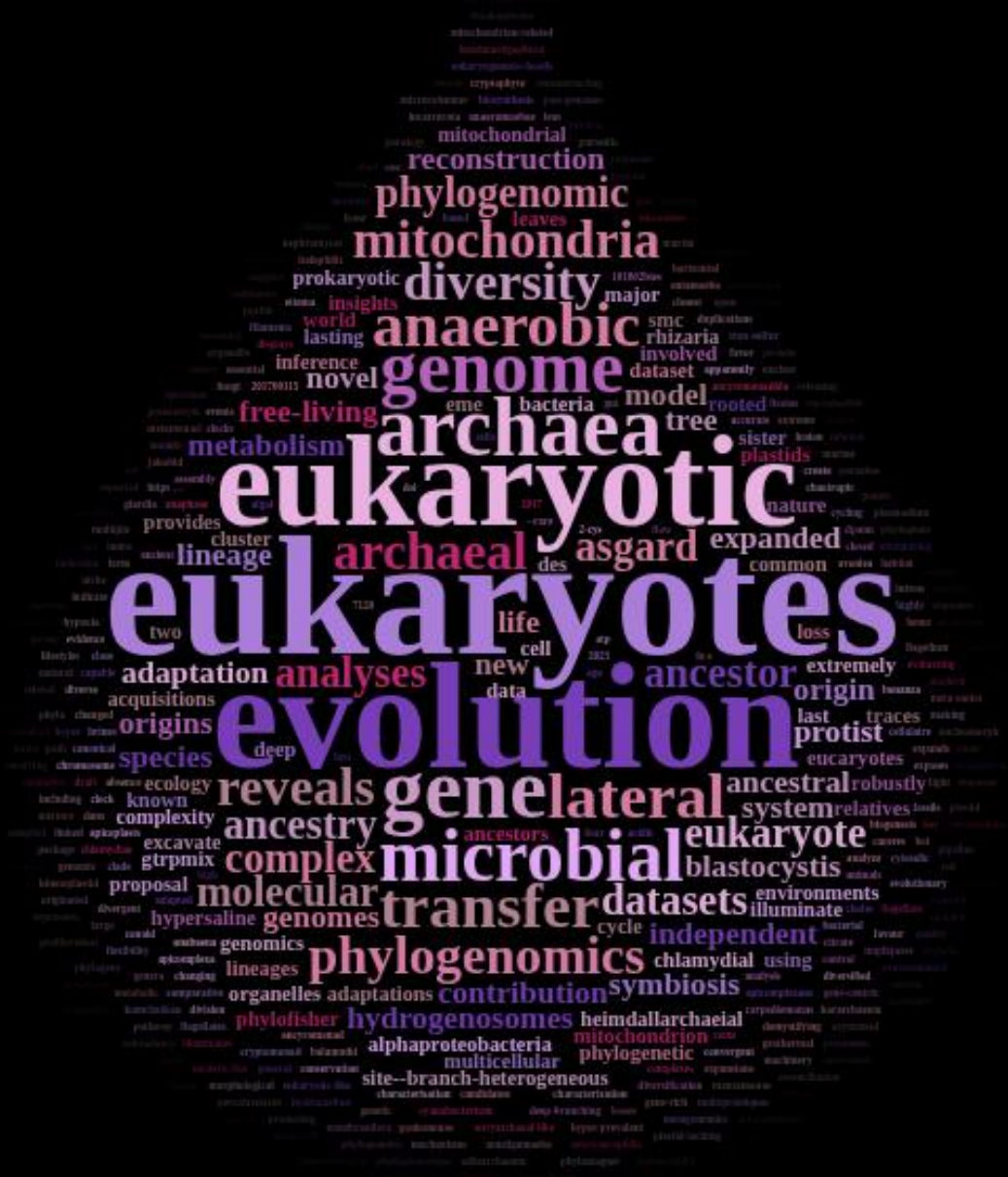
“Deep” phylogenetics

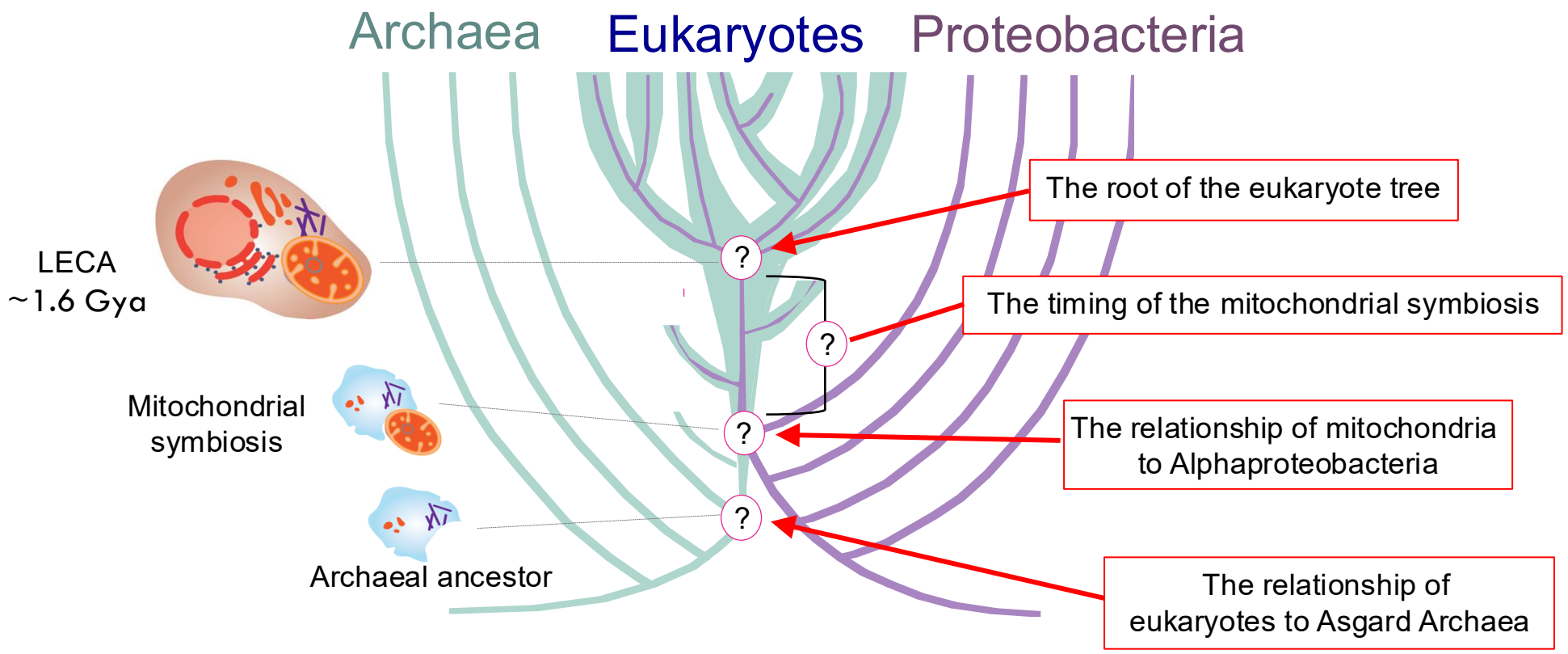
Laura Eme
Associate Professor
University of Rhode Island

@lauraeme.bsky.social



https://shiny.research.sfu.ca/u/rdmorin/scholargoggler10/

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

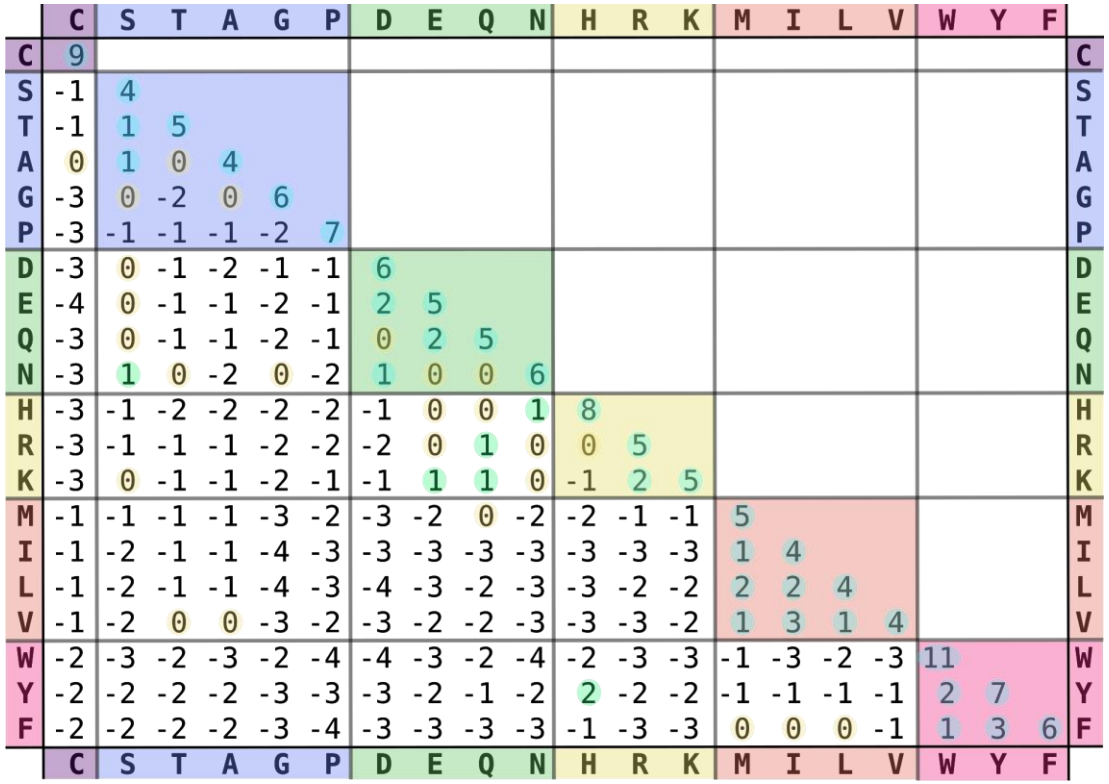
Code degeneracy

- 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



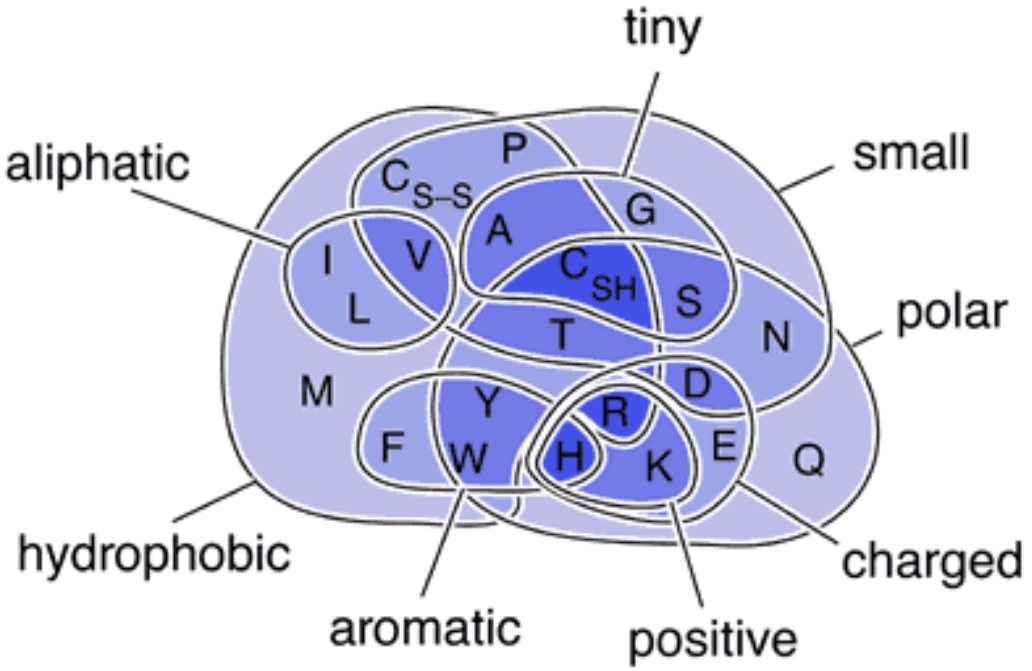
	C	S	T	A	G	P	D	E	Q	N	H	R	K	M	I	L	V	W	Y	F	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
A	0	1	0	4																	A
G	-3	0	-2	0	6																G
P	-3	-1	-1	-1	-2	7															P
D	-3	0	-1	-2	-1	-1	6														D
E	-4	0	-1	-1	-2	-1	2	5													E
Q	-3	0	-1	-1	-2	-1	0	2	5												Q
N	-3	1	0	-2	0	-2	1	0	0	6											N
H	-3	-1	-2	-2	-2	-2	-1	0	0	1	8										H
R	-3	-1	-1	-1	-2	-2	-2	0	1	0	0	5									R
K	-3	0	-1	-1	-2	-1	-1	1	1	0	-1	2	5								K
M	-1	-1	-1	-1	-3	-2	-3	-2	0	-2	-2	-1	-1	5							M
I	-1	-2	-1	-1	-4	-3	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-1	-4	-3	-4	-3	-2	-3	-3	-2	-2	2	2	4					L
V	-1	-2	0	0	-3	-2	-3	-2	-2	-3	-3	-3	-2	1	3	1	4				V
W	-2	-3	-2	-3	-2	-4	-4	-3	-2	-4	-2	-3	-3	-1	-3	-2	-3	11			W
Y	-2	-2	-2	-2	-3	-3	-3	-2	-1	-2	2	-2	-2	-1	-1	-1	-1	2	7		Y
F	-2	-2	-2	-2	-3	-4	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	1	3	6	F
	C	S	T	A	G	P	D	E	Q	N	H	R	K	M	I	L	V	W	Y	F	

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(π1, …, π20) = 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

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

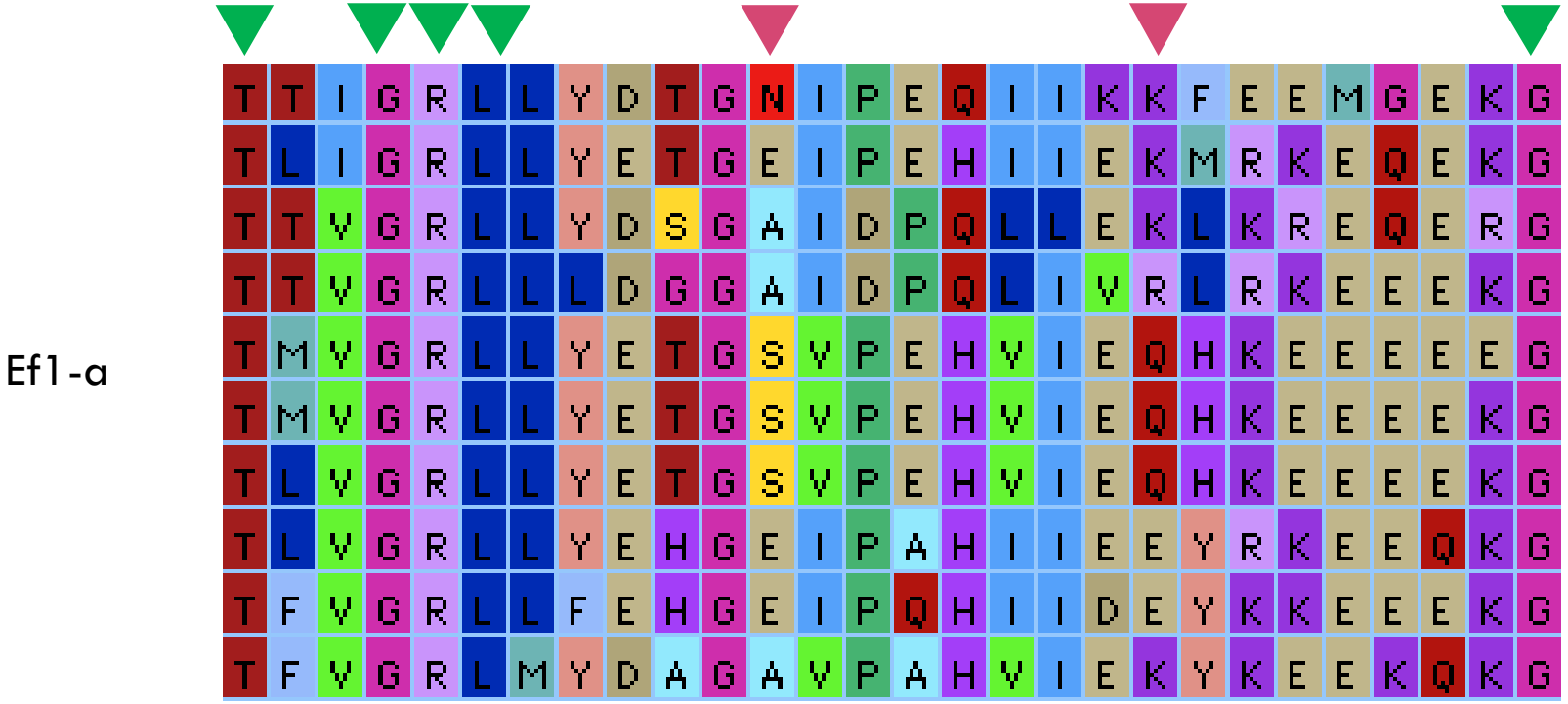
- With recent software, you can correct the πi values based on the observed frequencies in your data ("+F" option). E.g. LG+G+F

Warning *(Banos, Roger, Susko 2023): "it can decrease the probability of correct tree estimation, depending on the scenario, despite the fact that it tends to improve likelihood scores."*

Likelihood ≠ accuracy

Rate heterogeneity parameter

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



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
- IQ-TREE can pick for you, but worth knowing the difference.

1.2 Fully parameterized time-reversible model

What if your data doesn't fit any standard matrix?

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)

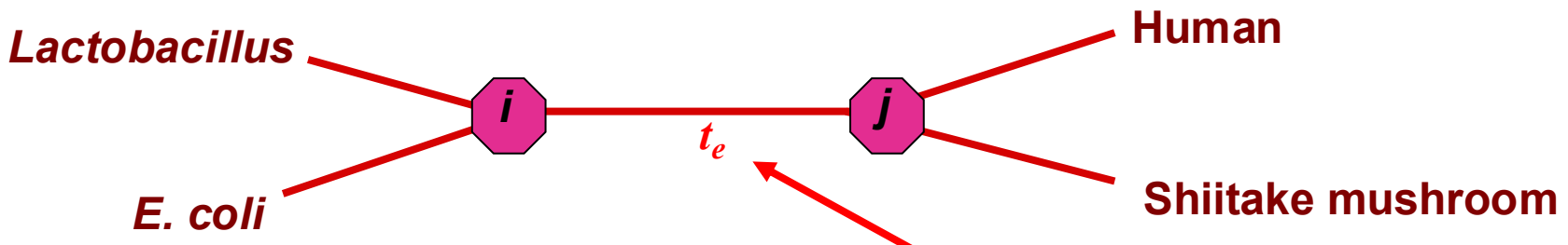
QMaker and dataset-specific empirical models

- Empirical matrices (LG, WAG, JTT) were estimated decades ago from databases dominated by a few well-studied taxa
- **QMaker:** a ML method to estimate a GTR Q matrix from a large set of alignments (available in IQ-TREE)
- Released a panel of new clade-specific empirical matrices:
  - **Q.pfam:** general (Pfam-trained), often outperforms LG
  - **Q.plant, Q.mammal, Q.bird, Q.insect, Q.yeast:** clade-specific
- In practice: just add -mset Q.pfam,LG,WAG,JTT to ModelFinder (let it pick)

Minh et al. 2021, *Syst Biol* 70:1046

1.3 Mixture models

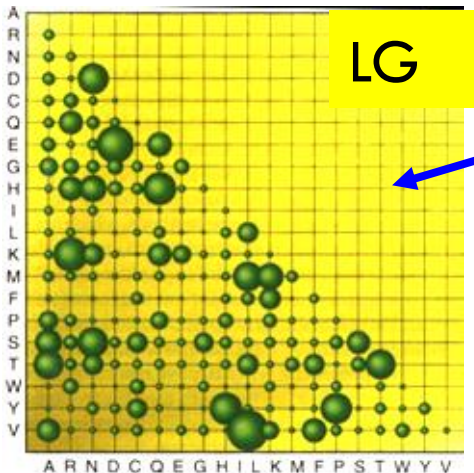
Your model is giving you the probability of going from amino acid *i* to *j* at site *x*, evolving at rate r_v on branch t_e



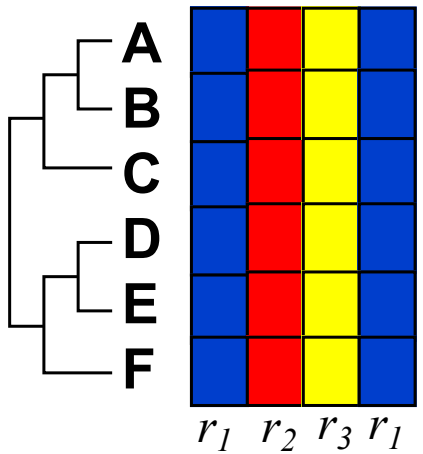
$$P(j \mid i;t) = \left[\exp(Q \times t_e \times r_v)\right]_{ij}$$

For $i \neq j$:

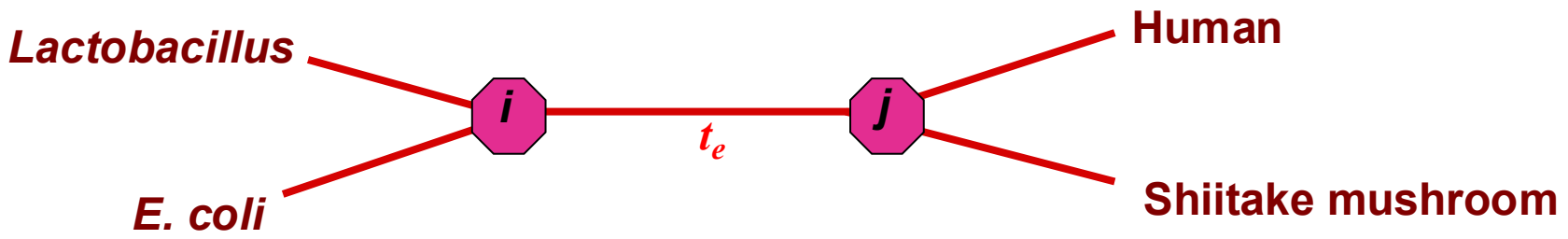
$$q_{ij} = r_{ij} \times \pi_j$$



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



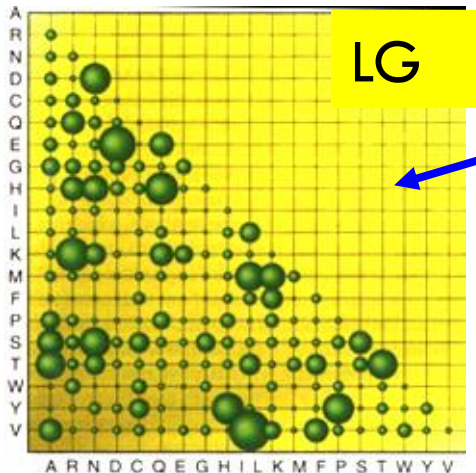
Your model is giving you the probability of going from amino acid *i* to *j* at site *x*, evolving at rate r_v on branch t_e



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

For $i \neq j$:

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



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

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

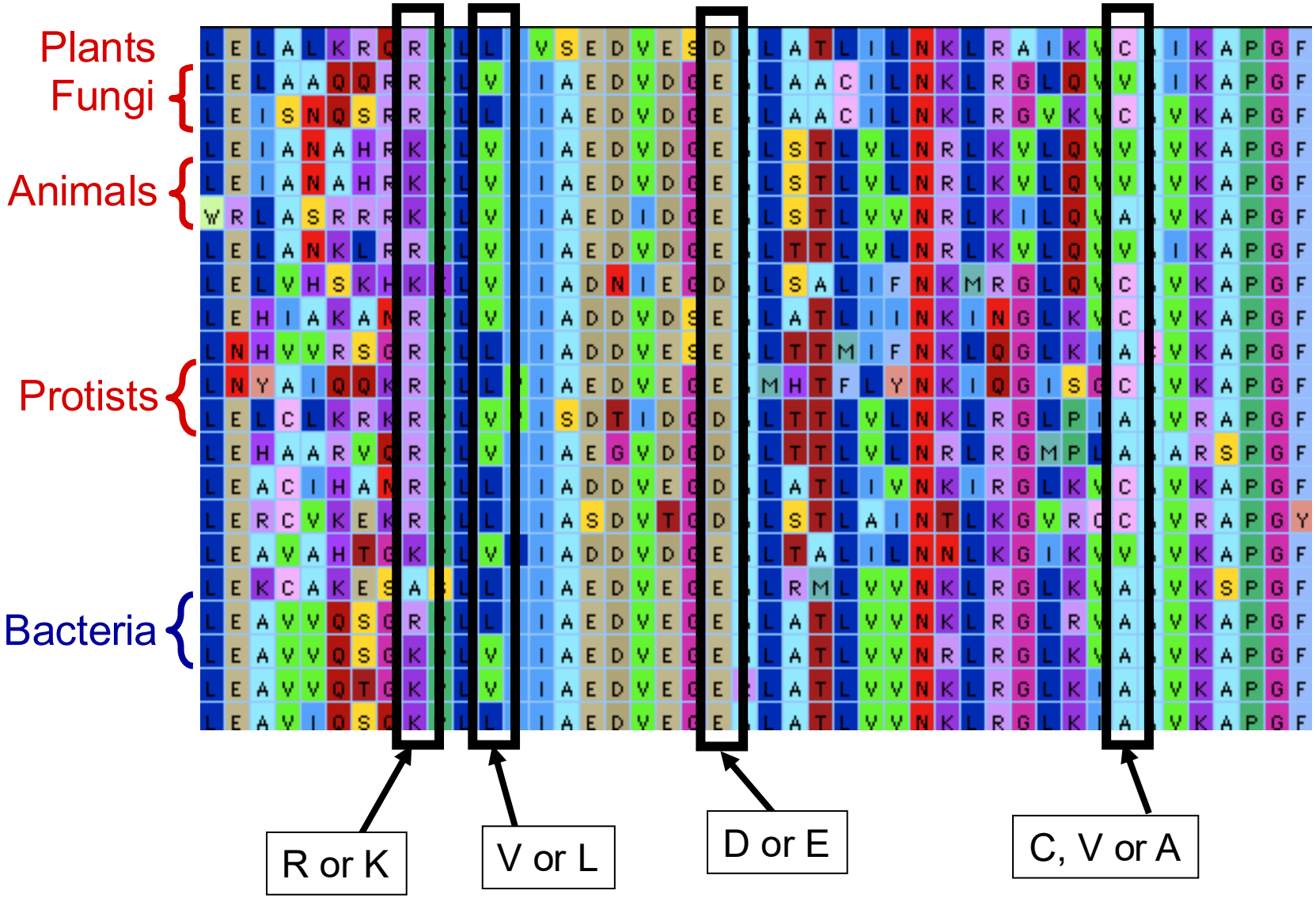
The problem…

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

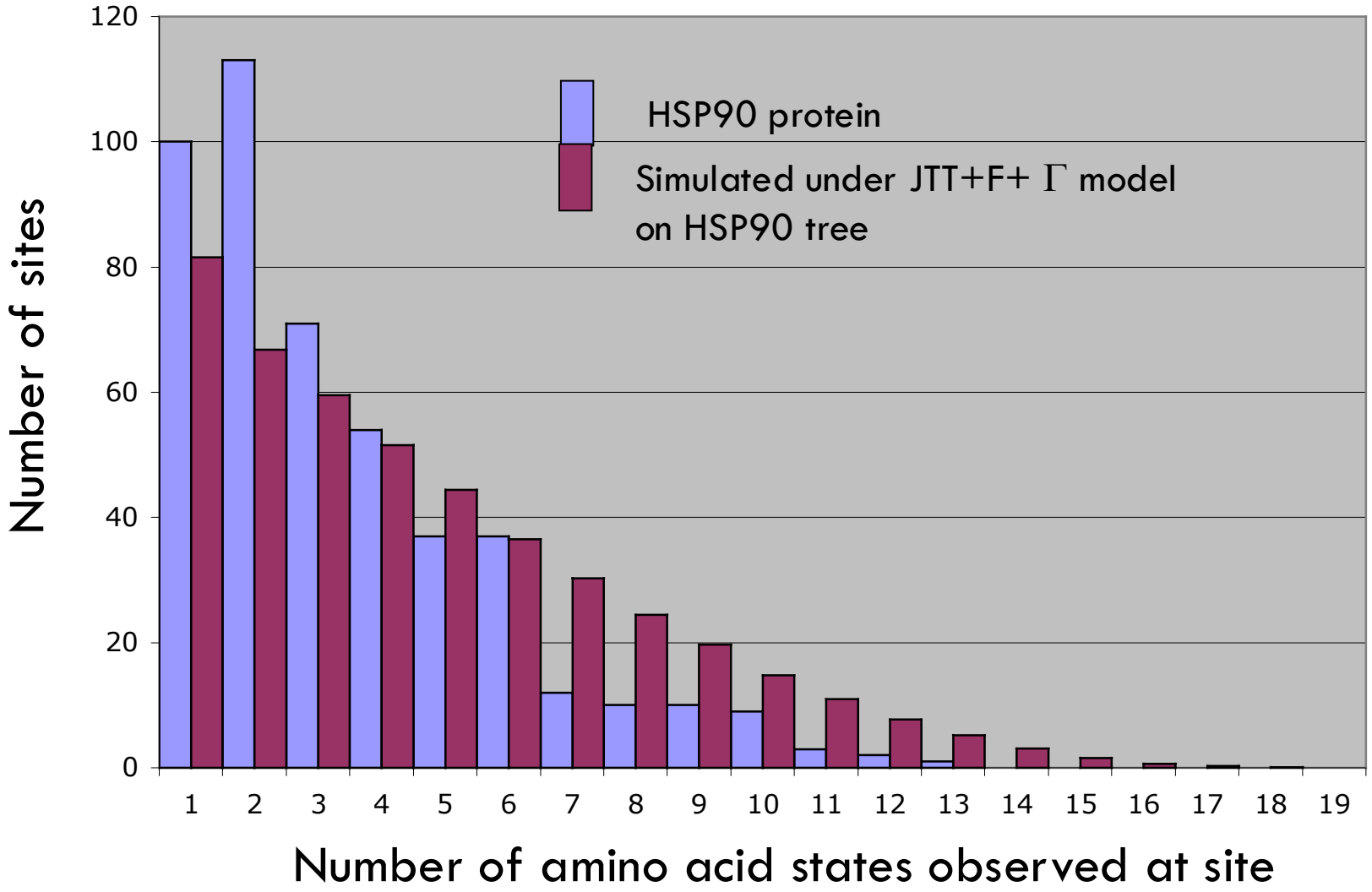
- Sites in proteins can change function over time
  - sites under negative selection ⟷ neutral ⟷ positive selection

- 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

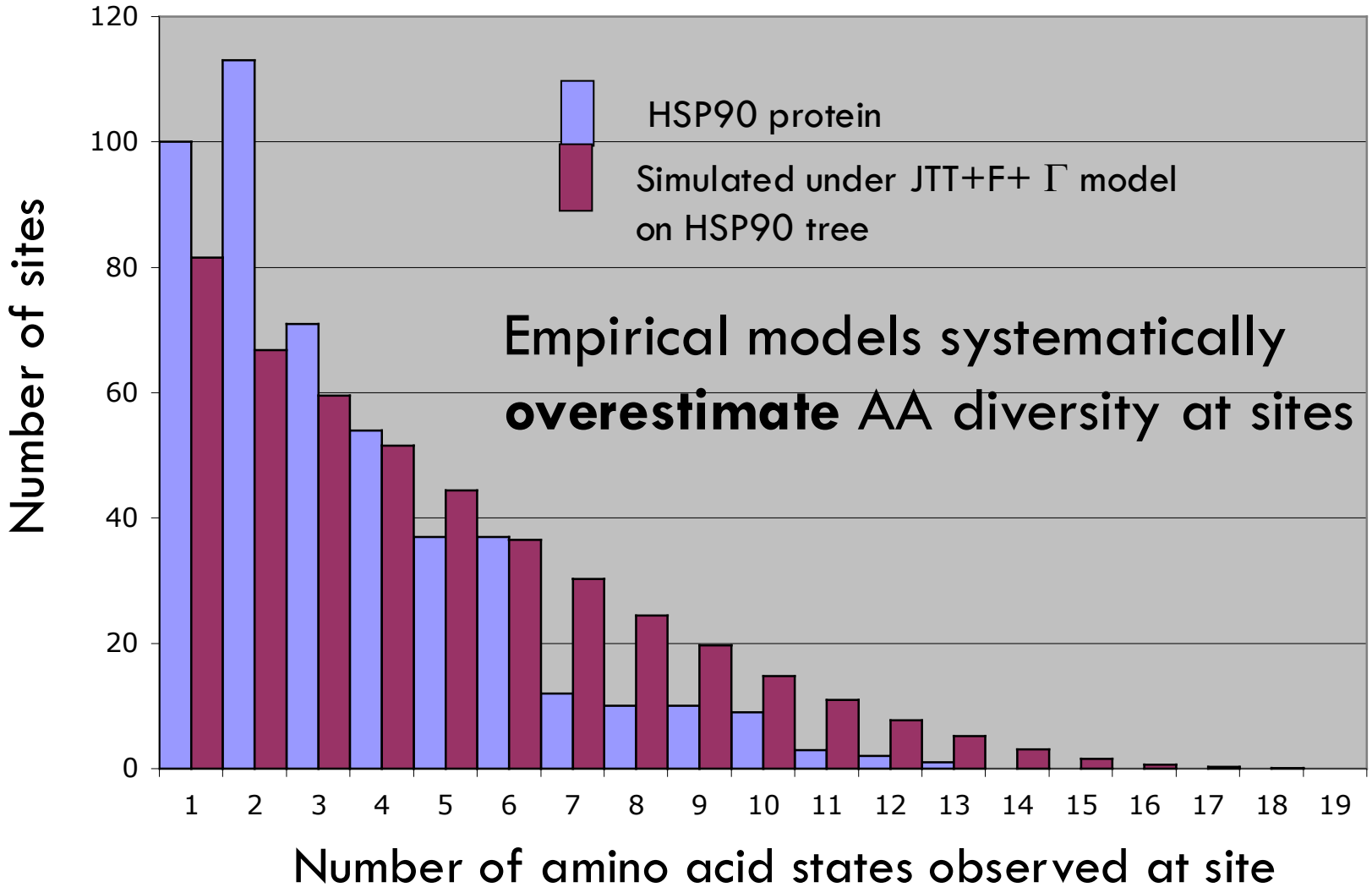


Distribution of the number of different amino acids at aligned sites



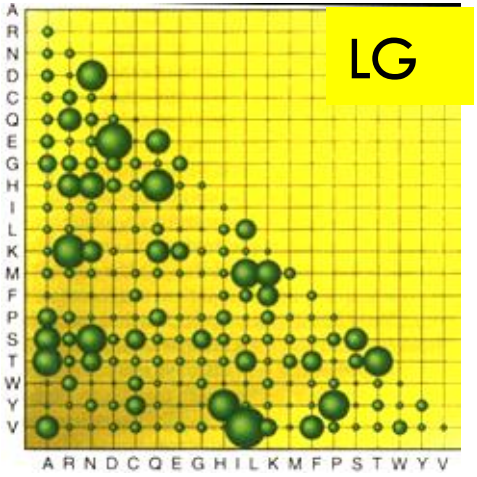
Wang *et al.* (2008) *BMC Evolutionary Biology* 8: 331

Distribution of the number of different amino acids at aligned sites

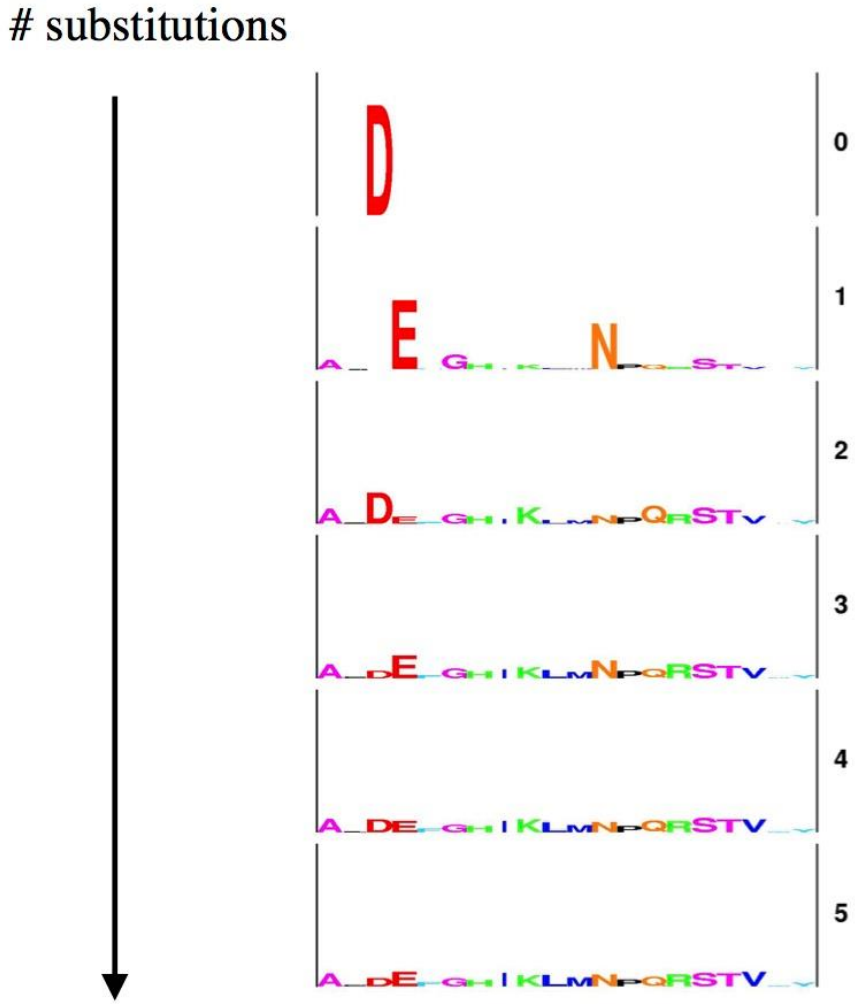


Wang *et al.* (2008) *BMC Evolutionary Biology* 8: 331

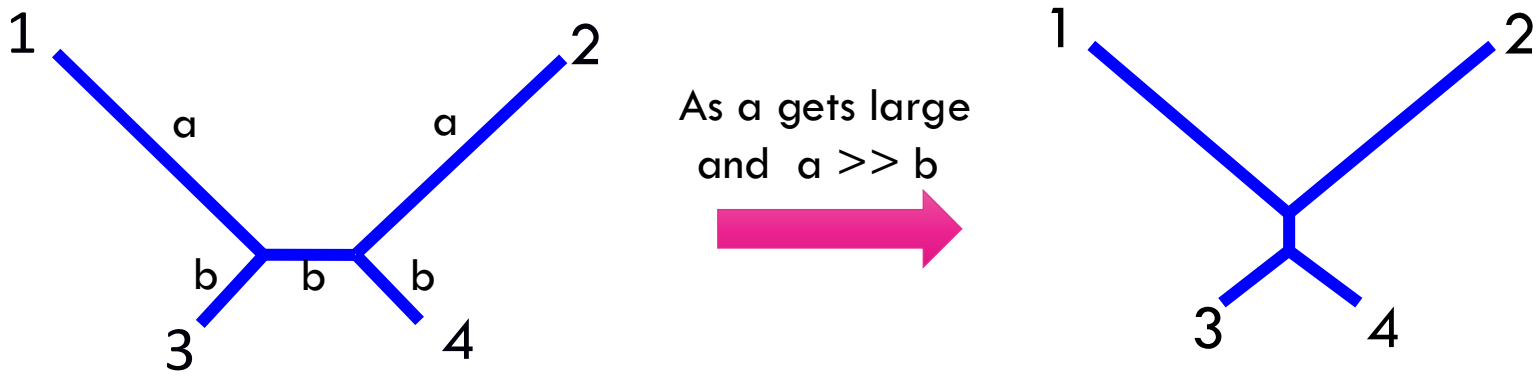
Starting at a D with a site homogeneous matrix (LG+F)



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



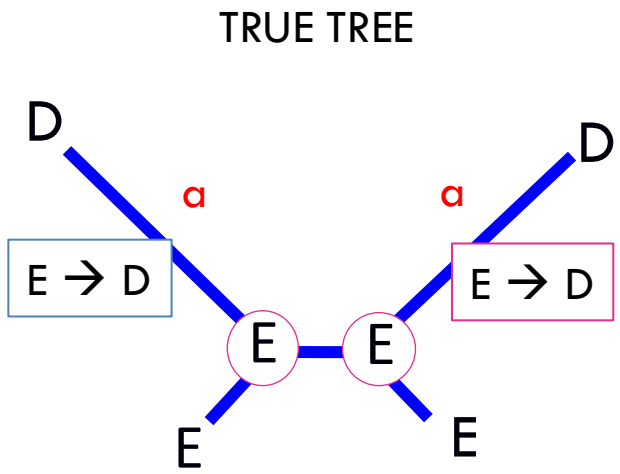
What happens to phylogenetic estimation when you ignore site-heterogeneity?

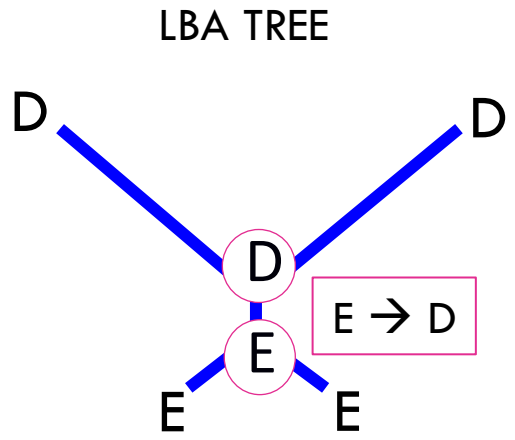


Long branch attraction

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)?

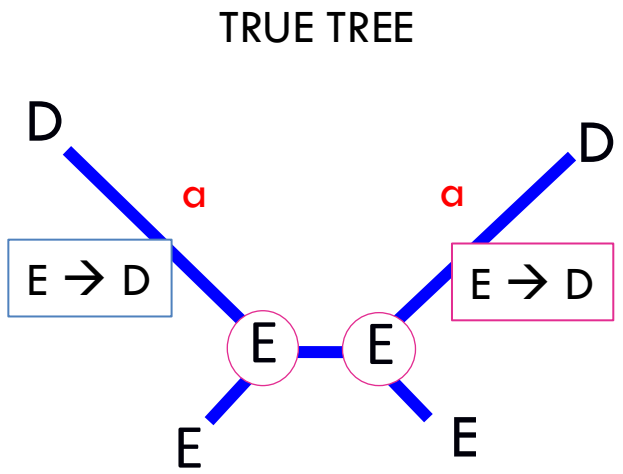


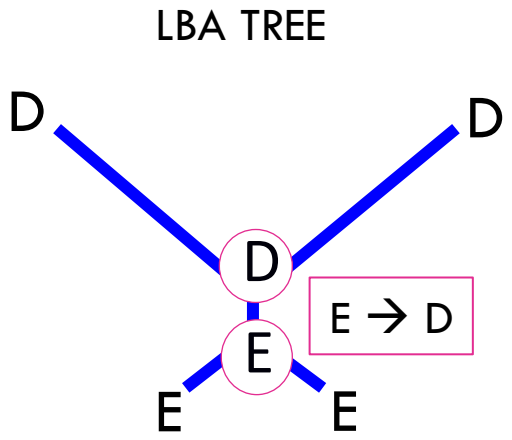


Under site homogeneous model (LG), the probability of converging on the same state (E → D) twice is pretty low:

➔ 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 part of the mixture model

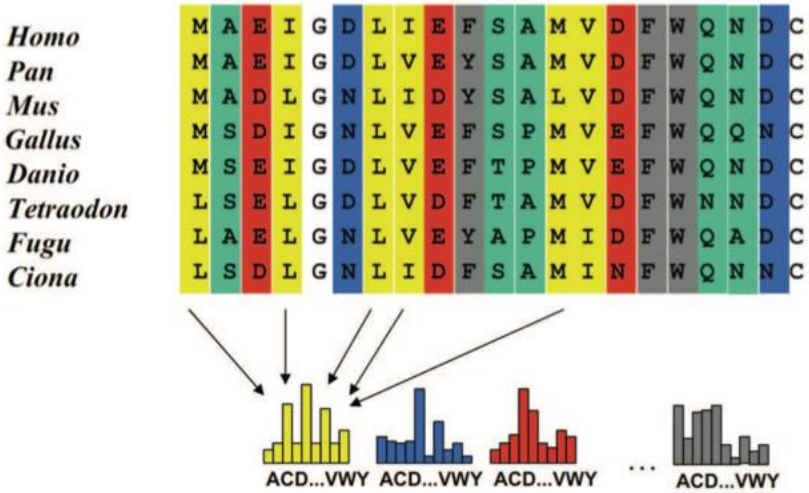
$$L_i = p(\mathbf{y}_i|r_1)p(r_1) + p(\mathbf{y}_i|r_2)p(r_2) + \cdots + p(\mathbf{y}_i|r_k)p(r_k)$$

Weight of the rate class

Mixture models: terminology warning

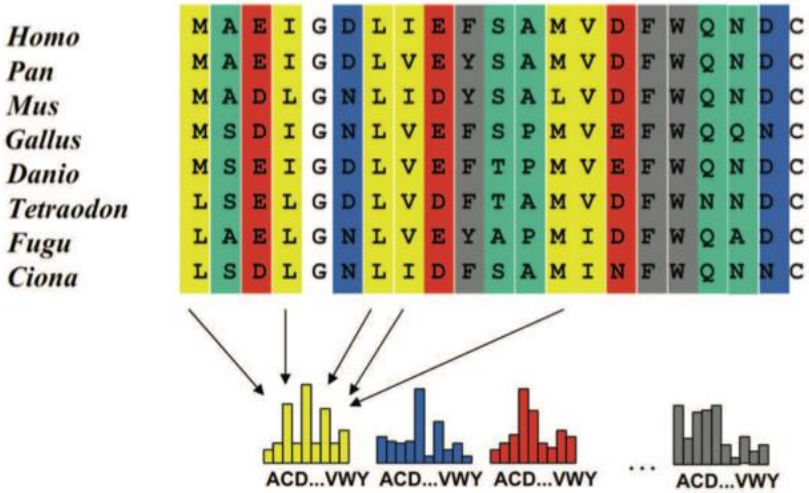
- Different kinds of mixture models!
- Rate-category mixture model
- **Usually people refer to mixture of amino-acid replacement matrices**
- Mixtures can be apply to any part of the model (e.g., branch lengths)

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

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

GTRpmix and the new linked mixture models

- LG was estimated assuming **one profile for all sites,** but we use it combined with C60 (which assumes 60 profiles)
- Mismatch: exchangeabilities and profiles are capturing some of the same signal twice
- **GTRpmix**: ML estimation of one set of exchangeabilities *jointly with* a profile mixture

Baños, Wong, Daneau, Susko, Minh, Lanfear, Brown, Eme, Roger, 2024 *Molecular Biology and Evolution*

GTRpmix and the new linked mixture models

- Two ready-to-use matrices, estimated under C60:
  - **ELM** (Eukaryotic Linked Mixture) — for eukaryote-only phylogenomics
  - **EAL** (Eukaryote-Archaea Linked) — for eukaryote-prokaryote questions

- Drop-in replacement: -m ELM+C60+G or -m EAL+C60+G in IQ-TREE

- **Warning:** ELM and EAL underperform LG when used alone (+F);
  they only work *with* a profile mixture

Baños, Wong, Daneau, Susko, Minh, Lanfear, Brown, Eme, Roger, 2024 *Molecular Biology and Evolution*

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 ($f_{j,x}$)

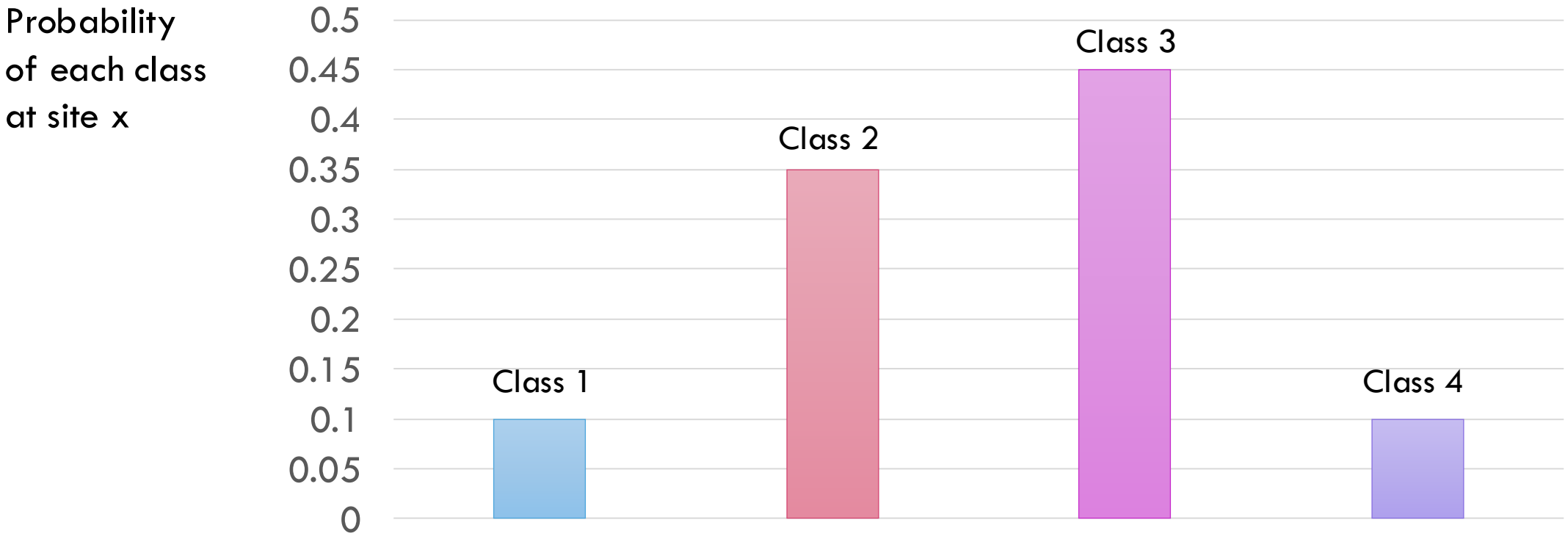
$$f_{j,x} = \sum_c f_{j,c} \times P(c|x)$$

Sum over all classes

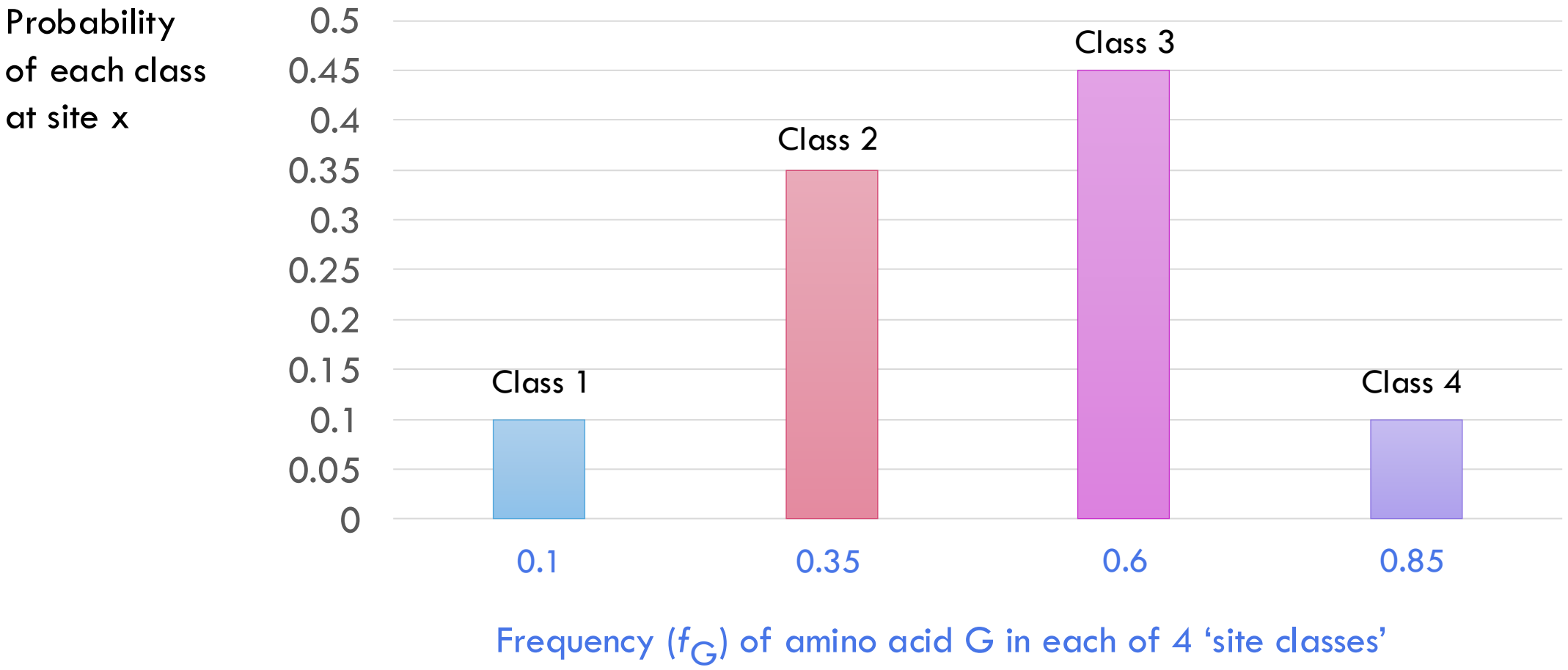
Freq of AA j for class c

Prob of class c at site x

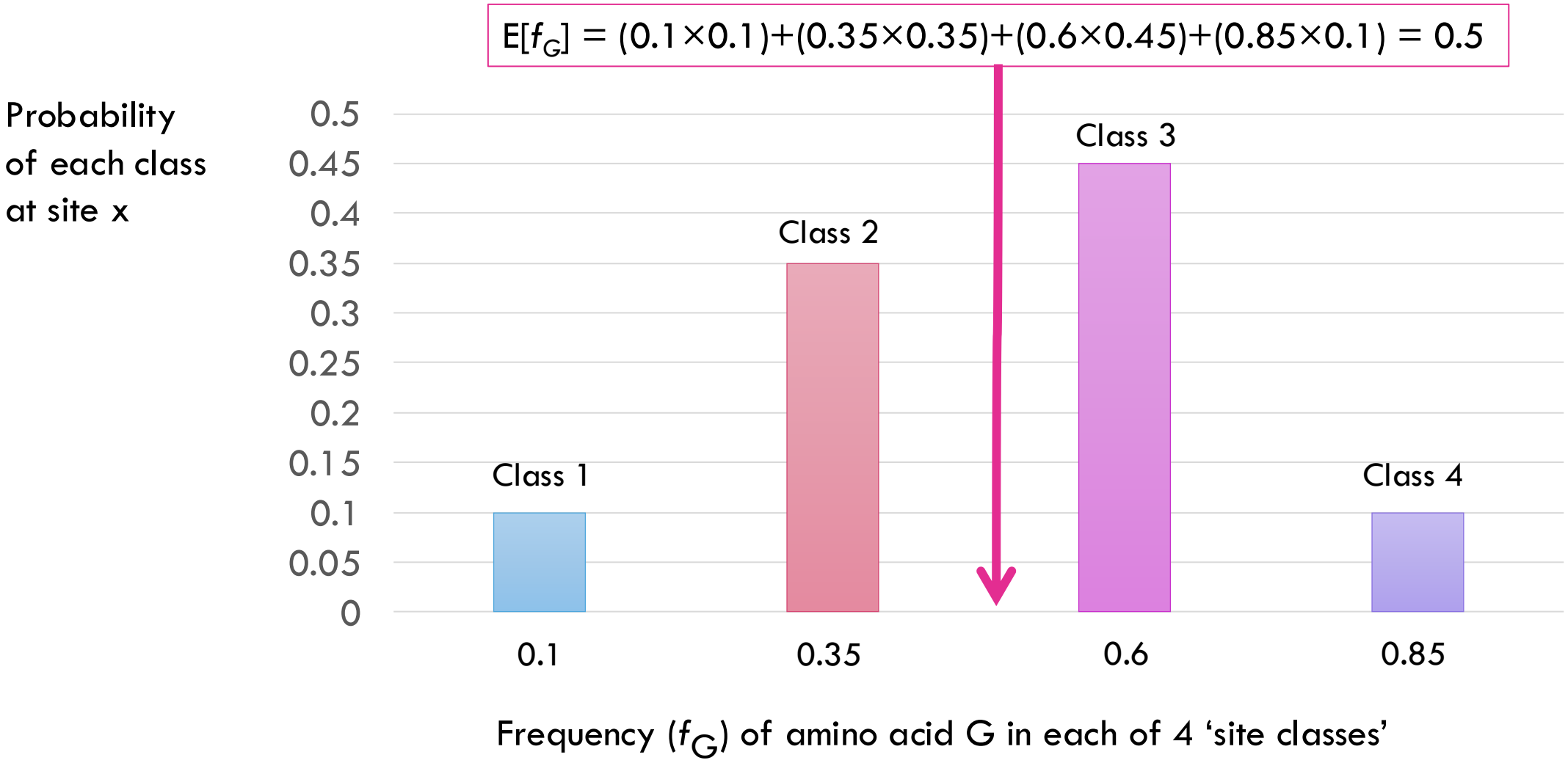
Example: Posterior mean site frequency for 'G' at a given site x, with a 4 class mixture model



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

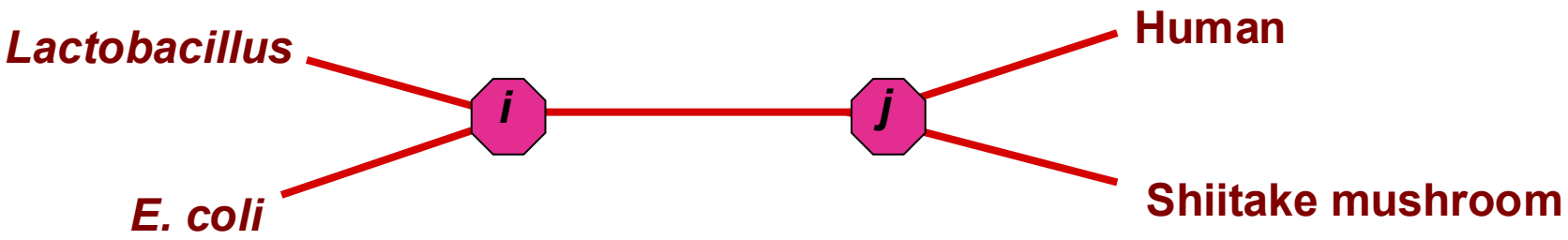
Take home (for this part)

- 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

(but really, this is heterotachy)

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

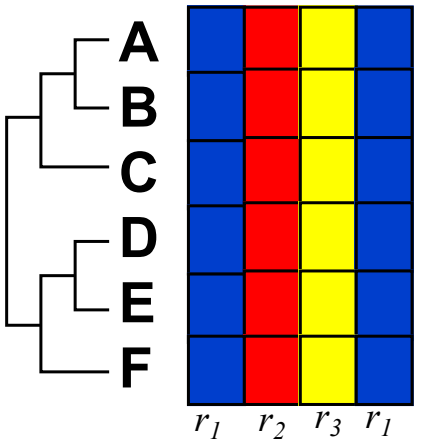


$$P(j \mid i;t) = \left[\exp(R \square \Pi \square t_e \square r_v)\right]_{ij}$$

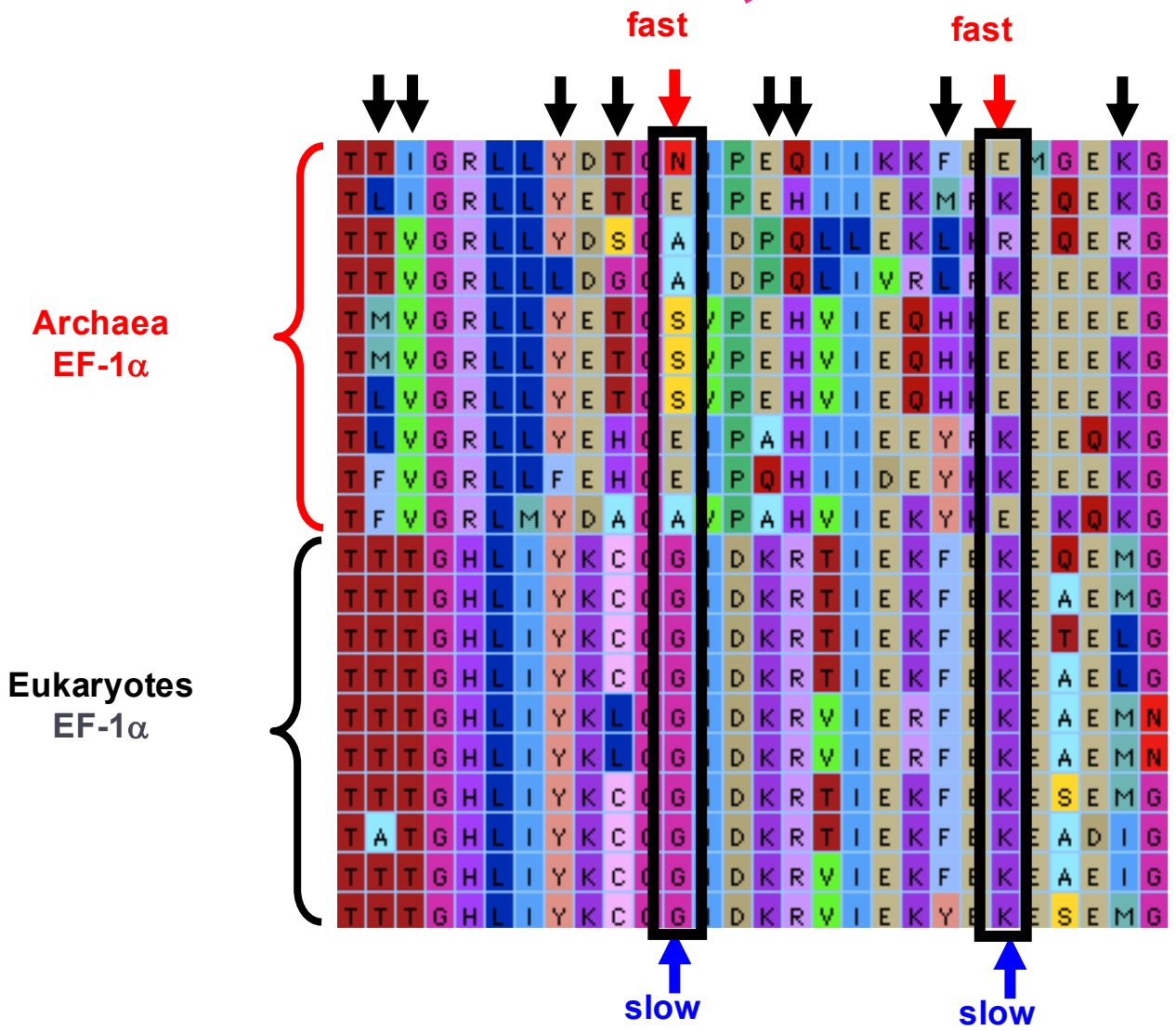
Assumptions

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

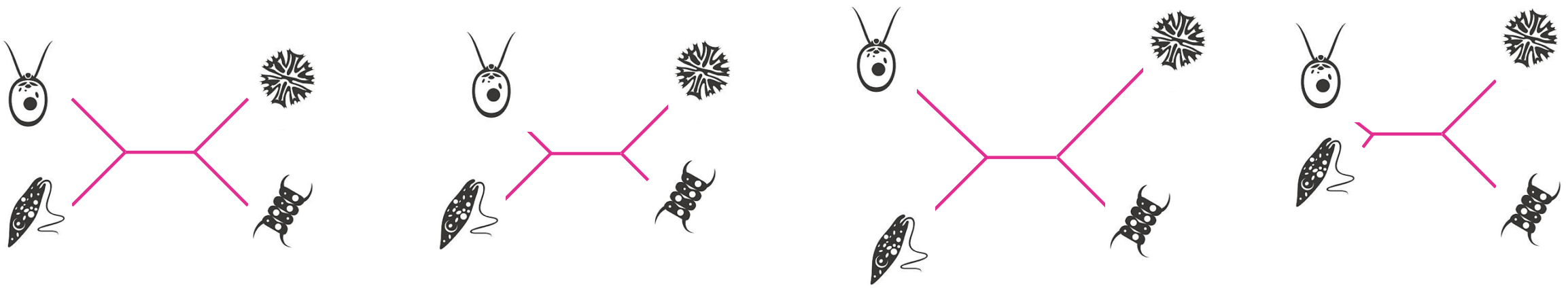
-Sites have the same rate of evolution (r_v) on different branches of tree



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
  - Allow the sites "switch" between high rates and low rates over the tree
  - Computationally intensive

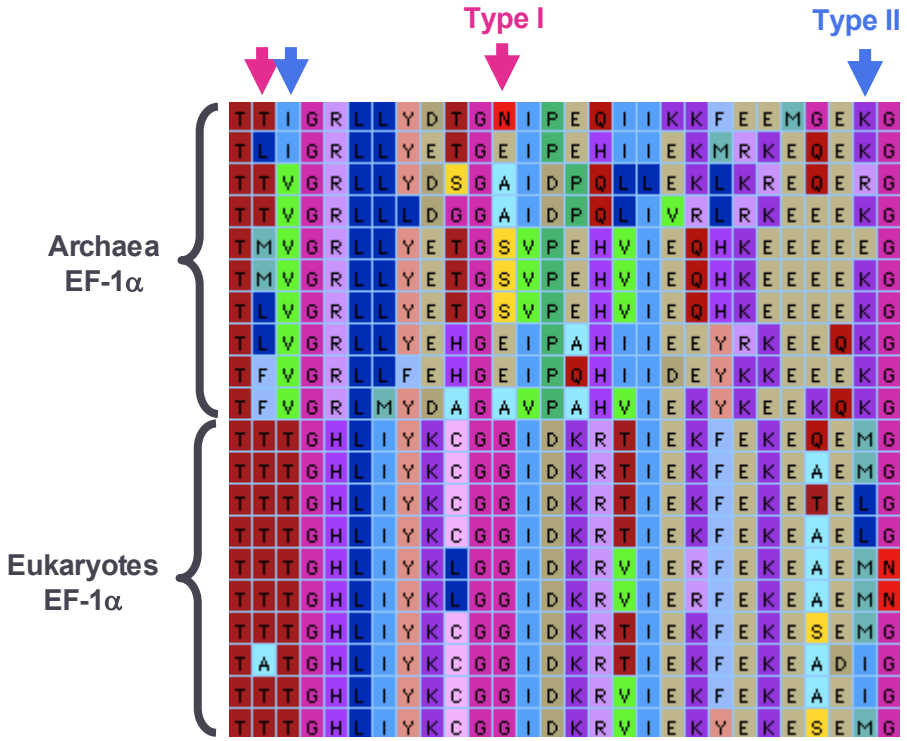
- Rate-shift models
  - Allows rates at many different sites to change abruptly on one branch

- 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).



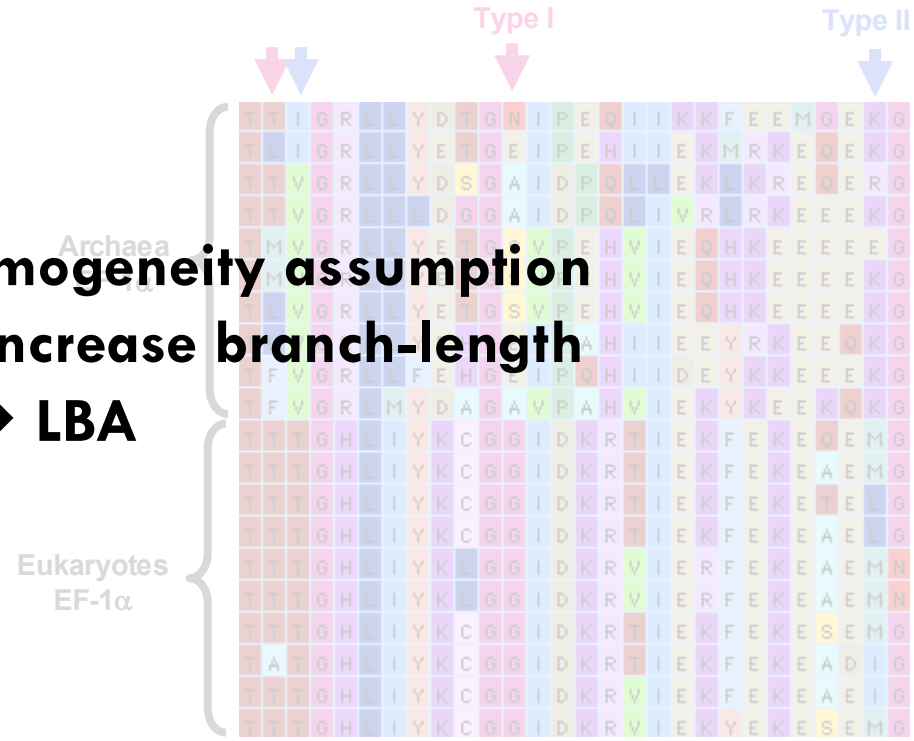
Functionally divergent sites generate heterotachy

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

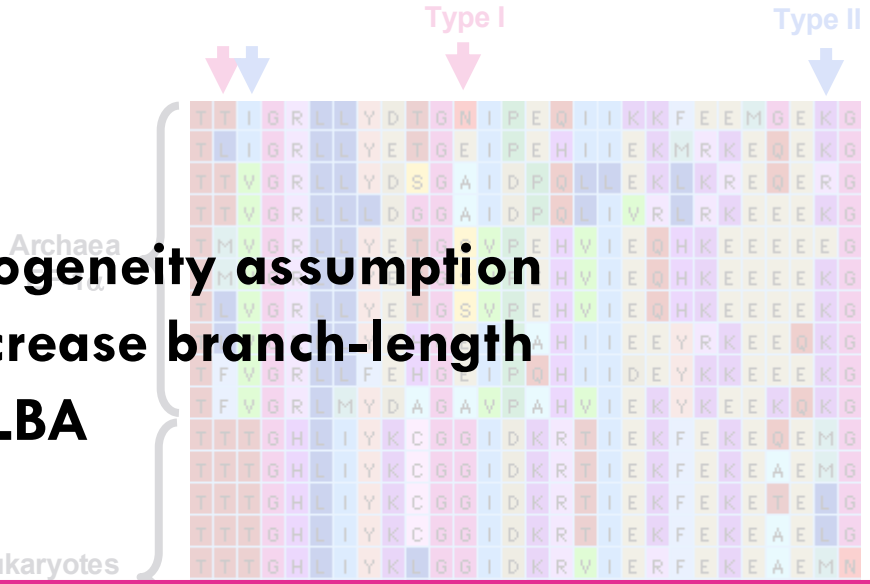
➔ **LBA**



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

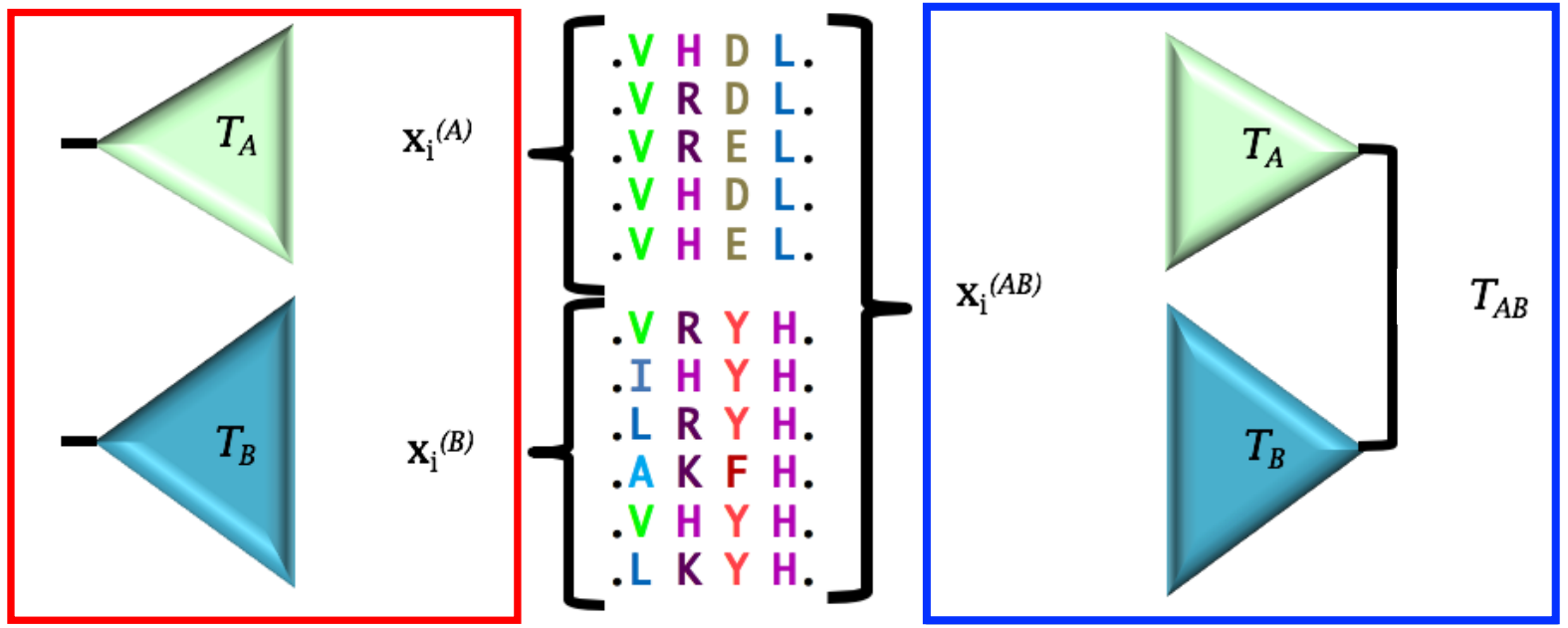
➔ **LBA**

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:

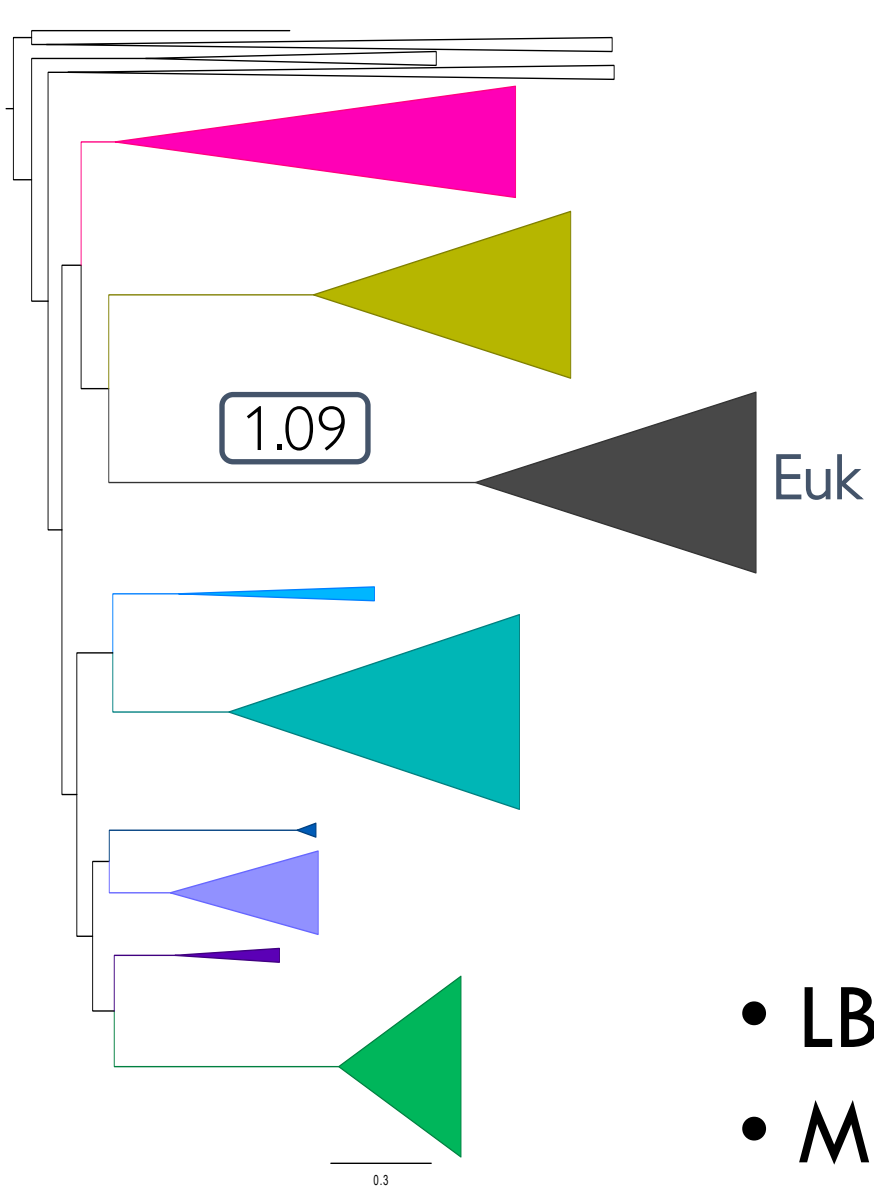
$$L_i(\theta, T_{AB}) = \rho \left(P(\mathbf{x_i}^{(A)}; \theta, T_A)\right) \left(P(\mathbf{x_i}^{(B)}; \theta, T_B)\right) + (1-\rho)\left(P(\mathbf{x_i}; \theta, T_{AB})\right)$$



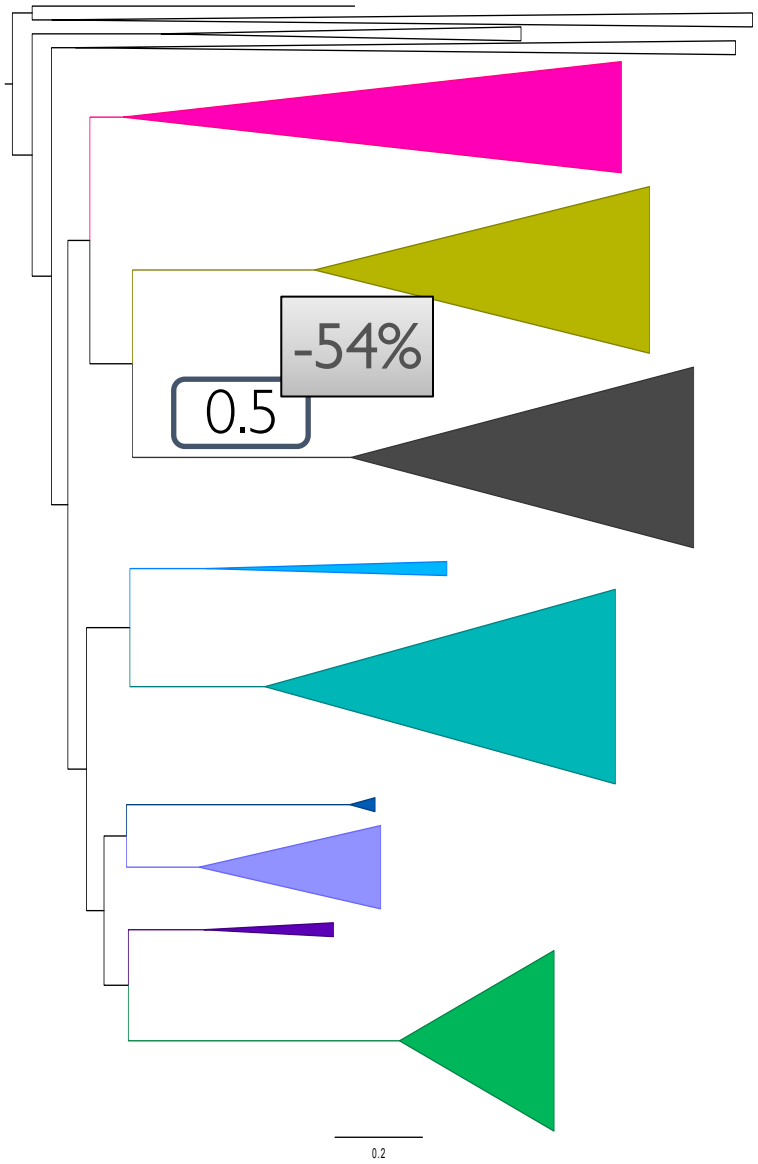
Independent across split (=FD)

Dependent across split (=non-FD)

Gaston, Susko and Roger (2011)



- ‘FD sites’

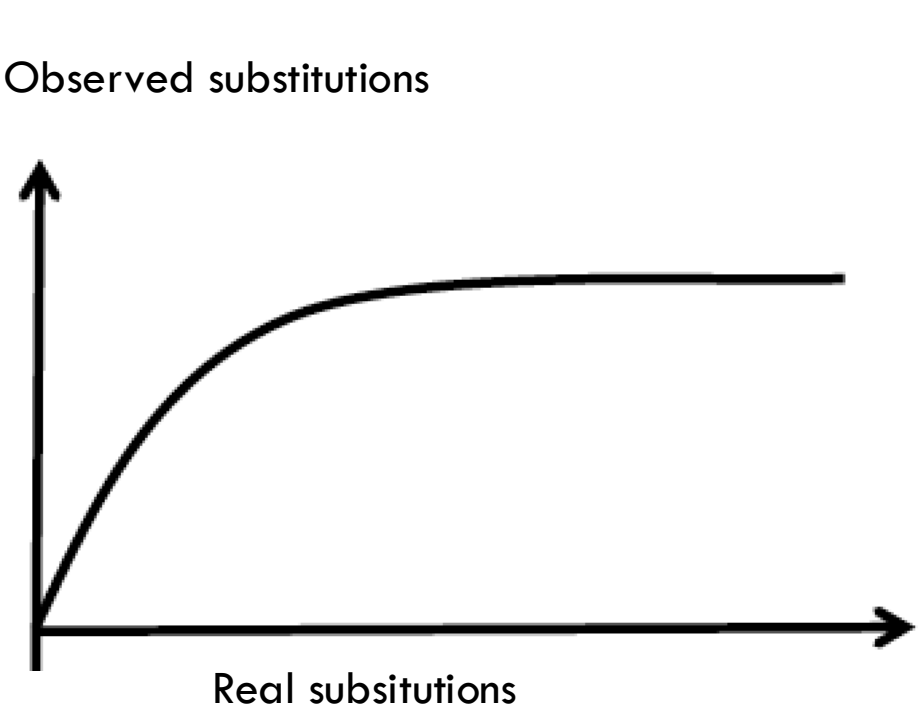


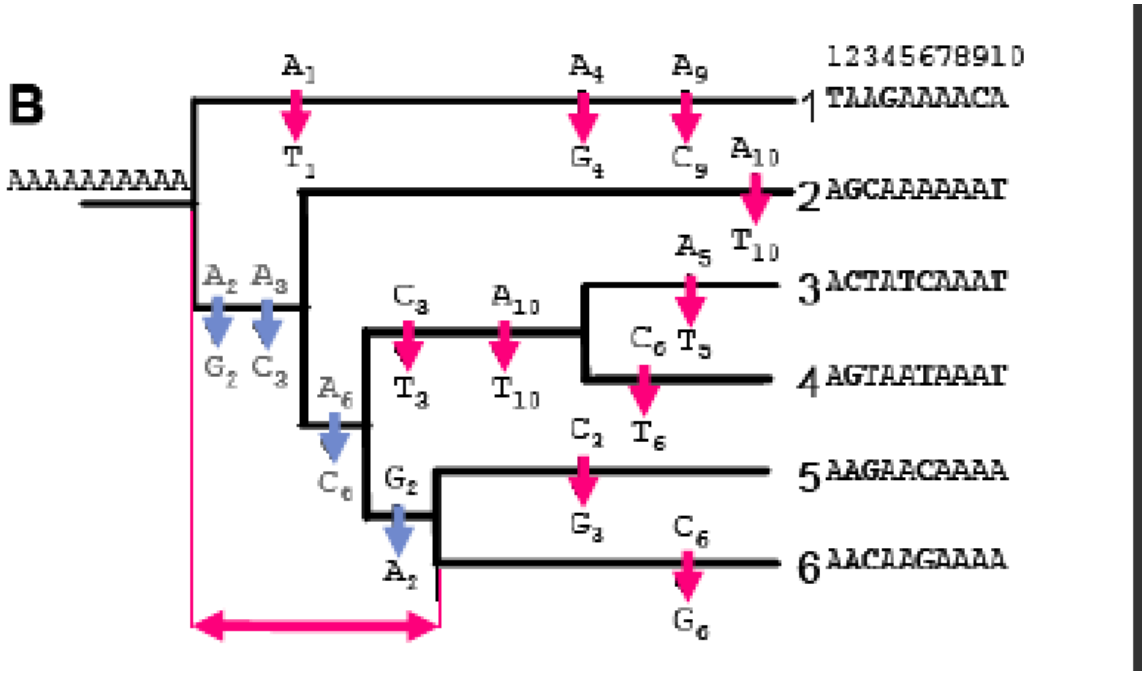
- LBA
- Molecular clock

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’

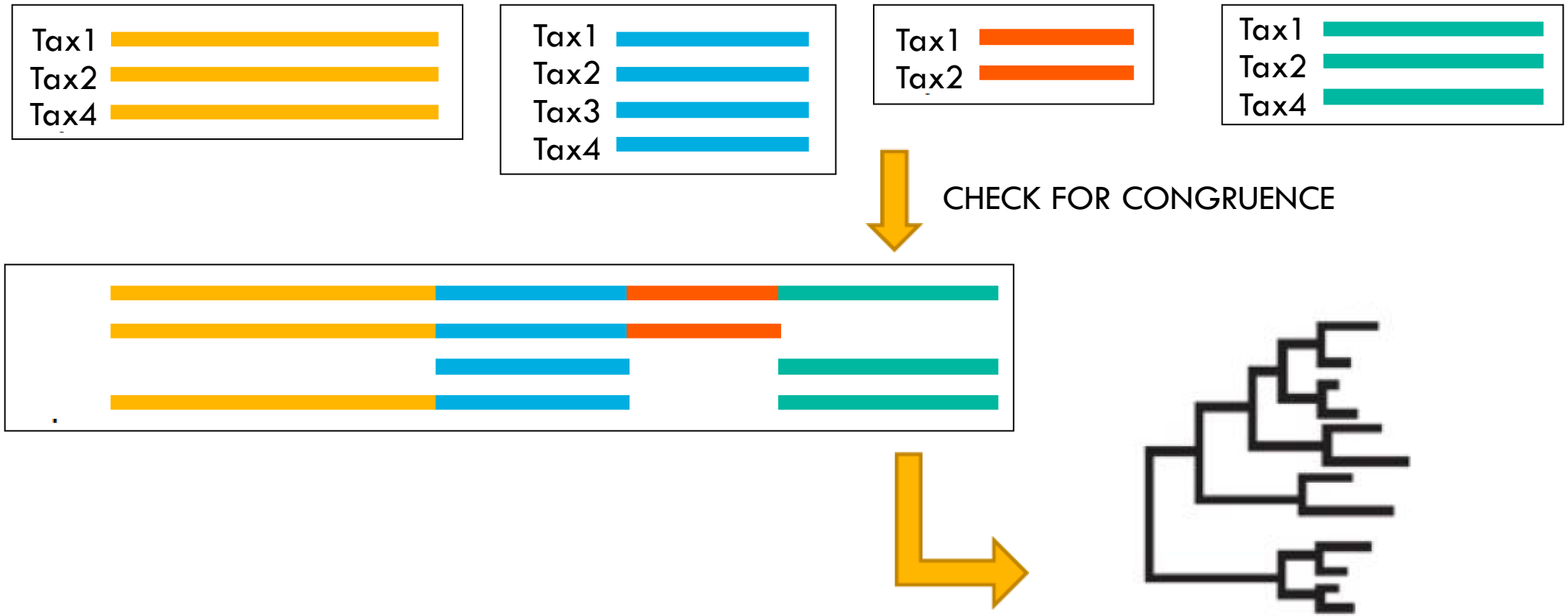
“Ancient” signal erased by more recent substitutions





Supermatrices

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



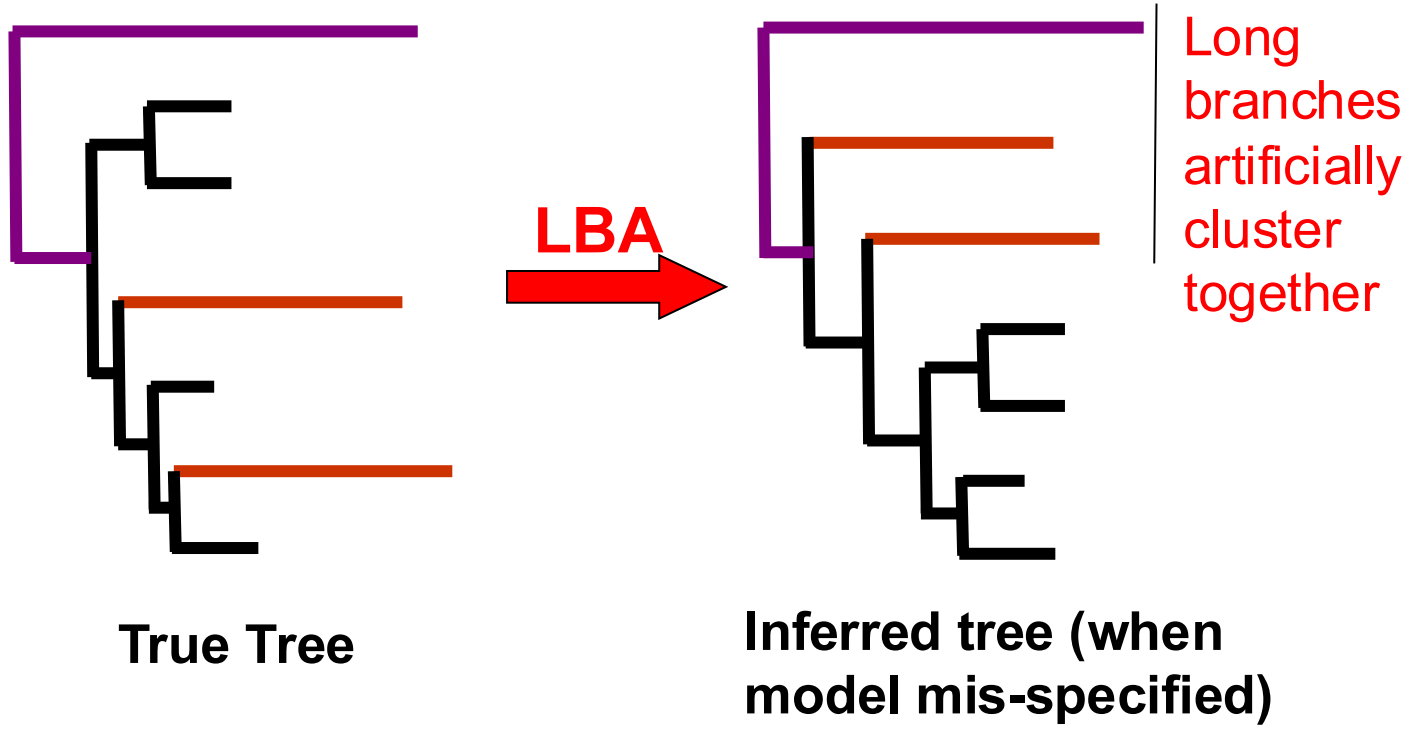
What can affect your topology

- 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

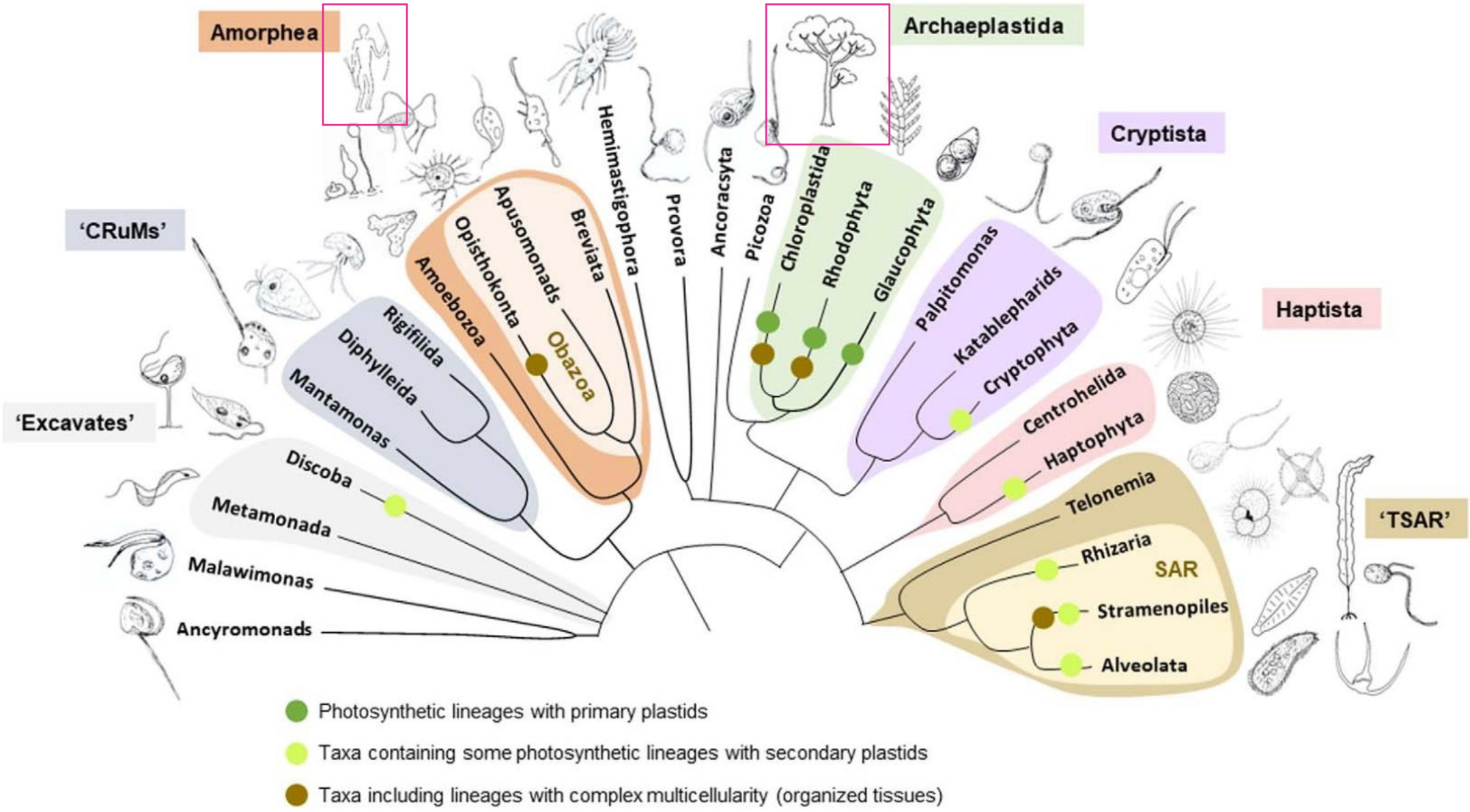
Model mispecification: statistical inconsistency

Long Branch Attraction (LBA) Artefact



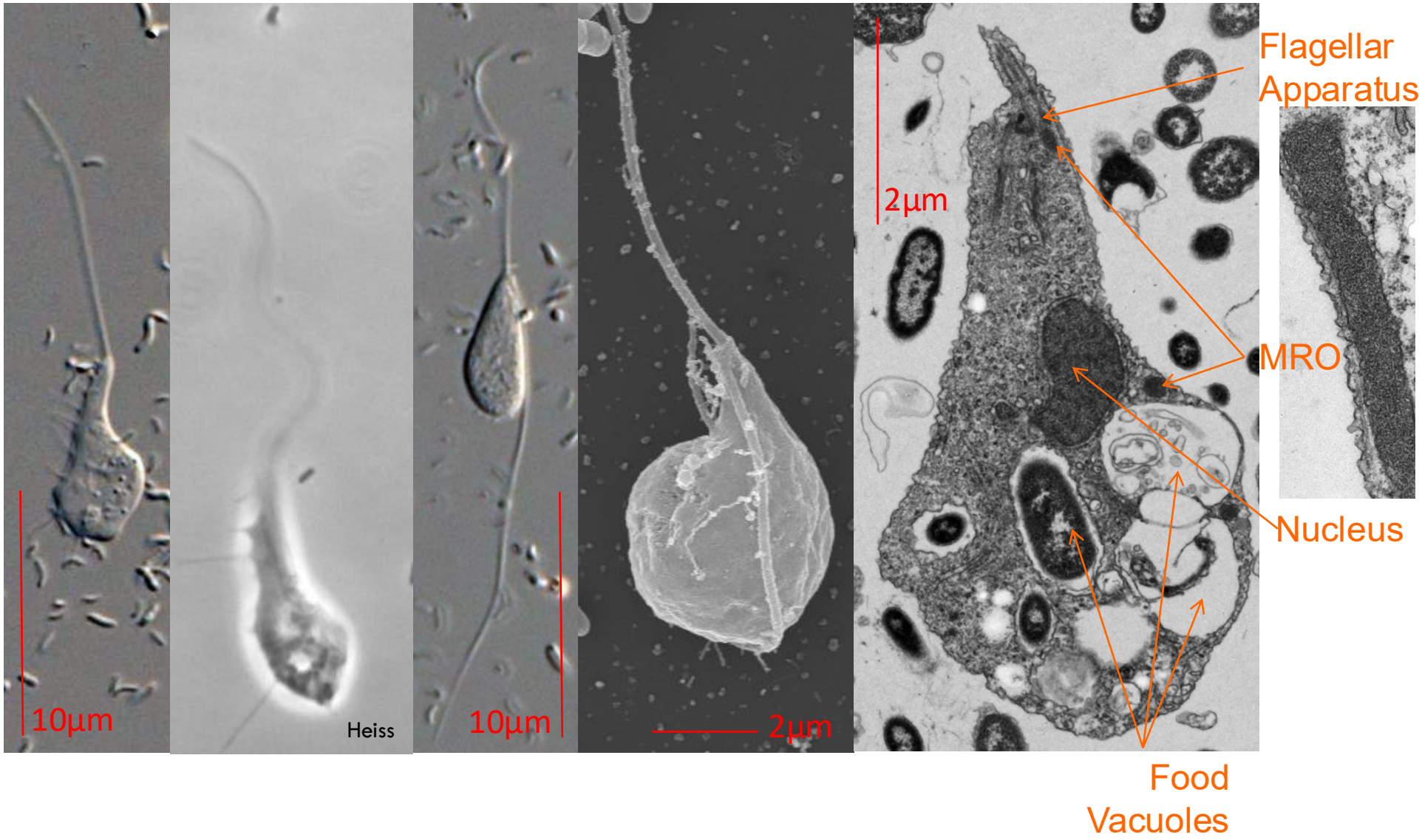
Adding more data *strengthens* artefact
→ statistical inconsistency

Tree of eukaryotes



Pygsuia biforma (Brown,…Roger 2013)

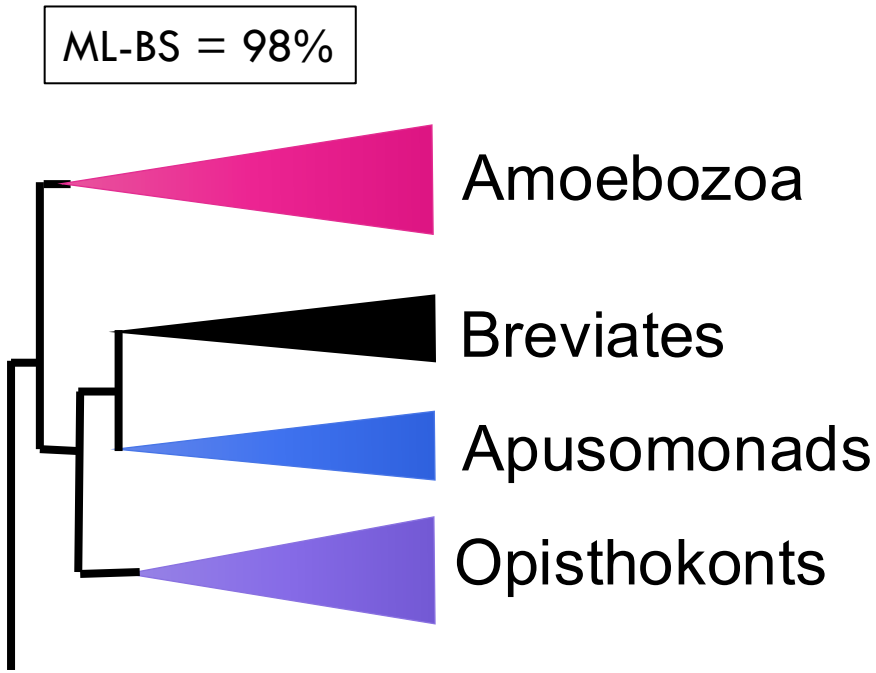
Isolated in Cape Cod!



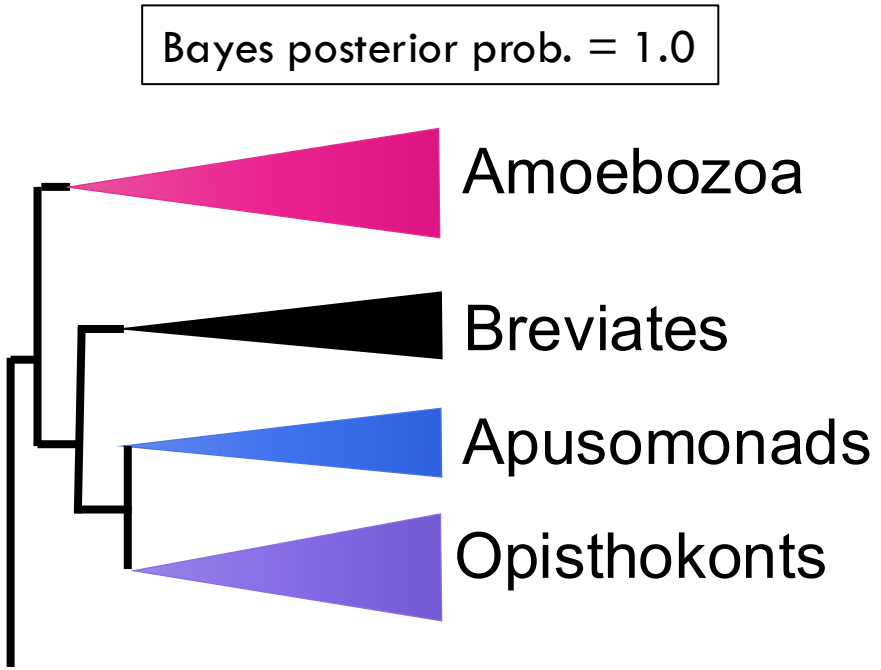
Two different topologies within Obazoa are supported by different phylogenetic models

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?

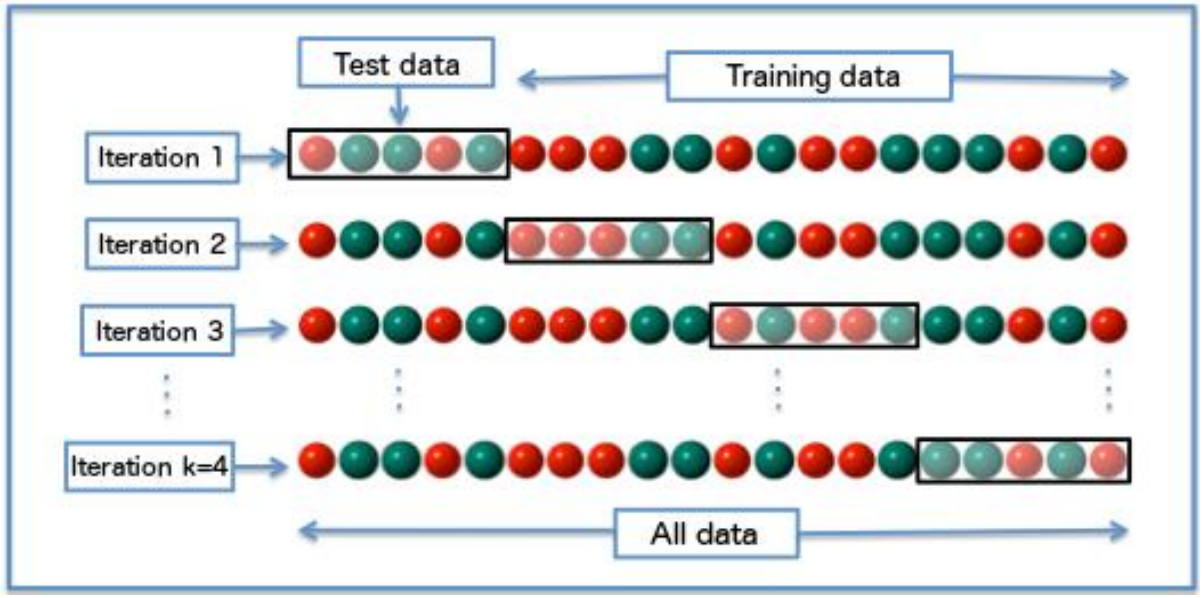
Cross-validation:

1) parameters of the model estimated on the learning set

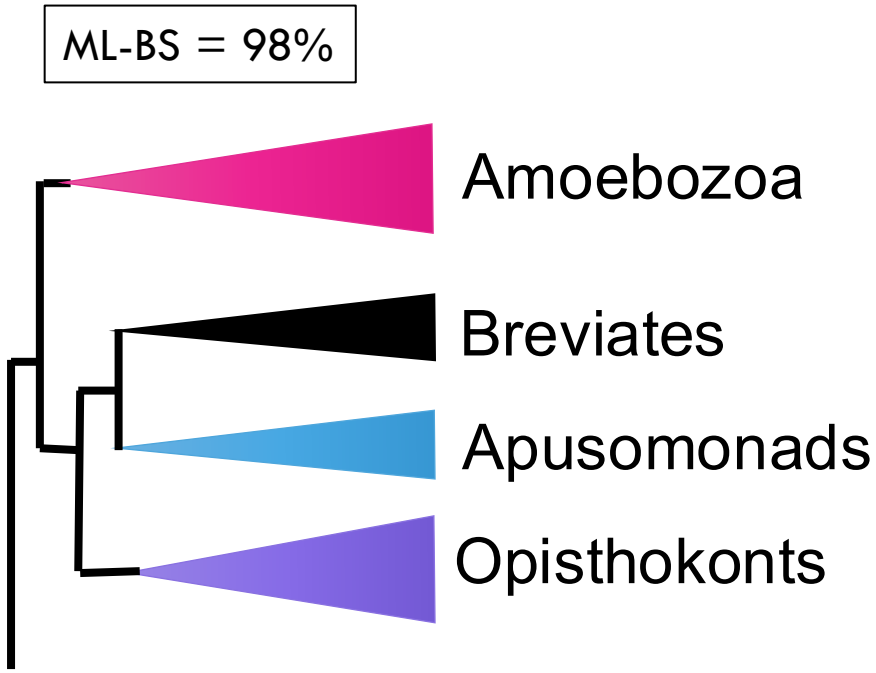
2) these parameter values are then used to compute the likelihood of the test set = **how well the test set is 'predicted' by the model?**

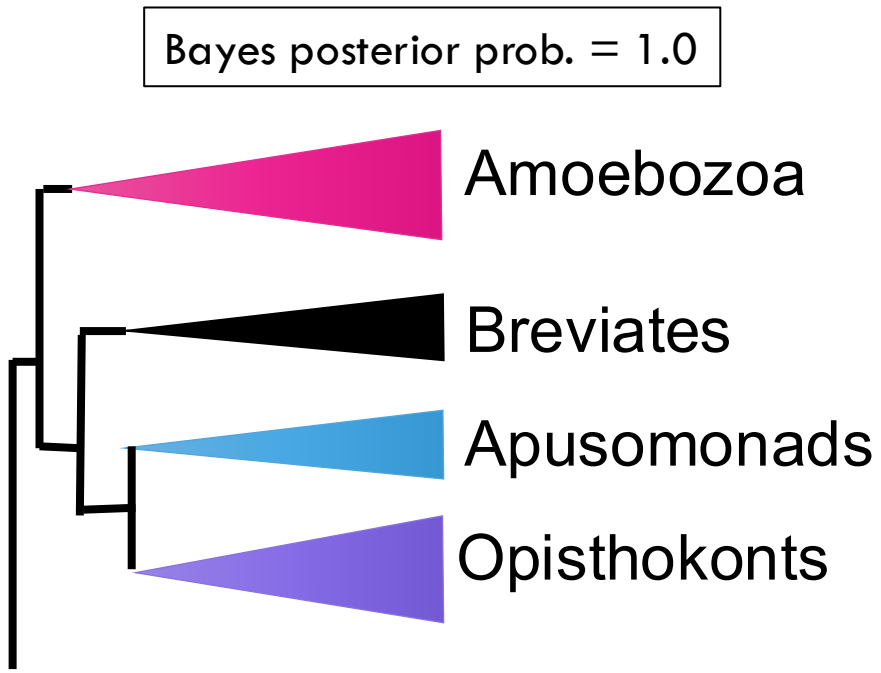
3) Repeat over all partitions and average the likelihood

4) Repeat for each model and compare



Cross-validation favors CAT-GTR over LG





ML – 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?

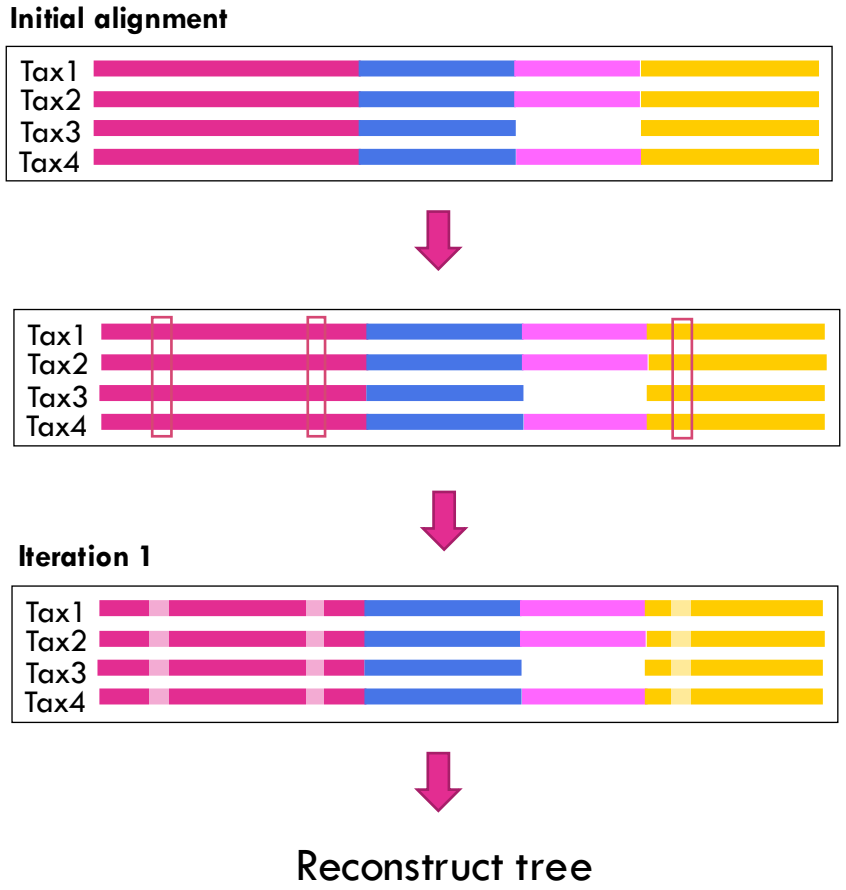
- One of two topologies is likely artefactual resulting from misspecified model
- Test which substitution model fits better
  - Cross-validation
- Try to eliminate ‘noisiest’ data
  - Fast-evolving site removal
  - Fast-evolving gene removal
  - Fast-evolving taxon removal
  - Recoding

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

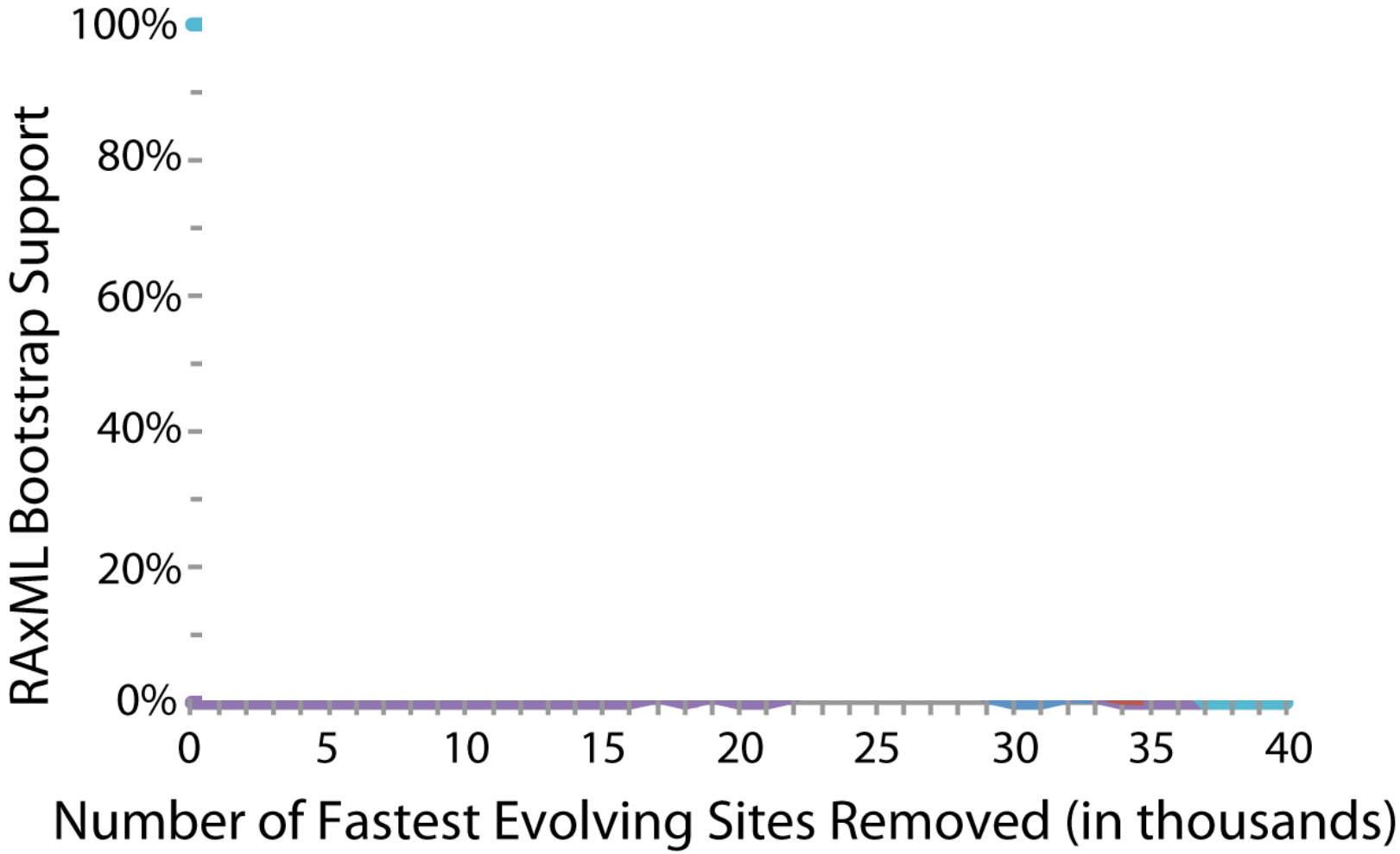
Fast-evolving sites : carry the 'noisiest' signal (most saturated sites)



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

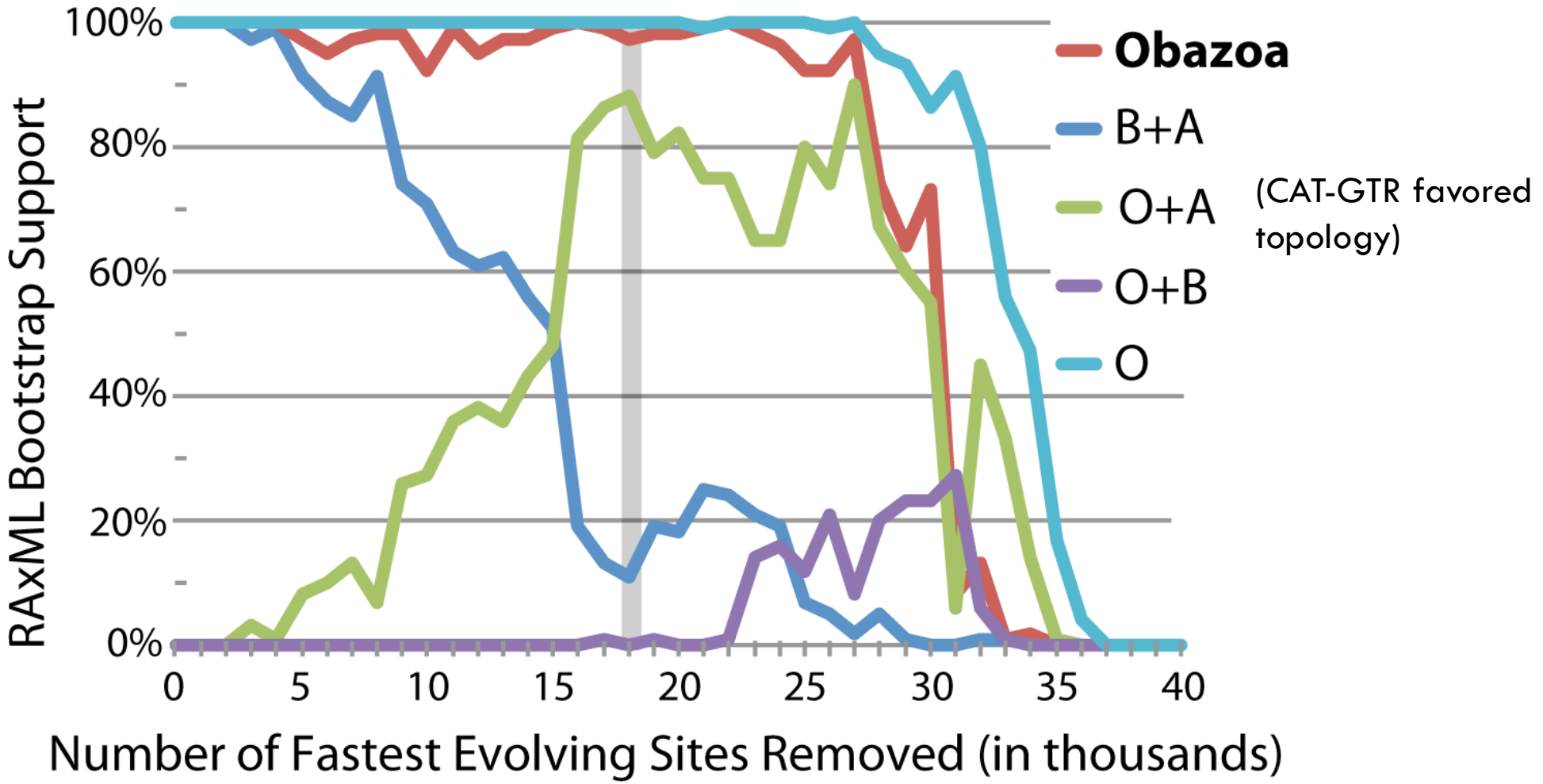
Step	# sites left	
1	43615	Tree1
2	42615	Tree 2
…		
4	40615	Tree 4
…		
40	3615	Tree 40

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



Brown et al. 2013 *PRSB*

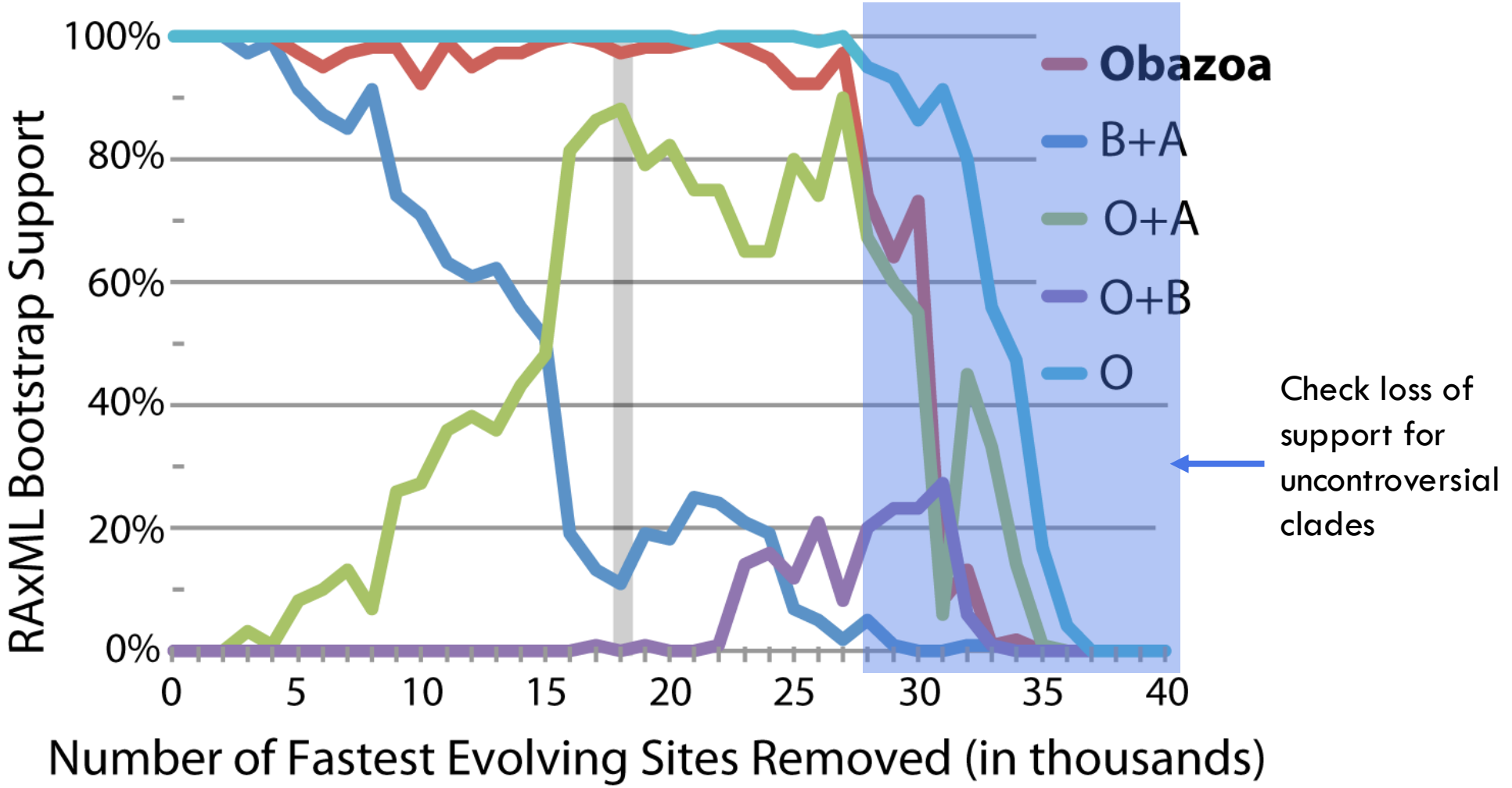
Estimate support for conflictual clades as we remove Fast-Evolving Sites (under LG+G)



(CAT-GTR favored topology)

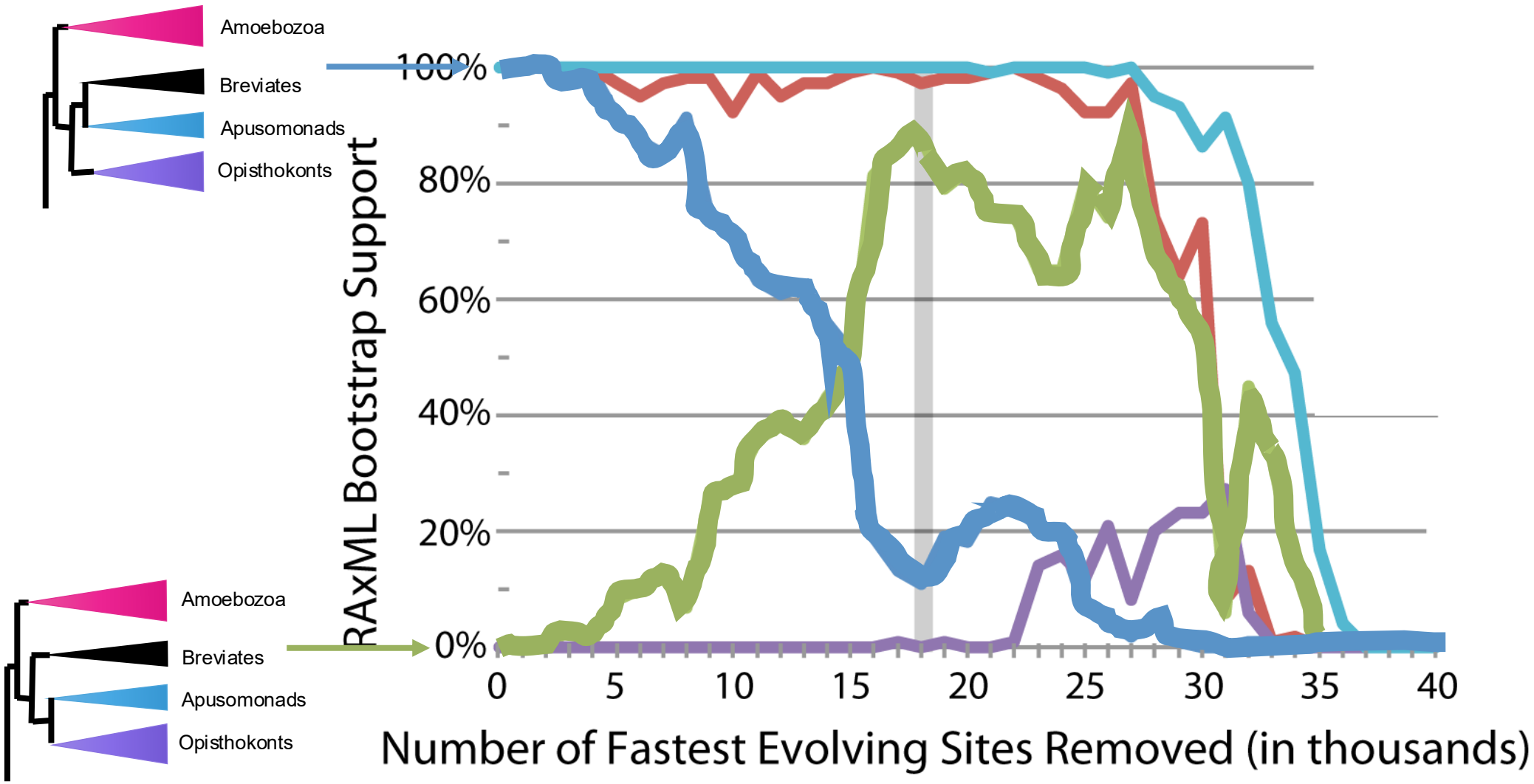
Brown et al. 2013 *PRSB*

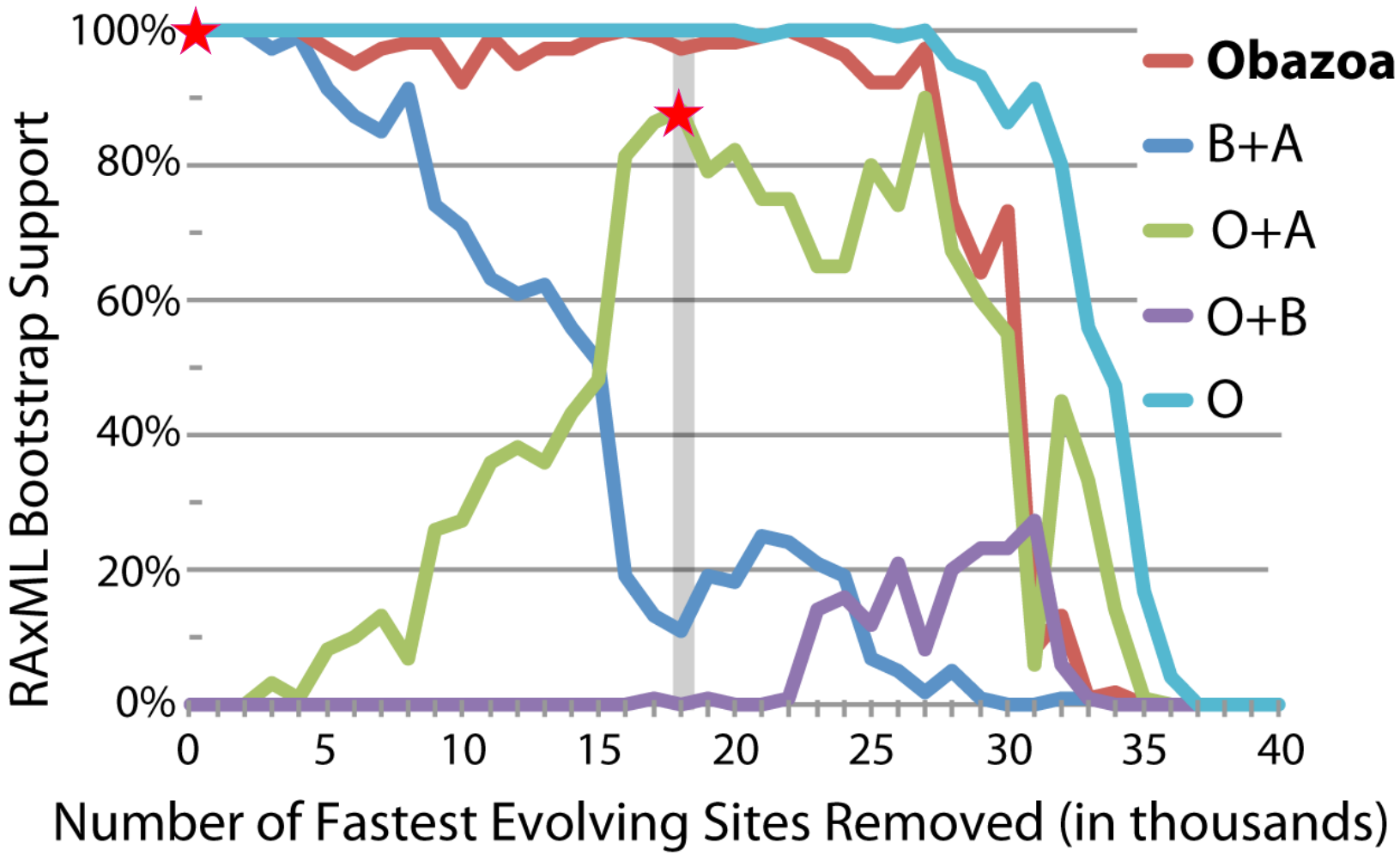
Estimate support for conflictual clades as we remove Fast-Evolving Sites (under LG+G)



Brown et al. 2013 *PRSB*

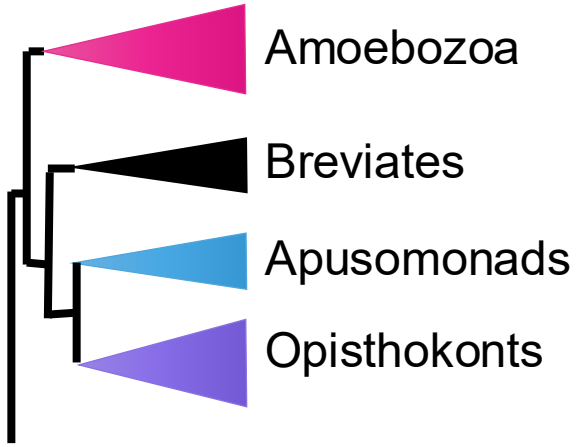
Breviates+Apusomonads (B+A) topology vs. Apusomonads+Opisthokonts (O+A)




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

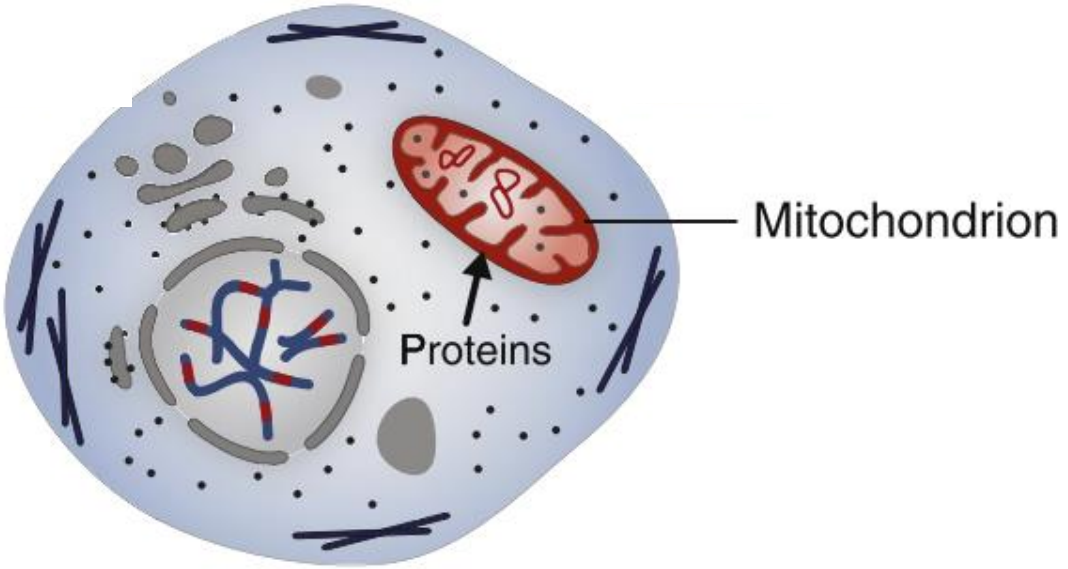
- In practice, fast sites seem to overlap to some extant with sites whose rate varies across the tree and are improperly modelled by most widely used models.

- You also remove the most saturated sites, which are usually poorly modelled

Break

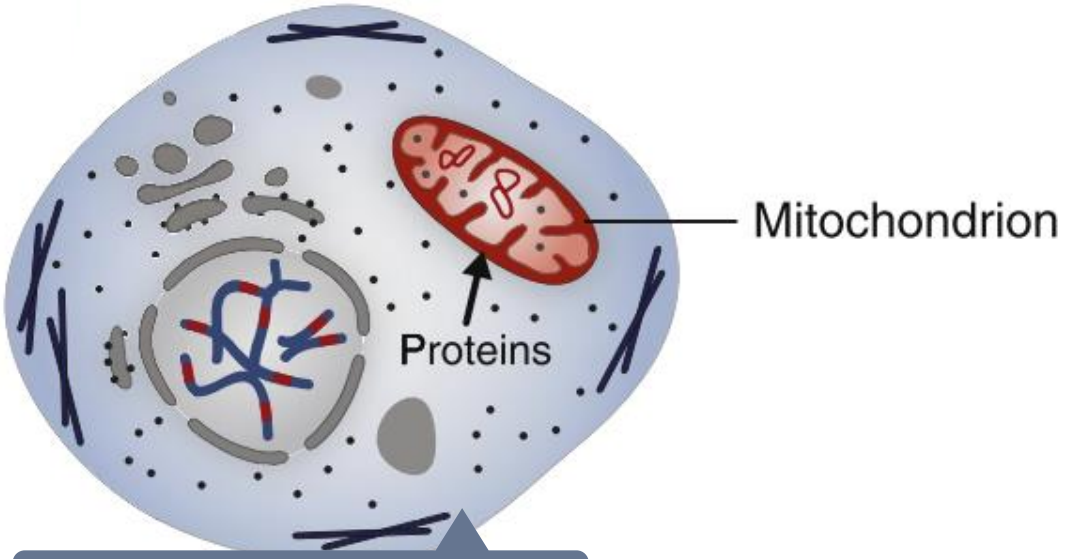
Example 2: Effect of amino-acid preference change over the tree

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



Complex eukaryotic cell

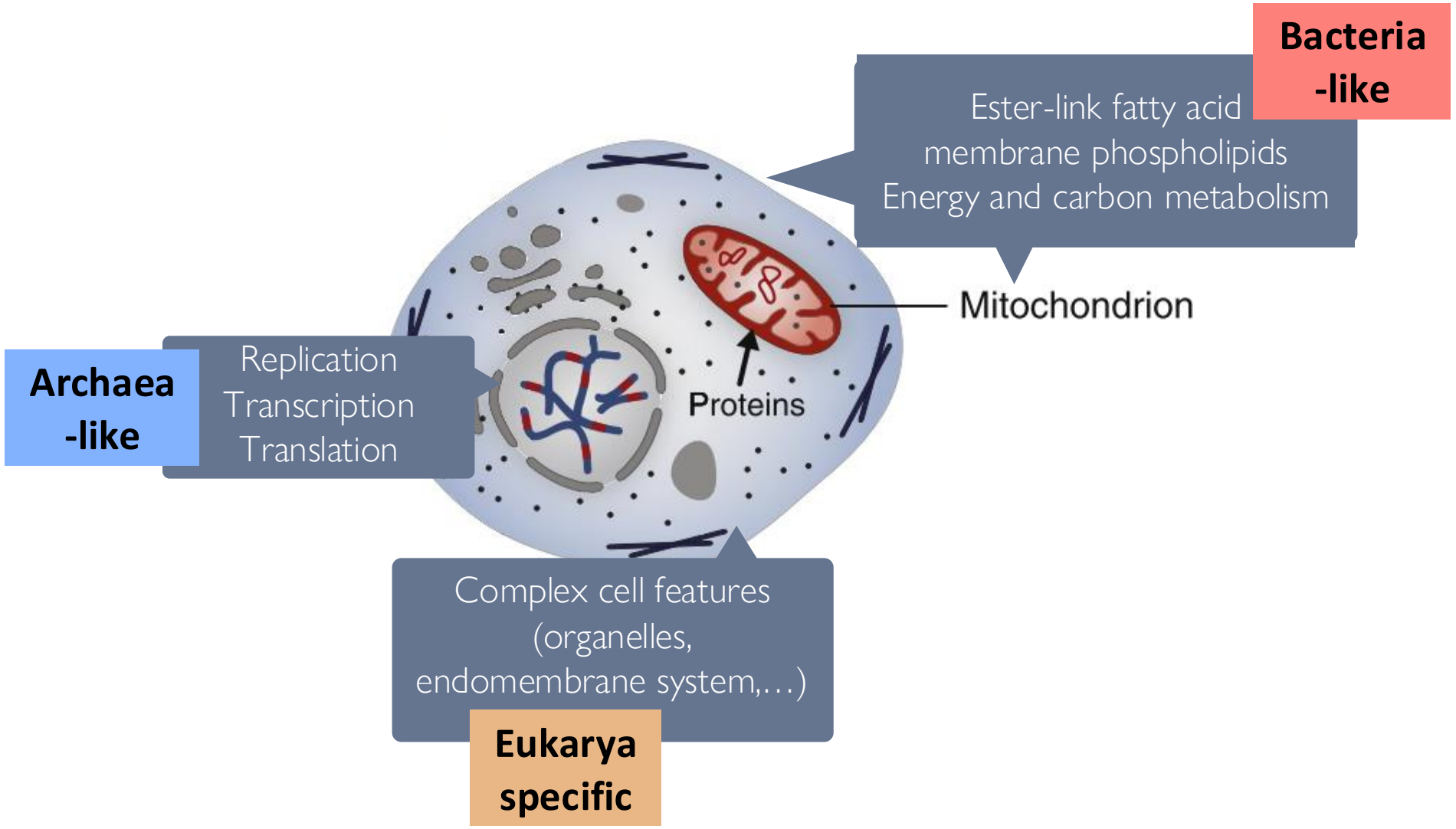
The chimeric nature of eukaryotes



Complex cell features (organelles, endomembrane system,…)

Eukarya specific

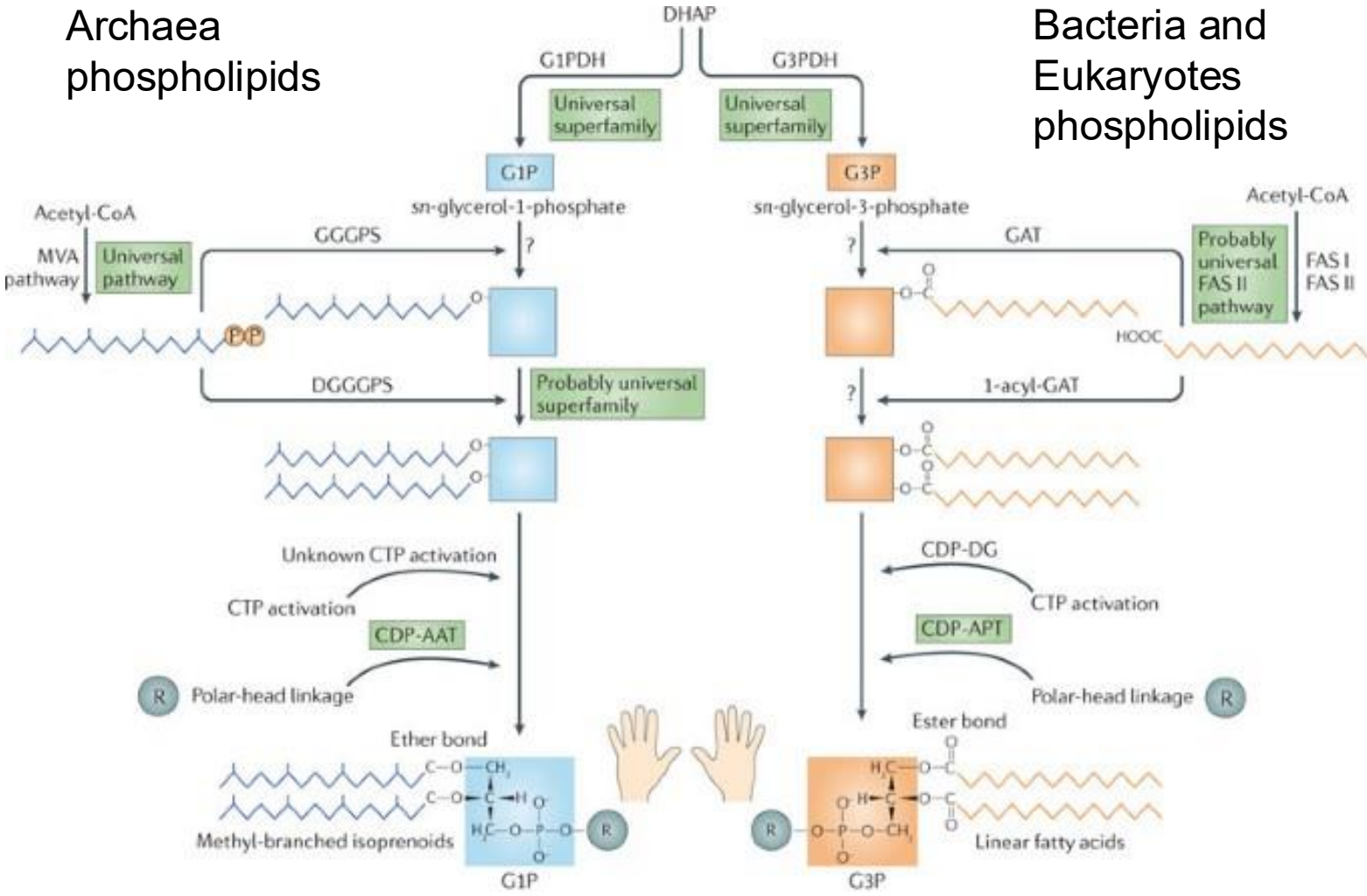
The chimeric nature of eukaryotes



Eukaryotic lipids resemble bacterial ones

Archaea phospholipids

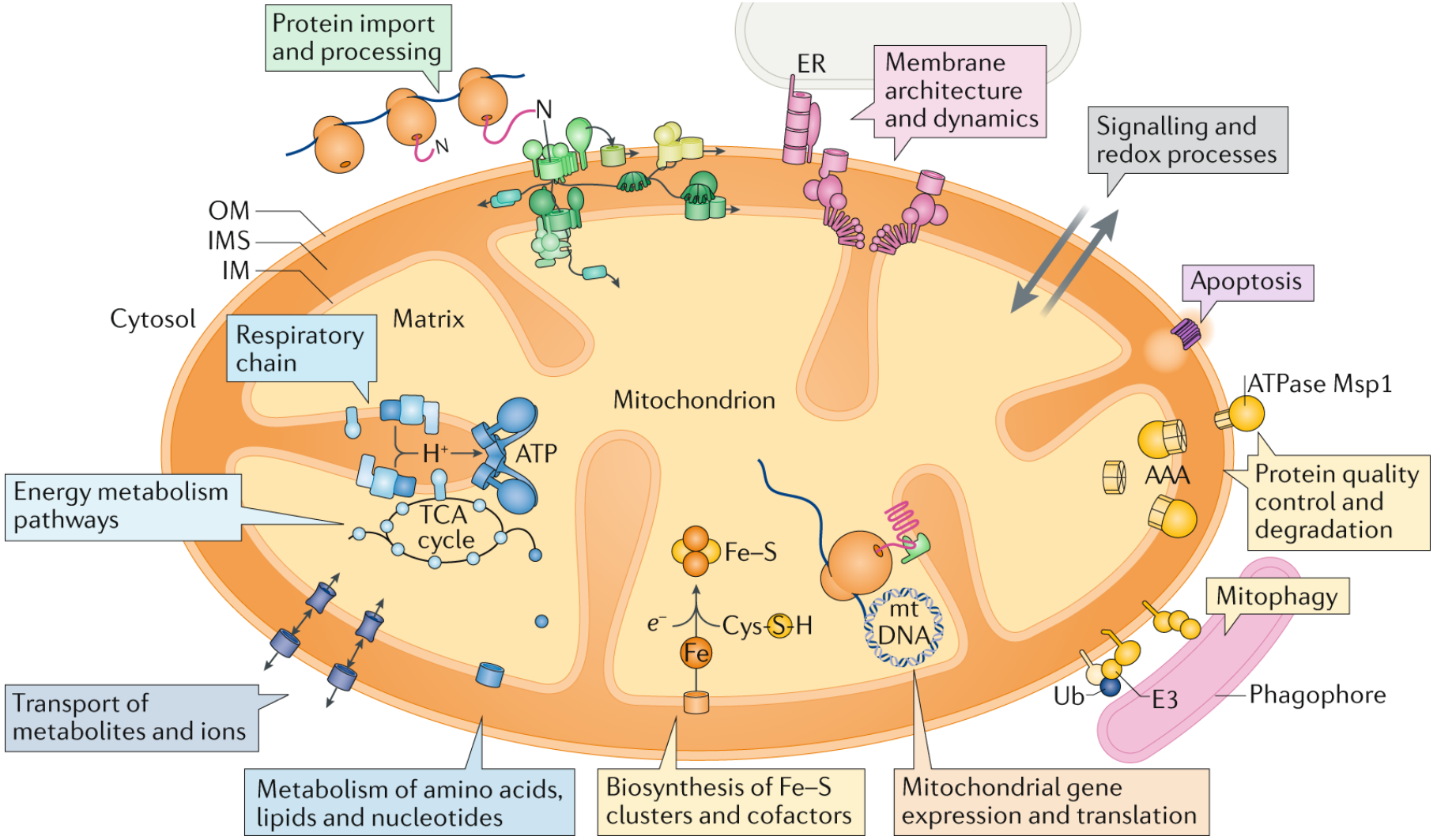
Bacteria and Eukaryotes phospholipids



Nature Reviews | Microbiology

Lombard et al, 2012

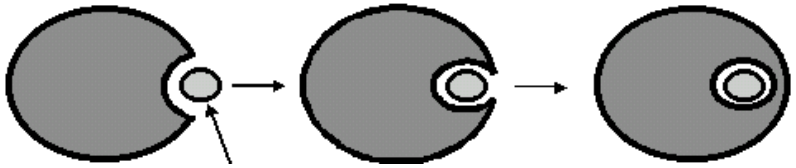
Mitochondria have diverse and crucial metabolic roles for the eukaryotic cell

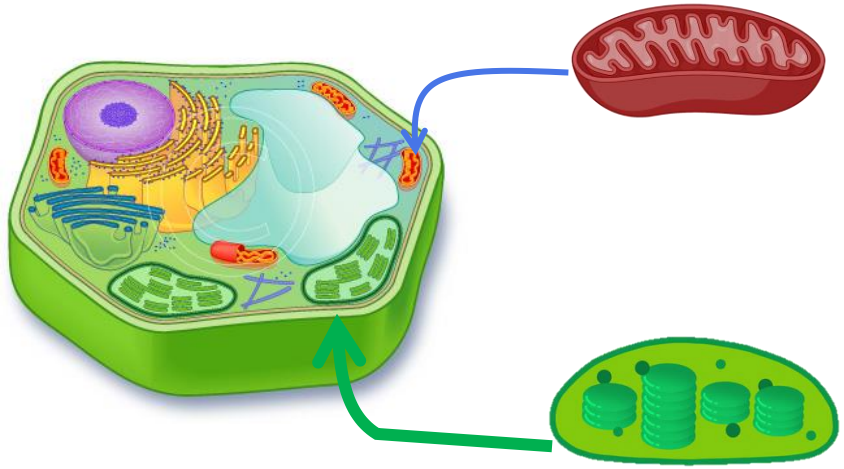


The notion that eukaryotes are chimeric in nature is not new

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

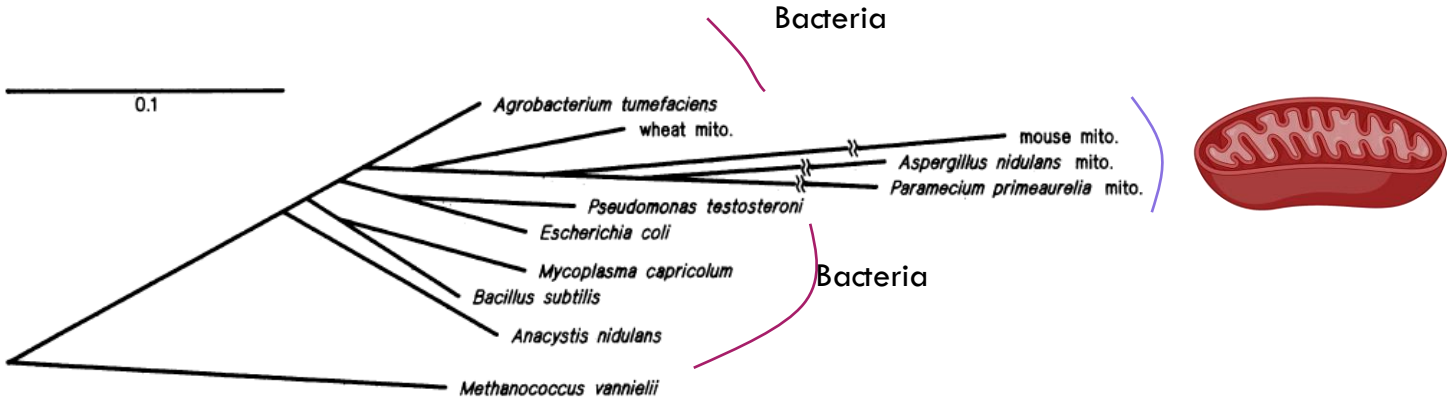




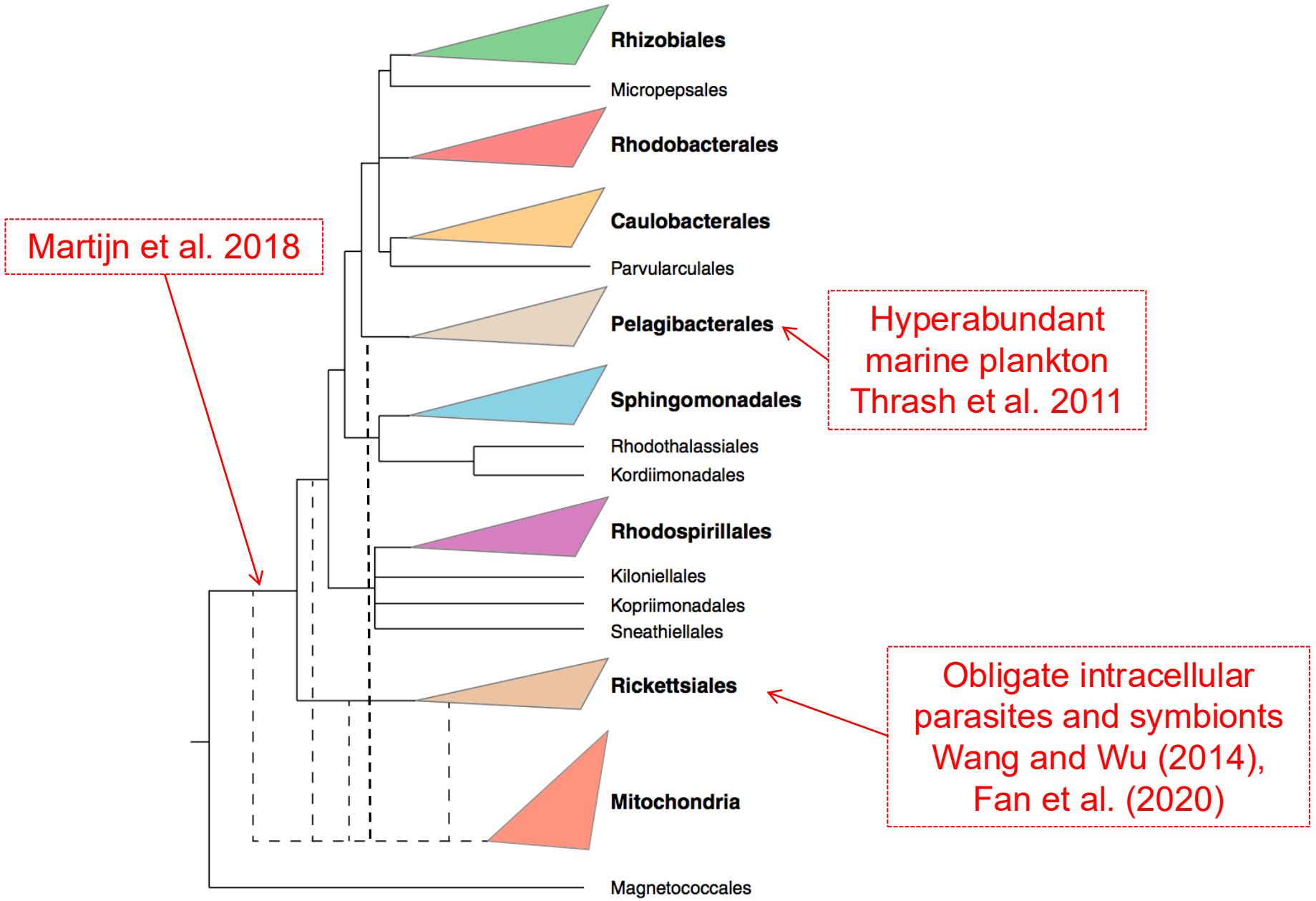
The notion that eukaryotes are chimeric in nature is not new

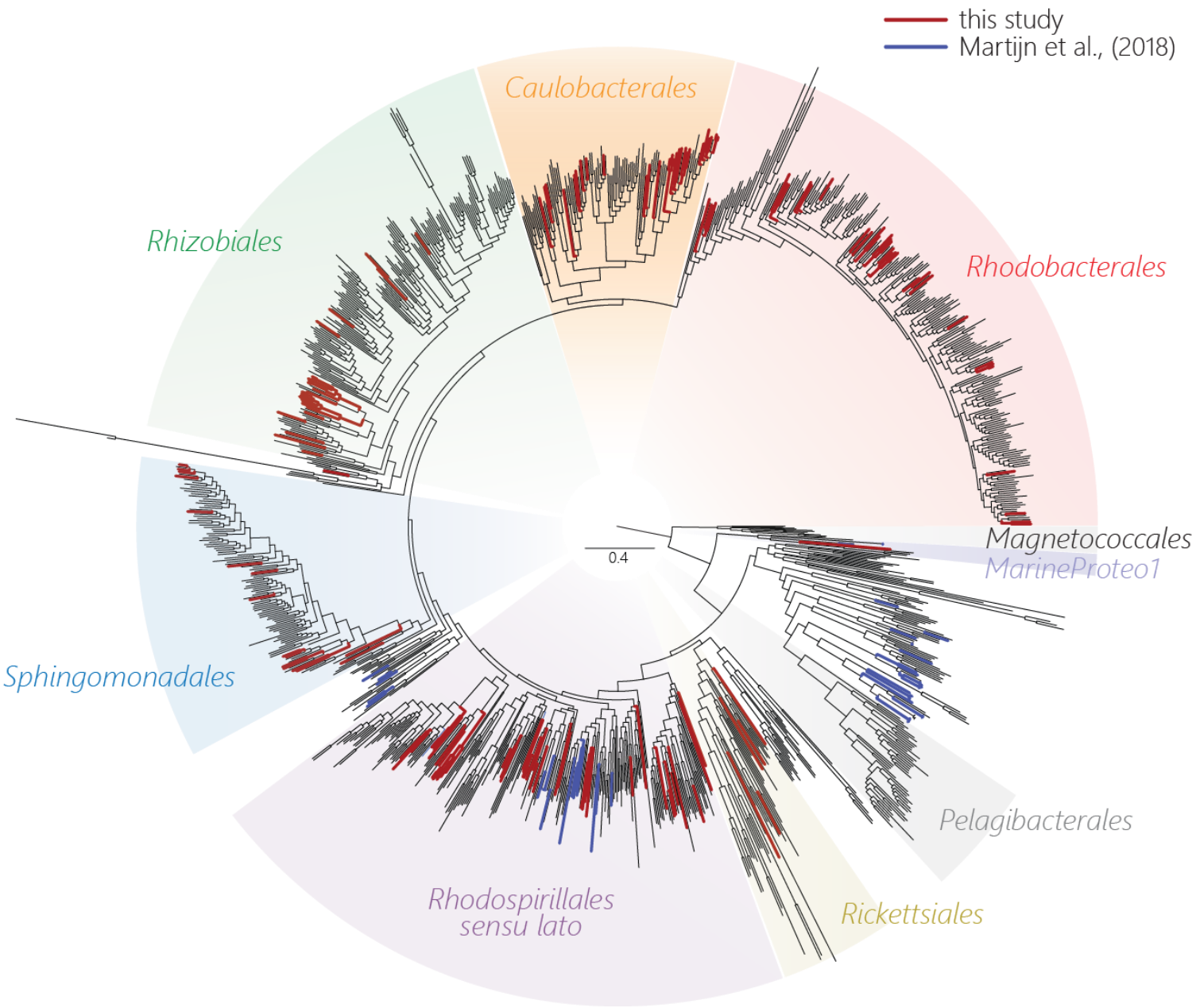
Lynn Margulis
(1938-2011)

- 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

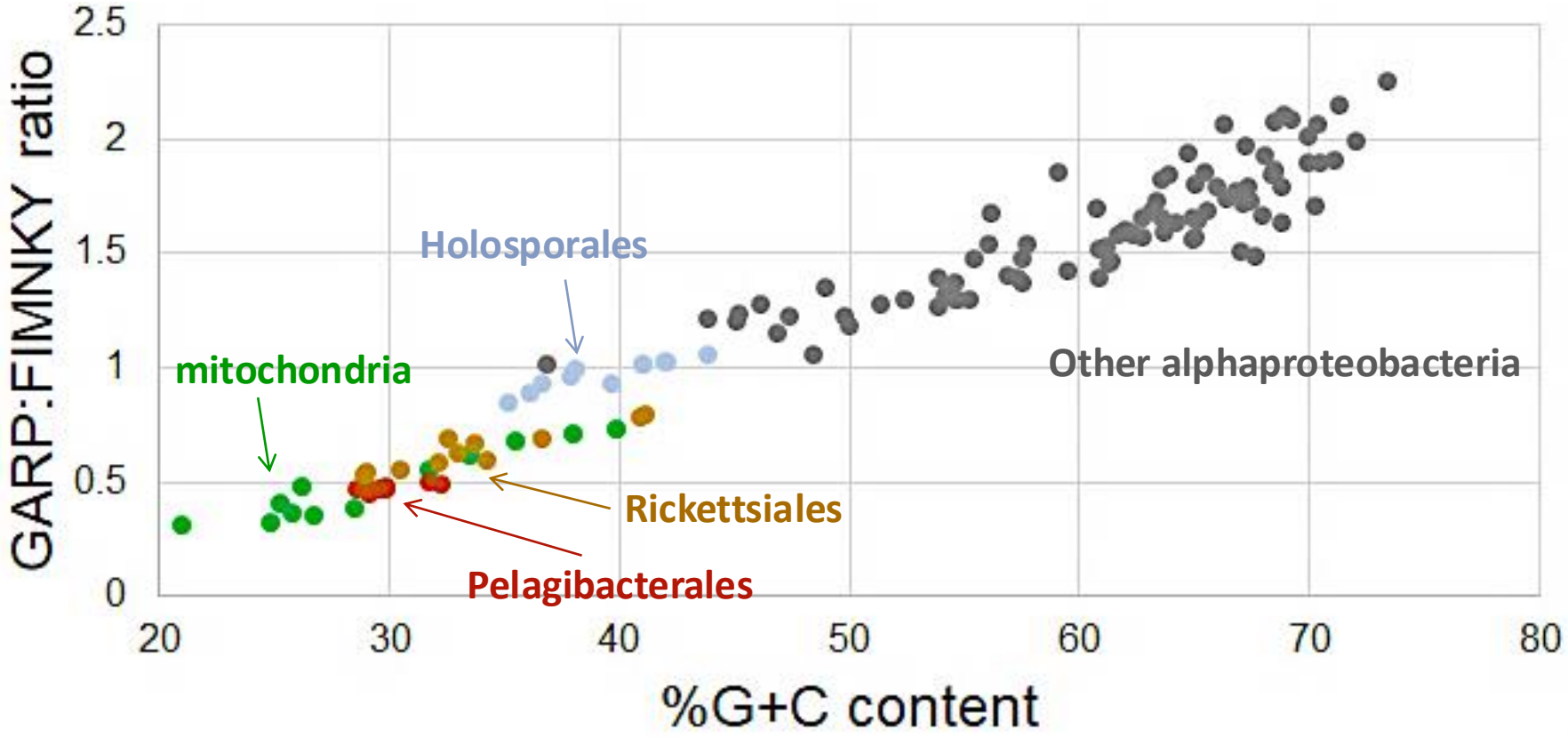
- More taxa:
150 non-marine alphaproteobacterial MAGs (microbial mats, microbialites and lake sediments)

- New model:
GFmix phylogenetic model

GFmix is now implemented in IQ-TREE 3 (Wong et al. 2025)!

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

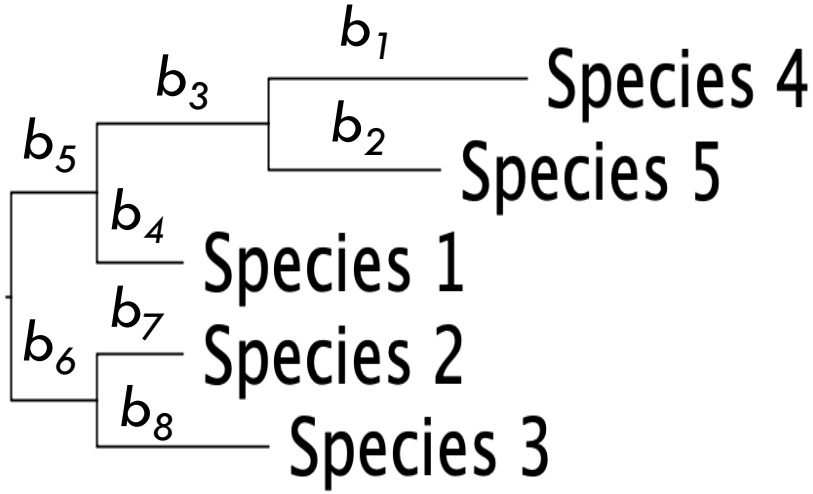
Edward Susko

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:

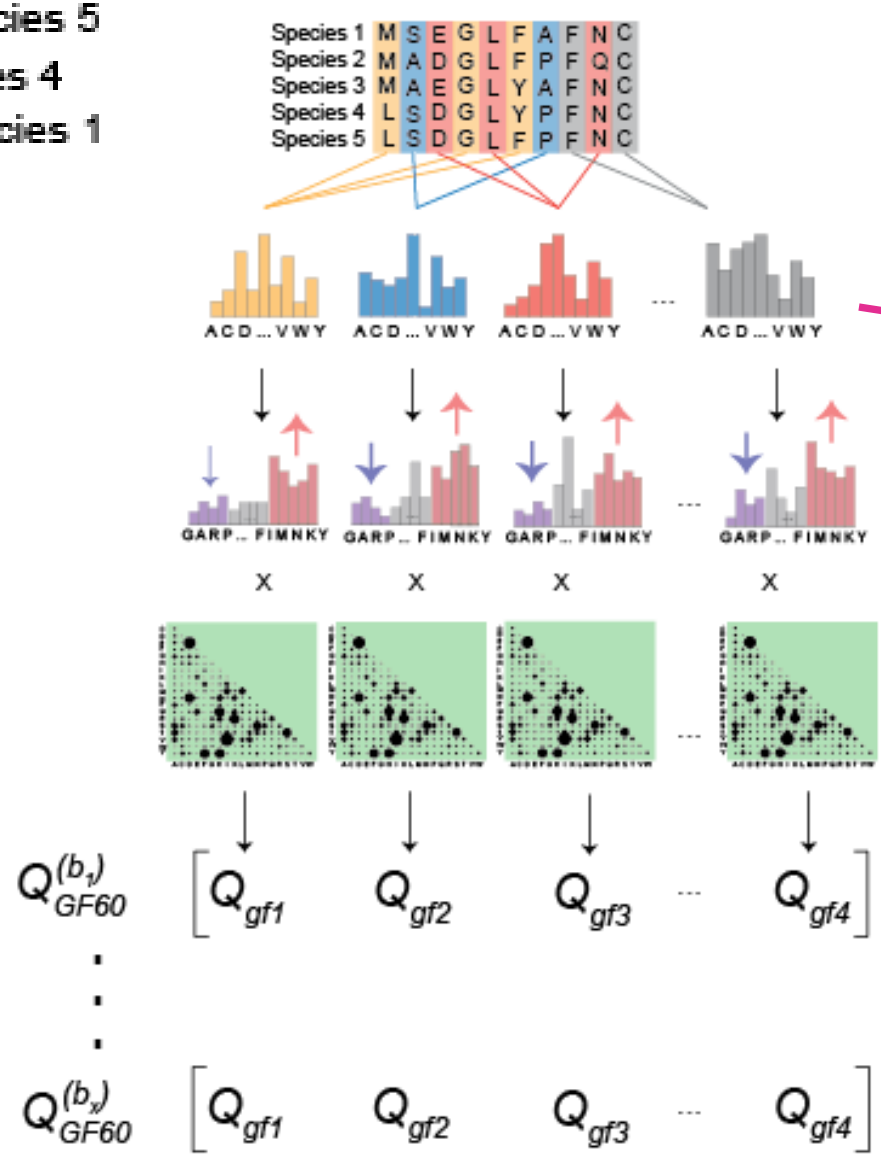


GFmix modifies the site profile mixture classes for each branch based on the *b* parameter

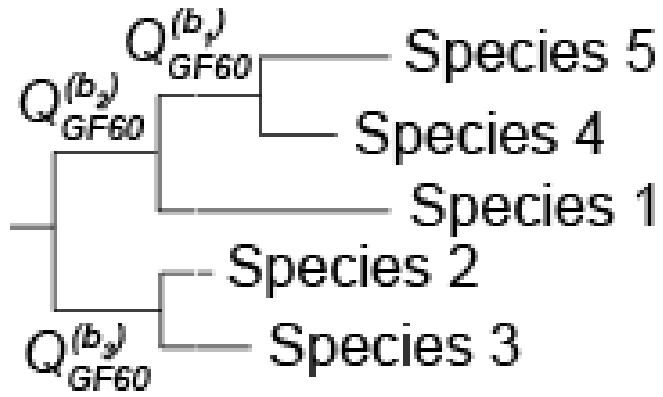
Muñoz-Gómez et al. (2022) *Nat. Ecol. Evol.*

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

Edward Susko

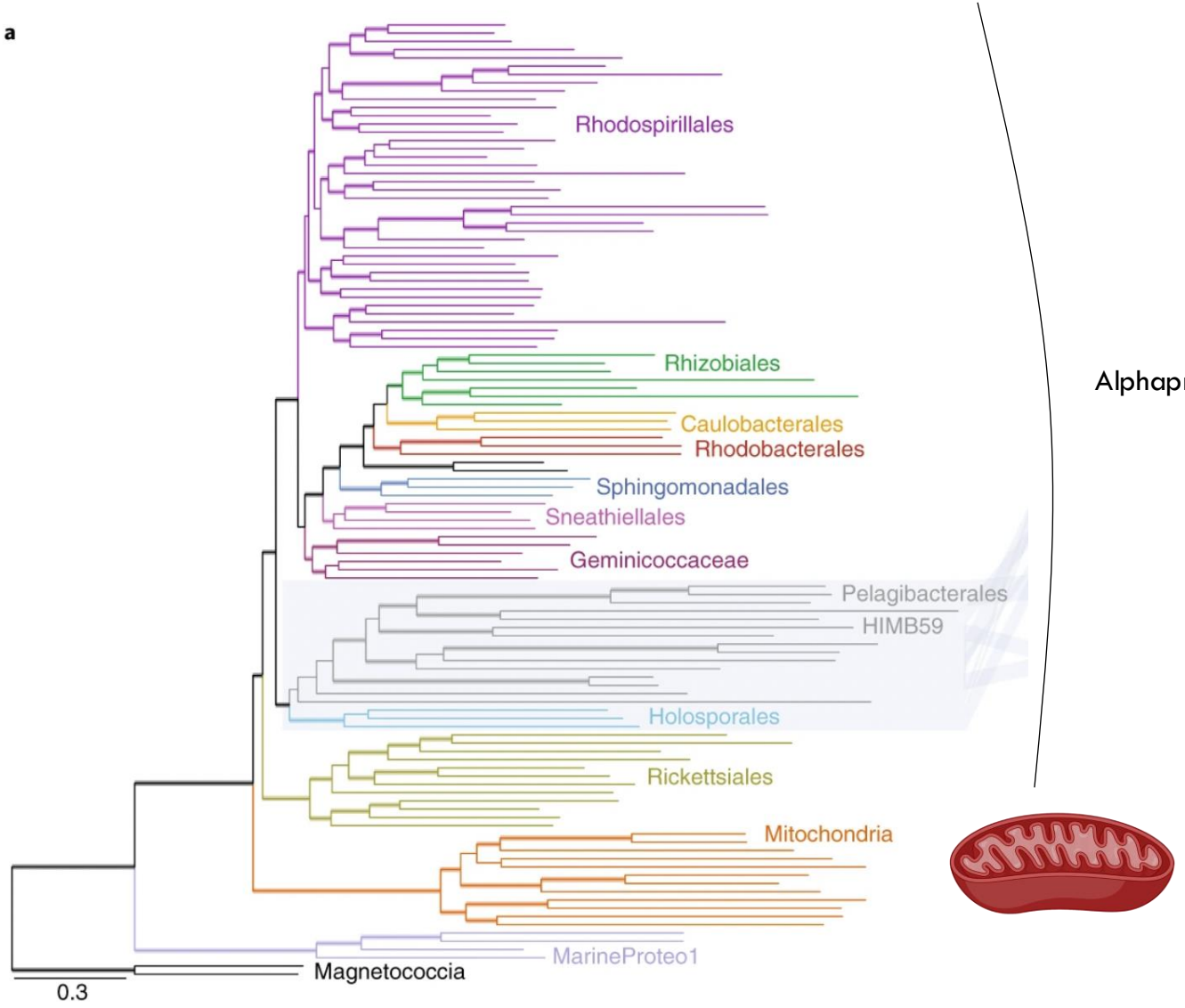


$$\frac{f_{G,A,R,P}}{f_{F,I,M,N,K,Y}} = b \ll 1$$



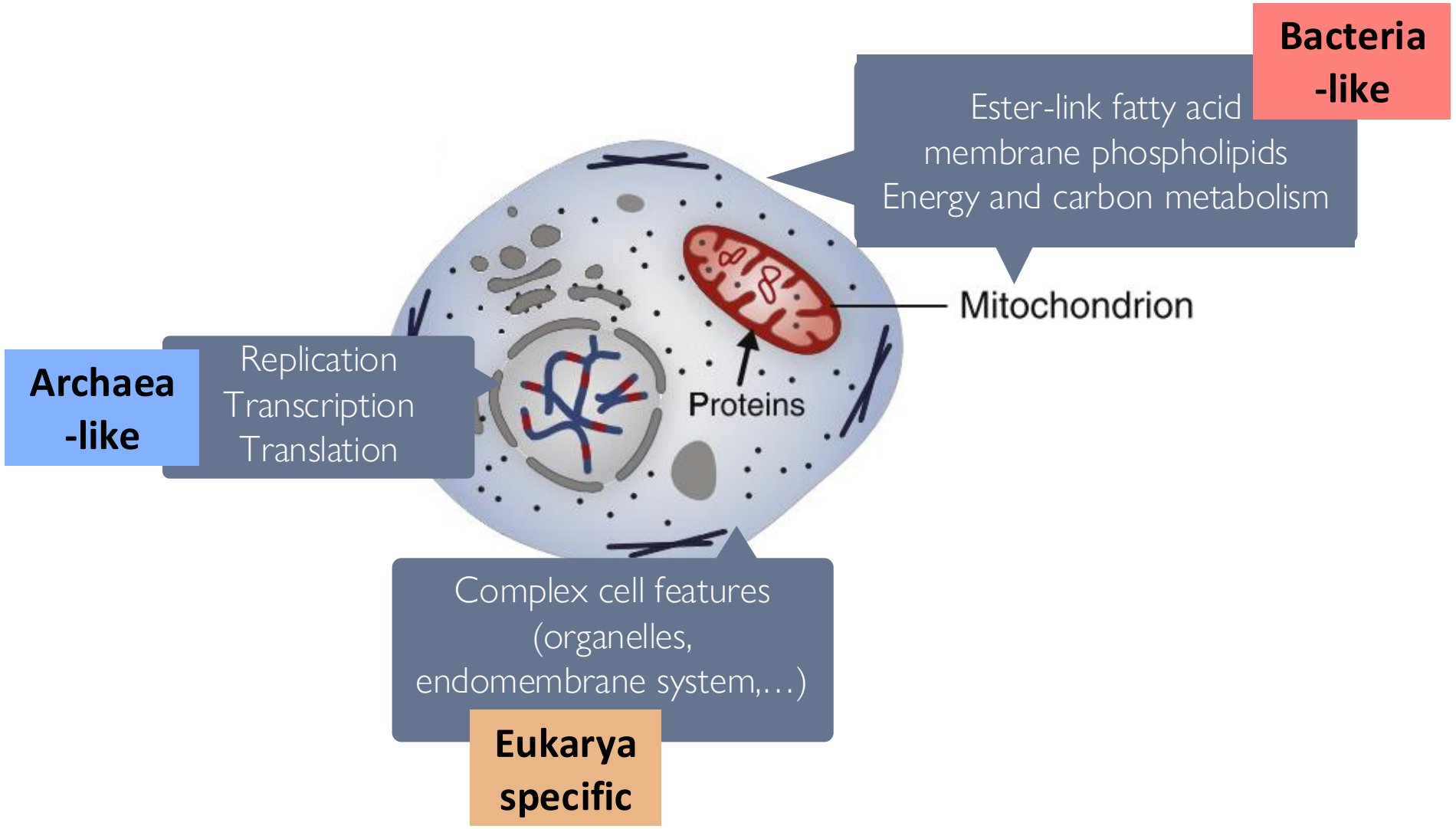
Muñoz-Gómez et al. (2022) *Nat. Ecol. Evol.*

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



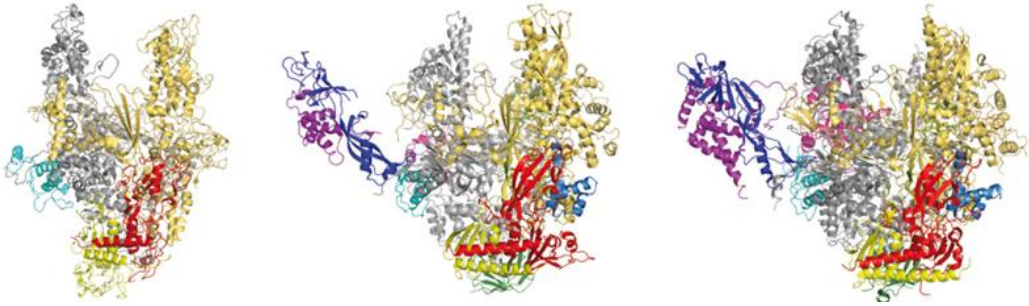
Munoz-Gomez et al., Nature Ecol Evol 2022

The chimeric nature of eukaryotes

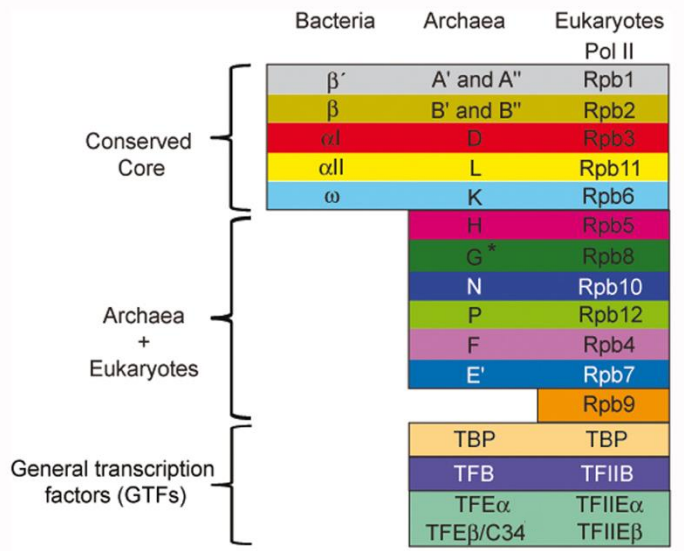


Informational systems suggest an archaeal connection

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

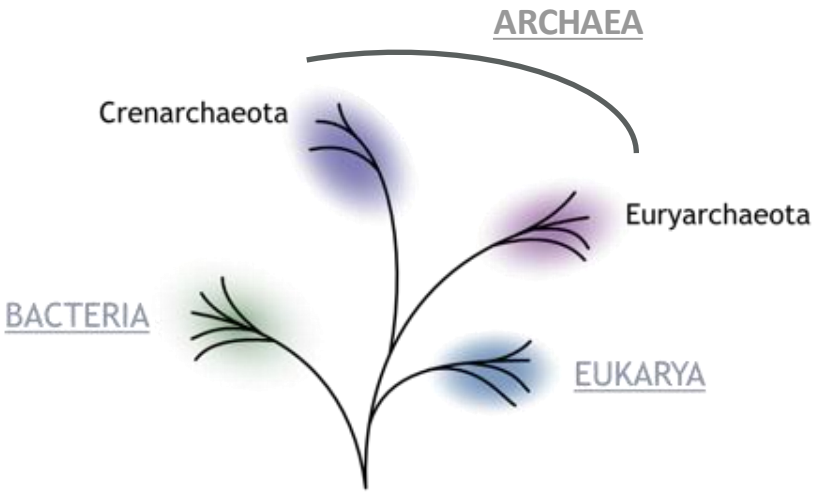


Overall architecture of RNA polymerases (RNAPs)

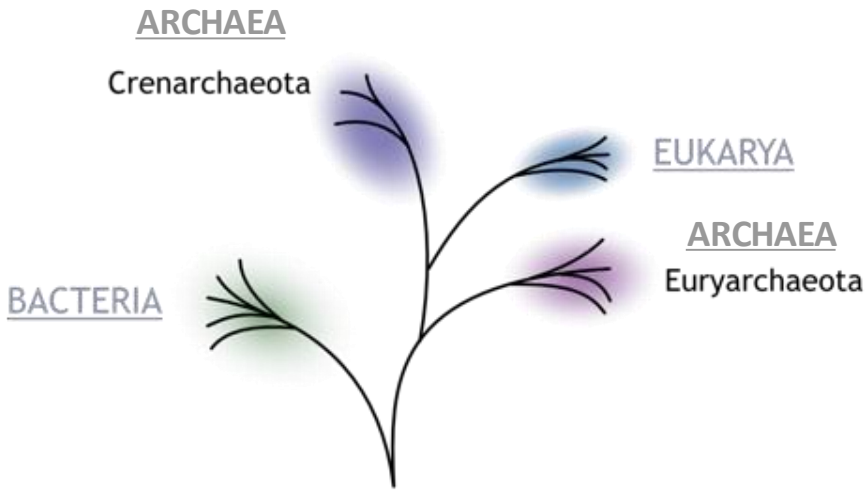


Subunit composition of the RNAPs

Archaea as sister-group or as ancestors of eukaryotes?



Three domain tree of Life



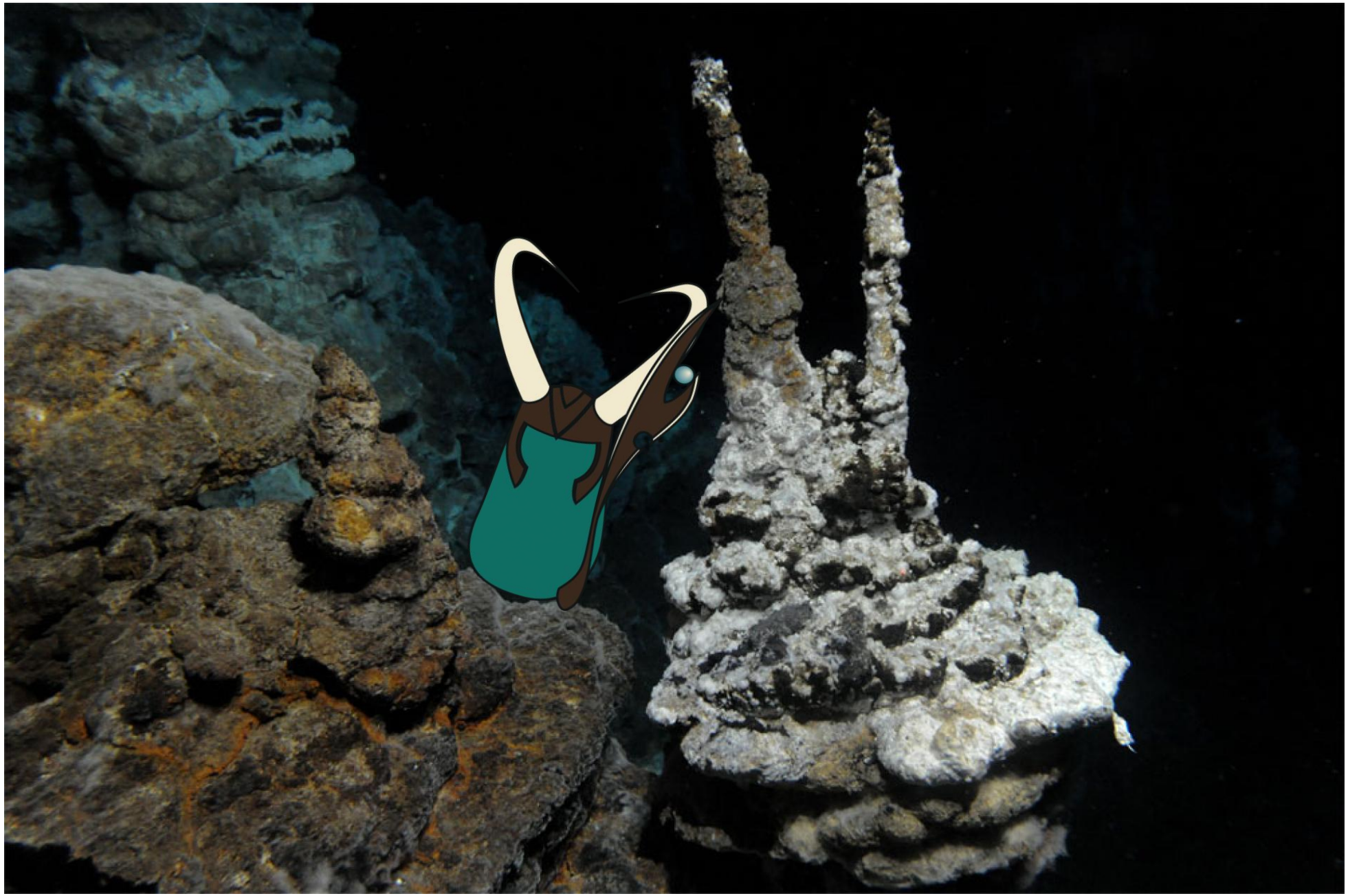
Two domain tree of Life

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*

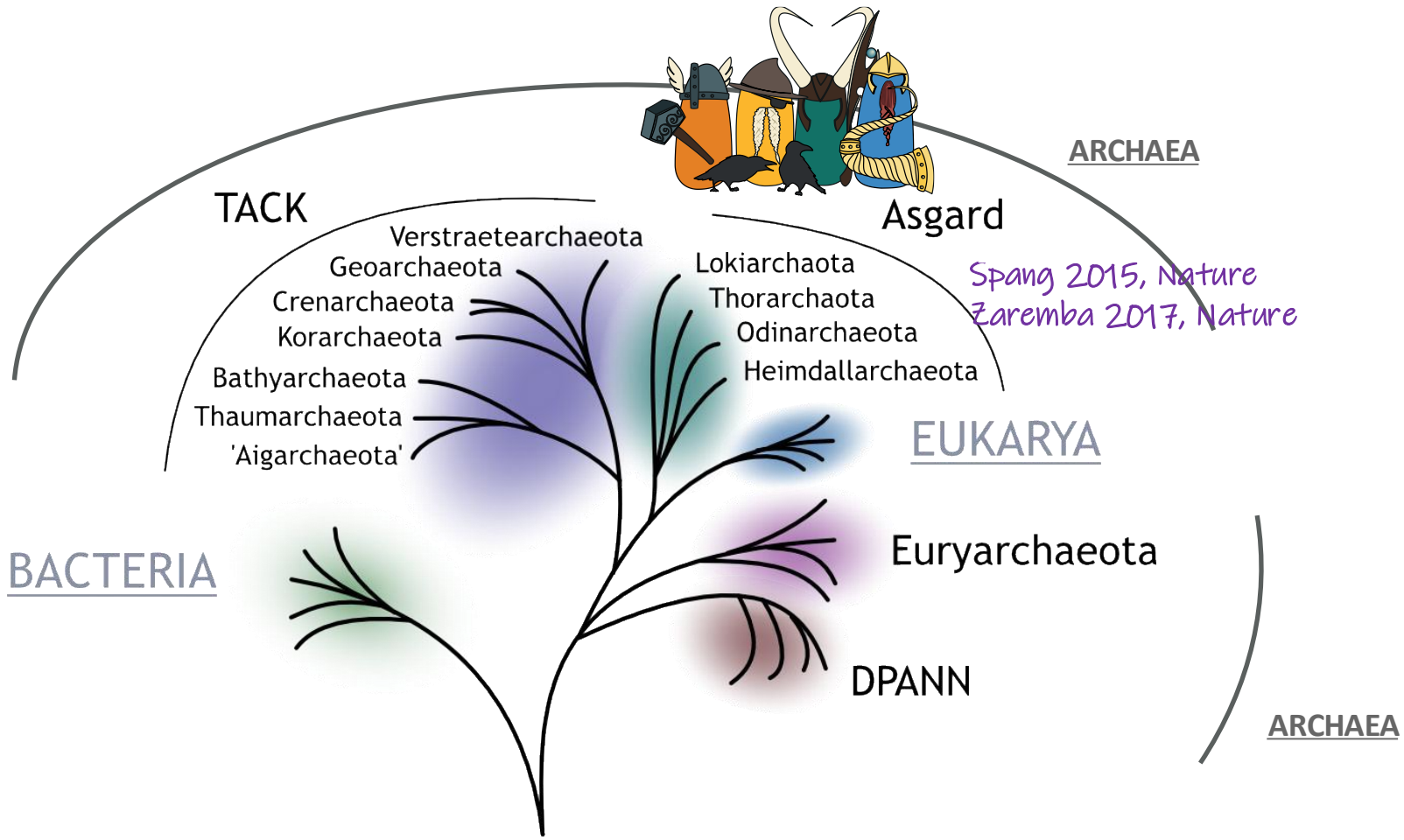


The unveiling of Asgard archaea through metagenomics

Asgard

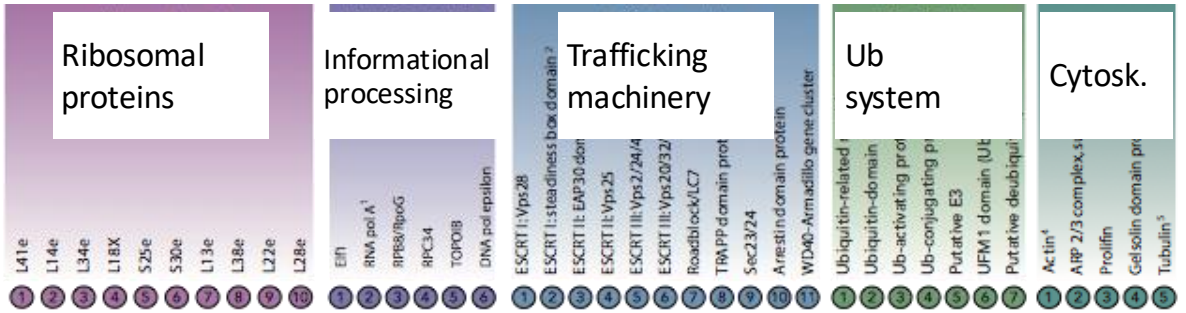
Spang 2015, Nature
Zaremba 2017, Nature

The unveiling of Asgard archaea through metagenomics



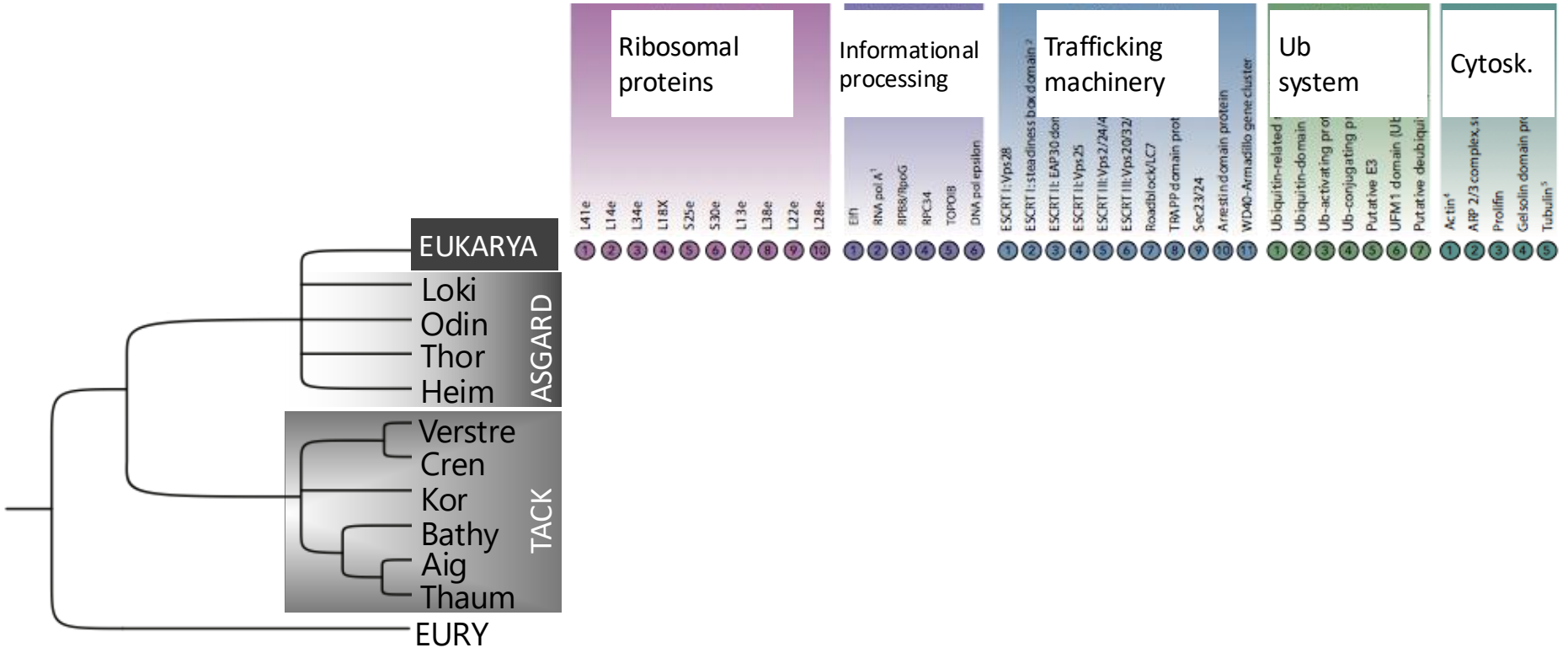
Eme et al. 2017, *Nat Rev Micro.*

Numerous Eukaryotic Signature Proteins (ESPs) in Asgard archaea



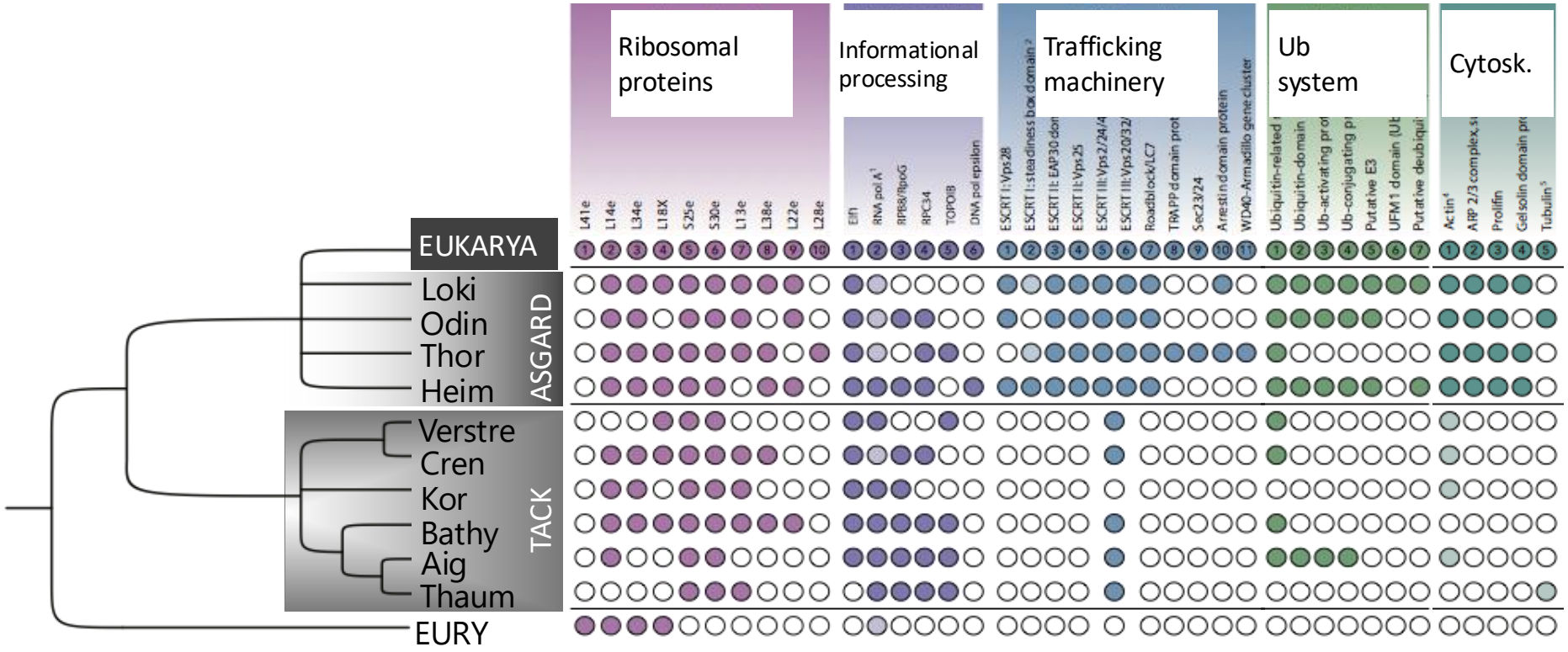
Eme et al. 2017, *Nat Rev Micro.*

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

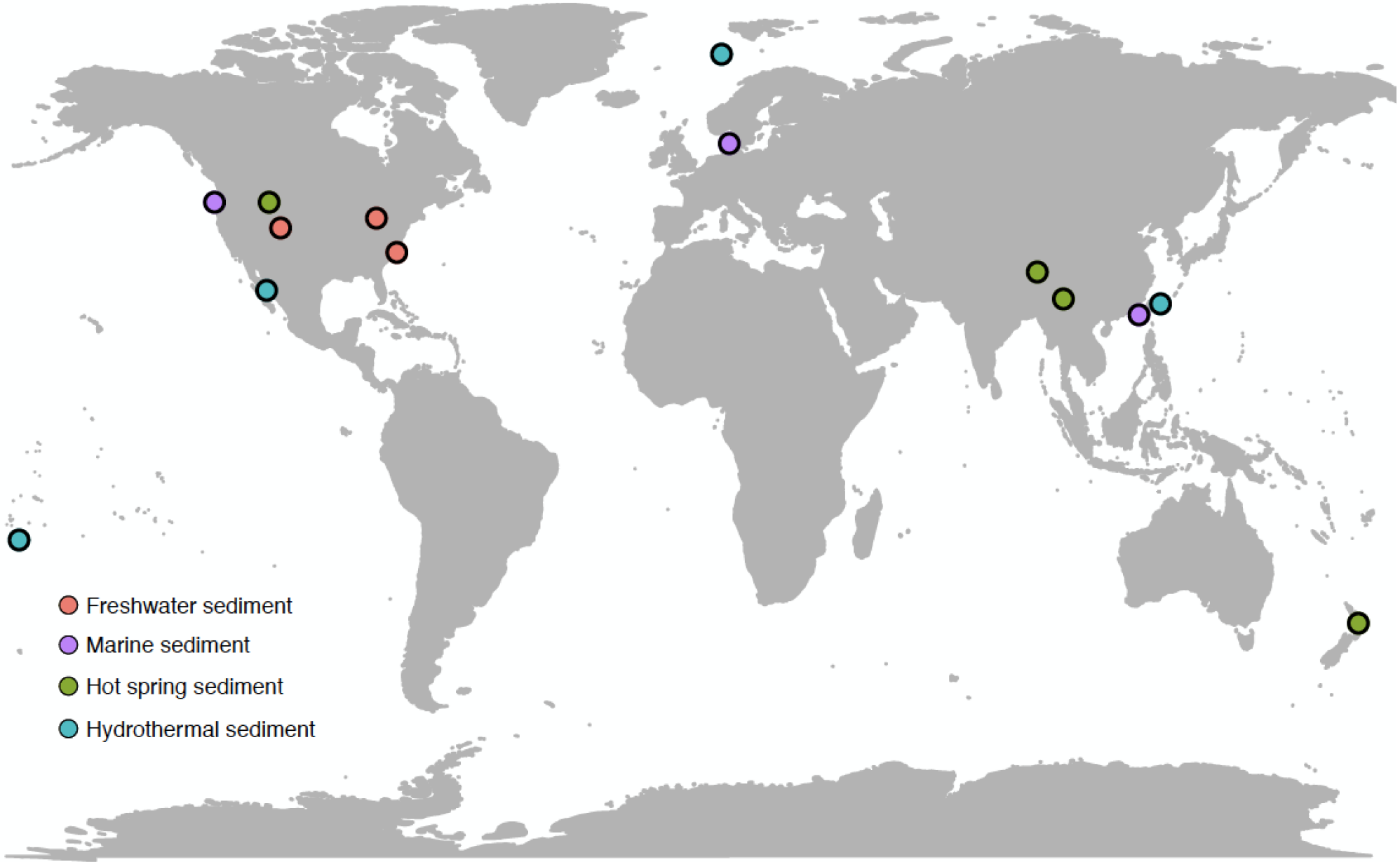
Received: 23 April 2021

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

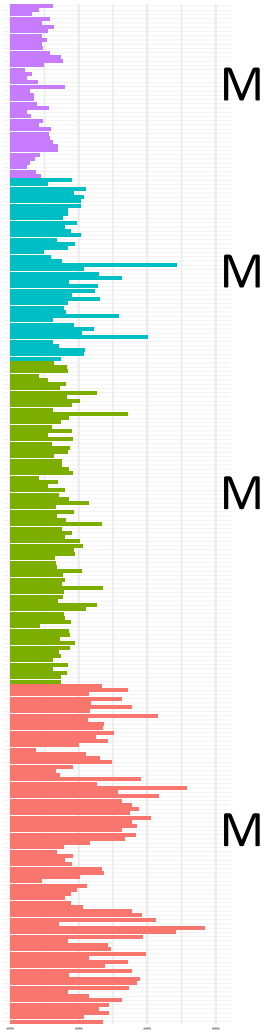
DPANN Median = 1.2 Mb

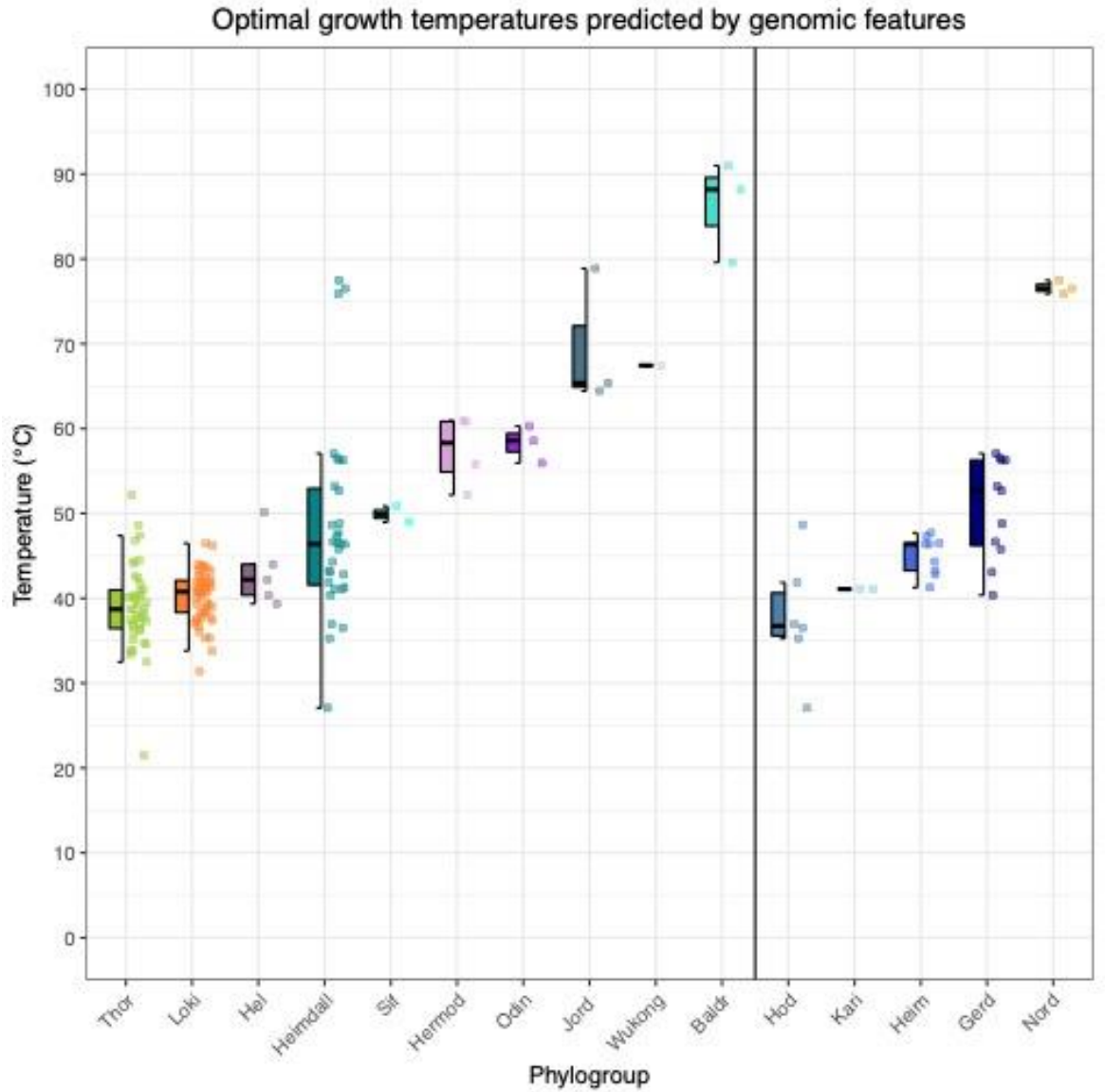
Euryarchaeota Median = 1.8 Mb

TACK Median = 1.8 Mb

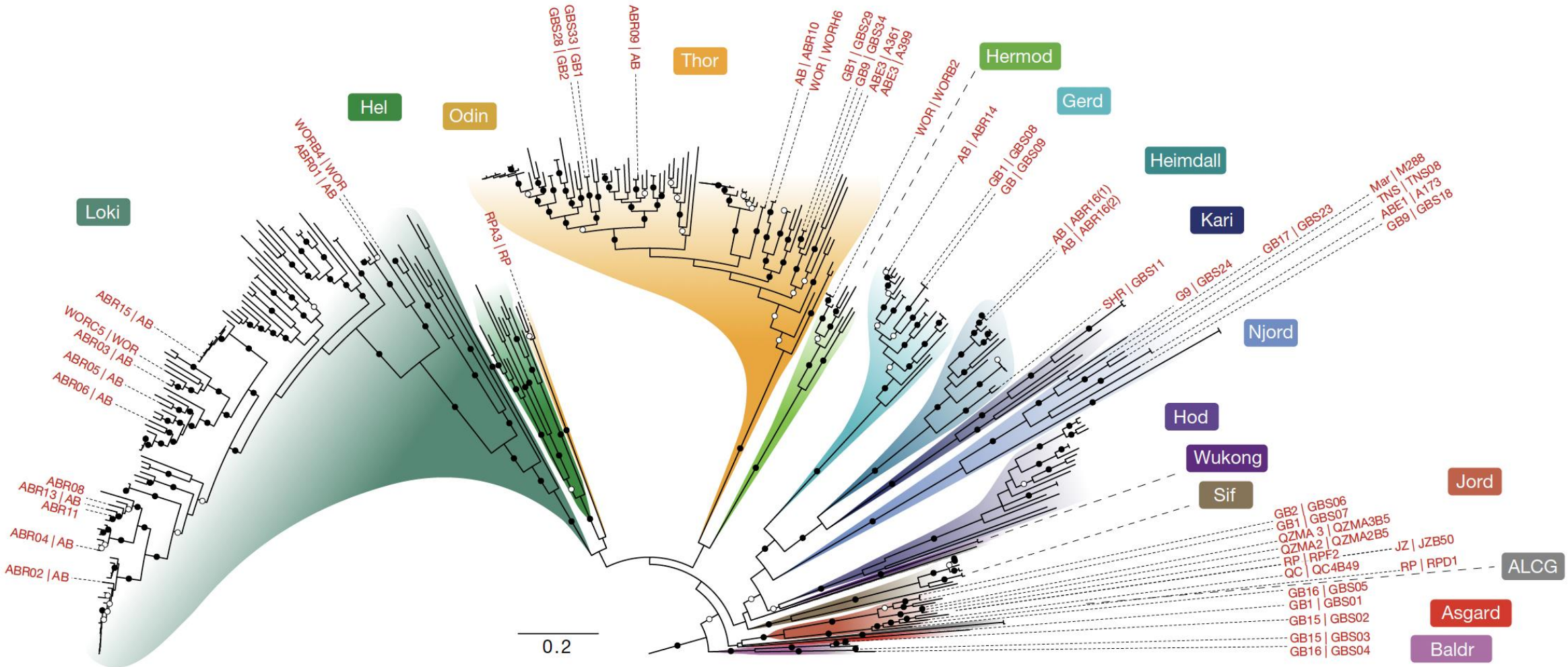
Asgard Median = 3.6 Mb

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?

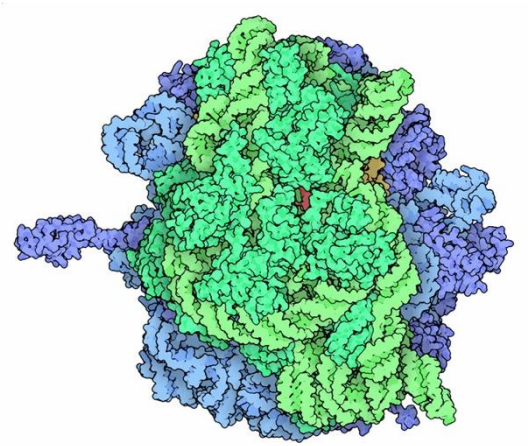
Ribosomal proteins:

- Slow evolving
- Universal
- Coevolve
- Short

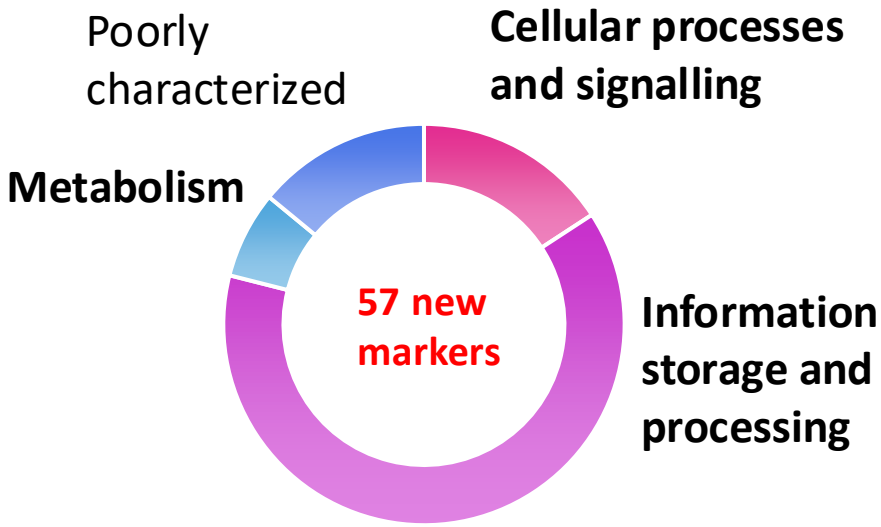
New markers:

- Conserved in eukaryote and archaea
- Not transferred horizontally

54 ribosomal proteins

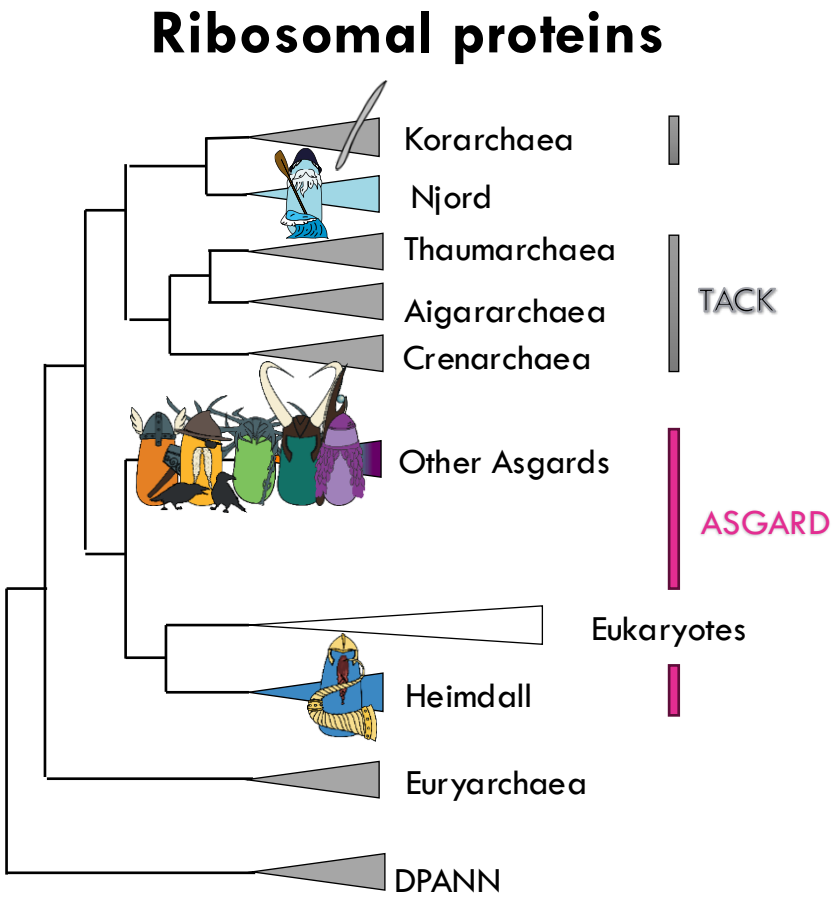


~6000 aa



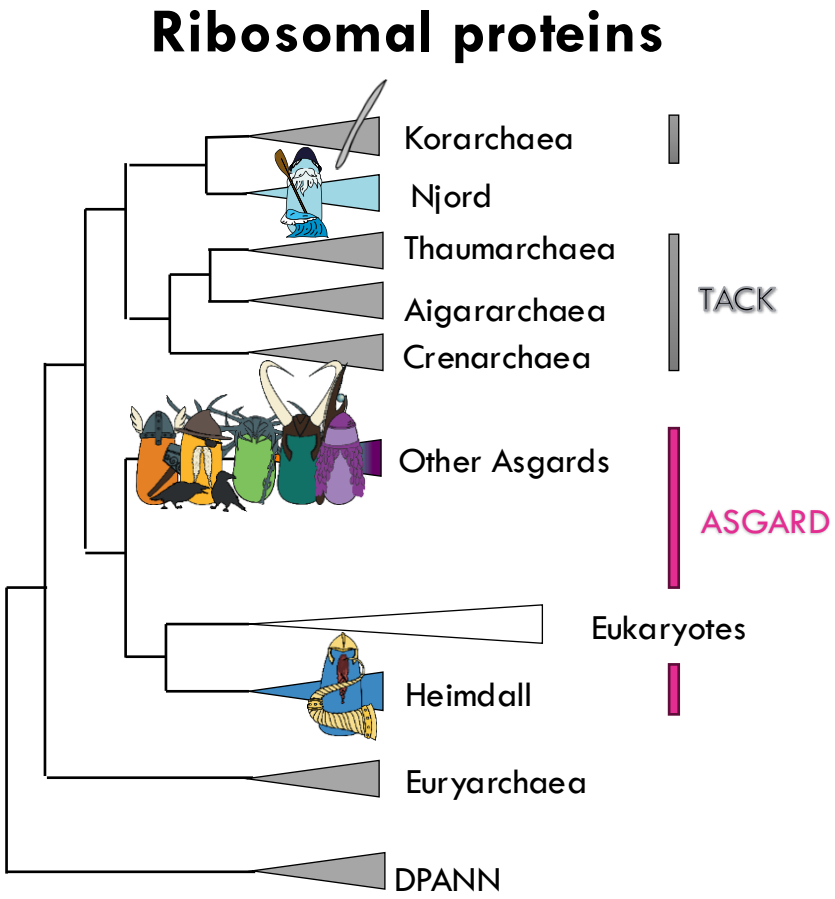
~14,000 aa

How do Eukaryotes relate to Asgards?



How do Eukaryotes relate to Asgards?

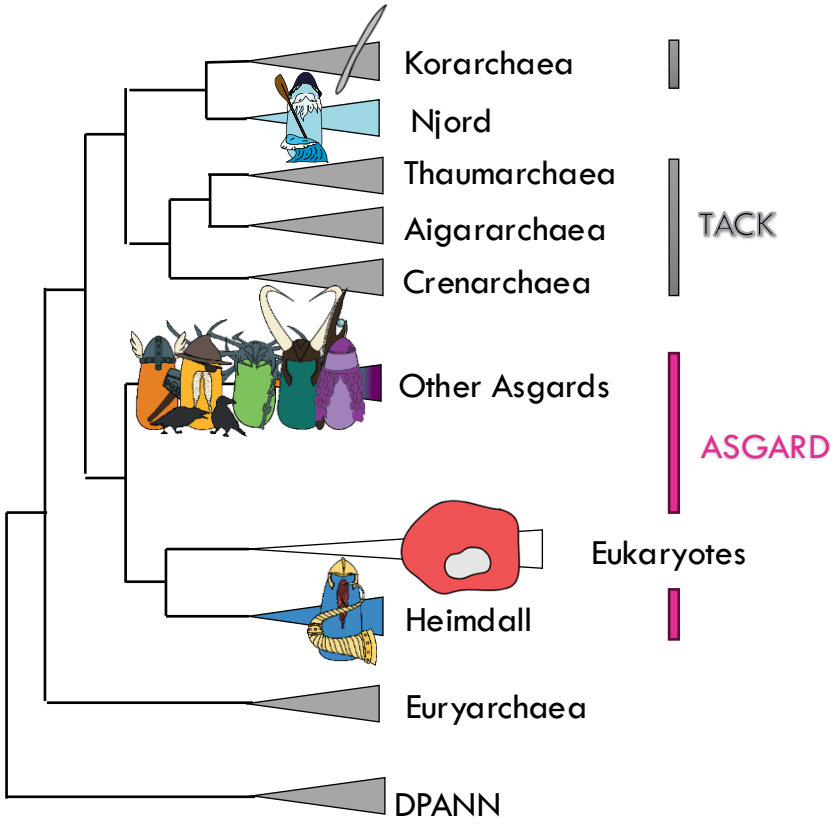
Many ESPs:
Actin,
RPL28e, …



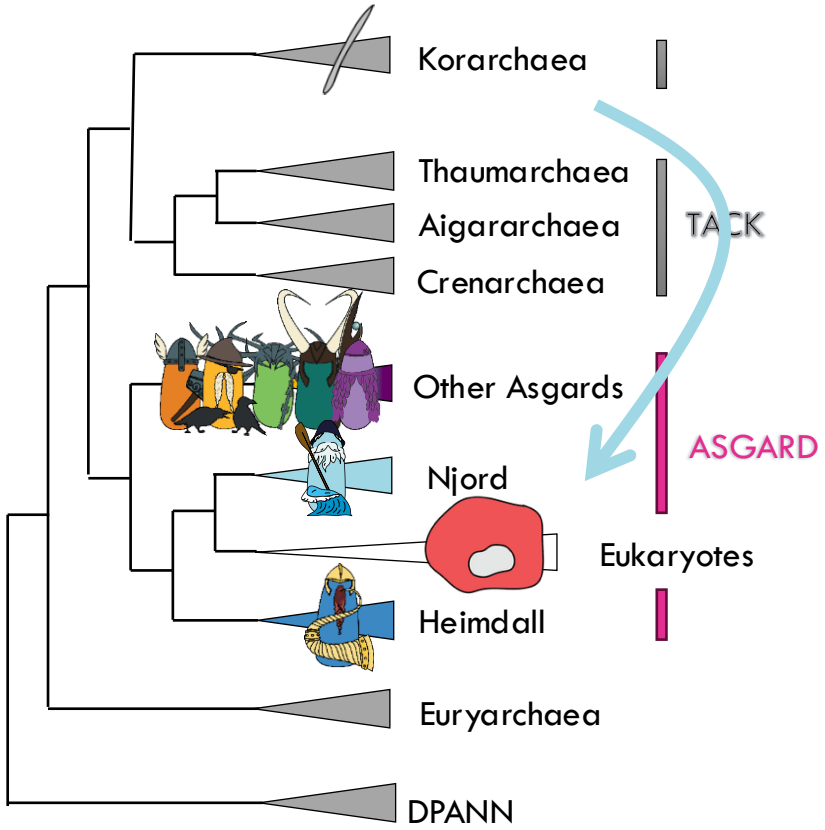
How do Eukaryotes relate to Asgards?

Many ESPs:
Actin,
RPL28e, …

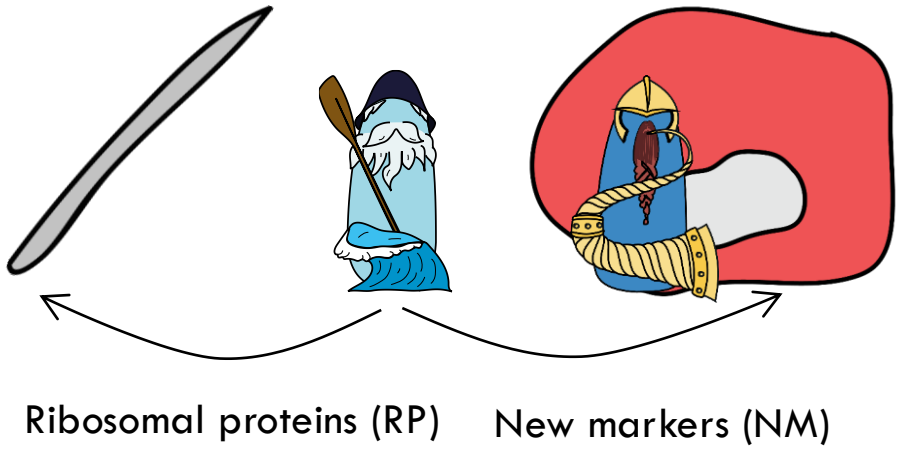
Ribosomal proteins



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)**

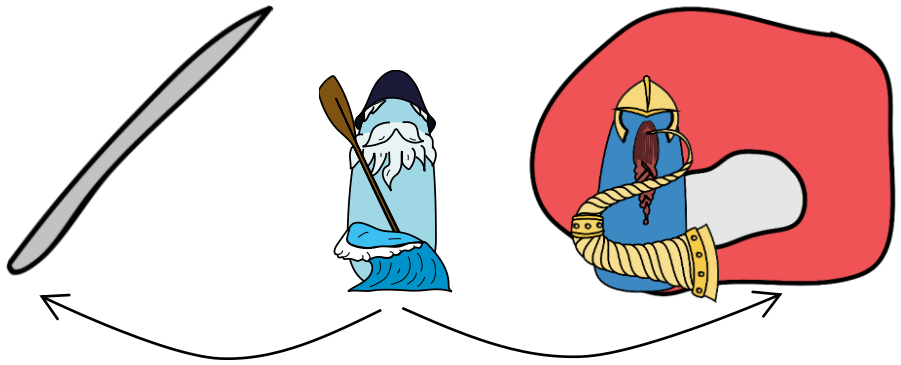
Taxon sampling

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

Fast-evolving site removal (FSR)

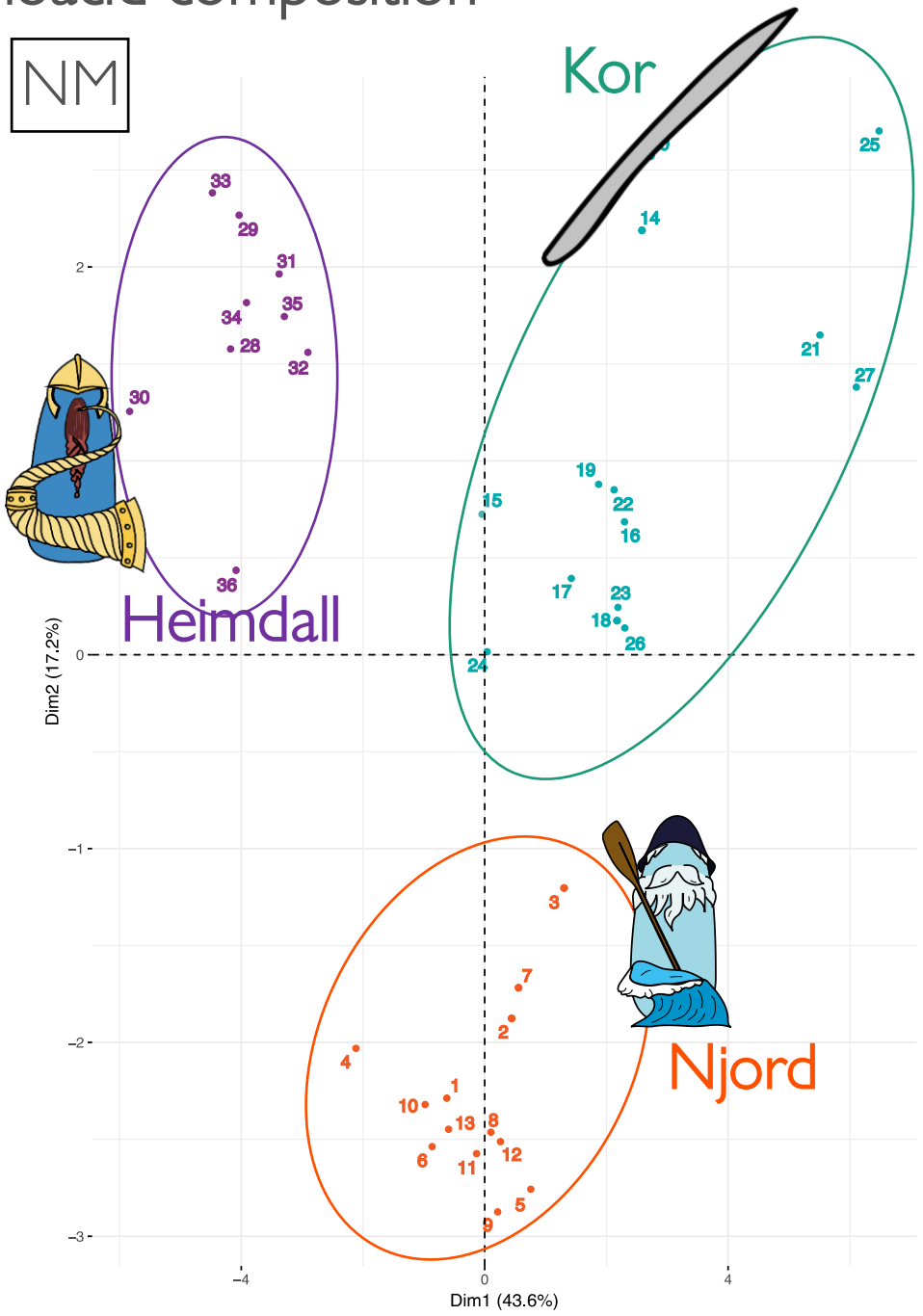
Recoding (SR4)

~800 phylogenies…

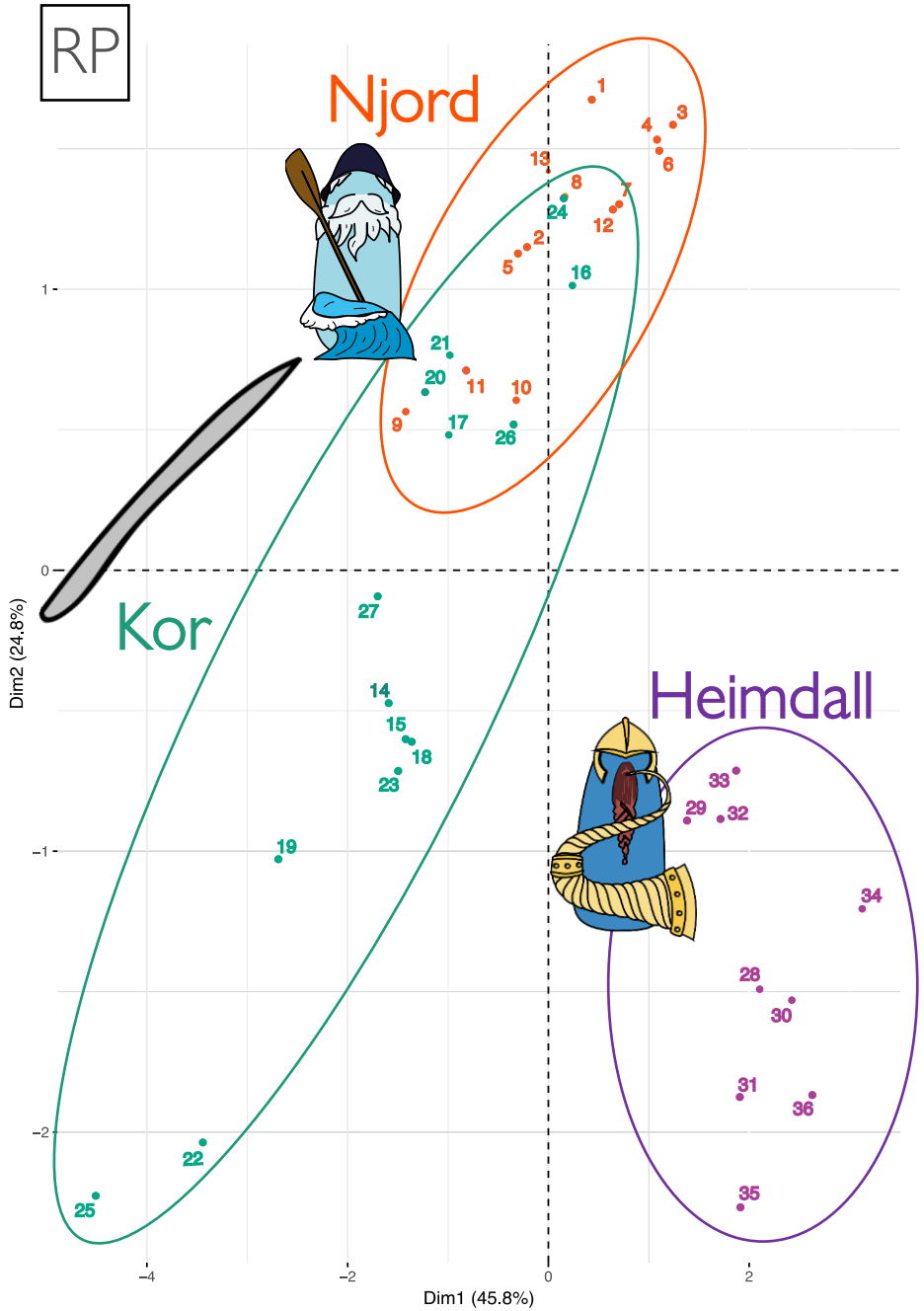


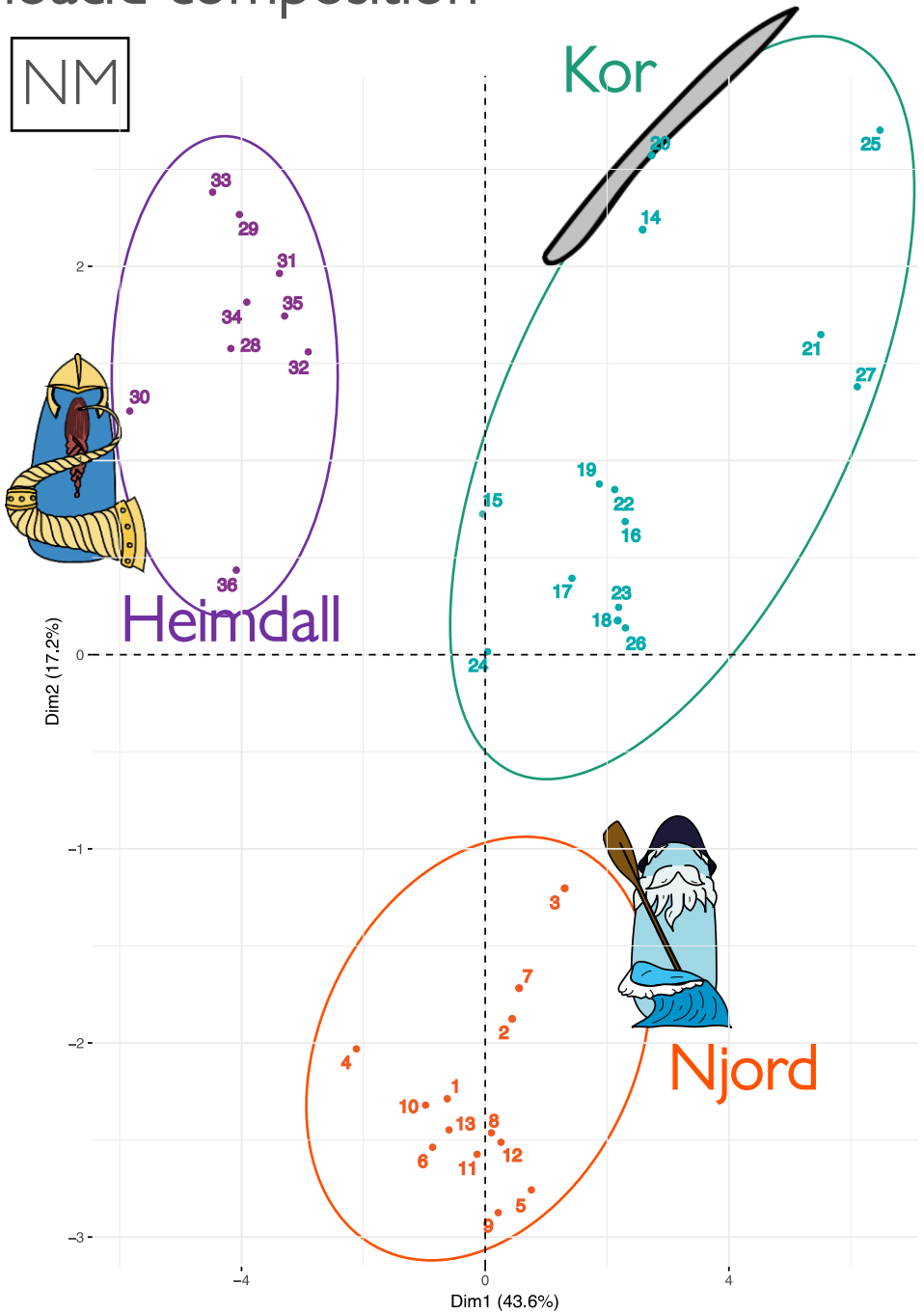
Ribosomal proteins (RP) New markers (NM)

PCA based on aminoacid composition



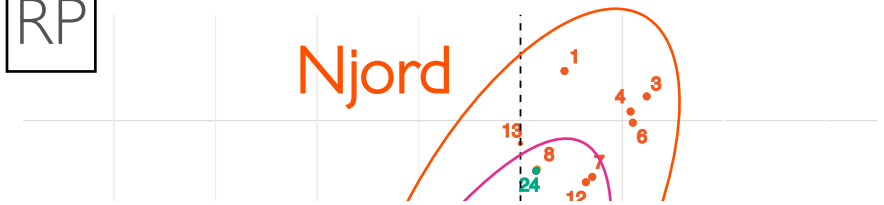
PCA based on aminoacid composition



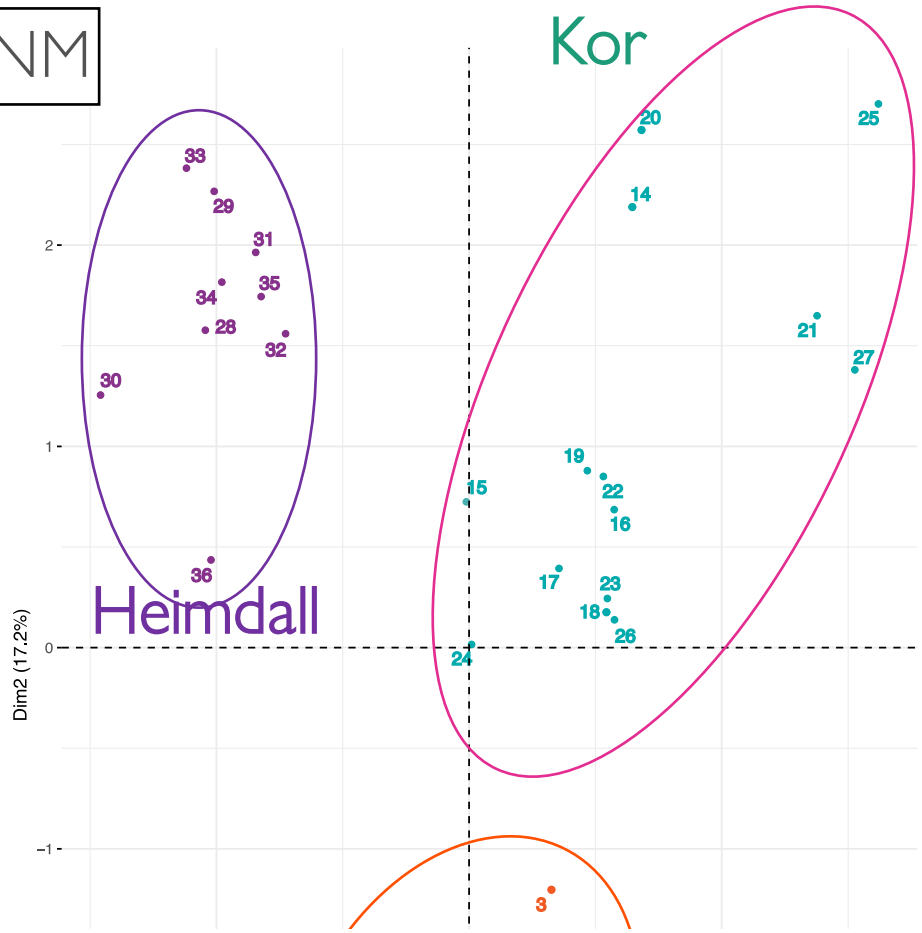


PCA based on aminoacid composition

RP



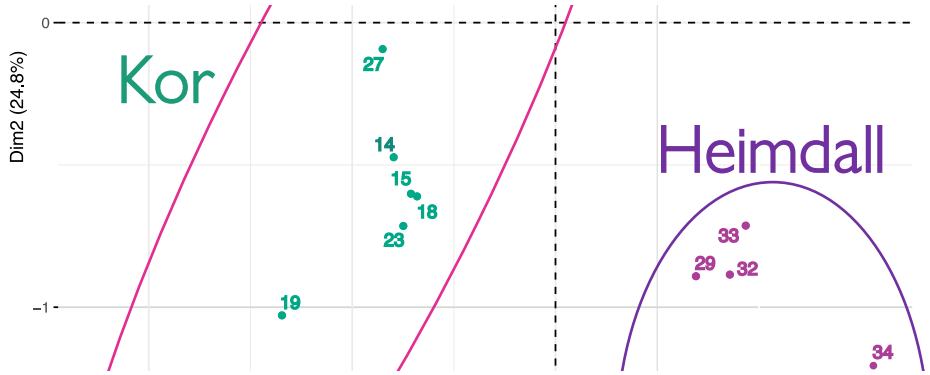
NM



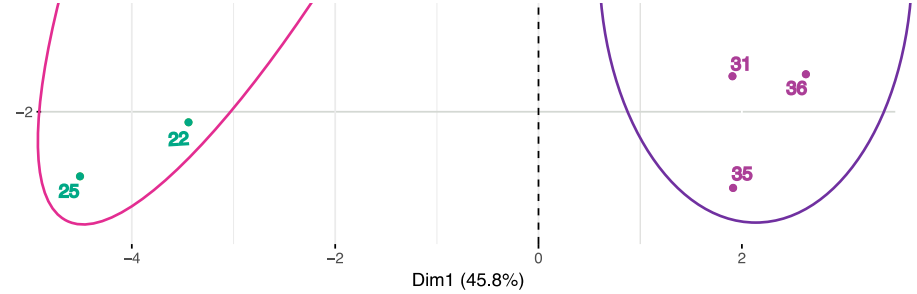
Njord and Kor hyperthermophiles:

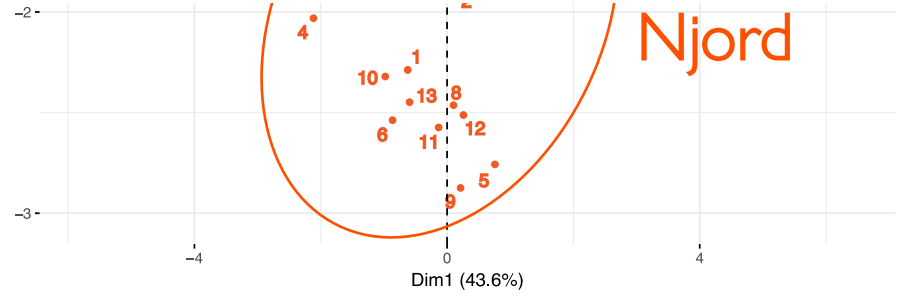
➔ rRNA composition bias

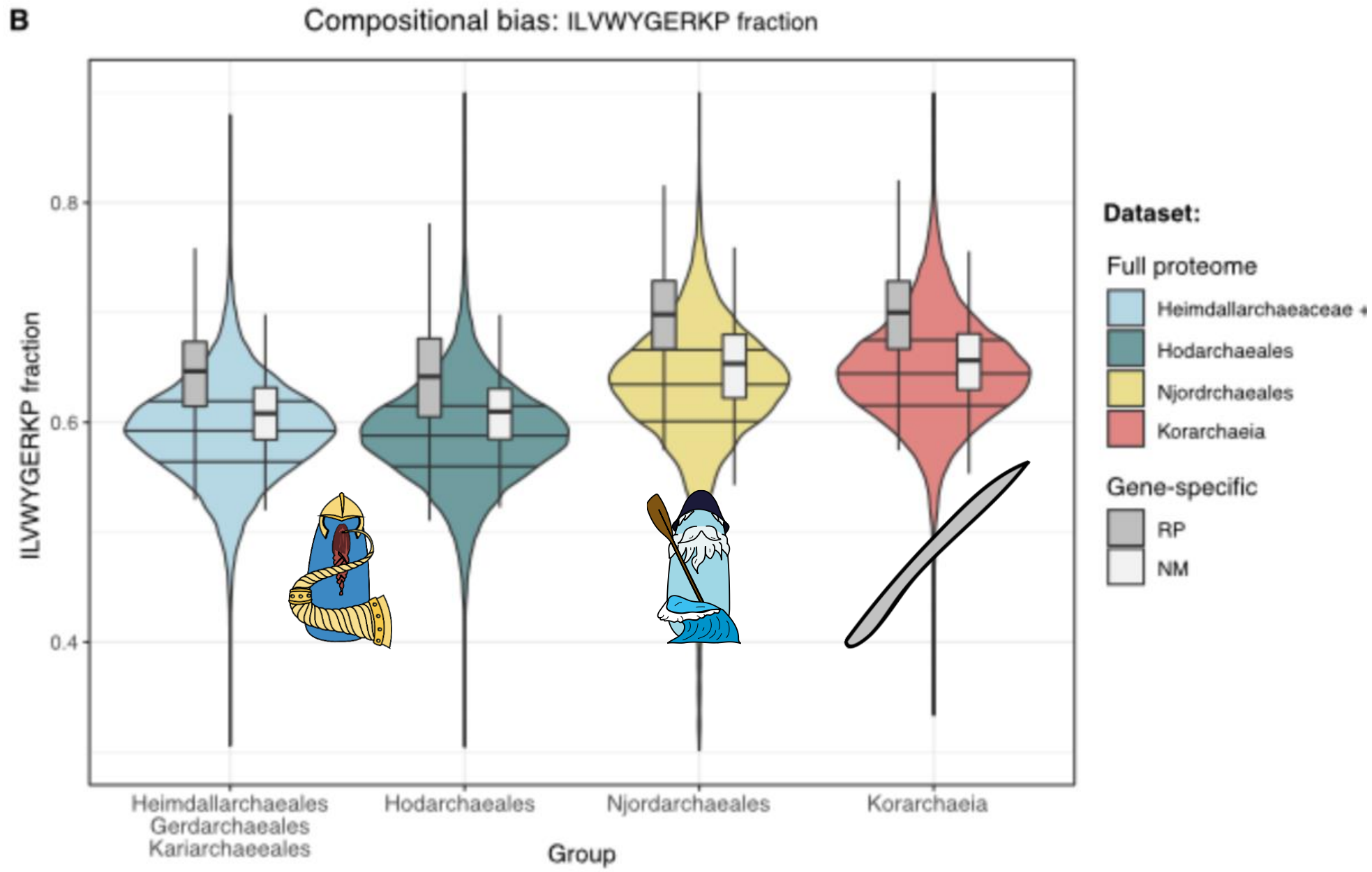
➔ structural constraints in the ribosome explaining convergent AA composition



➔ 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)**

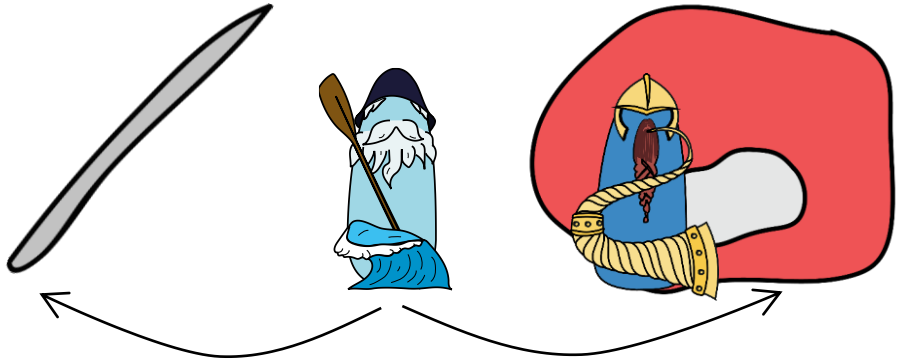
Taxon sampling

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

Fast-evolving site removal (FSR)

Recoding (SR4)

~800 phylogenies…

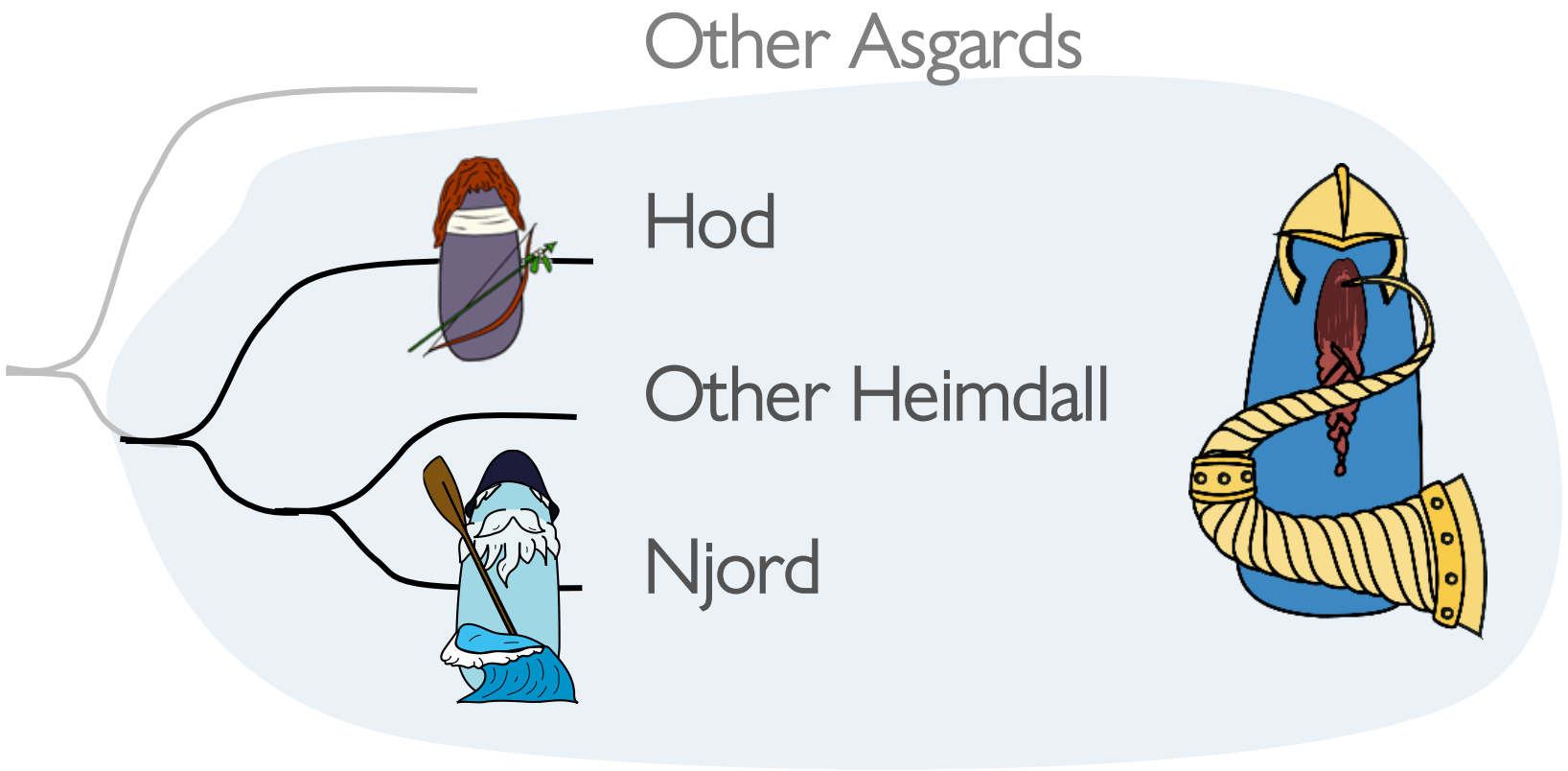


Ribosomal proteins (RP)

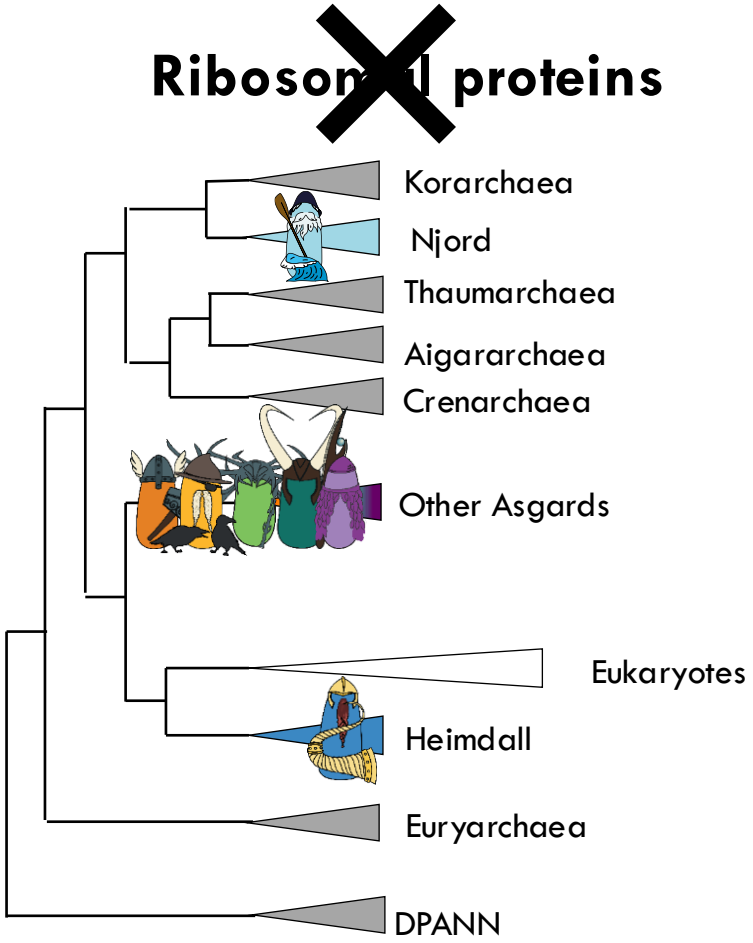
New markers (NM)

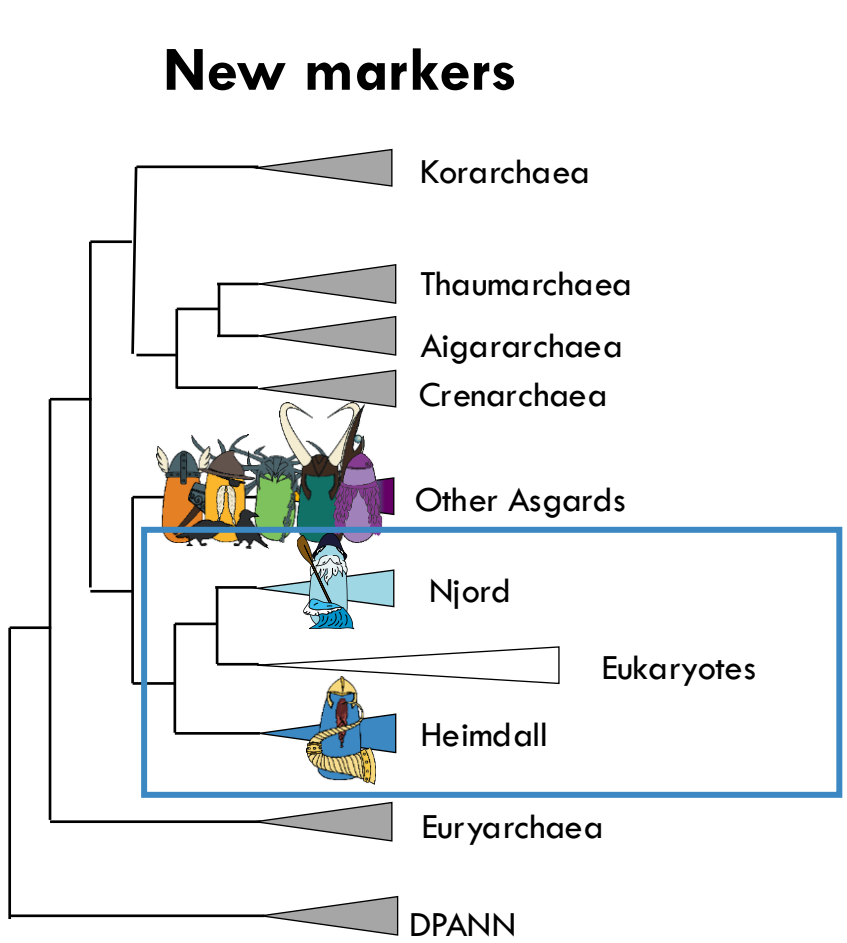
RP (without Korarchaea)

Njord are in fact a sub-clade of Heimdallarchaeia



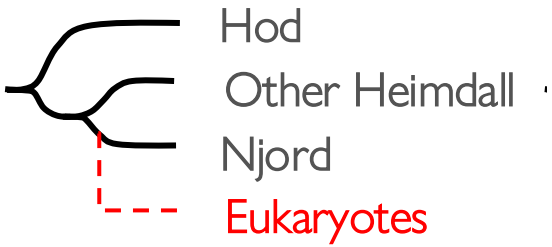
Why do we care, again?





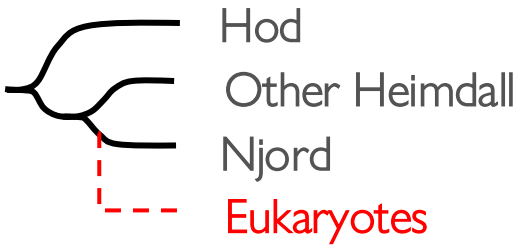
Placing Eukaryotes in the Asgard tree require careful phylogenetic investigations

Untreated

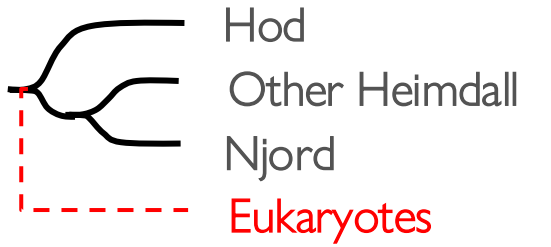


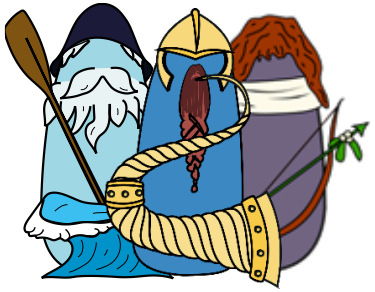
Placing Eukaryotes in the Asgard tree require careful phylogenetic investigations

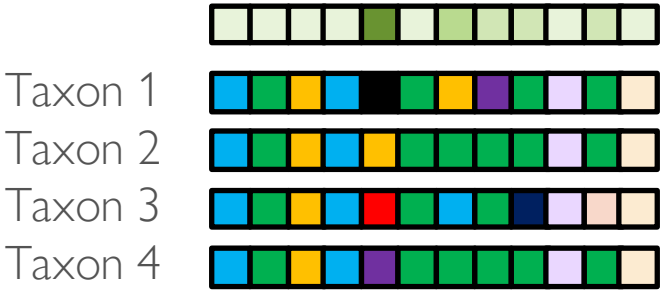
Untreated



Fast Site Removal

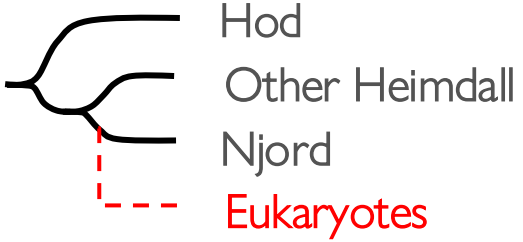




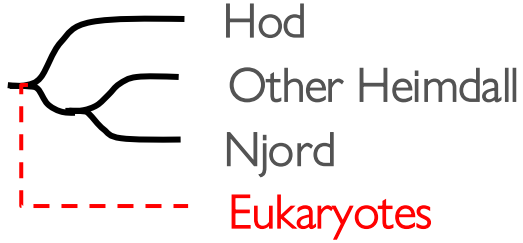


Placing Eukaryotes in the Asgard tree require careful phylogenetic investigations

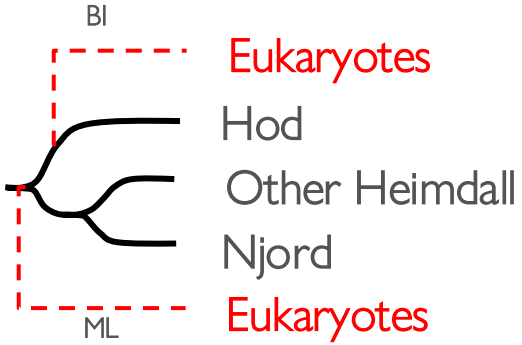
Untreated

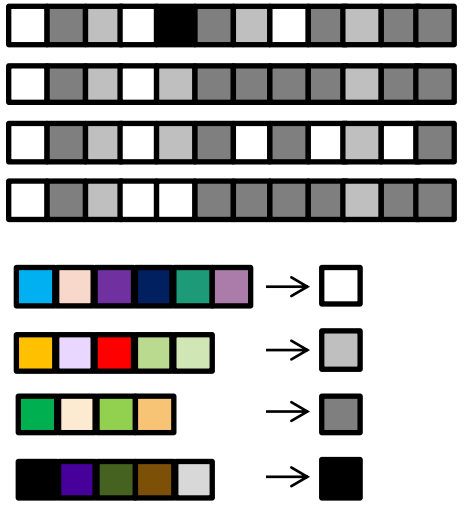


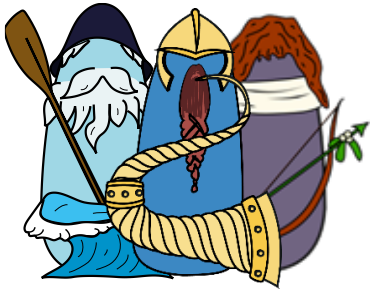
Fast Site Removal

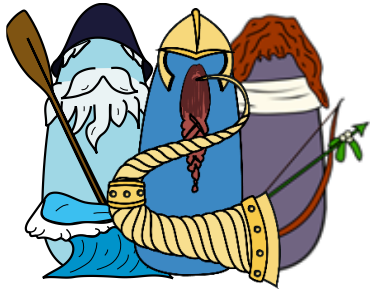


Recoding





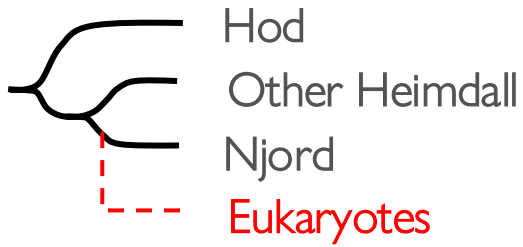




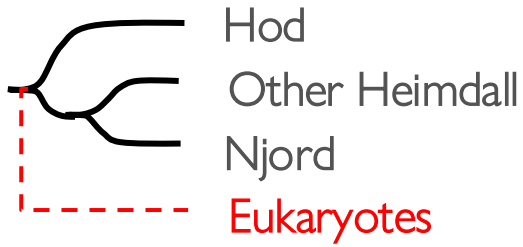
E.g.: SR4 recoding (Susko and Roger 2007)

Placing Eukaryotes in the Asgard tree require careful phylogenetic investigations

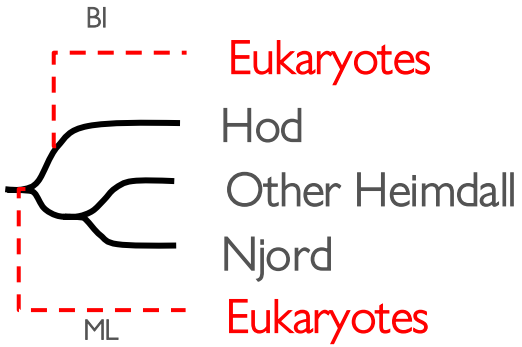
Untreated



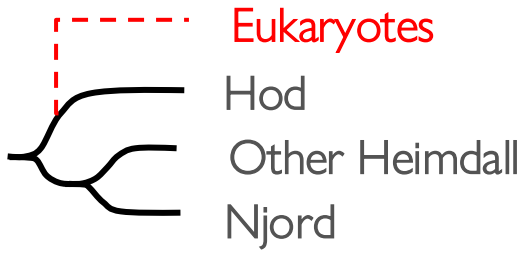
Fast Site Removal

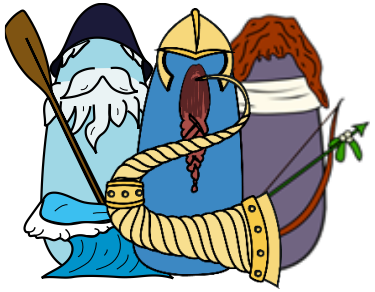


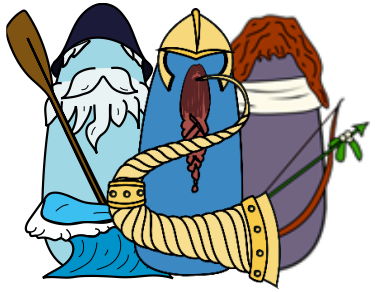
Recoding



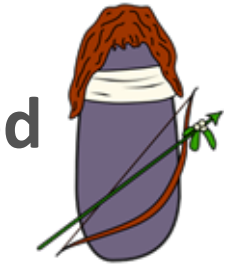
FSR+Recoding

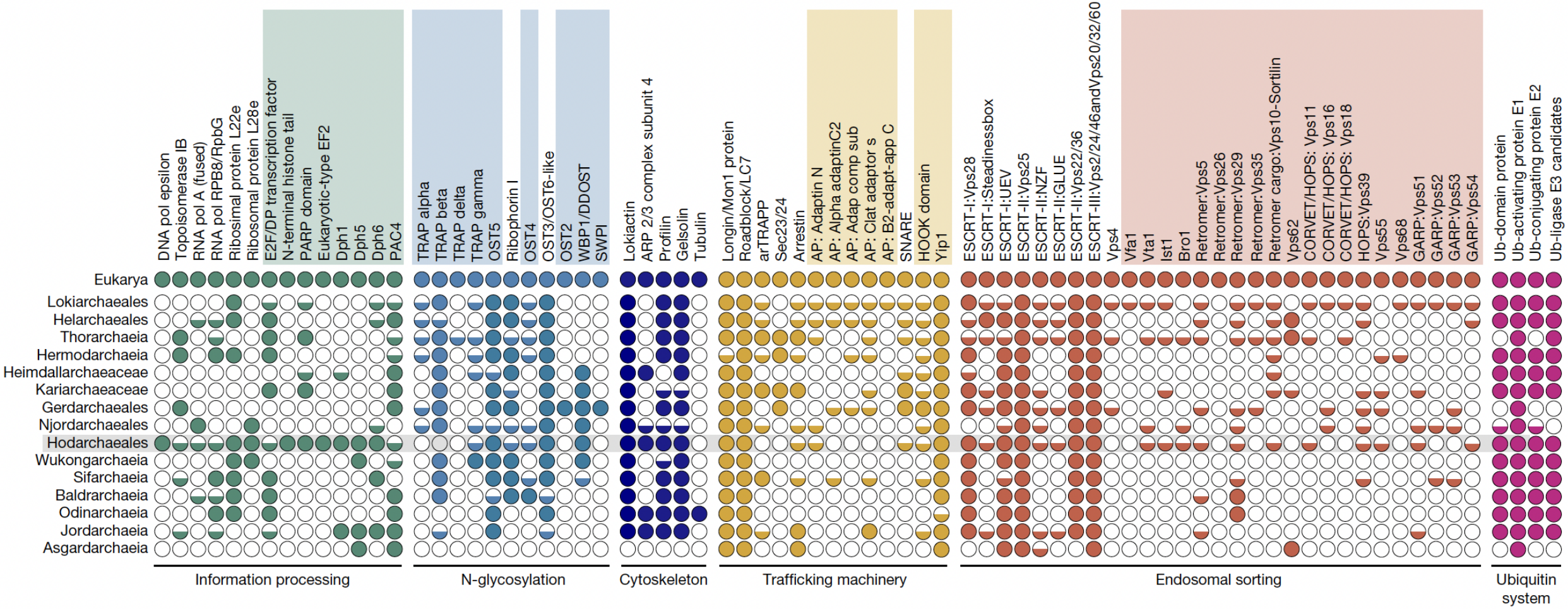




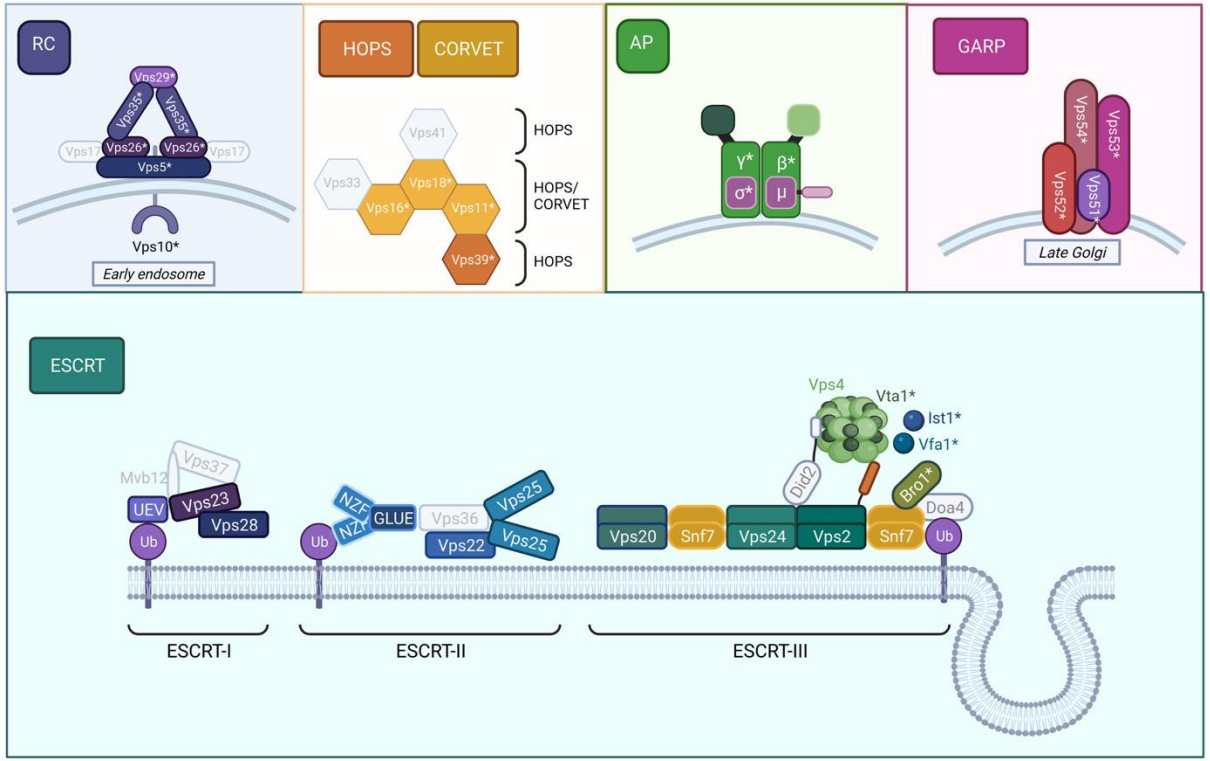


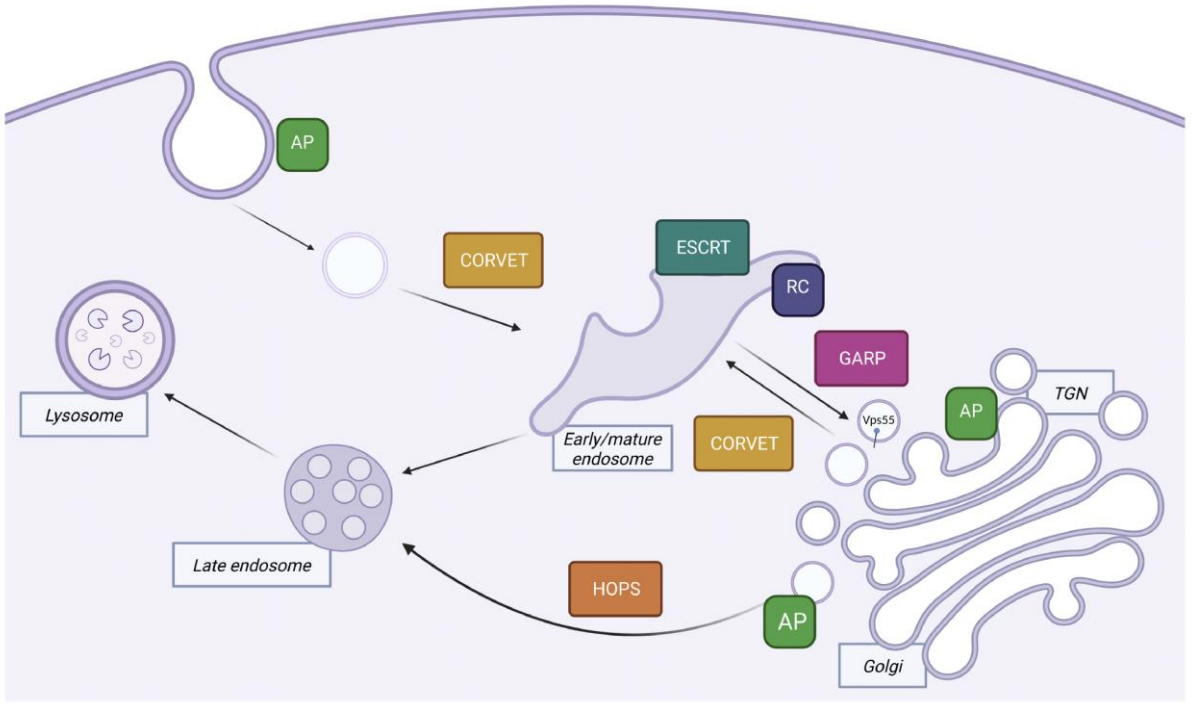
New ESPs support the relationship of Eukaryotes and Hod



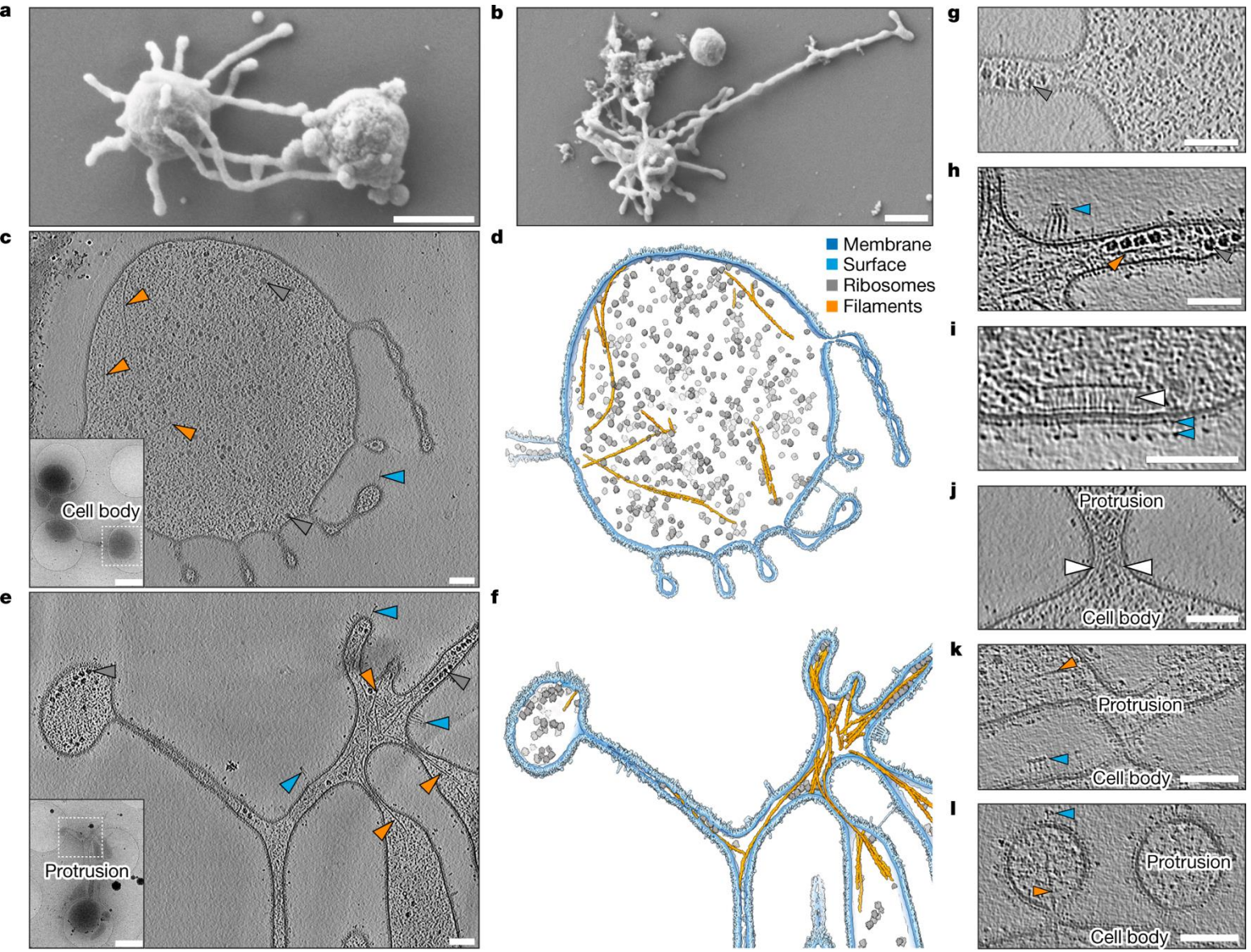


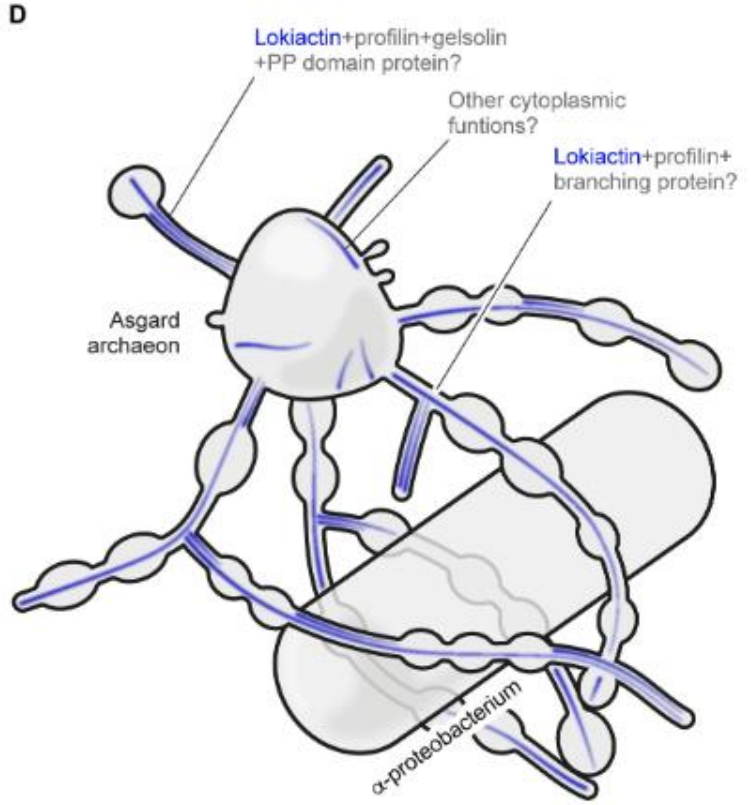
New ESPs involved in complex intracellular trafficking





Asgard archaea have an actin-based cytoskeleton





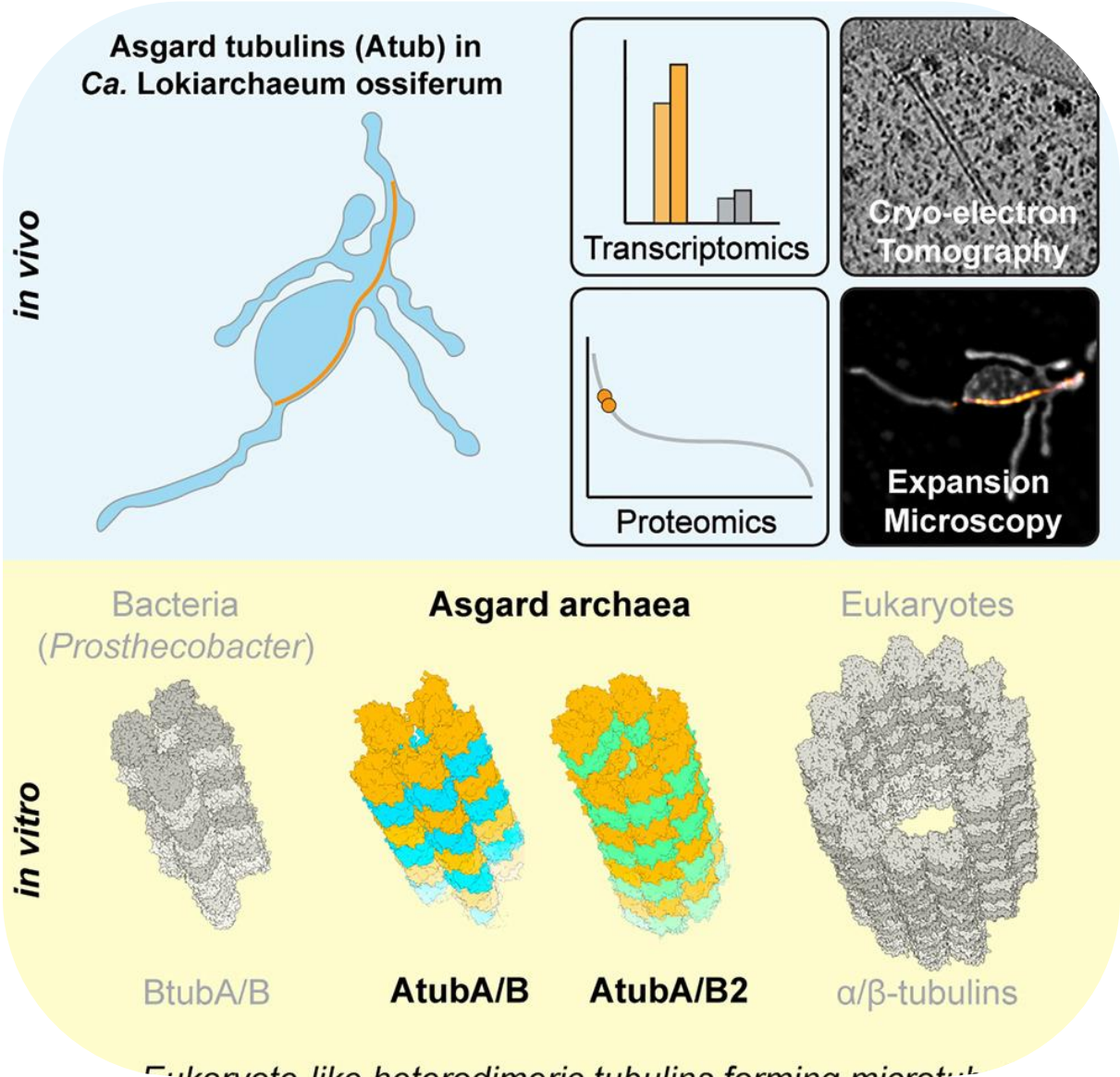
Charles-Orszag A, Petek-Seoane NA, Mullins RD. 2024. Archaeal actins and the origin of a multi-functional cytoskeleton. J Bacteriol

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).

(some) Asgard archaea have microtubule-like structures

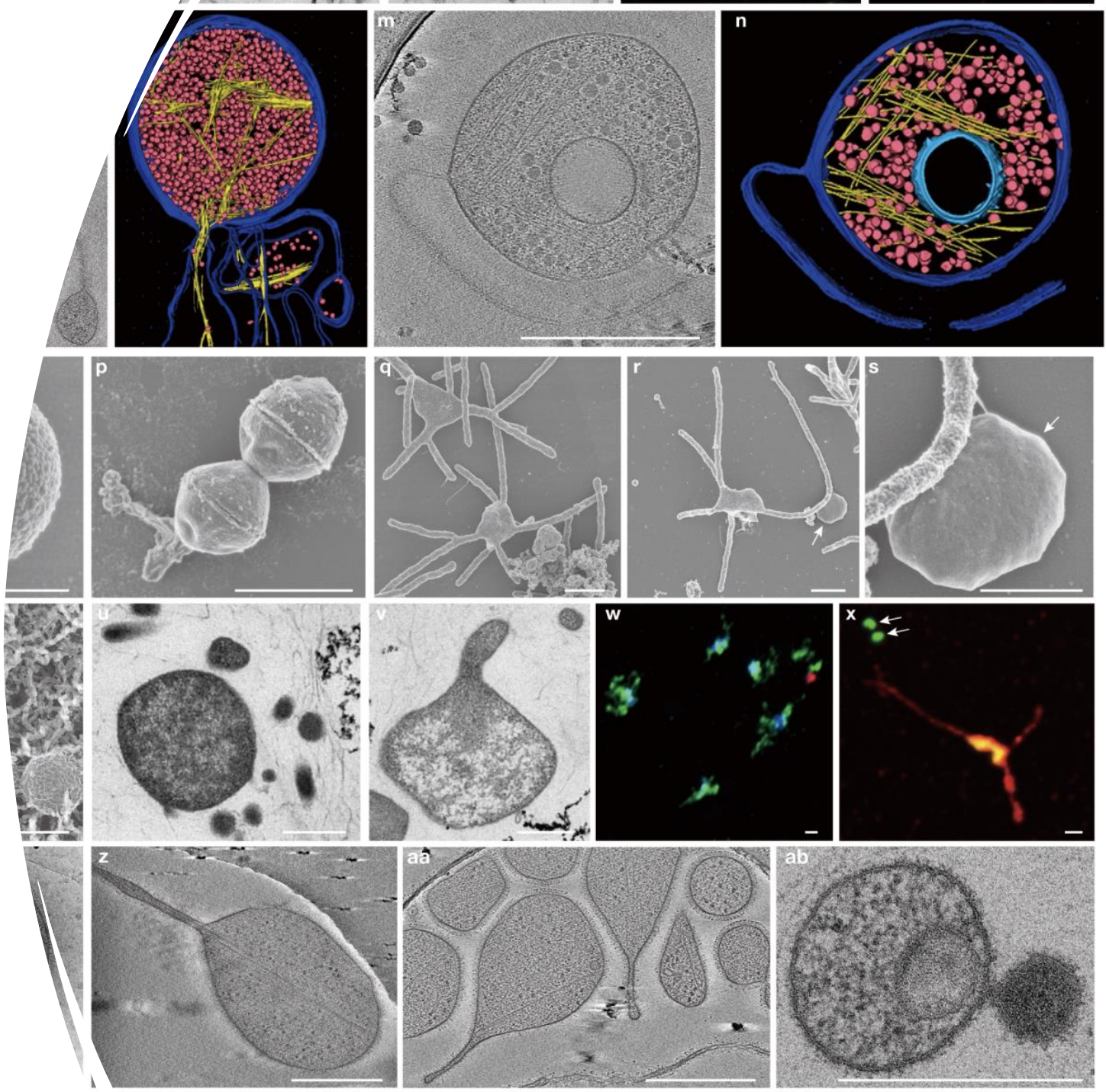
Wollweber et al. 2025:

- OdinTubulin forms eukaryote-like heterodimers that assemble into *bona fide* microtubules *in vitro*
- Canonical 5-protofilament forms + non-canonical 7-protofilament forms.
- Pre-eukaryotic origin for microtubules in addition to actin.



First cultured Hodarchaeal member

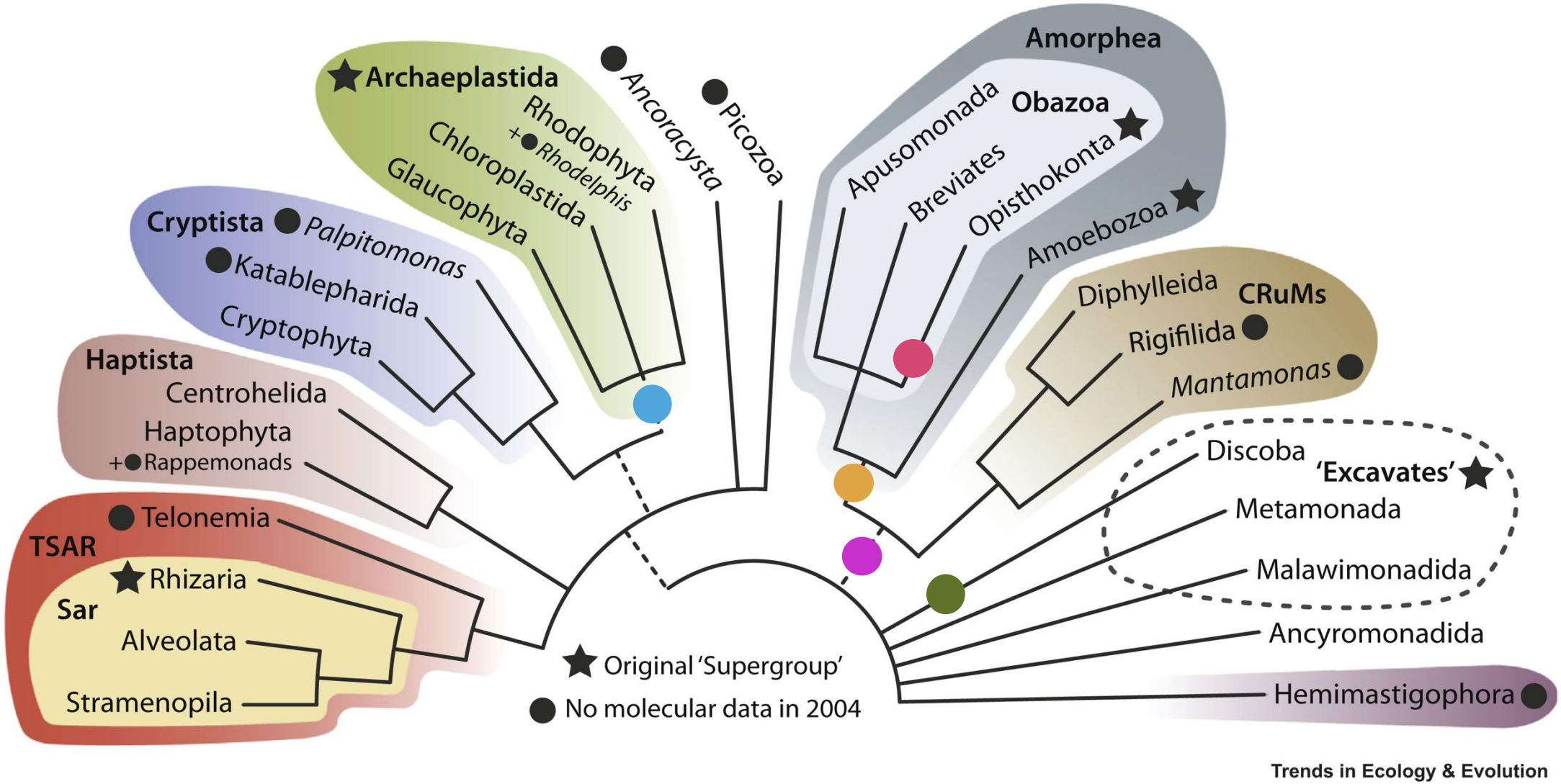
- Imachi et al. 2025: *Candidatus* Margulisarchaeum peptidophila,
- First cultured representative of what our 2023 paper put as eukaryotes' closest relatives.



Example 3

Rooting the tree of eukaryotes

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



Modified from Burki et al., 2020

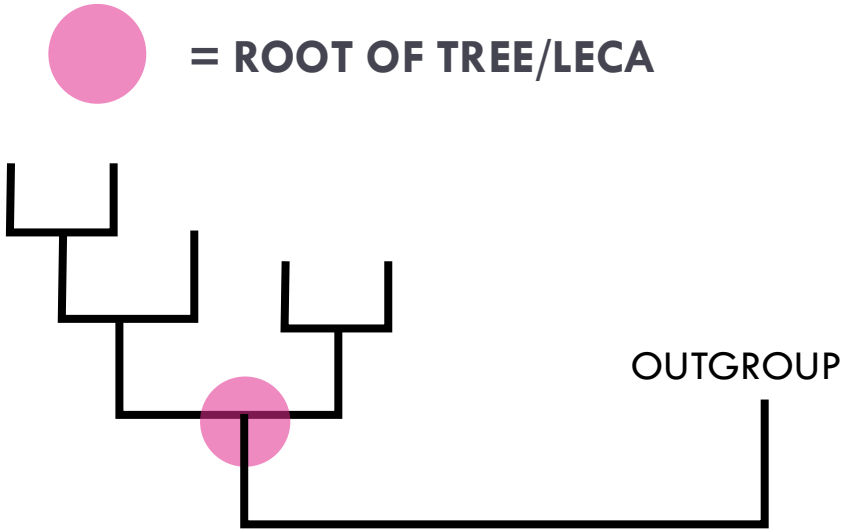
Gene tree parsimony, gene fusions and domain evolution, molecular and cellular features, replacement of highly conserved amino acids…

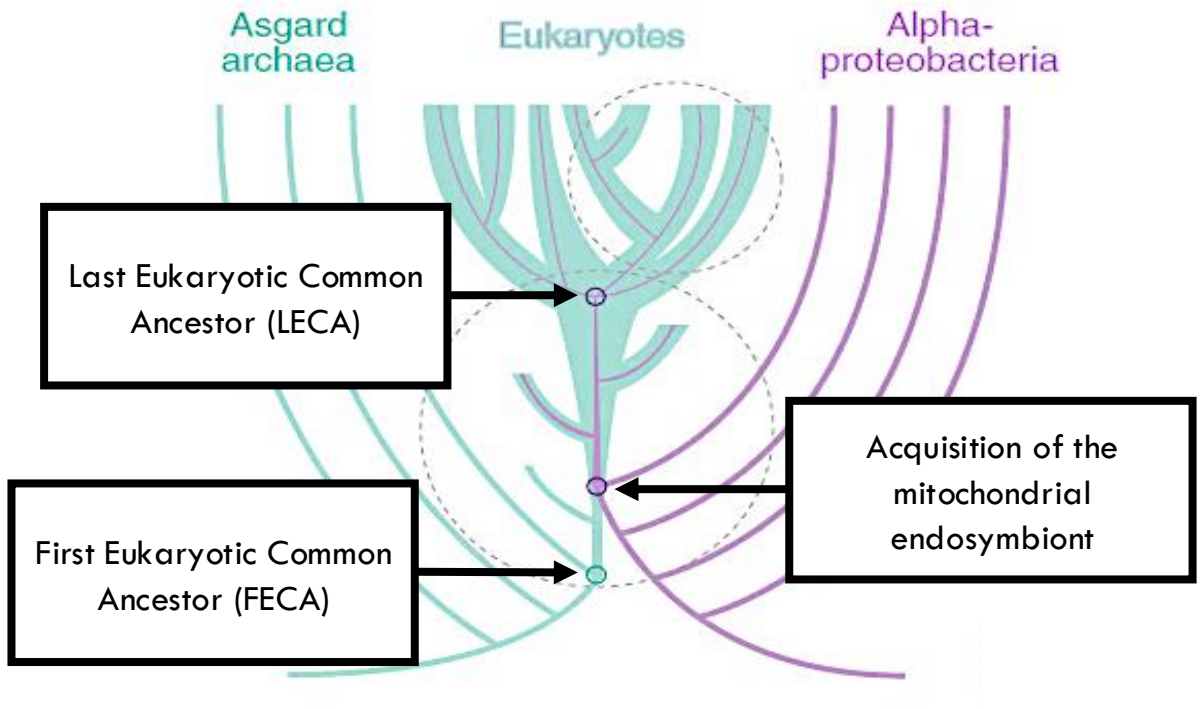
PROBLEMS:

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)

x5

Manual removal of non-orthologs and outliers

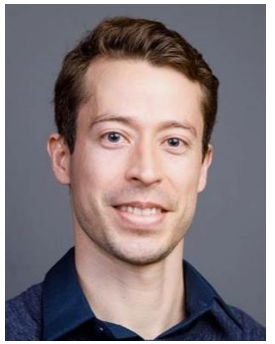
Select outgroup size and composition

FINAL DATASETS

93 marker genes
63 eukaryotes
37 alphaproteobacteria

Kelsey Williamson

Laura Eme



Sergio Muñoz-Gómez

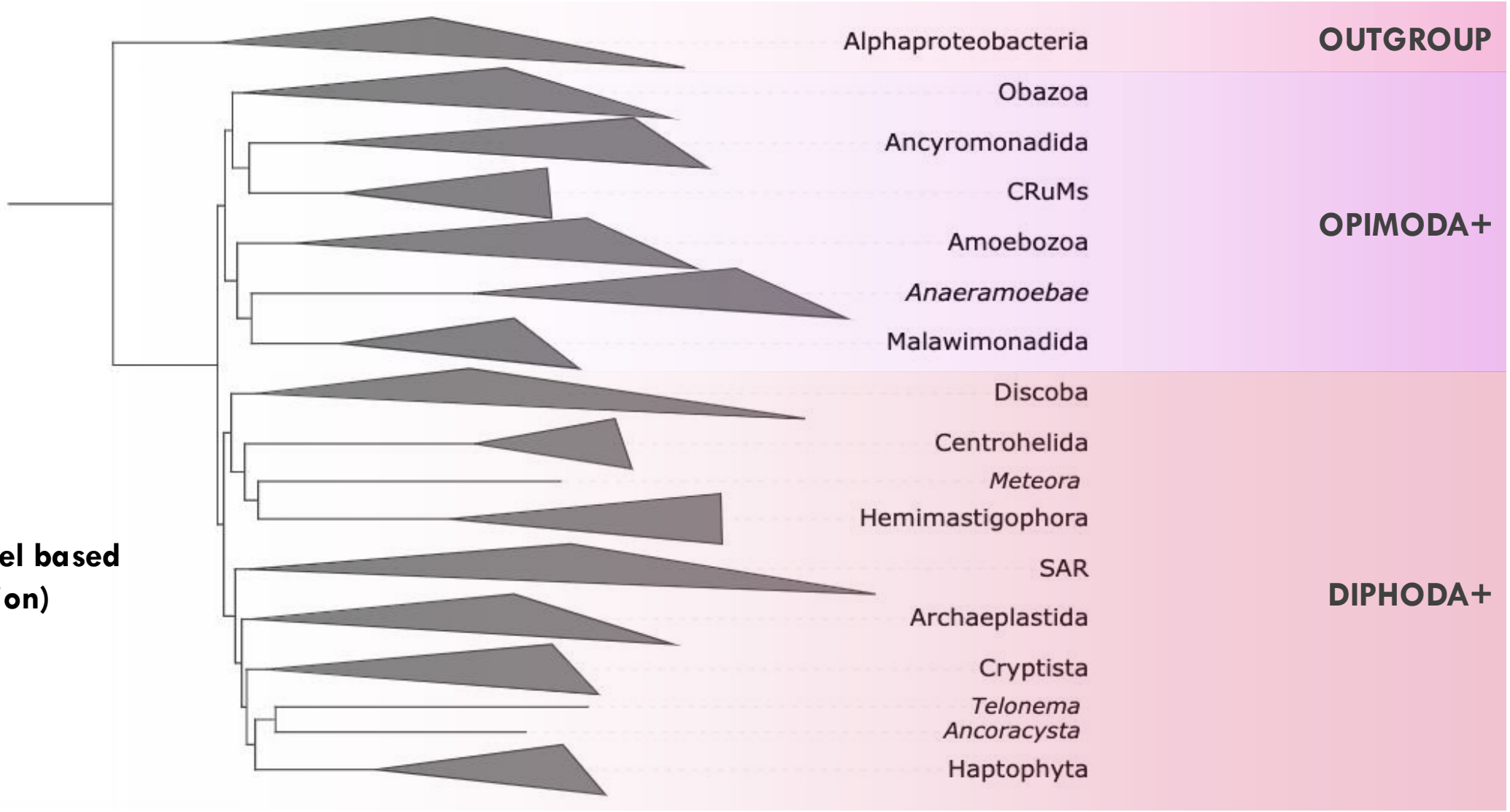
TESTING THE POSITION OF THE ROOT

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

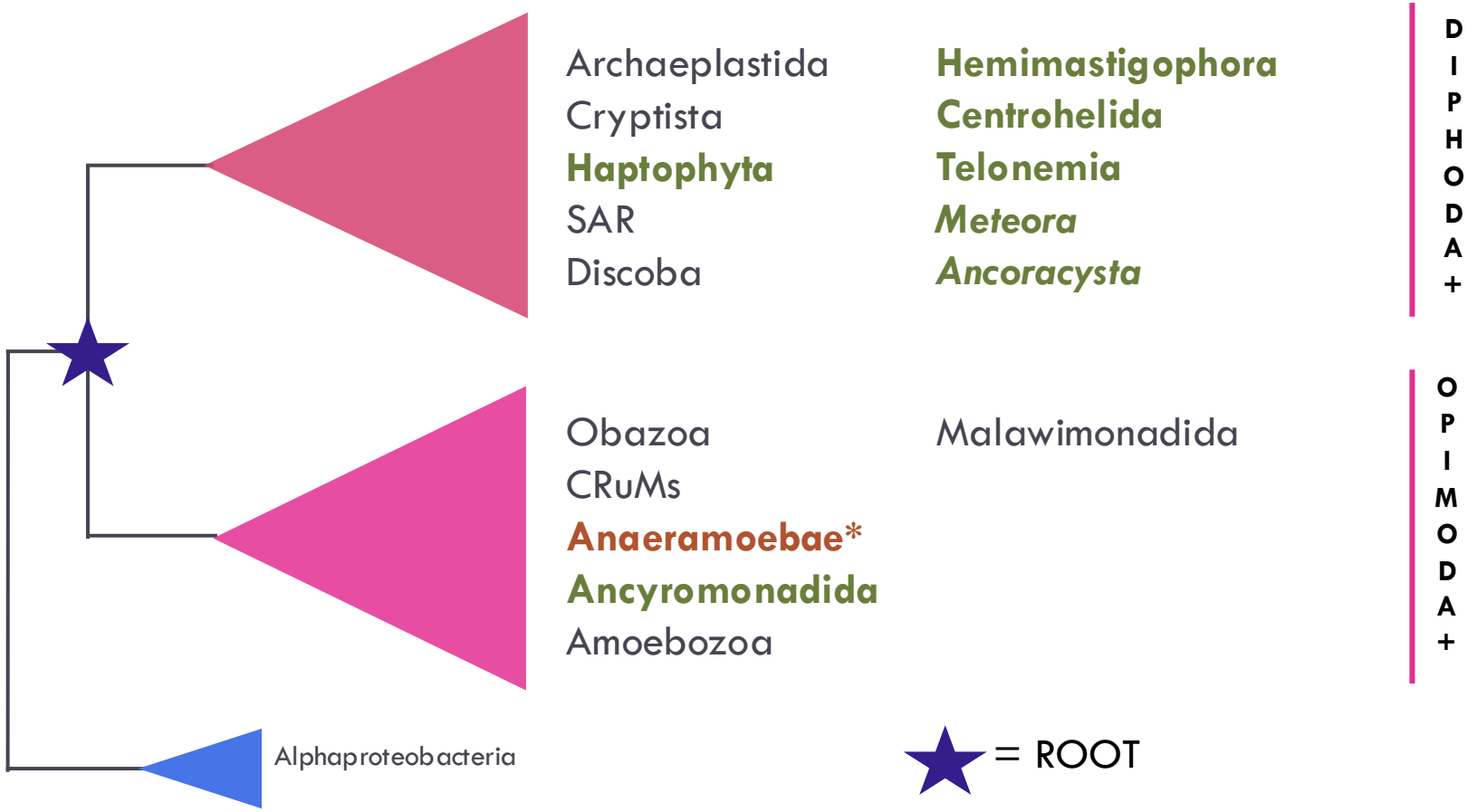
3. Evaluate all root positions under additional complex evolutionary models that account for different phenomena

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+’

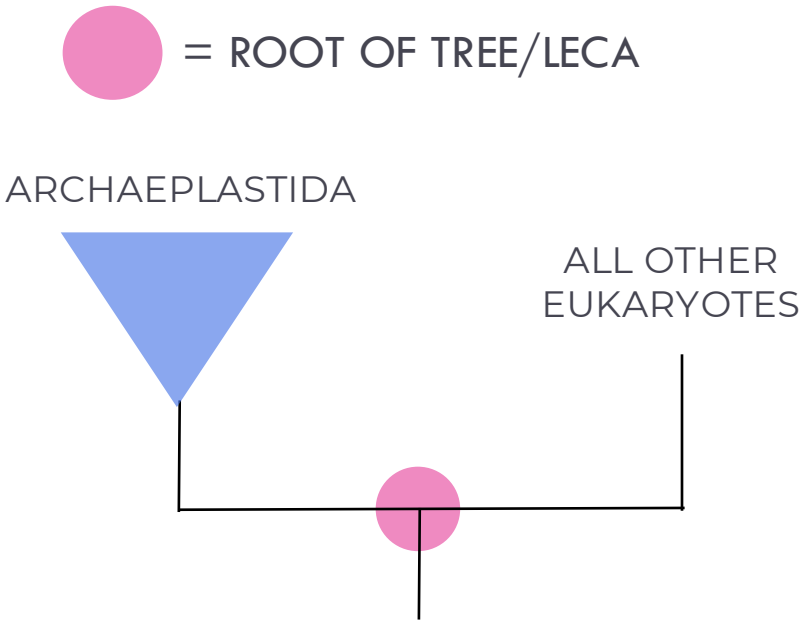


TESTING ALTERNATE ROOT POSITIONS

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)

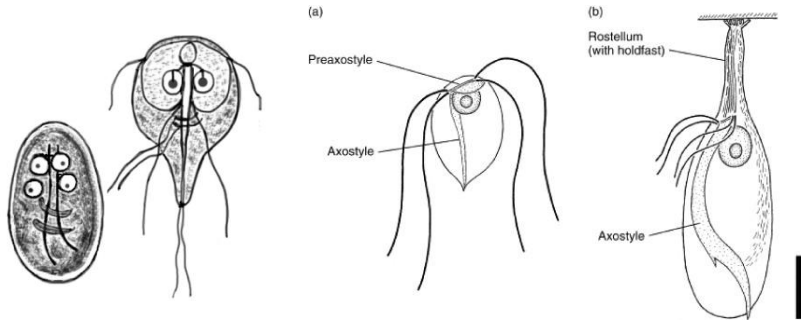
GF-MIX – accommodates **amino acid compositional biases** driven by changing GC content across the tree

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

Metamonads are problematic taxa for datasets of mitochondrial proteins

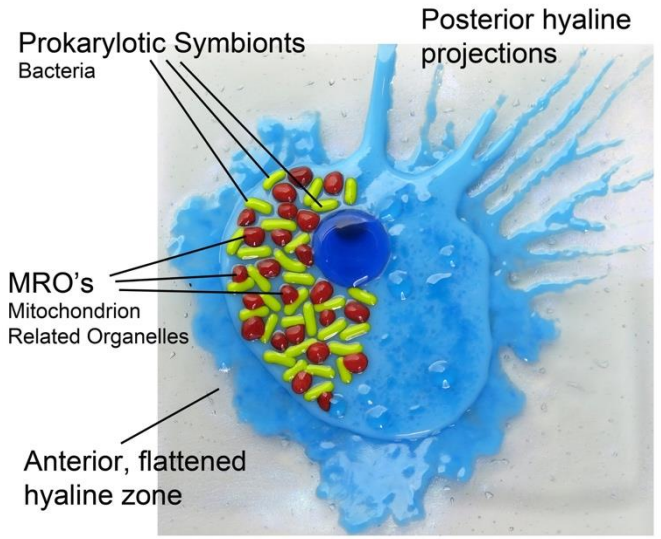


Metamonada consists entirely of **anaerobes:**
lack a mitochondrial genome;
have highly reduced mitochondrion-related organelles
➔ lack most of the proteins in the dataset

Metamonads with the most mitochondrial proteins were included:

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

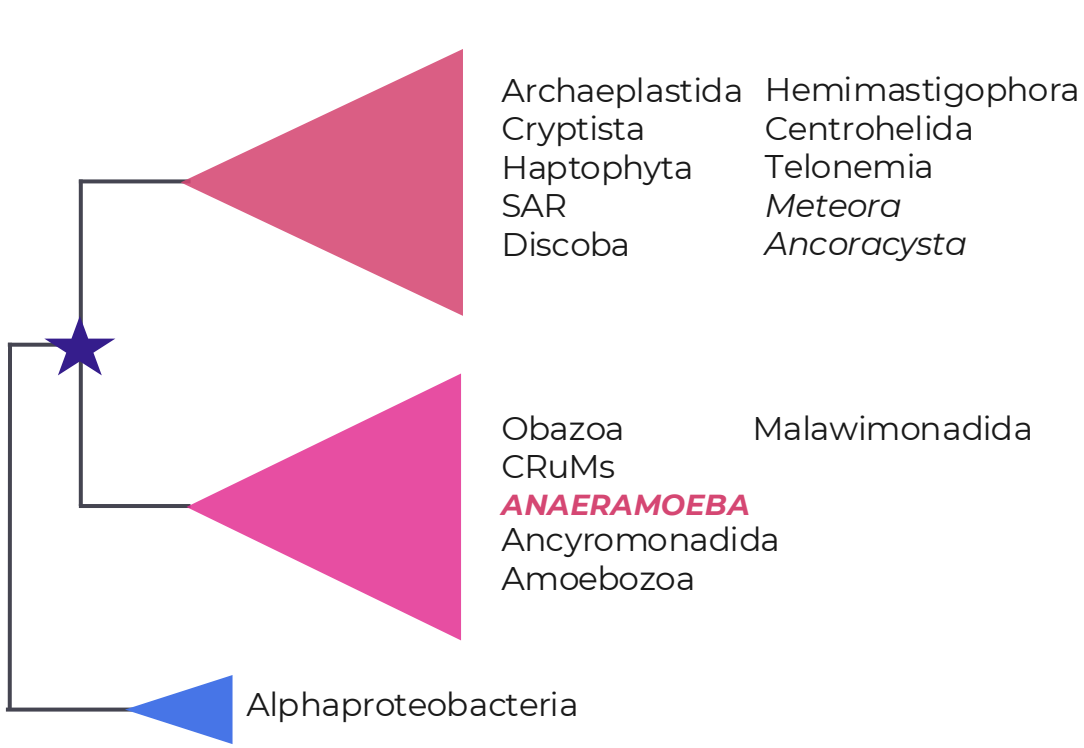
Datasets with and without metamonads were generated:
Anae+ and Anae-

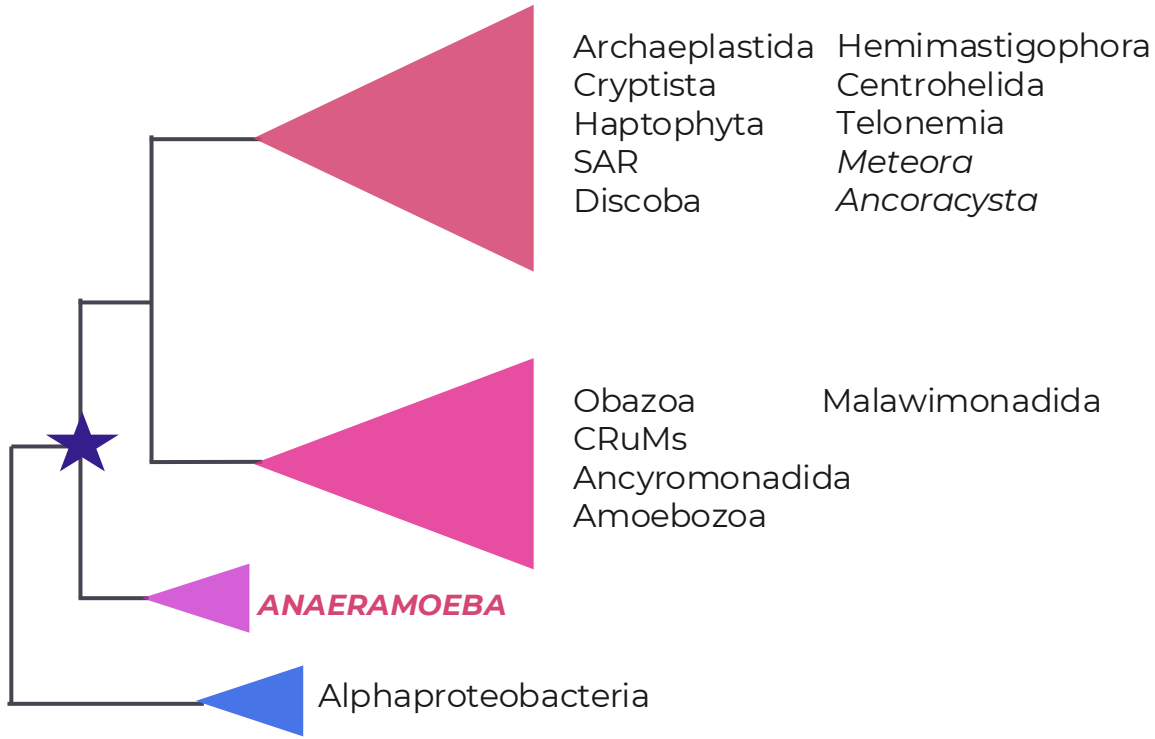


Anaeramoeba fused glass dish
Jane Hartman

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

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



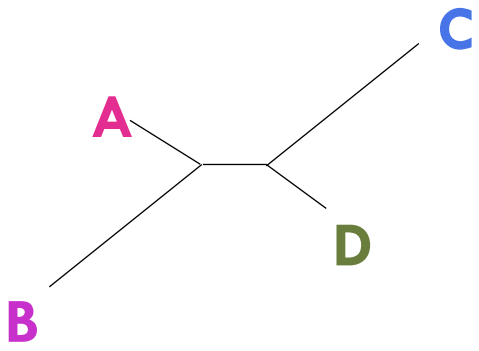


★ = ROOT

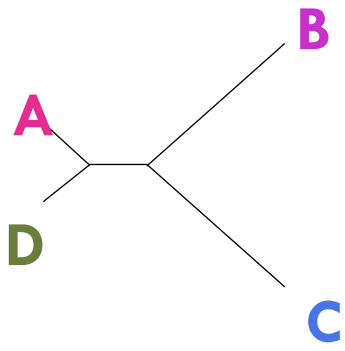
LONG BRANCH ATTRACTION (LBA)

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

TRUE TREE

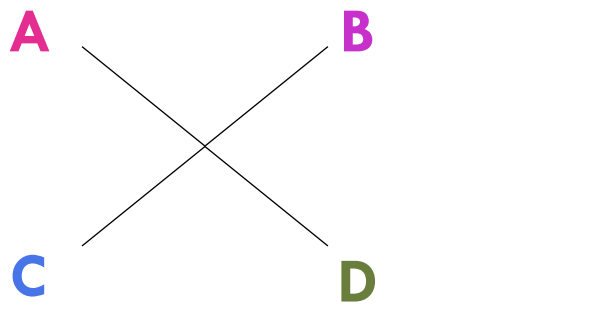


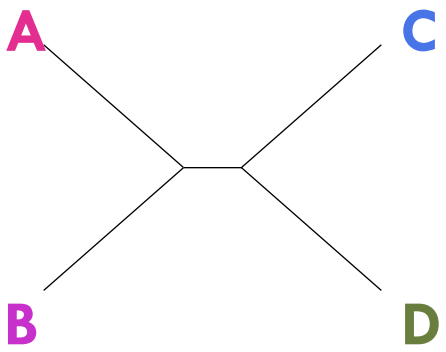
LBA TREE



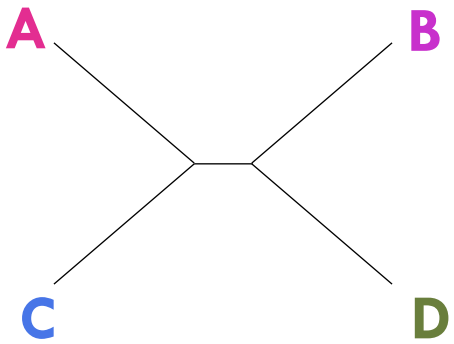
Simulation studies support an artefactual attraction of *Anaeramoeba* to the long branch connecting the outgroup

TRUE TREE

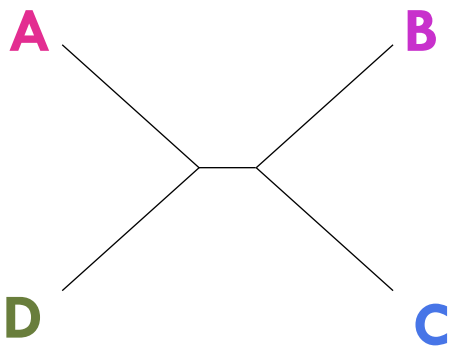




TREE 1



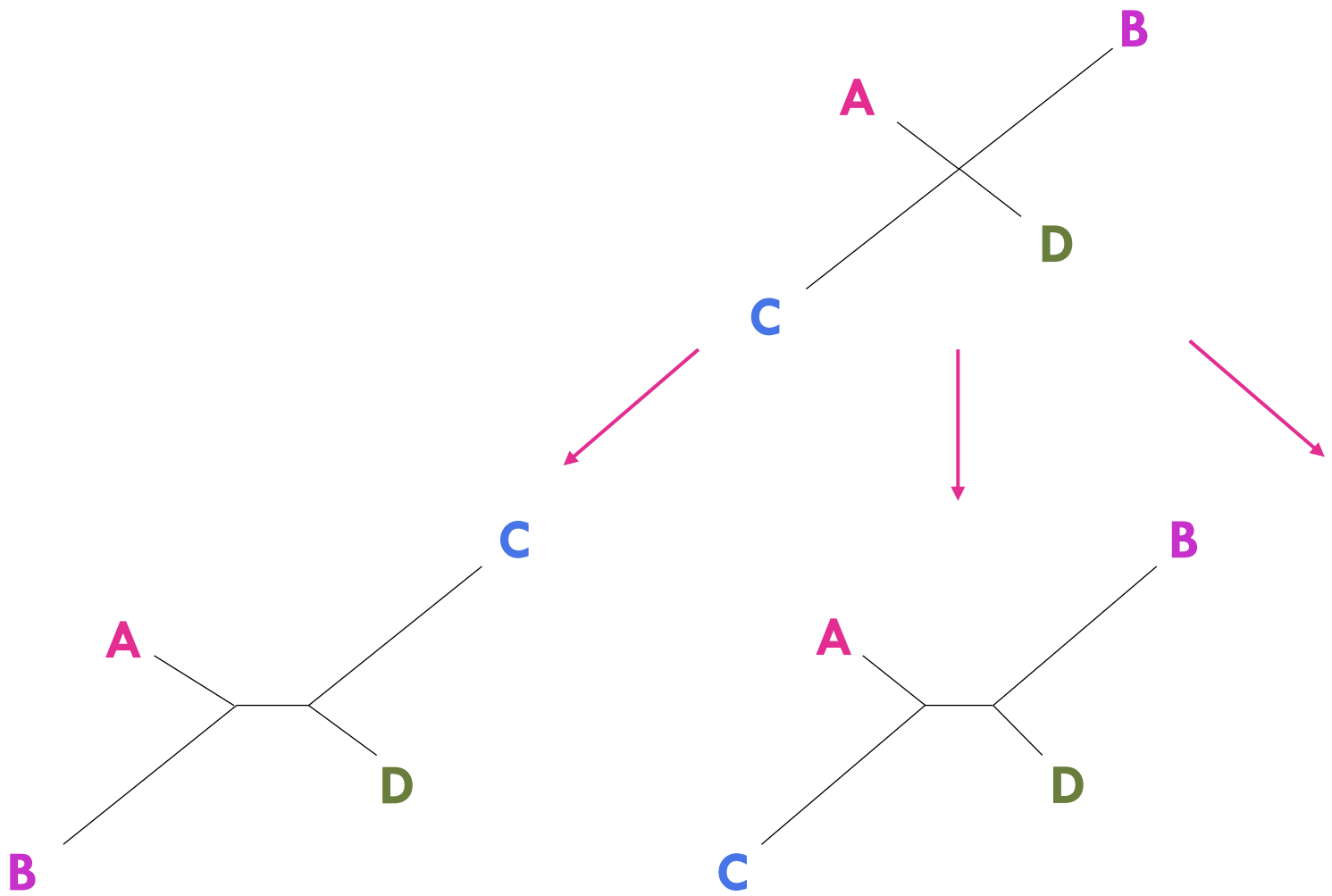
TREE 2



TREE 3

ALL TOPOLOGIES EQUALLY LIKELY

TRUE TREE



TREE 1

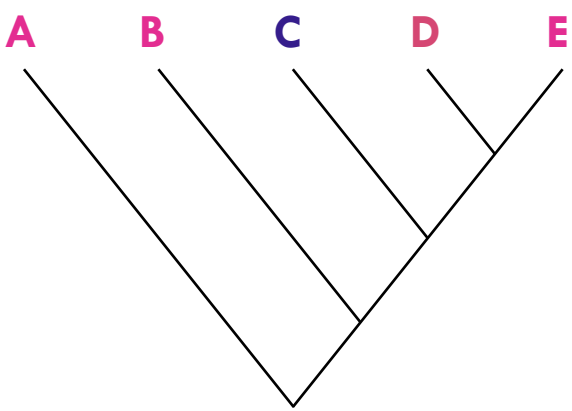
TREE 2

TREE 3

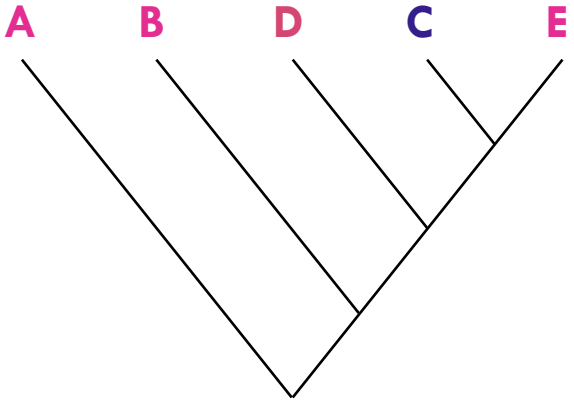
UNDER MODEL MISSPECIFICATION OR SMALL SAMPLE BIAS, LBA TREE IS PREFERRED

TESTING FOR LONG BRANCH ATTRACTION

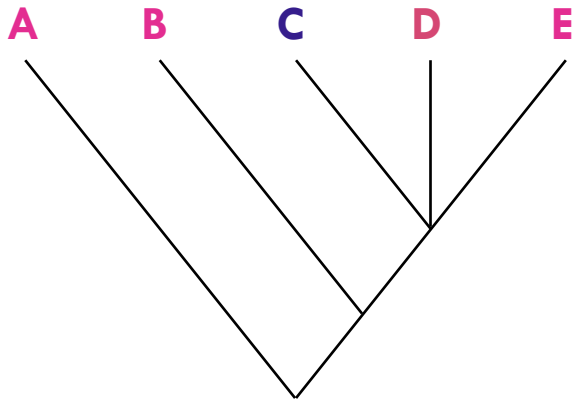
1. Create a consensus tree from two competing topologies



VS.



CONSENSUS:



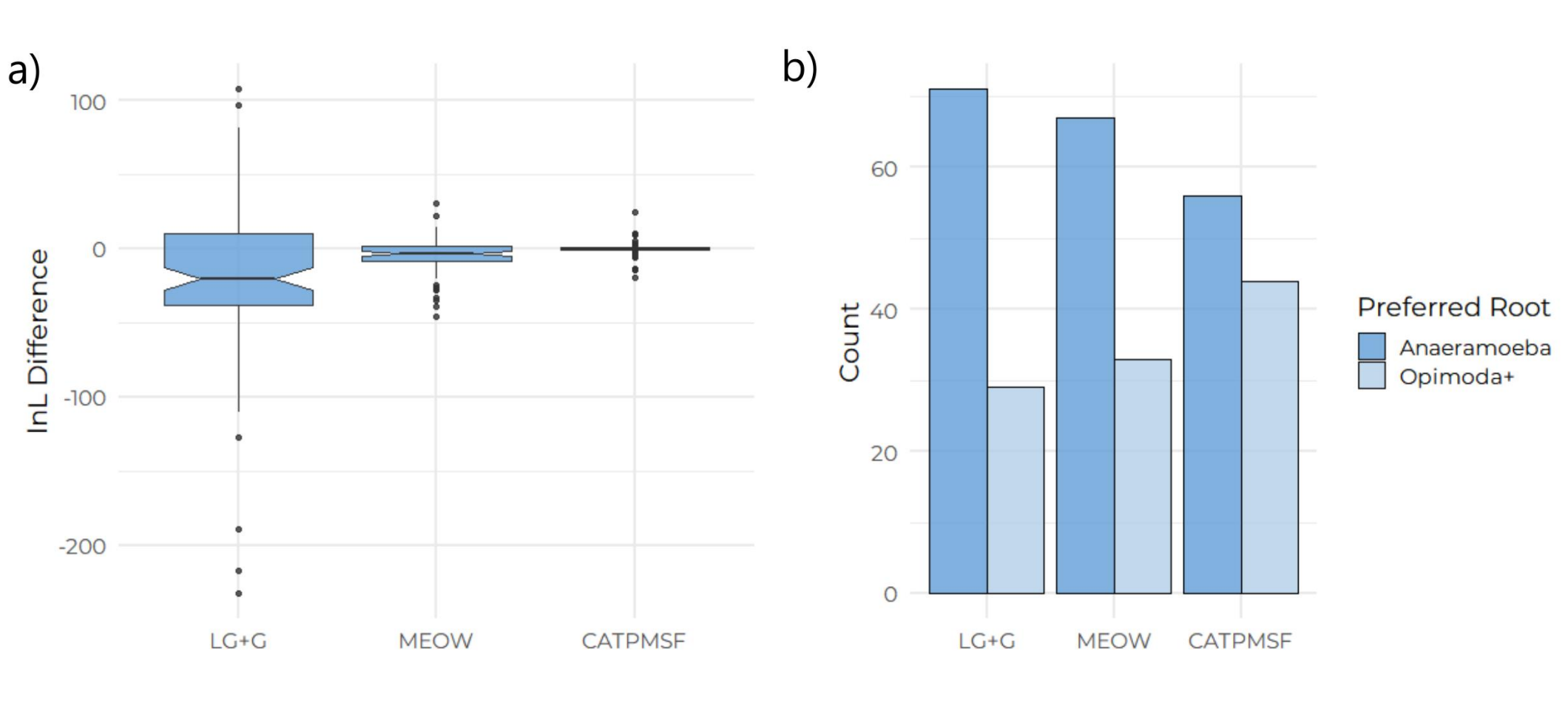
TESTING FOR LONG BRANCH ATTRACTION

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)

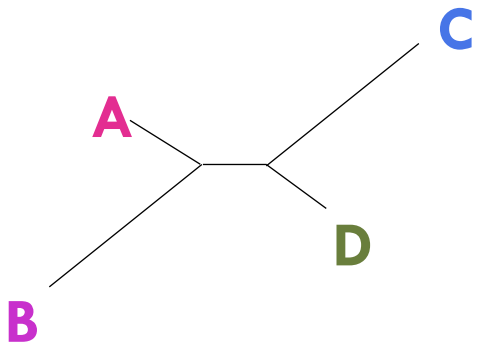
As model misspecification increases, preference for an *Anaeramoeba* root over an Opimoda+ root also increases



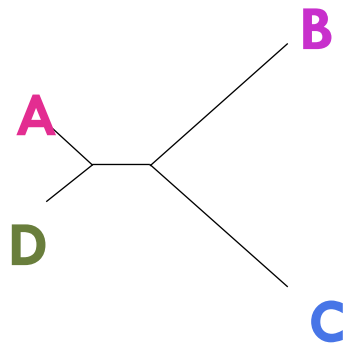
LONG BRANCH ATTRACTION (LBA)

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

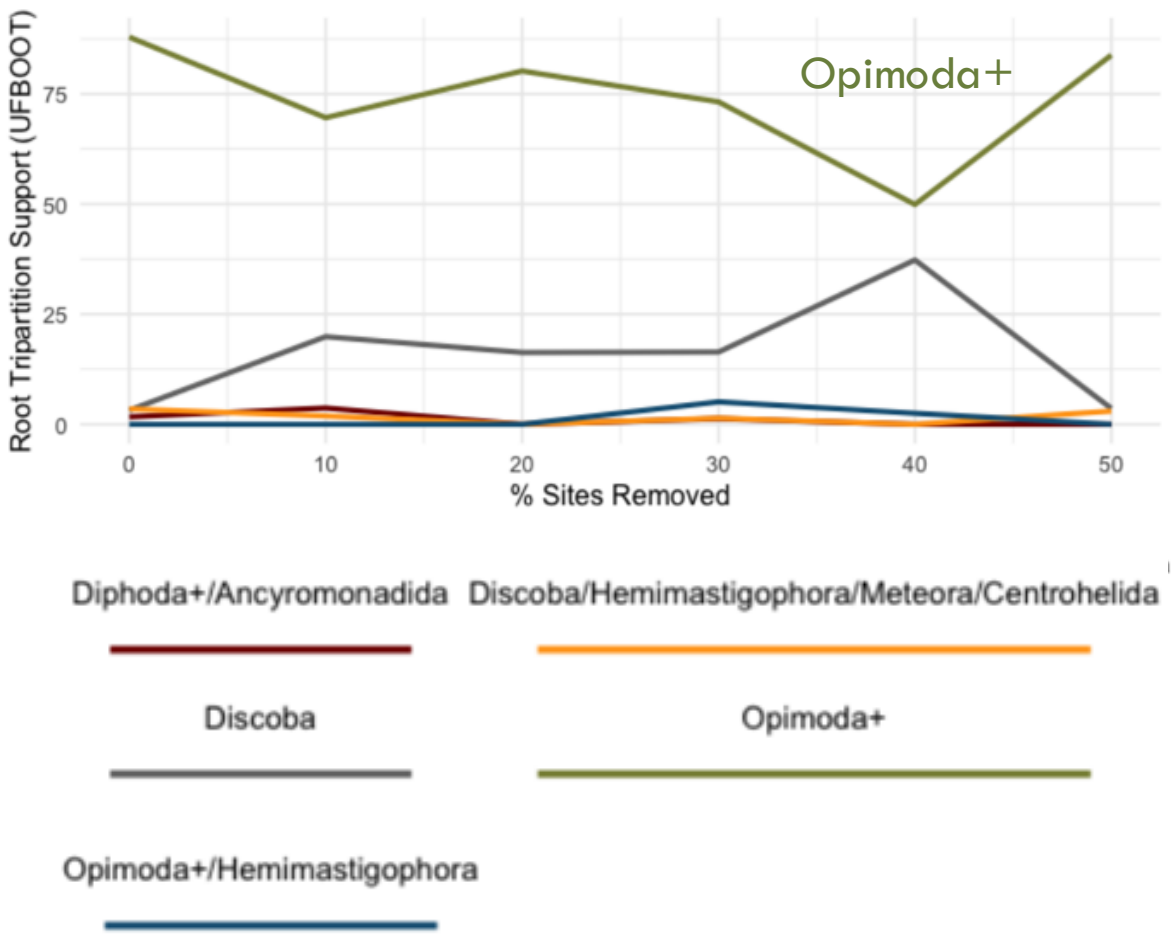
Removing the fastest-evolving sites, genes, and taxa

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

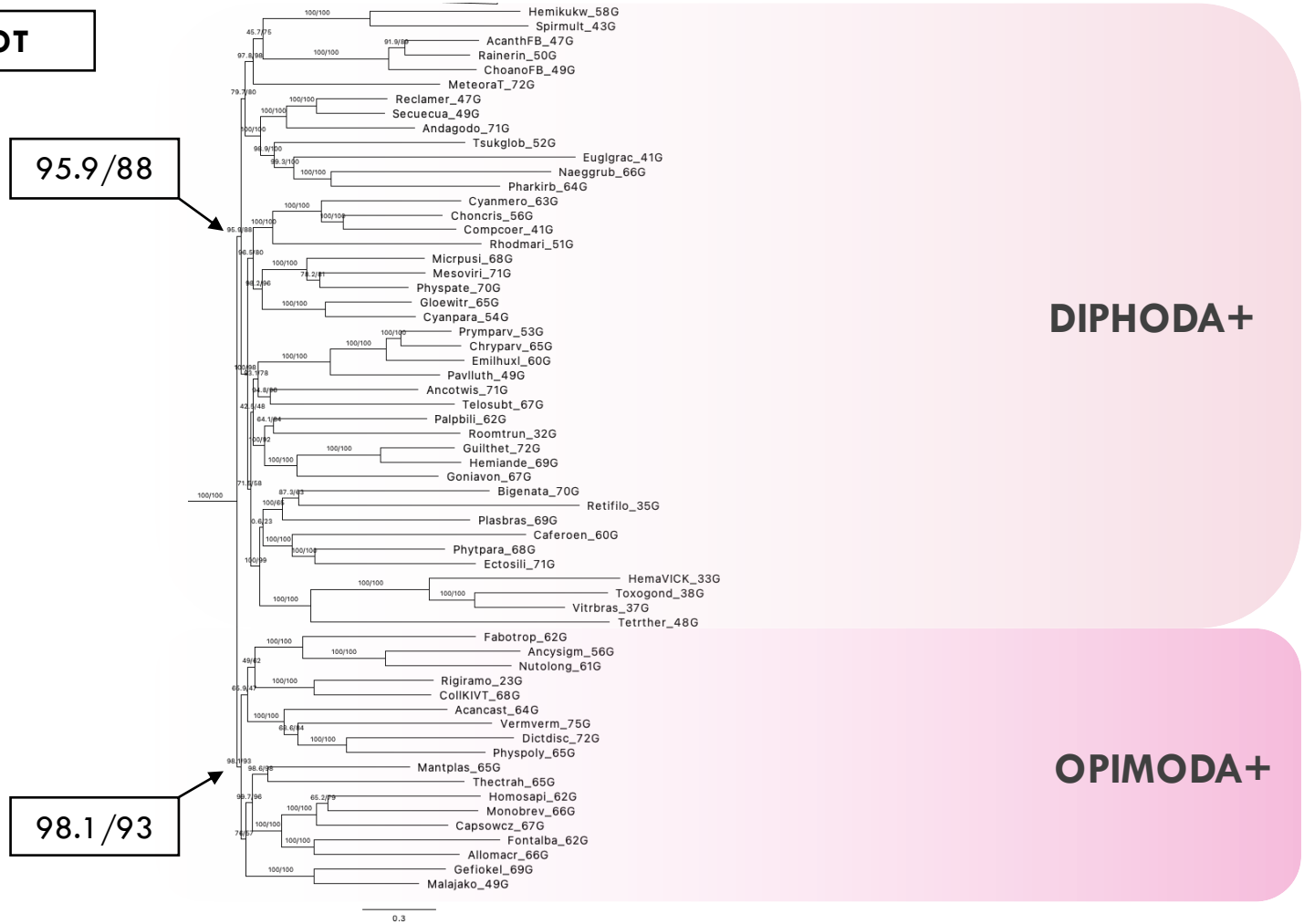


Williamson, Eme, et al. *Nature*, 2025

REMOVAL OF 13 MOST DIVERGENT GENES

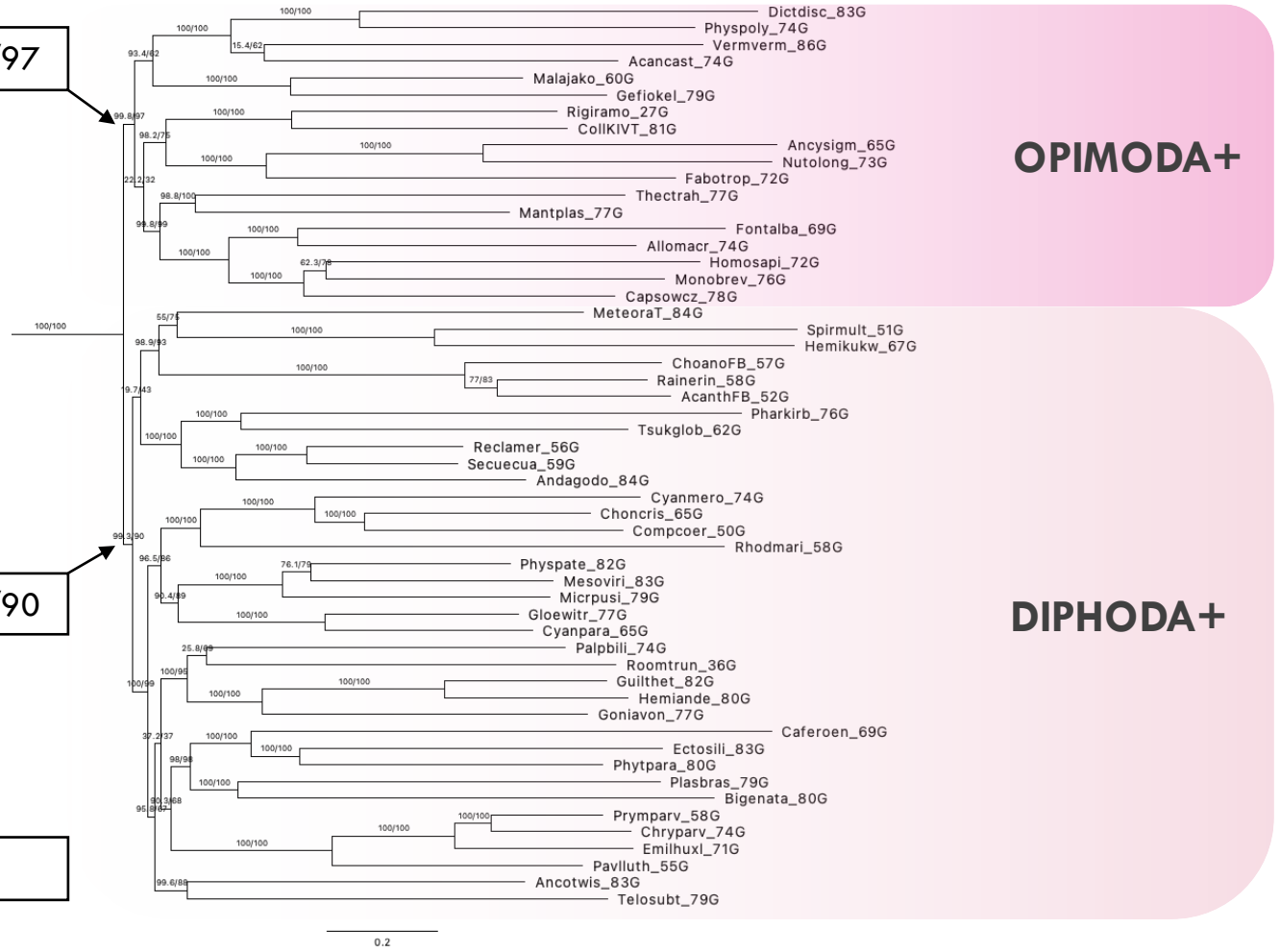
Based on the the **internal branch length** connecting eukaryotes and Alphaproteobacteria

SUPPORT: aLRT/UFBOOT

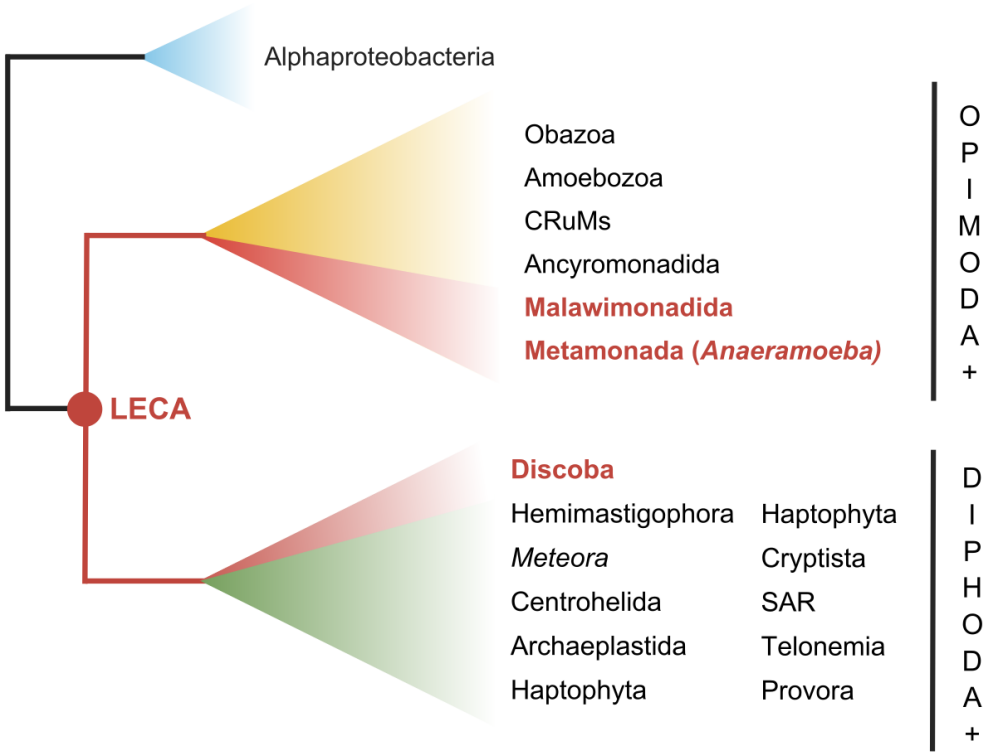


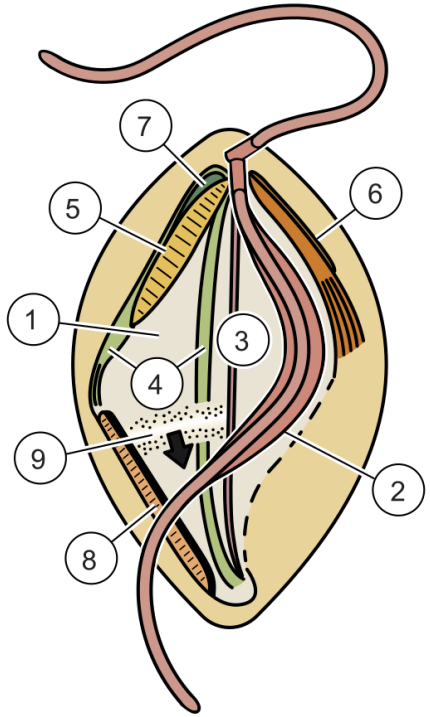
REMOVAL OF 7 FASTEST EVOLVING TAXA

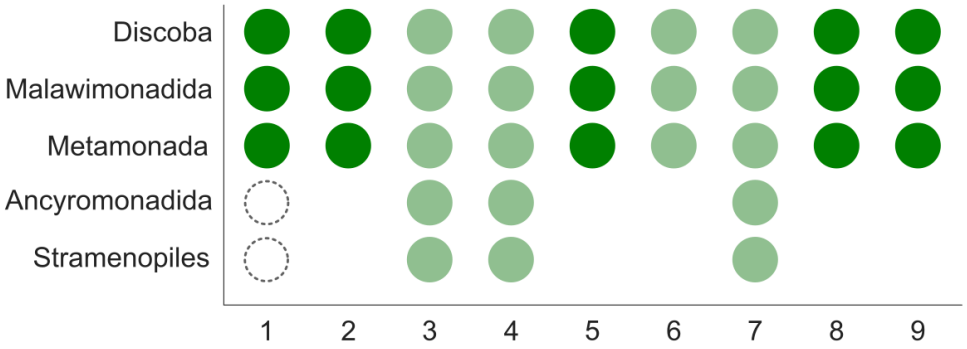
Based on tip-to-tip distance



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

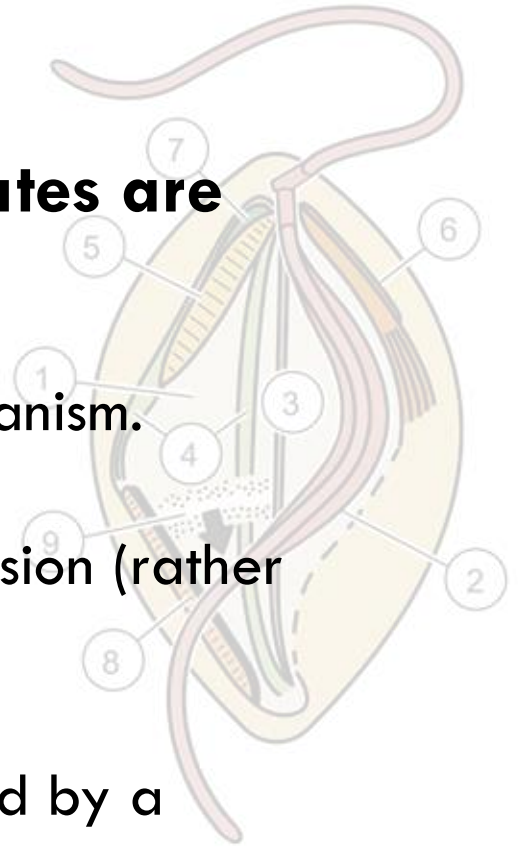
SUMMARY

- **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.

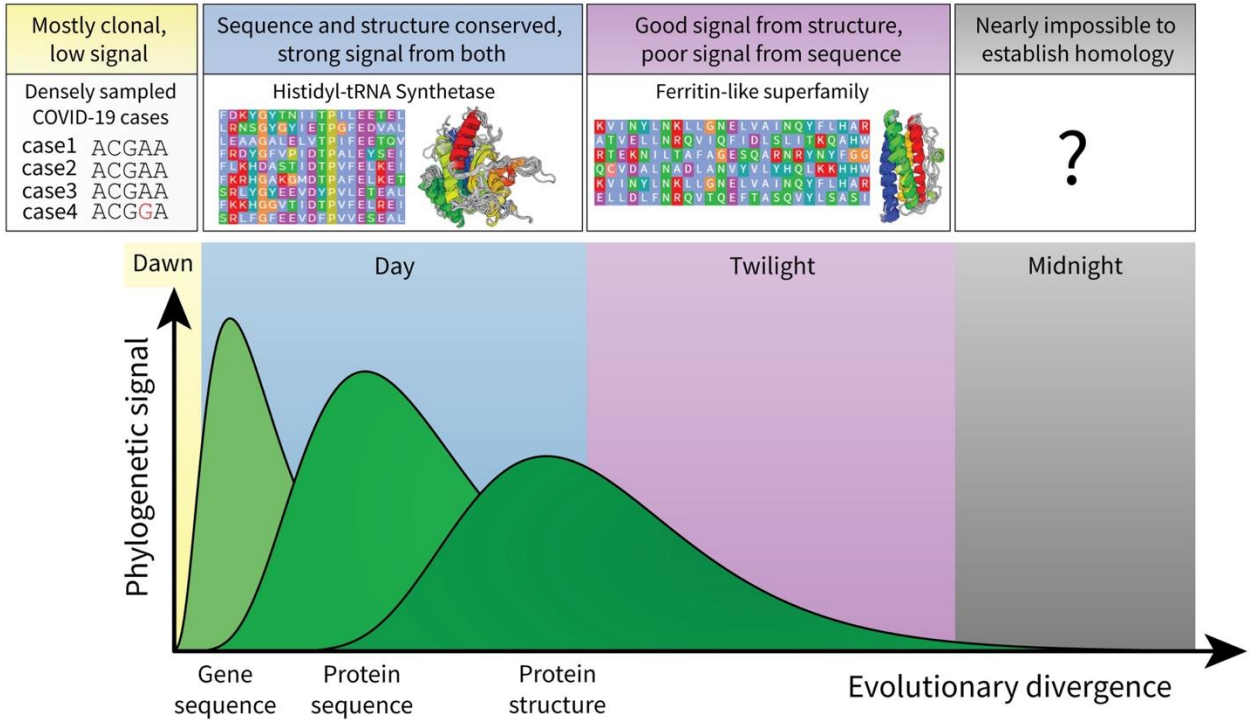
- Phagotrophic (likely bacterivorous), probably feeding on prey in suspension (rather than as a surface feeder like many amoebae).

- 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”.**

PART 3
Structural phylogenetics

Why structural phylogenetics?

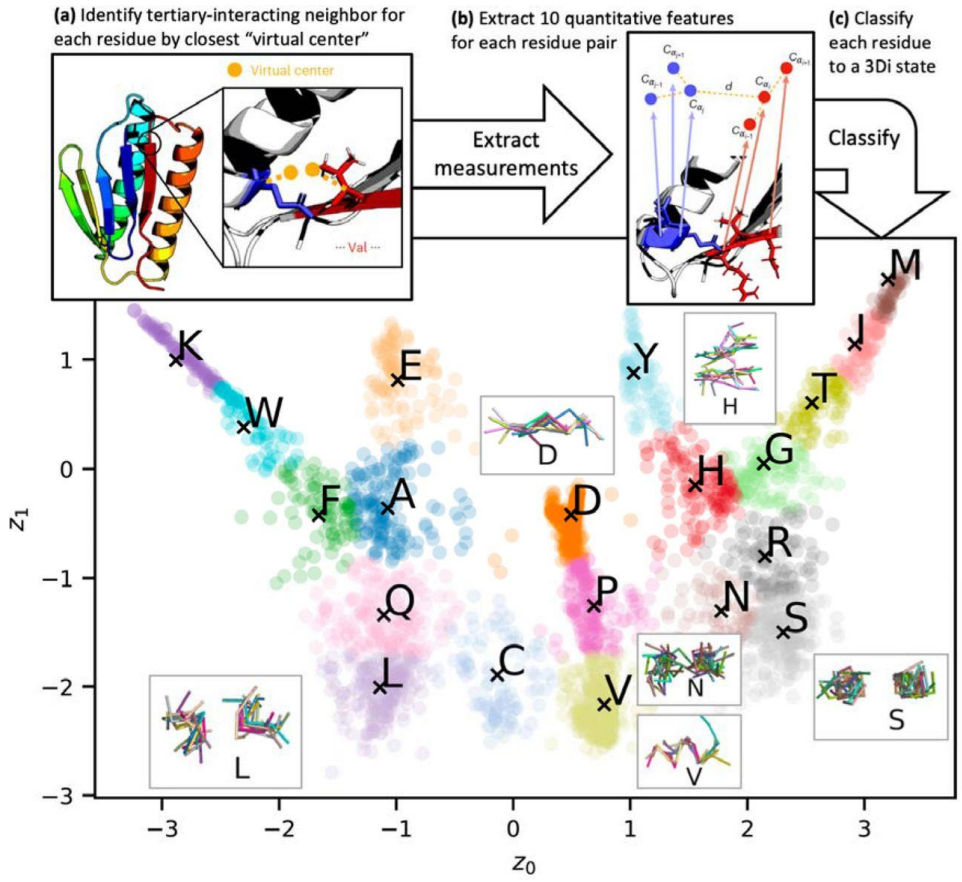


- Structures evolve more slowly than sequences
- Pre-AlphaFold: structural data scarce → mostly limited to a handful of well-studied folds
- AlphaFold2 + AlphaFold3 → predicted structures at scale for nearly all proteins in UniProt + MGnify (~1 billion structures)
- can 3D structure serve as phylogenetic data for the deep relationships we couldn't resolve with sequence alone?

review: Puente-Lelièvre et al. 2025 GBE 17:evaf139

The Foldseek/3Di alphabet

- **Foldseek** encode each residue by its tertiary interaction with the nearest neighboring residue in 3D
- 10 geometric features per residue → classifier → one of **20 "3Di" states**
- Confusingly, the 20 states are labeled A, C, D, ..., W, Y. Same letters as amino acids but **completely unrelated**
- The key trick: a protein of length L becomes a **3Di "sequence" of length L**
- All the standard alignment, ML, and Bayesian machinery *works*
- Partly why structural phylogenetics became practical in the last two years: structure could be made to look like sequence

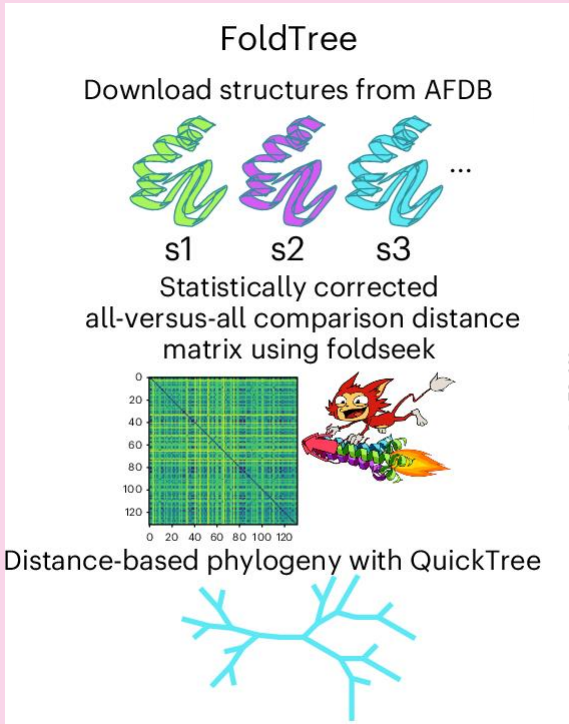


Foldseek: van Kempen et al. 2024 Nat Biotech 42:243; 3Di for phylogeny: Puente-Lelièvre et al. 2024 bioRxiv 571181

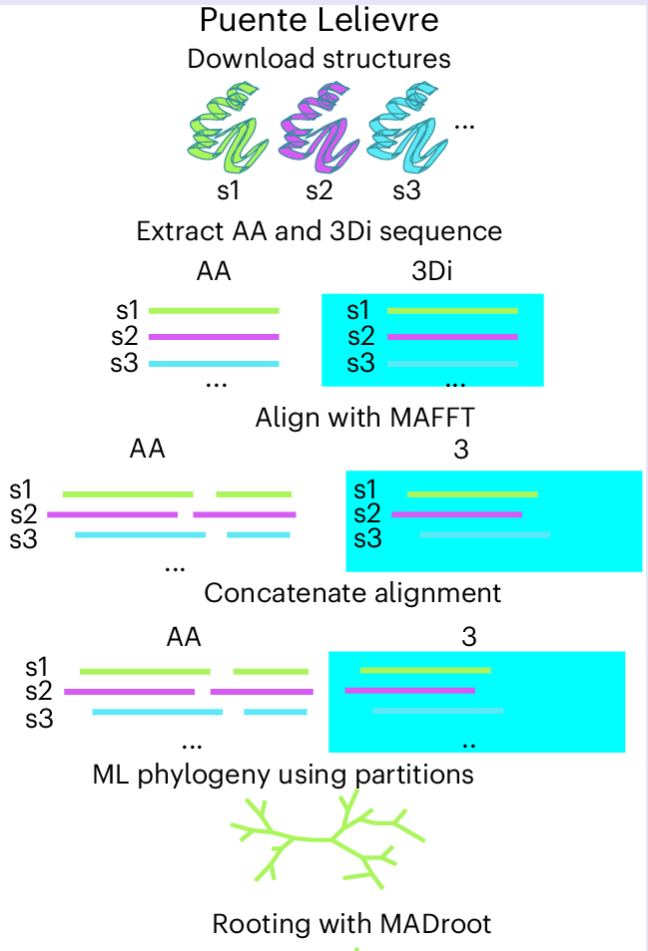
Distance vs ML-based methods

Distance-based: FoldTree (Moi et al. 2025)

- All-vs-all Foldseek comparison → statistically corrected distance matrix
- QuickTree (neighbor joining) + MAD rooting
- Fast, scales to thousands of proteins

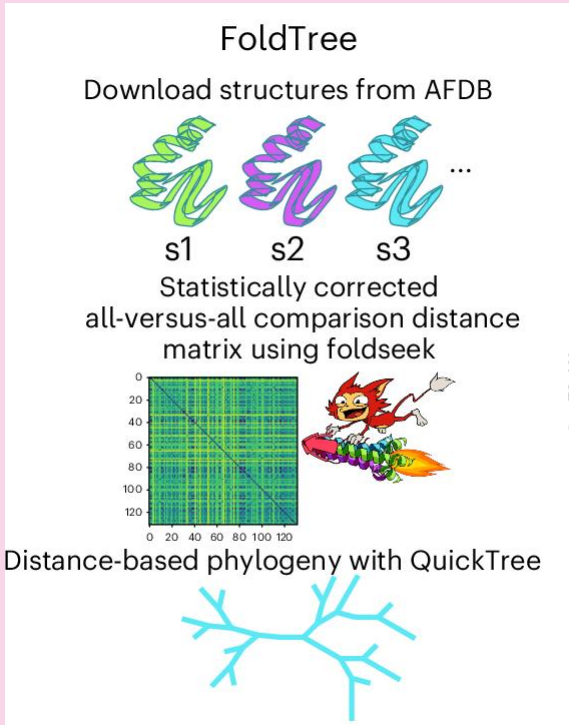


Distance vs ML-based methods



Distance-based: FoldTree (Moi et al. 2025)

- All-vs-all Foldseek comparison → statistically corrected distance matrix
- QuickTree (neighbor joining) + MAD rooting
- Fast, scales to thousands of proteins



ML-based: Partitioned 3Di + AA (Puente-Lelièvre et al. 2024)

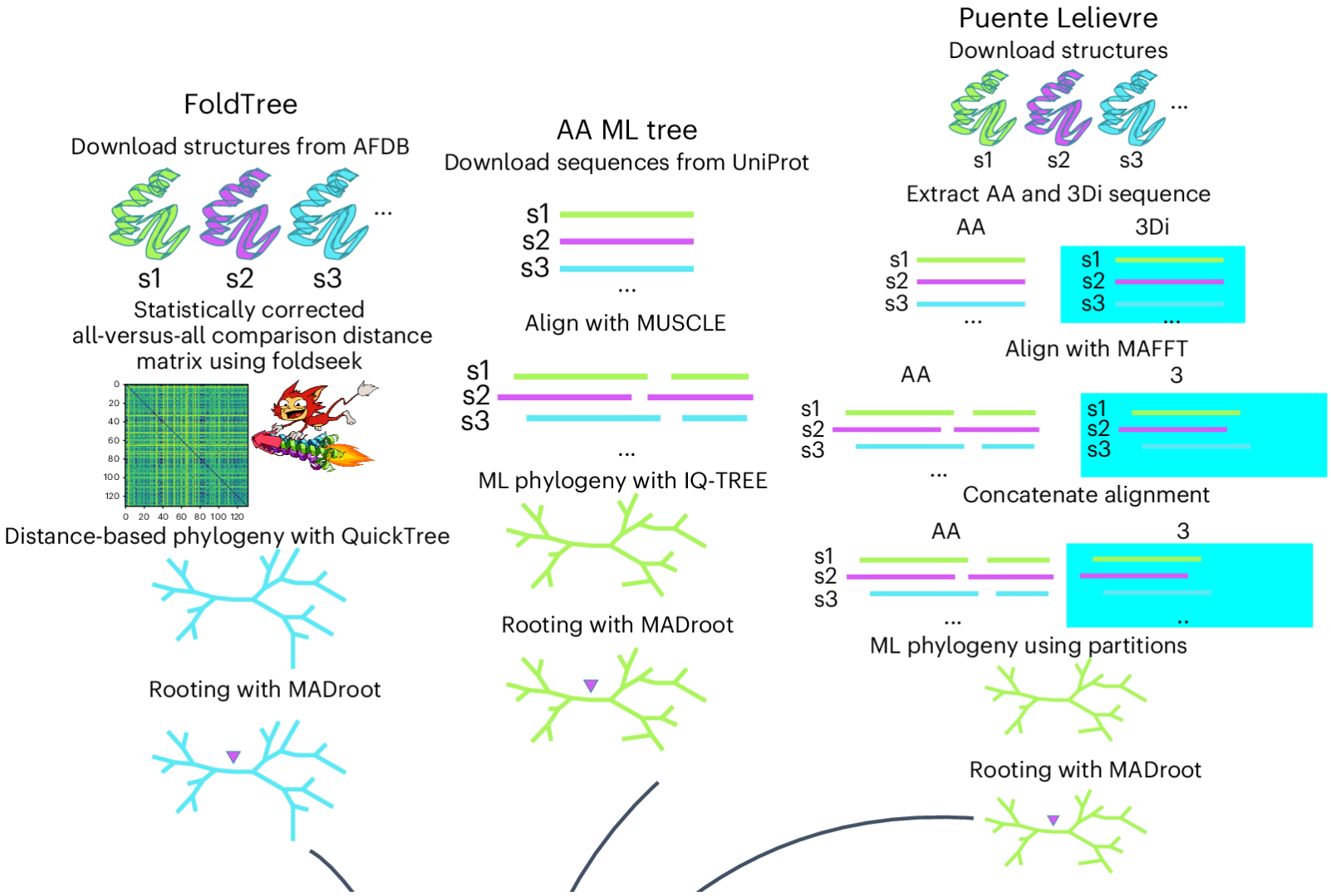
Align 3Di and AA sequences with MAFFT, concatenate with shared gap pattern

Partition model in IQ-TREE: LG (or similar) for AA, 3Di-specific Q matrix for the 3Di partition

Gets you bootstraps, model selection, and all the usual ML infrastructure

Custom 3Di Q matrices: Garg & Hochberg 2025 *MBE*

Distance vs ML-based methods



When does structure help?

- Moi et al. 2025 benchmarked on thousands of families using **taxonomic congruence score** as a proxy for tree accuracy
- **OMA dataset** (~4,600 relatively close families): FoldTree ≈ sequence ML: small advantage
- **CATH dataset** (~500 more divergent families): FoldTree wins by a much larger margin
- As mean pairwise %ID drops, the relative advantage of structure grows
- FoldTree shows the best molecular-clock-like behavior (lowest root-to-tip variance)
- Structure pays off most where sequence is most saturated (i.e., deep phylogeny problems)

Important caveats

- **3Di sites are not independent:** each character is defined relative to its nearest neighbor in 3D space, violating the site-independence assumption that *all* standard ML models rely on. We don't yet know how badly this breaks things.
- **Mutti et al. 2025 (*MBE*):** in a large-scale phylogenomic benchmark, **sequence-based ML still outperforms current structure-based methods** for typical gene-tree reconstruction
- **Tip-level resolution is worse:** 3Di compresses information into 20 states; close relationships at the tips of the tree are less reliable than with AA sequences
- **AlphaFold pLDDT varies:** poorly predicted structures = noisy data; filter or mask low-confidence regions
- **AlphaFold predictions are not ground truth:** they're statistical predictions, can miss conformational states, and quality is uneven for poorly-sampled lineages
- **Best current practice:** filter inputs by pLDDT, use partitioned 3Di + AA together rather than 3Di alone, treat as one source of evidence among many

Concrete impact in coming years

- Asgard ESPs (e.g., actin and tubulin) are exactly the kind of deep homologies where sequence is saturated
- Cryo-EM of AtubA/B confirmed at the structural level what sequence phylogeny only suggested: Asgard tubulin forms eukaryote-like α/β heterodimers and assembles into *bona fide* microtubules
- The next wave: structural phylogenomics applied to ESPs, to the eukaryote root, to deep gene families across the tree of life
- **Where I'd bet:** AA + 3Di partitioned ML, using sophisticated AA mixture models on the AA partition (EAL+C60+PMSF), will be the workhorse for deep questions in the next 2–3 years