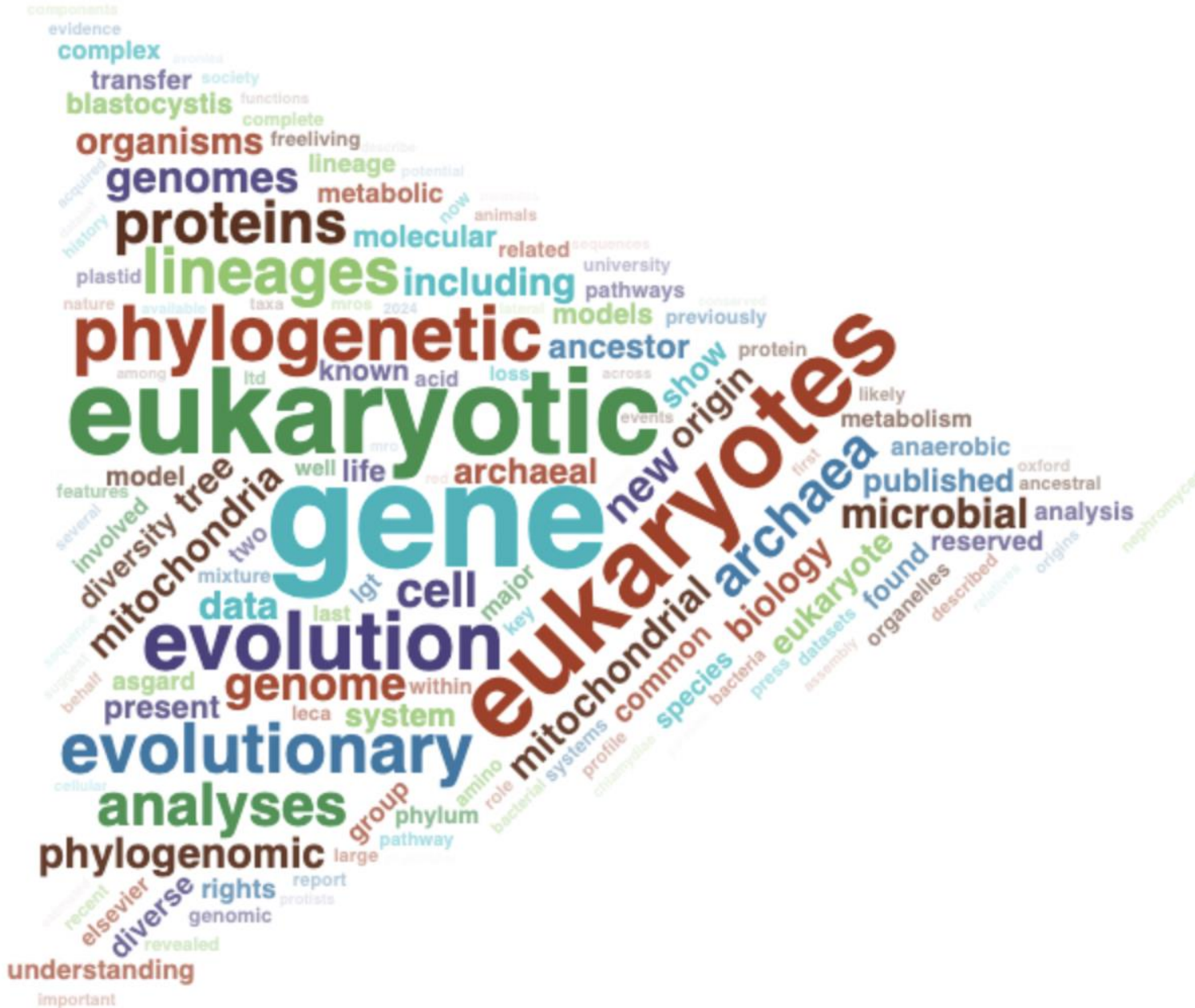
”Deep” phylogenetics

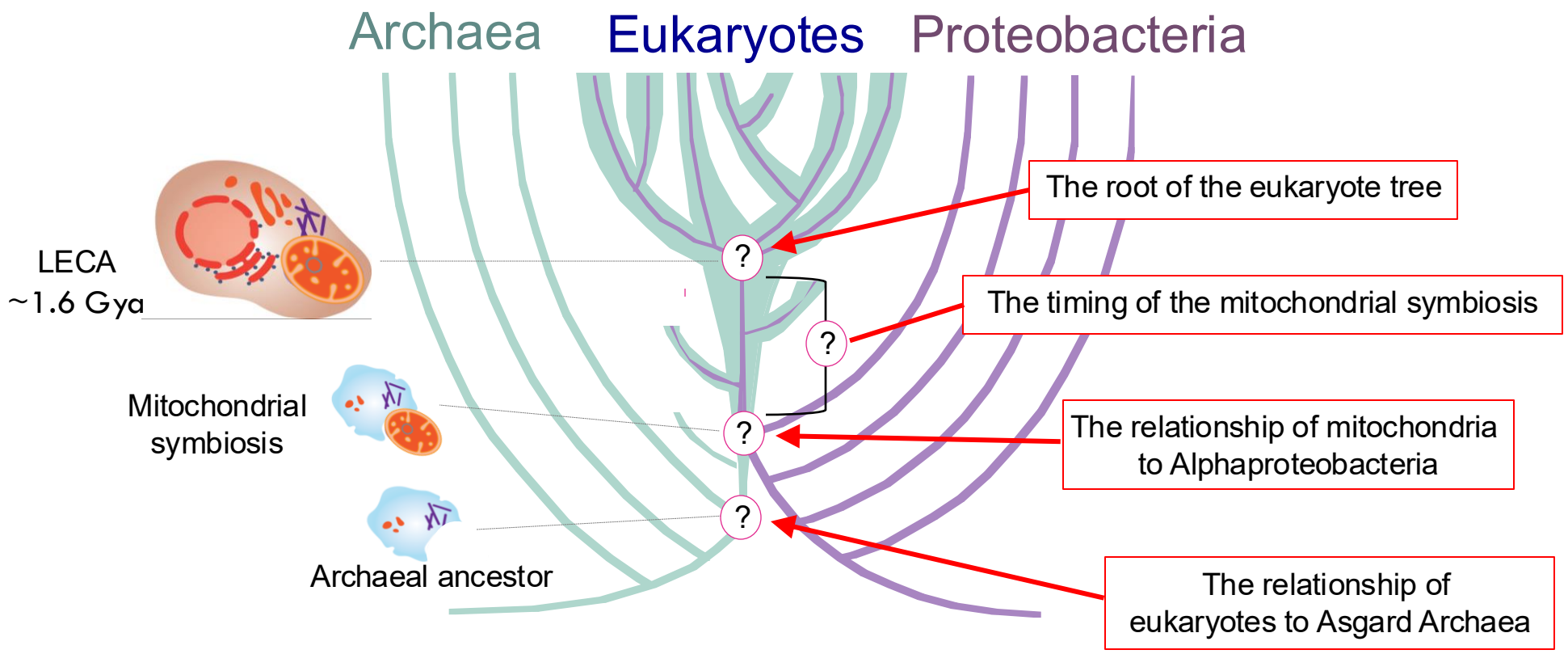
Laura Eme
Associate Professor
University of Rhode Island

@lauraeme.bsky.social



https://shiny.research.sfu.ca/

Four major phylogenetic problems of eukaryogenesis



Plagued by artefacts from use of overly simplistic phylogenetic models

1 - Protein models of evolution

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

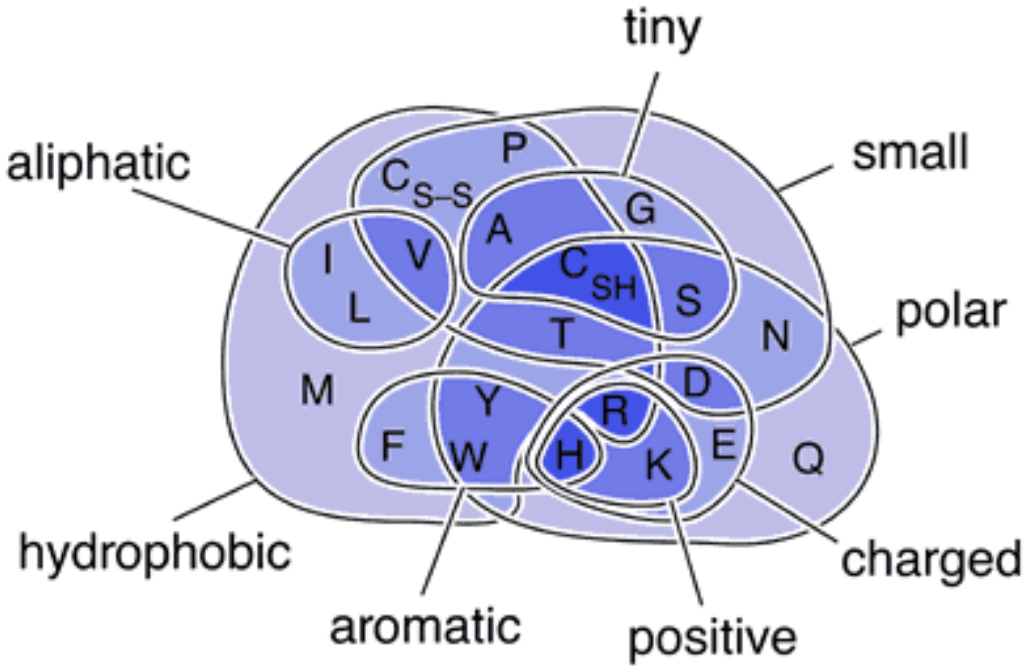
- A 20x20 rate matrix
- Assumes stationarity and reversibility

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

➔ 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 vs membrane proteins
    - Mitochondrial proteins, viral 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

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 (2001)

- 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 (2008)

- 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…)

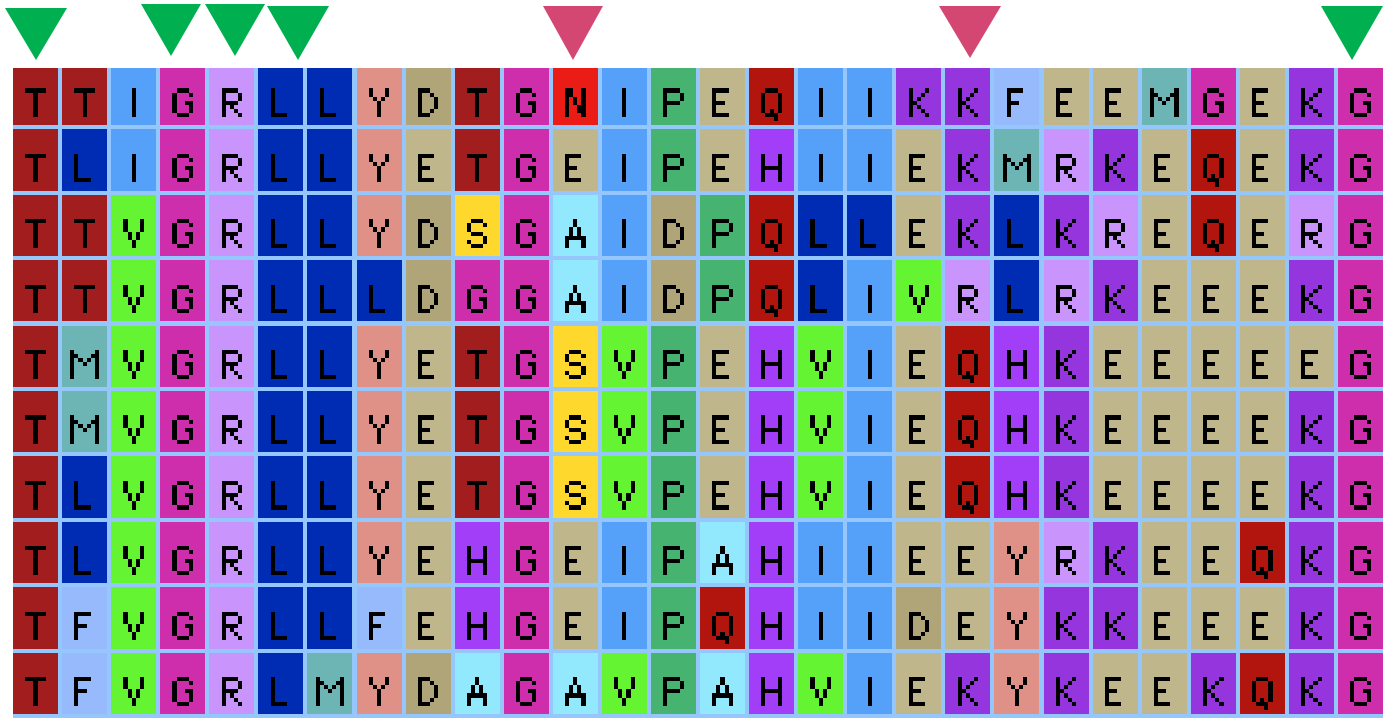
Summary

- 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

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

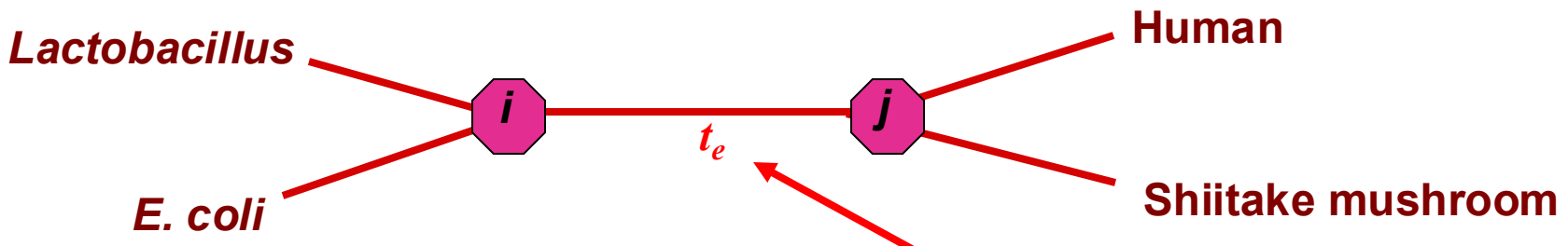
GTR (General time reversible)

- 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 information)

1.3 Mixture models

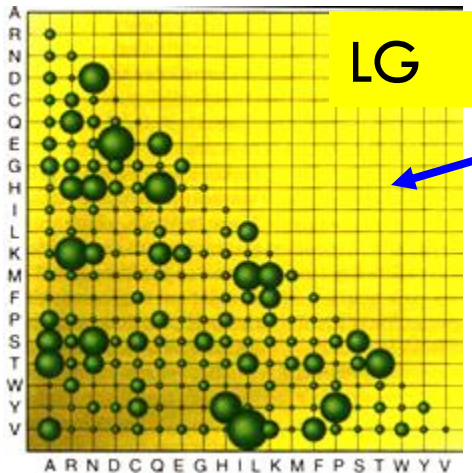
Your model is giving you the probability of going from amino acid *i* to *j* at site *x*, evolving at rate r_v on branch t_e



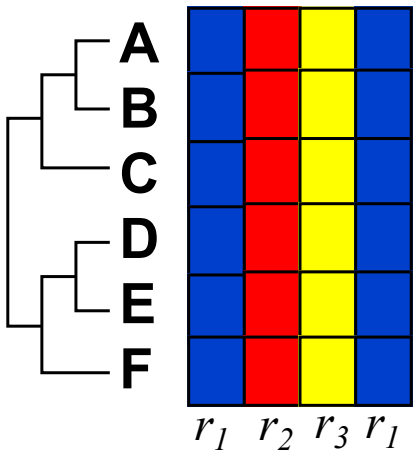
$$P(j \mid i;t) = \left[\exp(Q \cdot t_e \cdot r_v)\right]_{ij}$$

For $i \neq j$:

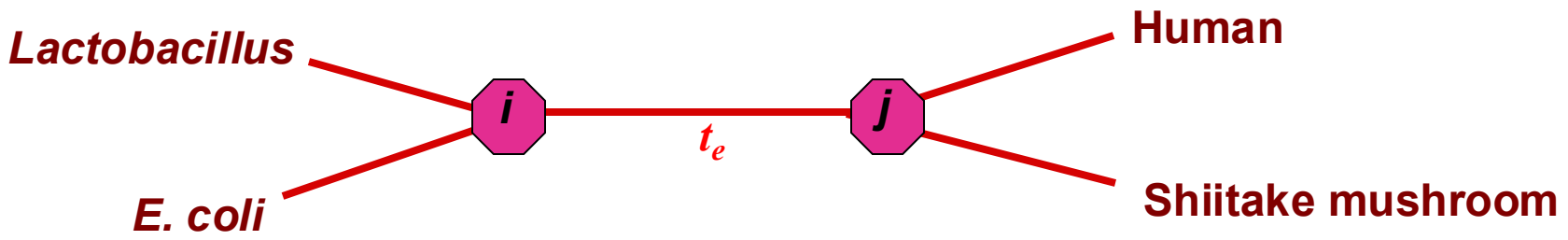
$$q_{ij} = r_{ij} \cdot \pi_j$$



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



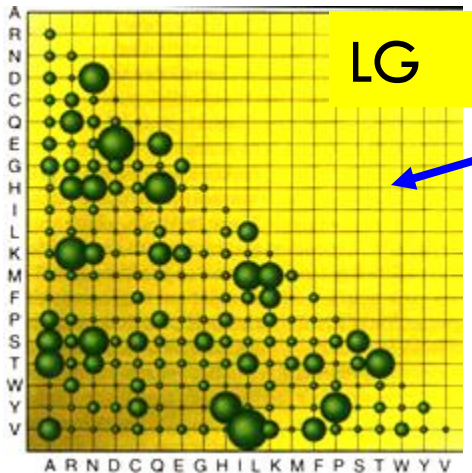
Your model is giving you the probability of going from amino acid *i* to *j* at site *x*, evolving at rate r_v on branch t_e



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

For $i \neq j$:

$$q_{ij} = r_{ij} \,\square\, \pi_j$$



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

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

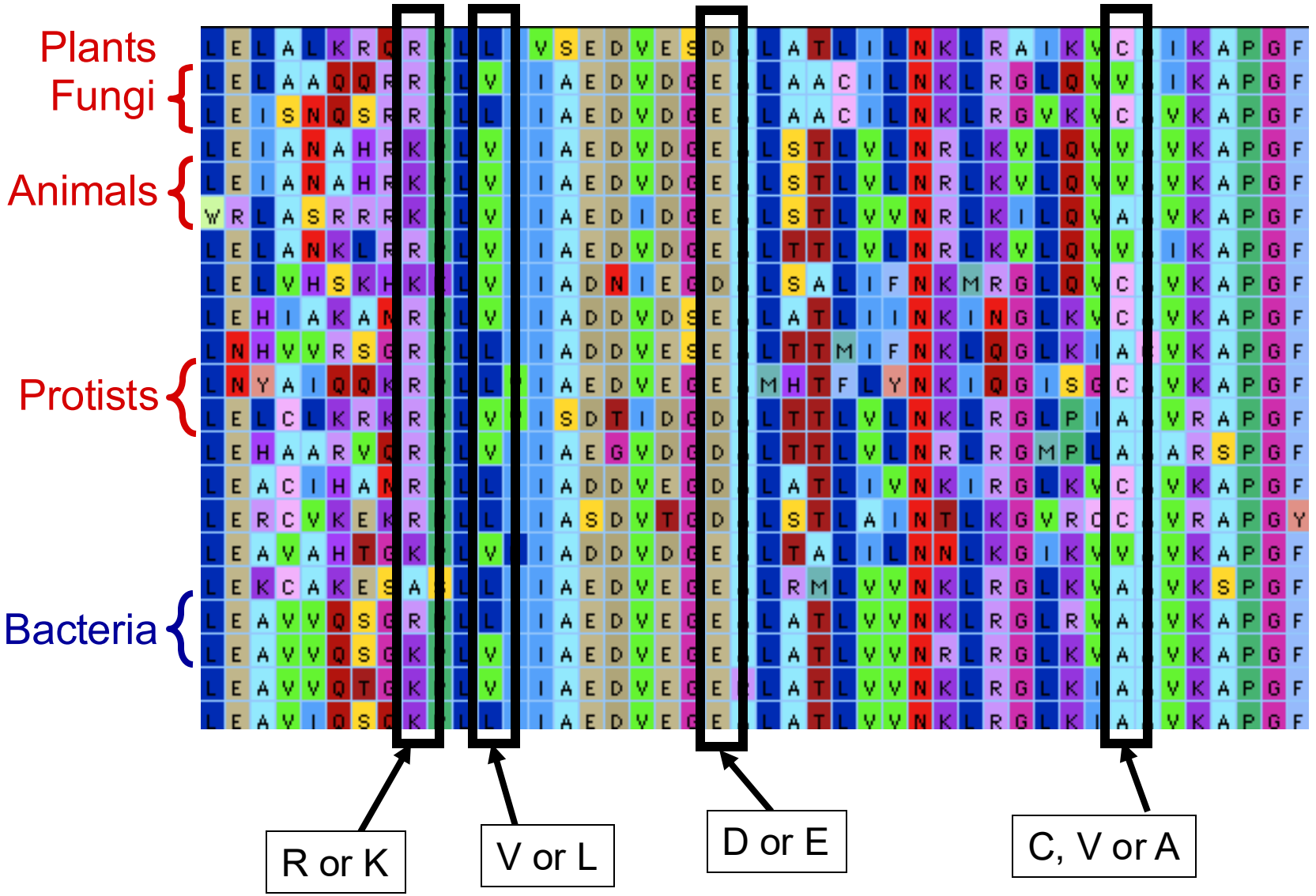
The problem…

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

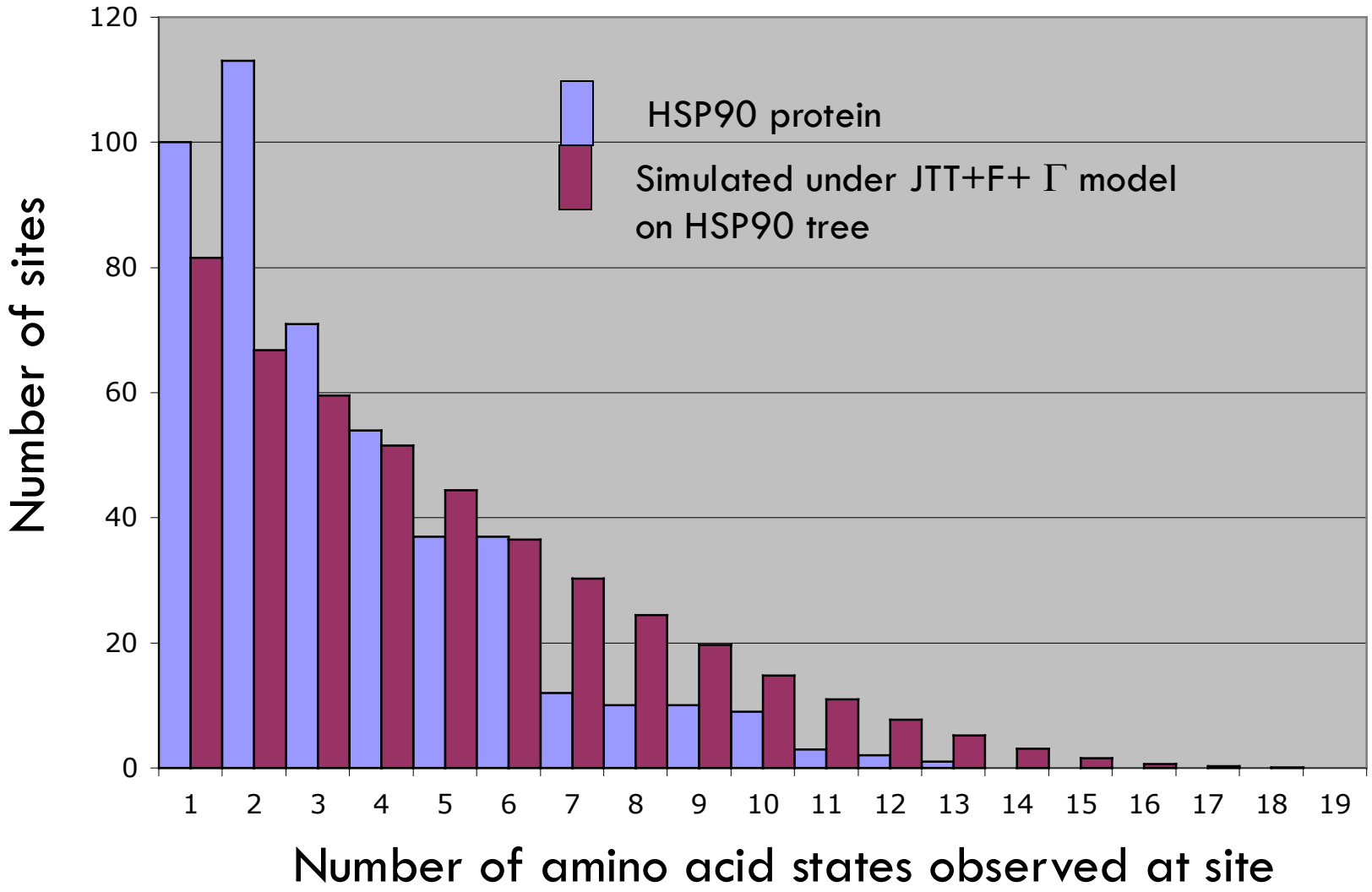
- Sites in proteins can change function over time
  - sites under negative selection ⟷ neutral ⟷ positive selection

- 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

Evolution of chaperonin 60 over ~1.5 billion years

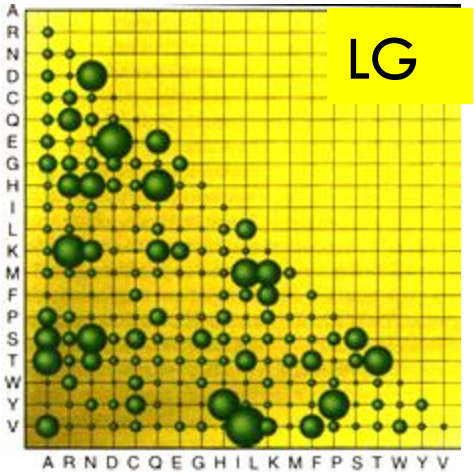


Distribution of the number of different amino acids at aligned sites

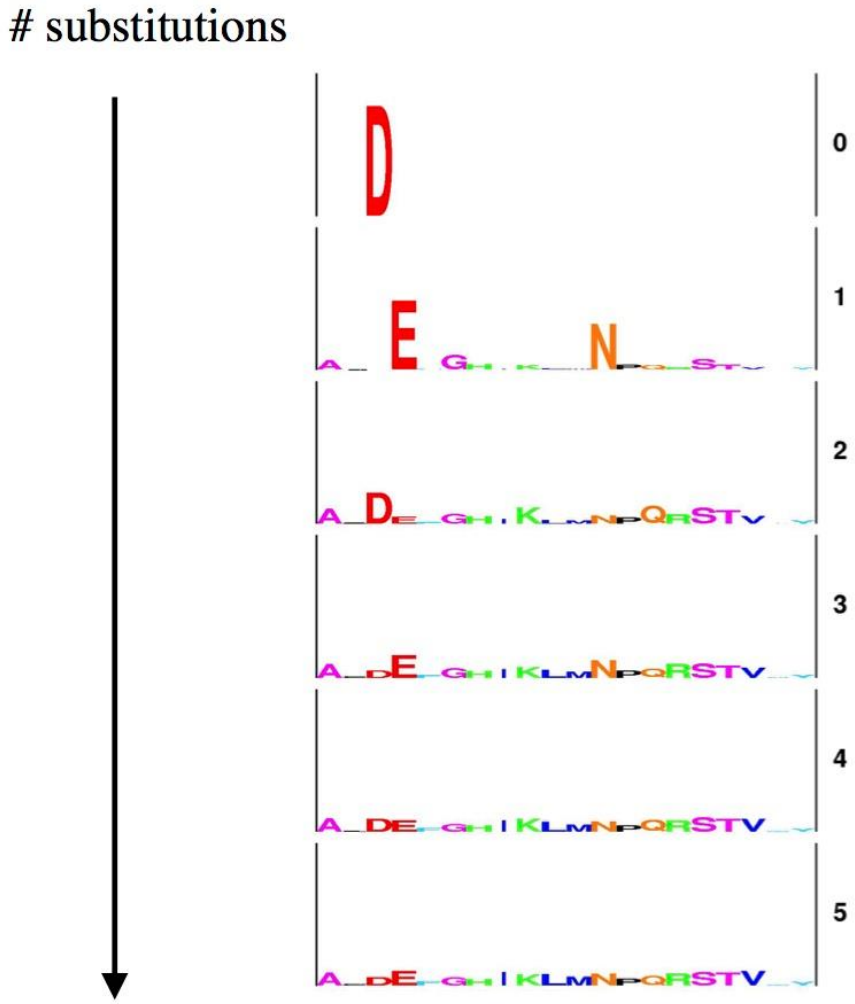


Wang *et al.* (2008) *BMC Evolutionary Biology* 8: 331

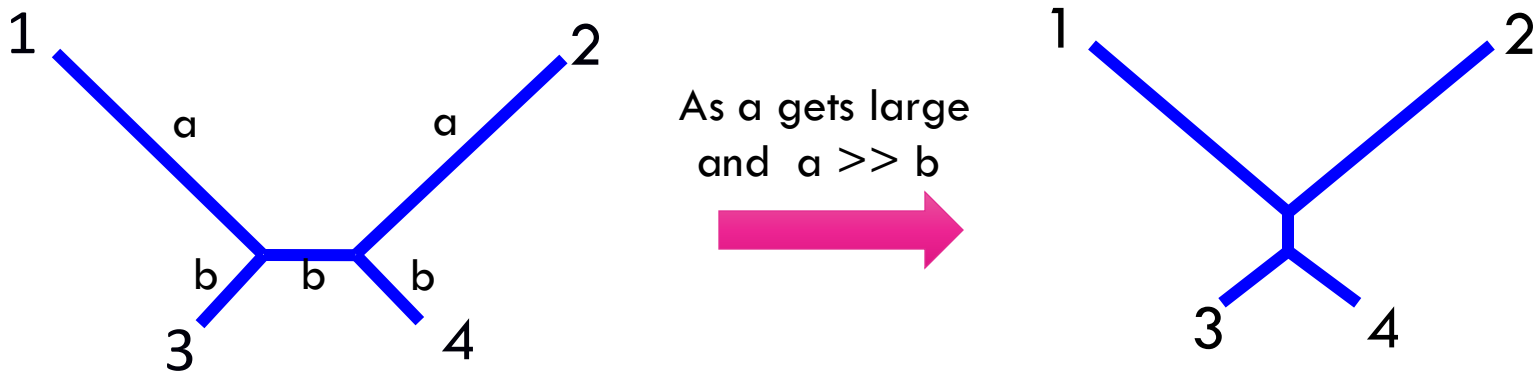
Starting at a D with a site homogeneous matrix (LG+F)



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



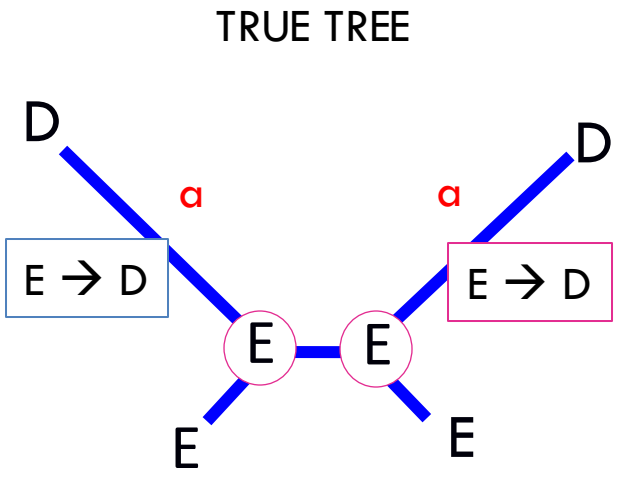
What happens to phylogenetic estimation when you ignore site-heterogeneity?

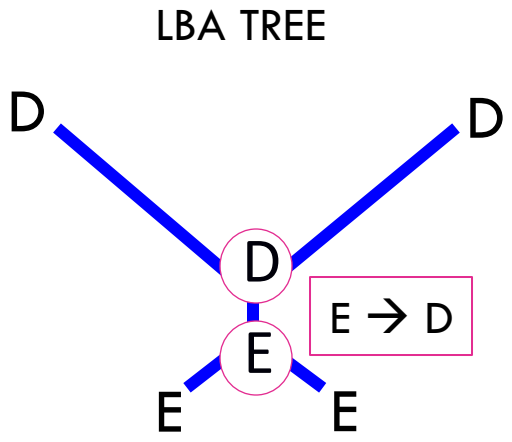


Long branch attraction

Susko et al. (2004) *Mol. Biol. Evol.*
Lartillot and Philippe (2007) *BMC Evol. Biol*
Wang et al. (2008) *BMC Evol. Biol.*
Roger and Susko (2021) *Systematic Bio*

Why long branch attraction (LBA)?

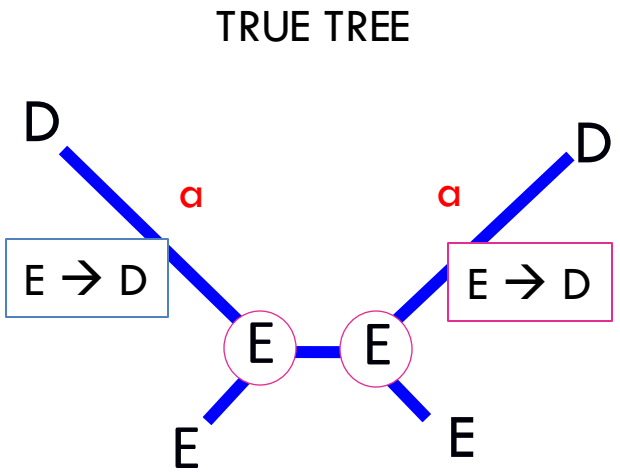


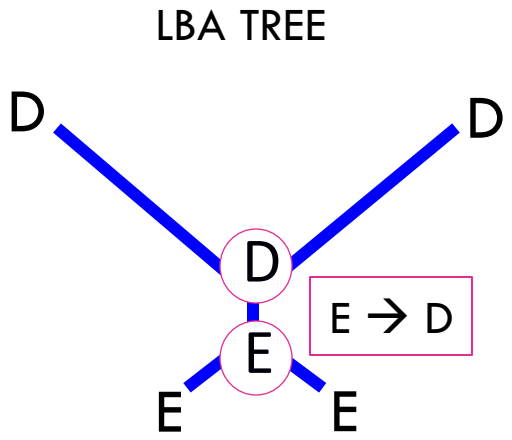


Under site homogeneous model (LG), the probability of converging on the same state (E → D) twice is pretty low:

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

Why long branch attraction (LBA)?





Under site homogeneous model (LG), the probability of converging on the same state (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):

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

Mixture models

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

$$L_i = p(\mathbf{y}_i|r_1)p(r_1) + p(\mathbf{y}_i|r_2)p(r_2) + \cdots + p(\mathbf{y}_i|r_k)p(r_k)$$

Weight of the rate class

Mixture models: terminology warning

- Different kinds of mixture models!
- Rate-category mixture model (see Dave's lecture)
- **Usually people refer to mixture of amino-acid replacement matrices**
- Mixtures can be apply to any part of the model (e.g., branch lengths)

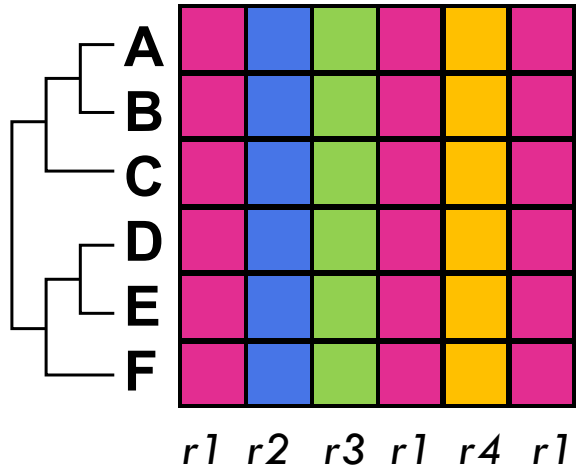
LG4M and LG4X mixture models

"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

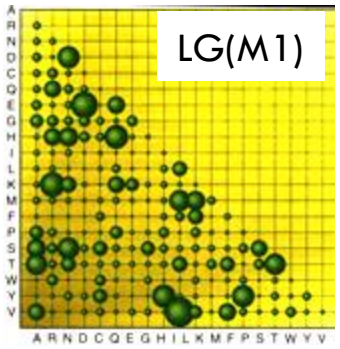
Standard LG+G model:
only the global rate differs from one category to another

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)

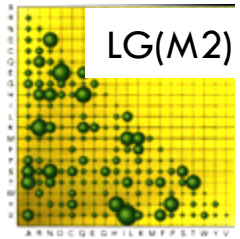


25%



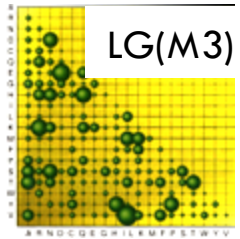
$$\Pi = \begin{bmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_R & 0 & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \pi_v \end{bmatrix}$$

25%



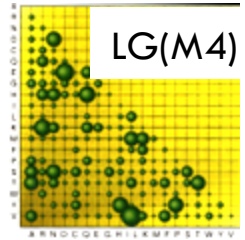
$$\Pi = \begin{bmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_R & 0 & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \pi_v \end{bmatrix}$$

25%



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

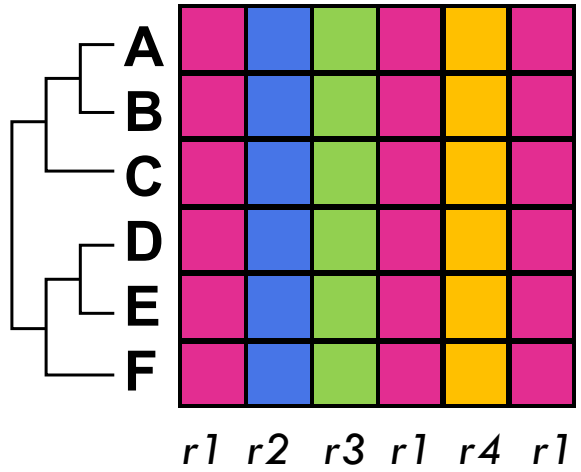
25%



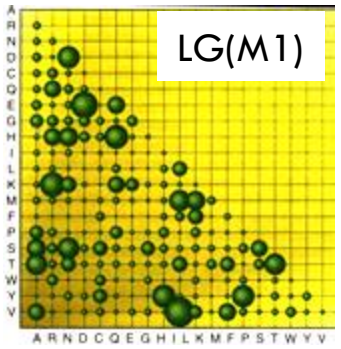
$$\Pi = \begin{bmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_R & 0 & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \pi_v \end{bmatrix}$$

LG4M and LG4X

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

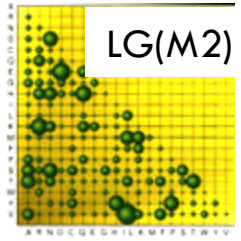


50%



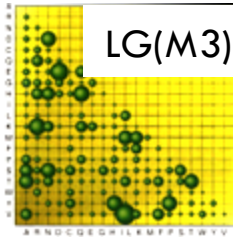
$$\Pi = \begin{bmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_R & 0 & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \pi_v \end{bmatrix}$$

15%



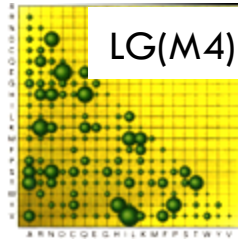
$$\Pi = \begin{bmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_R & 0 & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \pi_v \end{bmatrix}$$

30%



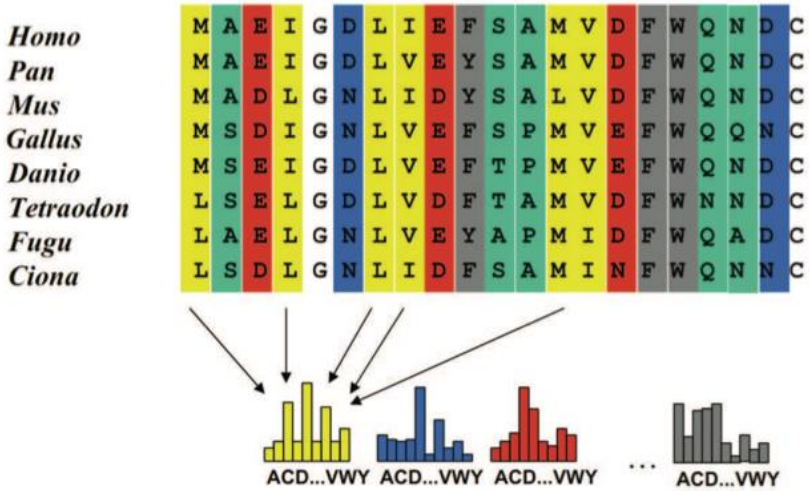
$$\Pi = \begin{bmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_R & 0 & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \pi_v \end{bmatrix}$$

5%



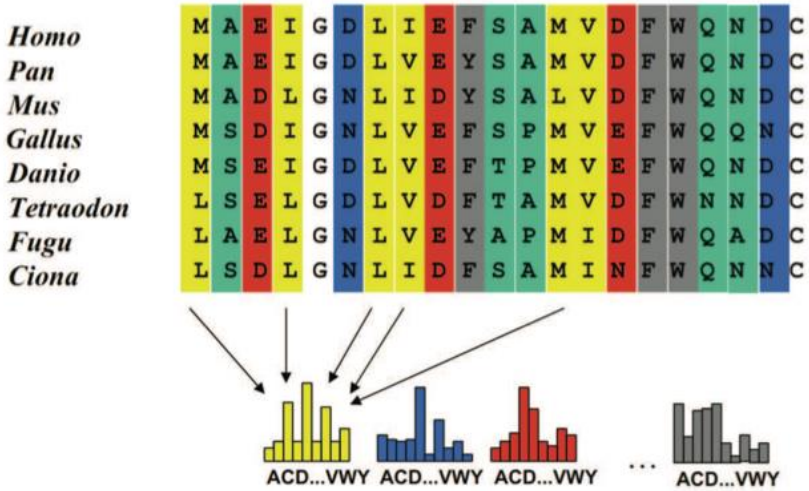
$$\Pi = \begin{bmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_R & 0 & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \pi_v \end{bmatrix}$$

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

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

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**

➔ PMSF (Posterior Mean Site Frequency) approximation: transforming a mixture model into a 'simple' model (Wang, Bui, Susko, Roger *Systematic Biology* 2018)

PMSF (Posterior Mean Site Frequency) model

Implemented in IQtree

1) Reconstruct an ML tree under a "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 ($f_{j,x}$)

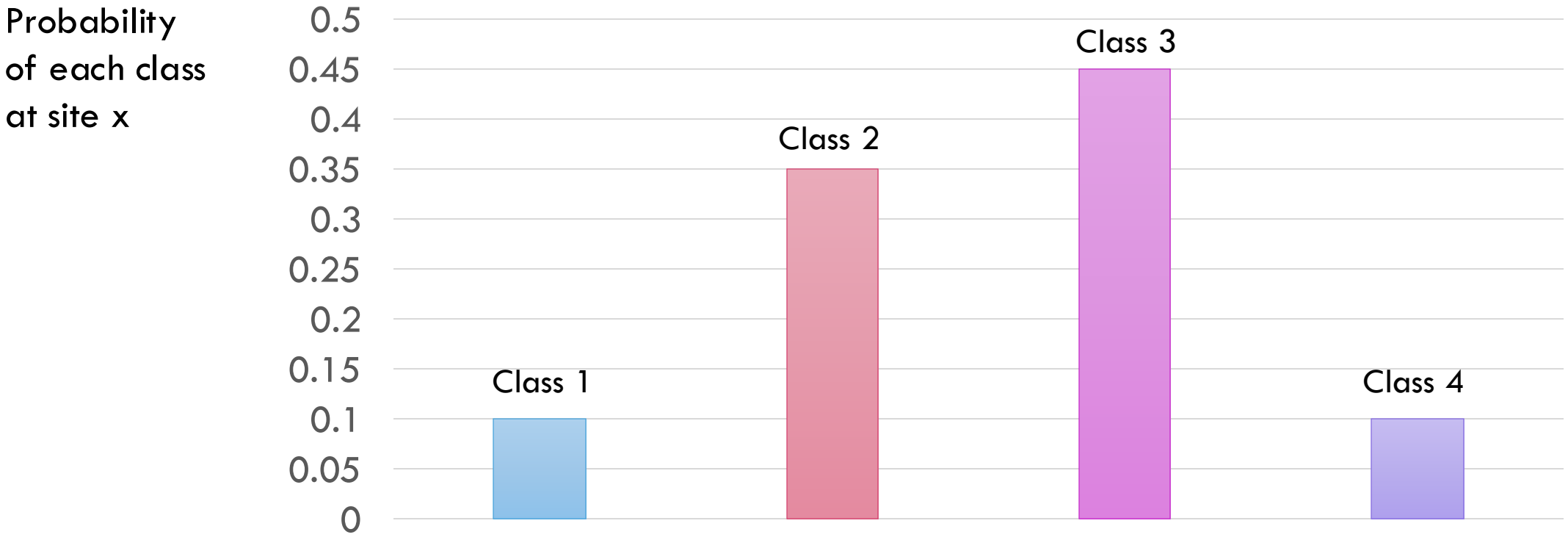
$$f_{j,x} = \sum_c f_{j,c} \times P(c|x)$$

Sum over all classes

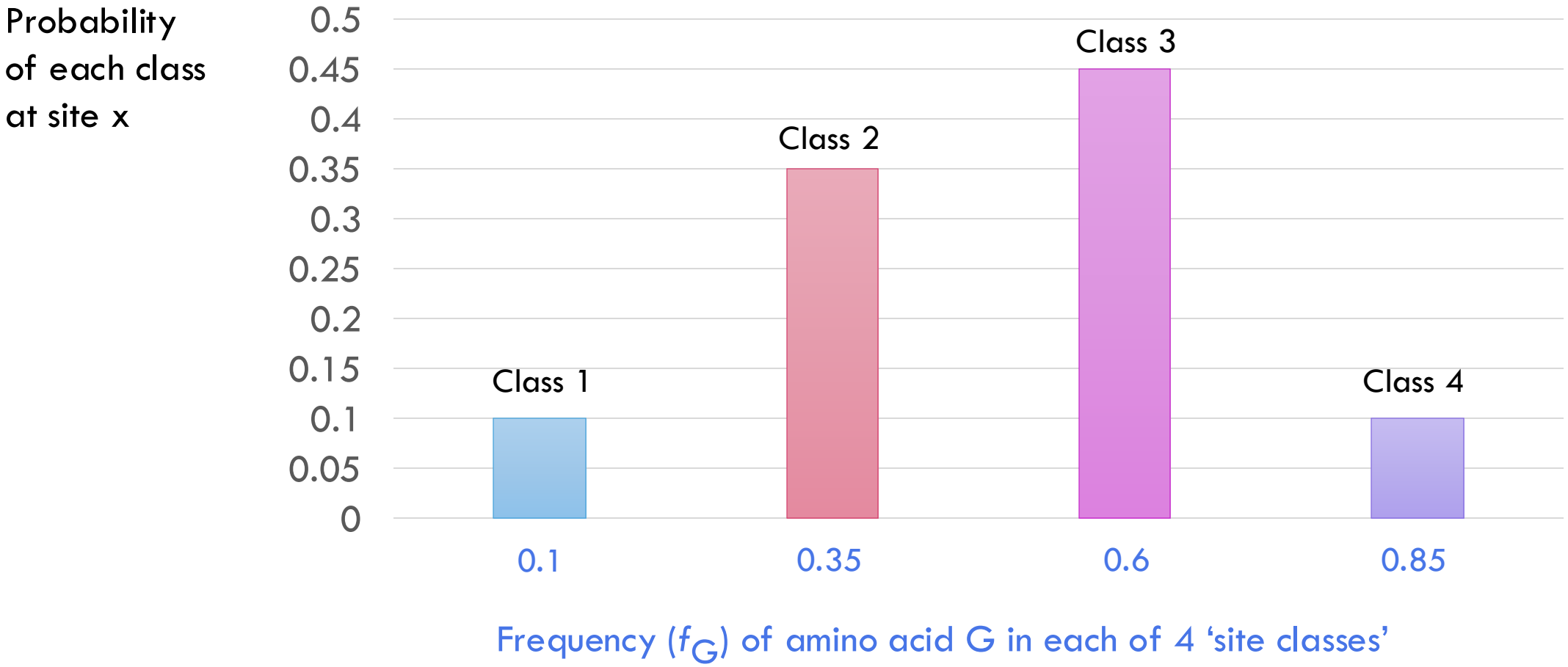
Freq of AA j for class c

Prob of class c at site x

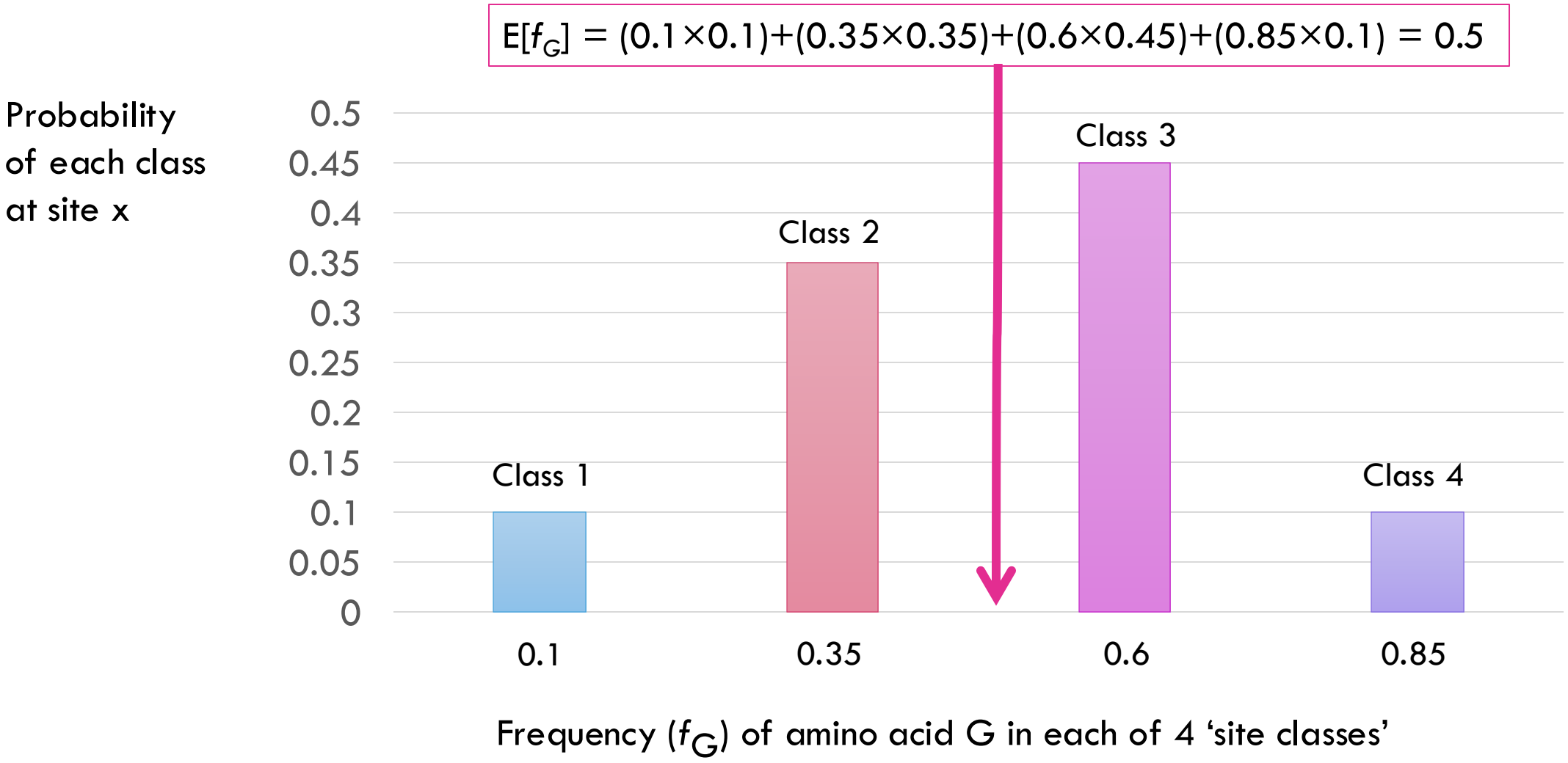
Example: Posterior mean site frequency for 'G' at a given site x, with a 4 class mixture model



Example: 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

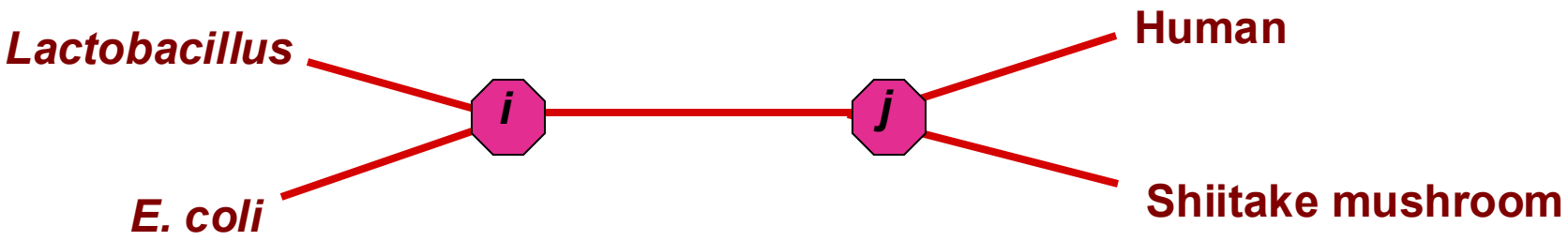
➔ Bootstrapping is dramatically faster

Take home (for this part)

- Models are idealizations of the actual process of protein evolution
- Model misspecification (e.g. 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

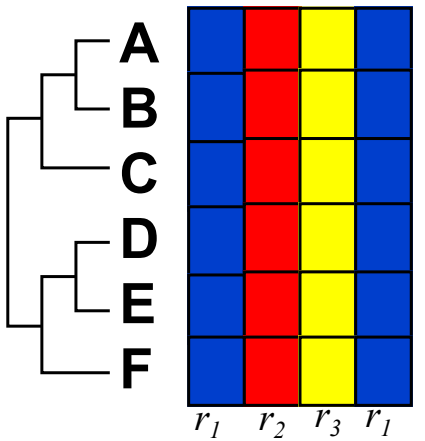


$$P(j \mid i;t) = \left[\exp(R \square \Pi \square t_e \square r_v)\right]_{ij}$$

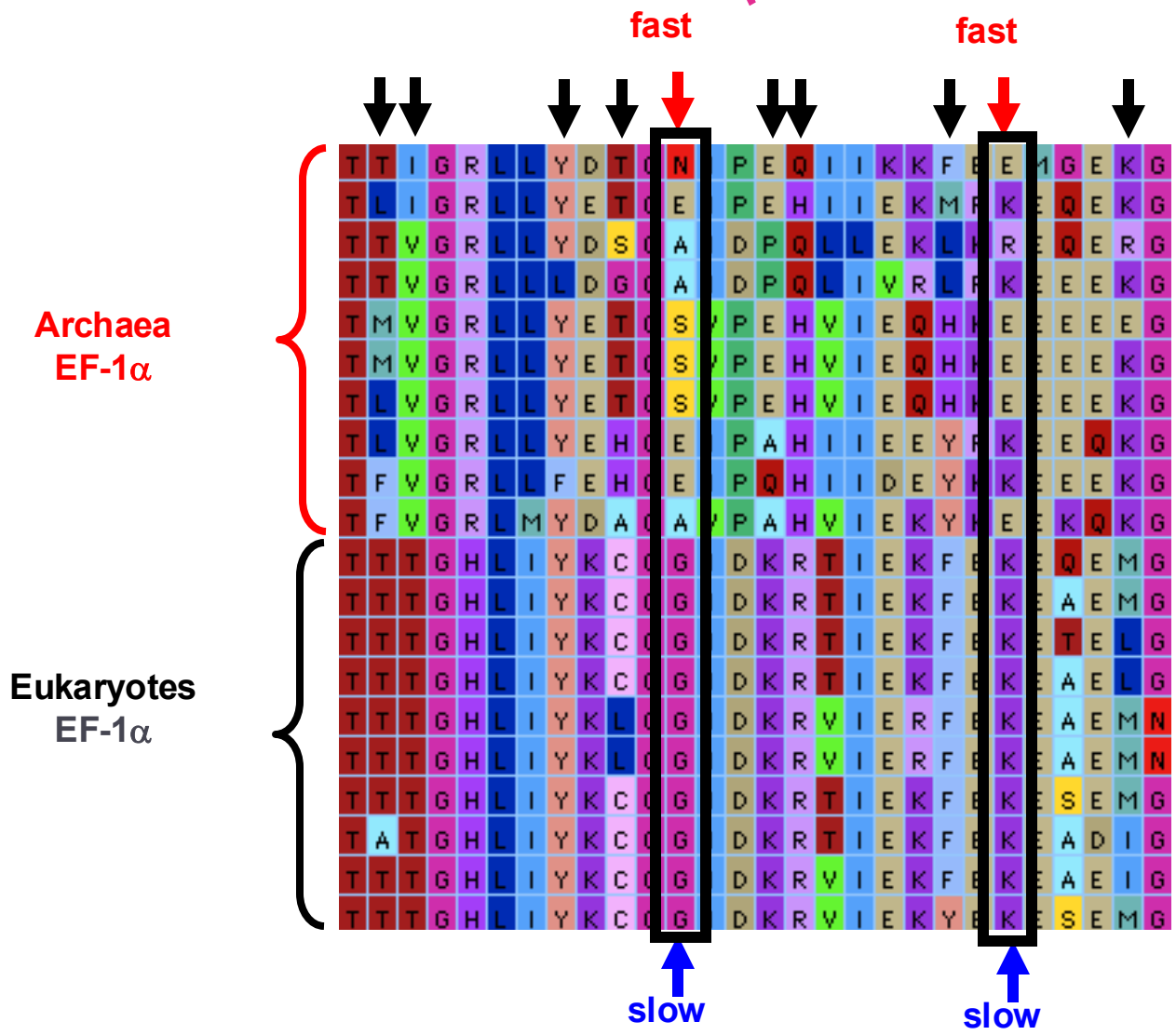
Assumptions

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

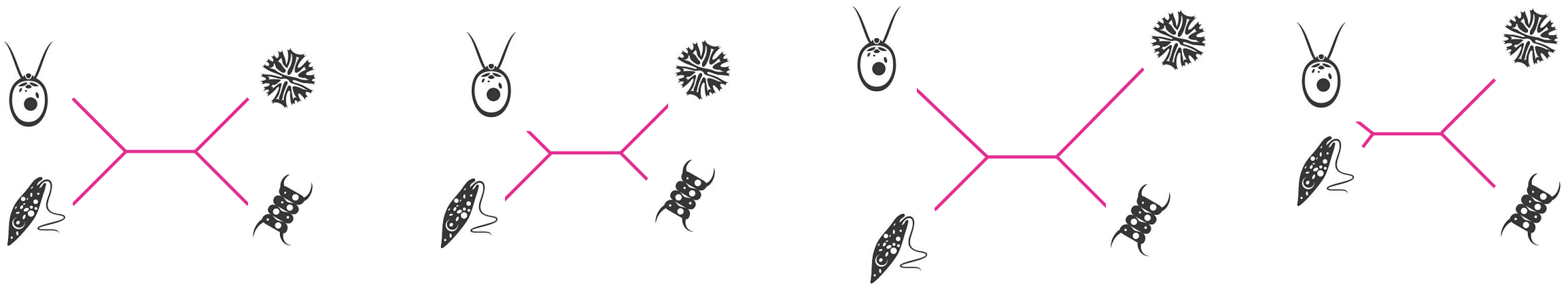
-Sites have the same rate of evolution (r_v) on different branches of tree



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

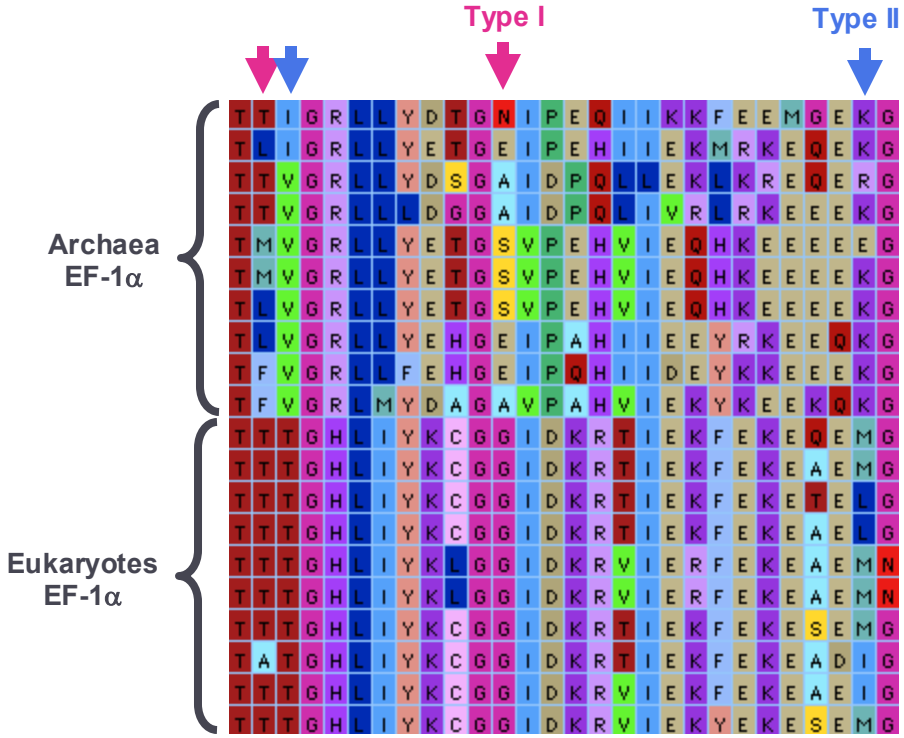
- Rate-shift models
  - Allows rates at many different sites to change abruptly on one branch

- Mixture of branch-length models
  - Allows different branch-lengths for different sites (e.g. GHOST model in IQtree)

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).



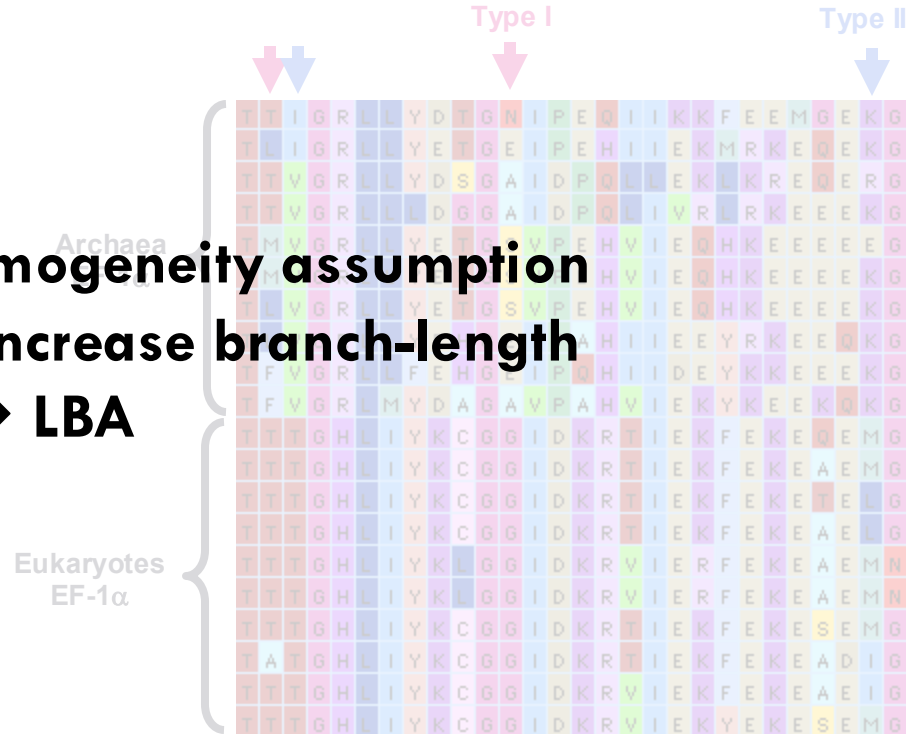
Functionally divergent sites generate heterotachy

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



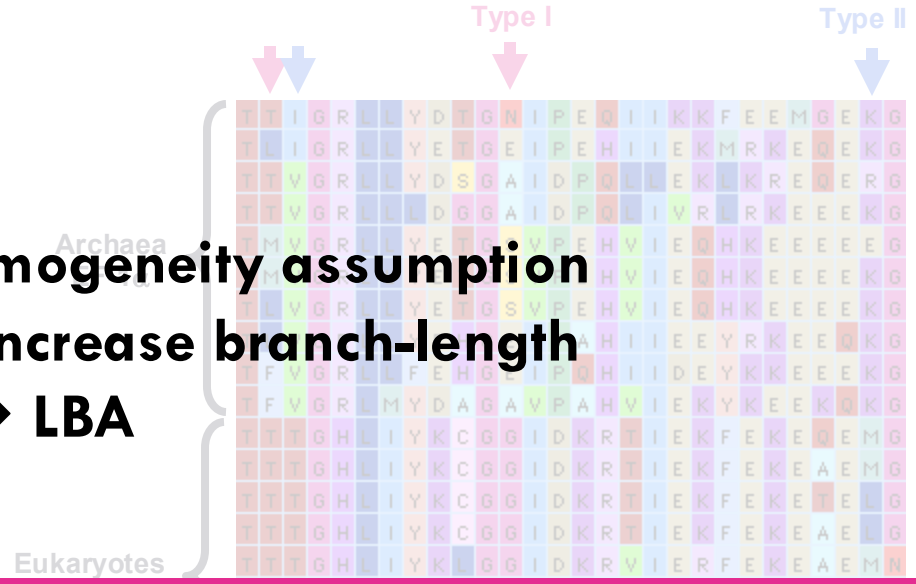
Functionally divergent sites generate heterotachy

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 phylogenetic tree but for

FD sites violate homogeneity assumption and artefactually increase branch-length

➔ LBA

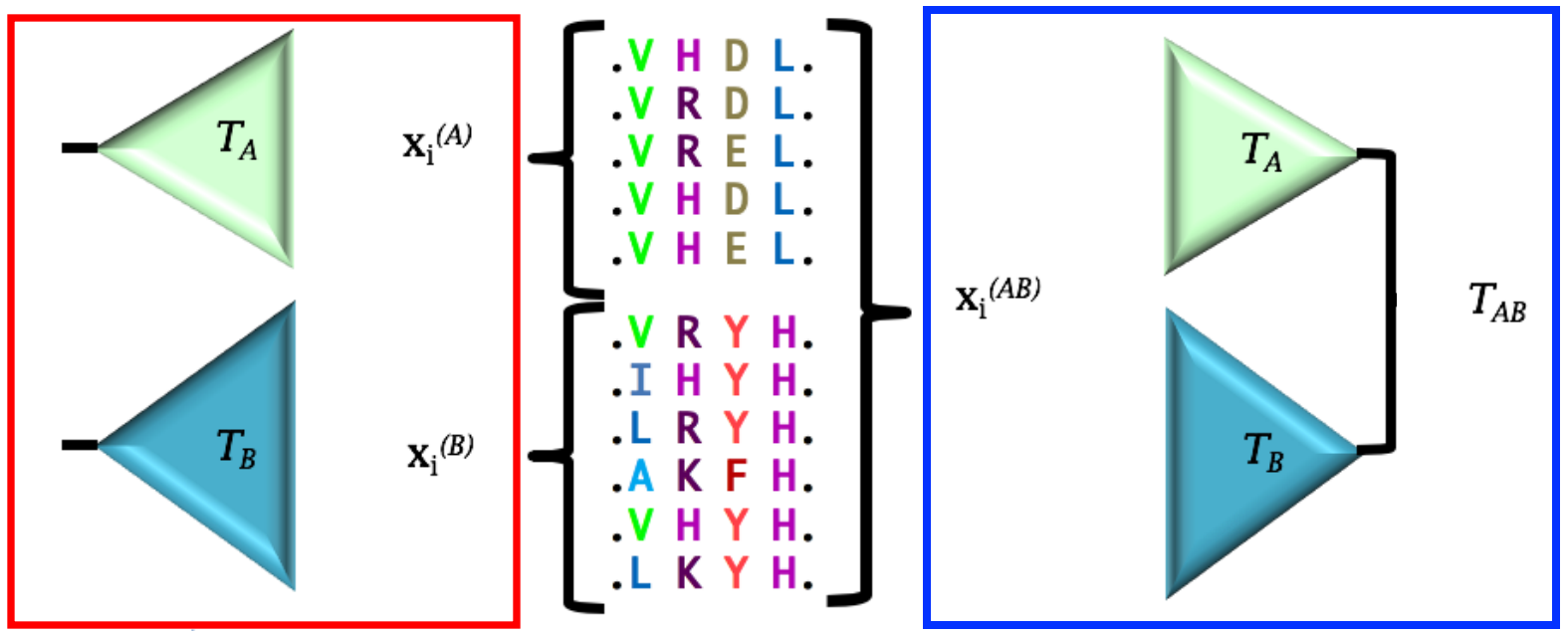


FunDi : identifies FD sites along a specific branch taking into account the phylogeny (ML framework)

FunDi mixture model

- For each site, FunDi allows for FD and non-FD evolution across a pre-specified split:

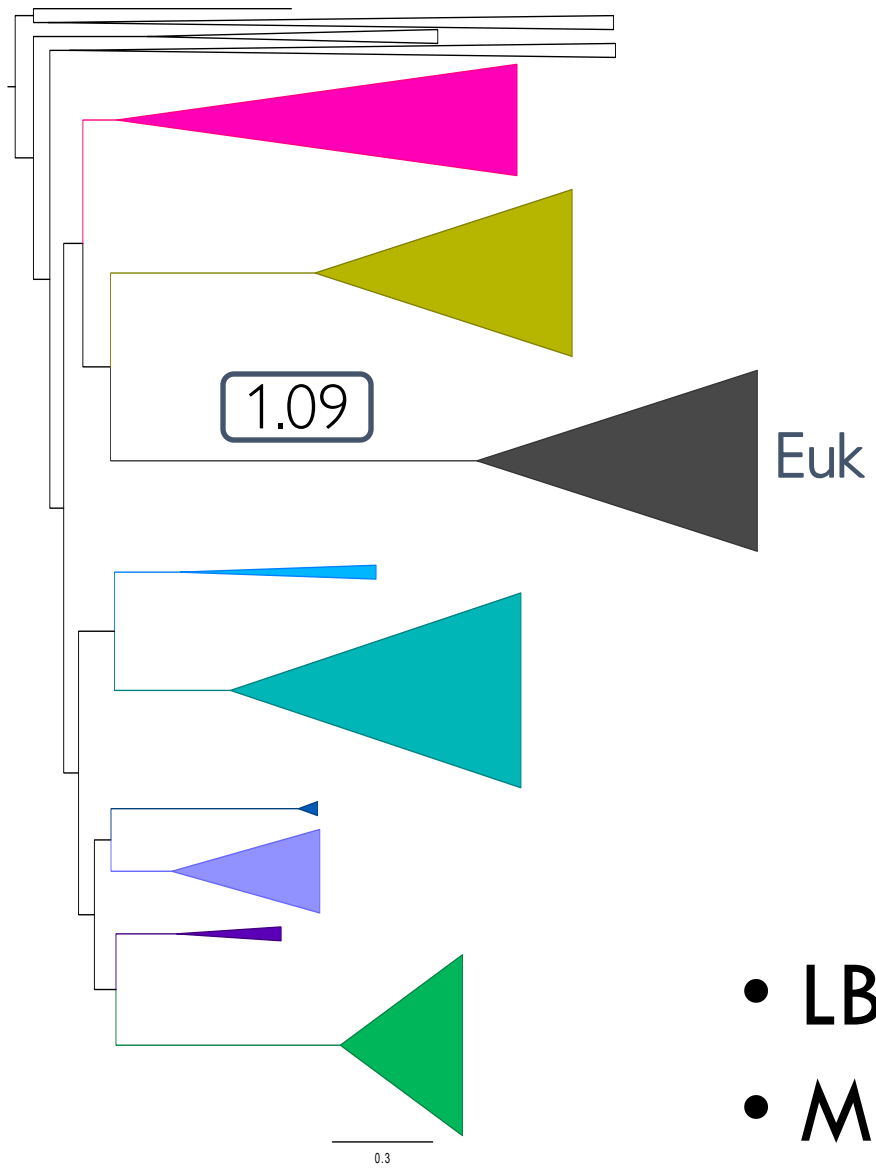
$$L_i(\theta, T_{AB}) = \rho \left(P(\mathbf{x_i}^{(A)}; \theta, T_A)\right) \left(P(\mathbf{x_i}^{(B)}; \theta, T_B)\right) + (1-\rho) \left(P(\mathbf{x_i}; \theta, T_{AB})\right)$$



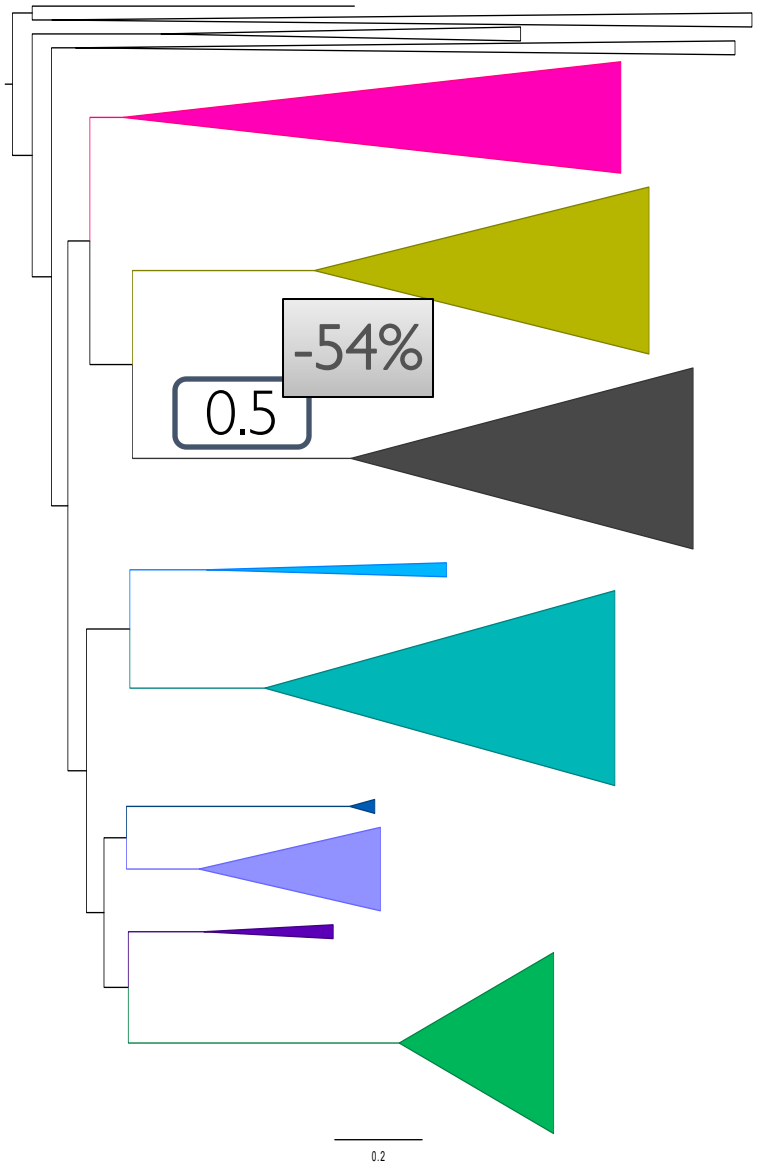
Independent across split (=FD)

Dependent across split (=non-FD)

Gaston, Susko and Roger (2011)



- ‘FD sites’



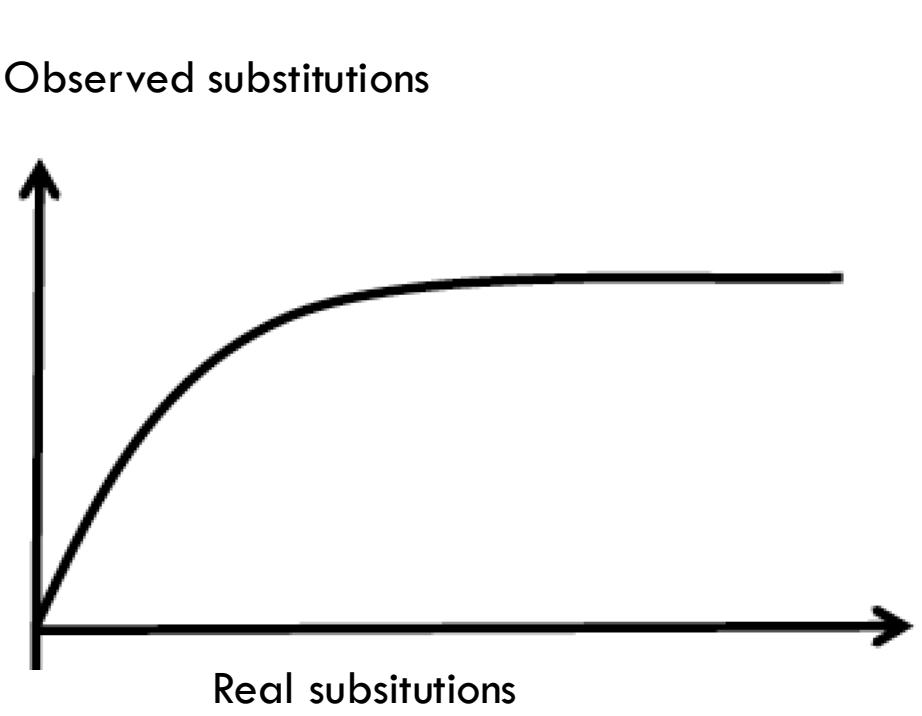
- LBA
- Molecular clock

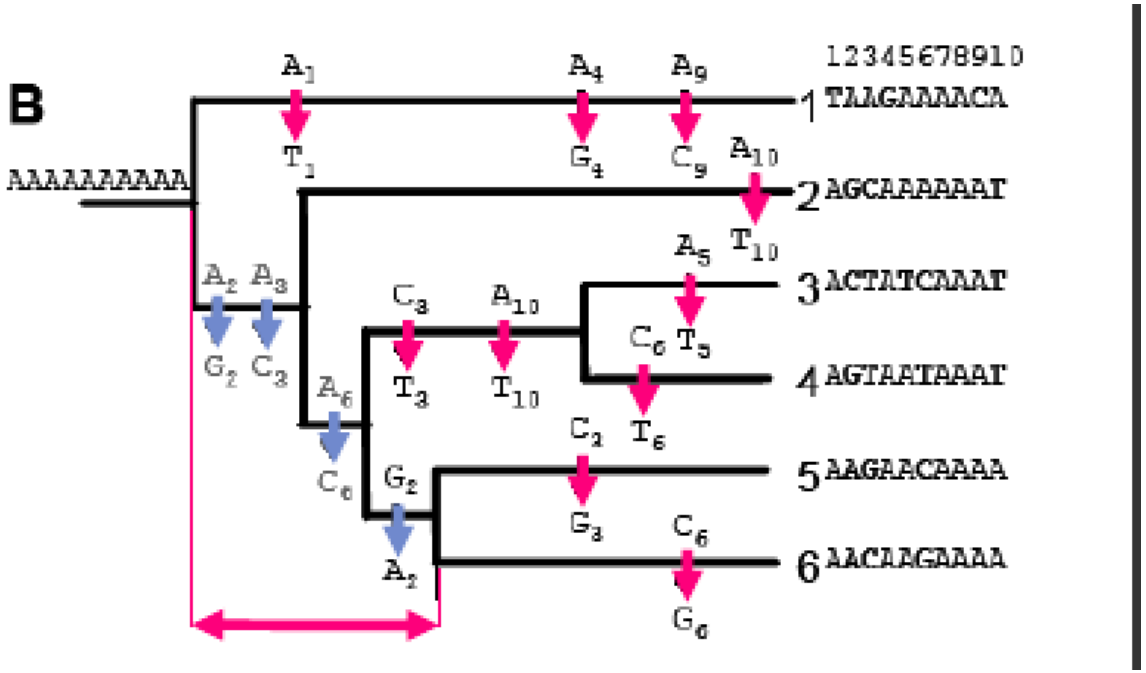
Break

PART 2
Real examples of ‘deep’ phylogenetic problems and how we tried to address them

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

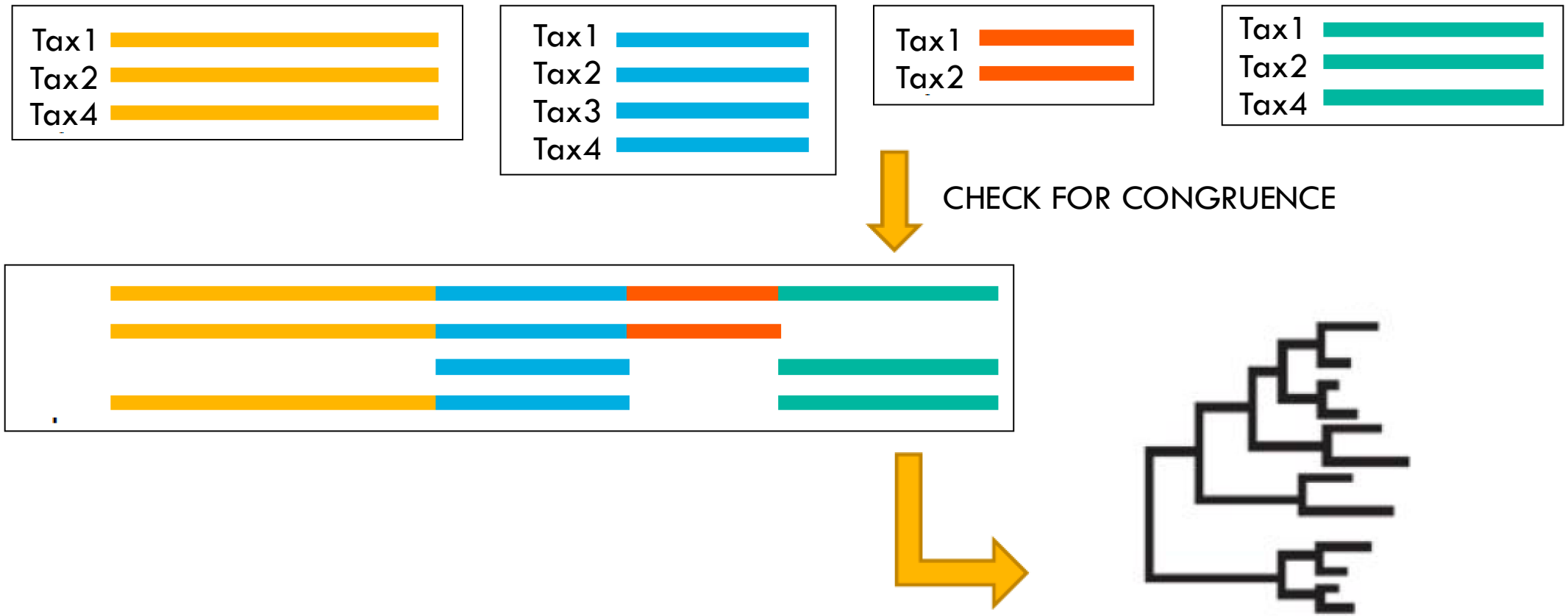
“Ancient” signal erased by more recent substitutions





Supermatrices

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



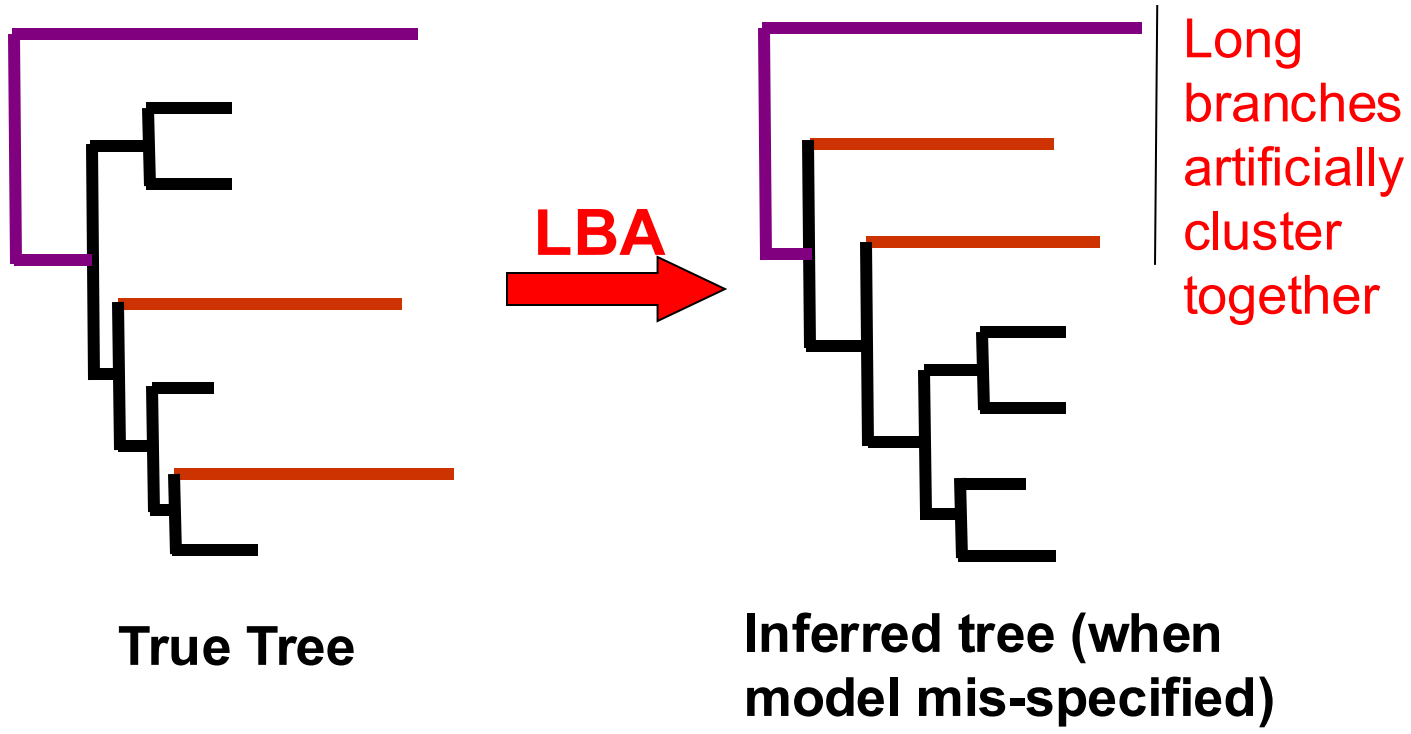
What can affect your topology

- 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
- (many other things…)

Example 1: Effect of model misspecification and of fast-evolving sites

Model mispecification: statistical inconsistency

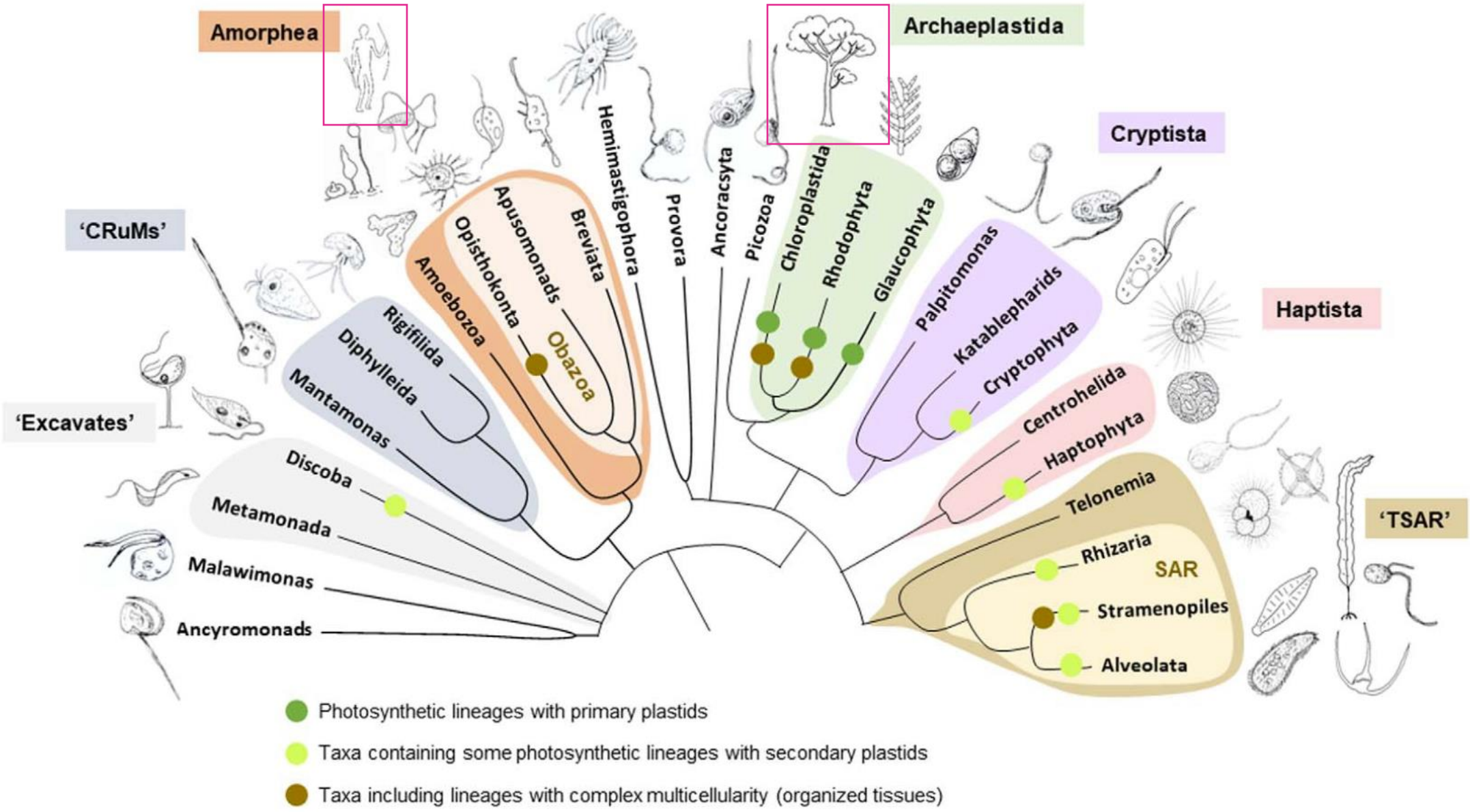
Long Branch Attraction (LBA) Artefact



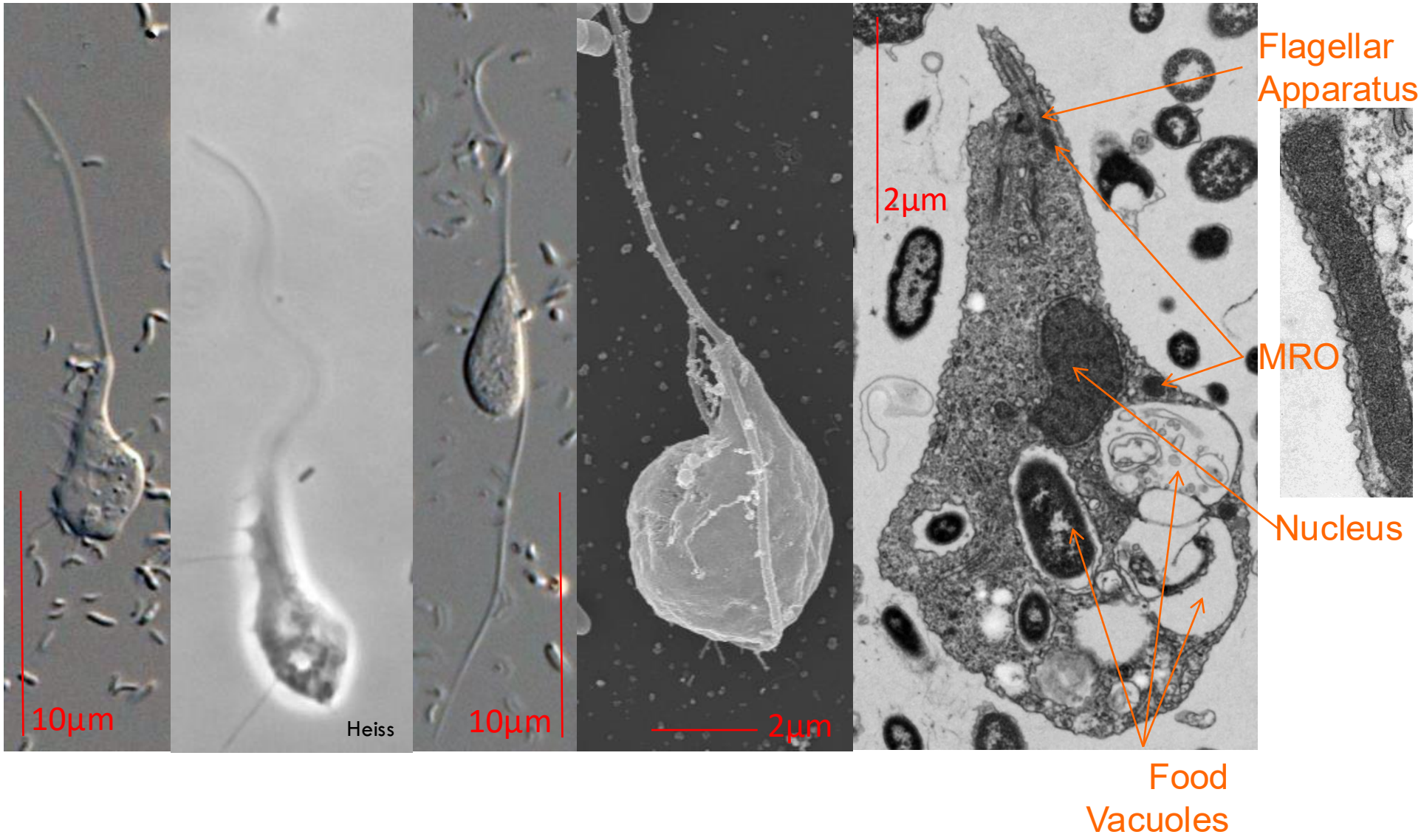
Adding more data *strengthens* artefact

→ statistical inconsistency

Tree of eukaryotes



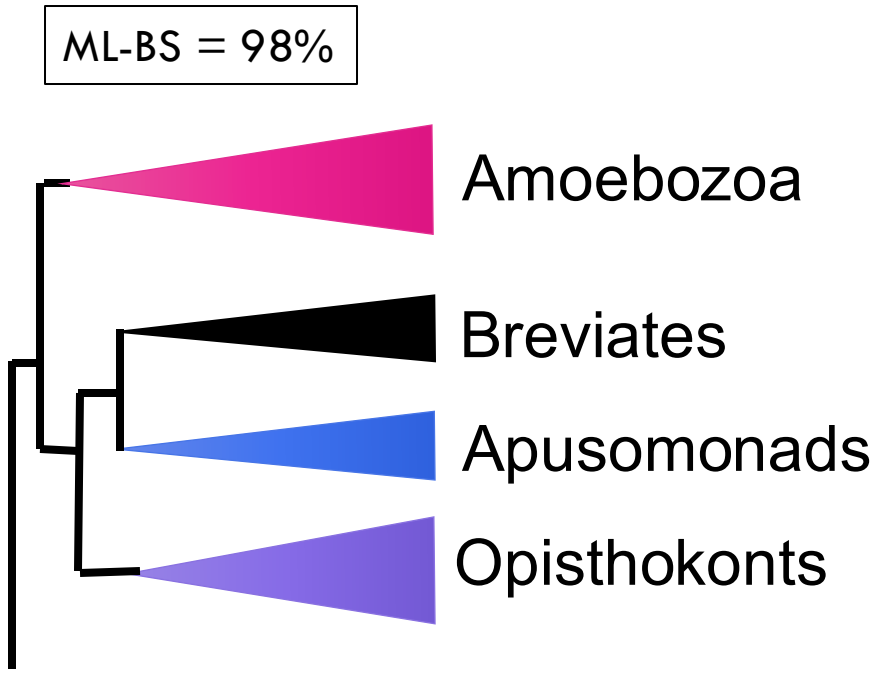
Pygsuia biforma (Brown,…Roger 2013)

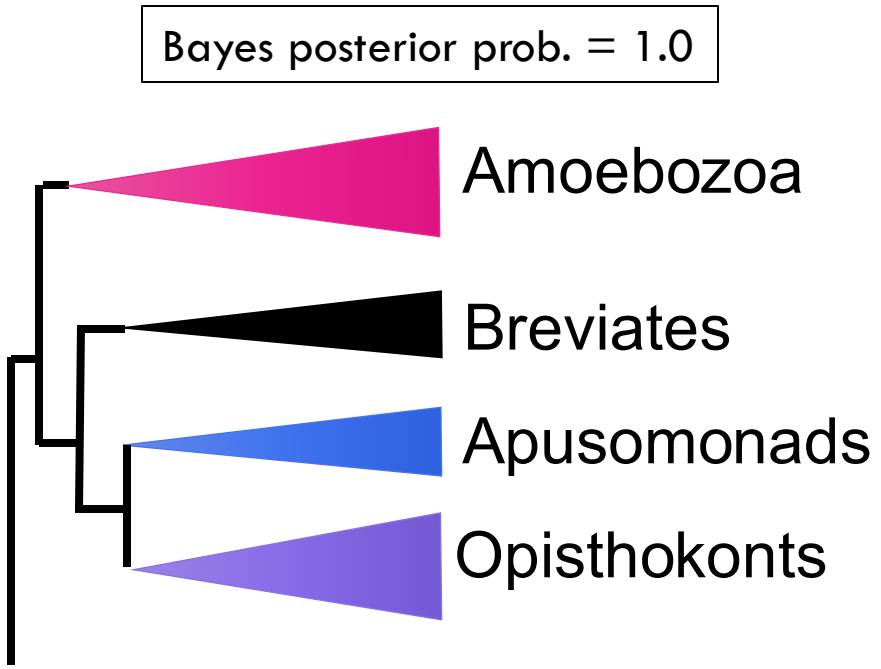


Two different topologies within Obazoa are supported by different phylogenetic models

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 mis-specified 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?

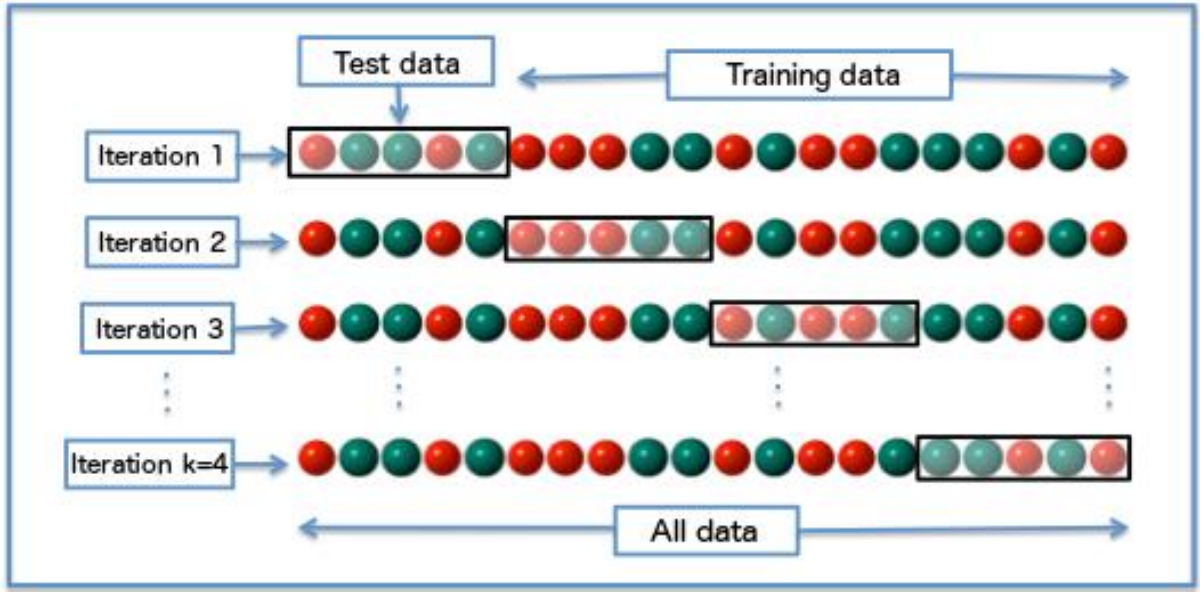
Cross-validation:

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?

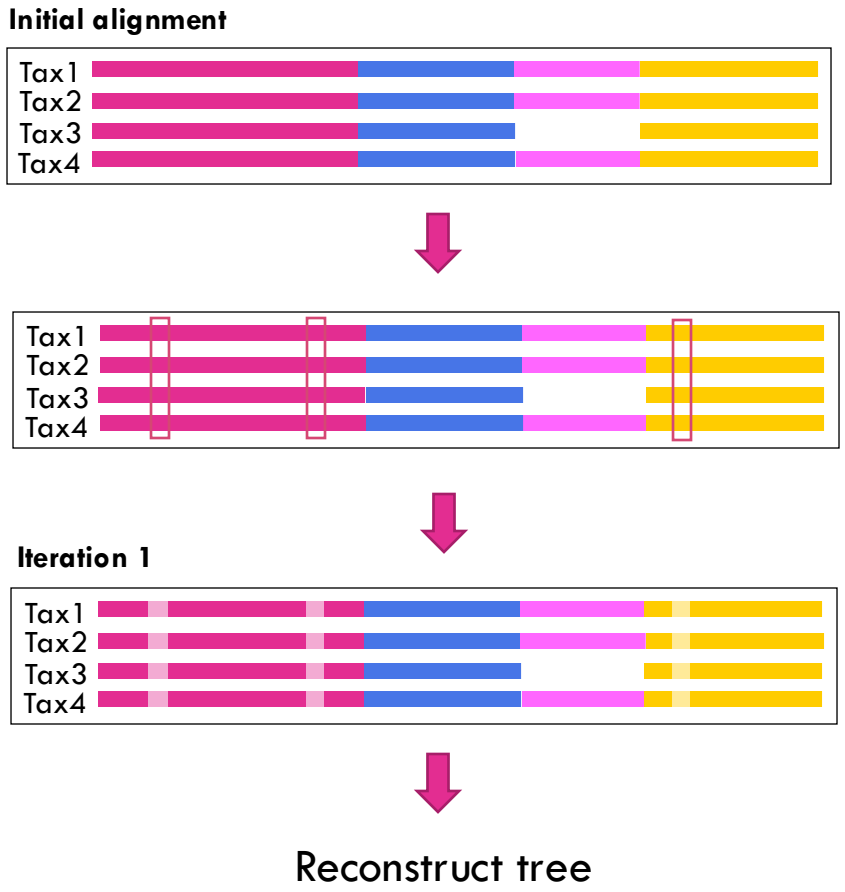
- One of two topologies is likely artefactual resulting from misspecified model
- Test which substitution model fits better
  - Cross-validation
- Try to eliminate ‘noisiest’ data
  - Fast-evolving site removal
  - Fast-evolving gene removal
  - Fast-evolving taxon removal
  - Recoding

Removal of fast-evolving sites

Goal: can we yield the same topology
under LG+G as under CAT+GTR
after we remove poorly modelled sites?

Fast Evolving Sites removal

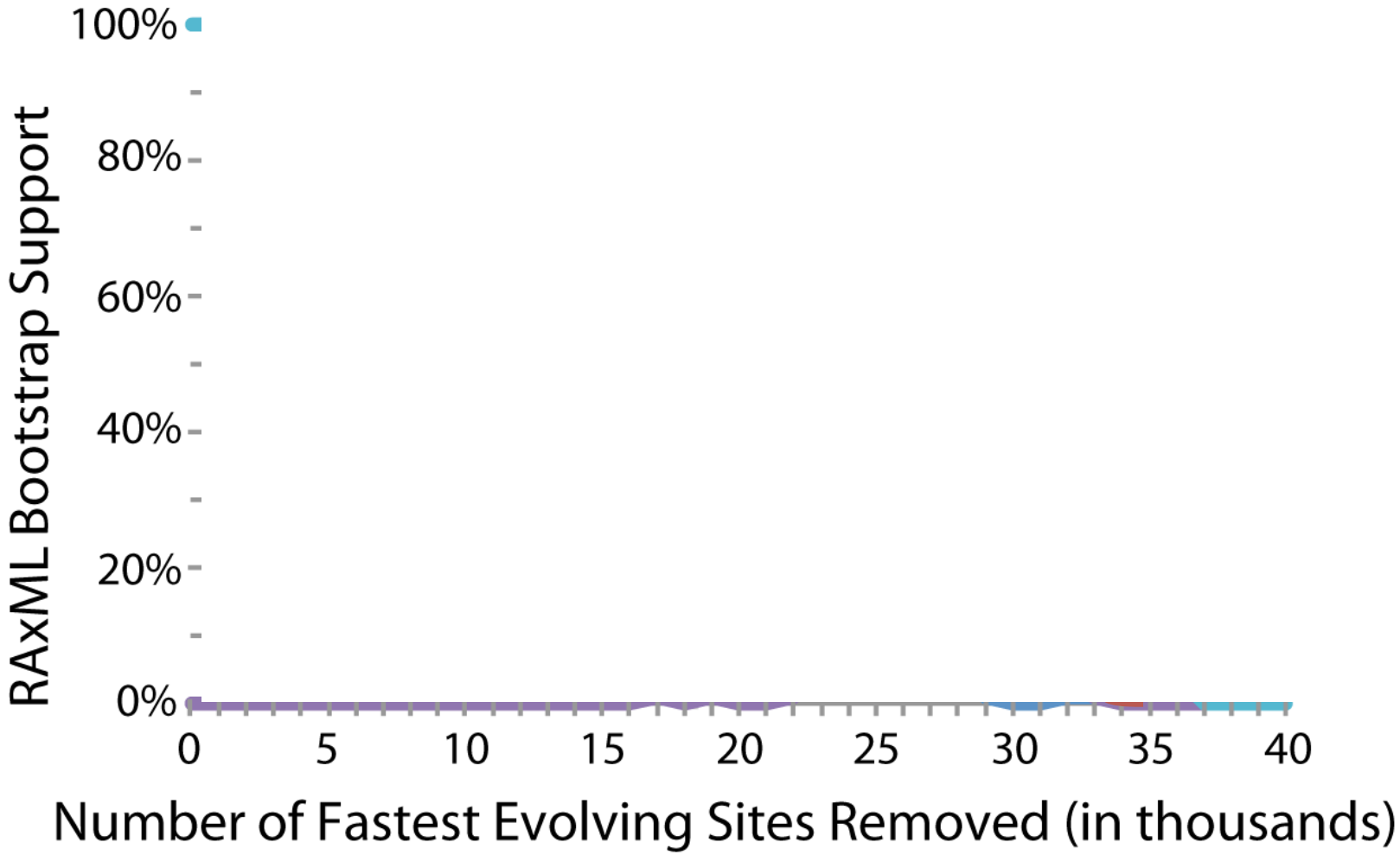
Fast-evolving sites : carry the 'noisiest' signal (most saturated sites)



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

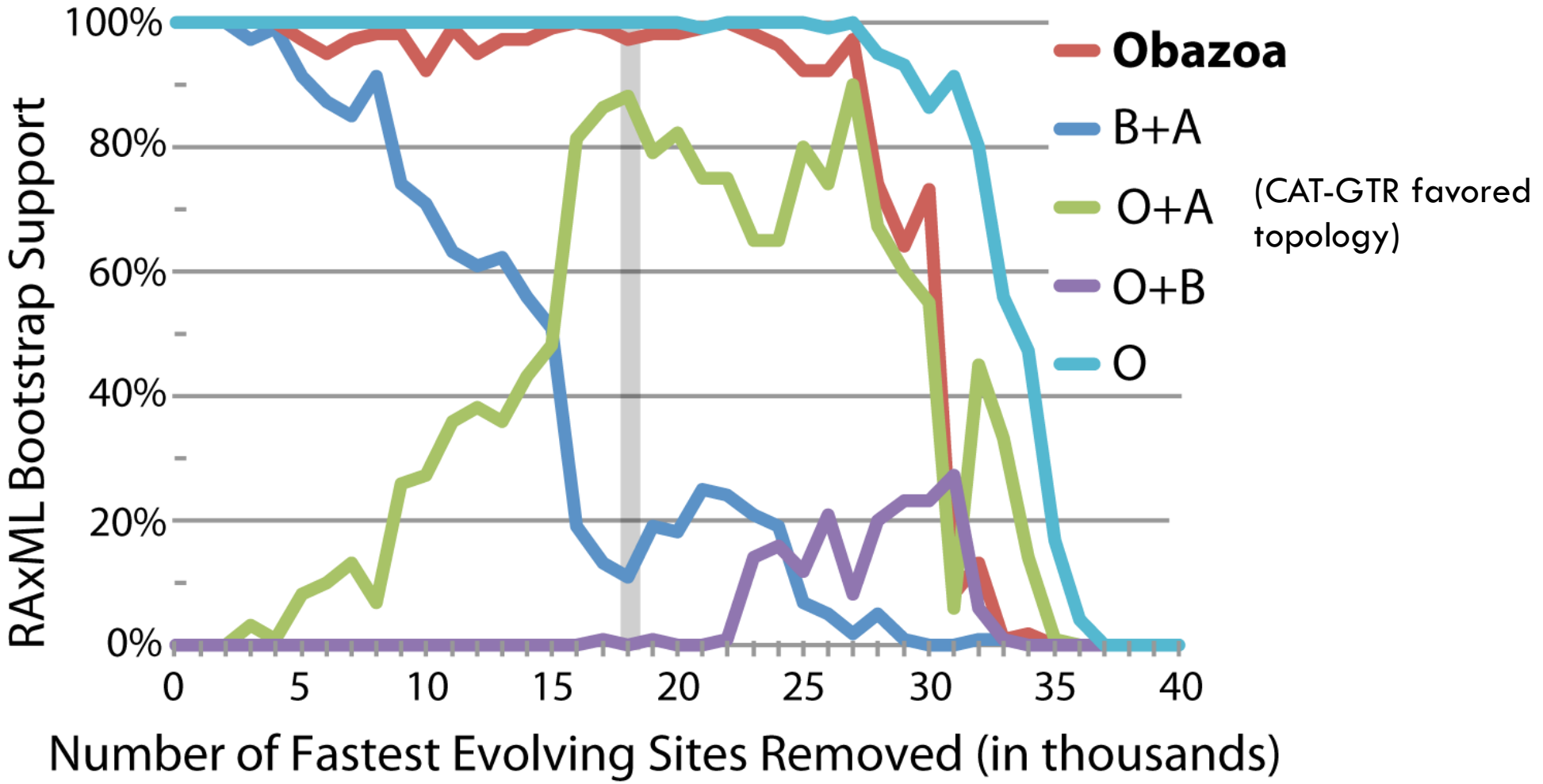
Step	# sites left	
1	43615	Tree1
2	42615	Tree 2
…		
4	40615	Tree 4
…		
40	3615	Tree 40

Estimate support for conflictual clades as we remove Fast-Evolving Sites (under LG+G)



Brown et al. 2013 *PRSB*

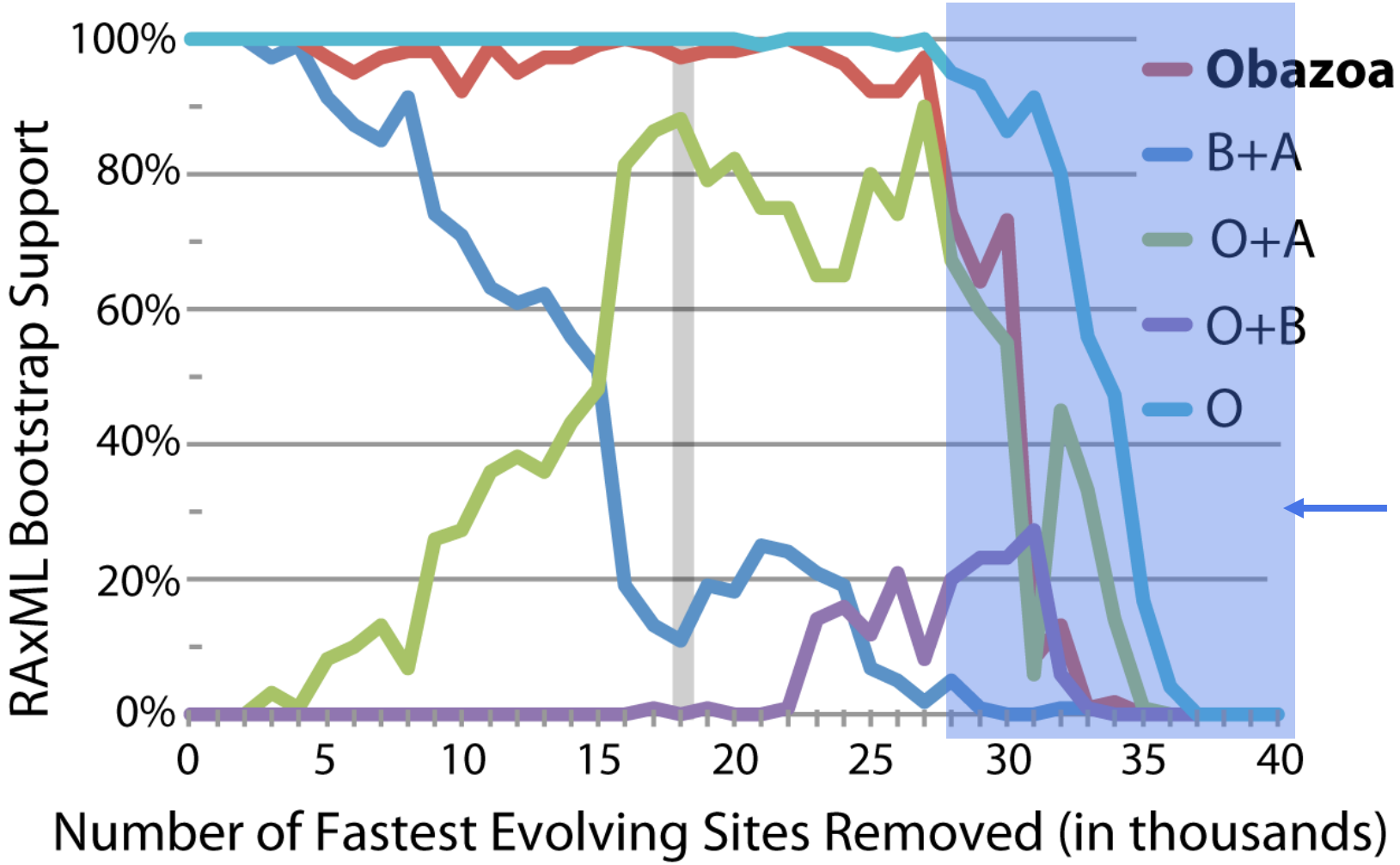
Estimate support for conflictual clades as we remove Fast-Evolving Sites (under LG+G)



(CAT-GTR favored topology)

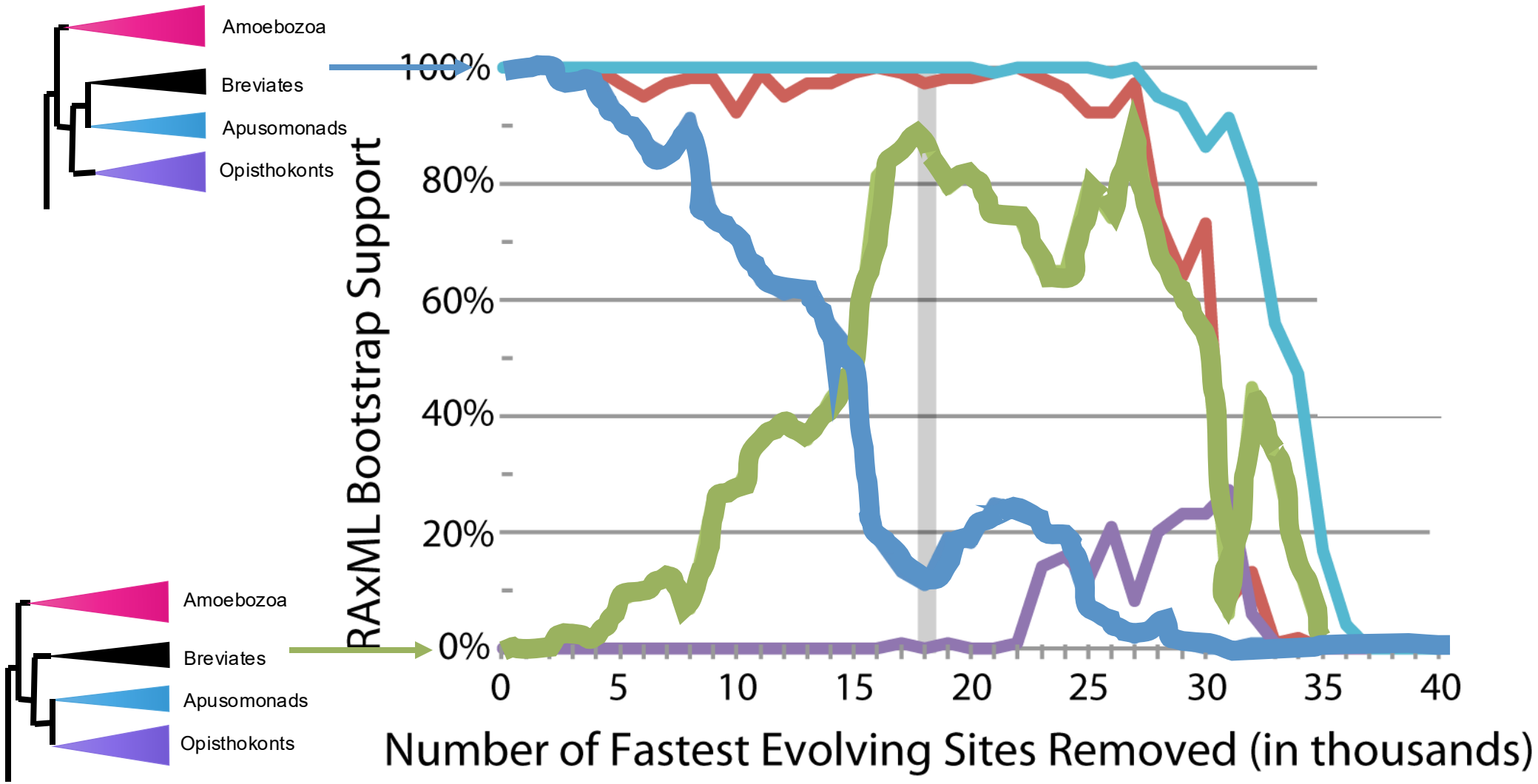
Brown et al. 2013 *PRSB*

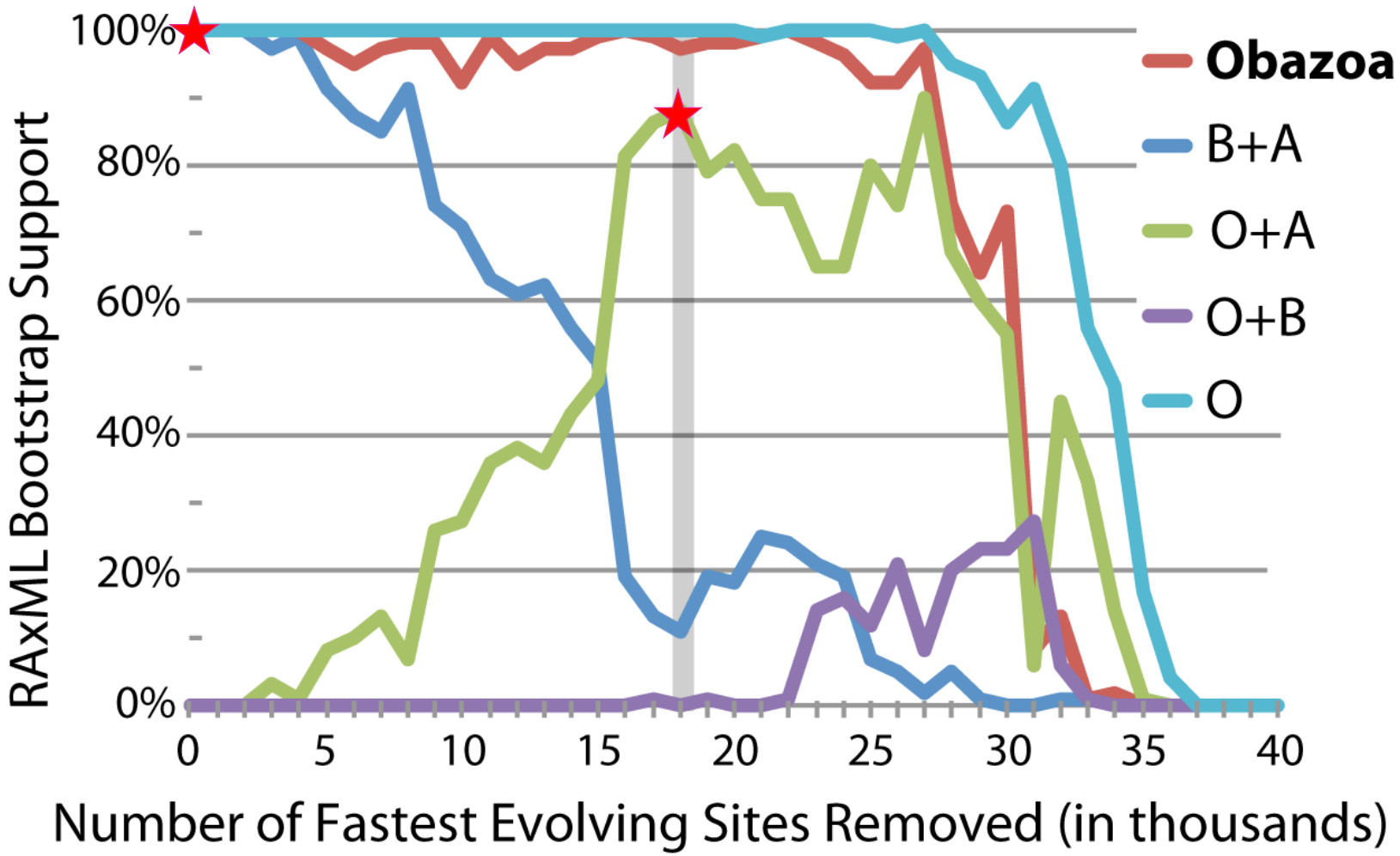
Estimate support for conflictual clades as we remove Fast-Evolving Sites (under LG+G)

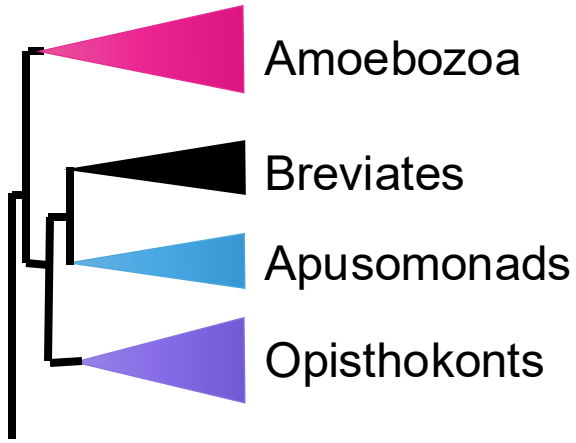


Brown et al. 2013 *PRSB*

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

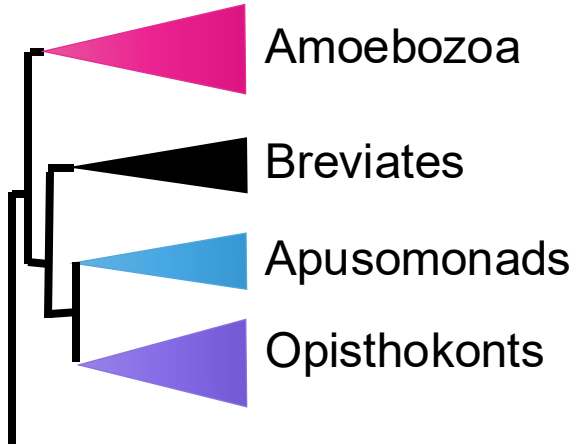



100%
80%
60%
40%
20%
0%
RAxML Bootstrap Support
0
5
10
15
20
25
30
35
40
Number of Fastest Evolving Sites Removed (in thousands)
Obazoa
B+A
O+A
O+B
O



ML – LG+Γ

Removal of 18,000 fastest-evolving sites



ML – LG+Γ

Fast-evolving site removal: warning

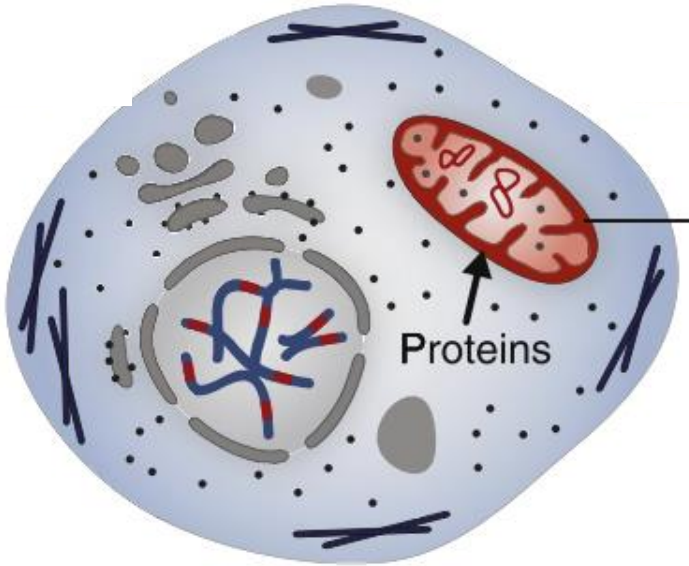
- Poor proxy for heterotacheous site removal
  - Fast sites in themselves are not necessarily a problem if they are fast across the entire tree

- 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

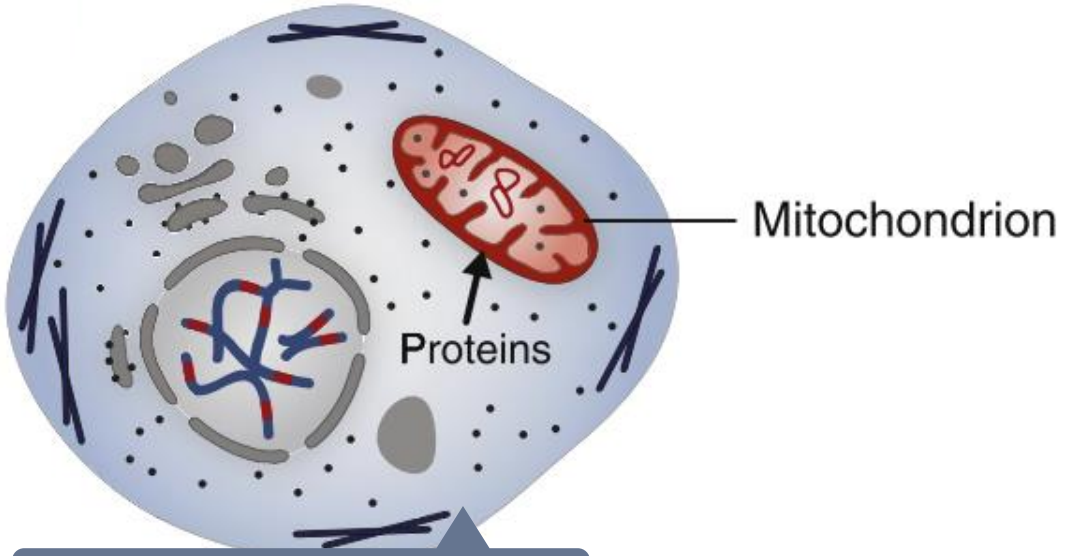
Example 2: Effect of amino-acid preference change over the tree

A brief account of the little we know about the origin of eukaryotes

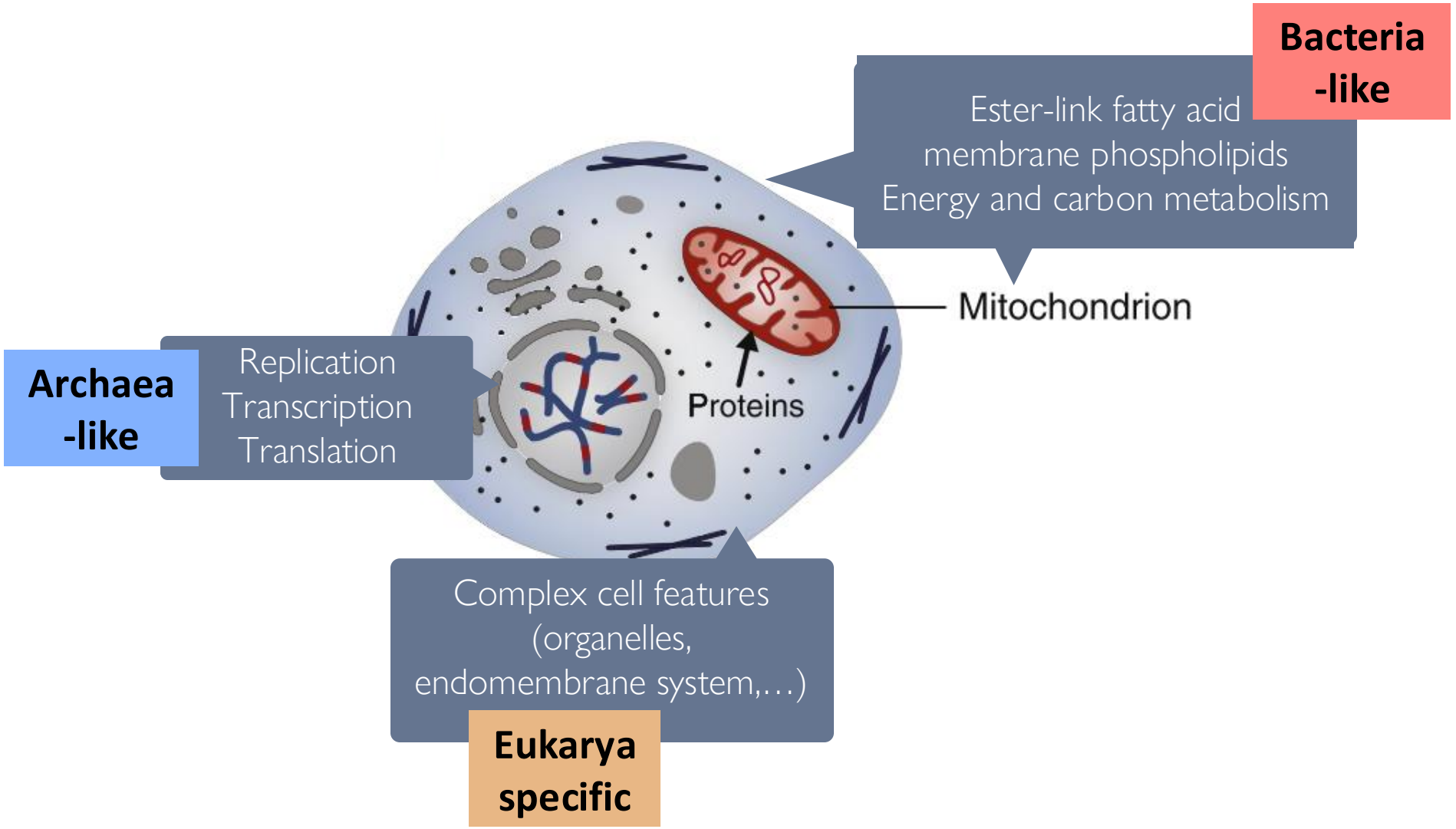


Complex eukaryotic cell

The chimeric nature of eukaryotes



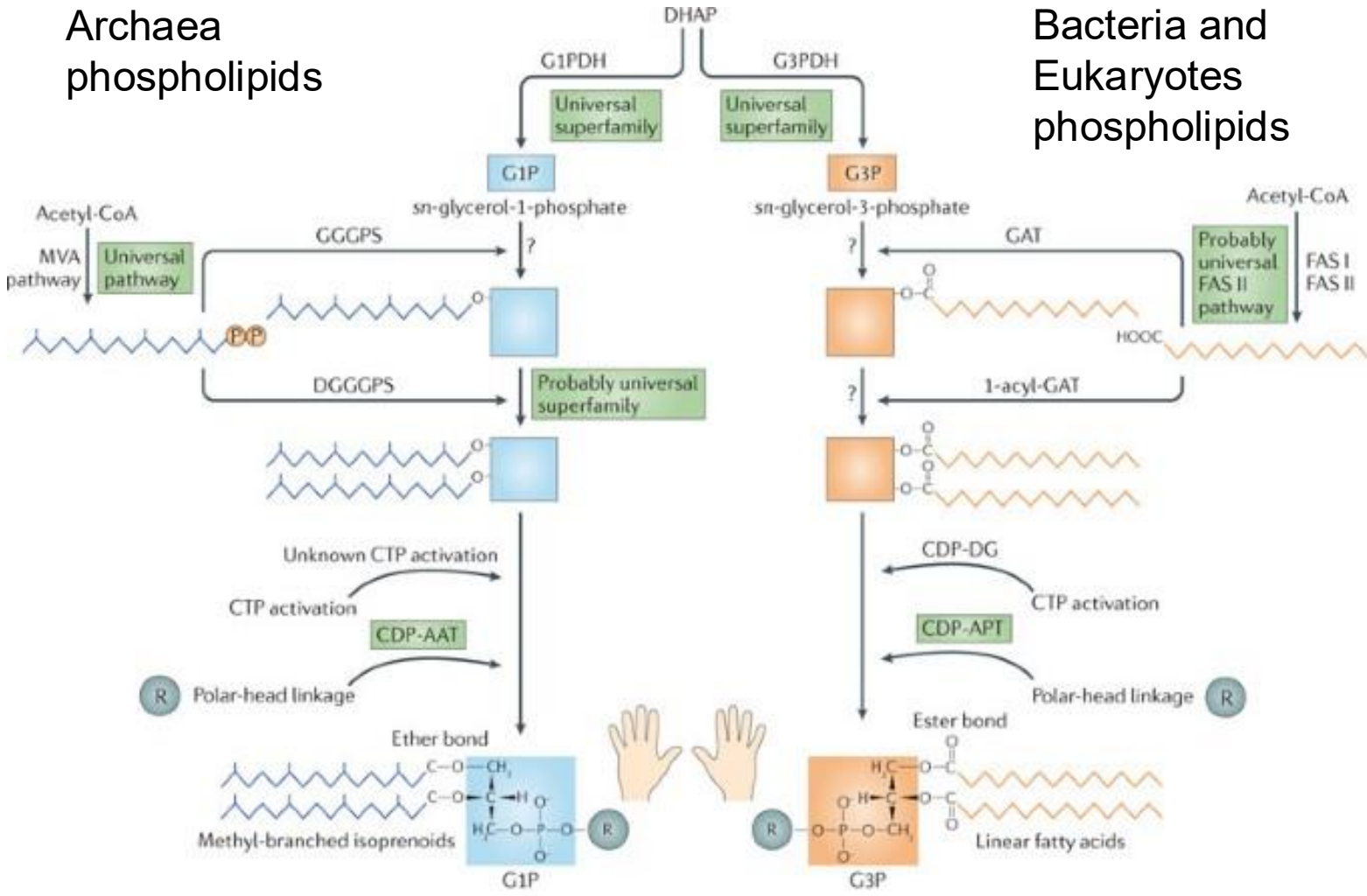
The chimeric nature of eukaryotes



Eukaryotic lipids resemble bacterial ones

Archaea phospholipids

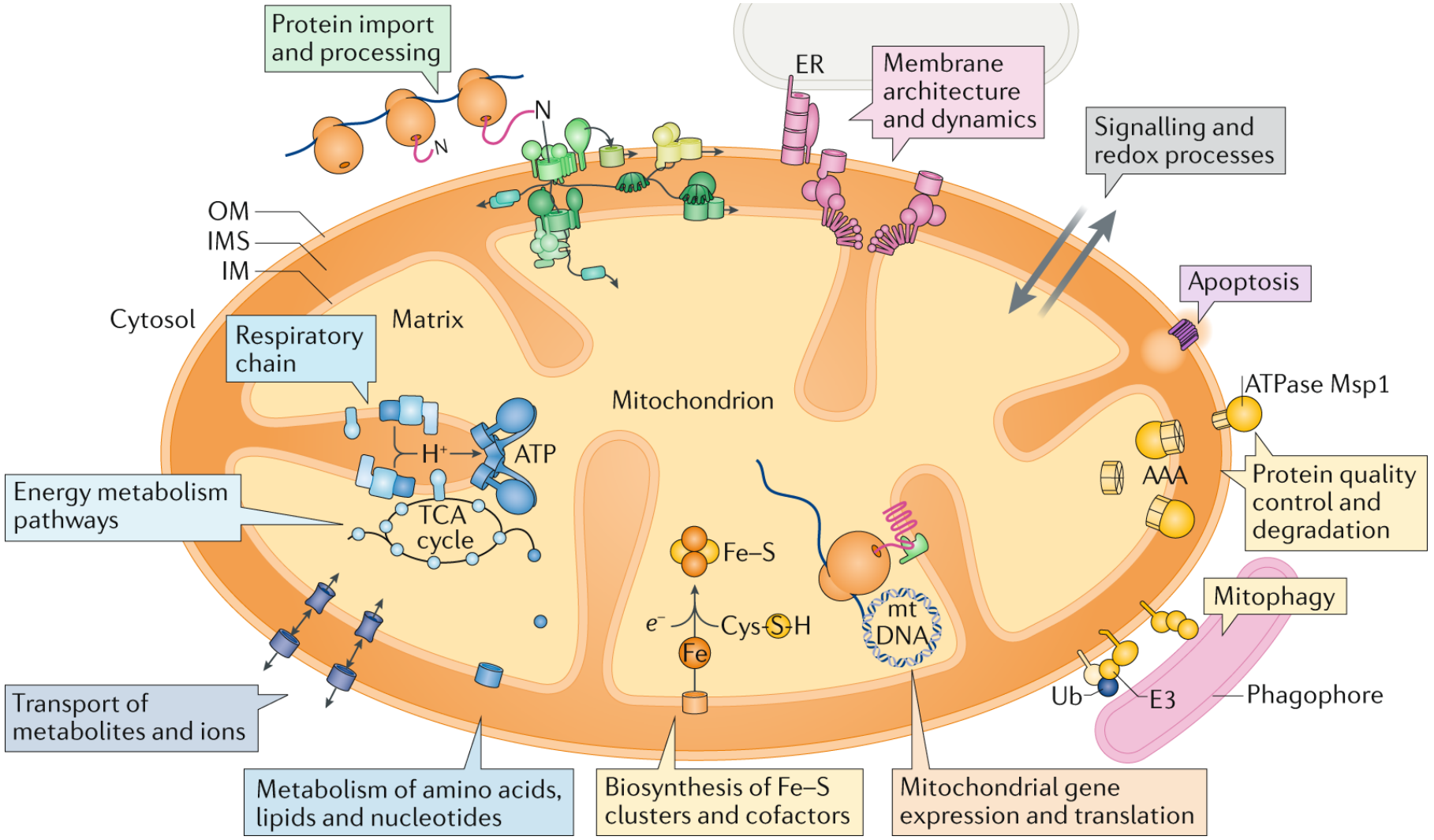
Bacteria and Eukaryotes phospholipids



Nature Reviews | Microbiology

Lombard et al, 2012

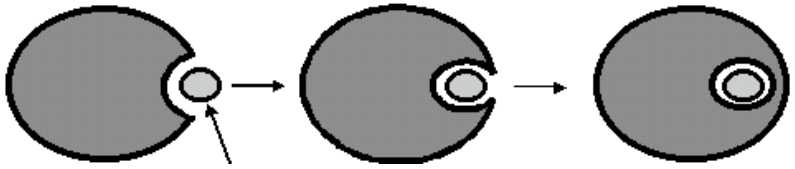
Mitochondria have diverse and crucial metabolic roles for the eukaryotic cell

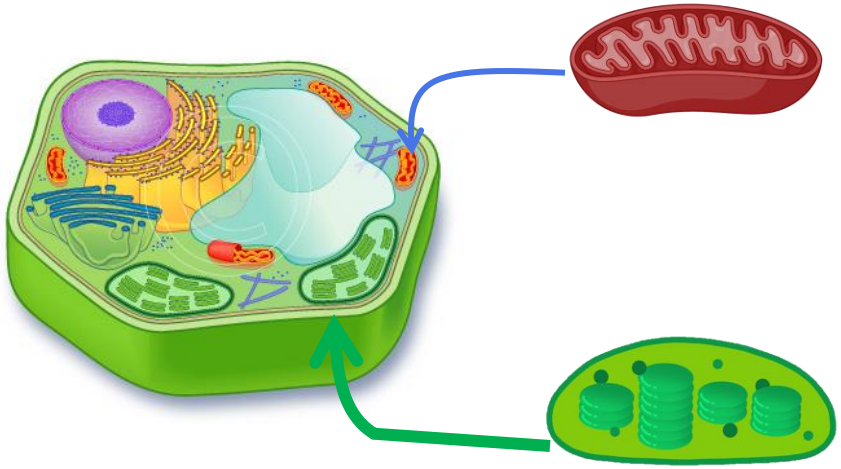


The notion that eukaryotes are chimeric in nature is not new

Lynn Margulis
(1938-2011)

- Endosymbiont theory: Mitochondria and chloroplasts were once free-living bacteria
- First proposed by Altmann (1890) and Mereschkowsky (1905), developed into modern form by Lynn Margulis

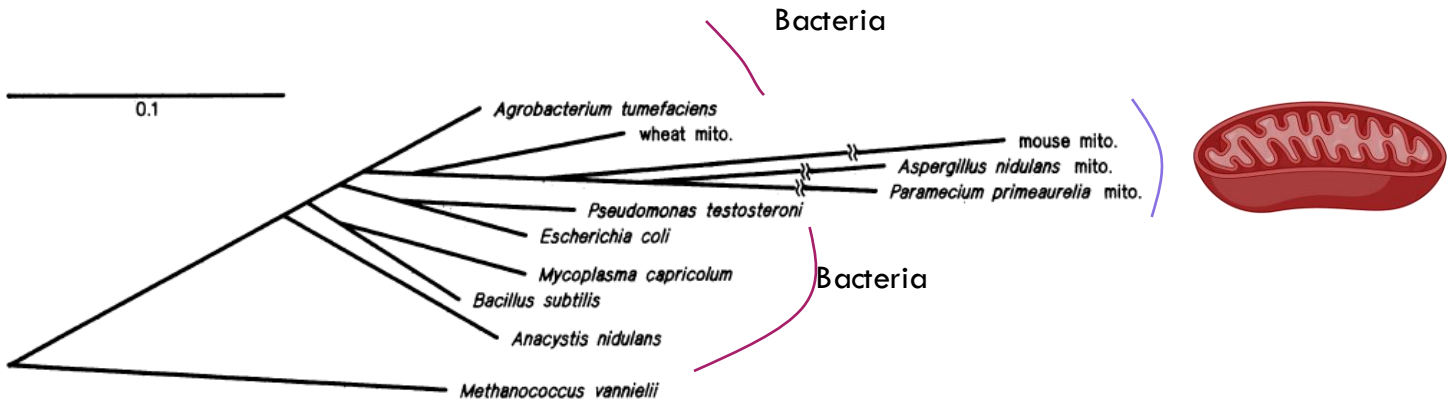




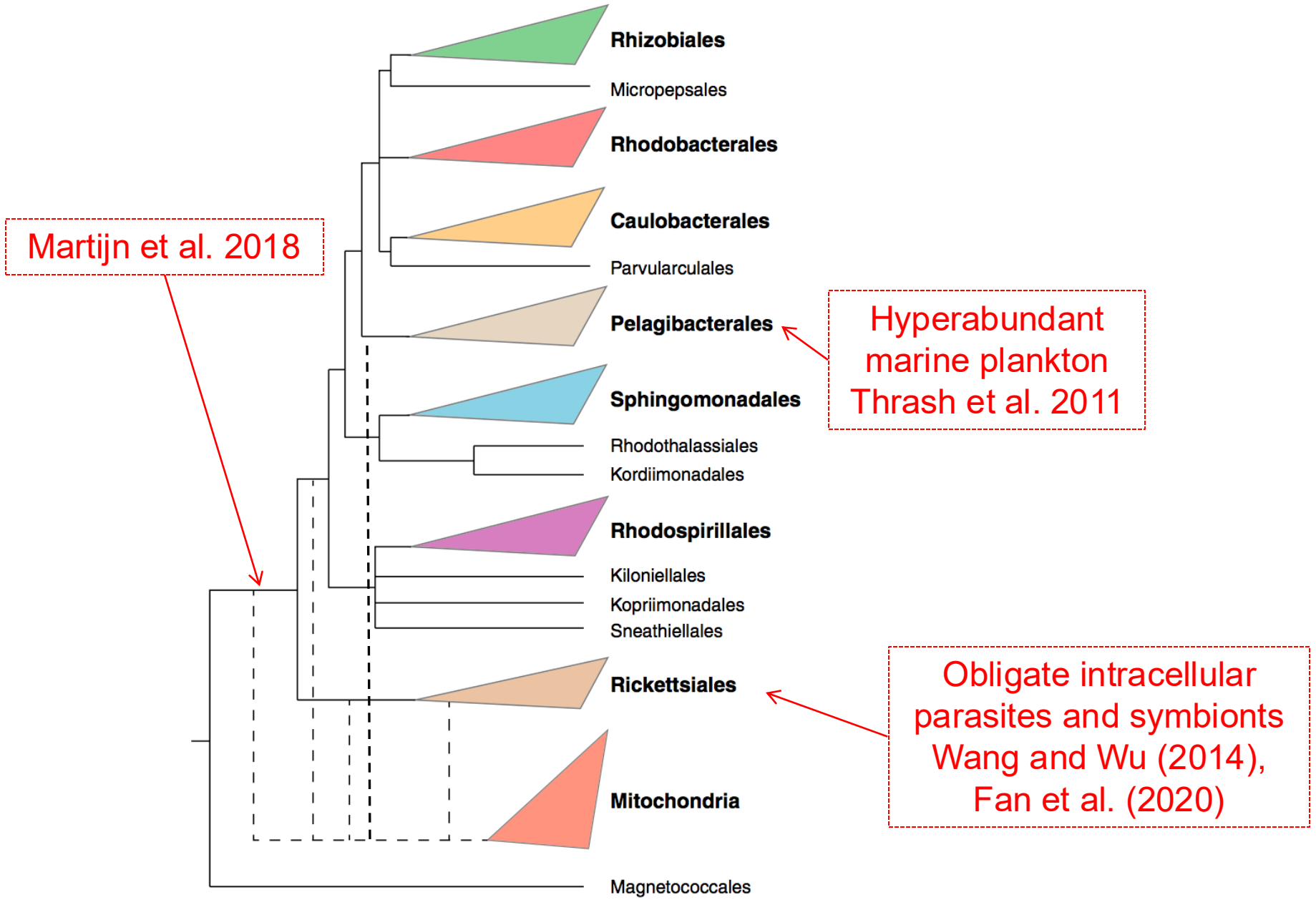
The notion that eukaryotes are chimeric in nature is not new

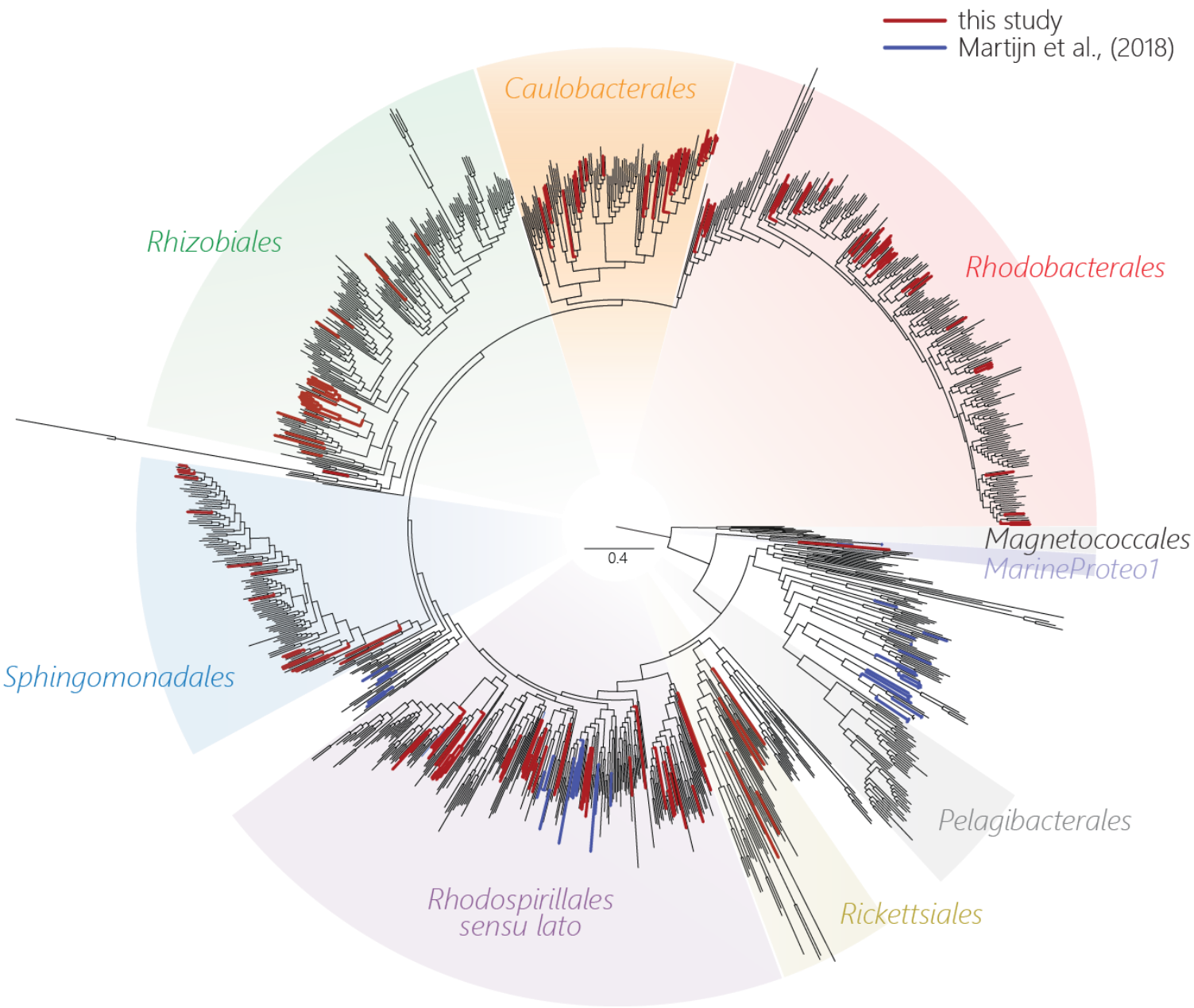
Lynn Margulis
(1938-2011)

- Endosymbiont theory: Mitochondria and chloroplasts were once free-living bacteria
- First proposed by Altmann (1890) and Mereschkowsky (1905), developed into modern form by Lynn Margulis



Controversy over position of mitochondria within Alphaproteobacteria





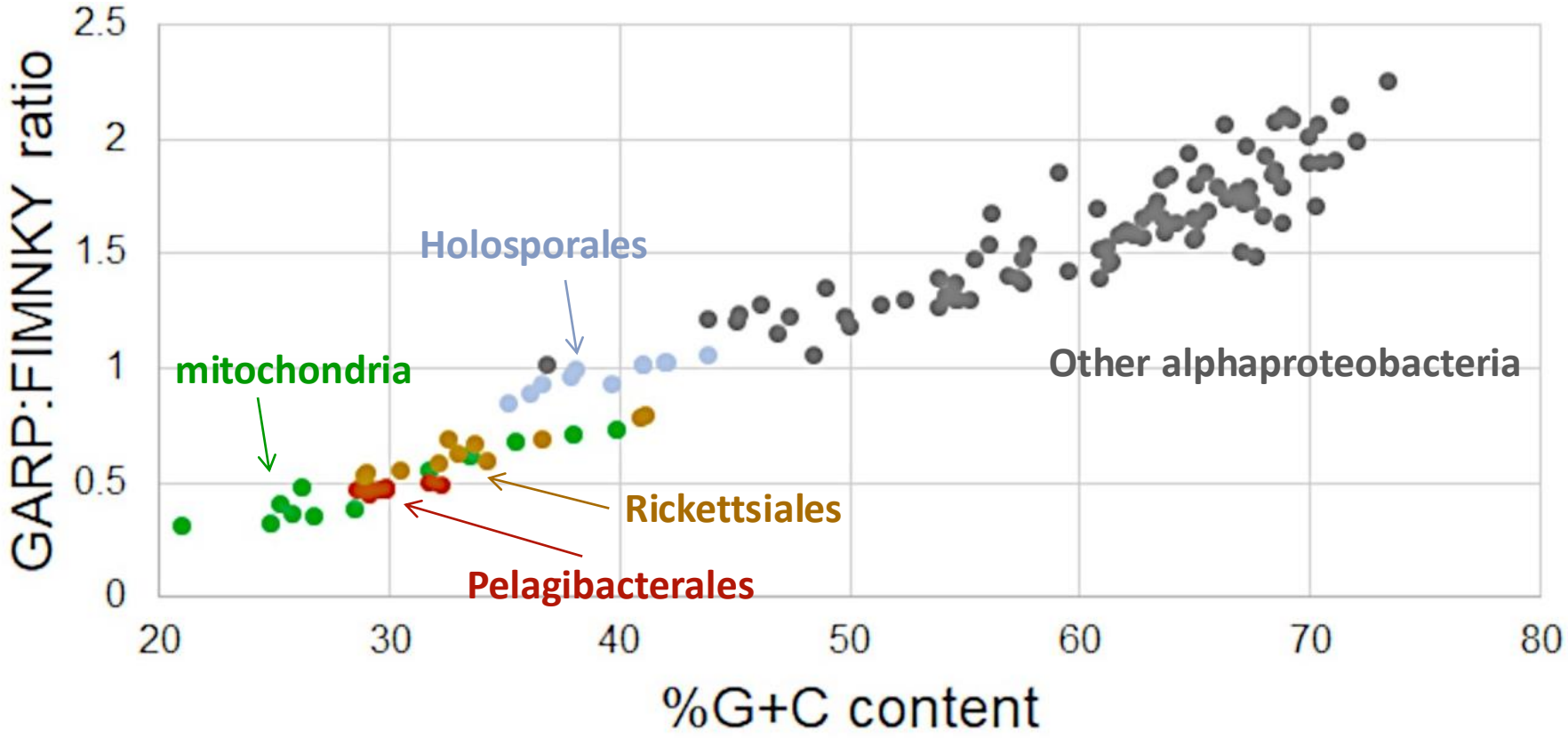
- More markers:
108 proteins of mitochondrial origin

- More taxa:
150 non-marine alphaproteobacterial MAGs (microbial mats, microbialites and lake sediments)

- New model:
GFmix phylogenetic model

Muñoz-Gómez, Susko, Williamson, Eme, … Roger (2022) *Nat. Ecol. Evol.*

The proteome AA composition varies with the genomic GC content



A site-and-branch-heterogeneous profile mixture model GFmix

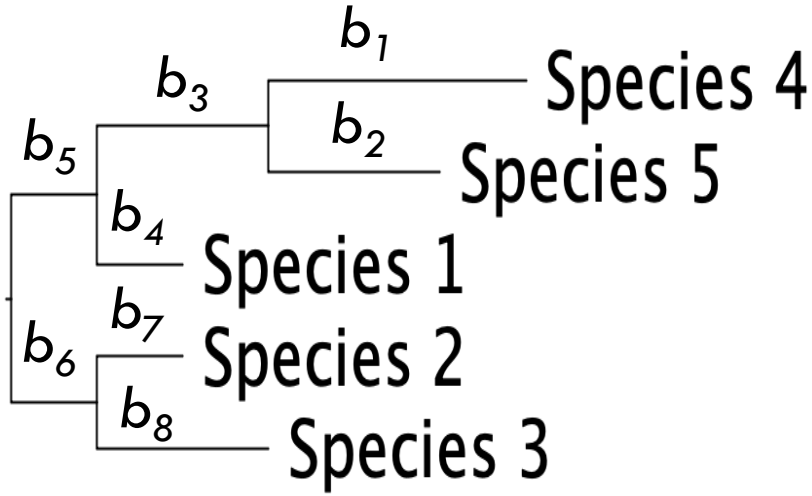
Edward Susko

For each branch assume there's a specific ratio of frequencies of GARP:FYMINK

- call this the 'b' parameter

$$b_x = \frac{f_{G,A,R,P}}{f_{F,I,M,N,K,Y}}$$

Then for the tree we have:

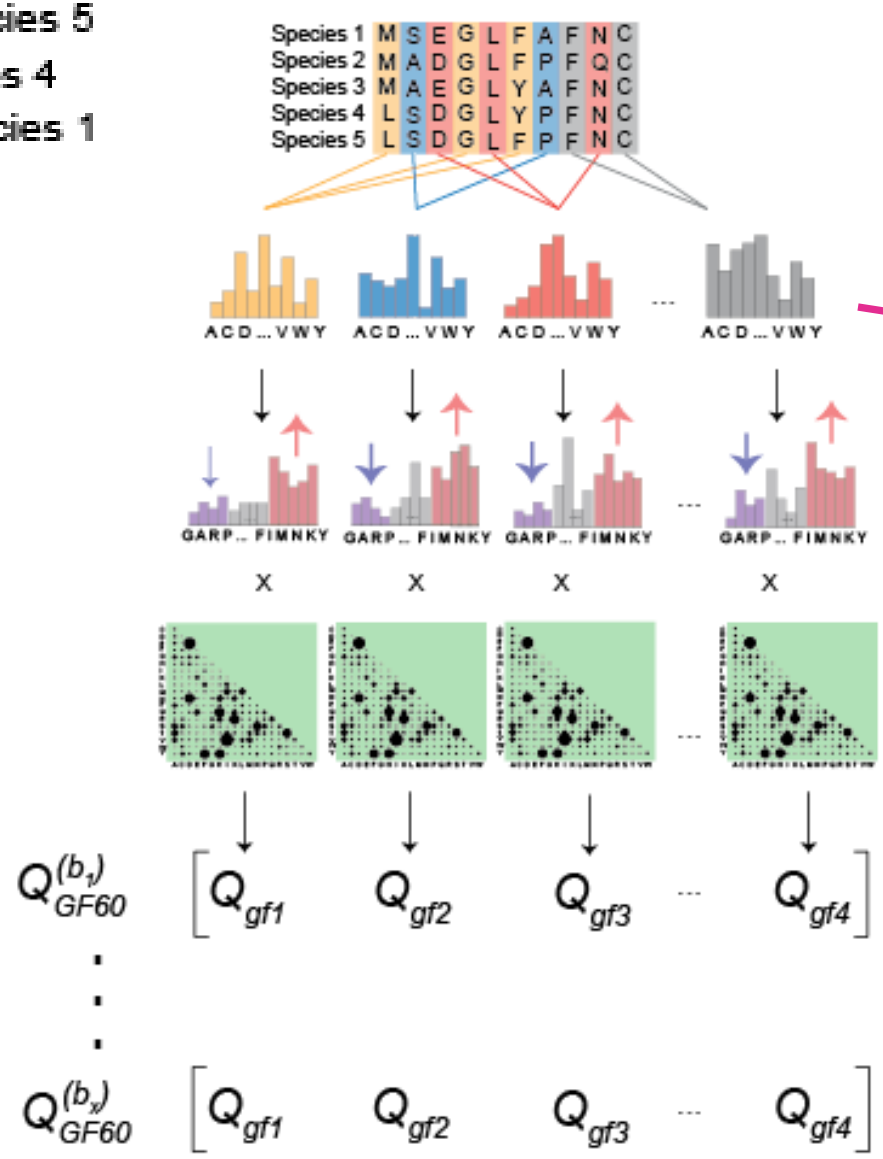


GFmix modifies the site profile mixture classes for each branch based on the *b* parameter

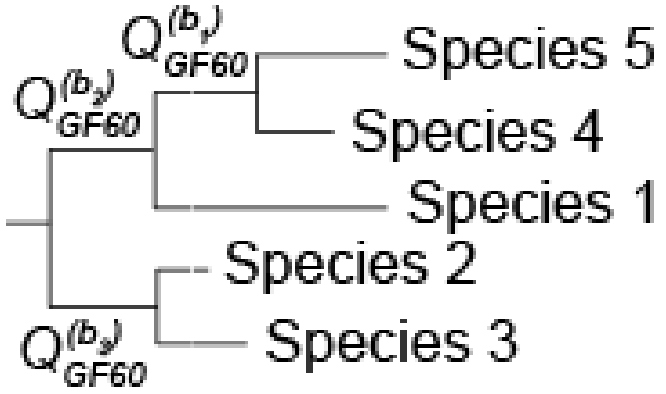
Muñoz-Gómez et al. (2022) *Nat. Ecol. Evol.*

A site-and-branch-heterogeneous profile mixture model GFmix

Edward Susko

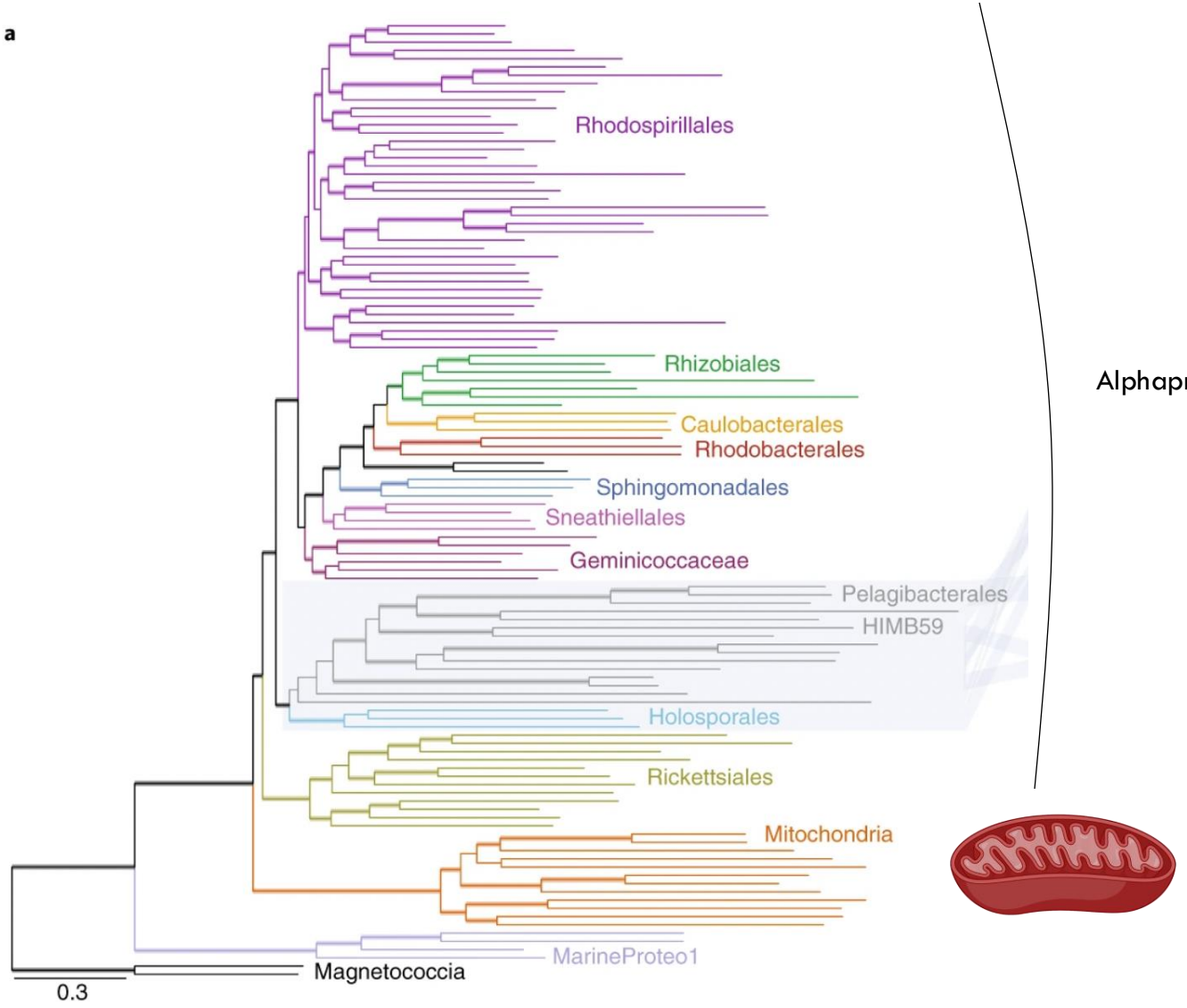


$$\frac{f_{G,A,R,P}}{f_{F,I,M,N,K,Y}} = b \ll 1$$



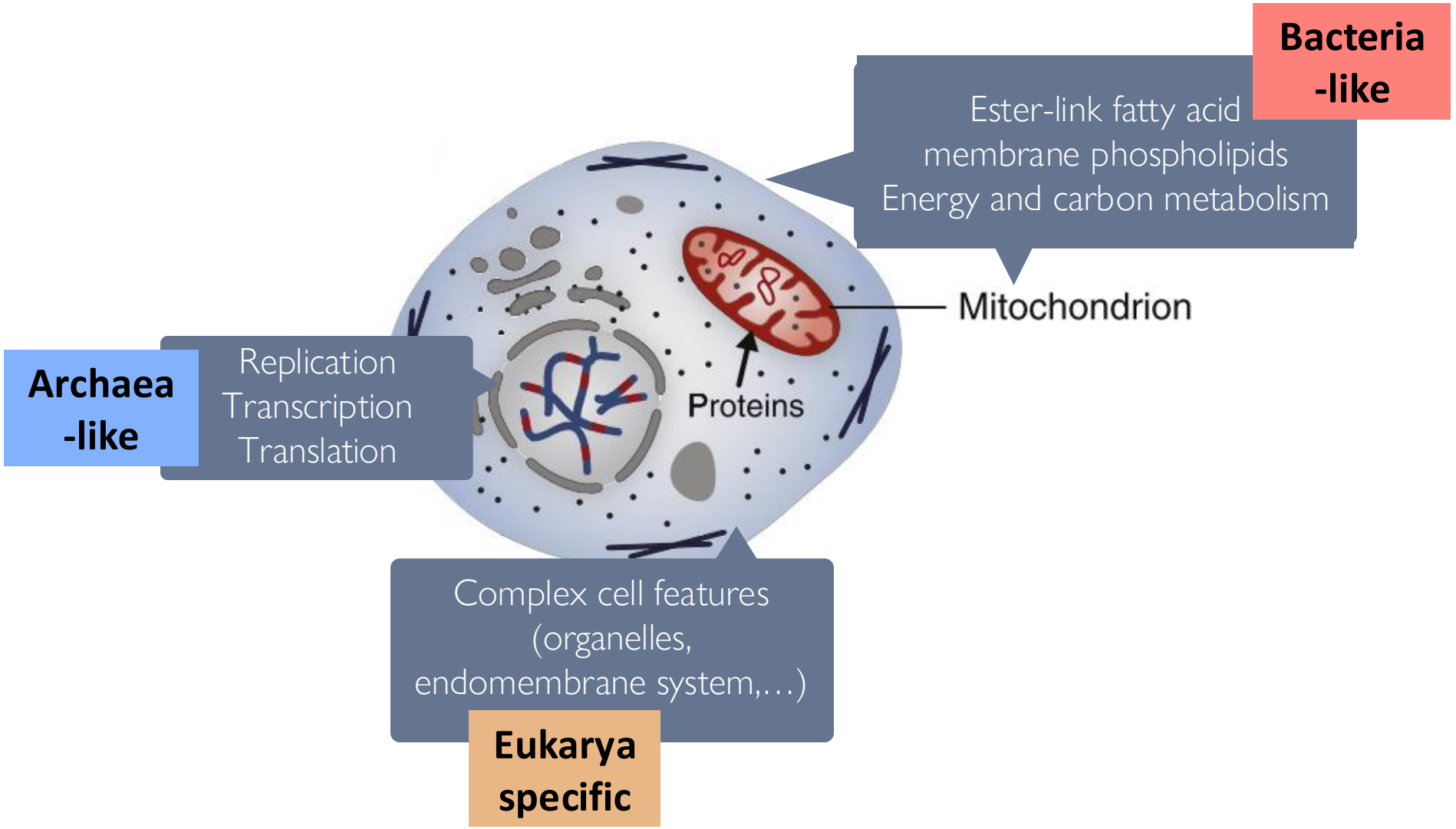
Muñoz-Gómez et al. (2022) *Nat. Ecol. Evol.*

The GFmix model supports the ancestry of mitochondria from outside known diversity of alphaproteobacteria



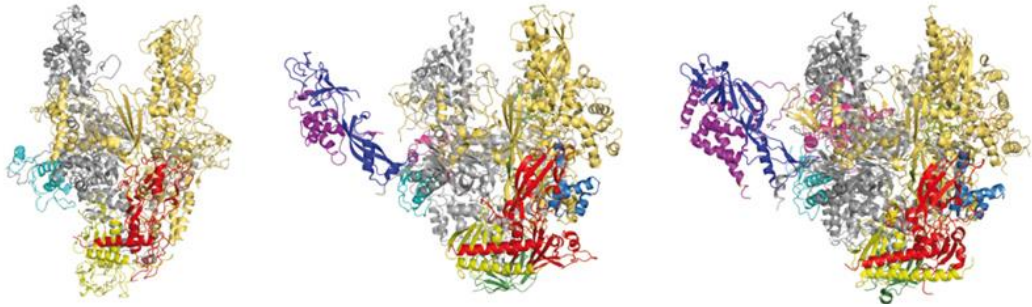
Munoz-Gomez et al., Nature Ecol Evol 2022

The chimeric nature of eukaryotes

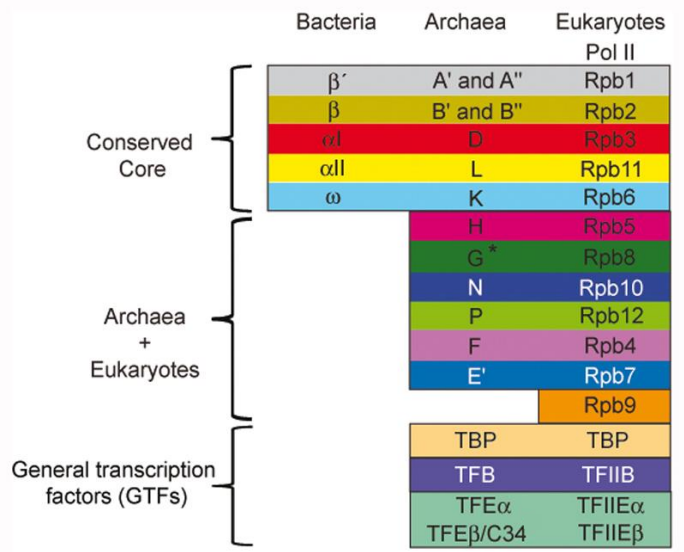


Informational systems suggest an archaeal connection

Transmission and expression of genetic information show a higher similarity between eukaryotes and Archaea than with Bacteria

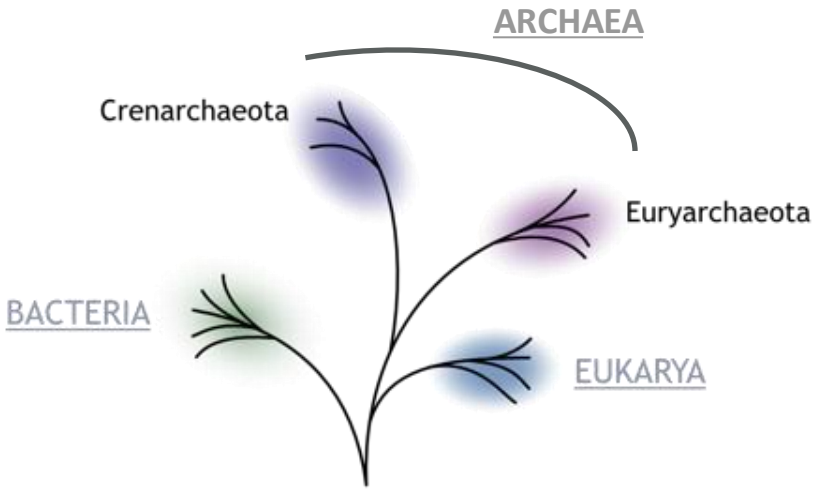


Overall architecture of RNA polymerases (RNAPs)

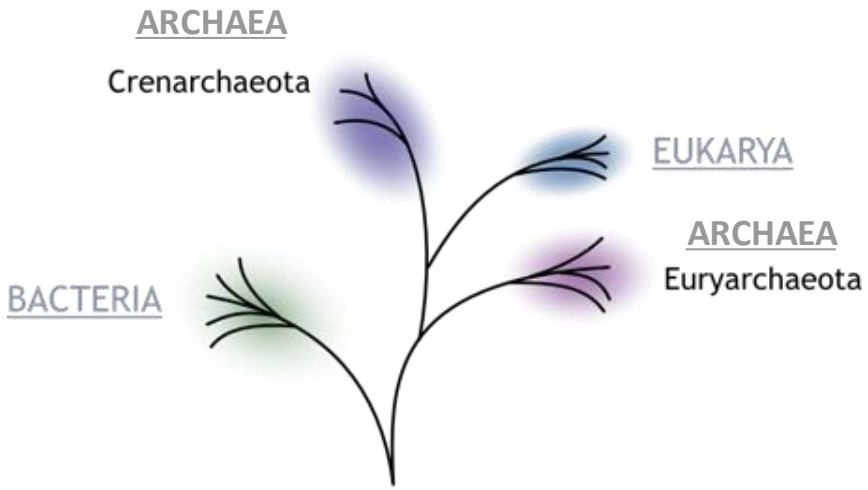


Subunit composition of the RNAPs

Archaea as sister-group or as ancestors of eukaryotes?



Three domain tree of Life



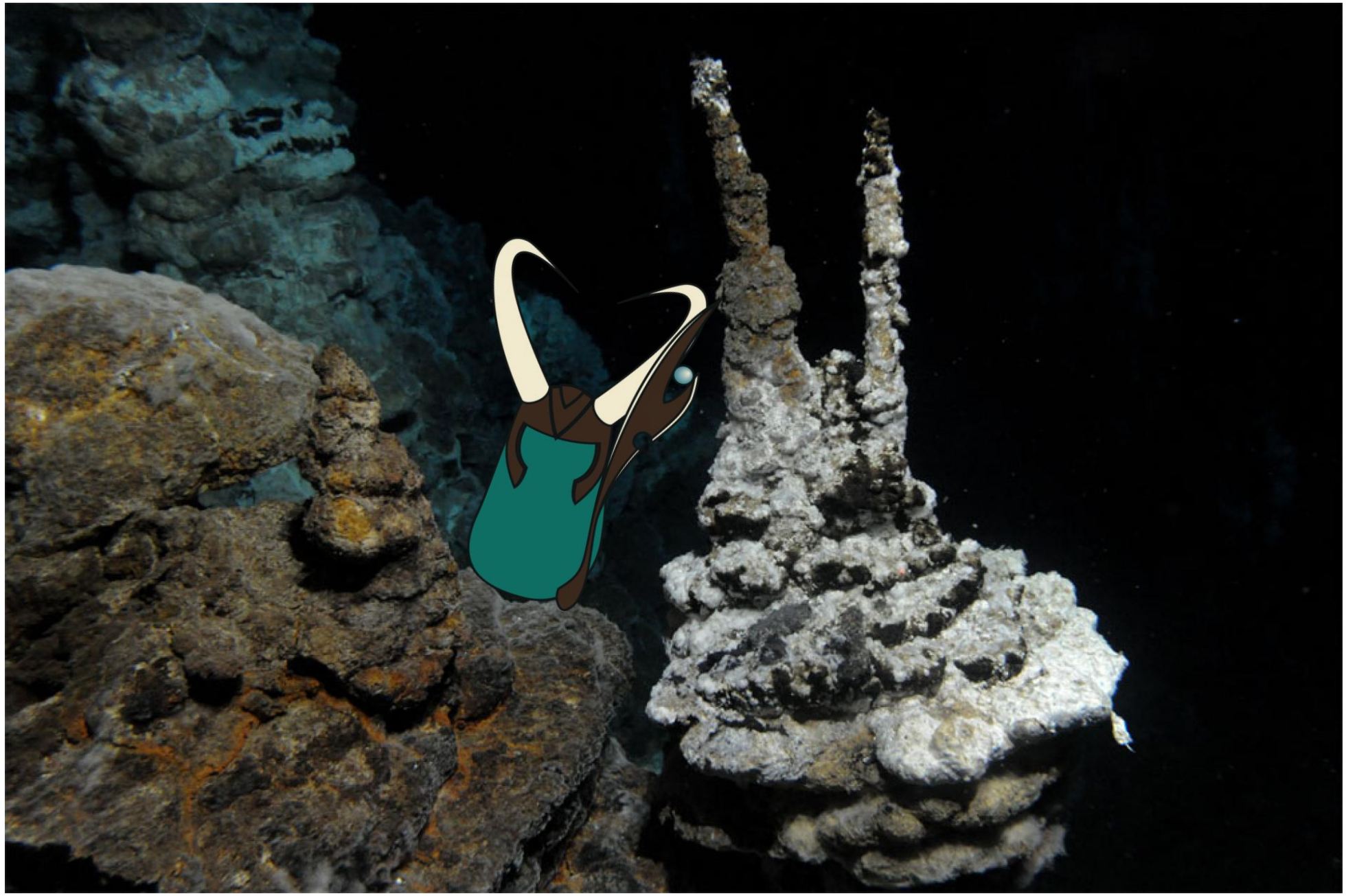
Two domain tree of Life

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

Eme et al. 2017, *Nat Rev Micro.*

Culture-independent genomics (e.g. metagenomics)

Spang et al. 2015, *Nature*

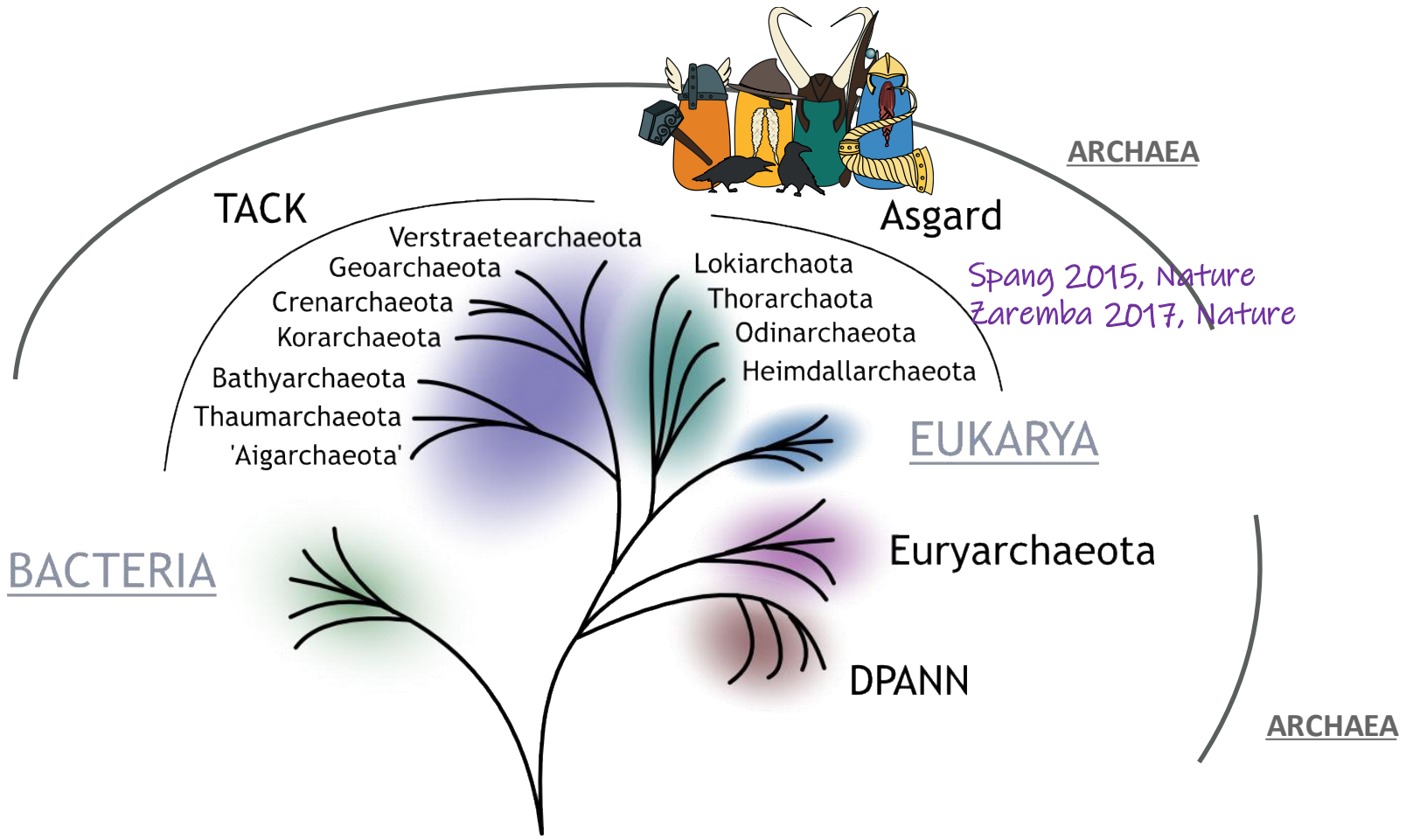


The unveiling of Asgard archaea through metagenomics

Asgard

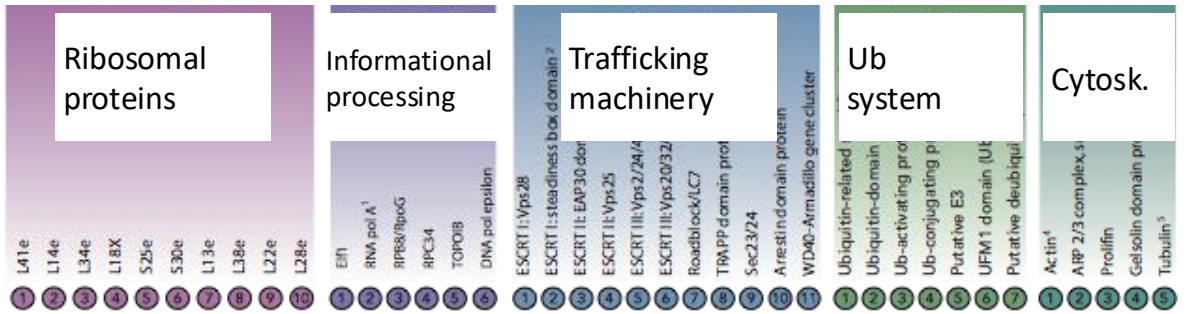
Spang 2015, Nature
Zaremba 2017, Nature

The unveiling of Asgard archaea through metagenomics



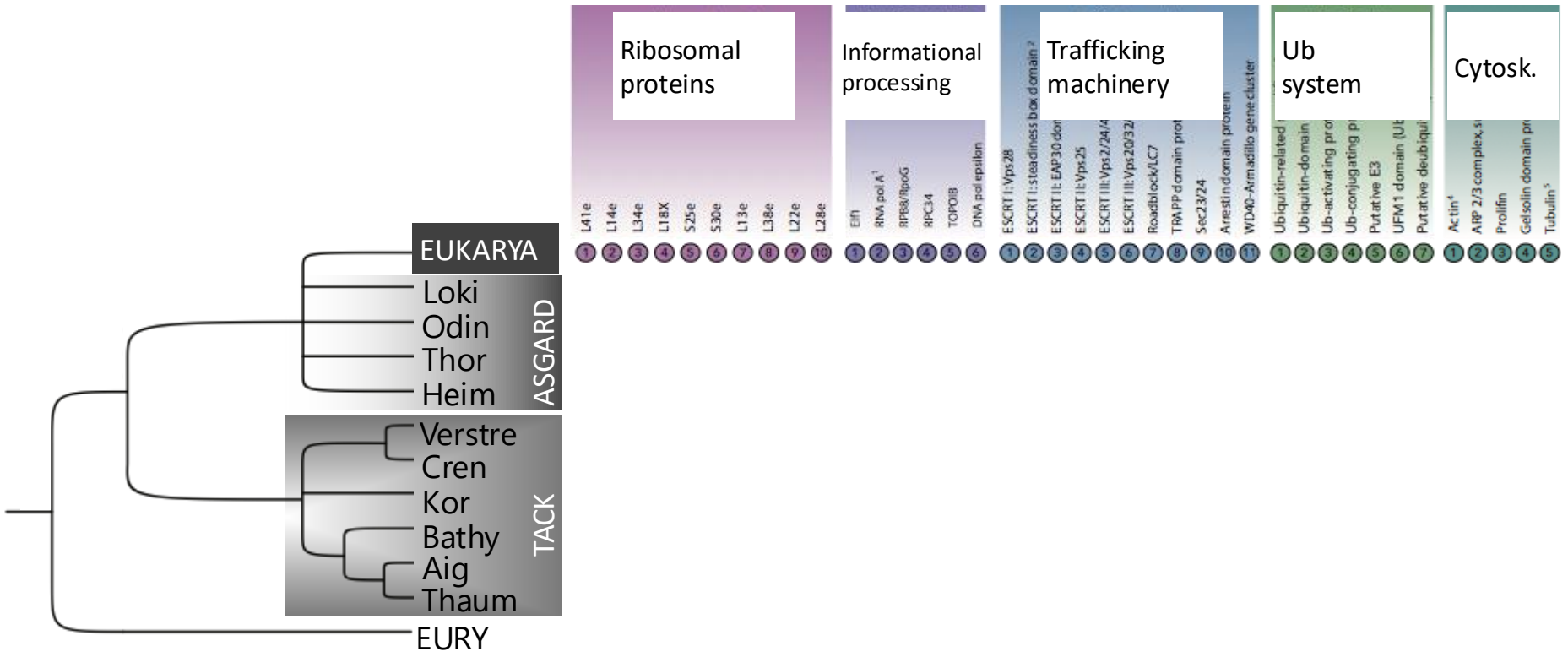
Eme et al. 2017, *Nat Rev Micro.*

Numerous Eukaryotic Signature Proteins (ESPs) in Asgard archaea



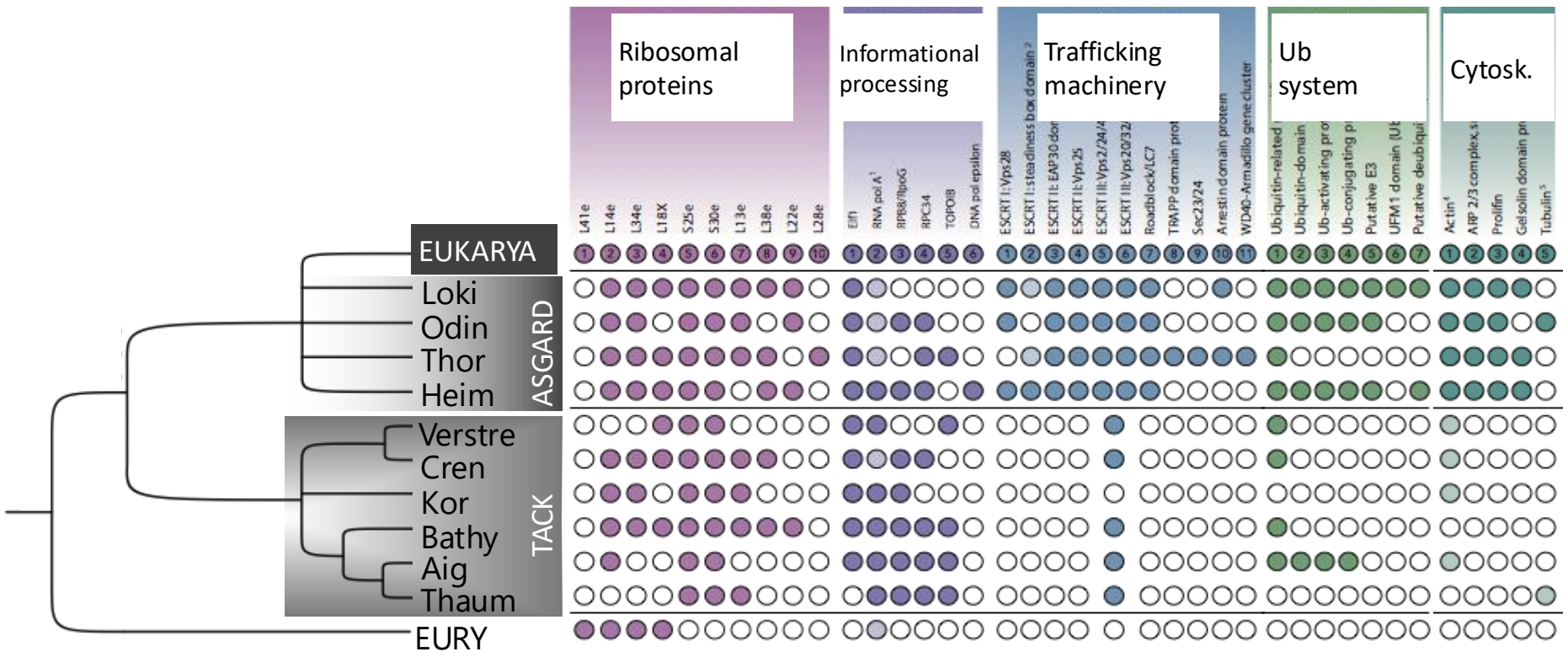
Eme et al. 2017, *Nat Rev Micro.*

Numerous Eukaryotic Signature Proteins (ESPs) in Asgard archaea



Eme et al. 2017, *Nat Rev Micro.*

Numerous Eukaryotic Signature Proteins (ESPs) in Asgard archaea



Eme et al. 2017, *Nat Rev Micro.*

Eme et al., *Nature* 2023

Article

Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes

https://doi.org/10.1038/s41586-023-06186-2

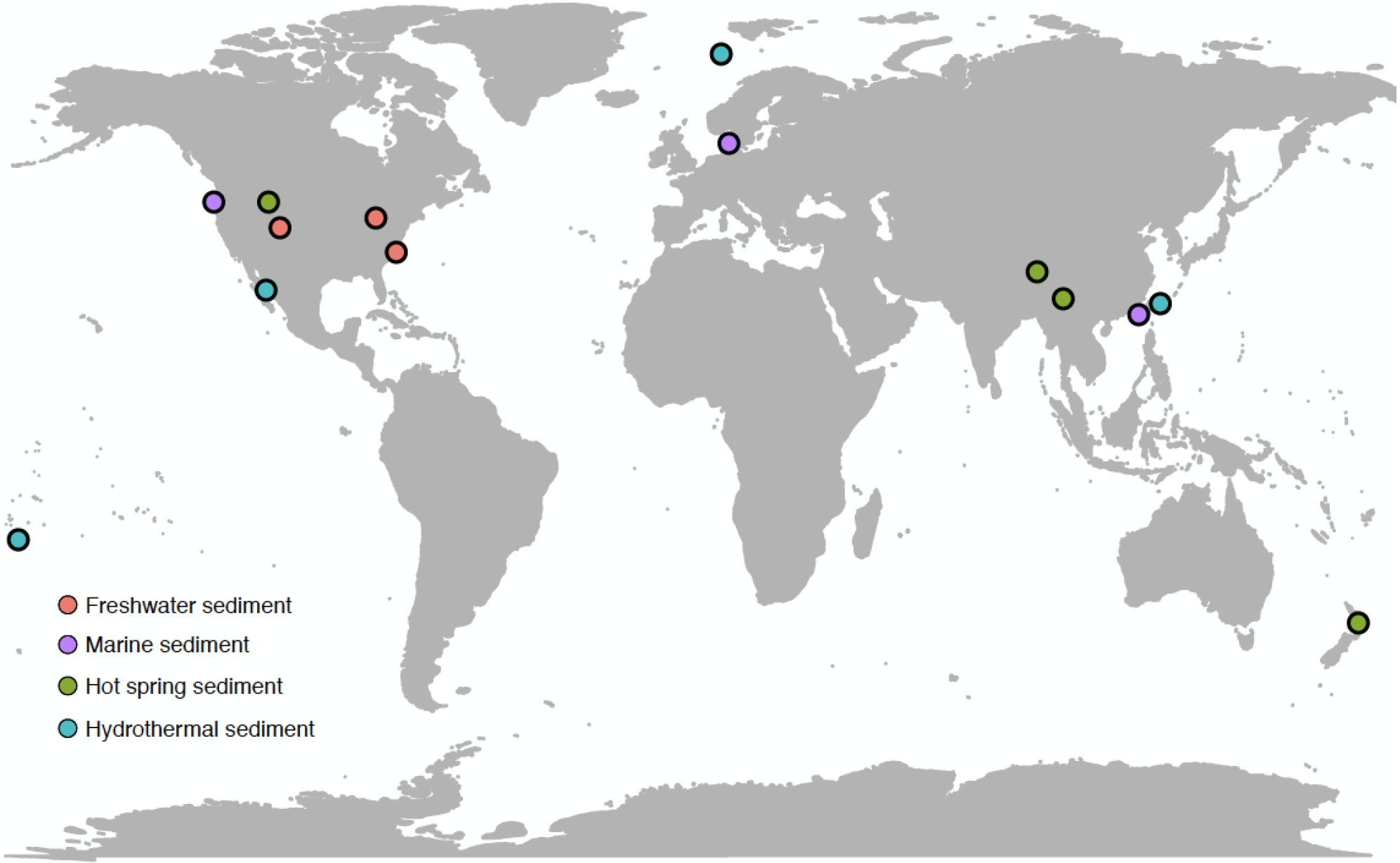
Received: 23 April 2021

Accepted: 10 May 2023

Open access

Laura Eme[1,2,21], Daniel Tamarit[1,3,4,15,21], Eva F. Caceres[1,3,21], Courtney W. Stairs[1,16], Valerie De Anda[5], Max E. Schön[1], Kiley W. Seitz[5,17], Nina Dombrowski[5,18], William H. Lewis[1,3,19], Felix Homa[3], Jimmy H. Saw[1,20], Jonathan Lombard[1], Takuro Nunoura[6], Wen-Jun Li[7], Zheng-Shuang Hua[8], Lin-Xing Chen[9], Jillian F. Banfield[9,10], Emily St John[11], Anna-Louise Reysenbach[11], Matthew B. Stott[12], Andreas Schramm[13], Kasper U. Kjeldsen[13], Andreas P. Teske[14], Brett J. Baker[5] & Thijs J. G. Ettema[1,3]✉

69 new Asgard genomes



Asgard genomes are substantially larger than most archaea

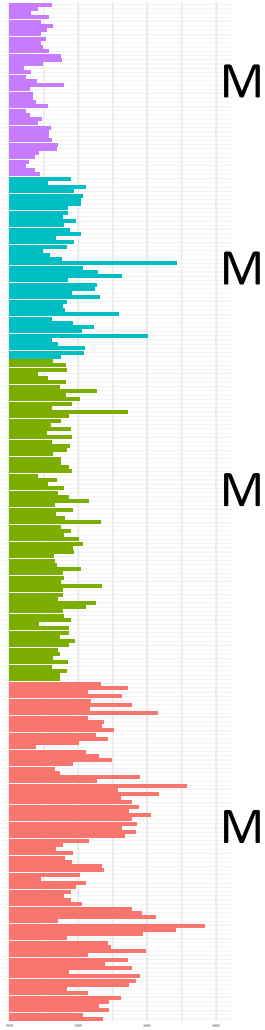
DPANN Median = 1.2 Mb

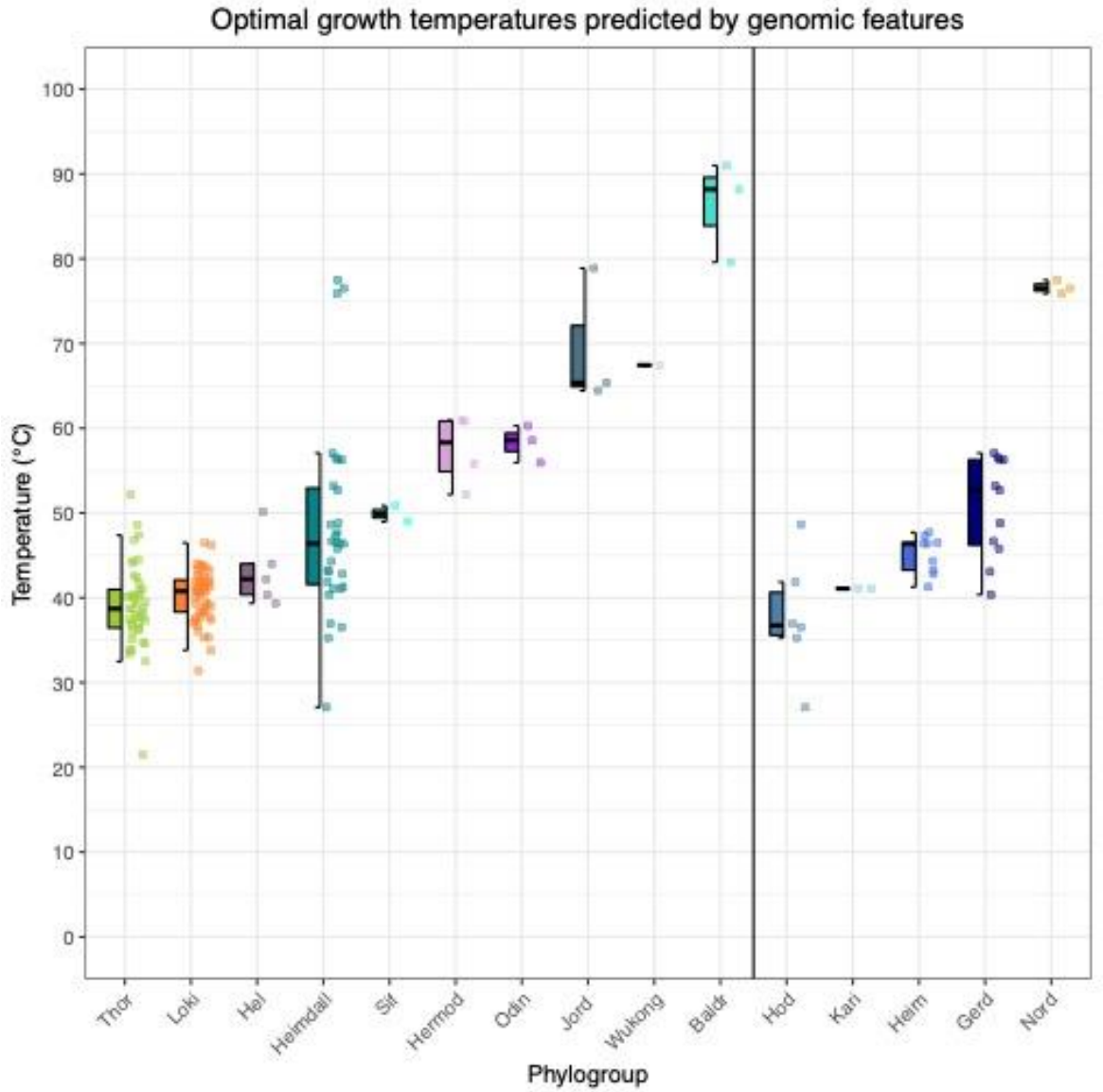
Euryarchaeota Median = 1.8 Mb

TACK Median = 1.8 Mb

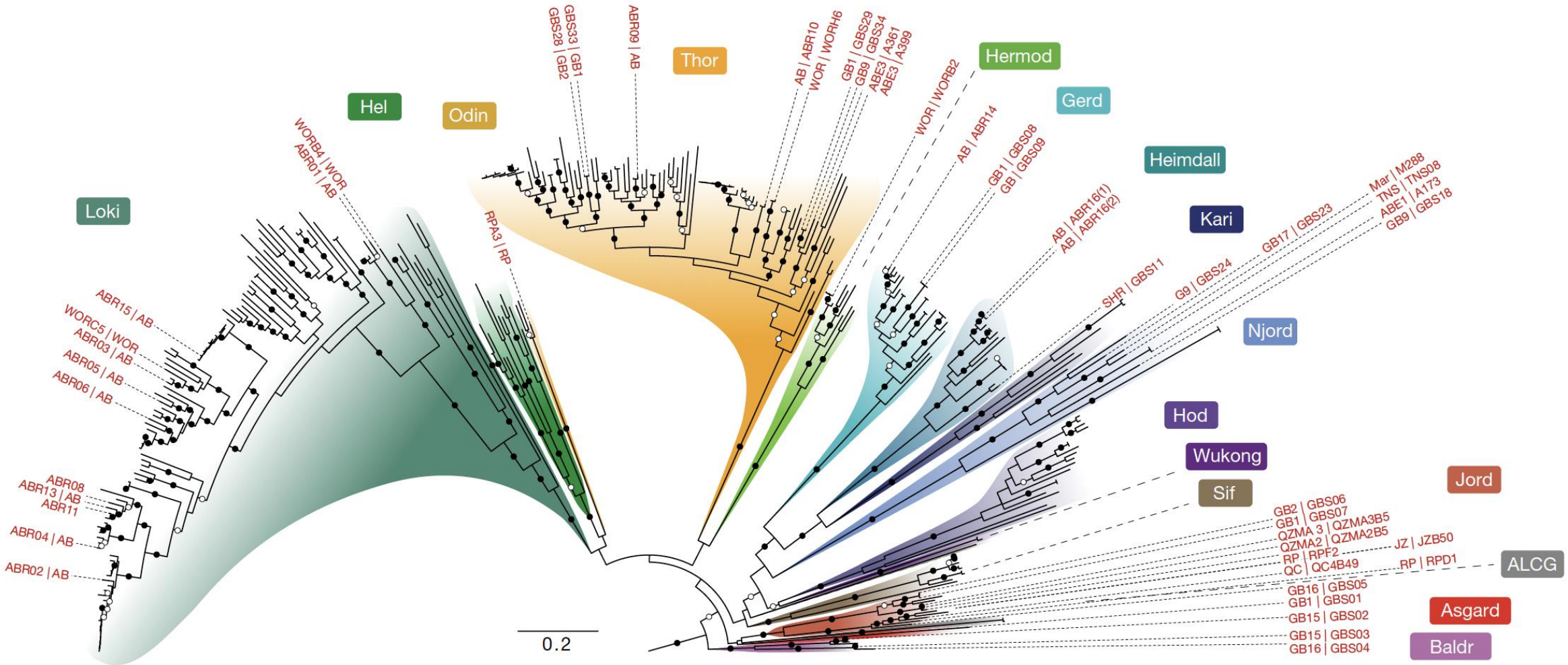
Asgard Median = 3.6 Mb

Min: 1.6 Mb
Max: 6.3 Mb





69 new Asgard MAGs



Phylogeny from a 15 ribosomal protein contig reveals new Asgard clades

How do Eukaryotes relate to Asgards?

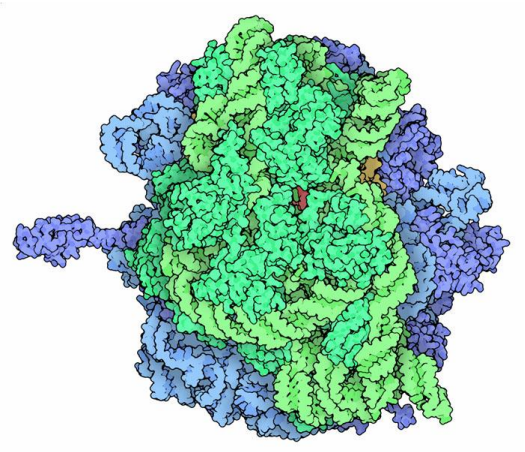
Ribosomal proteins:

- Slow evolving
- Universal
- Coevolve
- Short

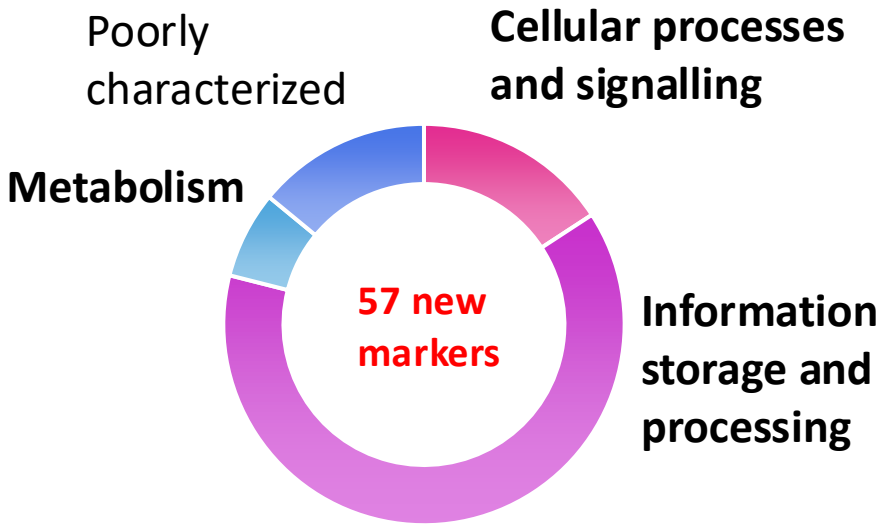
New markers:

- Conserved in eukaryote and archaea
- Not transferred horizontally

54 ribosomal proteins

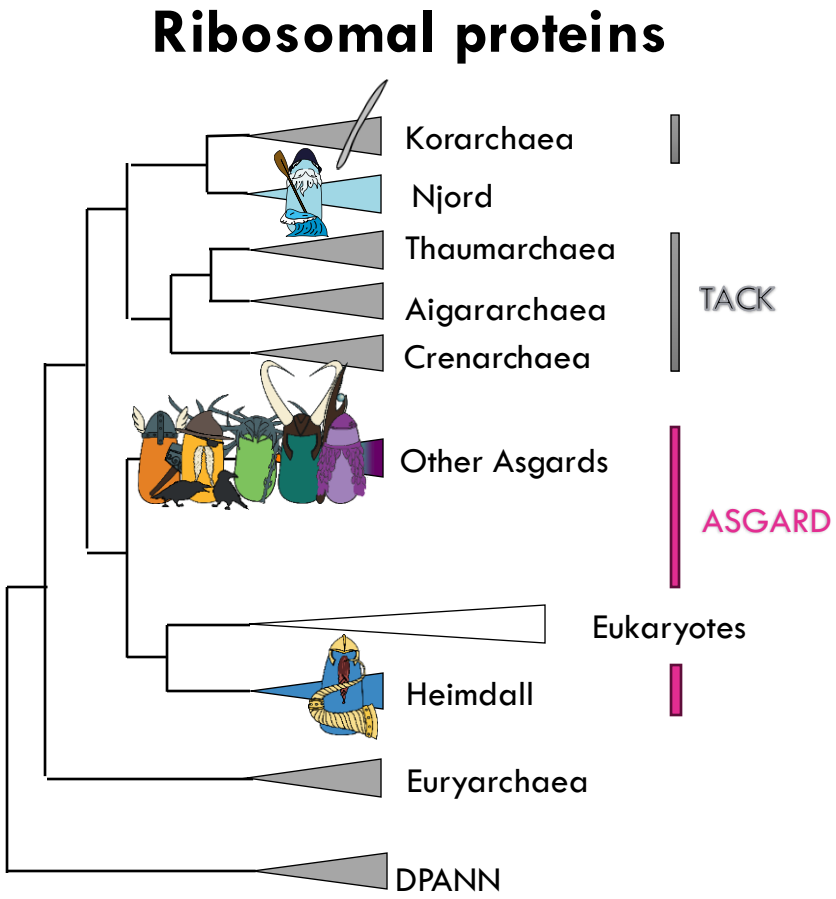


~6000 aa



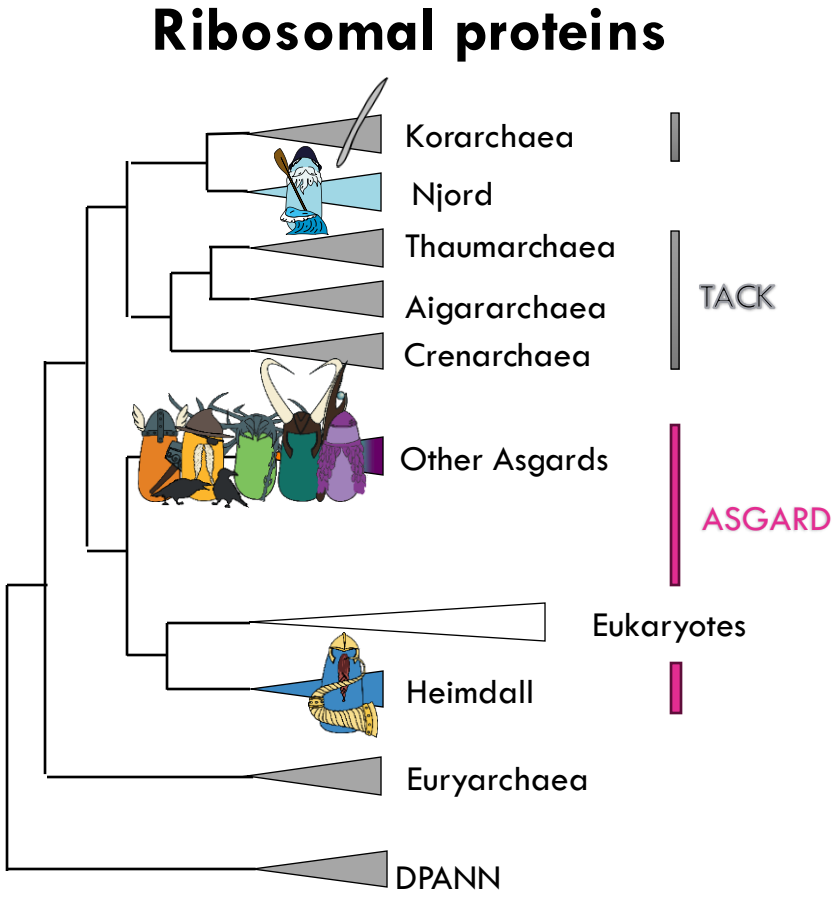
~14,000 aa

How do Eukaryotes relate to Asgards?



How do Eukaryotes relate to Asgards?

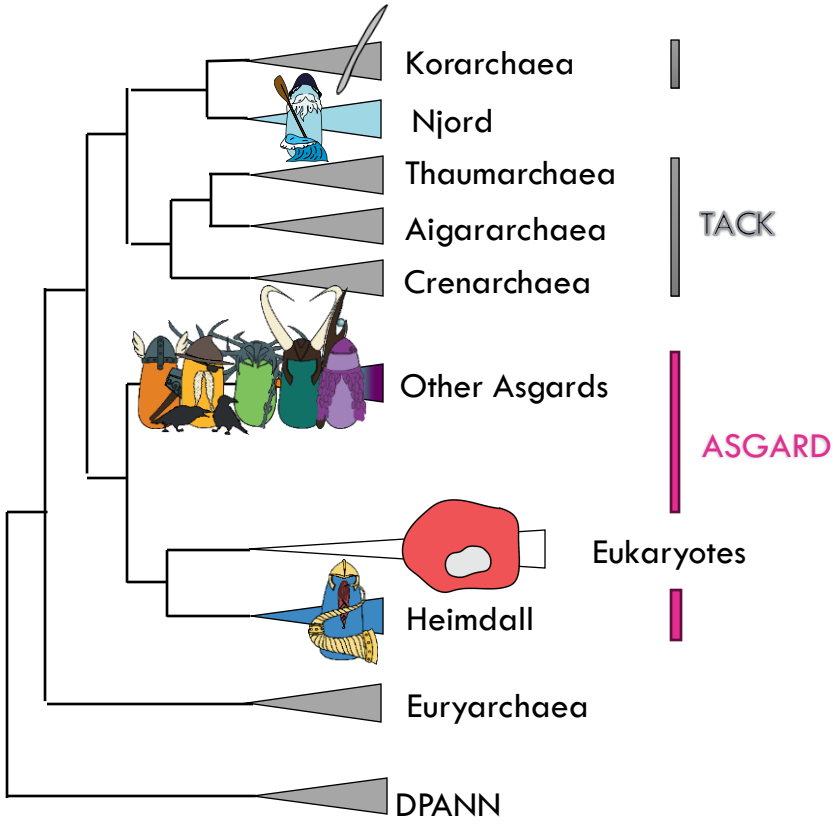
Many ESPs:
Actin,
RPL28e, …



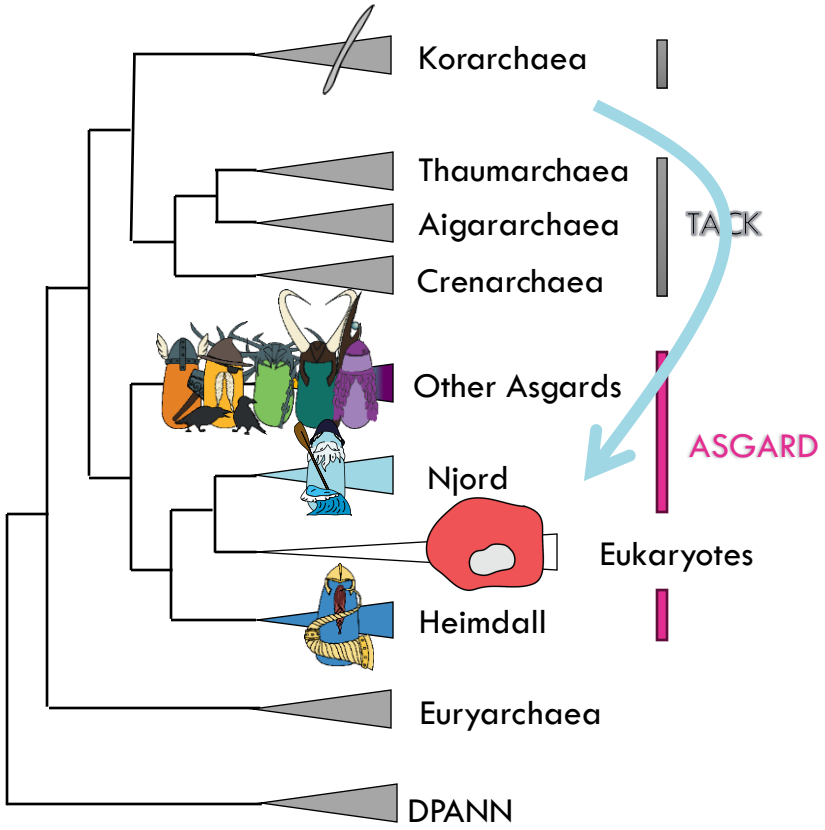
How do Eukaryotes relate to Asgards?

Many ESPs:
Actin,
RPL28e, …

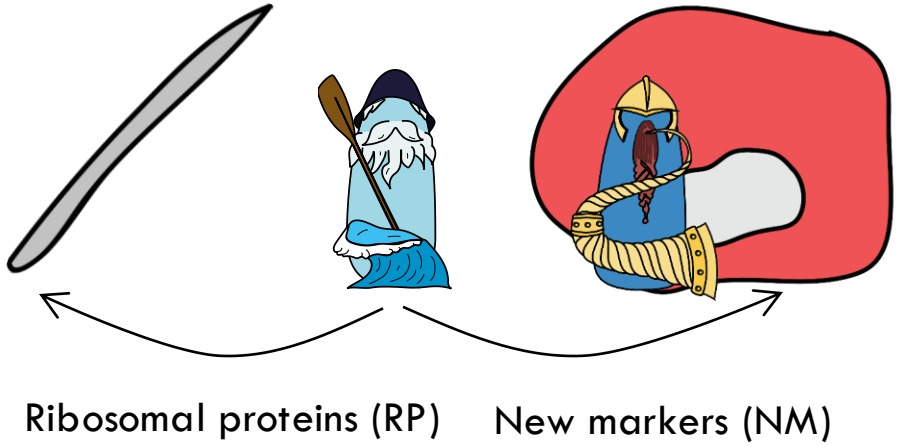
Ribosomal proteins



New markers



How do Eukaryotes relate to Asgards?



How do Eukaryotes relate to Asgards?

Dataset

- **Ribosomal proteins**
- **New markers**

Software

- **IQ-Tree (LG+G4+C60+F+PMSF)**
- **Phylobayes (CAT+LG+G4)**

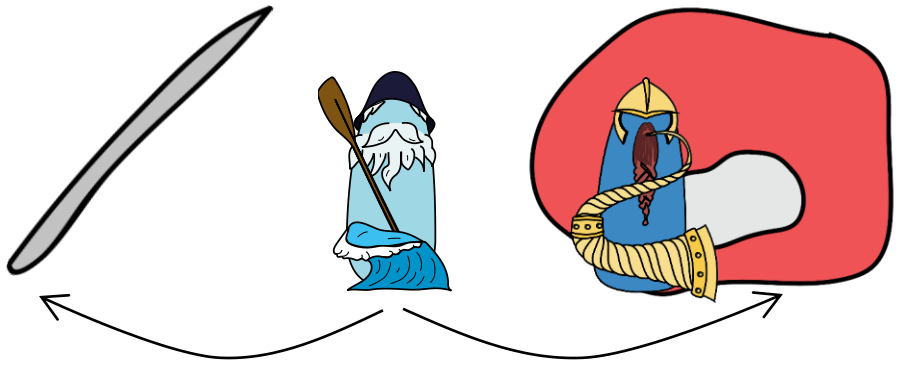
Taxon sampling

- **+/- DPANN**
- **+/- Eukaryotes**
- **+/- Korarchaea**

Fast-evolving site removal (FSR)

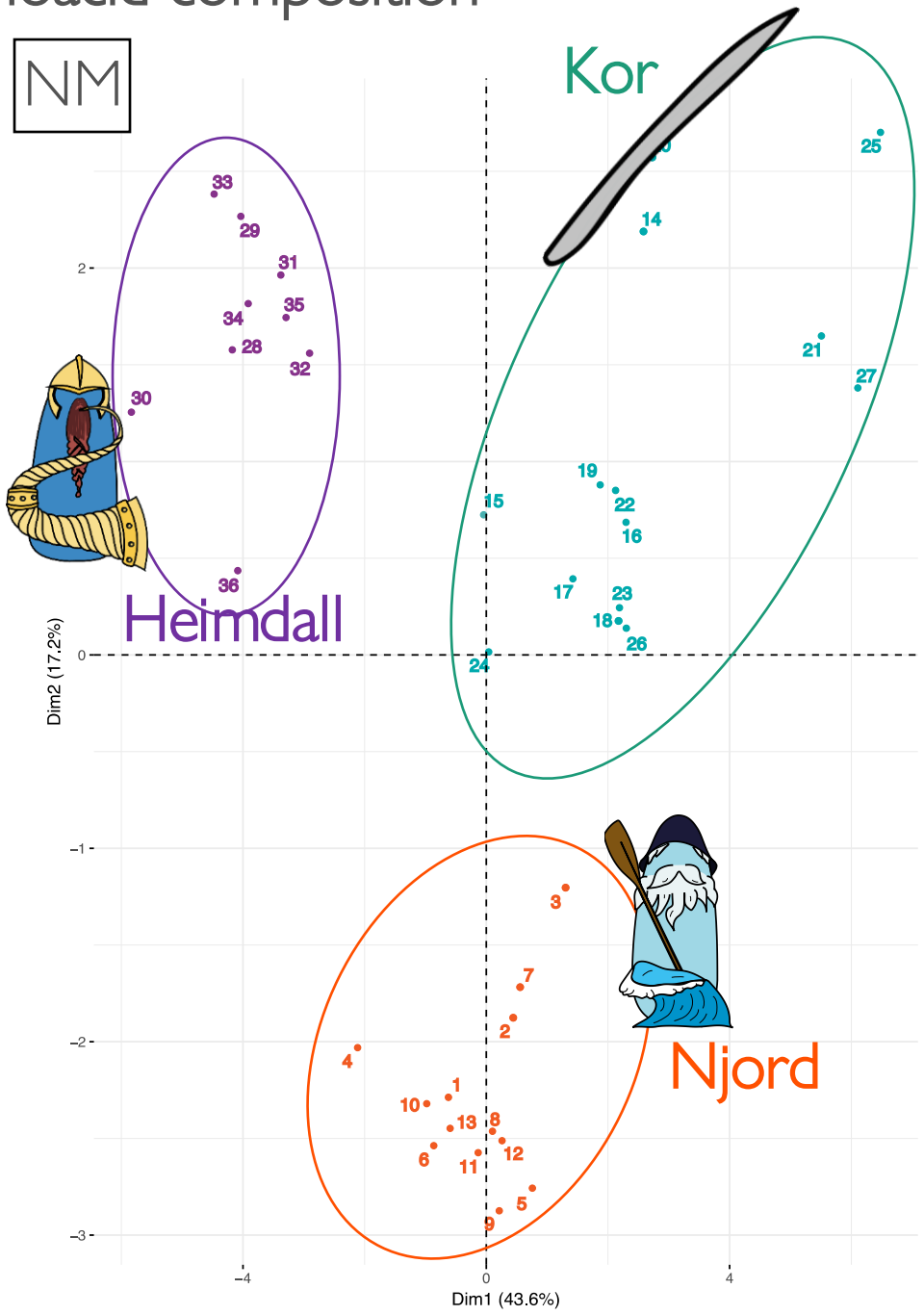
Recoding (SR4)

~800 phylogenies…

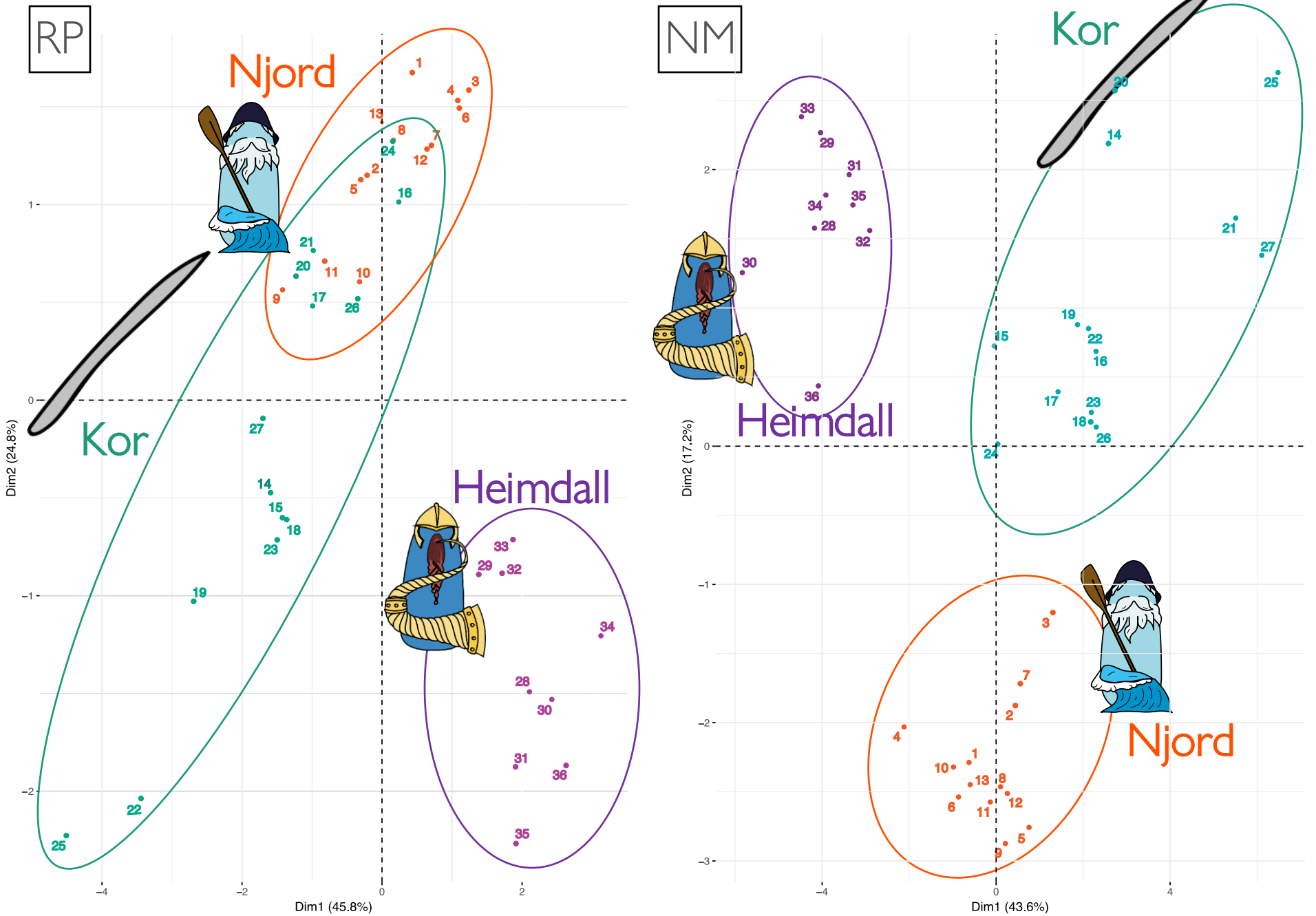


Ribosomal proteins (RP) New markers (NM)

PCA based on aminoacid composition

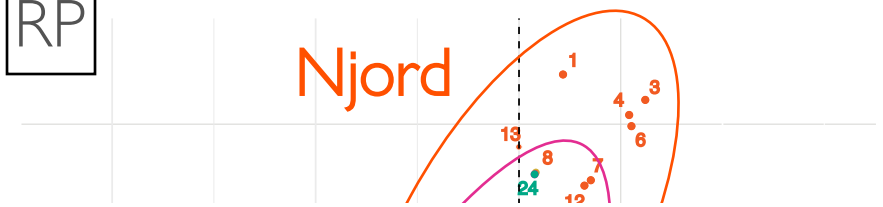


PCA based on aminoacid composition

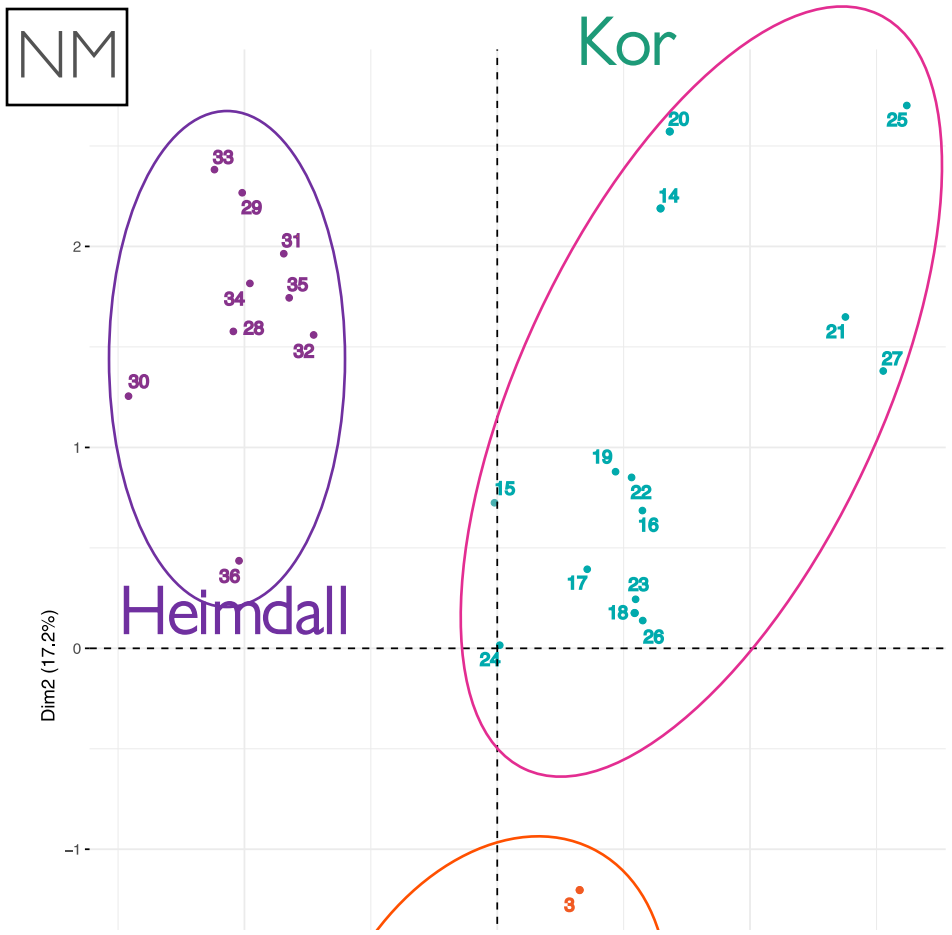


PCA based on aminoacid composition

RP



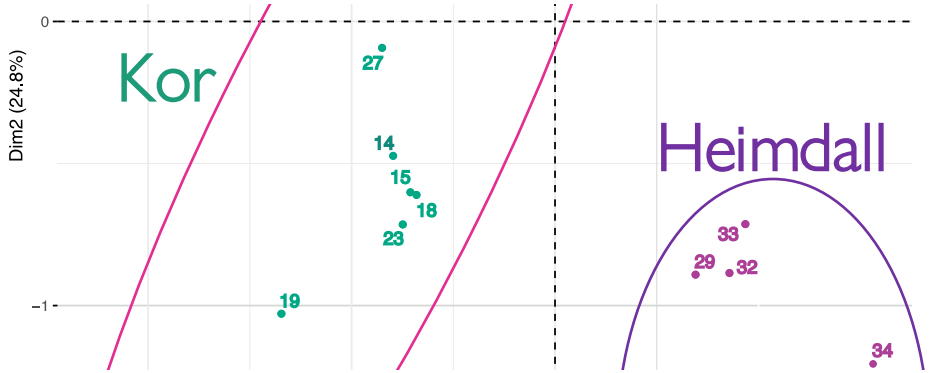
NM



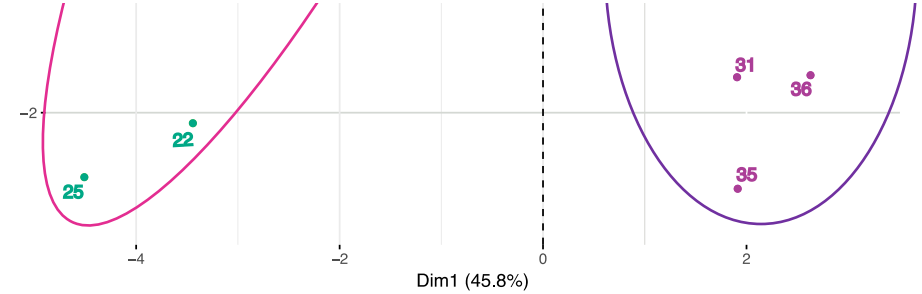
Njord and Kor hyperthermophiles:

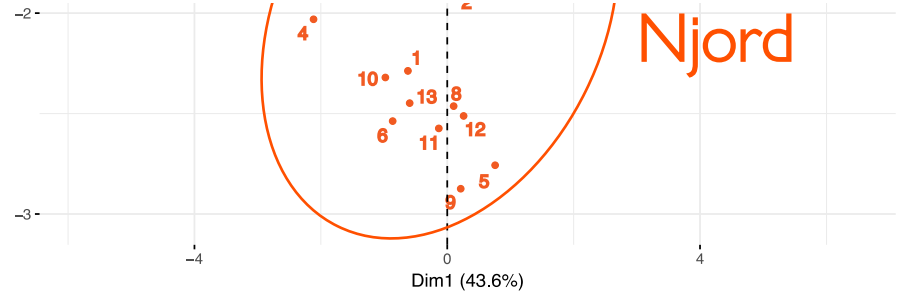
➔ rRNA composition bias

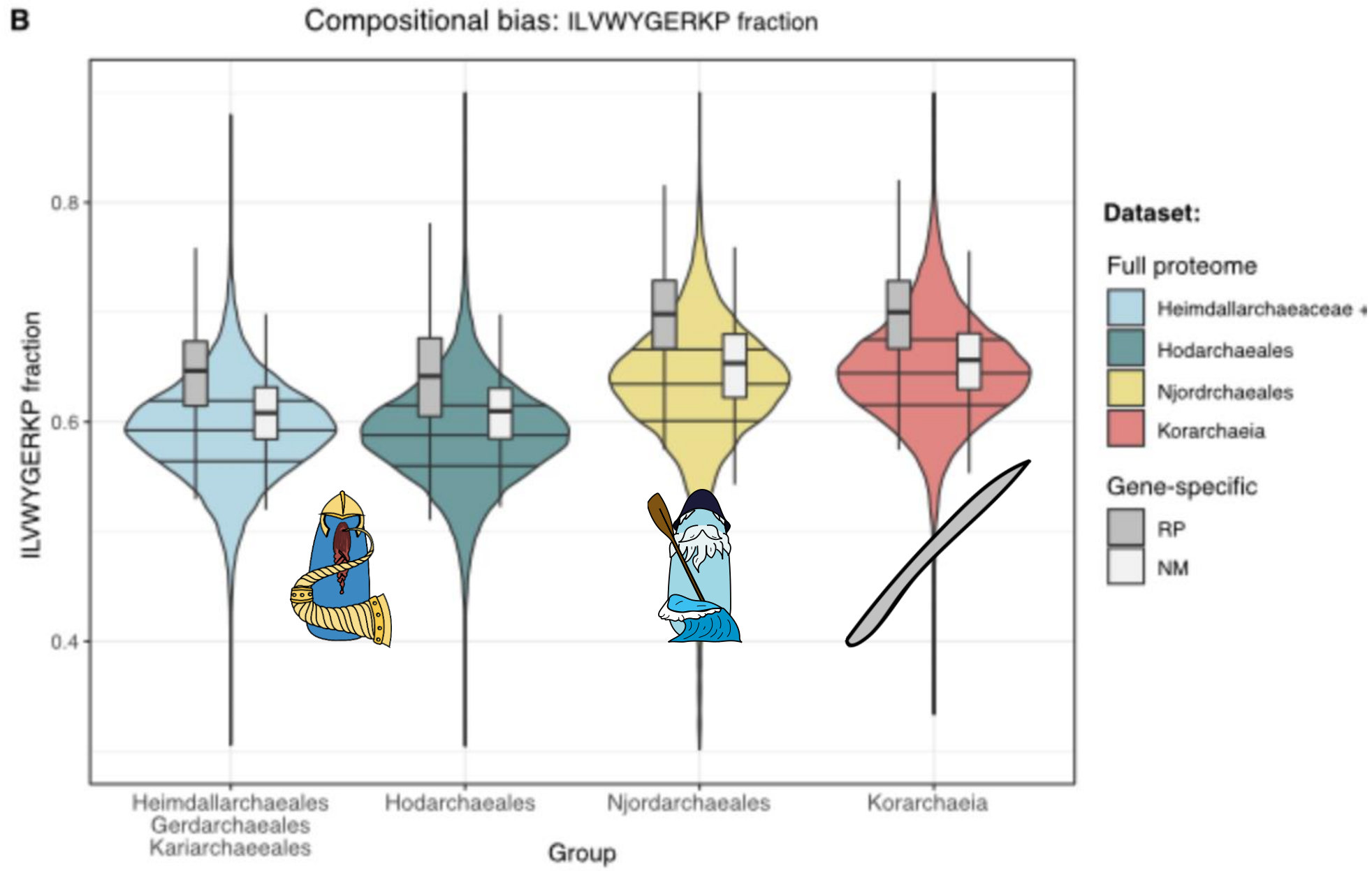
➔ structural constraints in the ribosome explaining convergent AA composition



➔ Small systematic bias leads to large artefacts in concatenations






B
Compositional bias: ILVWYGERKP fraction
ILVWYGERKP fraction
0.8
0.6
0.4
Heimdallarchaeales
Gerdarchaeales
Kariarchaeeales
Hodarchaeales
Njordarchaeales
Korarchaeia
Group
Dataset:
Full proteome
Heimdallarchaeaceae
Hodarchaeales
Njordrchaeales
Korarchaeia
Gene-specific
RP
NM

How do Eukaryotes relate to Asgards?

Dataset

- **Ribosomal proteins**
- **New markers**

Software

- **IQ-Tree (LG+G4+C60+F+PMSF)**
- **Phylobayes (CAT+LG+G4)**

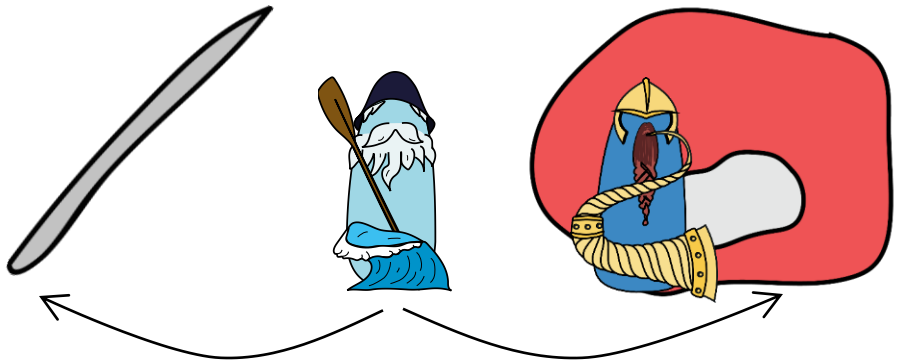
Taxon sampling

- **+/- DPANN**
- **+/- Eukaryotes**
- **+/- Korarchaea**

Fast-evolving site removal (FSR)

Recoding (SR4)

~800 phylogenies…

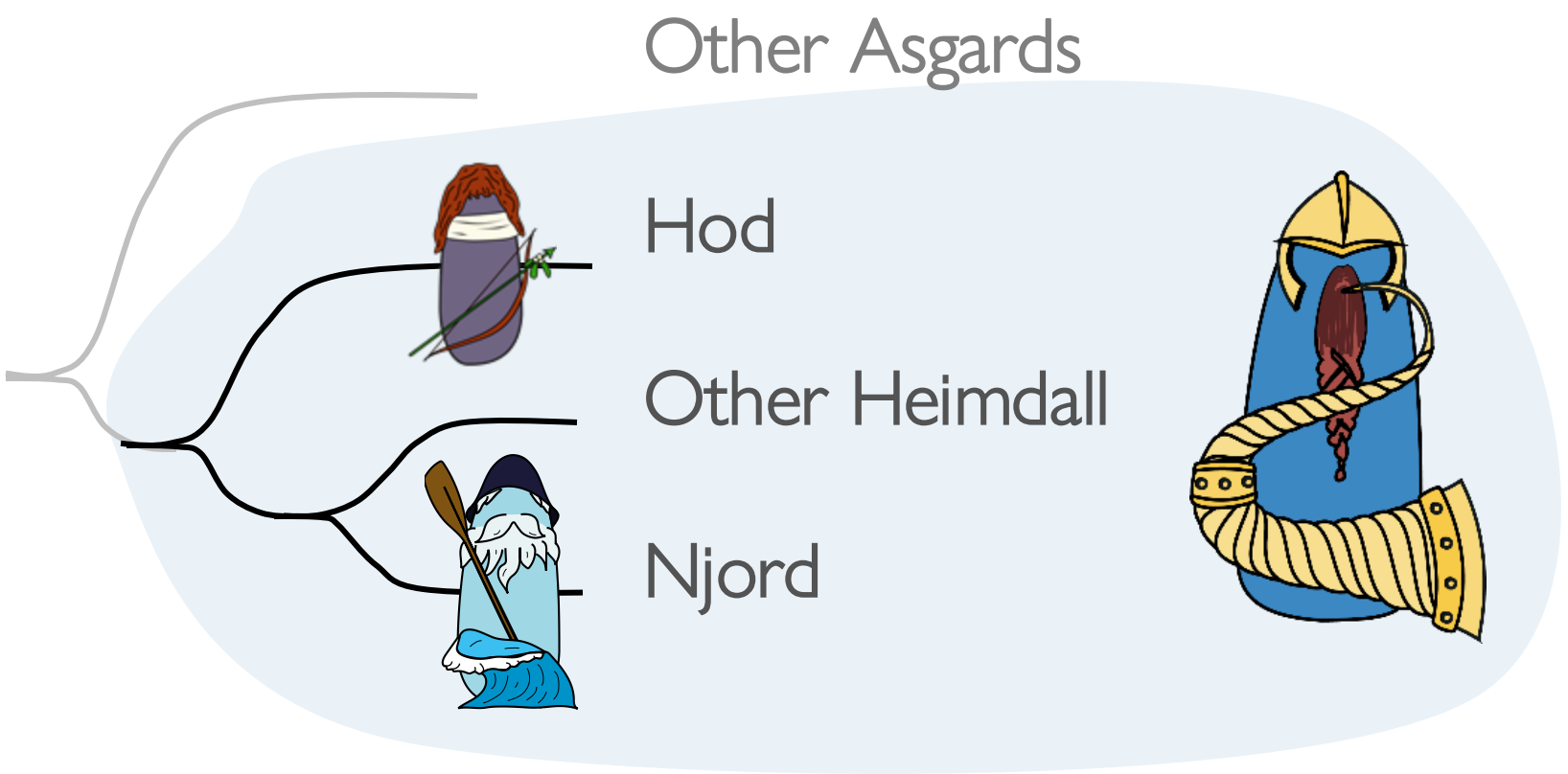


Ribosomal proteins (RP)

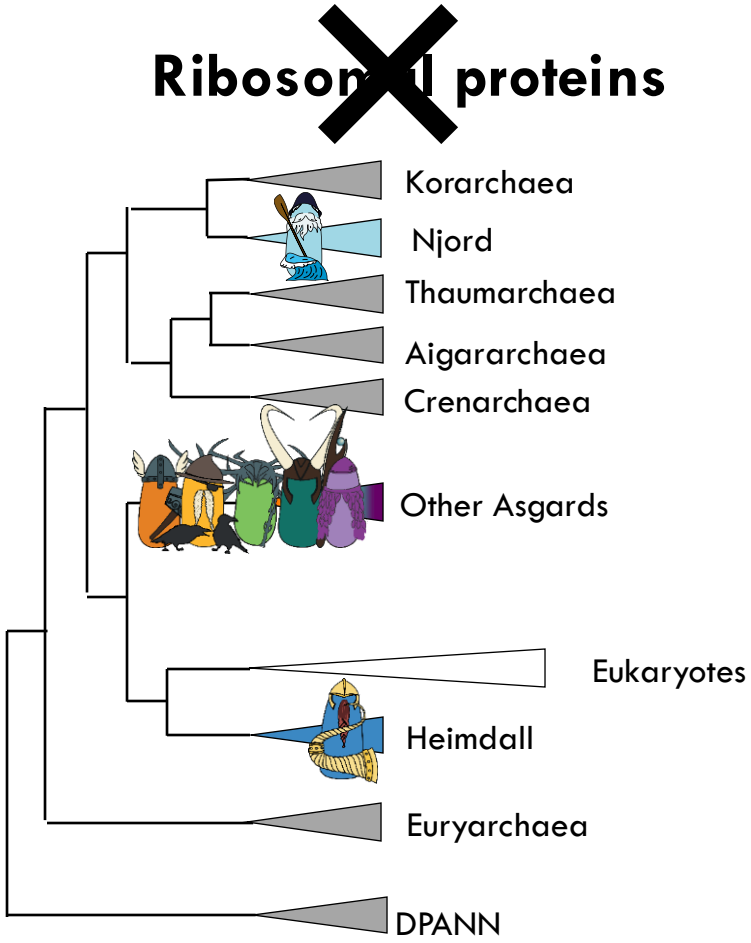
New markers (NM)

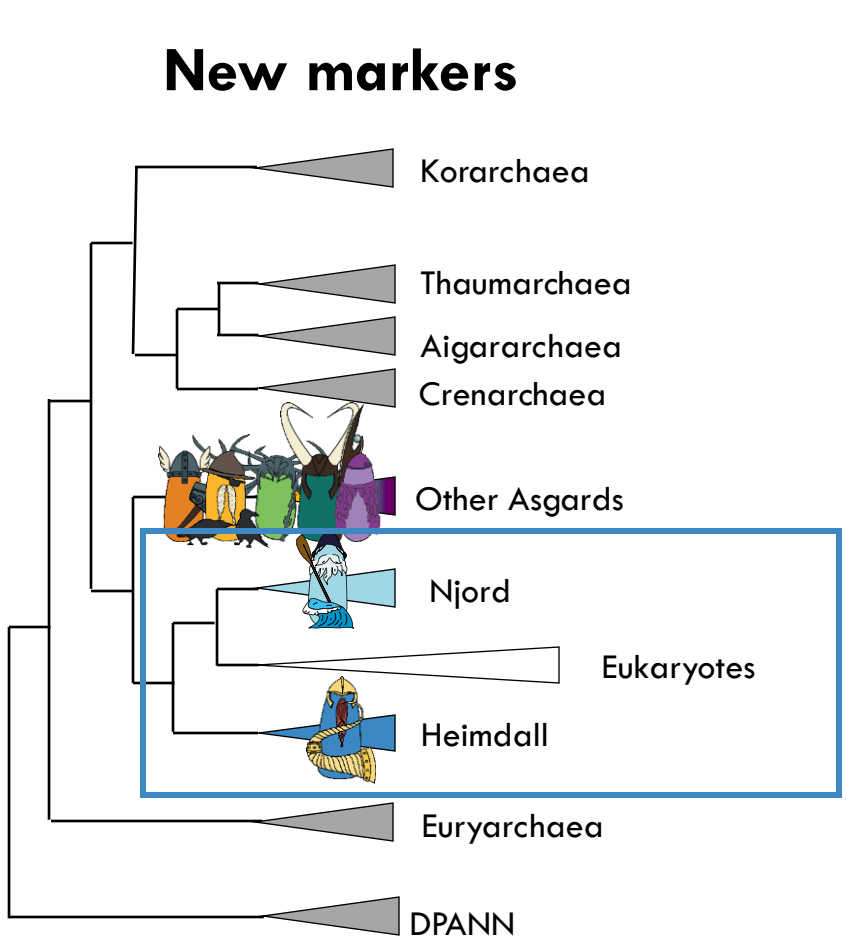
RP (without Korarchaea)

Njord are in fact a sub-clade of Heimdallarchaeia



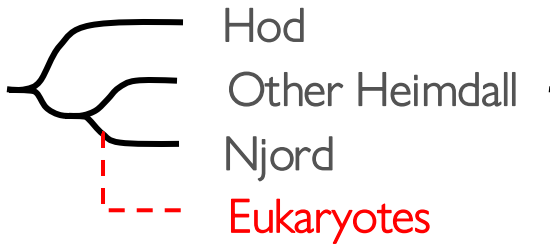
Why do we care, again?





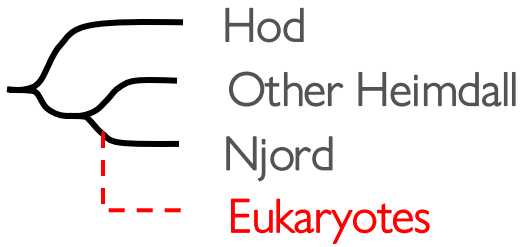
Placing Eukaryotes in the Asgard tree require careful phylogenetic investigations

Untreated

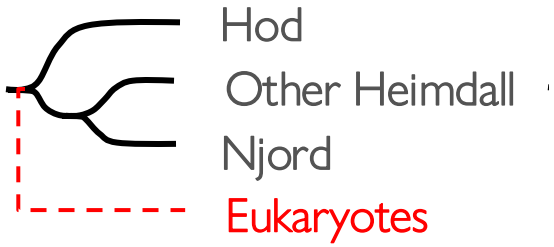


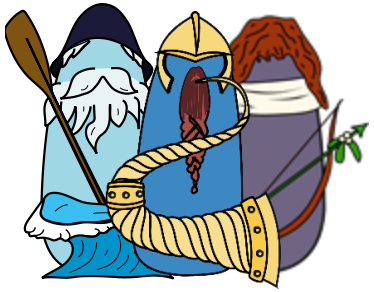
Placing Eukaryotes in the Asgard tree require careful phylogenetic investigations

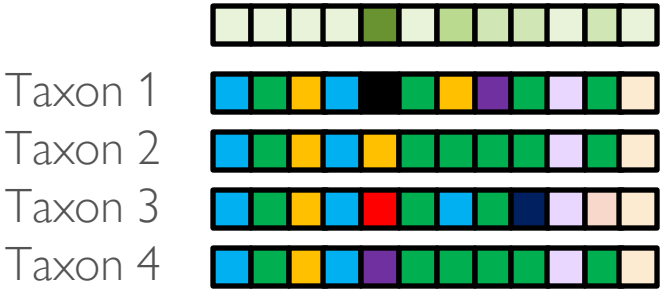
Untreated



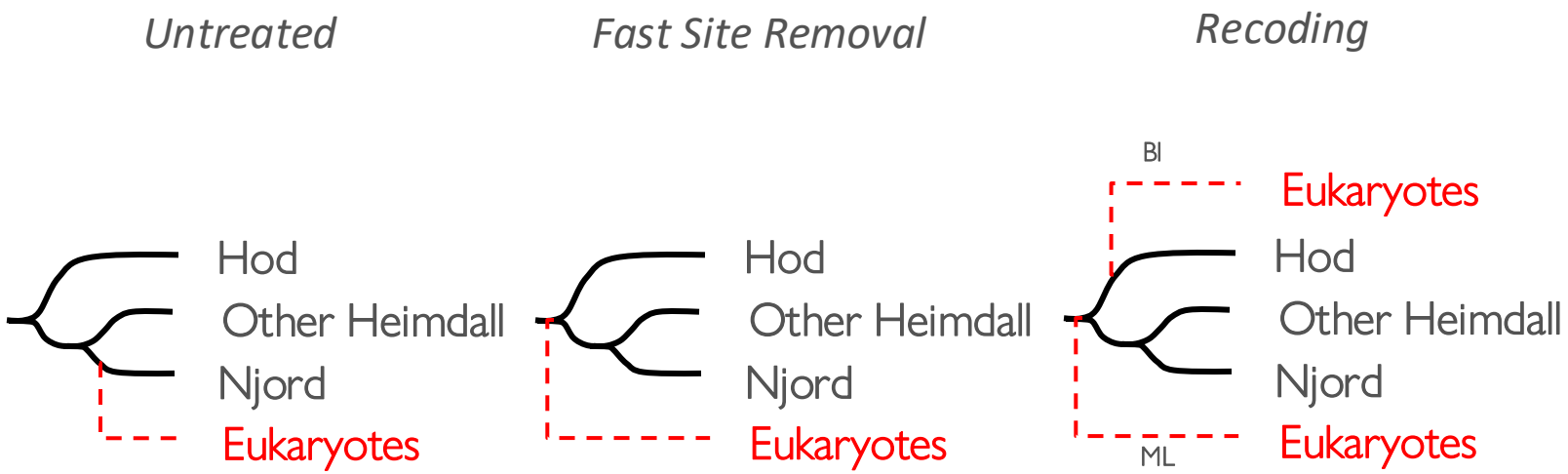
Fast Site Removal

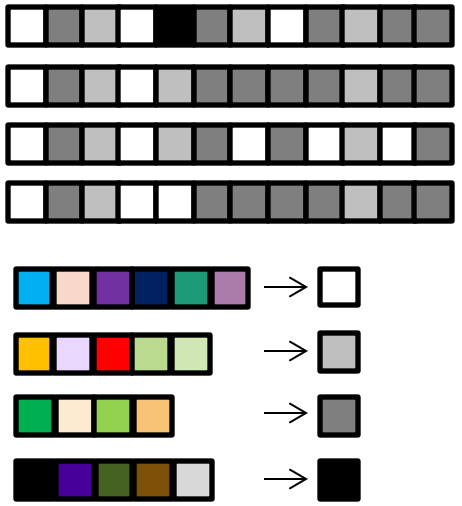


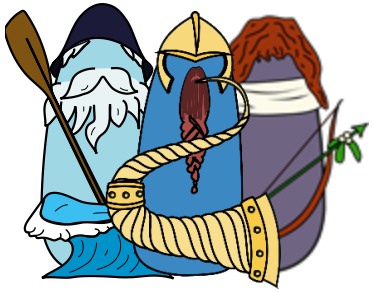


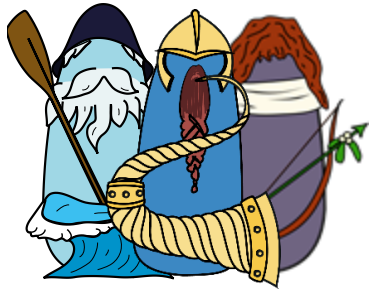


Placing Eukaryotes in the Asgard tree require careful phylogenetic investigations





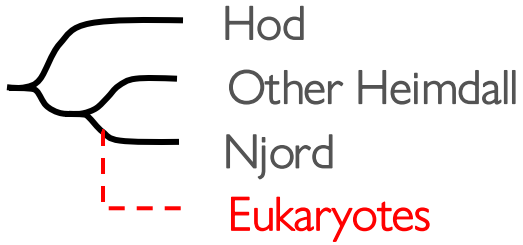




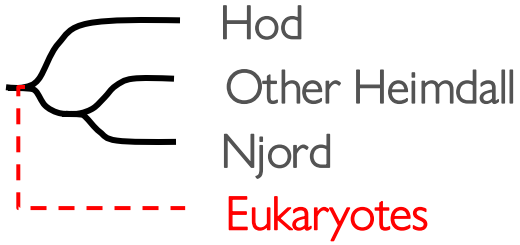
E.g.: SR4 recoding (Susko and Roger 2007)

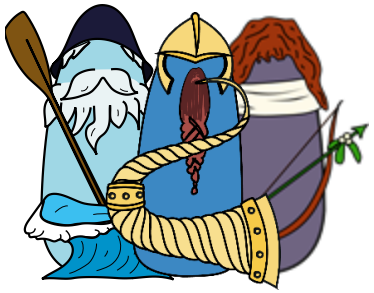
Placing Eukaryotes in the Asgard tree require careful phylogenetic investigations

Untreated

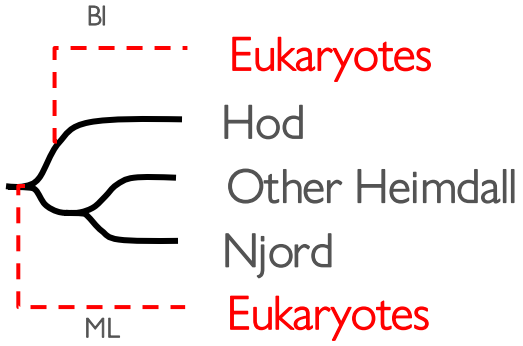


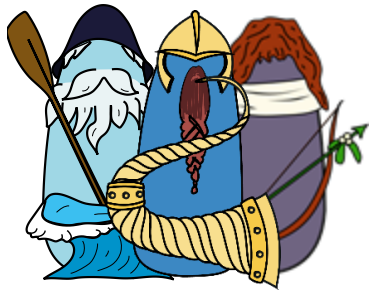
Fast Site Removal



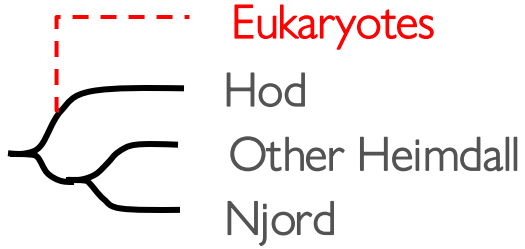


Recoding

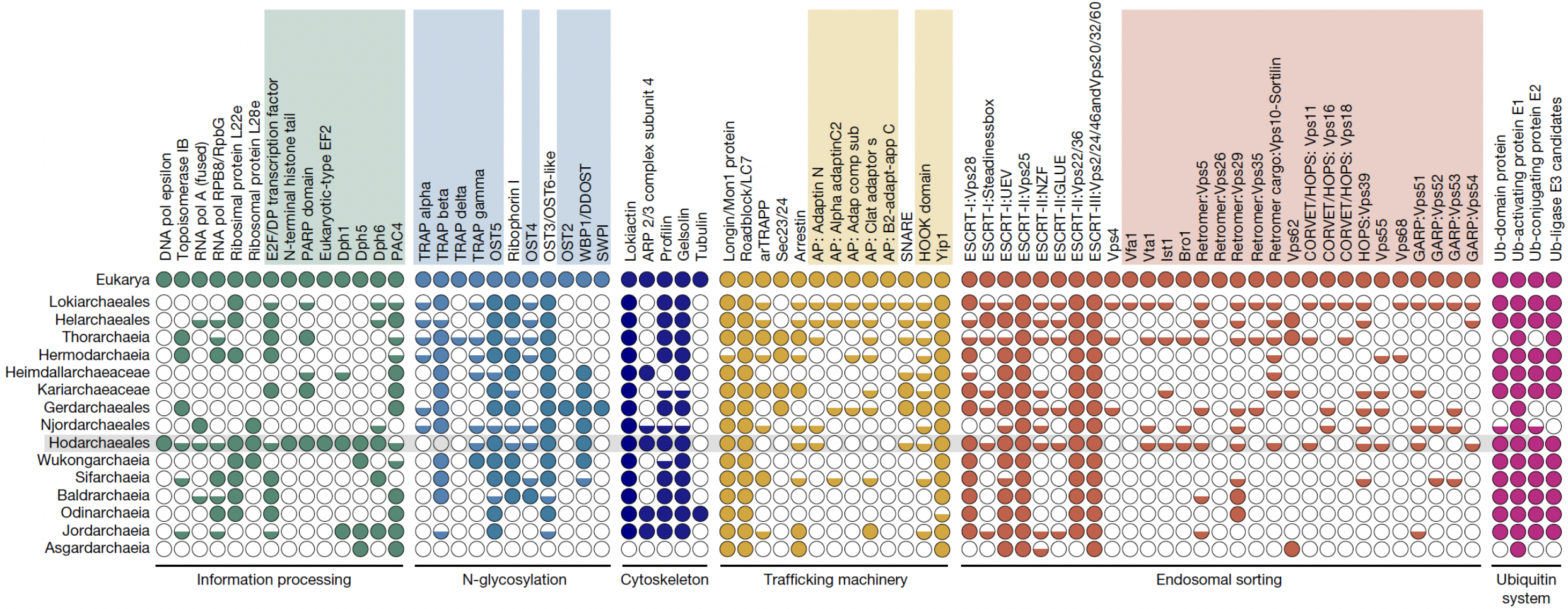




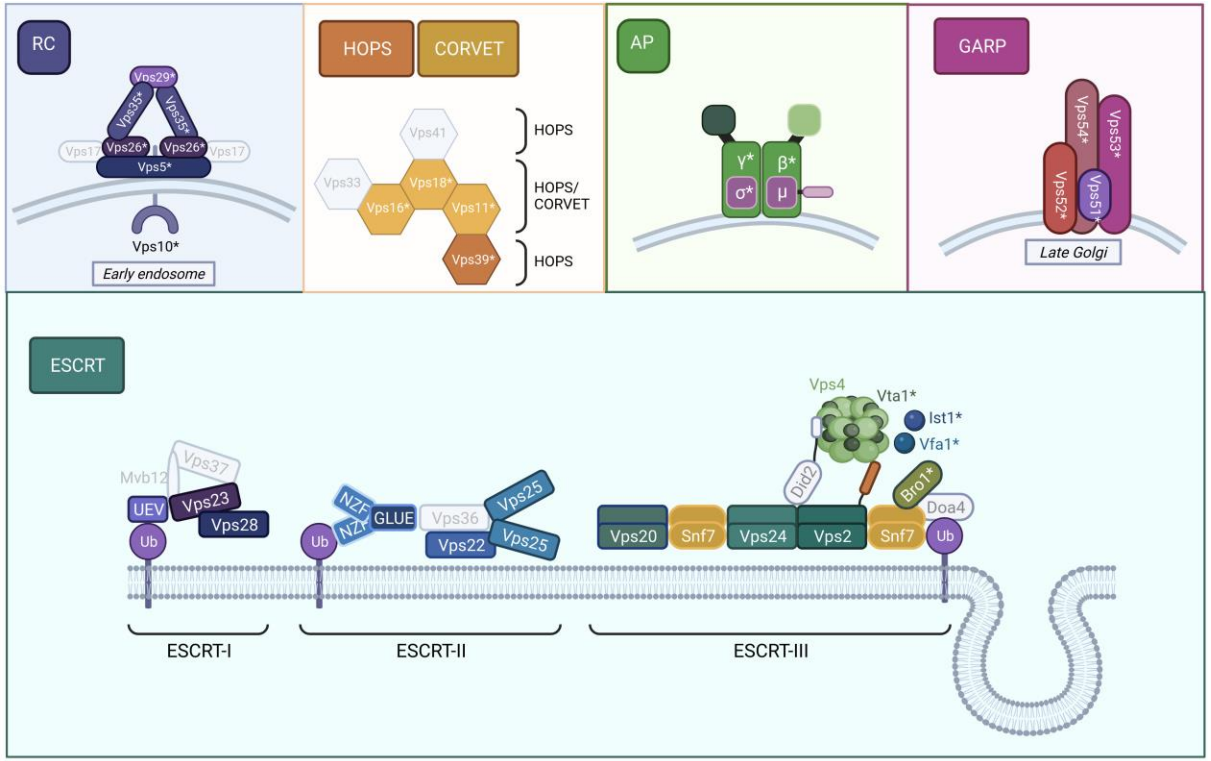
FSR+Recoding

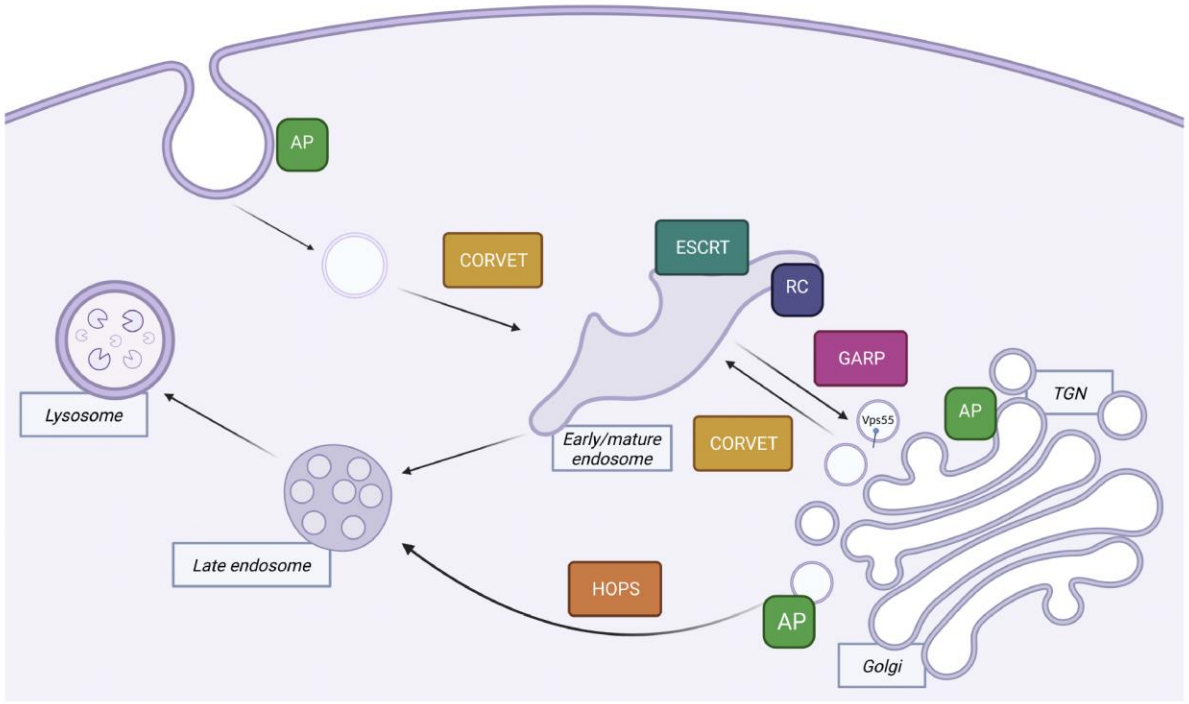


New ESPs support the relationship of Eukaryotes and Hod

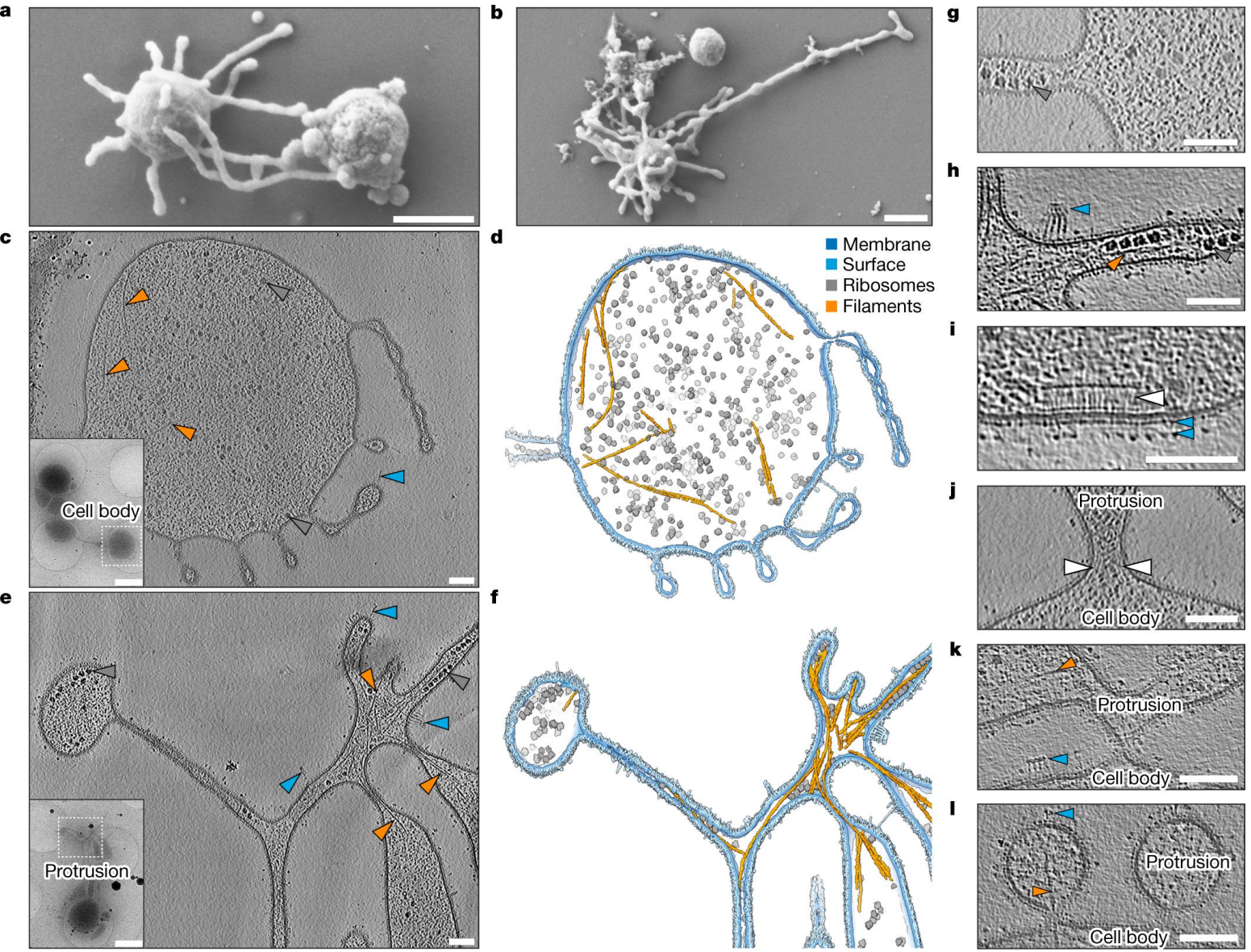


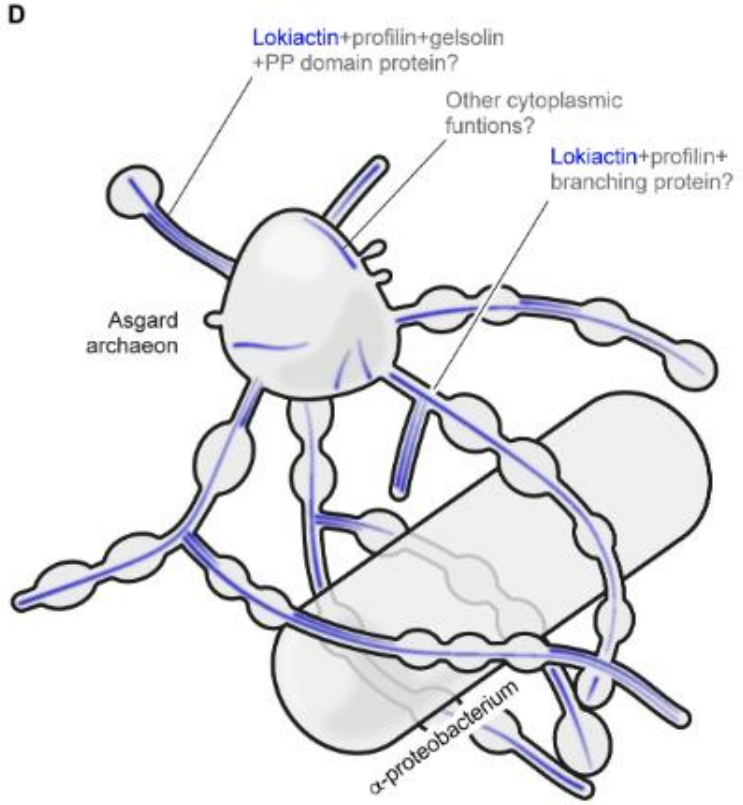
New ESPs involved in complex intracellular trafficking





Asgard archaea have an actin-based cytoskeleton





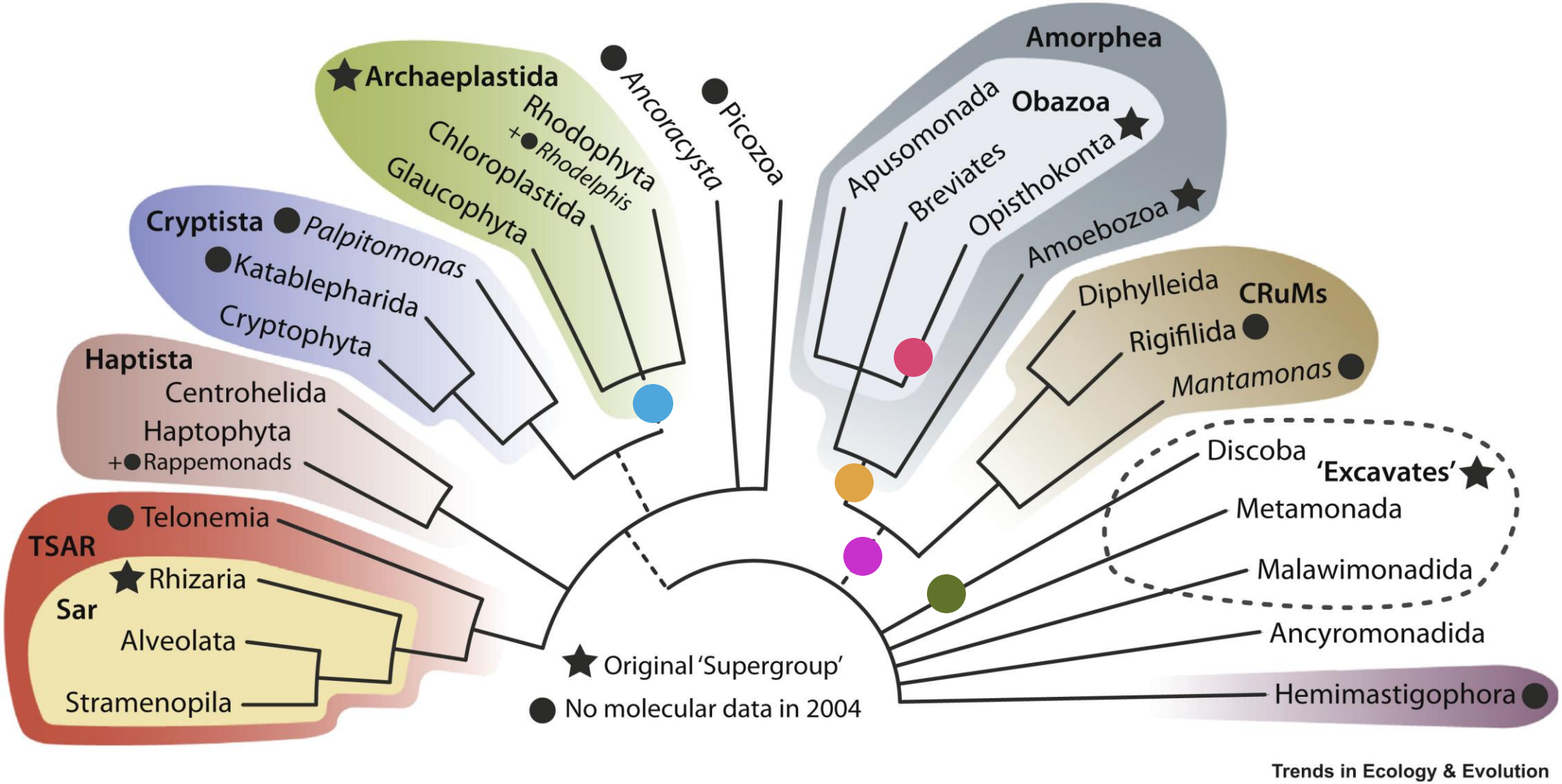
Charles-Orszag A, Petek-Seoane NA, Mullins RD. 2024. Archaeal actins and the origin of a multi-functional cytoskeleton. J Bacteriol

Rodrigues-Oliveira, T., Wollweber, F., Ponce-Toledo, R.I. *et al.* Actin cytoskeleton and complex cell architecture in an Asgard archaeon. *Nature* **613**, 332–339 (2023).

Example 3

Rooting the tree of eukaryotes

Williamson, Eme, et al. *Nature*, 2025



Modified from Burki et al., 2020

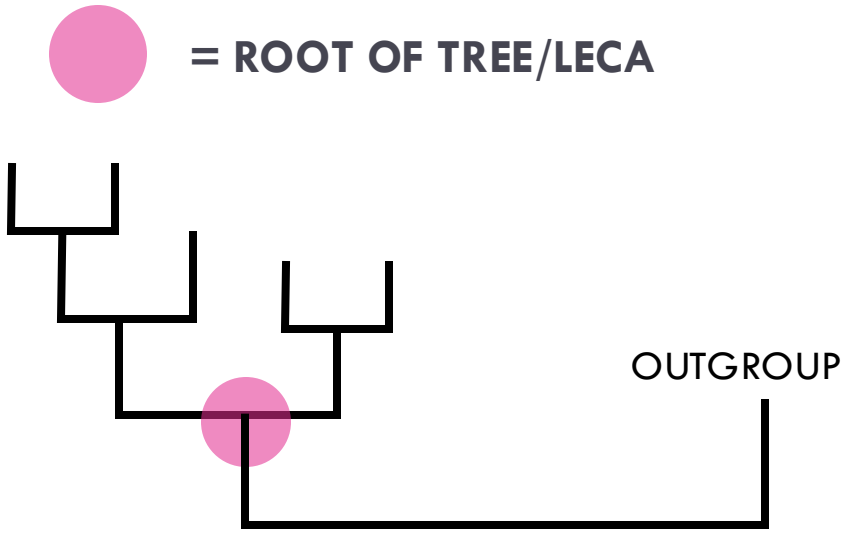
Gene tree parsimony, gene fusions and domain evolution, molecular and cellular features, replacement of highly conserved amino acids…

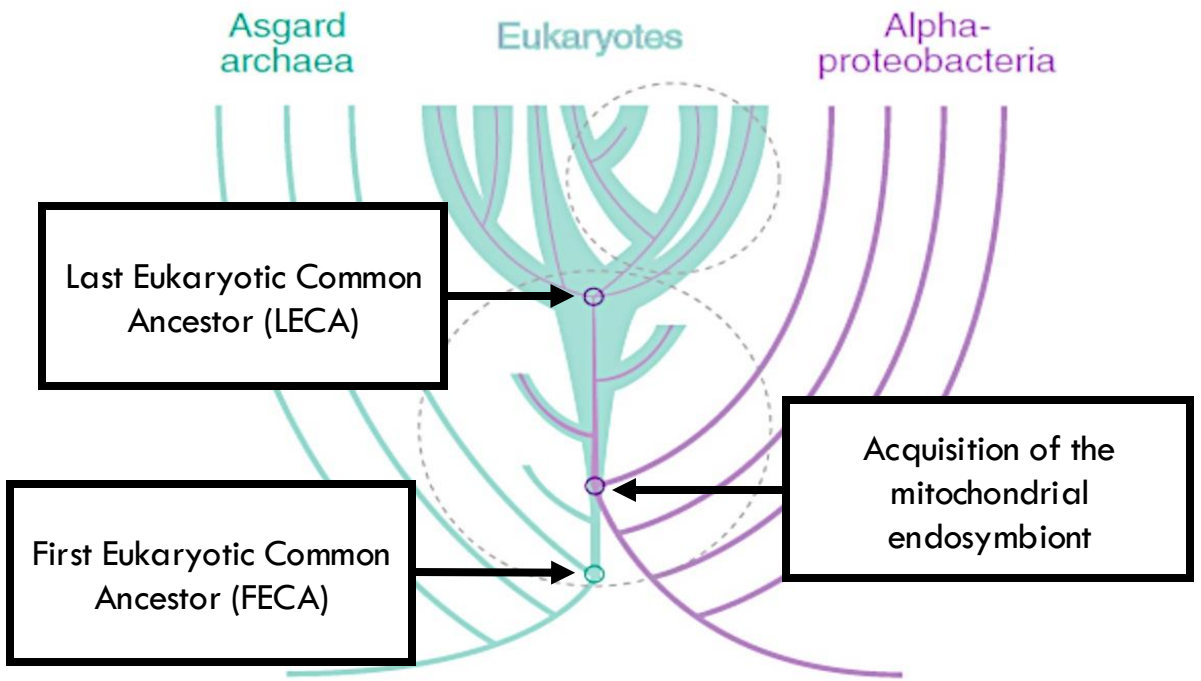
PROBLEMS:

Convergence/reversion, lack of evolutionary models, no tests for robustness, no consensus,
sensitive to missing data and incomplete taxonomic sampling

OUTGROUP ROOTING

Outgroups are lineages that fall outside of the group being studied, but are closely related enough to retain phylogenetic signal through orthologous genes





The evolutionary relationship of eukaryotes with archaea and alphaproteobacteria make them possible outgroups

Shorter branch length between alpha and eukaryotes

Modified from Roger et al., 2017

Identify proteins of alphaproteobacterial origin in a small set of eukaryotes

Retrieve homologs from selected eukaryotes and prokaryotes

Estimate single gene trees for each marker gene (IQ-TREE)

Manual removal of non-orthologs and outliers

x5

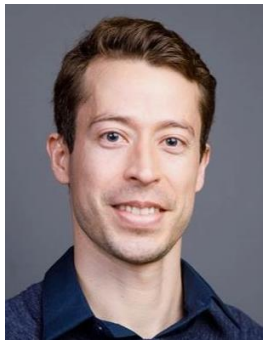
Select outgroup size and composition

FINAL DATASETS

93 marker genes
63 eukaryotes
37 alphaproteobacteria

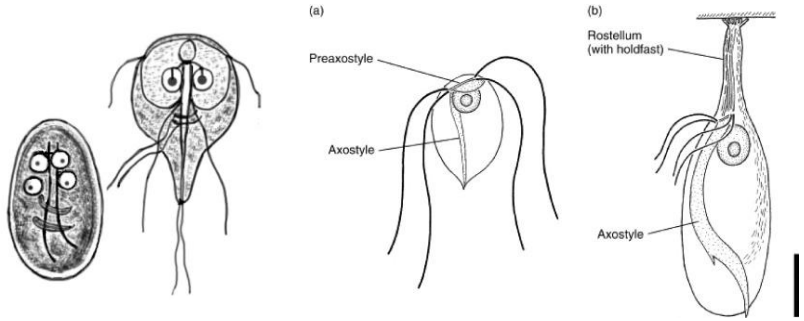
Kelsey Williamson

Laura Eme



Sergio Muñoz-Gómez

Metamonads are problematic taxa for datasets of mitochondrial proteins

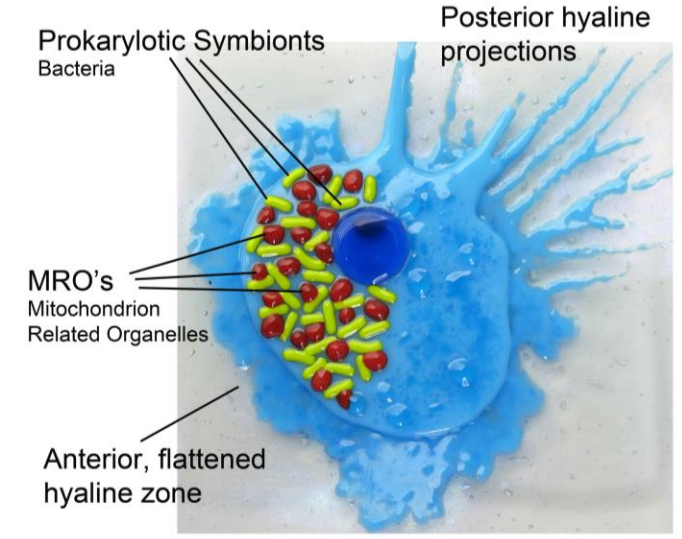


Metamonada consists entirely of **anaerobes:**
lack a mitochondrial genome;
have highly reduced mitochondrion-related organelles
➔ lack most of the proteins in the dataset

Metamonads with the most mitochondrial proteins were included:

Anaeramoeba ignava – 13 genes, 19% site occupancy
Anaeramoeba flamelloides – 11 genes, 16.5% site occupancy

Datasets with and without metamonads were generated:
Anae+ and Anae-



Anaeramoeba fused glass dish
Jane Hartman

TESTING THE POSITION OF THE ROOT

1. Estimate rooted phylogeny with site-heterogenous models

2. Create alternate root positions to evaluate likelihood differences between previously proposed roots and the optimal root estimated here

3. Evaluate all root positions under additional complex evolutionary models that account for different phenomena

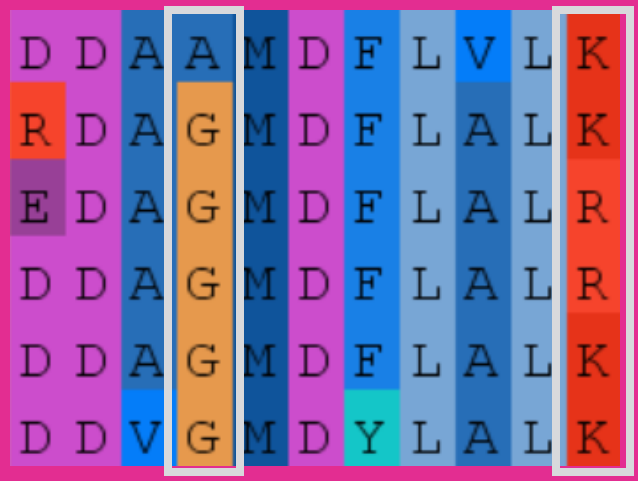
SITE-HETEROGENEOUS MODELS

Site heterogeneous models account for the fact that different sites evolve under different evolutionary constraints.

C-series models (E.g. C60)

UDM models (E.g. UDM64)

Set of amino acid frequency classes generated from public protein databases



SITE-HETEROGENEOUS MODELS

MEOW (**M**AMMaL **E**xtension **O**n **W**hole alignment)

- Amino acid frequency classes are estimated directly from the data
- Can use different proportions of high- versus low-rate sites to generate custom site classes

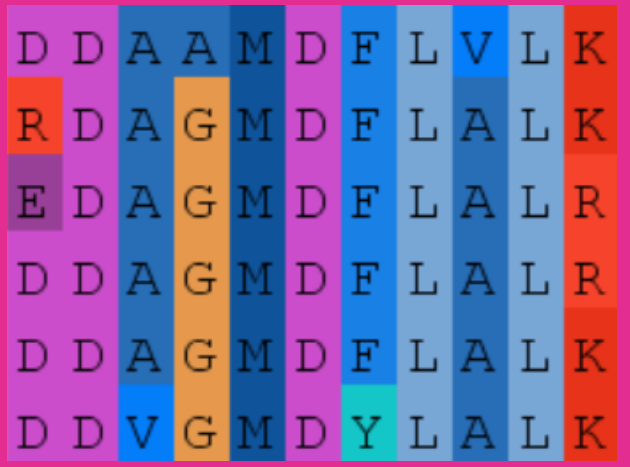
CAT-PMSF

- Generates set of site-specific amino acid frequencies
- Phylobayes is run on a fixed guide tree (ex. LG+G) to estimate frequencies

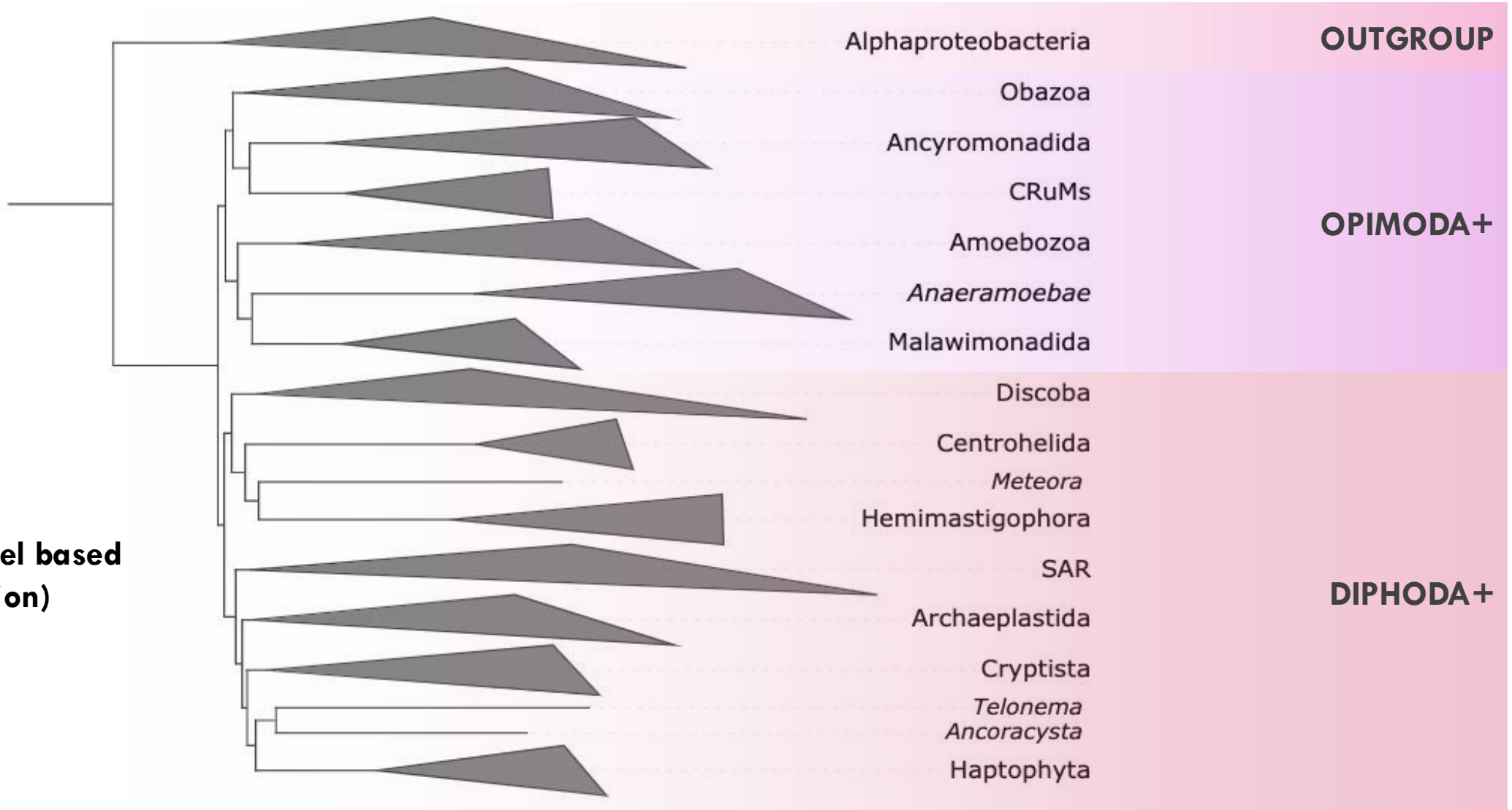
Hector Baños

Ed Susko

Andrew Roger



All models recover a root separating eukaryotes into ‘Opimoda+’ and ‘Diphoda+’



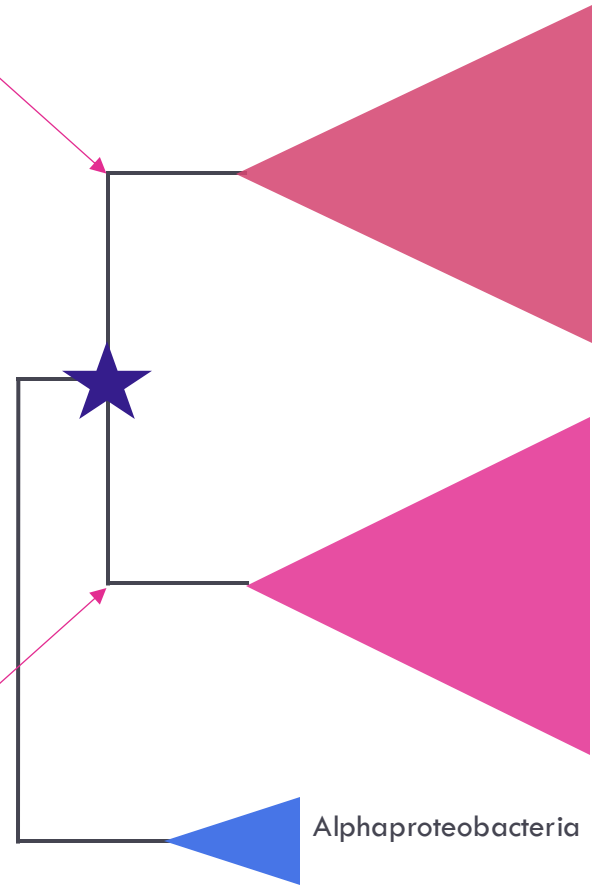
LG+MEOW80
(best fitting model based on cross-validation)

All models recover a root separating eukaryotes into ‘Opimoda+’ and ‘Diphoda+’

ANAE+: 94.9/87
ANAE-: 93/91

MEOW80 (aLRT/UFBOOT):

ANAE+: 99.9/55
ANAE-: 98.7/97



Archaeplastida
Cryptista
Haptophyta
SAR
Discoba

Hemimastigophora
Centrohelida
Telonemia
Meteora
Ancoracysta

DIPHODA+

Obazoa
CRuMs
Anaeramoebae*
Ancyromonadida
Amoebozoa

Malawimonadida

OPIMODA+

Alphaproteobacteria

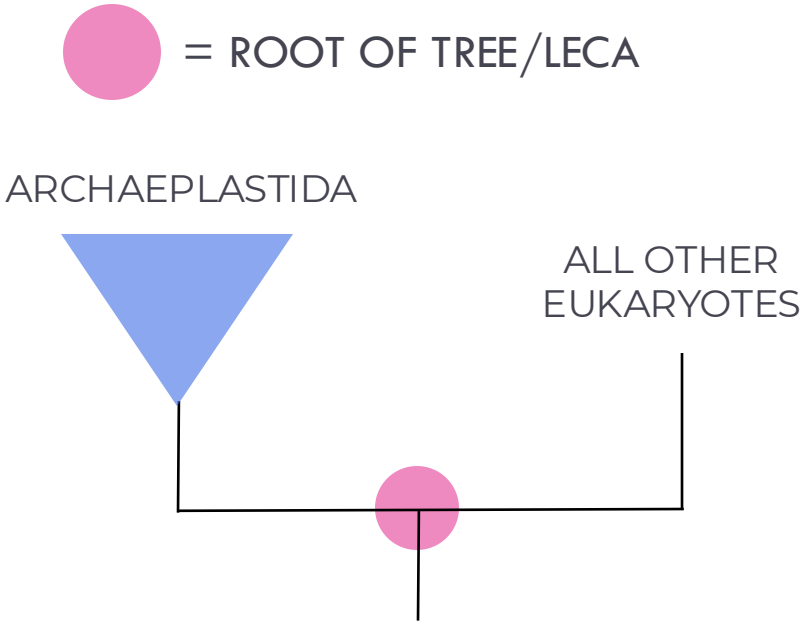
★ = ROOT

TESTING ALTERNATE ROOT POSITIONS

Alternative root topologies were generated by setting monophyly constraints in IQ-TREE

Additional roots tested:

1. **METAMONADA (ANAERAMOEBAE)**
2. **DISCOBA**
3. **OPISTHOKONTA**
4. **MALAWIMONADA**
5. **JAKOBIDA**
6. **EUGLENOZOA**
7. **ARCHAEPLASTIDA**



INCLUDING ADDITIONAL COMPLEXITY TO THE PHYLOGENETIC MODELS

FUNDI – models **functional divergence** (change of amino acid preference) at sites across a known branch (ie, the branch between alpha and euka)

GHOST – models **heterotachy** (changing rate of evolution at sites across the tree)

GF-MIX – accommodates **amino acid compositional biases** driven by changing GC content across the tree

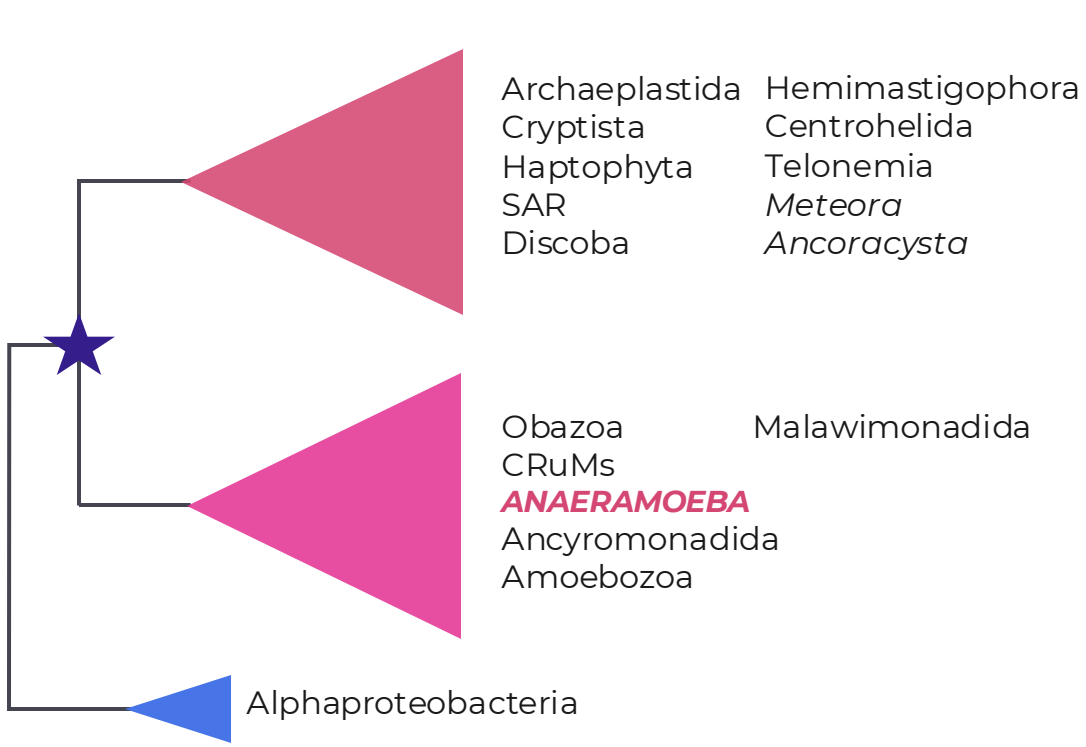
Likelihoods **improve** under more complex models, but order of root preference remains consistent.

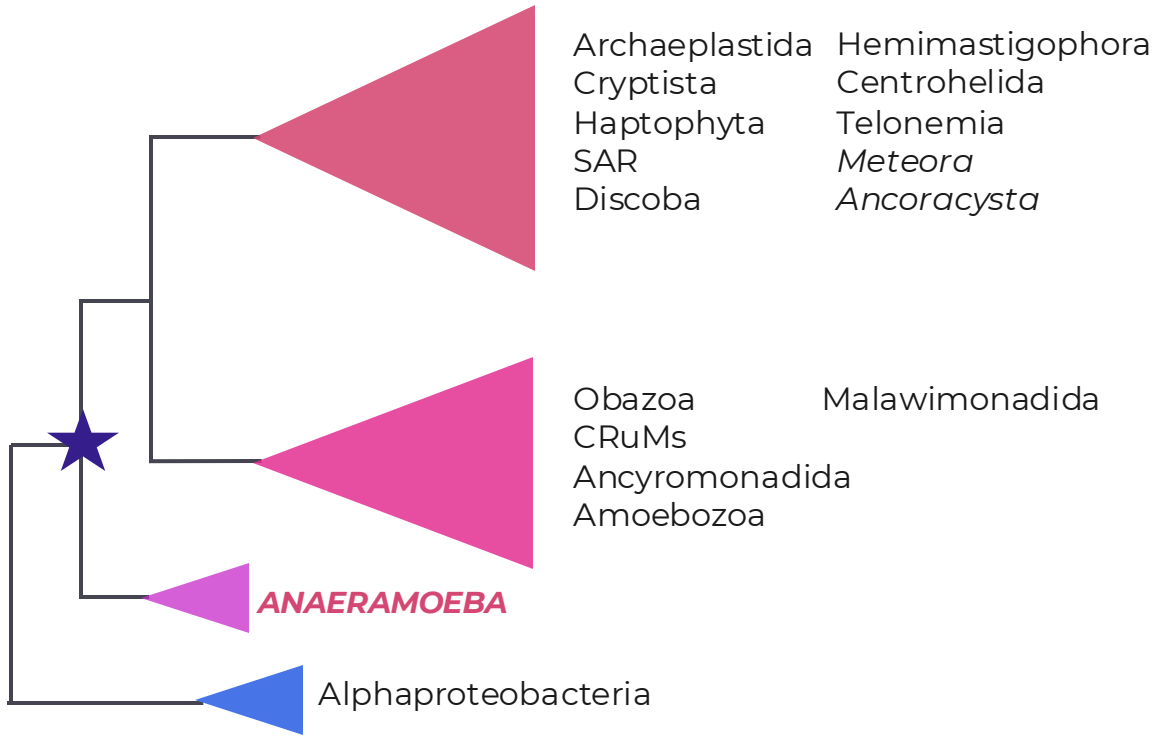
Neither an Opimoda+ root or an Anaeramoeba root are rejected by topology testing.

ANAE+ DATASET				
TOPOLOGY	**MEOW80**	**MEOW80+FUNDI**	**MEOW80+GHOST**	**MEOW80+GF-MIX**
OPIMODA+	**-1665043.854**	**-1657618.898**	**-1650367.99**	**-1655099.392**
ANAERAMOEBA	**-1665044.498**	**-1657619.576**	**-1650369.352**	**-1655103.092**
DISCOBA	-1665061.809	-1657635.485	-1650394.794	-1655134.492
MALAWIMONADIDA	-1665095.632	-1657650.686	-1650414.421	-1655148.935
OPISTHOKONTA	-1665138.368	-1657687.757	-1650454.564	-1655204.365
EUGLENOZOA	-1665146.579	-1657718.133	-1650475.147	-1655210.287
ARCHAEPLASTIDA	-1665170.708	-1657713.761	-1650467.552	-1655221.351
JAKOBIDA	-1665205.285	-1657772.885	-1650538.914	-1655281.622

The position of *Anaeramoeba*, the only representative of Metamonada, is not well-resolved

- **ANAEROBES – NO MITOCHONDRIAL GENOME**
- **SITE OCCUPANCY OF 18%**
- **LONG-BRANCHING**



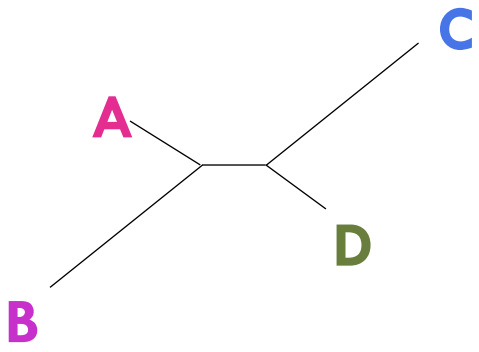


★ = ROOT

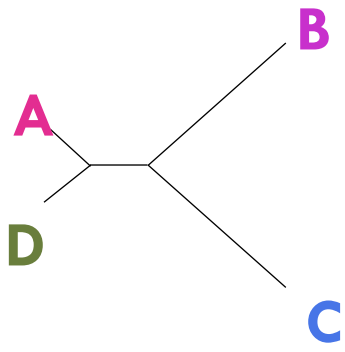
LONG BRANCH ATTRACTION (LBA)

Under model misspecification or small sample bias, long branches artefactually branch together

TRUE TREE

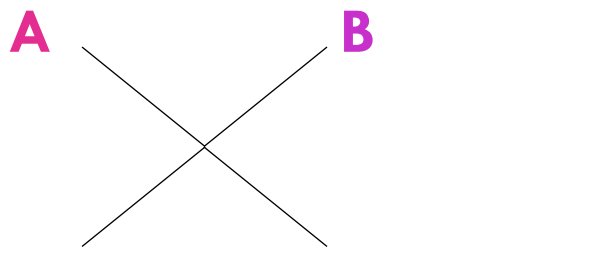


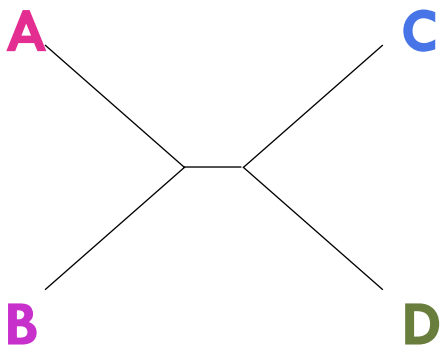
LBA TREE



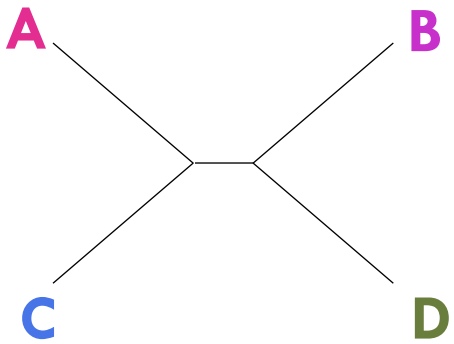
Simulation studies support an artefactual attraction of *Anaeramoeba* to the long branch connecting the outgroup

TRUE TREE

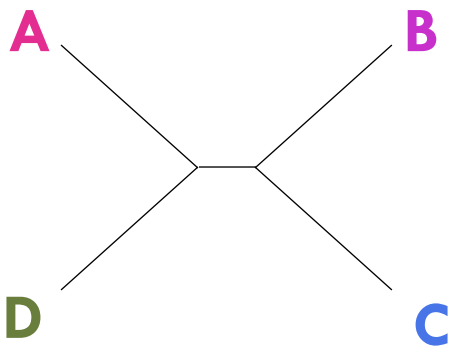




TREE 1

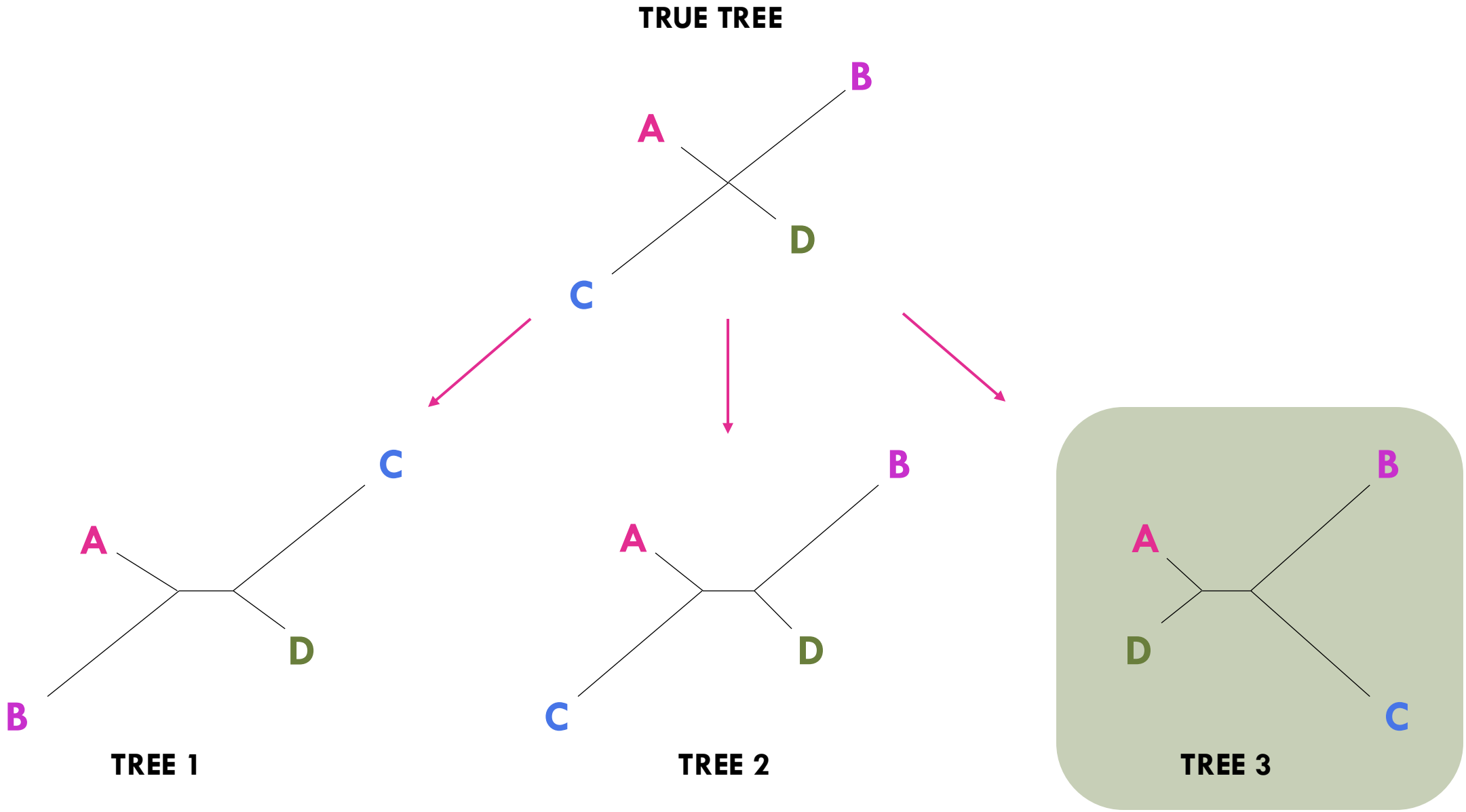


TREE 2



TREE 3

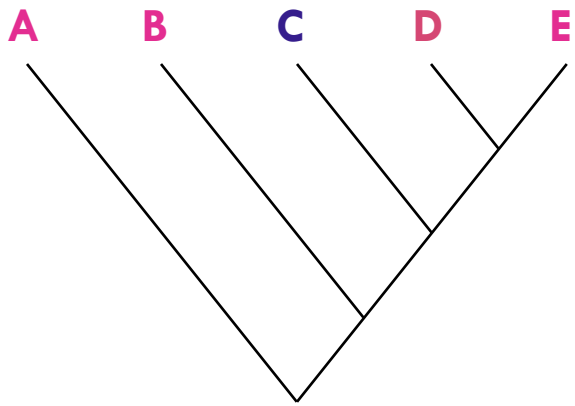
ALL TOPOLOGIES EQUALLY LIKELY



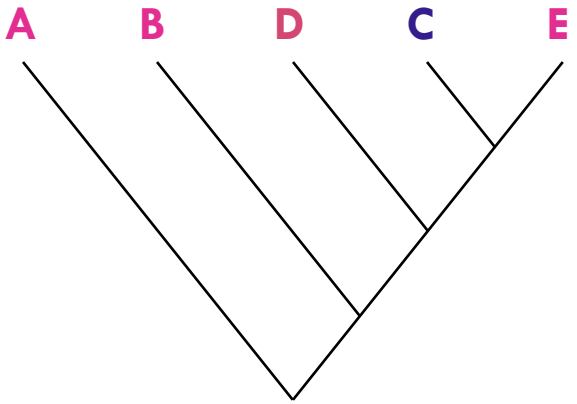
UNDER MODEL MISSPECIFICATION OR SMALL SAMPLE BIAS, LBA TREE IS PREFERRED

TESTING FOR LONG BRANCH ATTRACTION

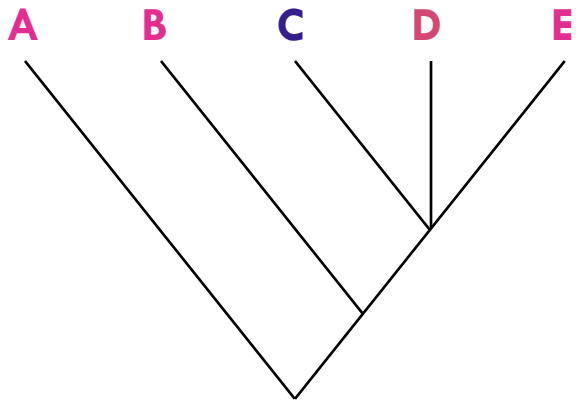
1. Create a consensus tree from two competing topologies



VS.



CONSENSUS:



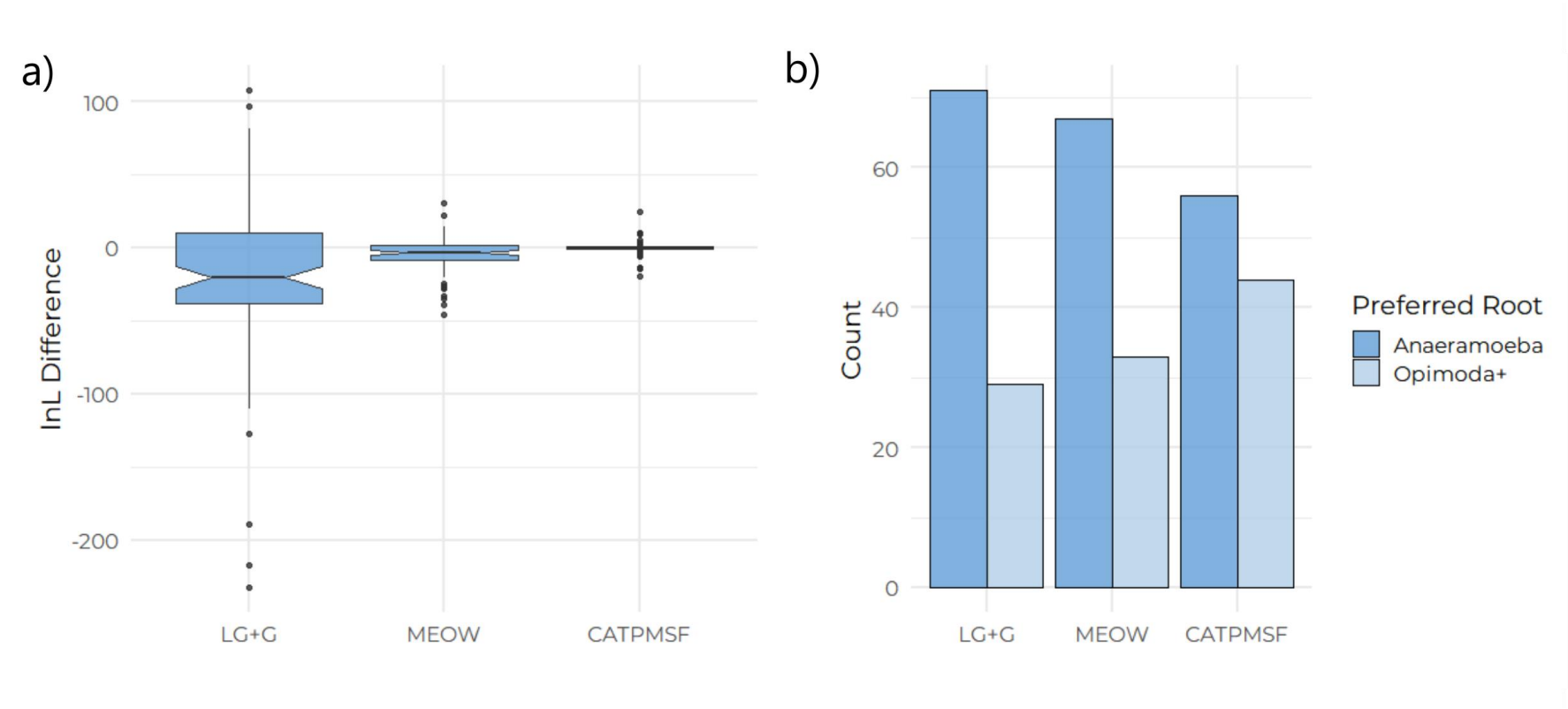
TESTING FOR LONG BRANCH ATTRACTION

1. Create a consensus tree from two competing topologies – **consider this the 'true' topology**
2. Simulate 100 alignments based on the consensus tree and CAT-PMSF site profiles (ALISIM)
3. Run IQ-TREE on all alignments to calculate the likelihood for both resolved topologies (Opimoda-like and *Anaeramoeba*) given the following models:
    1. CAT-PMSF
    2. MEOW
    3. LG+G

If there is **no LBA,** there should be no preference for one topology over the other (ie. Both topologies are equally likely to have the better likelihood in a given replicate)

If there **is LBA,** there will be a bias toward the *Anaeramoeba* root topology under model misspecification (ie. *Anaeramoeba* topology will have **better** likelihood)

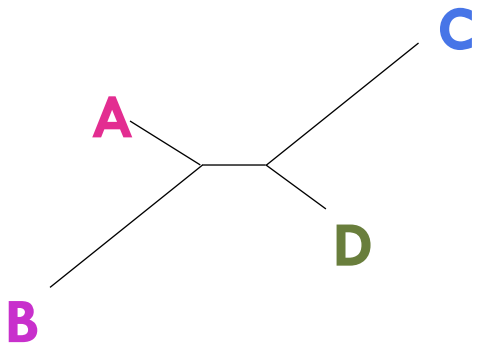
As model misspecification increases, preference for an *Anaeramoeba* root over an Opimoda+ root also increases



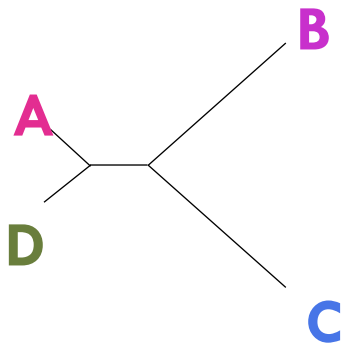
LONG BRANCH ATTRACTION (LBA)

Under model misspecification or small sample bias, long branches artefactually branch together

TRUE TREE



LBA TREE



Simulation studies support an artefactual attraction of *Anaeramoeba* to the long branch connecting the outgroup

➔ *Anaeramoeba* is sensitive to LBA – not enough information to place them confidently in the tree

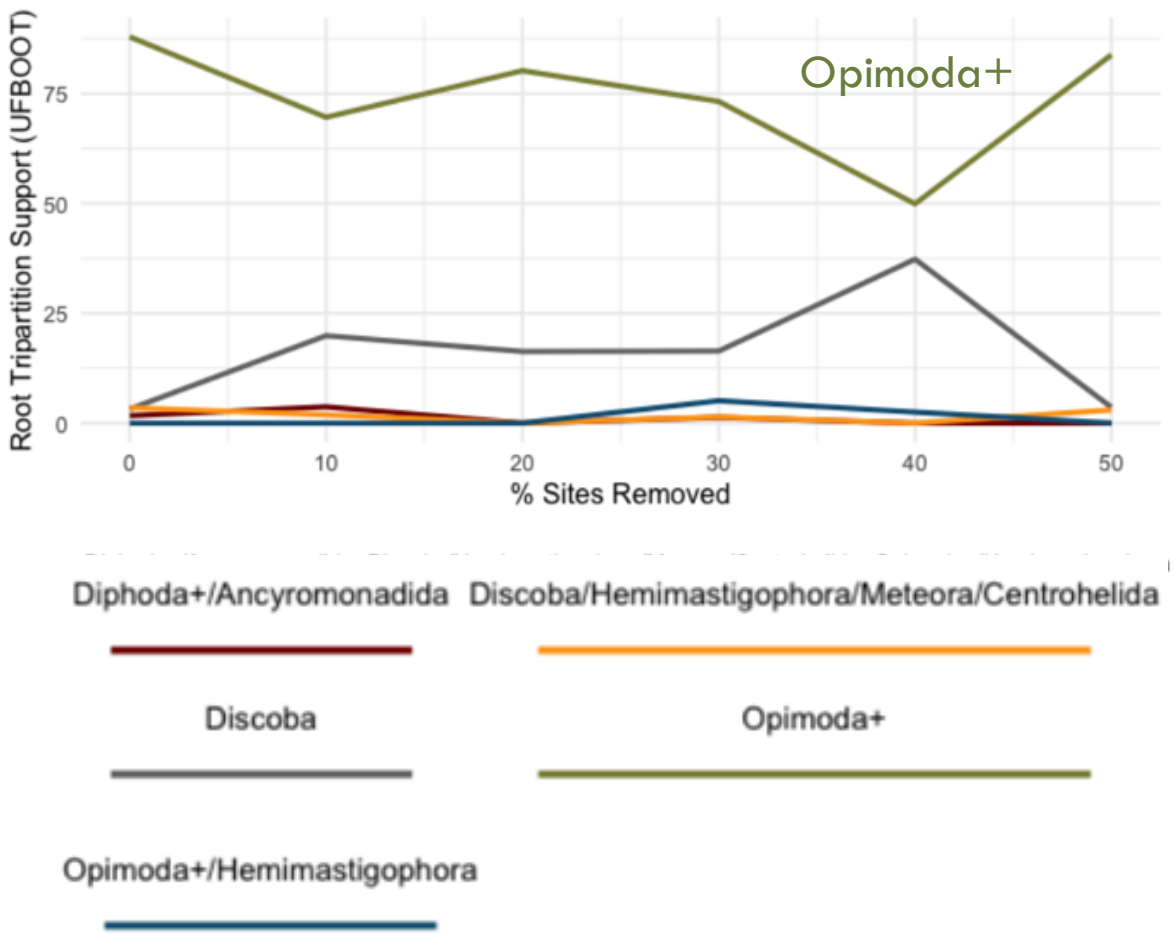
Removing the fastest-evolving sites, genes, and taxa

REMOVAL OF FASTEST EVOLVING SITES

A

% Sites Removed	aLRT/UFBOOT Support	
	Opimoda+	Diphoda+
10	98.5/93	53.2/75
20	99.4/98	78.7/82
30	99.2/93	50.3/75
40	94.6/93	30.2/55
50	98.0/97	95.8/87

B

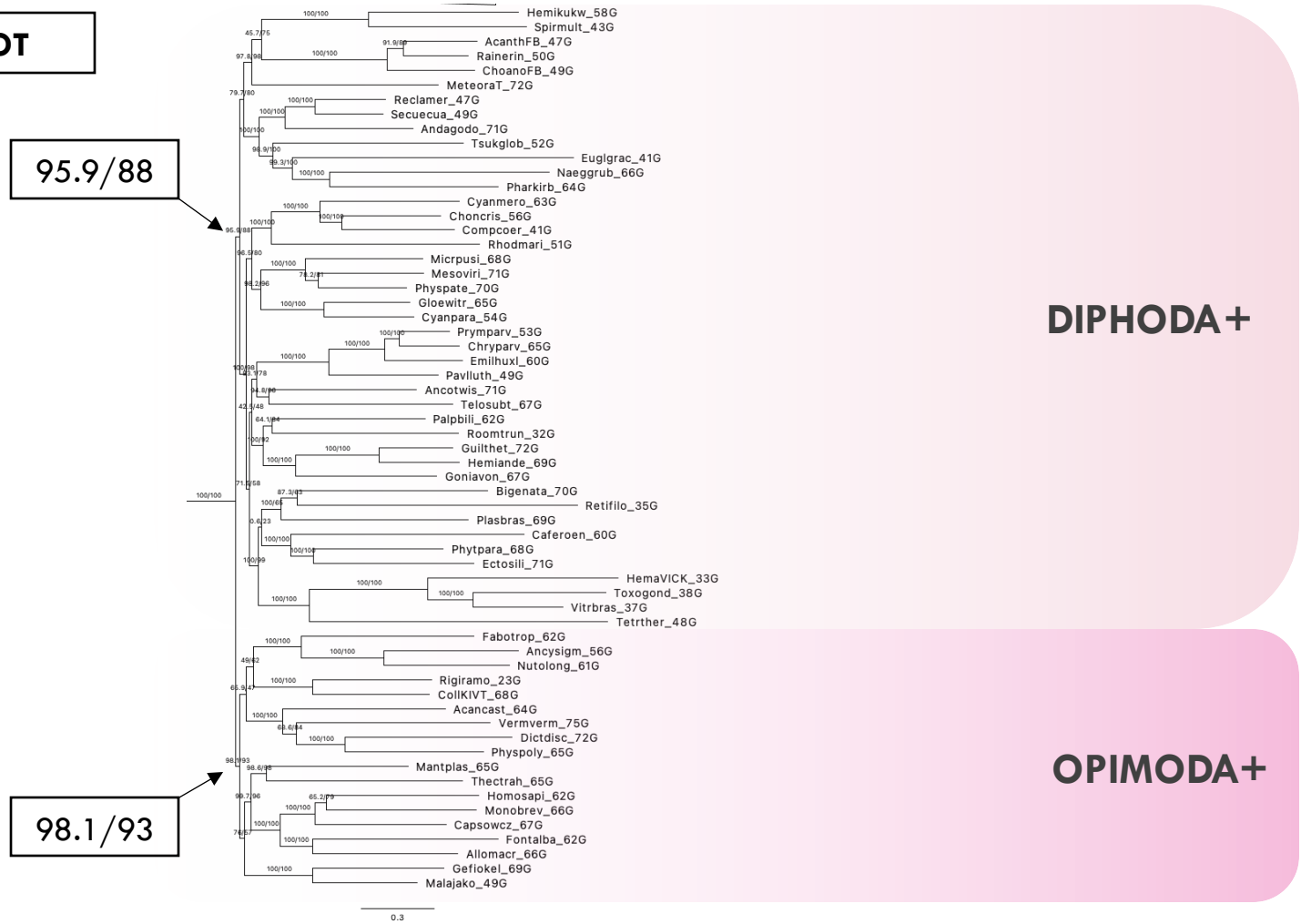


Williamson, et al. (under review)

REMOVAL OF 13 MOST DIVERGENT GENES

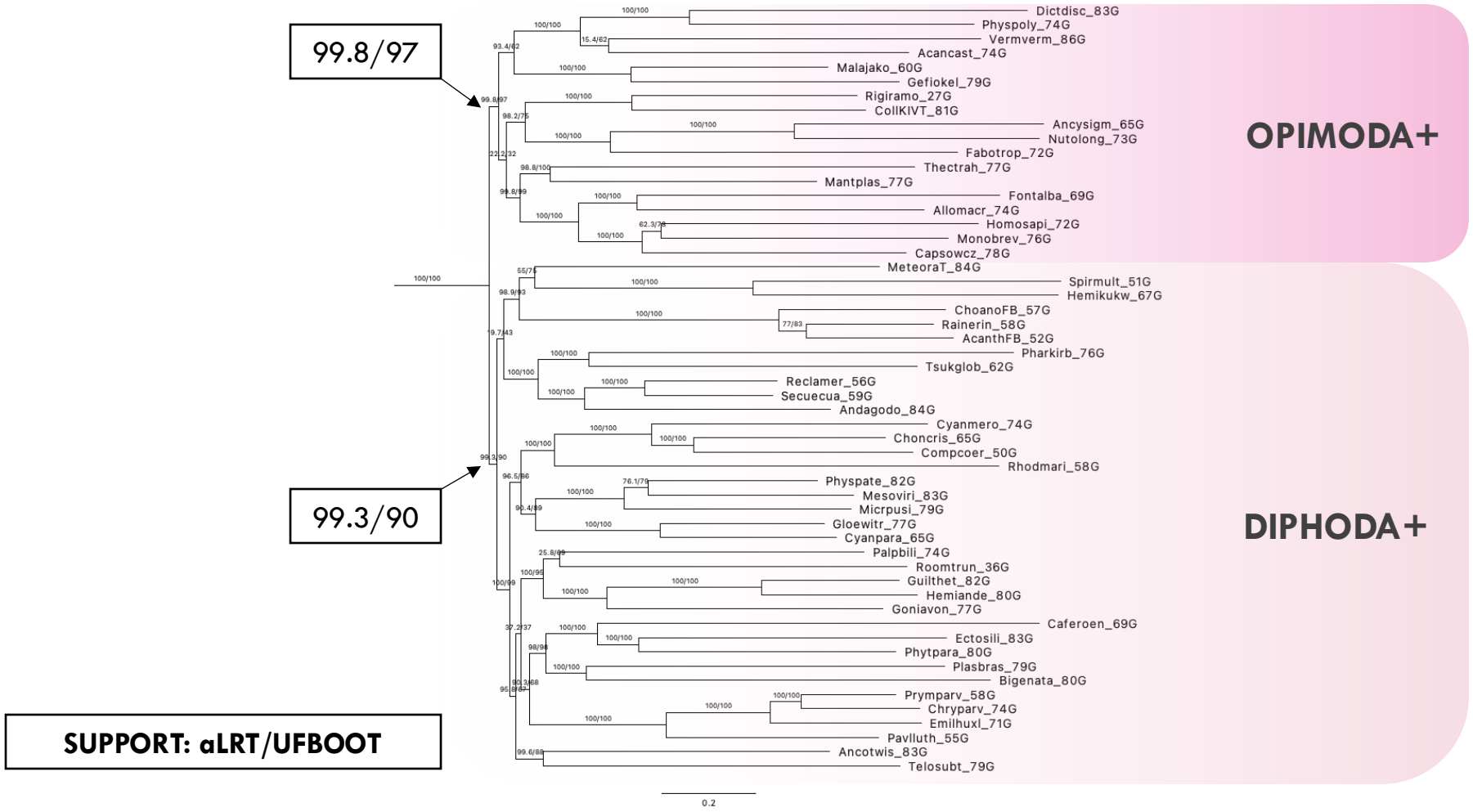
Based on the the **internal branch length** connecting eukaryotes and Alphaproteobacteria

SUPPORT: aLRT/UFBOOT

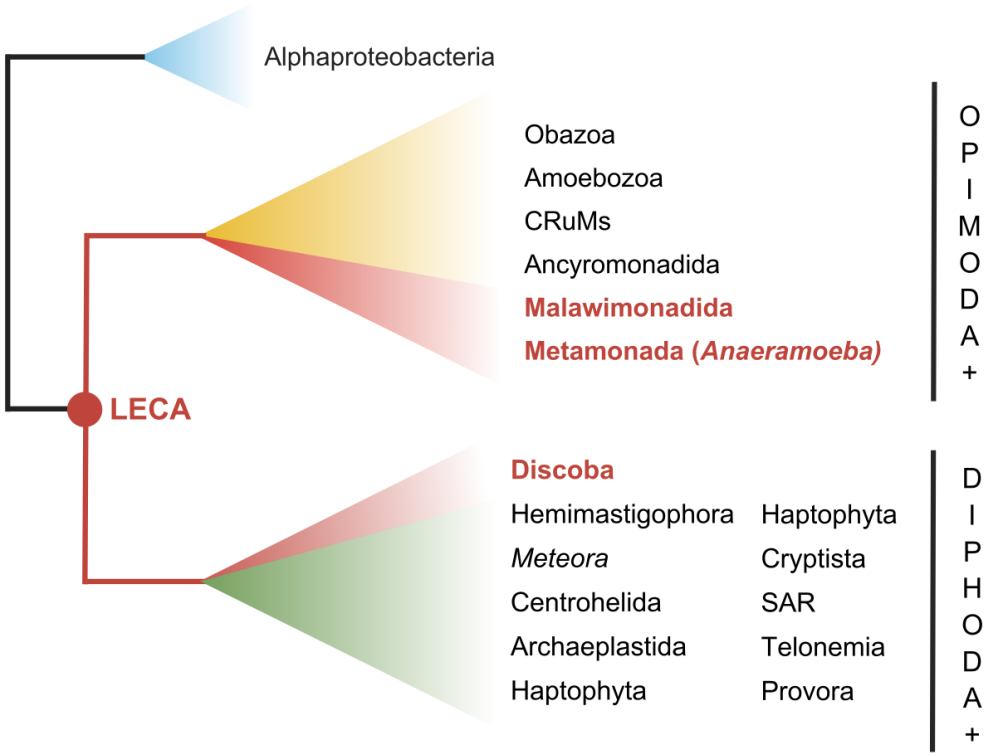


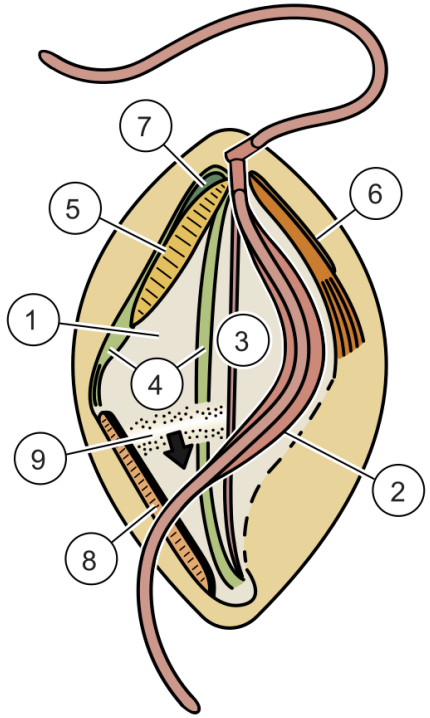
REMOVAL OF 7 FASTEST EVOLVING TAXA

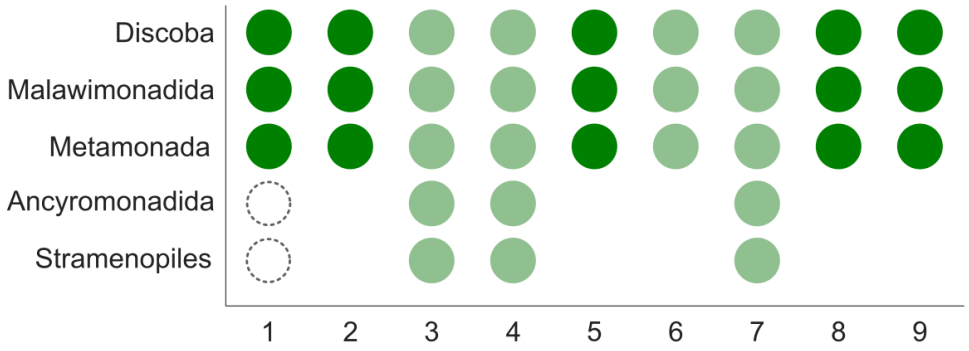
Based on tip-to-tip distance



The recovered root position suggests LECA was an excavate-like organism







1 - Suspension-feeding groove
2 - Flagellar vane
3 - Singlet root
4 - Split 'R2' root
5 - B fibre
6 - C fibre
7 - I fibre
8 - Composite fibre
9 - Wave

Williamson, Eme, et al. *Nature*, 2025

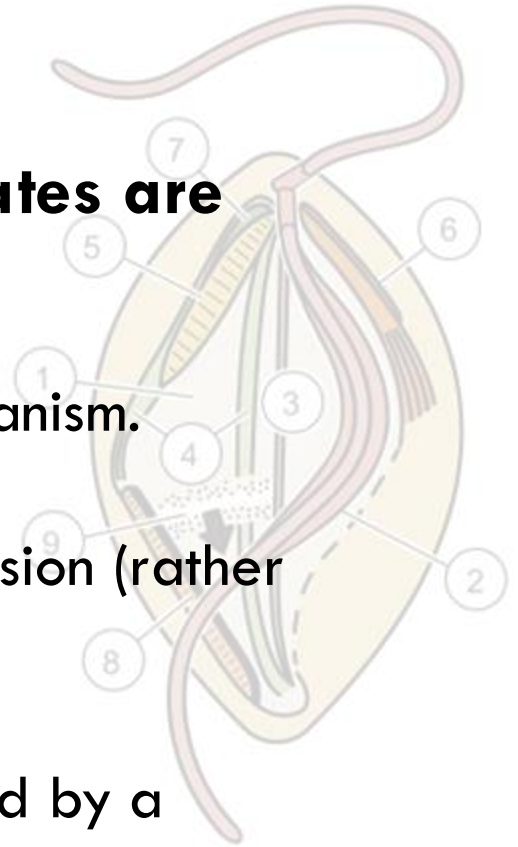
SUMMARY

- **New, much larger mitochondrial protein dataset** for rooting the eukaryote tree

- **All tested models recover an 'Opimoda+' root** in both maximum likelihood and Bayesian frameworks

- Root position is **robust to the removal of the fastest evolving sites, genes, and taxa**

Complex cytoskeletal elements common to ‘typical’ excavates are features of LECA that were lost in other lineages



- LECA was likely a small (25 µm or less), unicellular, and flagellated organism.

- Phagotrophic (likely bacterivorous), probably feeding on prey in suspension (rather than as a surface feeder like many amoebae).

- Had a defined cell shape crucial to its motility and feeding, underpinned by a complex cytoskeletal organisation, including microtubular roots of diverse sizes, functions, and associated non-microtubular elements.

- **Models of eukaryogenesis need to have such a form of eukaryote as their “endpoint”.**

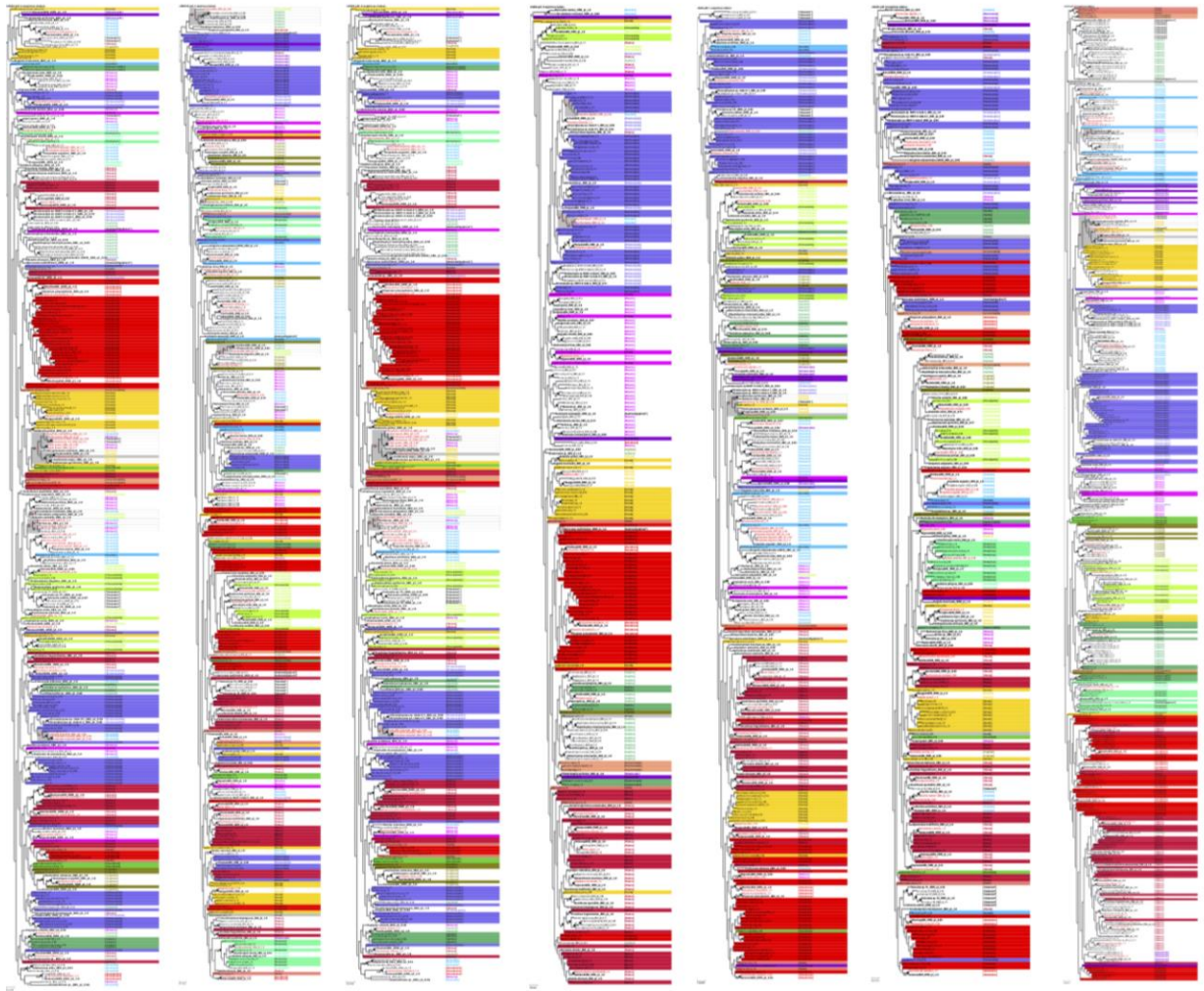
Useful tools

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

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

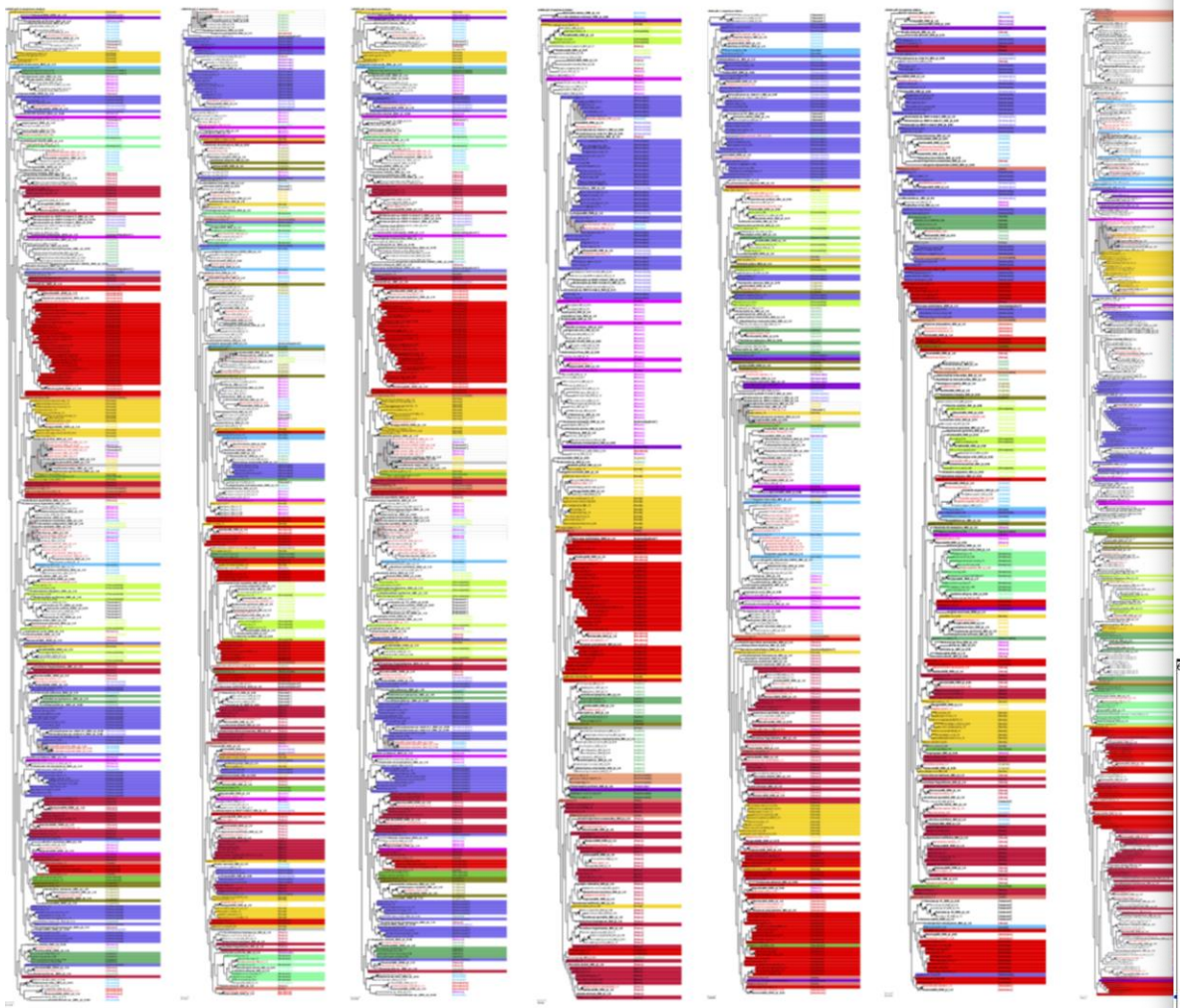
PHYLOFISHER

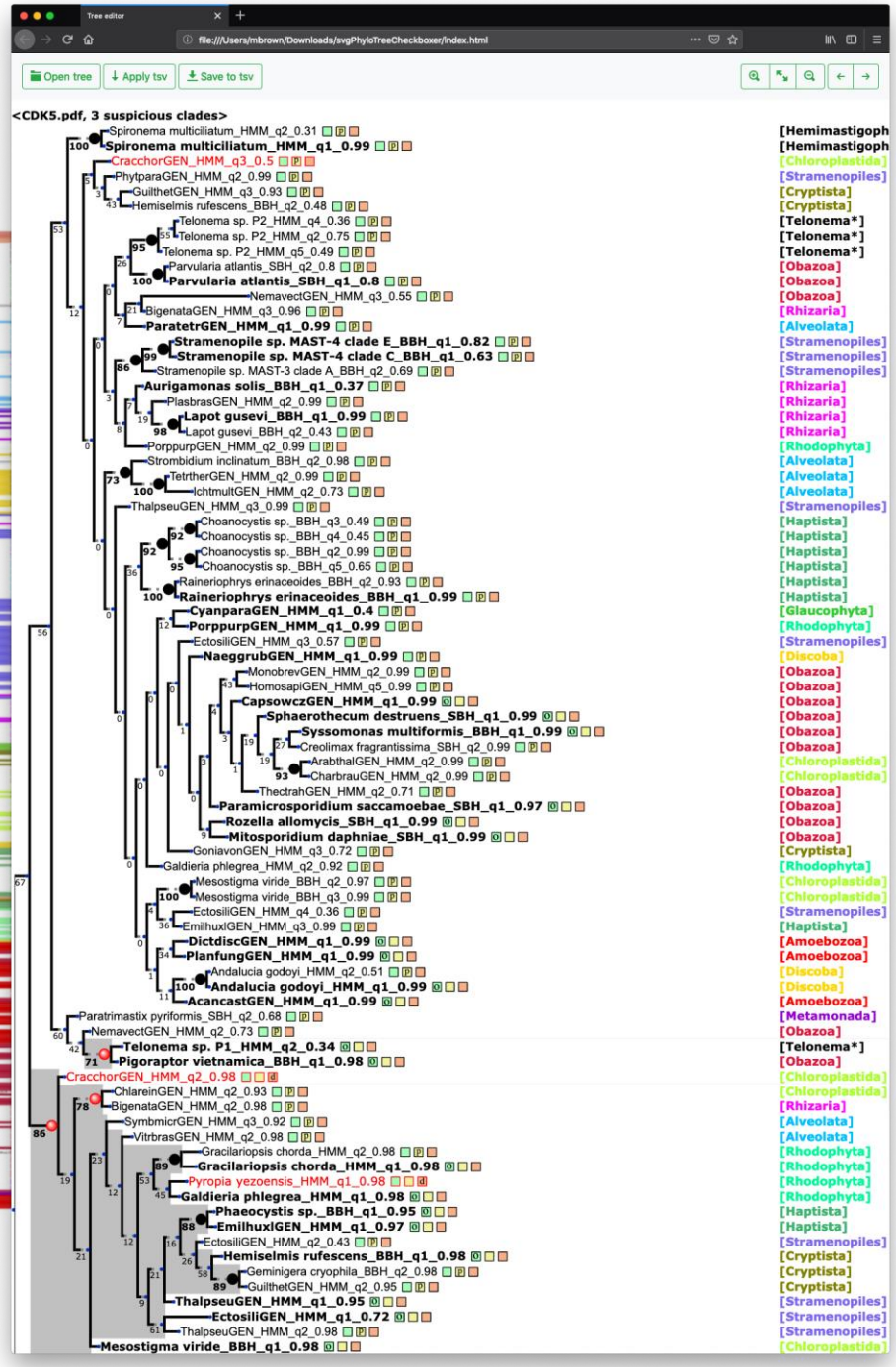
New method (and tool) allows for others to simply do phylogenomics

- 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

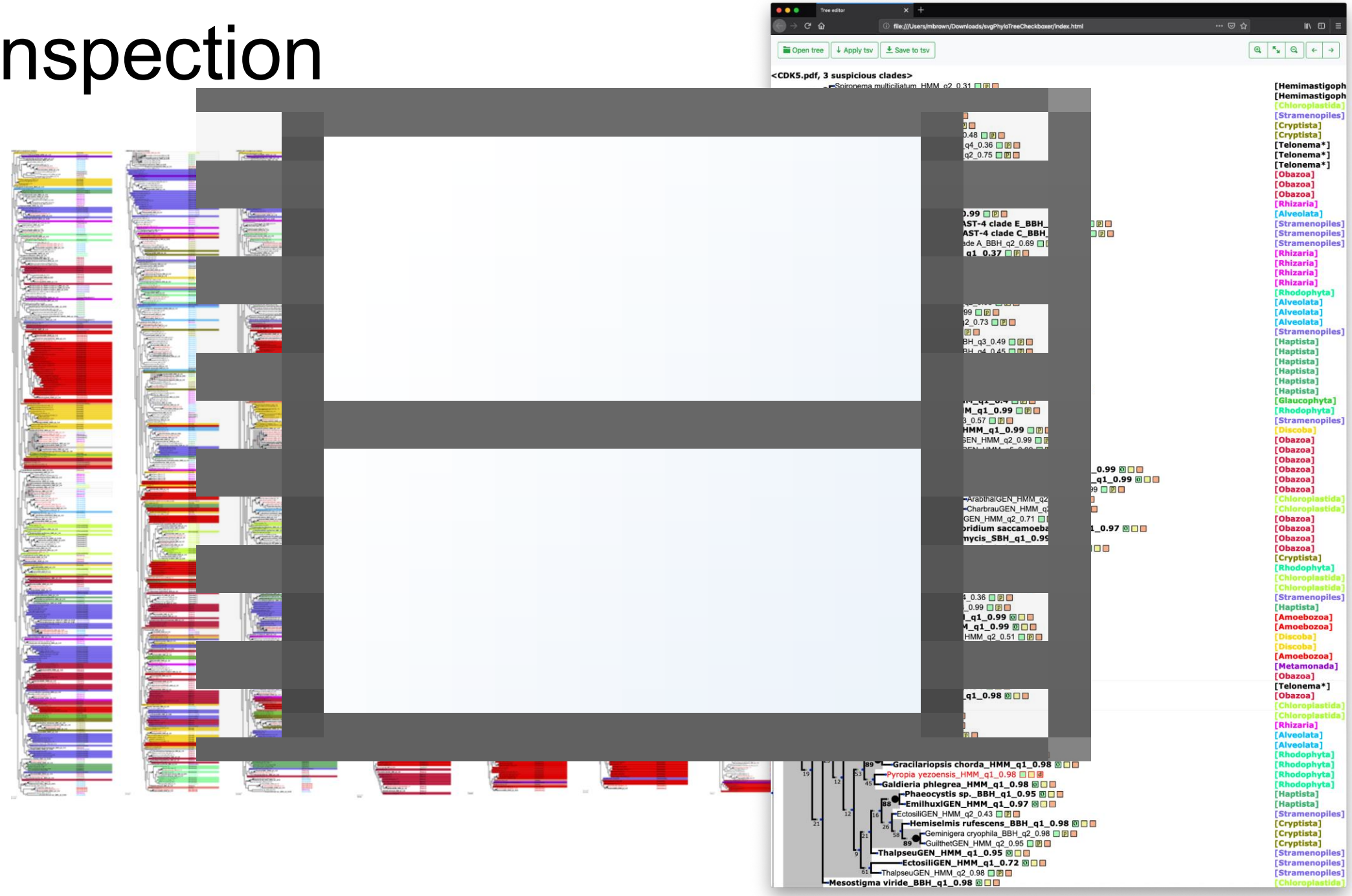

```
conda env create -f fisher_env.yml
```

Tree Inspection



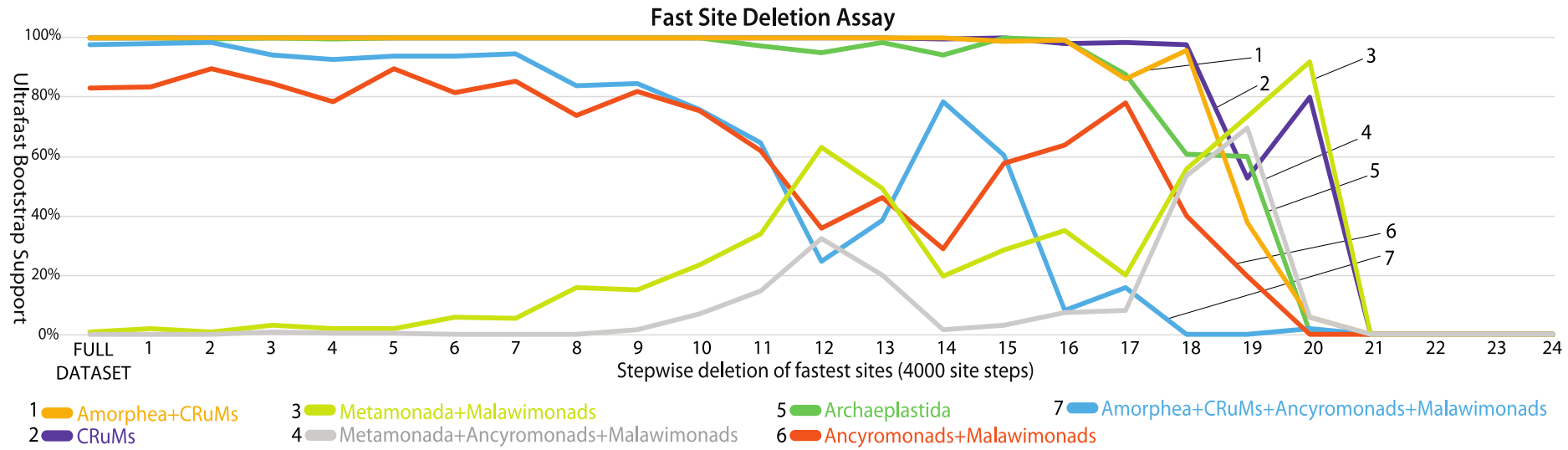


Tree Inspection



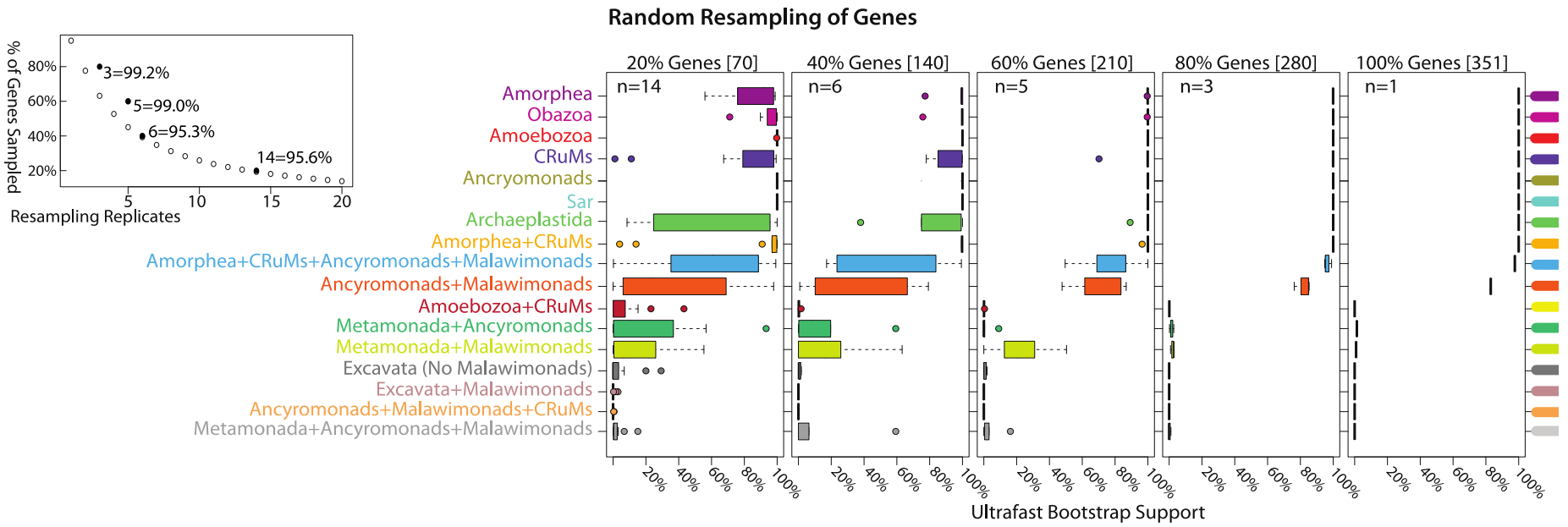
PHYLOFISHER

- 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

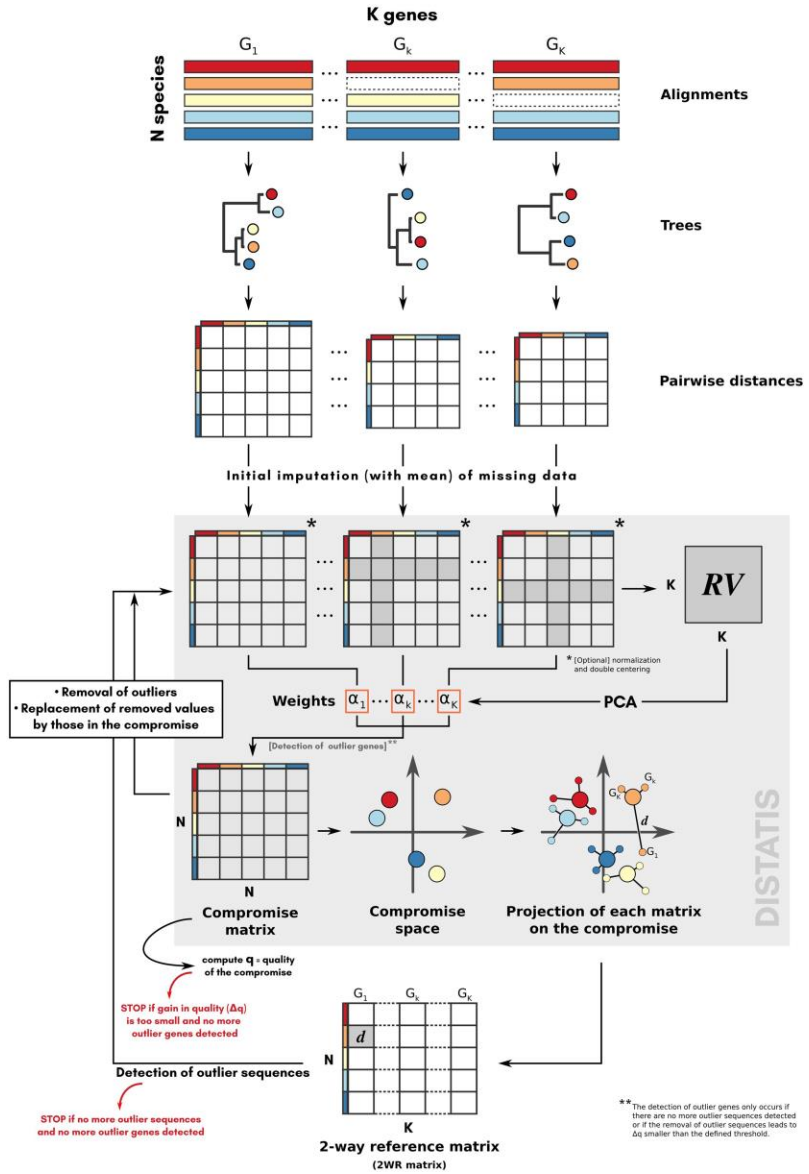


PHYLOFISHER

- 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



PhylteR


K genes
N species
Alignments
Trees
Pairwise distances
Initial imputation (with mean) of missing data
RV
* [Optional] normalization and double centering
• Removal of outliers
• Replacement of removed values by those in the compromise
Weights
PCA
[Detection of outlier genes]**
Compromise matrix
Compromise space
Projection of each matrix on the compromise
DISTATIS
compute q = quality of the compromise
STOP if gain in quality (Δq) is too small and no more outlier genes detected
Detection of outlier sequences
STOP if no more outlier sequences and no more outlier genes detected
2-way reference matrix
(2WR matrix)
** The detection of outlier genes only occurs if there are no more outlier sequences detected or if the removal of outlier sequences leads to Δq smaller than the defined threshold.

ETE3

ETE Toolkit

A Python framework to work with trees

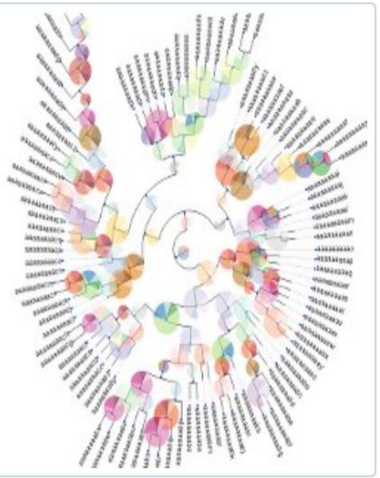
A Python framework for the analysis and visualization of trees.

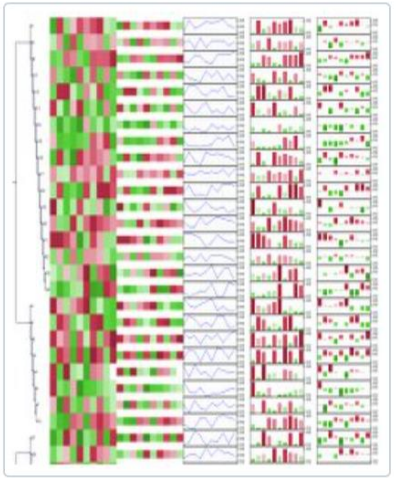
Download | Python API | Cookbook | Phylogenomic tools | Contribute

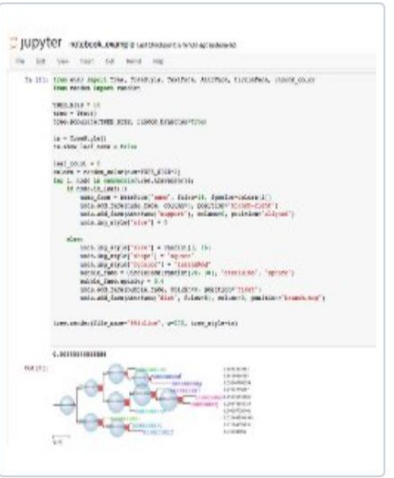
```
from ete3 import Tree
tree = Tree('((A,B), D);')

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.