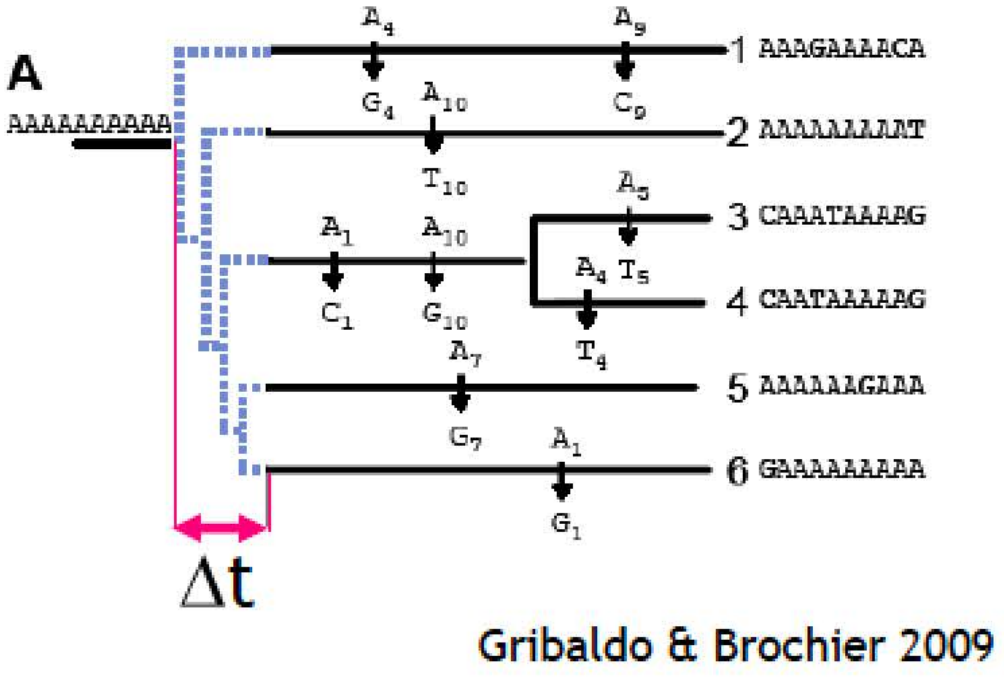
PART 2
Reconstructing ‘deep’ phylogenies

(aka large-scale species trees)

Single gene trees are not enough to resolve ‘ancient relationships’

Rapid radiation: little signal recorded during diversification

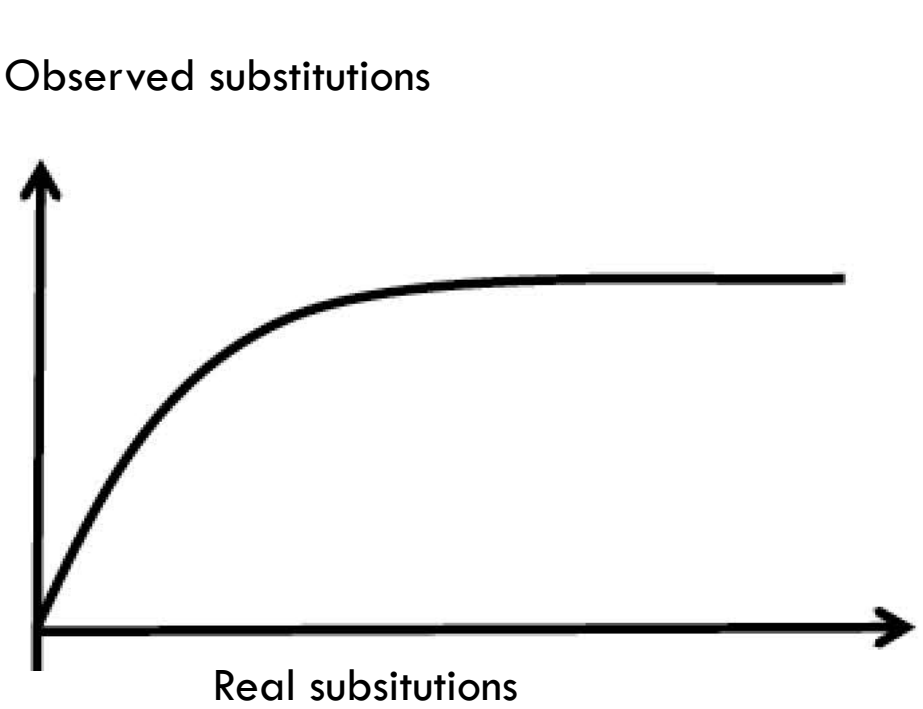


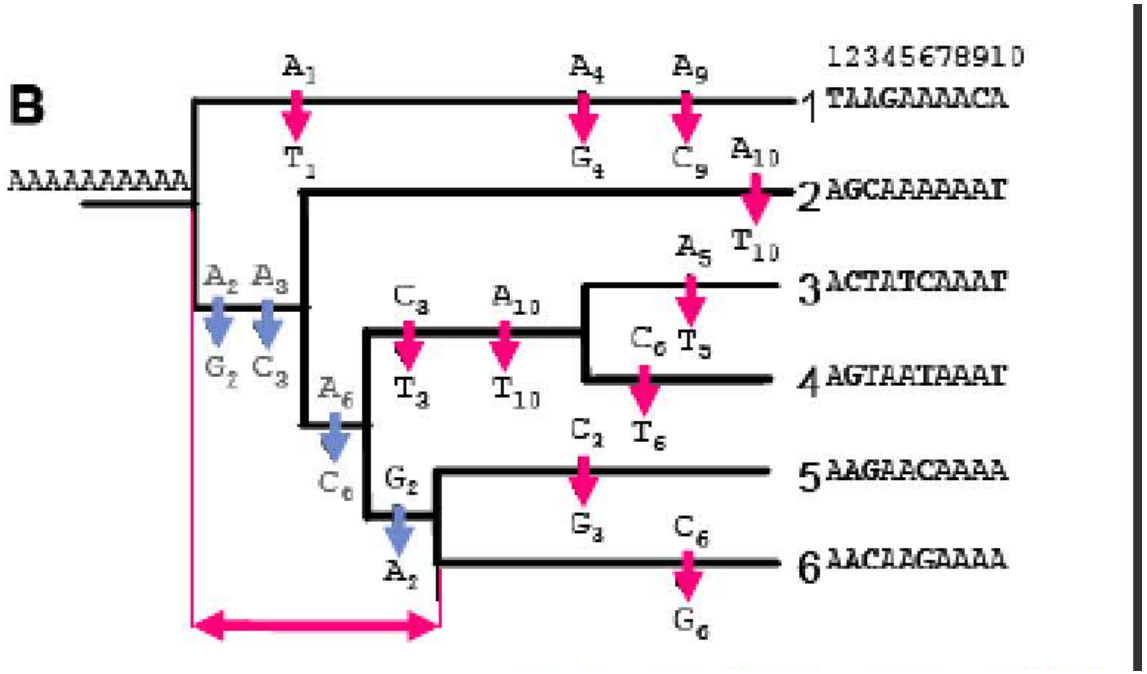


Gribaldo & Brochier 2009

Single gene trees are not enough to resolve ‘ancient relationships’

“Ancient” signal erased by more recent substitutions





How to improve phylogenetic signal

- Improve models
- Identify ‘rare’ genomic events (indels, gene fusions) used as synapormophy (be weary of convergences)
- Improve taxonomic sampling
- Increase number of analyzed sites (multi-gene analyses) 

How to combine phylogenetic signal from several genes?

- Supermatrices
- Supertrees
- Reconciliation methods

Supermatrices

Typical phylogenetic analysis (one protein):

TAXA	**SEQUENCES**
Species 1	GOODMORNING
Species 2	GODMORGON
Species 3	GOEDEMORGEN
Species 4	GUTEMORGEN

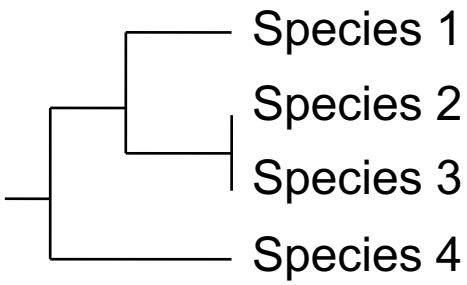
ALIGNMENT

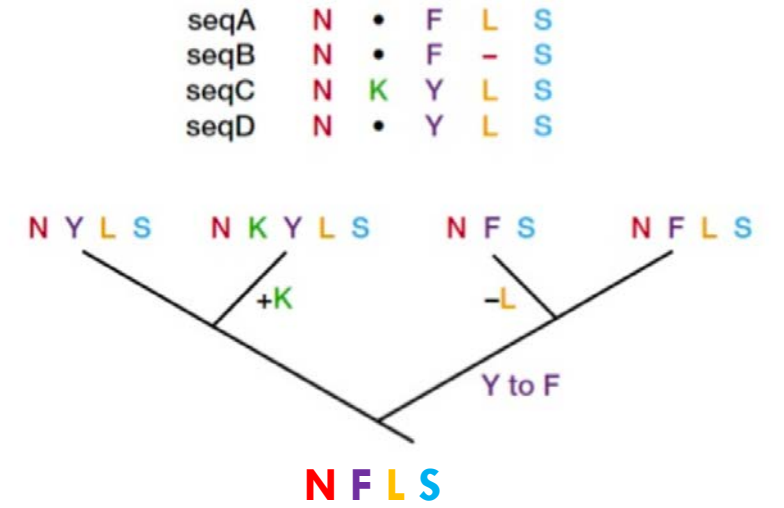
GOOD-MORNING
GO-D-MORGON-
GOEDEMORGEN-
GU-TEMORGEN-

SITE SELECTION

GODMORNN
GODMORGN
GODMORGN
GUTMORGN

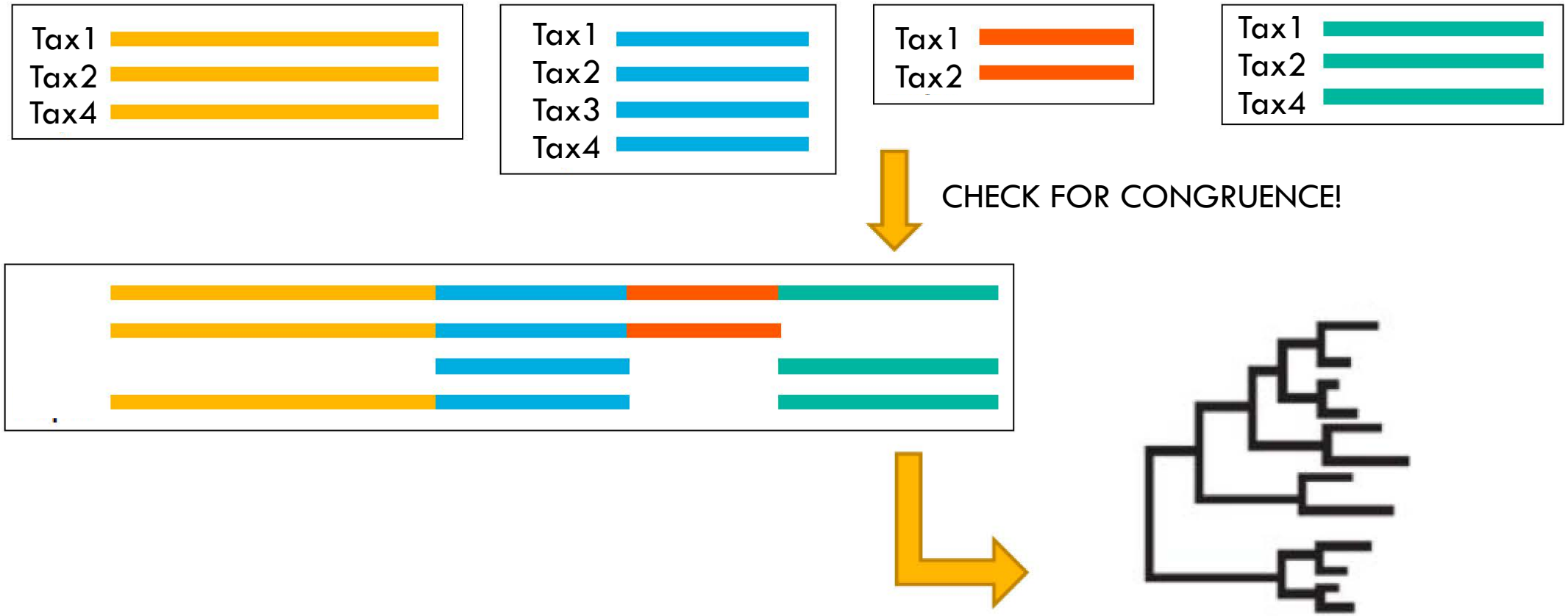
PHYLOGENY



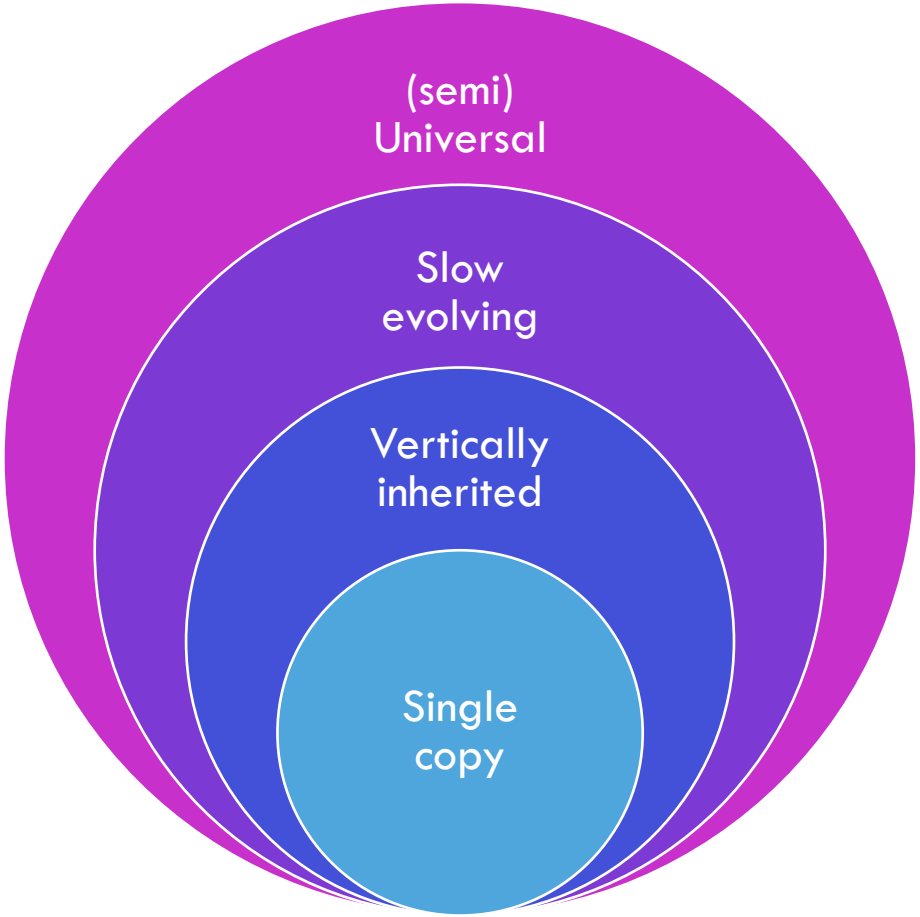


Supermatrices

Combine weak phylogenetic (historical) signal from many genes
Attenuate individual bias (IF RANDOM)



How to select multiple markers



Phylogenomic analysis "pipeline" software tools:

- PhyloGenie
- Scafos
- AMPHORA
- Orthoselect
- iPhy
- PhyloTOL

Precomputed ‘orthologous’ databases

- EggNOG
- HOMEOGEN
- OrthoMCL

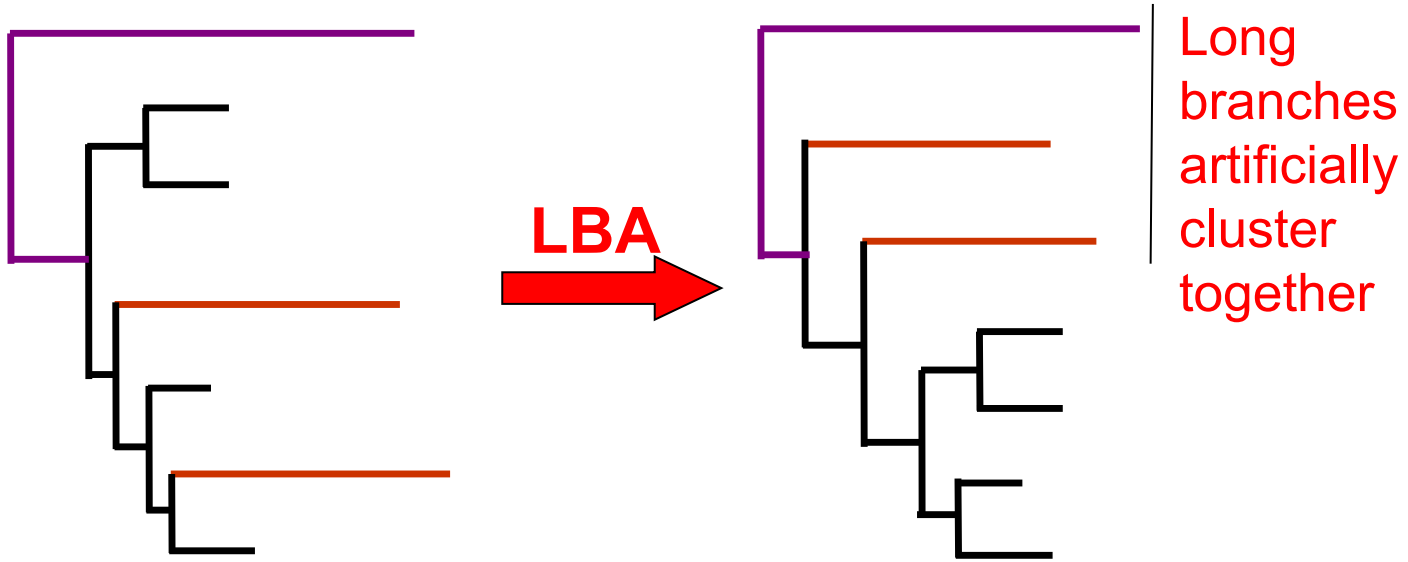
What can affect your topology

- Taxon sampling
  - Long branching taxa
  - Taxa with compositional bias
  - Contaminated data
- Gene/site sampling
  - Heterotachy
  - Saturated sites
- Model misspecification
  - LBA
- Highways of HGT
  - Consistently conflicting with vertical signal
- (and many other things…)

Minimazing potential artefacts

Model mispecification or over-simplistic: statistical inconsistency

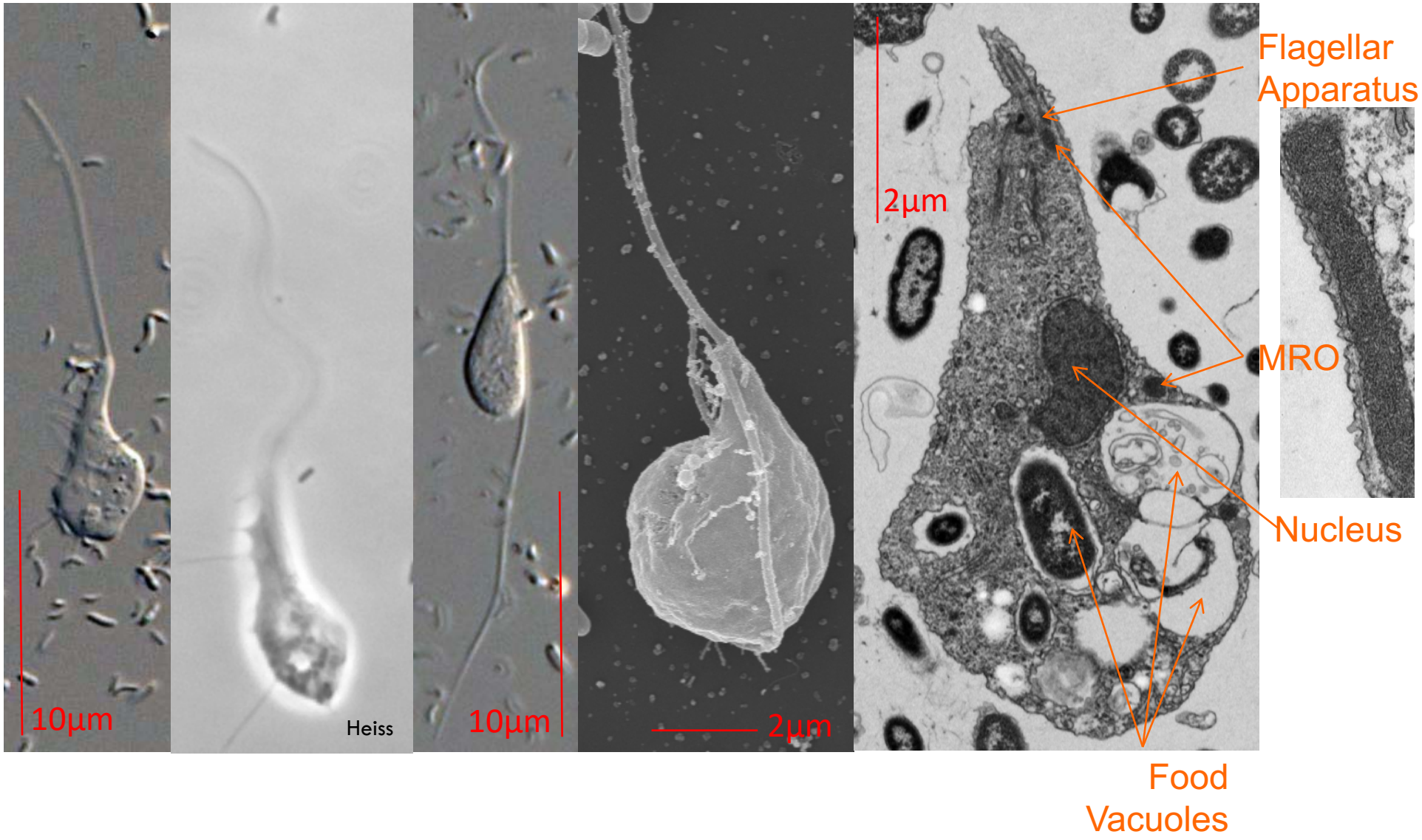
Long Branch Attraction (LBA) Artefact



Adding more data *strengthens* artefact

→ statistical inconsistency

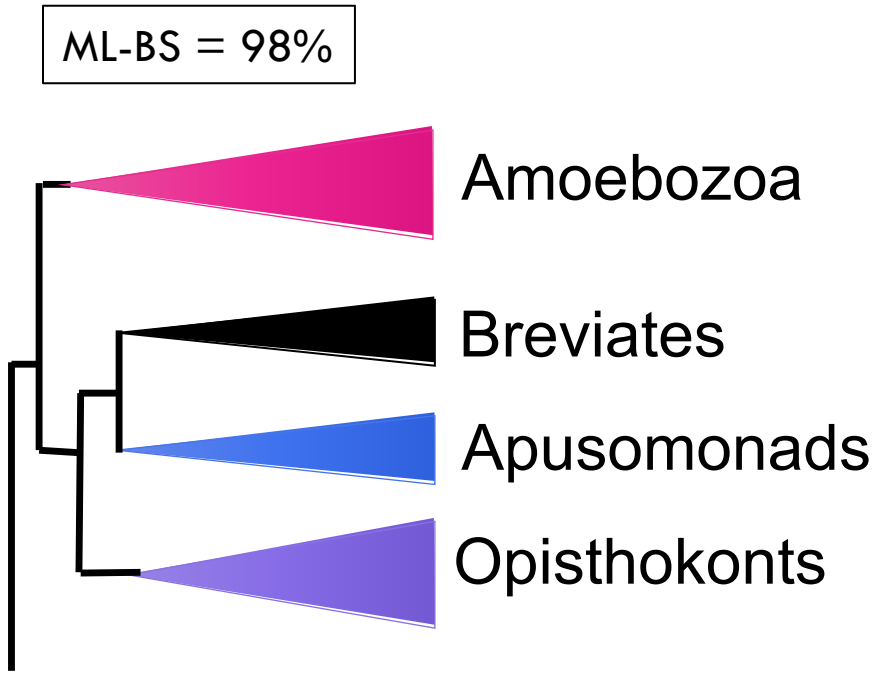
Pygsuia biforma n. gen. n. sp.

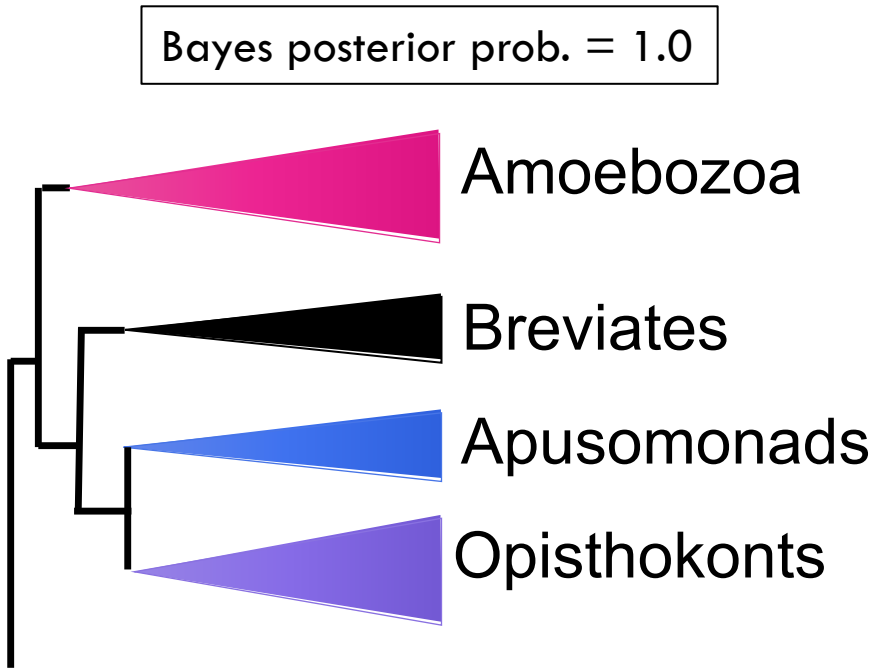


Two different topologies within Obazoa are supported by different phylogenetic models

Opisto + Breviates + Apusomonads = OBAzoa

Two different topologies within Obazoa are supported by different phylogenetic models





ML – LG+Γ
Bayes – LG+Γ

Bayes – CAT-Poisson+Γ
Bayes – CAT-GTR+Γ

How to decide which is real and which is artefact?

- One of two topologies is likely artefactual resulting from misspecified model

- Test which substitution model fits better
  - E.g., Cross-validation, Bayes factors, Posterior prediction

How to decide which is real and which is artefact?

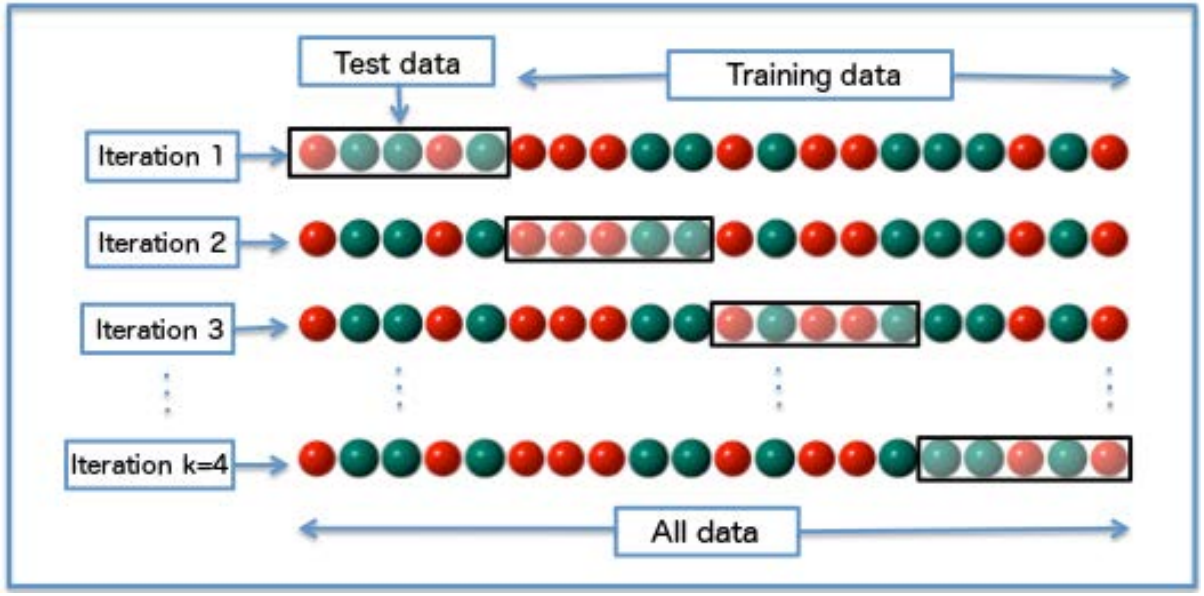
Cross-validation:

1) parameters of the model estimated on the learning set

2) these parameter values are then used to compute the likelihood of the test set = **how well the test set is 'predicted' by the model?**

3) Repeat over all partitions and average the likelihood

4) Repeat for each model and compare



Cross-validation favors CAT-GTR over LG

ML-BS = 98%

Bayes posterior prob. = 1.0

ML – LG+Γ

Bayes – LG+Γ

Bayes – CAT-Poisson+Γ

Bayes – CAT-GTR+Γ

Cross validation

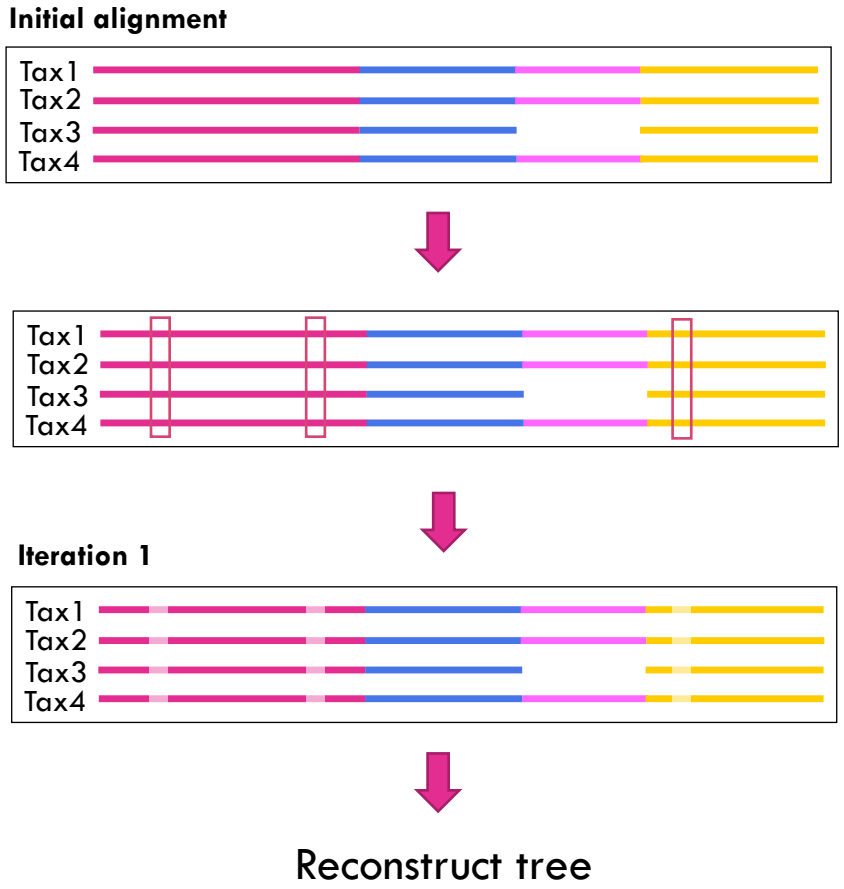
How do decide which is real and which is artefact?

- One of two topologies is likely artefactual resulting from misspecified model
- Test which substitution model fits better
  - Cross-validation
- Try to eliminate ‘noisiest’ data
  - Fast-evolving site removal
  - Fast-evolving gene removal
  - Fast-evolving taxon removal
  - Recoding

Removal of fast-evolving sites

Fast Evolving Sites removal

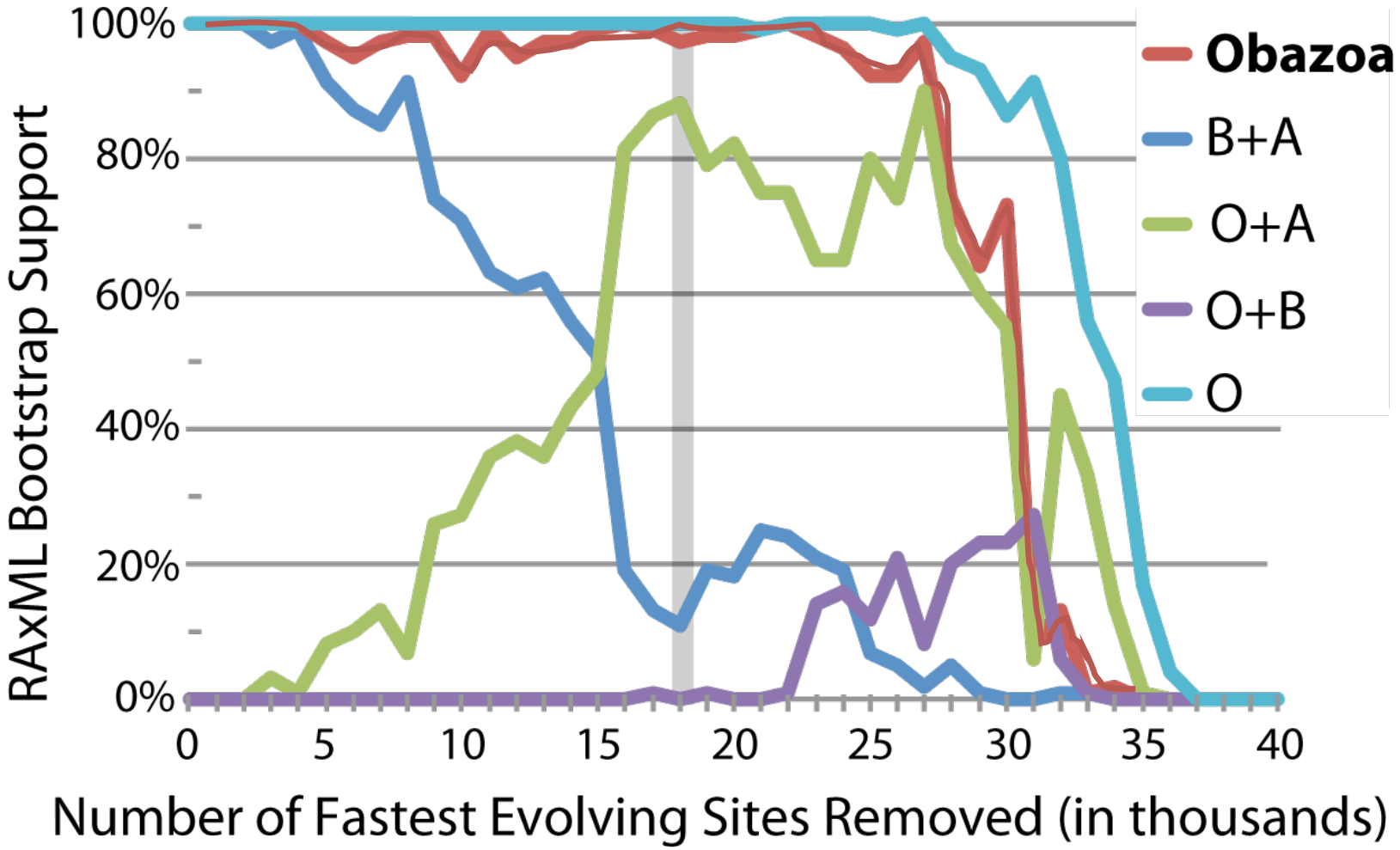
Fast-evolving sites : carry the 'noisiest' signal (most saturated sites)



➔ 40 steps of removal of 1000 sites *(evolutionary rates estimated by IQTREE for example)*

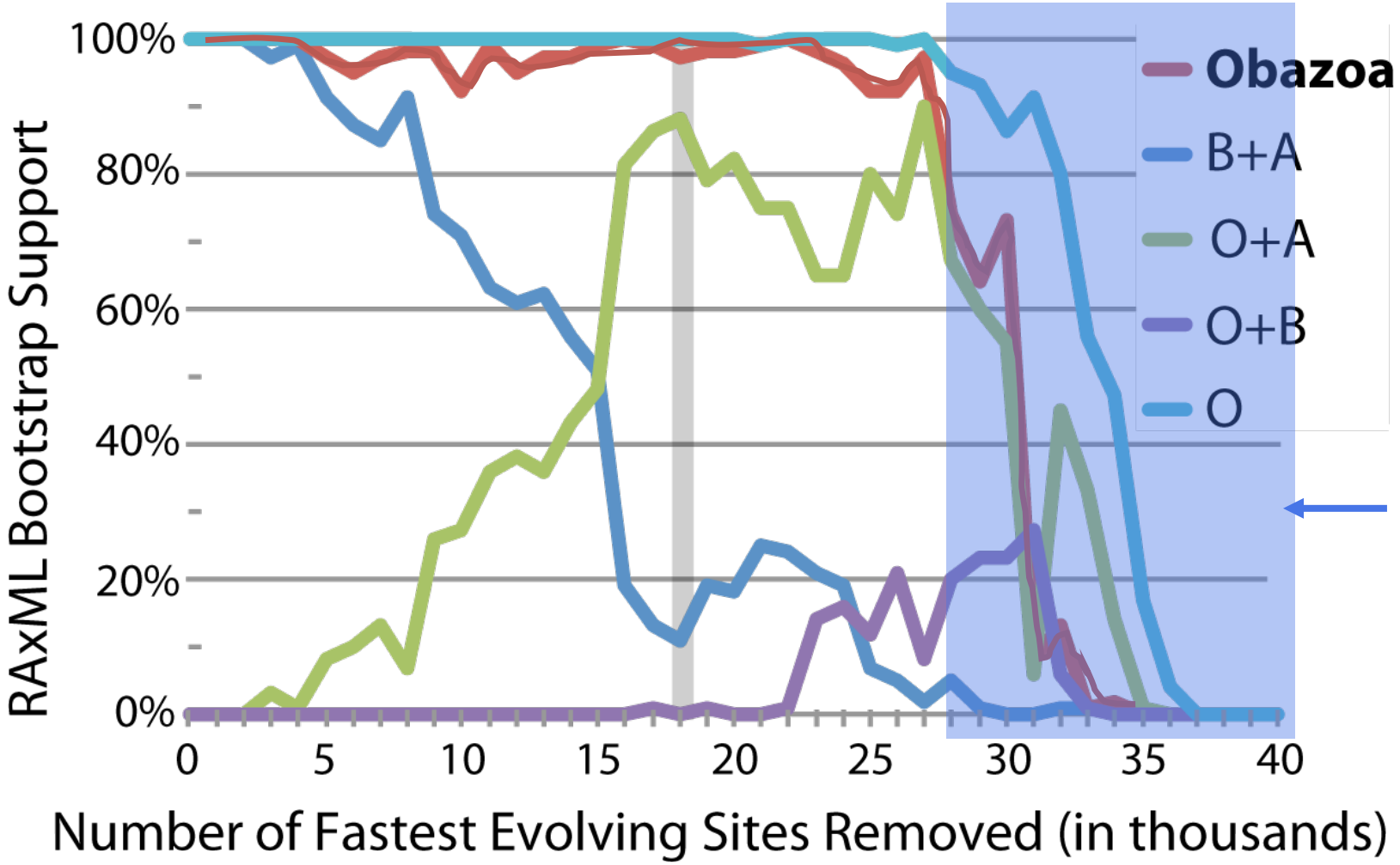
Step	# sites left	
1	43615	Tree1
2	42615	Tree 2
…		
4	40615	Tree 4
…		
40	3615	Tree 40

Estimate BS for various clades as we remove Fast-Evolving Sites



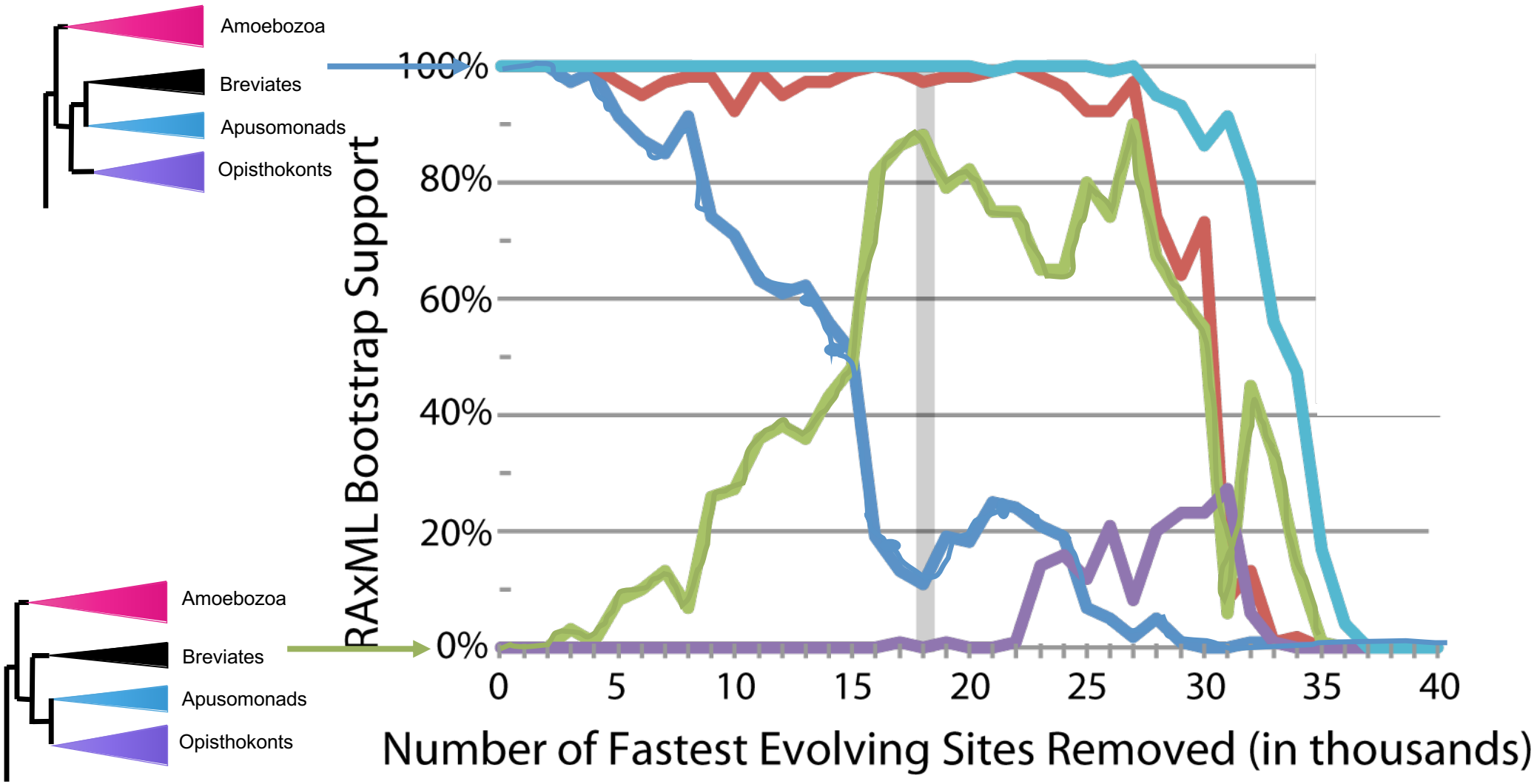
Brown et al. 2013 *PRSB*

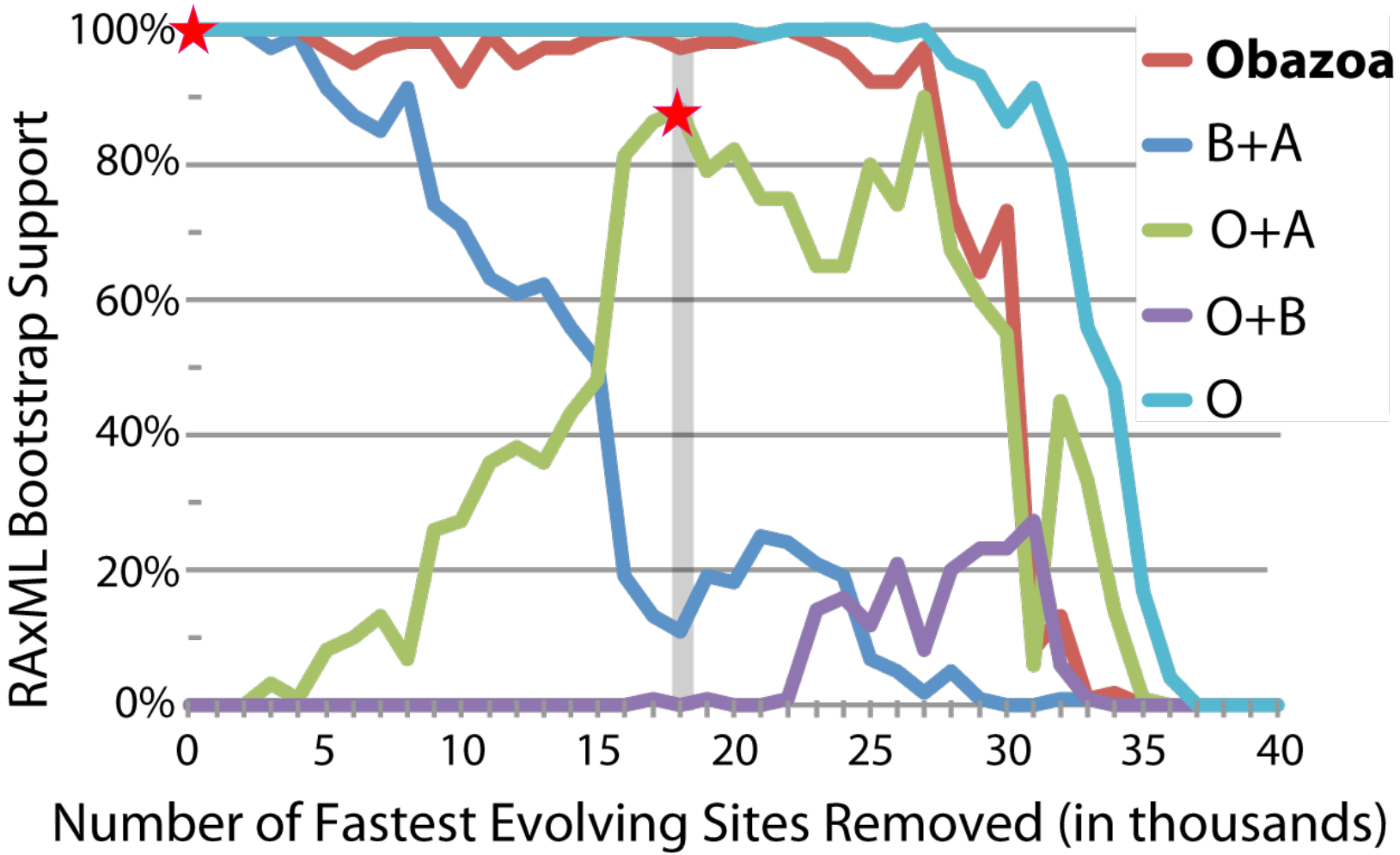
Estimate BS for various clades as we remove Fast-Evolving Sites

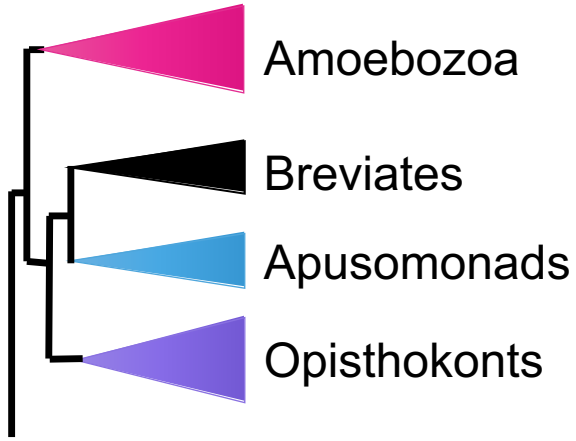


Brown et al. 2013 *PRSB*

Breviates+Apusomonads (B+A) topology vs. Apusomonads+Opisthokonts (O+A)

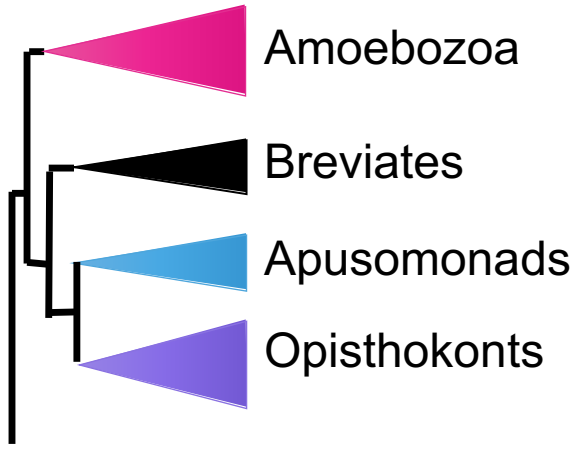



100%
80%
60%
40%
20%
0%
RAxML Bootstrap Support
0
5
10
15
20
25
30
35
40
Number of Fastest Evolving Sites Removed (in thousands)
Obazoa
B+A
O+A
O+B
O



ML – LG+Γ

Removal of 18,000 fastest-evolving sites



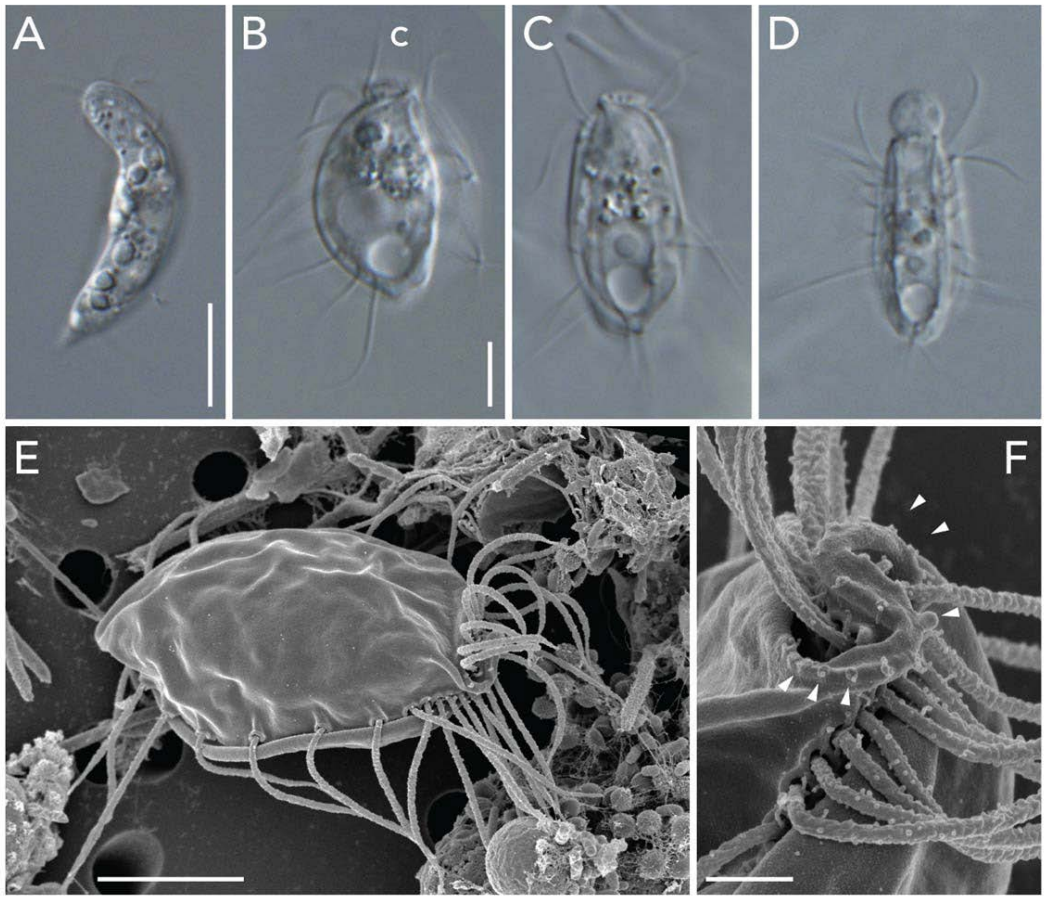
ML – LG+Γ

Fast-evolving site removal: warning

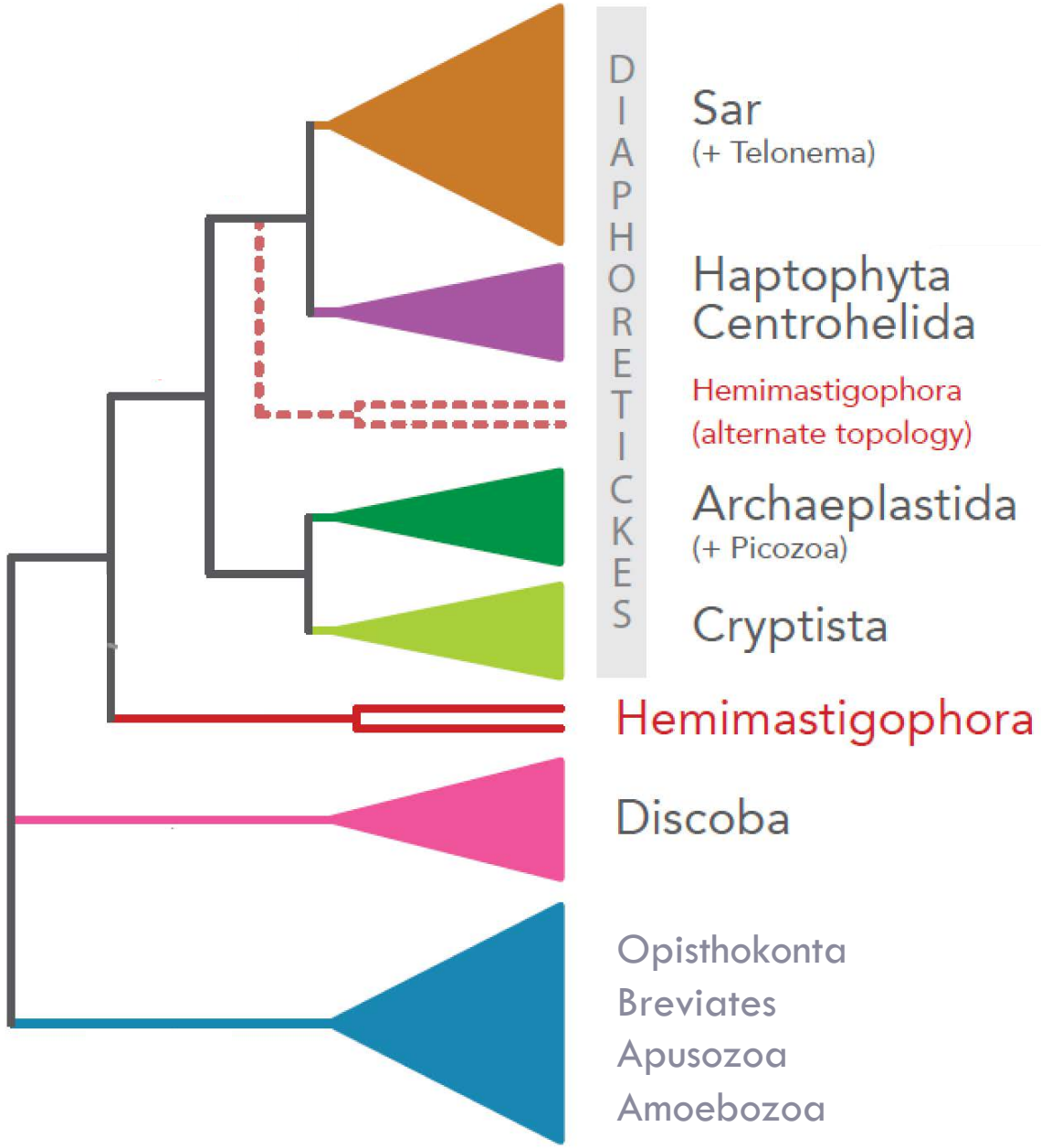
- Very poor proxy for heterotacheous site removal
  - Fast sites in themselves are not a problem if they are fast across the entire tree

- In practice, fast sites seem to overlap to some extant with sites whose rate varies across the tree and are improperly modelled by most widely used models.
- You also remove the most saturated sites, which are usually poorly modelled

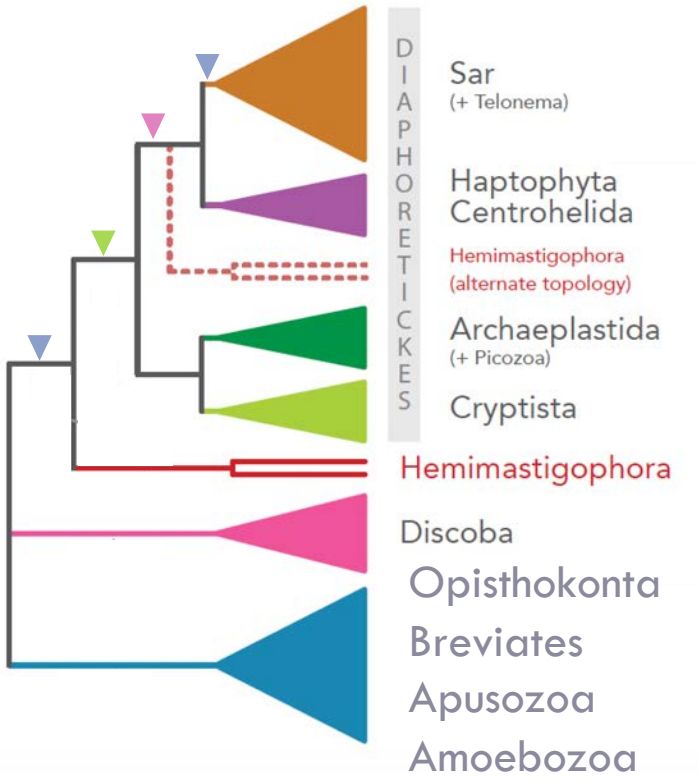
FES removal is not the key to all conflictual signal

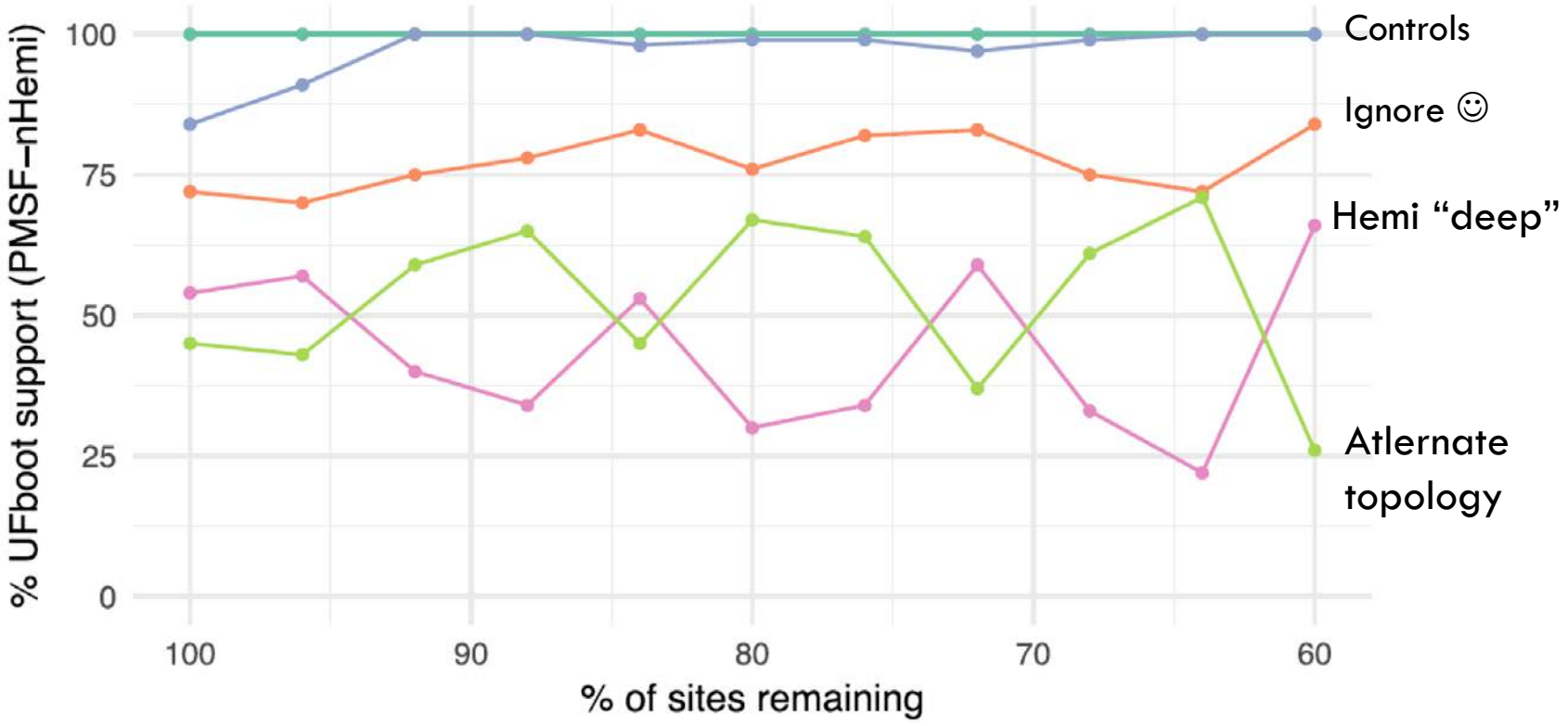


Lax*, Eglit*, Eme* et al., *Nature* 2018



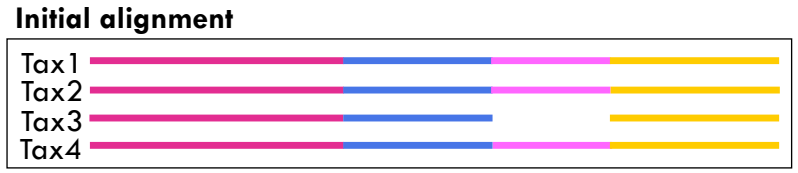
FES removal is not the key to all conflictual signal

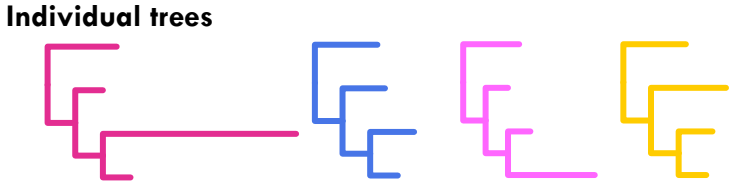




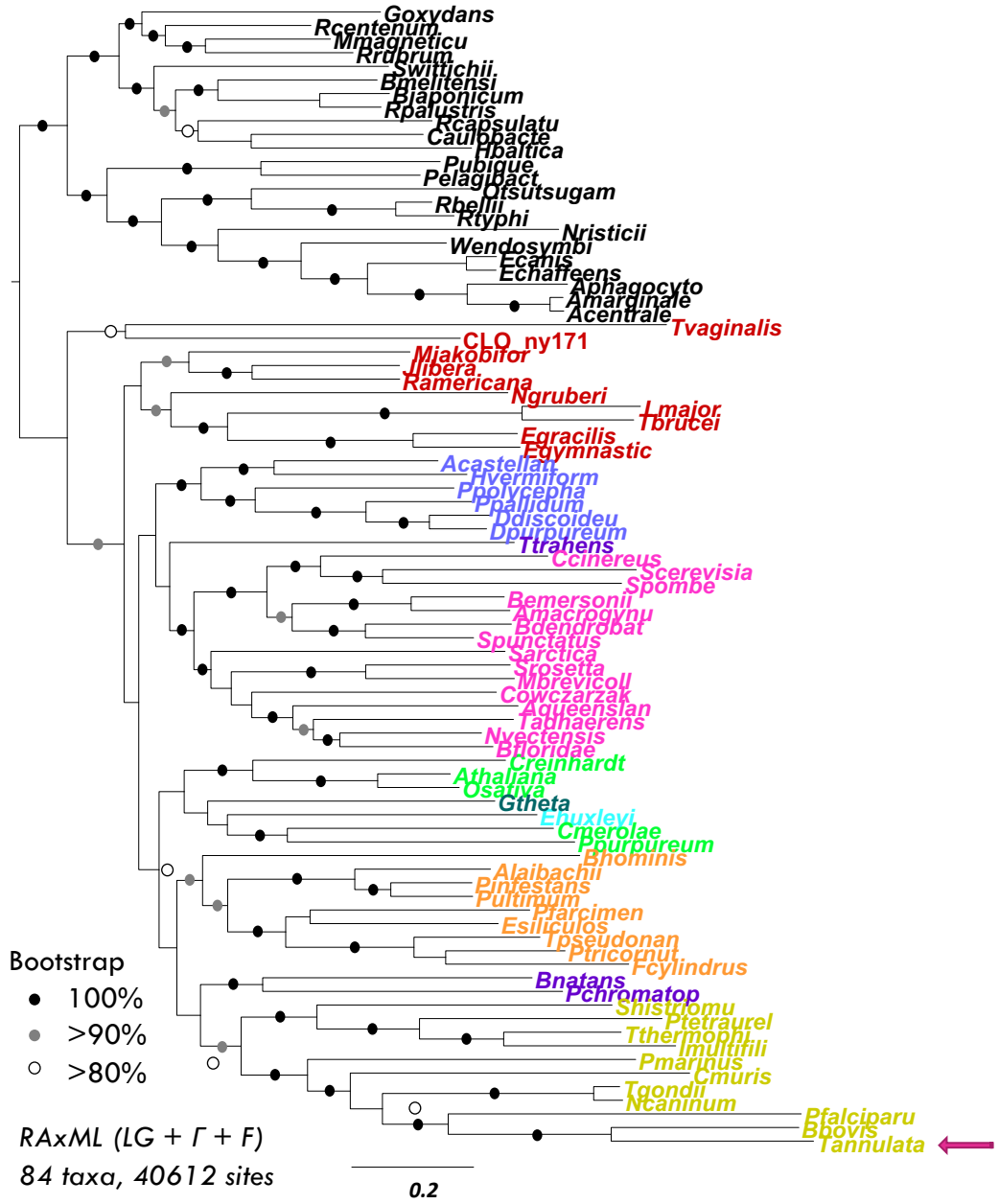
Removal of long-branching genes

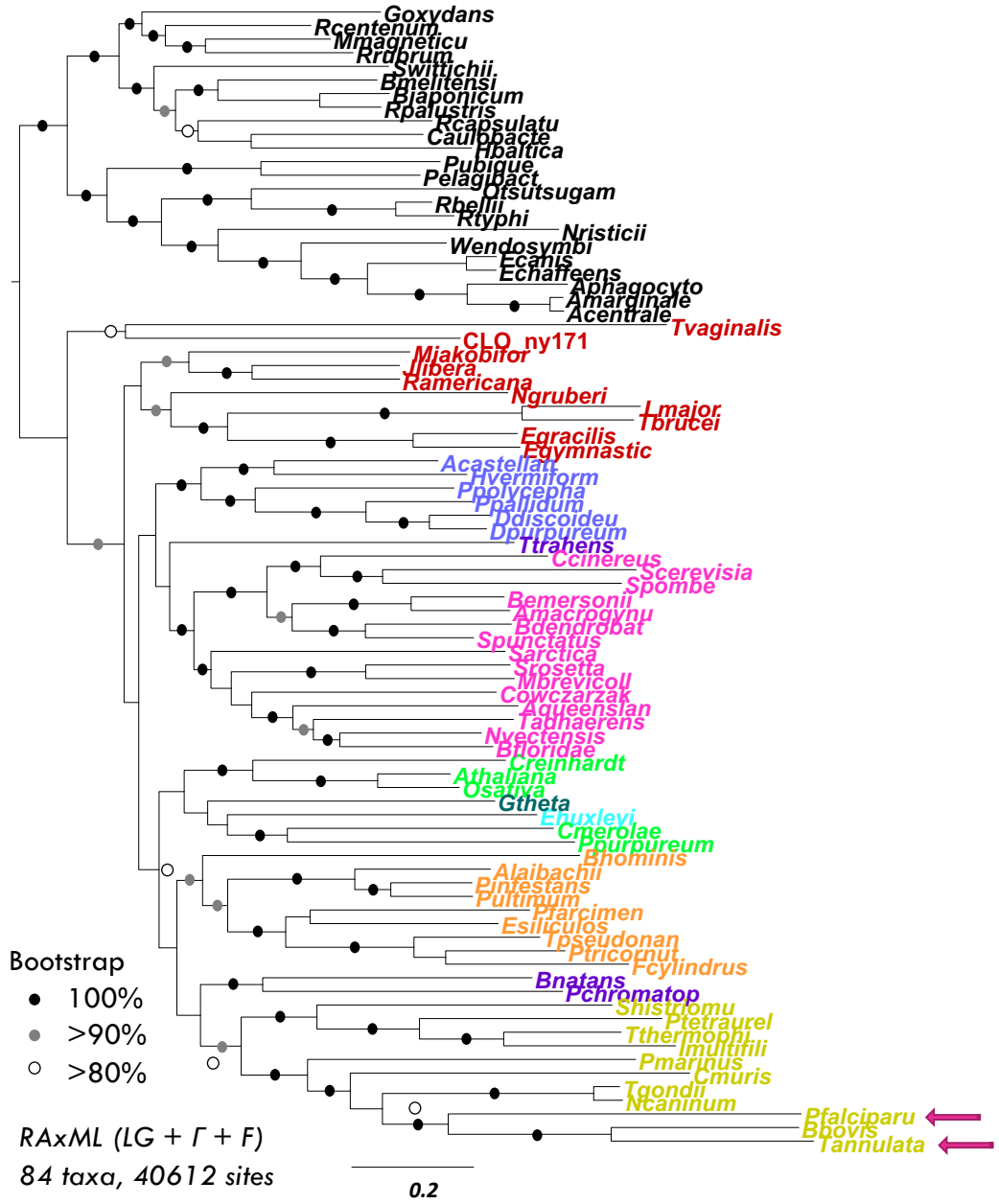
Long-branching genes removal

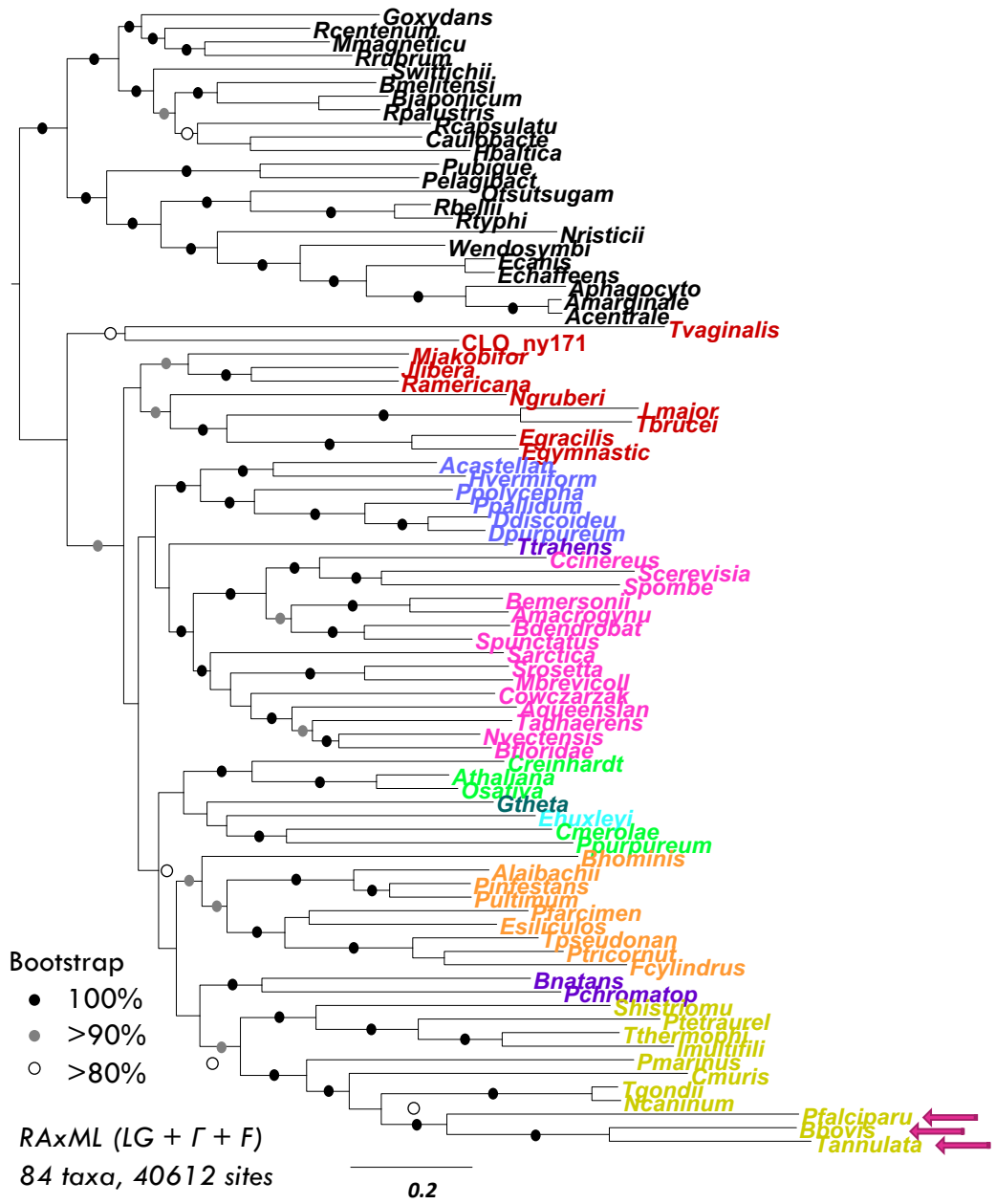




Removal of long-branching taxa


Goxydans
Rcentenum
Mmagneticu
Rrubrum
Swittichii
Bmelitensi
Bjaponicum
Rpalustris
Rcapsulatu
Caulobacte
Hbaltica
Pubique
Pelagibact
Orsutsugam
Rbellii
Rtyphi
Nristicii
Wendosymbi
Ecanis
Echaffeens
Aphagocyto
Amarginale
Acentrale
Tvaginalis
CLO_ny171
Miakobifor
Jlibera
Ramericana
Ngruberi
Lmajor
Tbrucei
Egracilis
Egymnastic
Acastellati
Hvermiform
Ppolycepha
Pnalijdum
Ddiscoideu
Dpurpureum
Ttrahens
Ccinereus
Scerevisia
Spombe
Bemersonii
Amacrogynu
Bdendrobat
Spunctatus
Sarctica
Srosetta
Mbrevicoll
Cowczarzak
Aqueenslan
Tadhaerens
Nvectensis
Bfloridae
Creinhardt
Athaliana
Osativa
Gtheta
Ehuxlevi
Cmerolae
Pourpureum
Bhominis
Alaibachii
Pinfestans
Pultimum
Pfarcimen
Esiliculos
Tpseudonan
Ptricornut
Fcylindrus
Bnatans
Pchromatop
Shistromu
Ptetraurel
Tthermophi
Imultifili
Pmarinus
Cmuris
Tgondii
Ncaninum
Pfalciparu
Bbovis
Tannulata
Bootstrap
100%
>90%
>80%
RAxML (LG + Γ + F)
84 taxa, 40612 sites
0.2


Goxydans
Rcentenum
Mmagneticu
Rrubrum
Swittichii
Bmelitensi
Bjaponicum
Rpalustris
Rcapsulatu
Caulobacte
Hbaltica
Pubique
Pelagibact
Orsutsugam
Rbellii
Rtyphi
Nristicii
Wendosymbi
Ecanis
Echaffeens
Aphagocyto
Amarginale
Acentrale
Tvaginalis
CLO_ny171
Miakobifor
Jlibera
Ramericana
Ngruberi
Lmajor
Tbrucei
Egracilis
Egymnastic
Acastellati
Hvermiform
Ppolycepha
Pnalijdum
Ddiscoideu
Dpurpureum
Ttrahens
Ccinereus
Scerevisia
Spombe
Bemersonii
Amacrogynu
Bdendrobat
Spunctatus
Sarctica
Srosetta
Mbrevicoll
Cowczarzak
Aqueenslan
Tadhaerens
Nvectensis
Bfloridae
Creinhardt
Athaliana
Osativa
Gtheta
Ehuxlevi
Cmerolae
Pourpureum
Bhominis
Alaibachii
Pinfestans
Pultimum
Pfarcimen
Esiliculos
Tpseudonan
Ptricornut
Fcylindrus
Bnatans
Pchromatop
Shistromu
Ptetraurel
Tthermophi
Imultifili
Pmarinus
Cmuris
Tgondii
Ncaninum
Pfalciparu
Bbovis
Tannulata
Bootstrap
100%
>90%
>80%
RAxML (LG + Γ + F)
84 taxa, 40612 sites
0.2


Goxydans
Rcentenum
Mmagneticu
Rrubrum
Swittichii
Bmelitensi
Bjaponicum
Rpalustris
Rcapsulatu
Caulobacte
Hbaltica
Pubique
Pelagibact
Orsutsugam
Rbellii
Rtyphi
Nristicii
Wendosymbi
Ecanis
Echaffeens
Aphagocyto
Amarginale
Acentrale
Tvaginalis
CLO_ny171
Miakobifor
Jlibera
Ramericana
Ngruberi
Lmajor
Tbrucei
Egracilis
Egymnastic
Acastellati
Hvermiform
Ppolycepha
Pnalijdum
Ddiscoideu
Dpurpureum
Ttrahens
Ccinereus
Scerevisia
Spombe
Bemersonii
Amacrogynu
Bdendrobat
Spunctatus
Sarctica
Srosetta
Mbrevicoll
Cowczarzak
Aqueenslan
Tadhaerens
Nvectensis
Bfloridae
Creinhardt
Athaliana
Osativa
Gtheta
Ehuxlevi
Cmerolae
Pourpureum
Bhominis
Alaibachii
Pinfestans
Pultimum
Pfarcimen
Esiliculos
Tpseudonan
Ptricornut
Fcylindrus
Bnatans
Pchromatop
Shistromu
Ptetraurel
Tthermophi
Imultifili
Pmarinus
Cmuris
Tgondii
Ncaninum
Pfalciparu
Bbovis
Tannulata
Bootstrap
100%
>90%
>80%
RAxML (LG + Γ + F)
84 taxa, 40612 sites
0.2

Observing a shift of topology when removing long branching taxa suggests a potential reconstruction artefact

Identifying conflict between single-gene trees

Detecting conflict between single-gene trees

Characteristics of Popular Congruence Tests

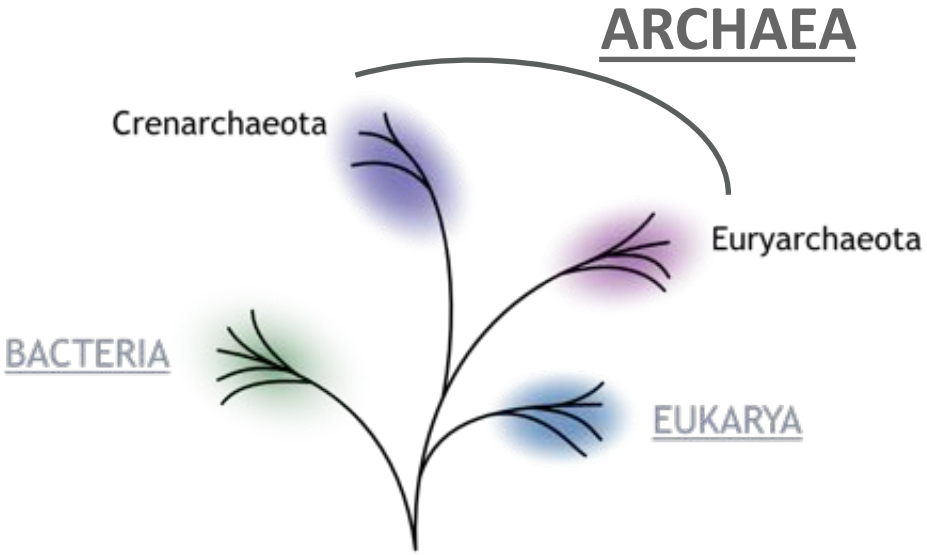
Test	H_0	Algorithmic Complexity[a]	Identification of Multiple Subsets?	Interpretation of Missing Taxa
MAST (Lapointe and Rissler 2005; de Vienne et al. 2007)	Incongruence	$O(n)$	Yes[b]	Pruned and ignored[b]
CADM (Campbell et al. 2009)	Incongruence	$O(n^2)$	Yes	N/A
ILD (Farris et al. 1994)	Congruence	$O(n)$[c]	No	N/A
Multiple ILD (Planet and Sarkar 2005)	Congruence	$O(n^2)$	Yes	Pruned and ignored
LRT (Huelsenbeck and Bull 1996)	Congruence	$O(n)$[c]	No	N/A
Concaterpillar hierarchical LRT (Leigh et al. 2008)	Congruence	$O(n^2)$	Yes	Pruned and ignored
LRT (Waddell et al. 2000)	Congruence	$O(nm)$	No	N/A
Likelihood-based topology tests	Congruence	$O(nm)$	No	Pruned and ignored
Principal component analysis	Congruence	$O(nm)$	No	Pruned and ignored
Heatmaps	Congruence	$O(nm)$	Yes	Pruned and ignored
Likelihood-based topology tests	Congruence[d]	$O(nm)$	No	N/A

Leigh et al., GBE 2011

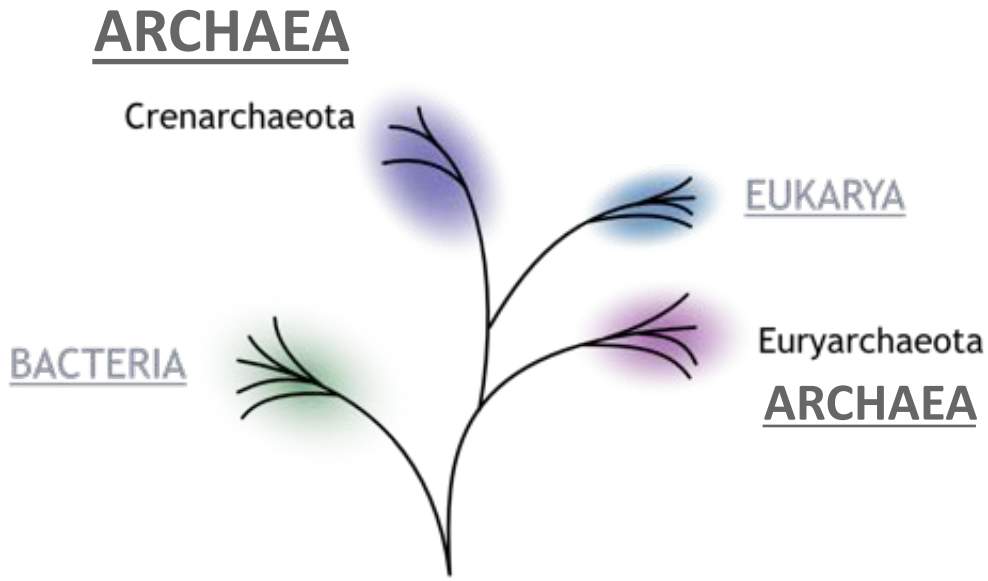
Small systematic bias leads to large artefacts in concatenations

An example from my recent research

Archaea as **sister-group** or as **ancestors** of eukaryotes?



Three domain
tree of Life

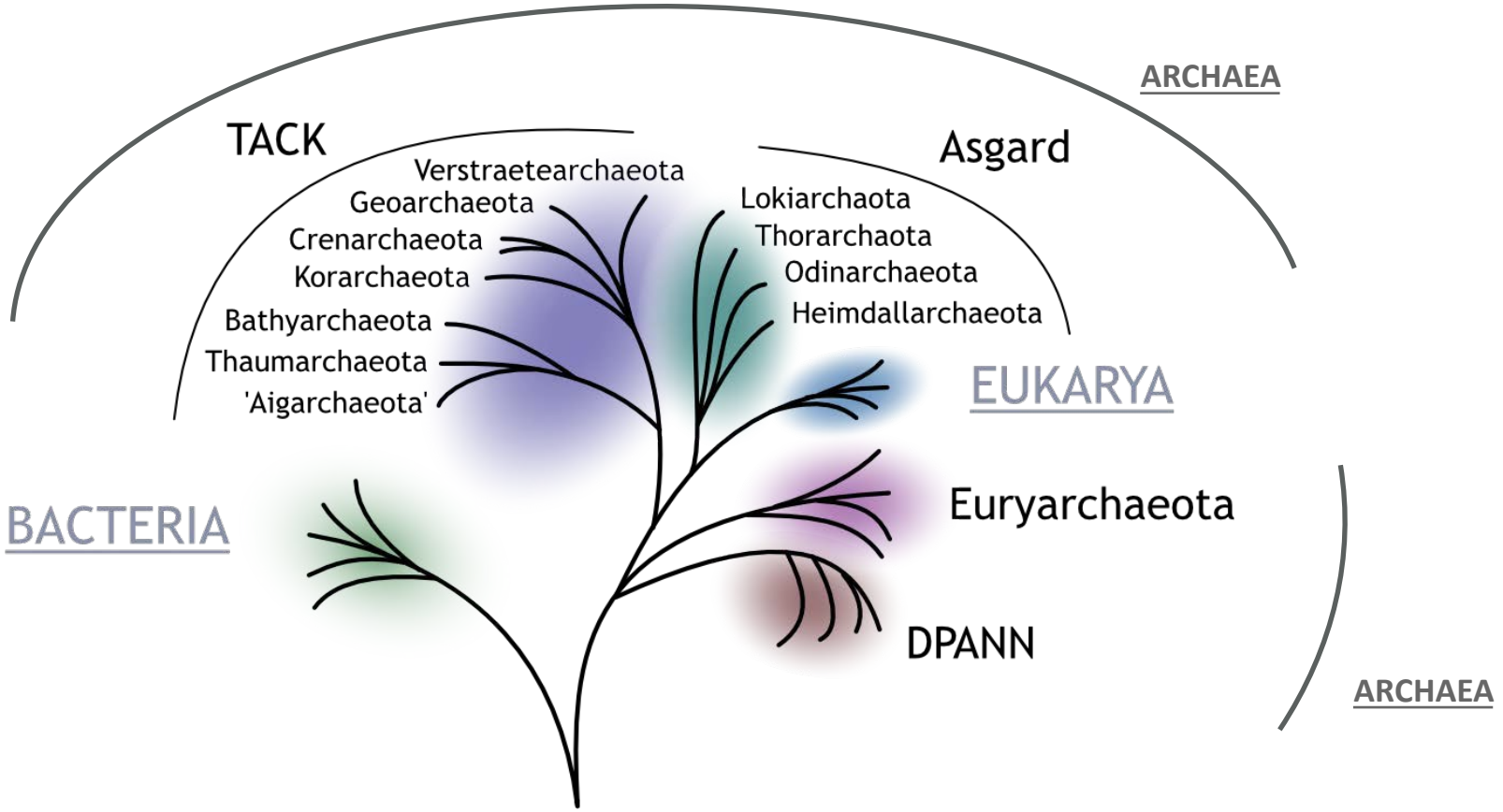


Two domain
tree of Life

1990s-2000s: Phylogenetic analyses: few genes; few cultivated organisms

Culture-independent genomics (e.g. metagenomics)

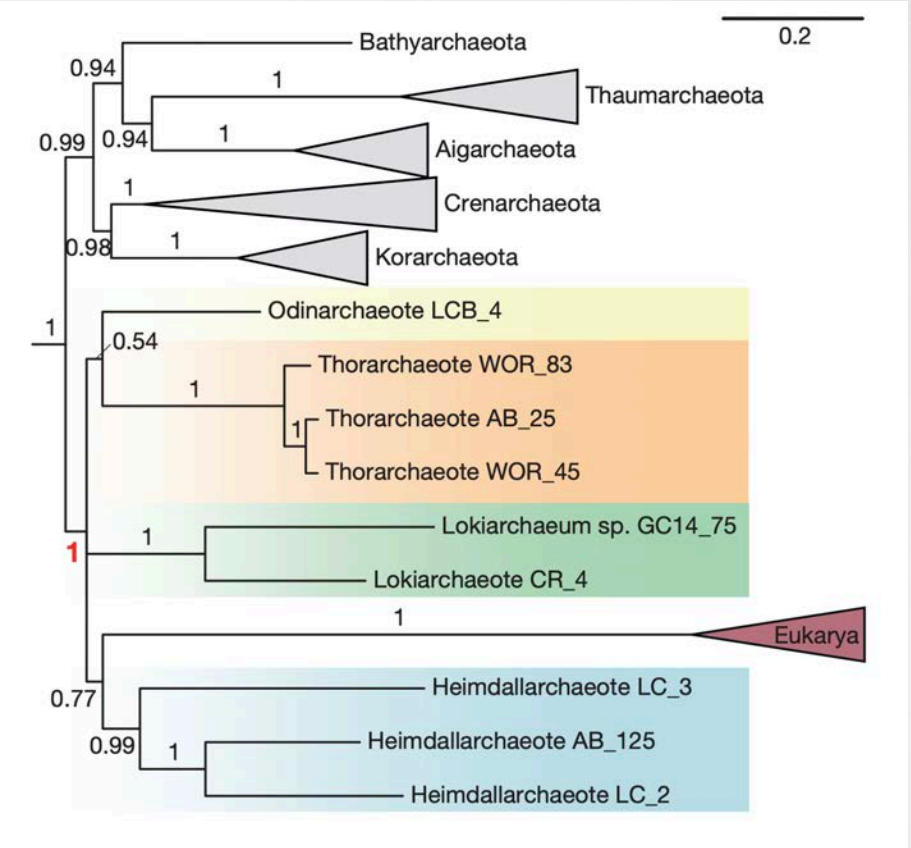
Archaea as **ancestors** of eukaryotes



2017

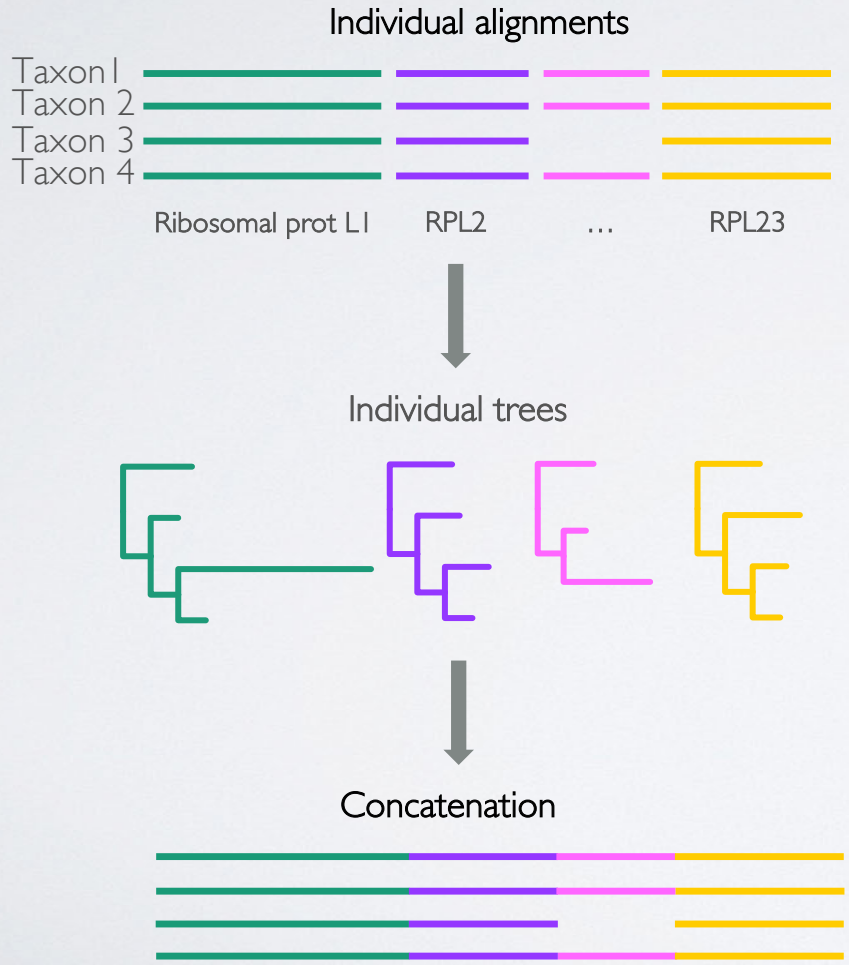
Culture-independent genomics (e.g. metagenomics)

Eukaryotes:
within or sister to Asgard ?

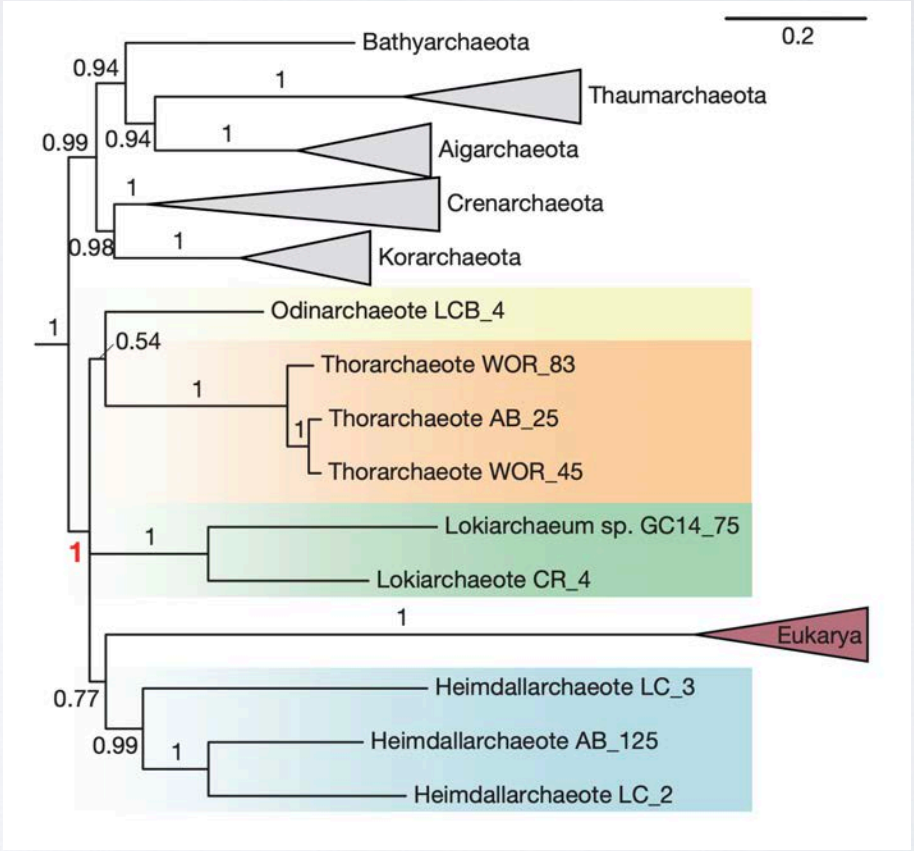


Zaremba-Niedzwiedzka *et al.*, Nature 2017

9 asgard genomes
56 ribosomal proteins



Eukaryotes:
within or sister to Asgard ?

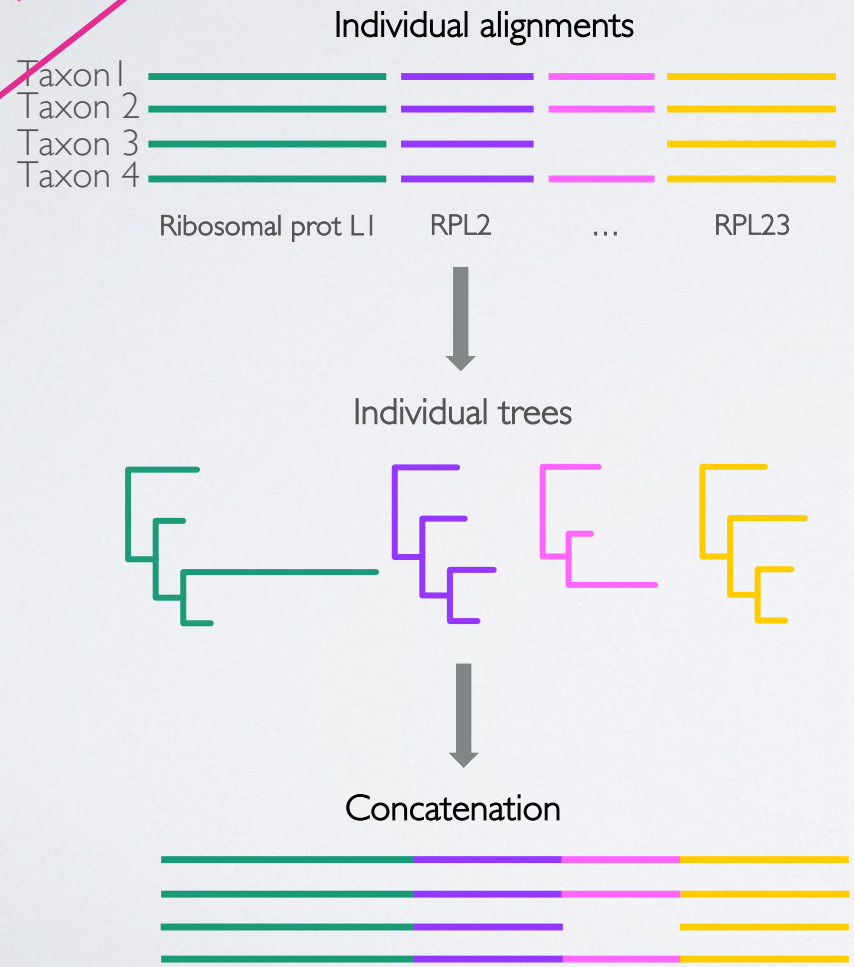


Zaremba-Niedzwiedzka *et al.*, Nature 2017

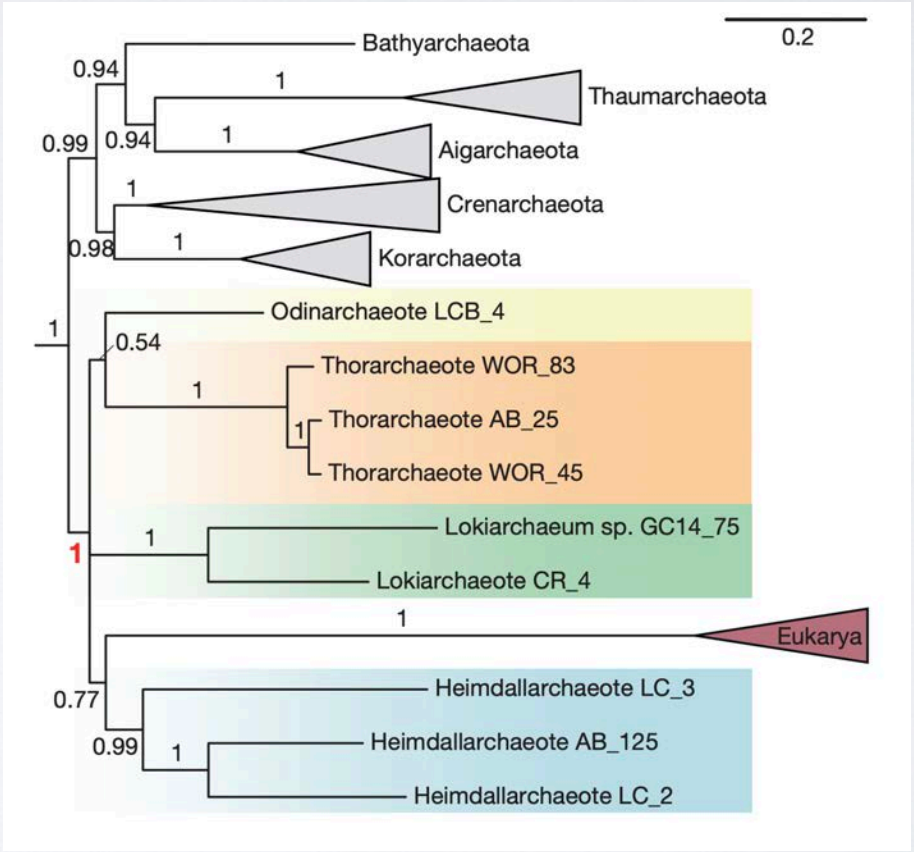
Increase taxon sampling

Increase concatenation size

9 asgard genomes
56 ribosomal proteins

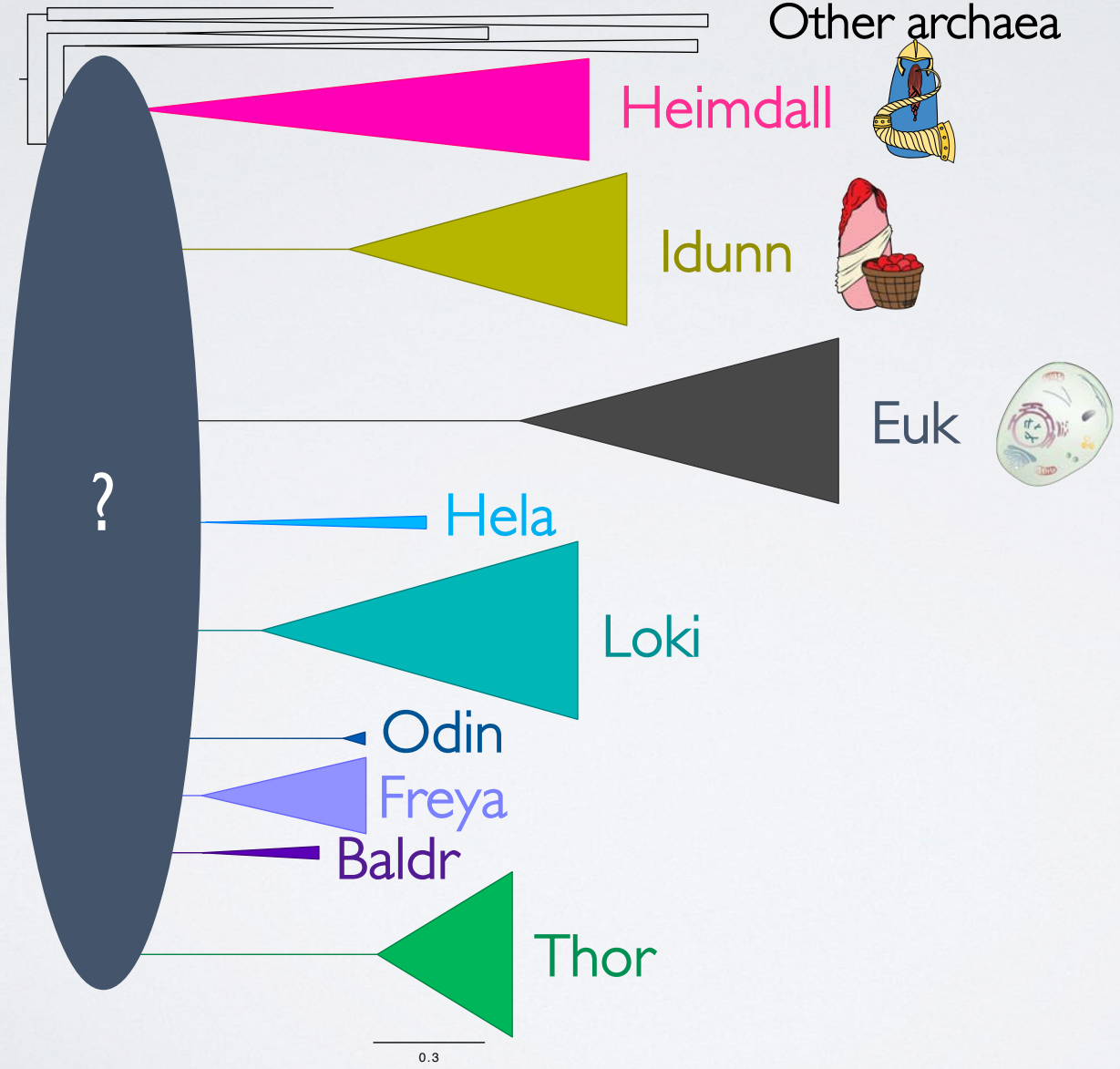


Eukaryotes: within or sister to Asgard ?



Zaremba-Niedzwiedzka *et al.*, Nature 2017

60 new Asgard genomes and many new major lineages



➔ Identifying new phylogenomic markers

Ribosomal proteins (RP) *vs* New markers (NM)

Ribosomal proteins:

Slow evolving
Universal
Short
Functional divergence

Ribosomal proteins (RP) *vs* New markers (NM)

Ribosomal proteins:

Slow evolving
Universal
Short
Functional divergence

New markers:

200 archaeal markers (Petitjean *et al.*, MBE 2015)

>= 10/14 euks
Of archaeal origin in euks
Present in all Asgard phyla
Manual + automated check for HGT

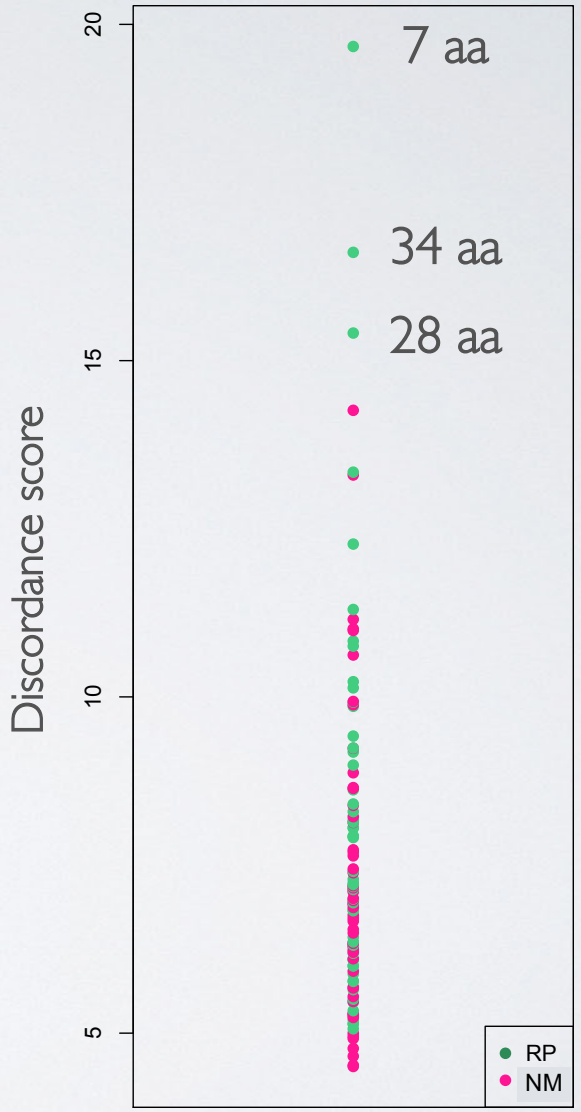
57 new markers

Testing for congruence: Discordance score

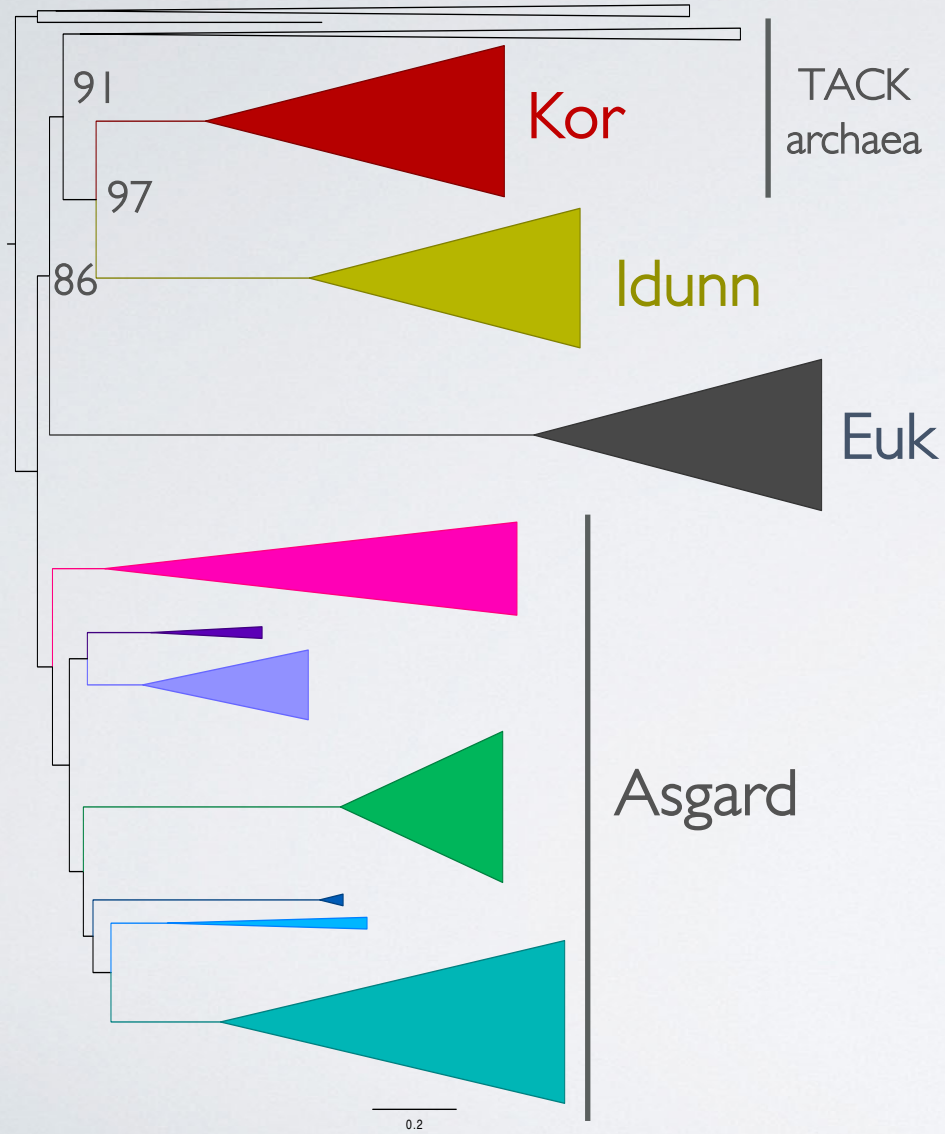
Discordance score:

~proportional to the frequency of incompatible (highly supported) bipartitions between a tree and all others.

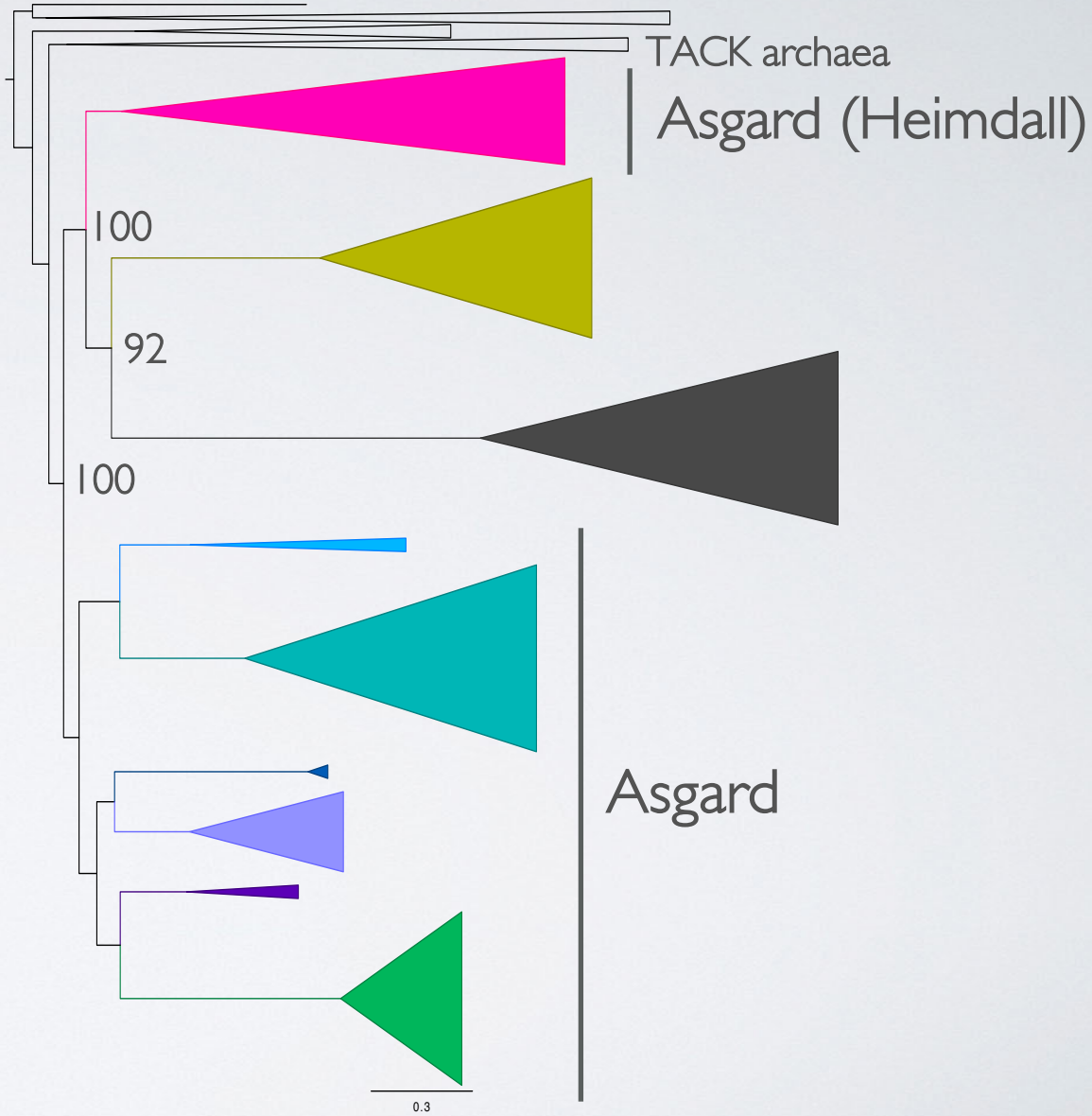
~All trees appear congruent
(or more so, "not incongruent")



56 ribosomal proteins (5647 aa), 195 taxa
IQ-tree C60+LG+F+G

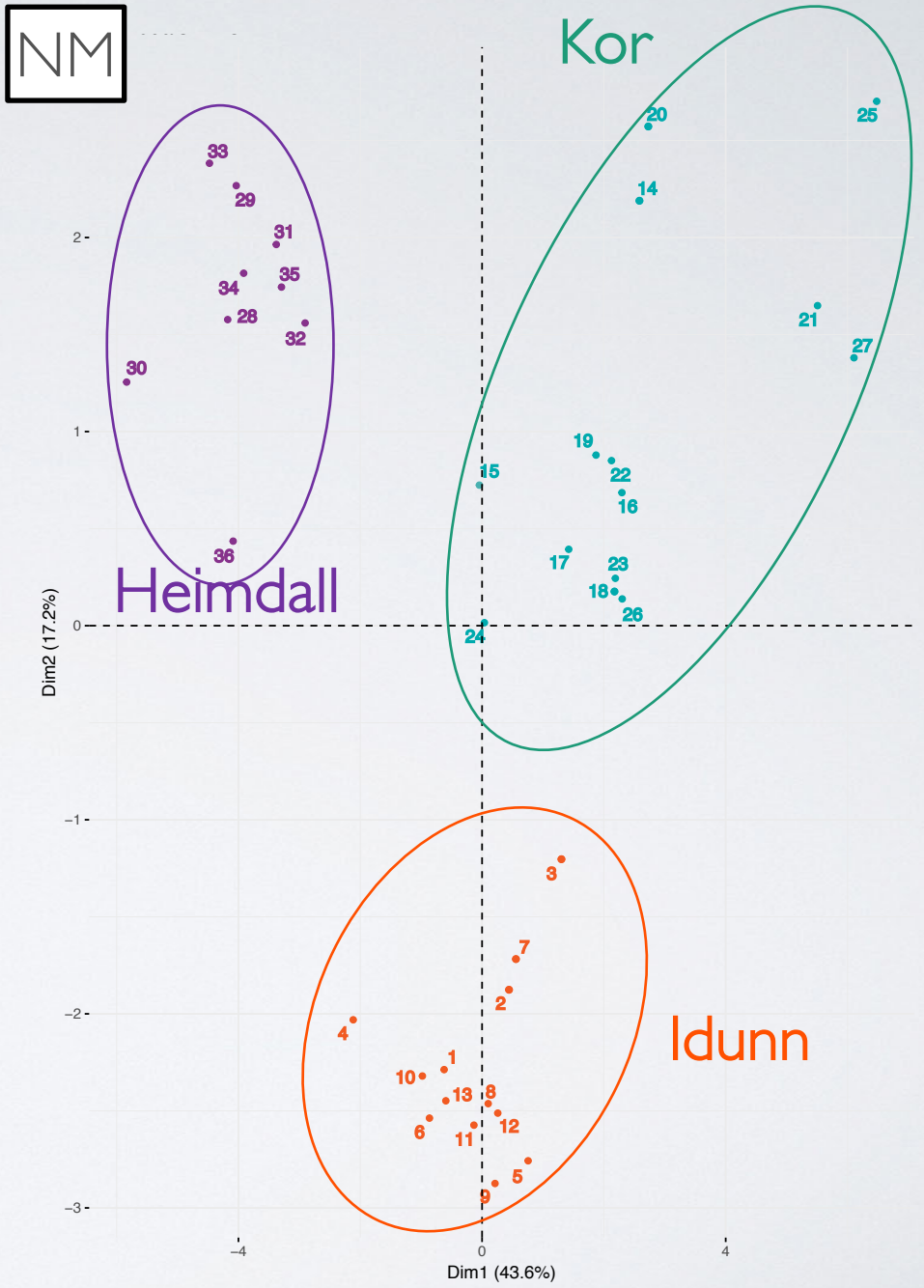


57 new markers (13485 aa), 195 taxa
IQ-tree C60+LG+F+G

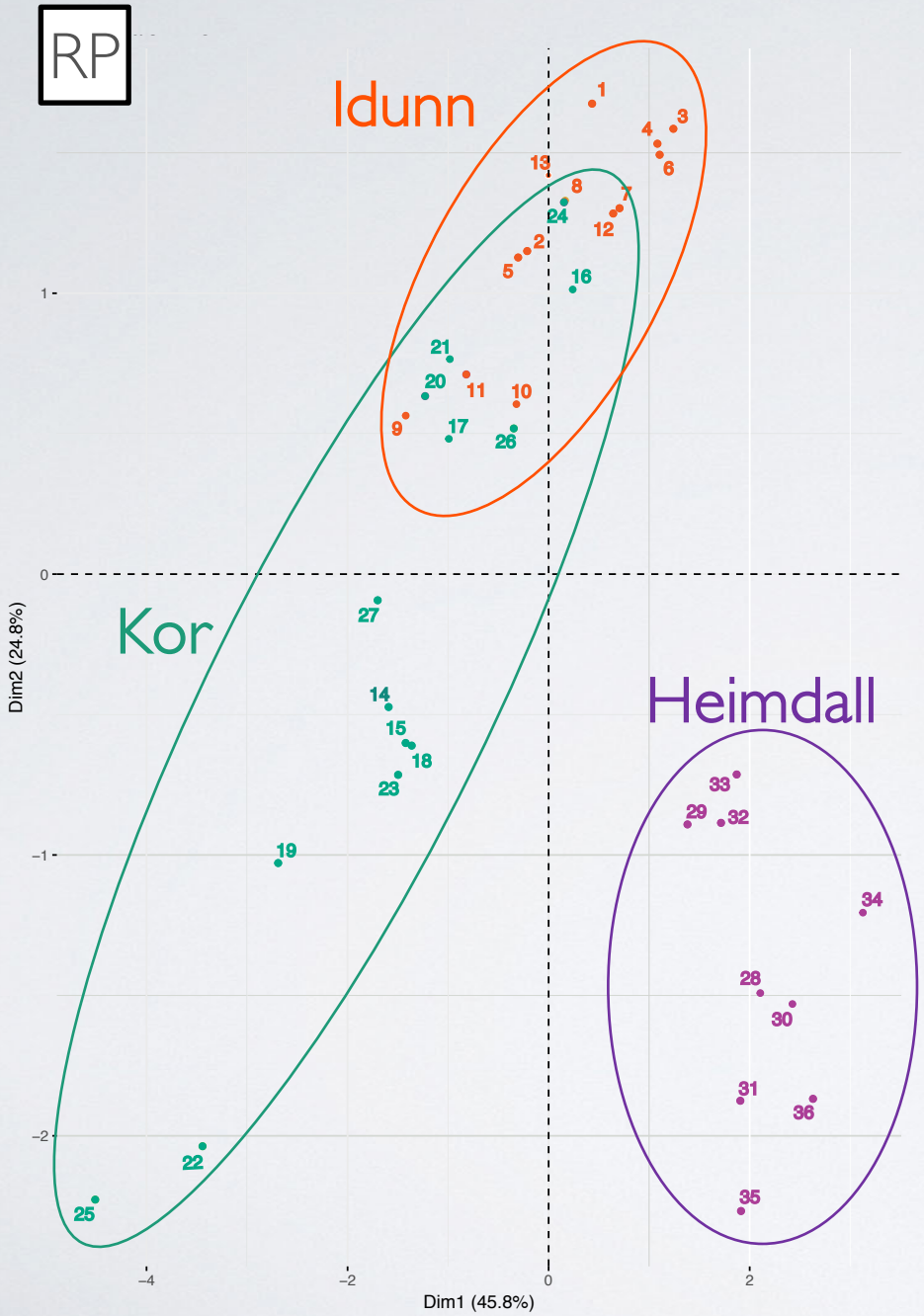


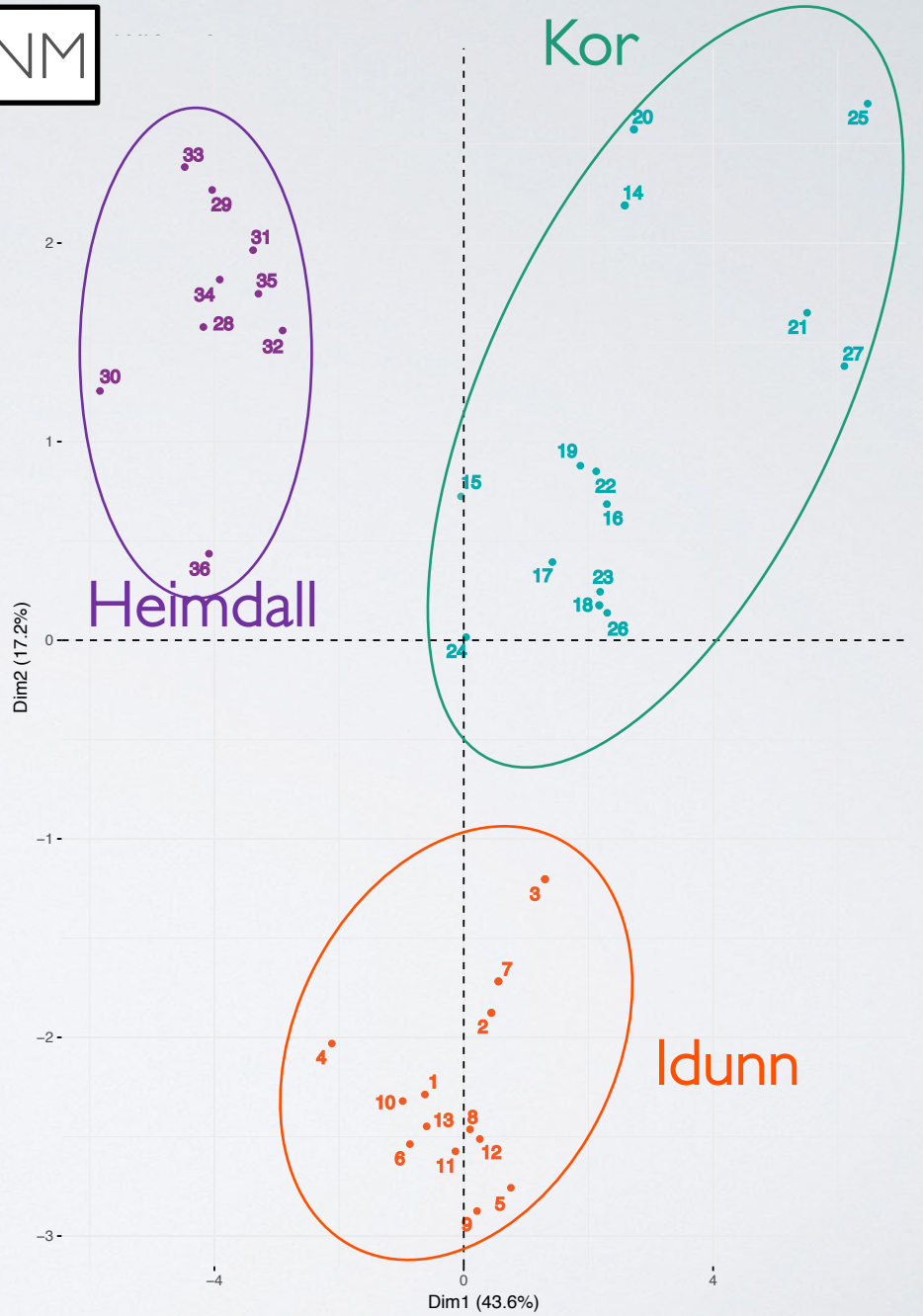
Investigating the discrepancy between RP and NM

PCA based on aminoacid composition

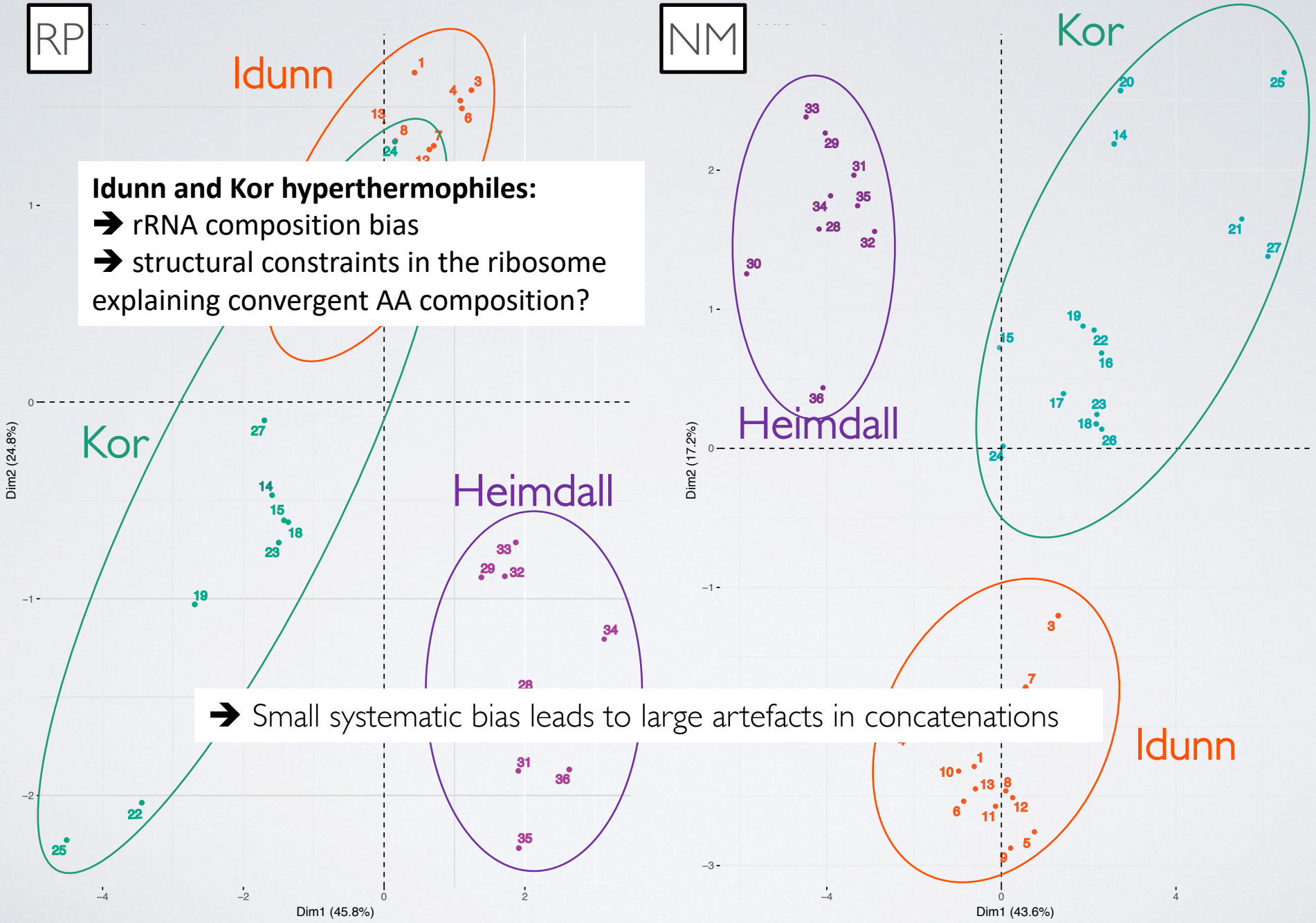


PCA based on aminoacid composition





PCA based on aminoacid composition



Idunn and Kor hyperthermophiles:

➔ rRNA composition bias

➔ structural constraints in the ribosome explaining convergent AA composition?

➔ Small systematic bias leads to large artefacts in concatenations

Recoding

Recoding AA

- Recoding with the same state AAs which often substitutes for one another. Eg: SR4 recoding scheme: AGNPST CHWY DEKQR FILMV

- Accommodates compositional biases and saturation
- Allows the use of more complex models because it reduces computational intensity

But...

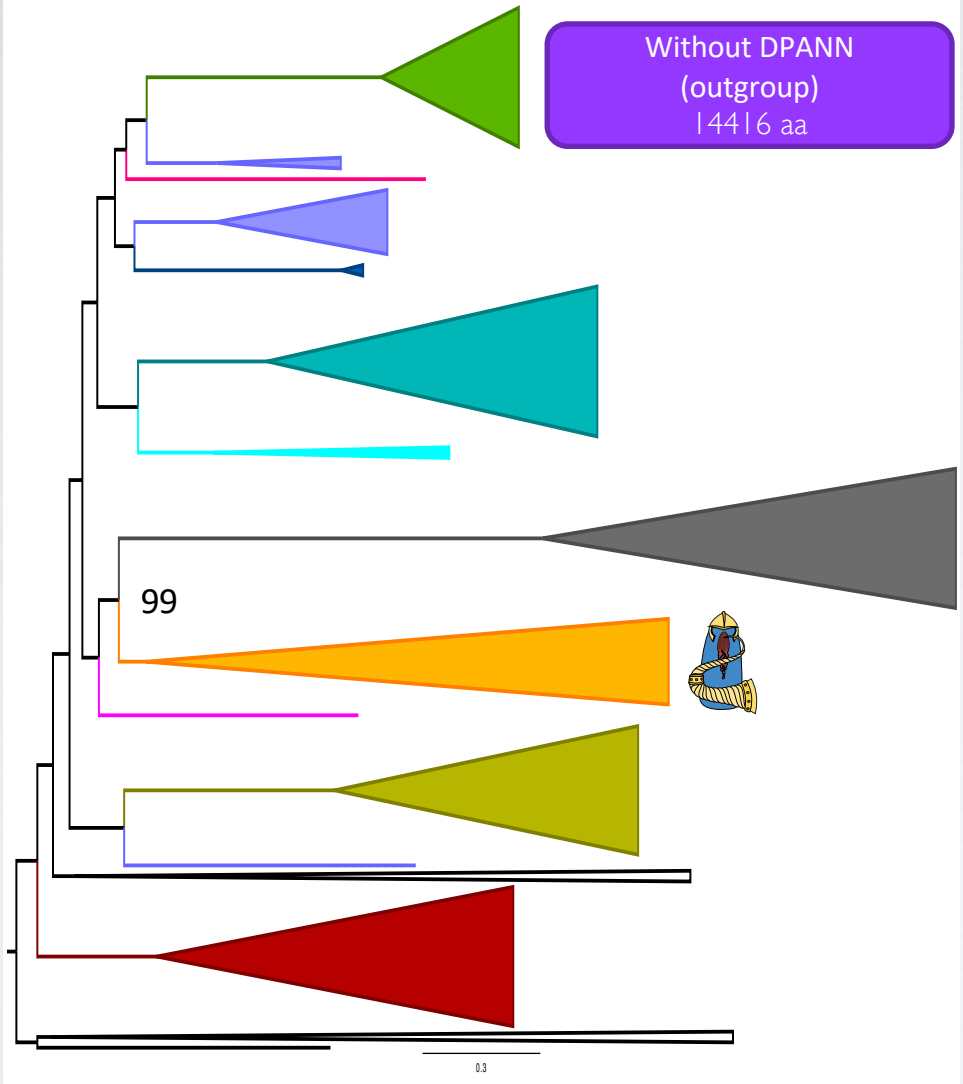
- Reduces the amount of phylogenetic information

Other elements that impact your tree reconstruction

Alignment and trimming

The effect of realigning

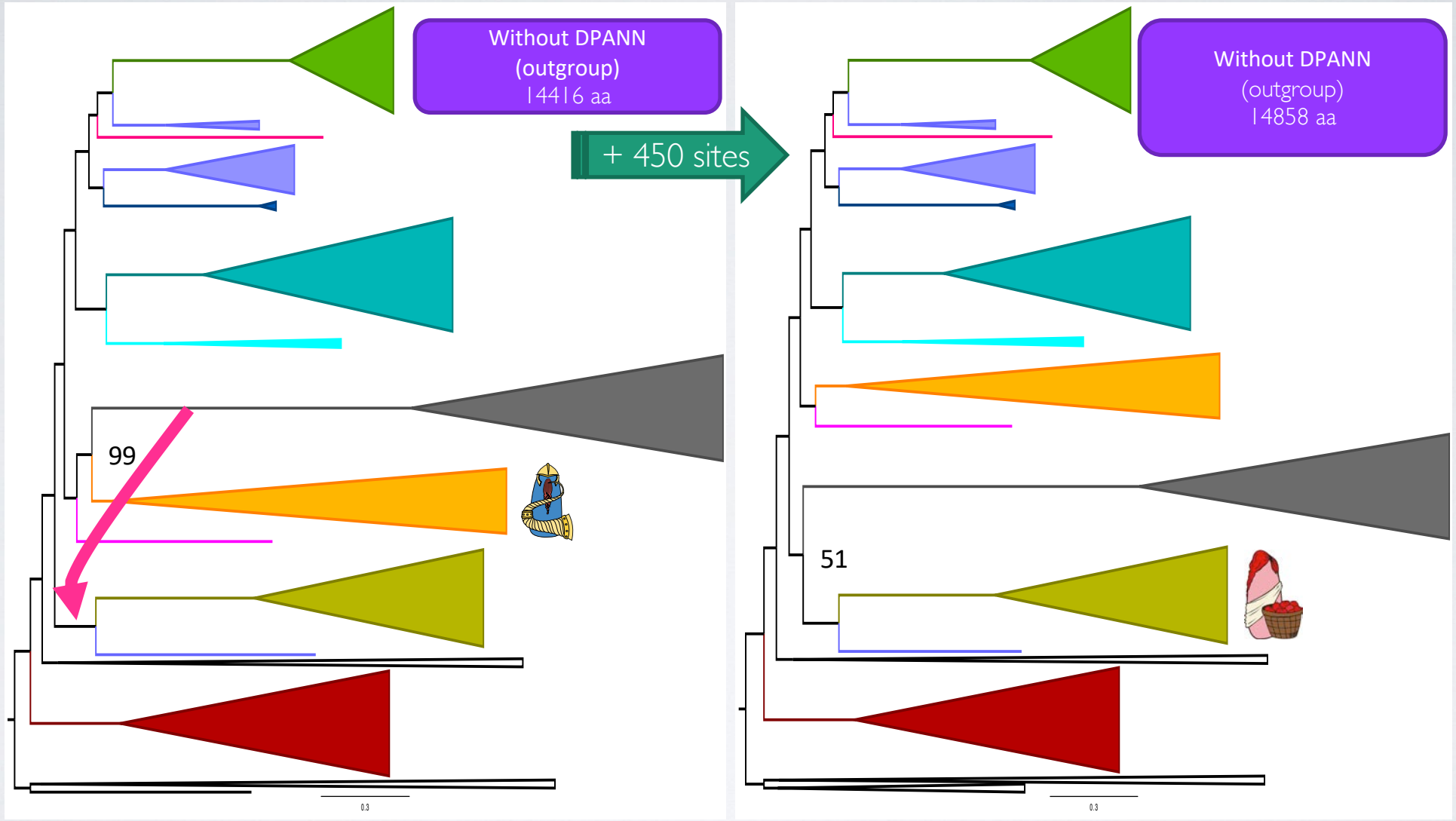
Might be obvious but…
At this evolutionary scale, we need to realign for each taxon sampling, even for "well conserved" markers
➔ Better quality alignments
➔ More sites after trimming



In agreement with Zaremba 2017

IQ-TREE (LG+C60+F+G+PMSF)

Realigned + trimmed after removing outgroups



IQ-TREE (LG+C60+F+G+PMSF)

If 3% of sites make such a difference…

ALIGNMENT AND TRIMMING STRATEGIES

MAFFT-LINSI

T-COFFEE

PROBCONS

Forward

MAFFT-LINSI

T-COFFEE

PROBCONS

Reverse

ALIGNMENT AND TRIMMING STRATEGIES

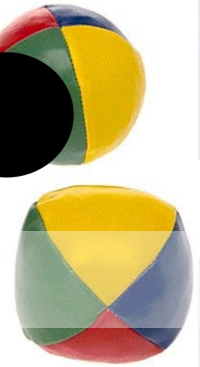
MAFFT-LINSI

T-COFFEE

PROBCONS

Forward

MAFFT-LINSI



T-COFFEE

PROBCONS

Reverse

ALIGNMENT AND TRIMMING STRATEGIES

I decided not to decide…

ALIGNMENT AND TRIMMING STRATEGIES

I decided not to decide…

- Trimal consensus mode (-ct 0.95)

➔ Selects for ‘consensually aligned sites’ across the 6 alignments

- Allows to be less stringent about gaps and block sizes

➔ Important a this evolutionary scale

parts of the backbone)

ALIGNMENT AND TRIMMING STRATEGIES

I decided not to decide…

- Trimal consensus mode (-ct 0.95)

➔ Selects for 'consensually aligned sites' across the 6 alignments

- Allows to be less stringent about gaps and block sizes

➔ Important at this evolutionary scale

➔ Allows to keep putative synapormophies (indels supporting specific parts of the backbone)

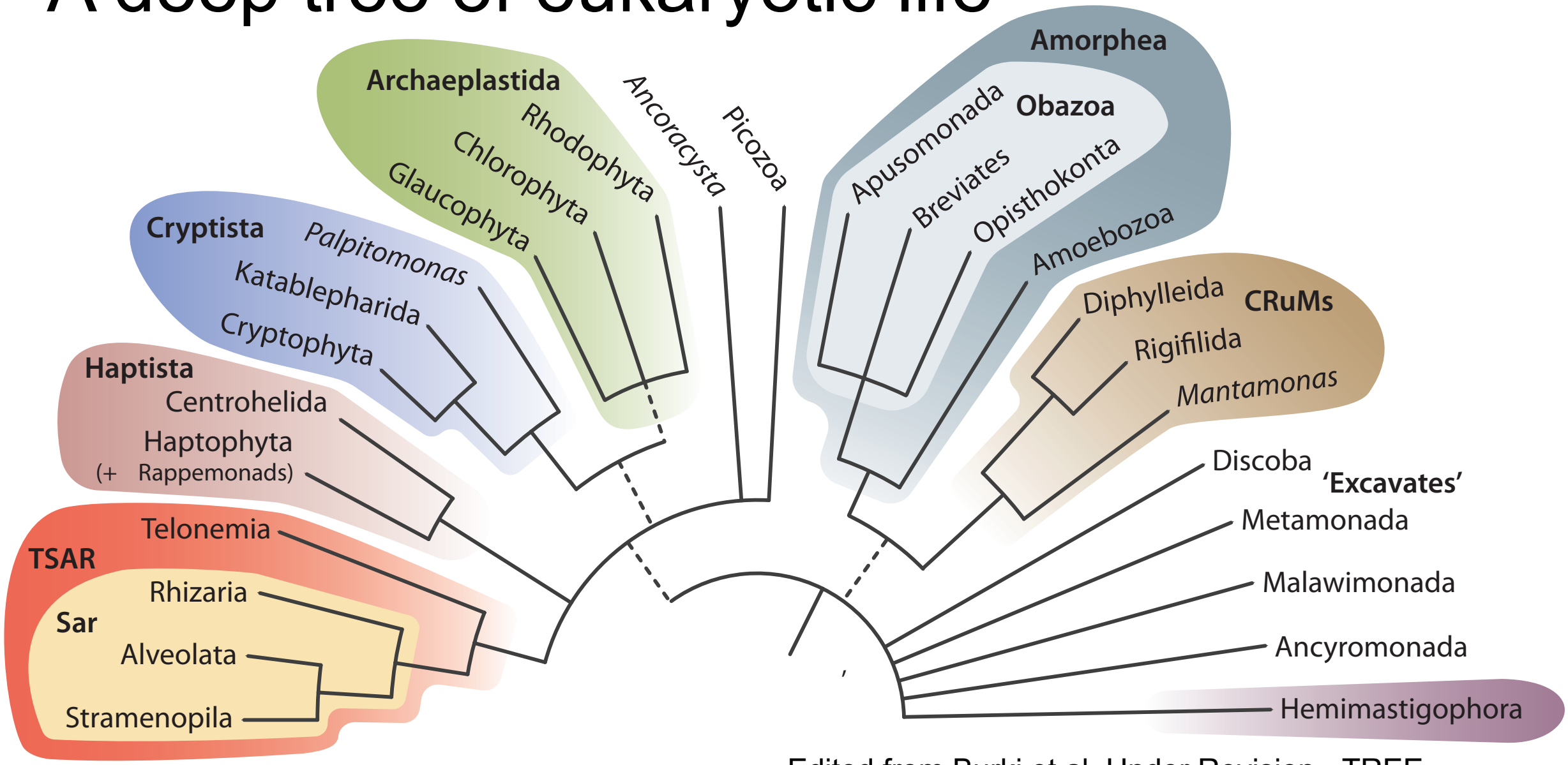
Useful tools

PHYLOFISHER

A phylogenetically aware pipeline for phylogenomic dataset construction

David Žihala, Alexander K. Tice, Tomáš Pánek, Serafim Nenarokov, Eric Salomaki, Andrew J. Roger, Martin Kolísko, Fabien Burki, Laura Eme, Marek Eliáš, Matthew W. Brown

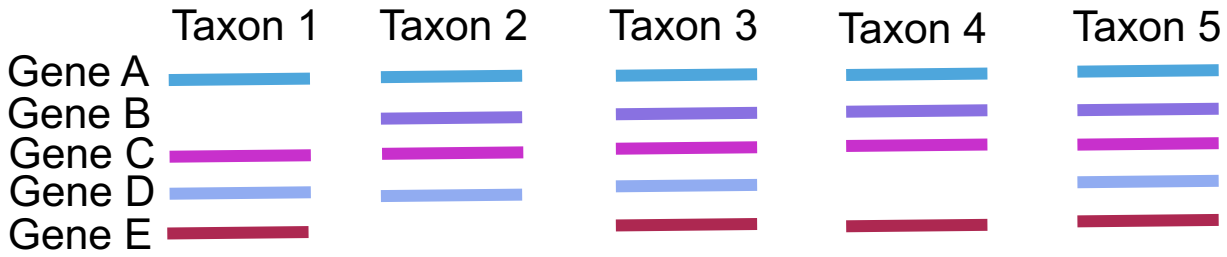
A deep tree of eukaryotic life



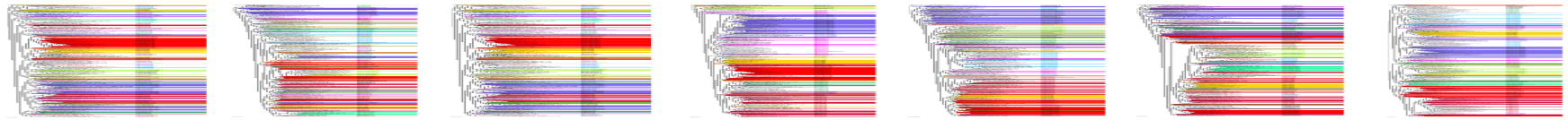
Edited from Burki et al. Under Revision - TREE

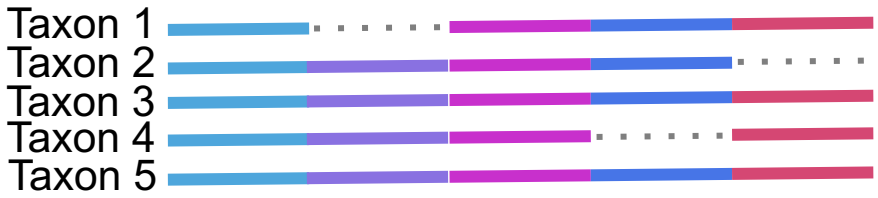
Multigene Phylogenetics – Phylogenomics

- Take multiple genes and infer a phylogeny
  - Genes A+B+C+D+E
- Offers more data
- Can handle missing data from taxa
- Concatenation
  - implies 1 taxon = 1 orthologous sequence

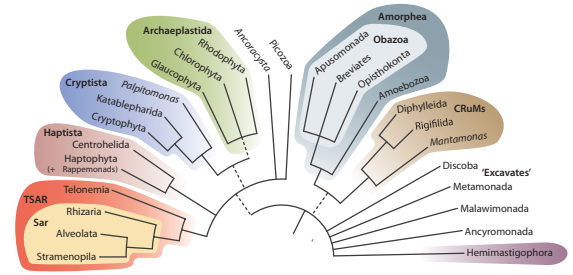


GENE SAMPLING





CONCATENTATION
(Gene Matrix)



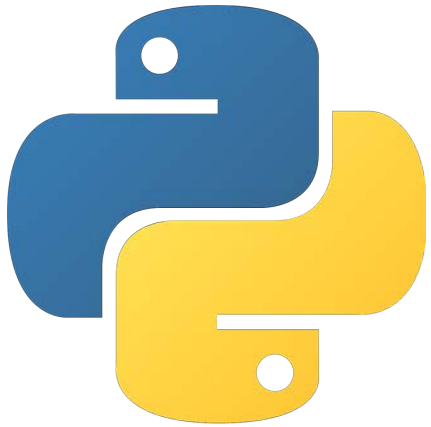
Selection of orthologs and orthologous sequences in them is critical

- Contamination
  - On sequencer
  - Endosymbionts
  - Prey (or predators)
- Paralogs
  - Genomic duplication
    - Deep- (i.e., a- vs b-tubulin)
    - Mid- (within a group)
    - In- (within a species (or genus))
- Phylogenetically informative
  - Broad taxonomic sampling
- To do this requires trees and careful consideration of them
  - Eyes

PHYLOFISHER

New method (and tool) allows for others to simply do phylogenomics

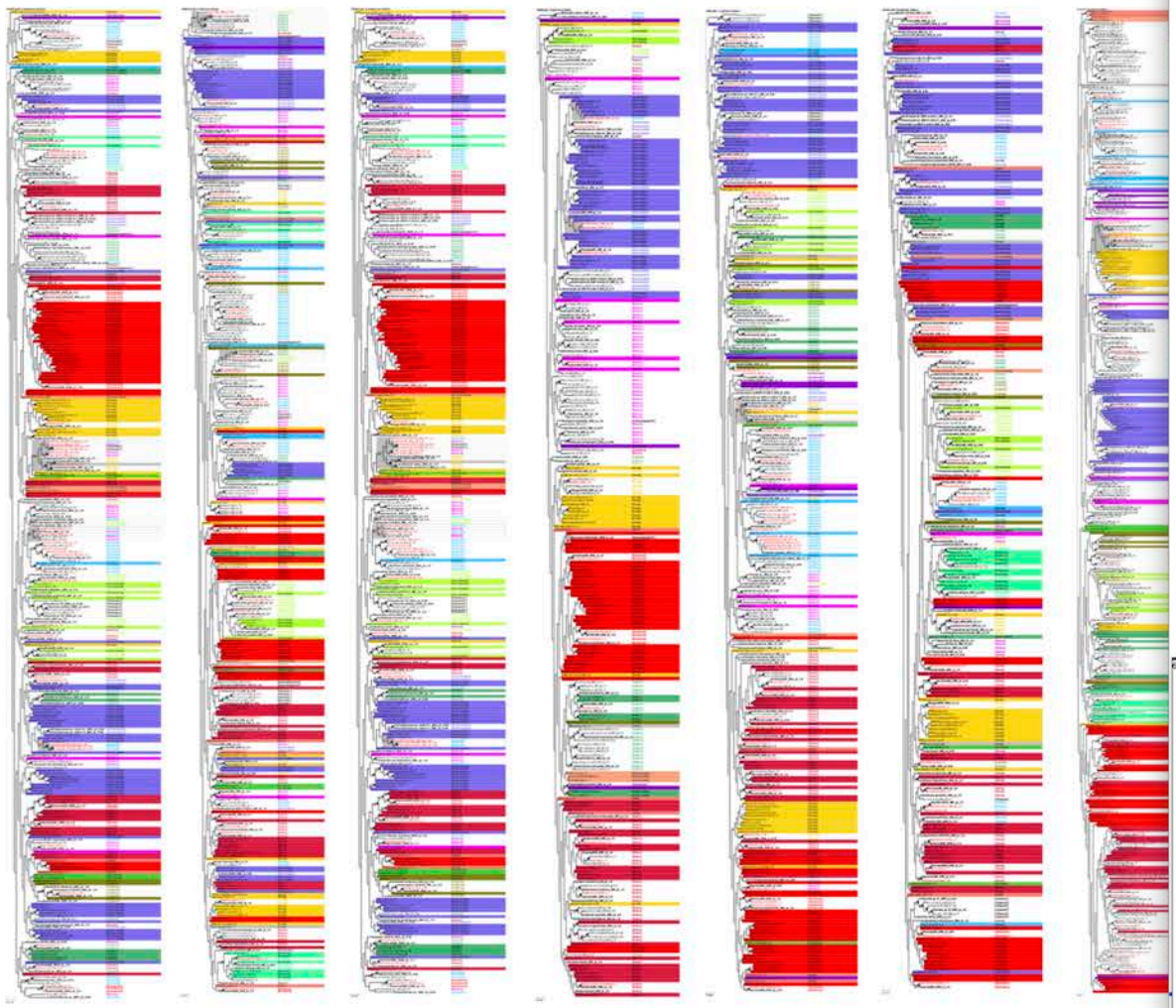
- Ships with a phylogenomic matrix and tool (via GitHub)
  - 310 Taxa, covering all deep eukaryotic groups
  - 240 Orthologs (whittled down from Brown et al. 2018)
- Coded in Python in a easy to to install CONDA environment
  - All dependencies are automatically installed



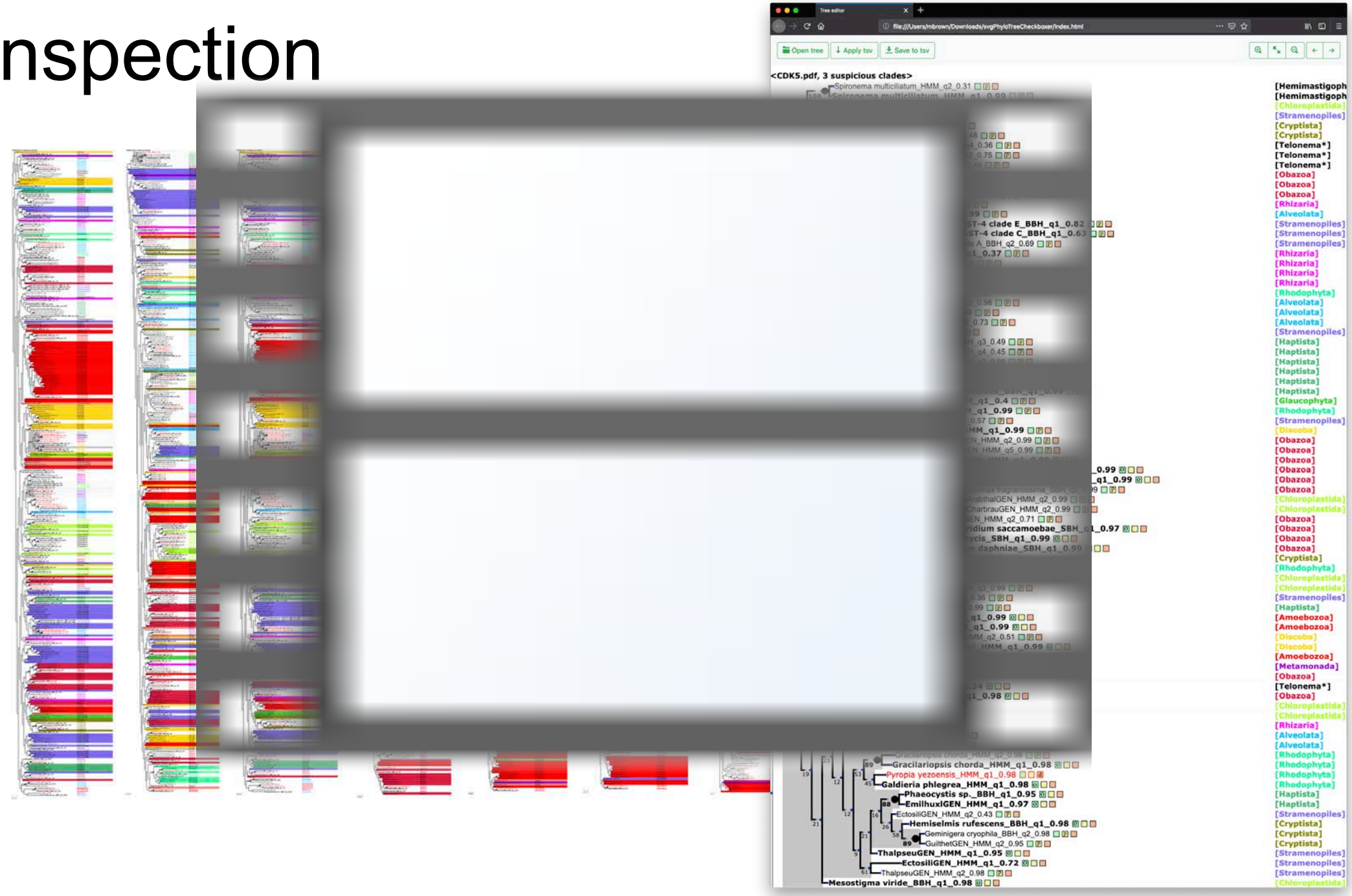
CONDA

```
conda env create -f fisher_env.yml
```

Tree Inspection

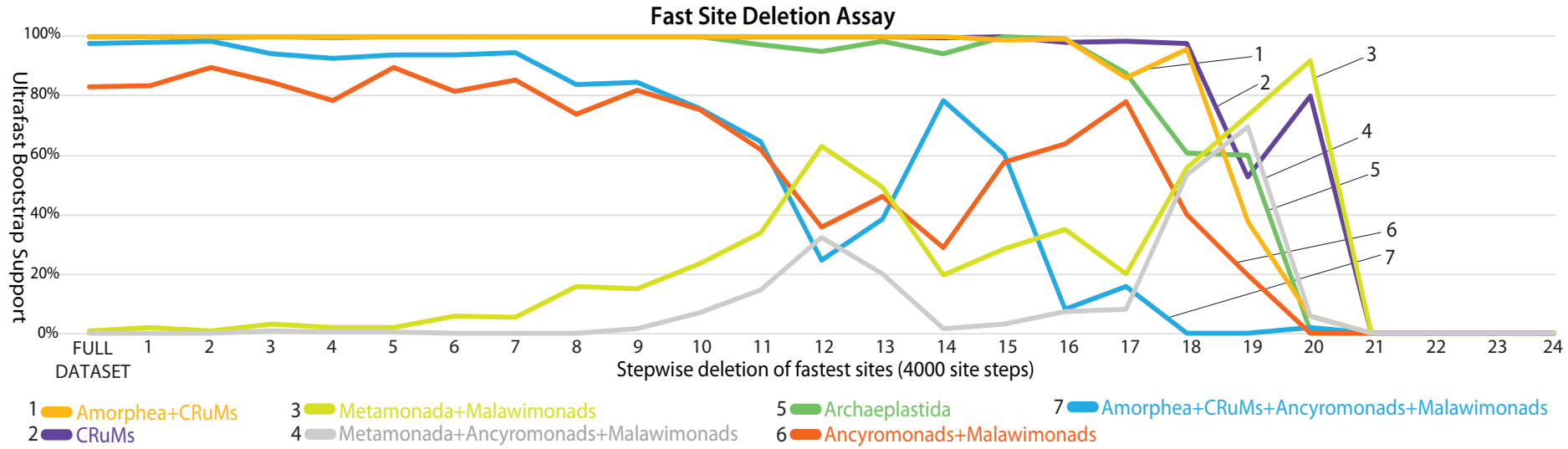


Tree Inspection



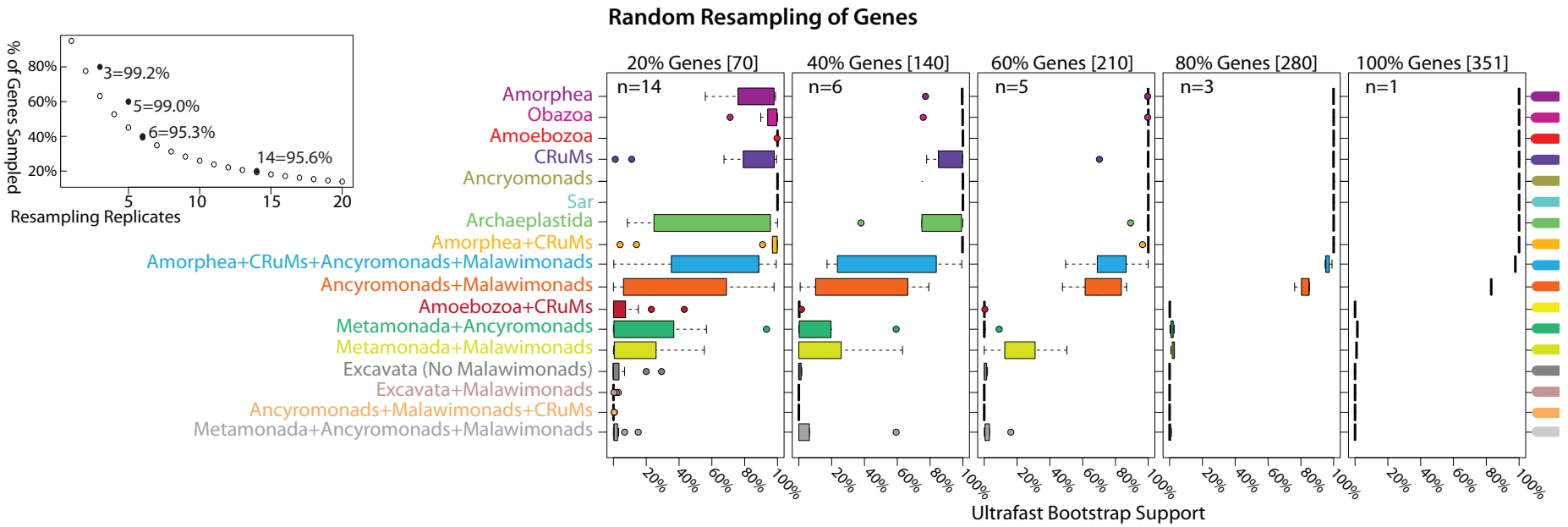
PHYLOFISHER

- Easily installed and simple usage
- Ships with our dataset
  - Includes Paralogs for tree building, more accurate identification
- Your own gene sets can be incorporated or used independently
- Tools for post-phylogenomic analyses



PHYLOFISHER

- Easily installed and simple usage
- Ships with our dataset
  - Includes Paralogs for tree building, more accurate identification
- Your own gene sets can be incorporated or used independently
- Tools for post-phylogenomic analyses



ETE3

ETE Toolkit

A Python framework to work with trees

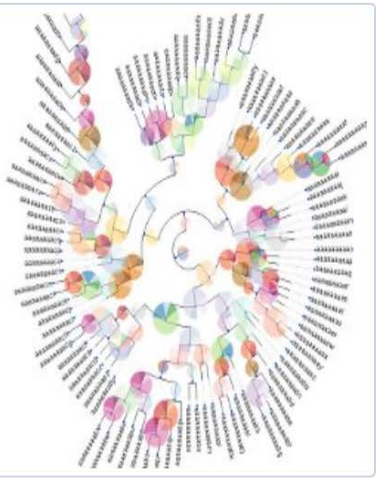
A Python framework for the analysis and visualization of trees.

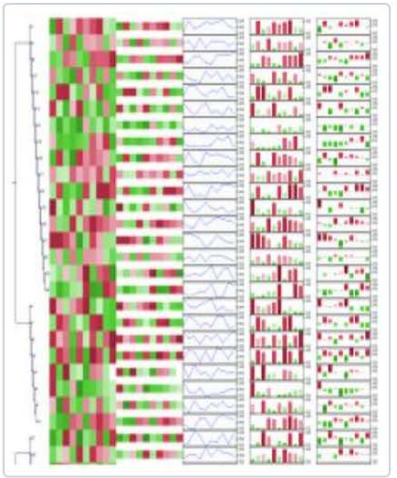
Download | Python API | Cookbook | Phylogenomic tools | Contribute

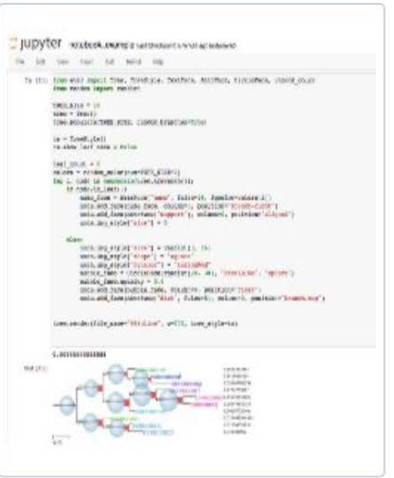
```
from ete3 import Tree
tree = Tree('((A,B), D);')

print tree
#        /-A
#    /-|
# --|    \-B
#    \-D

A = tree & "A"
A.up.show()
```







Trees as Python objects

Load, create, traverse, search, prune, or modify hierarchical tree structures with ease using the ETE Python API.

Programmatic tree visualization

Get full control of your tree images. Browse them interatively or render SVG, PNG of PDF images.

Tree annotation

Custom node attributes can be rendered as graphical elements. Choose among external images, charts, symbols, text labels, and

Jupyter notebook support

Prototype your methods using the Jupyter notebook framework including inline visualization of trees.

Signal in your data can be ‘real’ (non artefactual) and still not reflect the species tree

Gene trees may or may not = Species trees

Three main reasons:

(A) Deep coalescence of alleles (but usually ignored at this evolutionary scale)

(B) paralogy and orthology

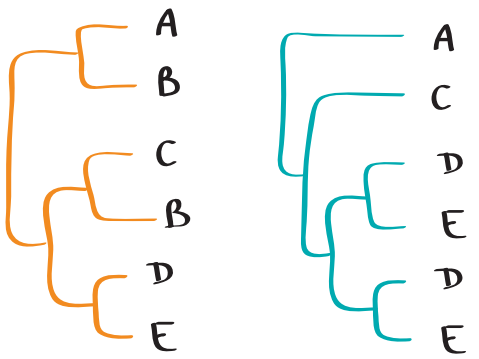
(C) lateral gene transfer (xenology)

Throw out data that is inconsistent

Try to model these events

Gene tree/Species tree reconciliation methods

Reconciliation approaches

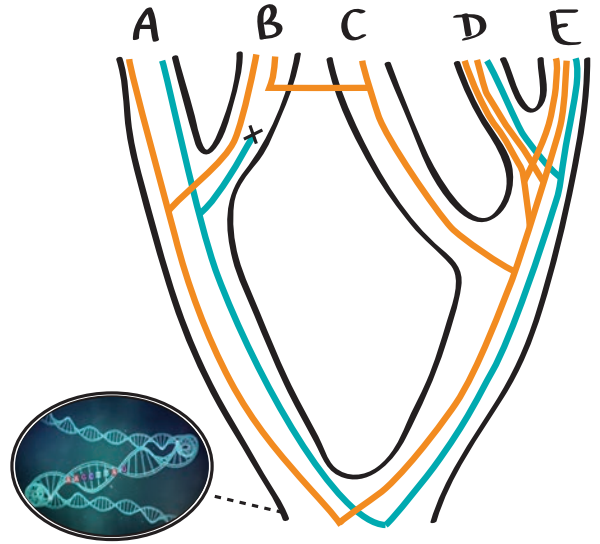


Thousands of gene trees

ALE
Phyldog
BUCKy
ecceTERA

Reconciliation

- ➔ Rooted species tree
- ➔ Losses and/or duplications and/pr transfers and/or coalescence
- ➔ Ancestral gene content
- ➔ Use HGT to date clades



Species tree

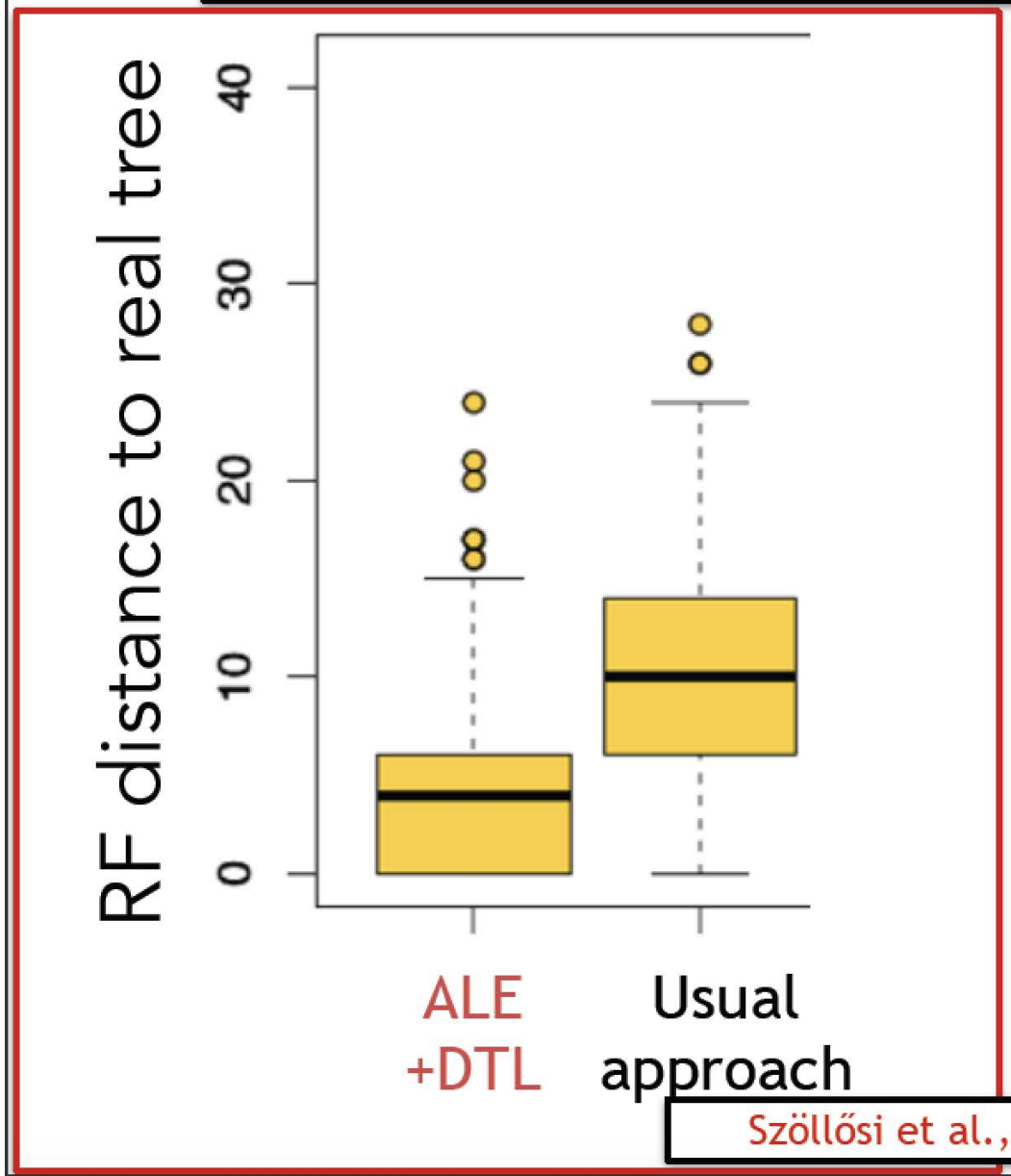
exODT: a model of gene duplication, transfer, and loss

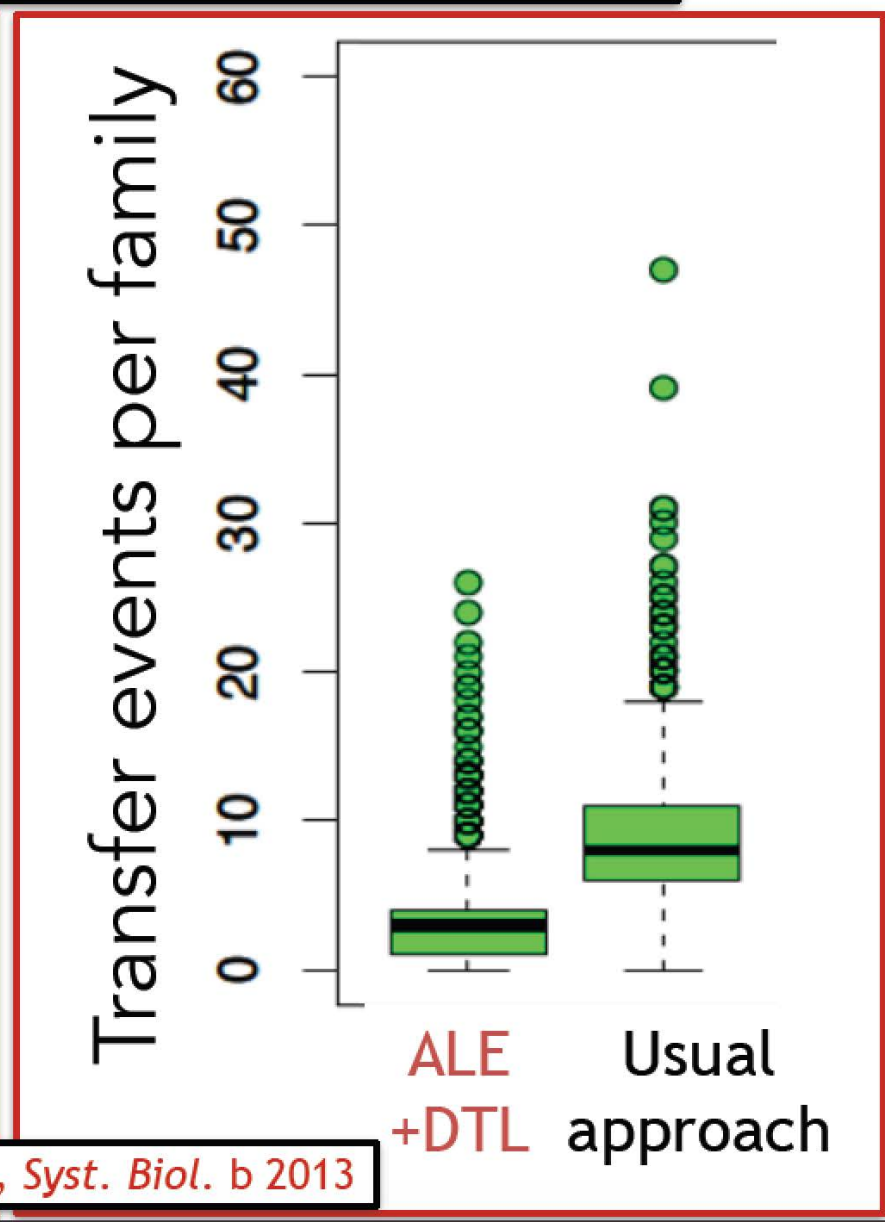
Assumptions

- Genes evolve along the species tree:
  - birth events:
    - duplications (rate of duplication)
    - transfers (rate of receiving a gene)
  - death events:
    - losses (rate of loss)
- Each gene family is independent of other genes
- Each gene copy is independent of other copies
- Transfers can go through unsampled/extinct species

Credit: B. Boussau

Better gene trees, fewer transfers

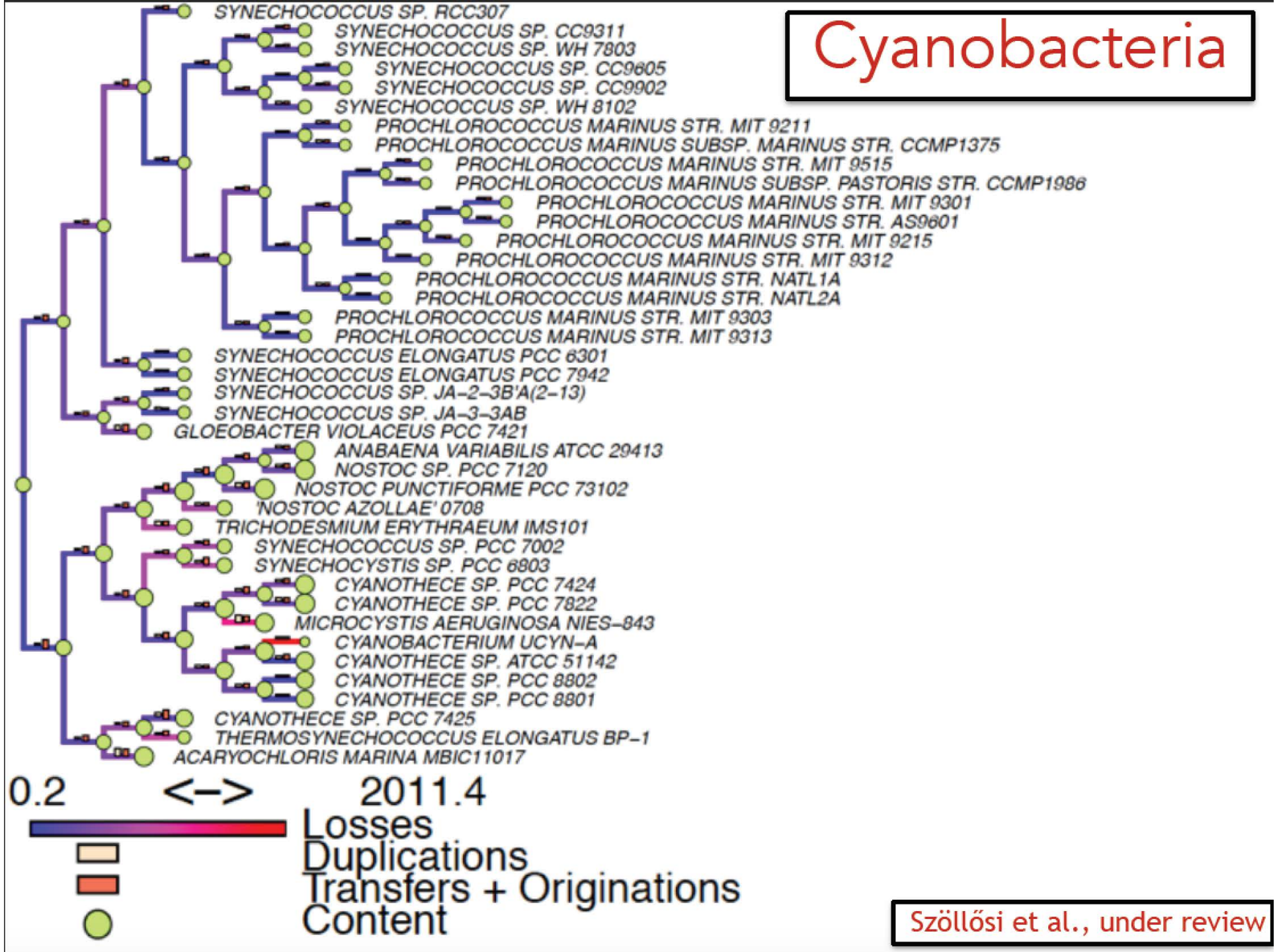




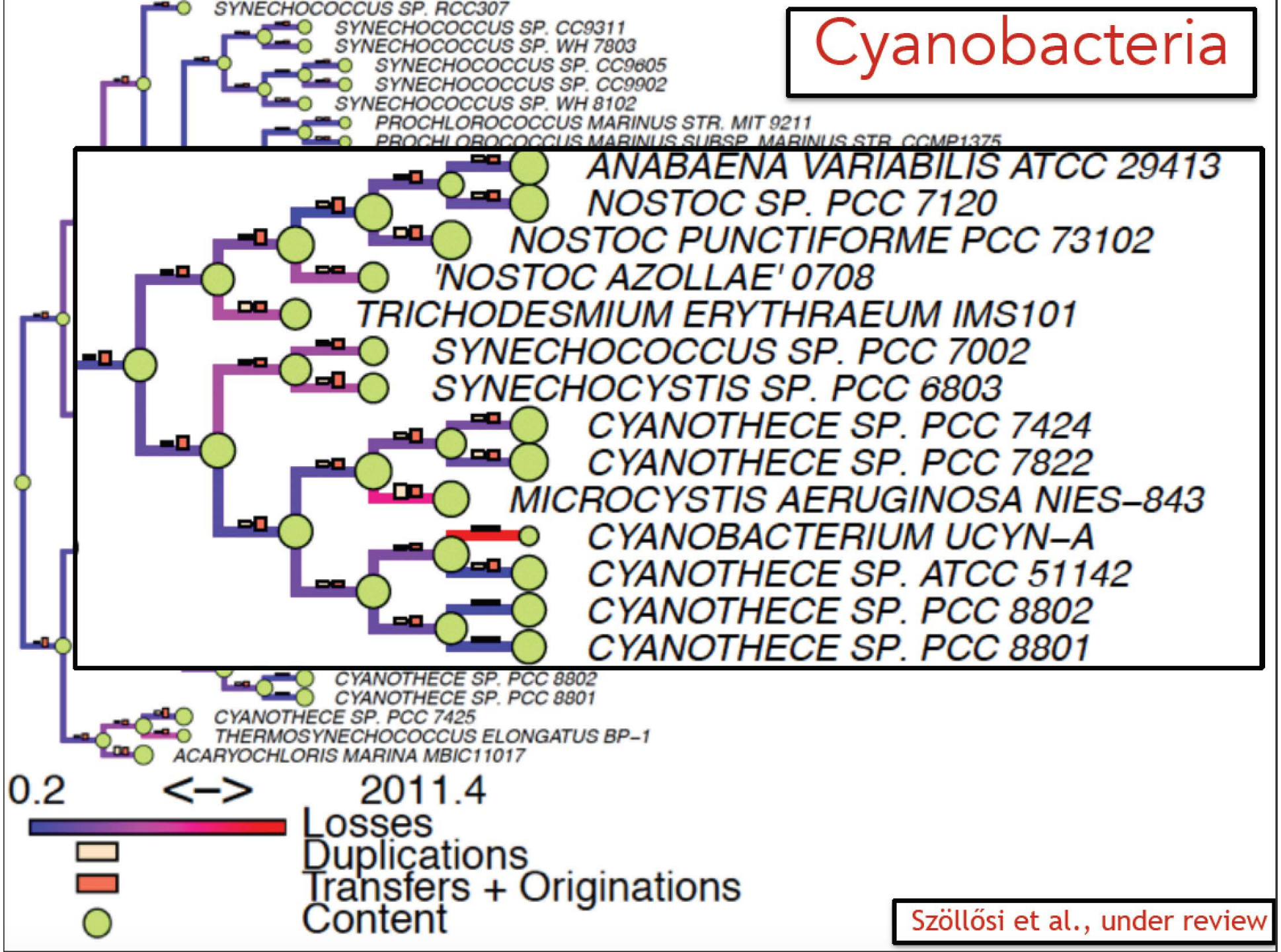
Szöllősi et al., *Syst. Biol.* b 2013

Credit: B. Boussau

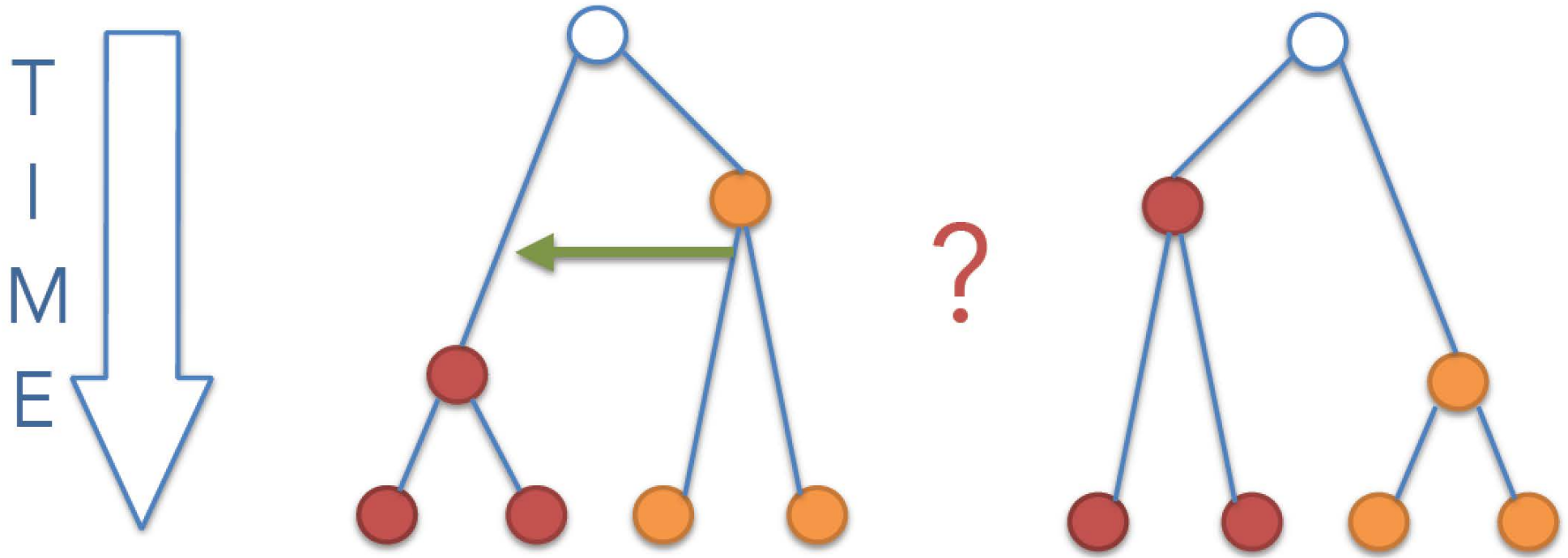
Cyanobacteria



Szöllősi et al., under review


Cyanobacteria
SYNECHOCOCCUS SP. RCC307
SYNECHOCOCCUS SP. CC9311
SYNECHOCOCCUS SP. WH 7803
SYNECHOCOCCUS SP. CC9605
SYNECHOCOCCUS SP. CC9902
SYNECHOCOCCUS SP. WH 8102
PROCHLOROCOCCUS MARINUS STR. MIT 9211
PROCHLOROCOCCUS MARINUS SUBSP. MARINUS STR. CCMP1375
ANABAENA VARIABILIS ATCC 29413
NOSTOC SP. PCC 7120
NOSTOC PUNCTIFORME PCC 73102
'NOSTOC AZOLLAE' 0708
TRICHODESMIUM ERYTHRAEUM IMS101
SYNECHOCOCCUS SP. PCC 7002
SYNECHOCYSTIS SP. PCC 6803
CYANOTHECE SP. PCC 7424
CYANOTHECE SP. PCC 7822
MICROCYSTIS AERUGINOSA NIES–843
CYANOBACTERIUM UCYN–A
CYANOTHECE SP. ATCC 51142
CYANOTHECE SP. PCC 8802
CYANOTHECE SP. PCC 8801
CYANOTHECE SP. PCC 8802
CYANOTHECE SP. PCC 8801
CYANOTHECE SP. PCC 7425
THERMOSYNECHOCOCCUS ELONGATUS BP–1
ACARYOCHLORIS MARINA MBIC11017
0.2
<–>
2011.4
Losses
Duplications
Transfers + Originations
Content
Szöllősi et al., under review

Using transfers to date clades



Credit: B. Boussau