Phylogenetic Placement: Computation, Analysis, and Visualization

Lucas Czech
Carnegie Institution for Science
Stanford, USA

2021-05-06
Guest Lecture
University of Oslo

Overview

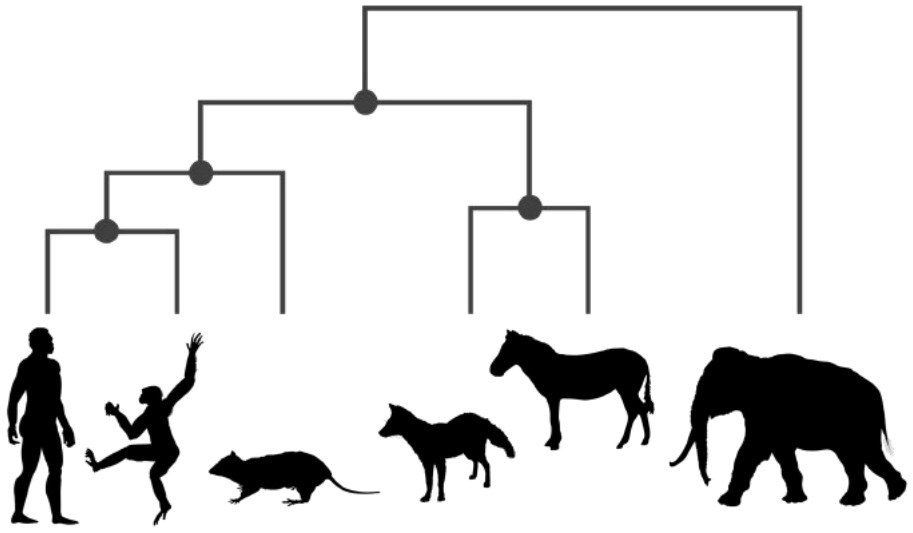
- Maximum Likelihood Tree Inference
- Phylogenetic Placement Computation
- Placement Analysis and Visualization

Tree Inference

MSA and Phylogenetic Tree

Multiple Sequence Alignment (MSA)

Human	C	A	A	A	T	C	C	A	C	A	T	A	C	A	A
Chimp	C	A	C	A	C	C	C	A	A	A	C	A	A	A	C
Mouse	C	C	T	A	C	C	A	A	C	T	C	C	C	A	T
Dog	C	A	C	A	T	C	C	A	A	A	C	G	A	A	C
Horse	C	A	C	A	T	G	C	A	C	G	C	G	C	A	C
Elephant	C	C	T	A	C	C	C	A	A	T	T	C	A	A	T



Maximum Likelihood Tree Inference

Find the phylogenetic tree which maximizes the likelihood that it produces the given MSA

$$\mathcal{L}(\ \mathrm{MSA} \mid T, \bar{b}, M, \bar{\theta}\ )$$

with

- Tree T
- Branch lengths $\bar{b}$
- Model of evolution M
- Model parameters θ

Note that this is the reverse of the intuitive computation!

Tree Search

- Basic strategy: Try out trees until we find a good one
- Optimize:
  - Tree topology itself
  - Branch lengths
  - Model parameters

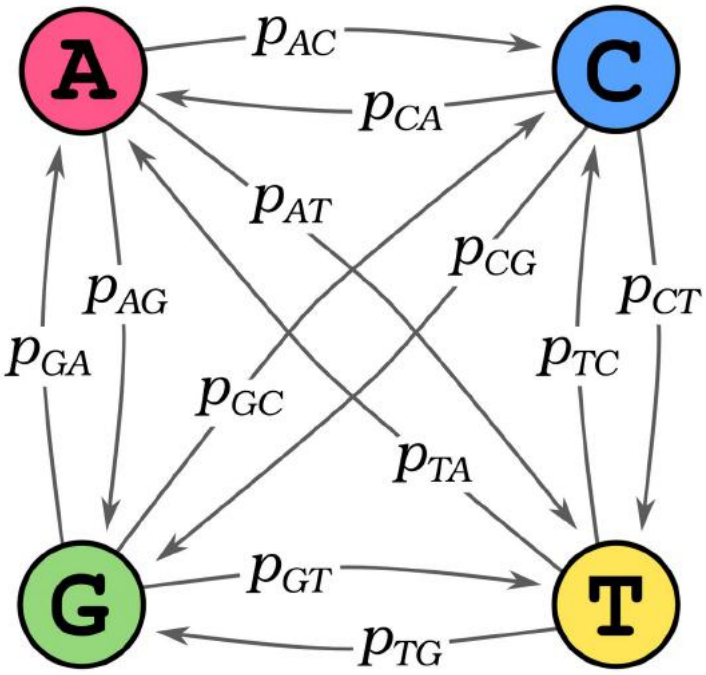
  → Computationally expensive!
- Many heuristics and methods developed over the years
  - Greedy hill-climbing from a (reasonable) starting tree
  - Felsenstein Pruning Algorithm
  - ...

Model of Nucleotide Substitution

- Assumption: Columns/sites of the MSA evolved independently!
- Assumption: (Evolutionary) time is reversible!
- How did the sequences evolve?
  - Need a model to estimate of the evolutionary distance between sequences
  - As we assume homologous traits (columns of the MSA), we only consider mutations (no insertions or deletions, although gaps are possible in the MSA, but ignored)
  - We use a continuous-time Markov chain (MC) model

Model of Nucleotide Substitution

- States of the Markov chain are the 4 nucleotides
- Transition probabilities p allow changes between states
- They depend on the evolutionary time t between the sequences, using evolutionary rate r and branch length b
- t = r * b
- Rate r can differ between sites, and typically is modeled via an additional distribution



Likelihood Computation

- Assume:
  - Given the MSA
  - Fixed (given) tree topology T
  - Fix branch lengths $\bar{b}$, fixed evolutionary rate r
  - Model of sequence evolution M with parameters θ
- Compute: $\mathcal{L}(\ \mathrm{MSA} \mid T, \bar{b}, M, \bar{\theta}\ )$
- Account for unknown states at inner nodes of the tree

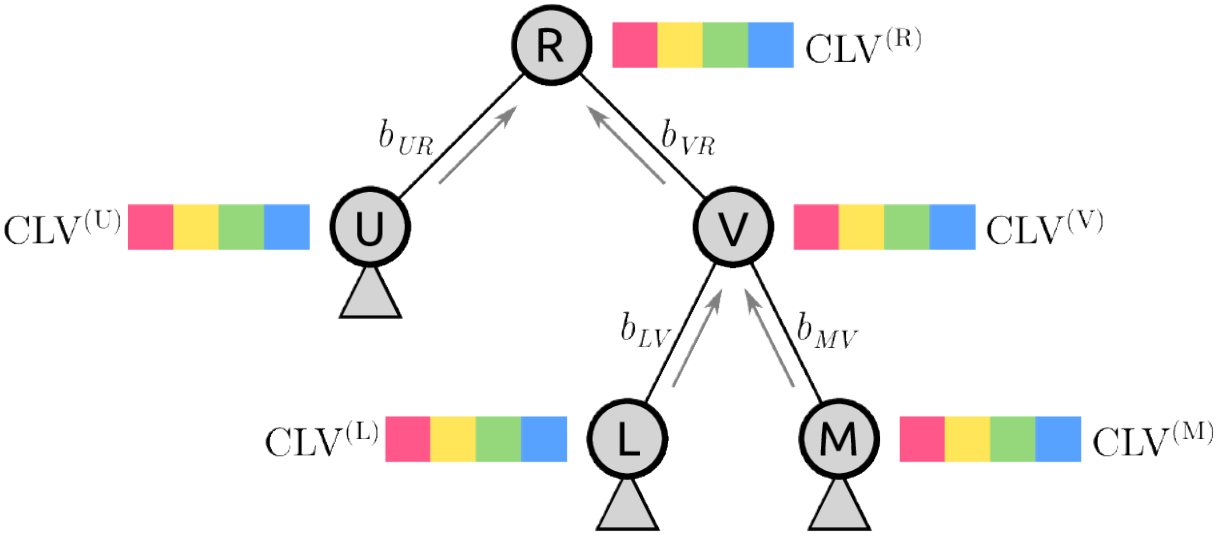
  → Sum over probabilities of every possible states

  → Felsenstein pruning algorithm

Felsenstein Pruning Algorithm

- At each node, compute a *conditional likelihood vector* (CLV)
- The CLV “summarizes” the subtree below its node:
  - For each site and each state (ACGT), it gives the *conditional likelihood* that this site is in that state at the node
  - Conditional on: subtree topology and branch lengths
- For tree tips (leaves), the state is simply the observed nucleotide of the sequence (e.g., for G: 0,0,1,0)
- We work from the tips of the tree inwards, called post-order traversal of the tree

Felsenstein Pruning Algorithm



$$\mathrm{CLV}_{s,c}^{(V)} = \left(\sum_{j \in N} p_{cj}(r \cdot b_{LV}) \cdot \mathrm{CLV}_{s,j}^{(L)} \right) \left(\sum_{k \in N} p_{ck}(r \cdot b_{MV}) \cdot \mathrm{CLV}_{s,k}^{(M)} \right)$$

s	alignment site	p	probability of state transition
c	state ∈ N (out of 4 nucleobases)	r*b = t	time between two nodes

Felsenstein Pruning Algorithm

$$\mathrm{CLV}_{s,c}^{(V)} = \left(\sum_{j \in N} p_{cj}(r \cdot b_{LV}) \cdot \mathrm{CLV}_{s,j}^{(L)} \right) \left(\sum_{k \in N} p_{ck}(r \cdot b_{MV}) \cdot \mathrm{CLV}_{s,k}^{(M)} \right)$$

s	alignment site	p	probability of state transition
c	state ∈ N (out of 4 nucleobases)	r*b = t	time between two nodes

- Inner product p * r * b * CLV: change from state c to state j
- Sum over all j in {ACGT}: Account for all possible inner states
- Product of these sums: conditional likelihood of node V being in state c at site s, given its two subtrees
- Repeat for all states c and all sites s

Likelihood Computation

- Due to our assumption of time reversibility,
  it does not matter which node we use as root,
  (or can place a *virtual root* node wherever we want)
- Compute all CLVs up to that root node
- Use base frequencies π (they are part of our probabilities
  in the Markov model) to compute likelihood for site s:

$$\mathcal{L}_s = \sum_{i \in N} \pi_i \cdot \mathrm{CLV}_{s,i}^{(R)}$$

- Due to assumption of independent sites, the total likelihood is:

$$\mathcal{L} = \prod_{s=1}^{m} \mathcal{L}_s$$

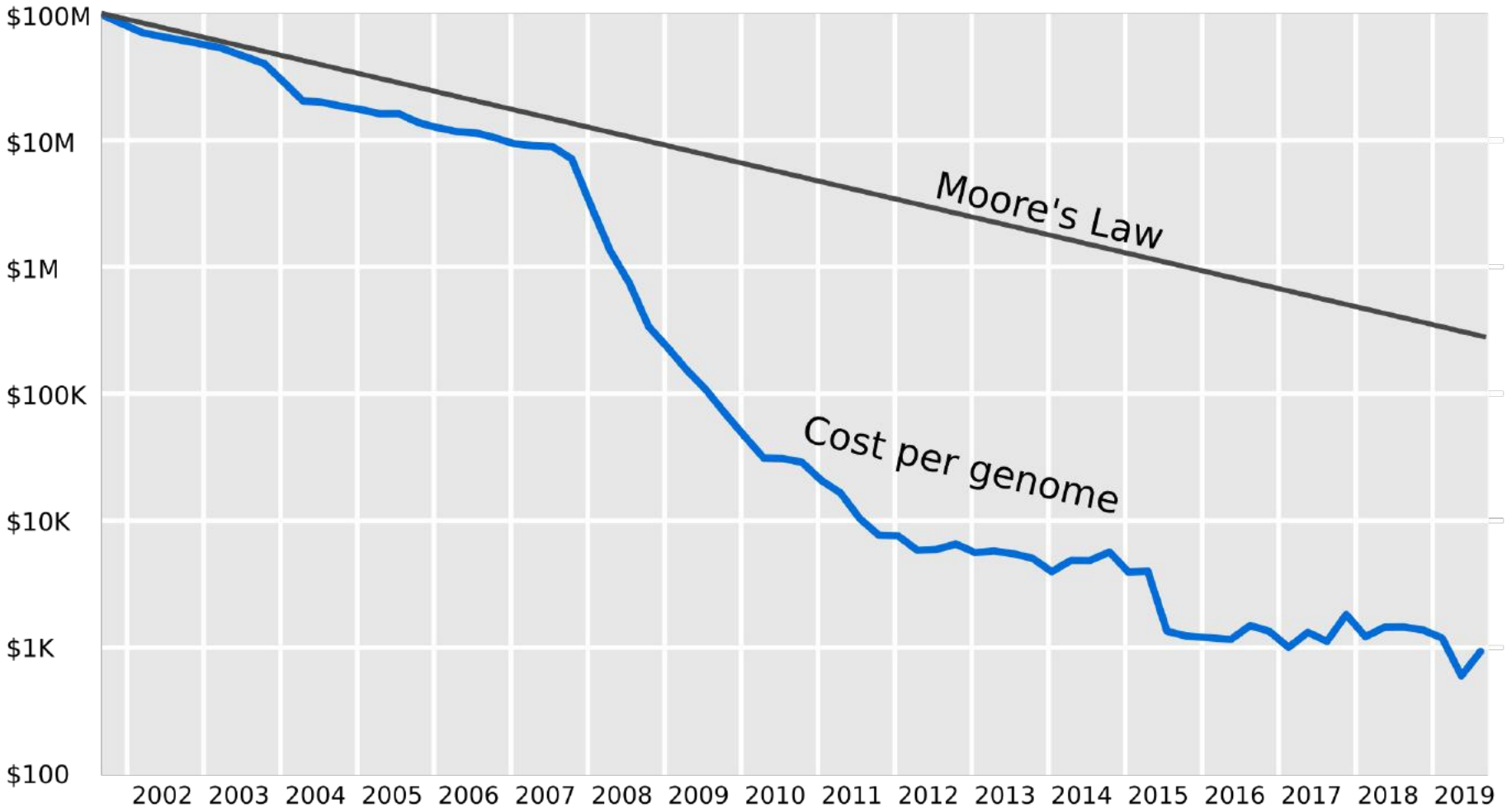
Likelihood Computation

- We now have the likelihood of a given tree

$$\mathcal{L}(\ \mathrm{MSA} \mid T, \bar{b}, M, \bar{\theta}\ )$$

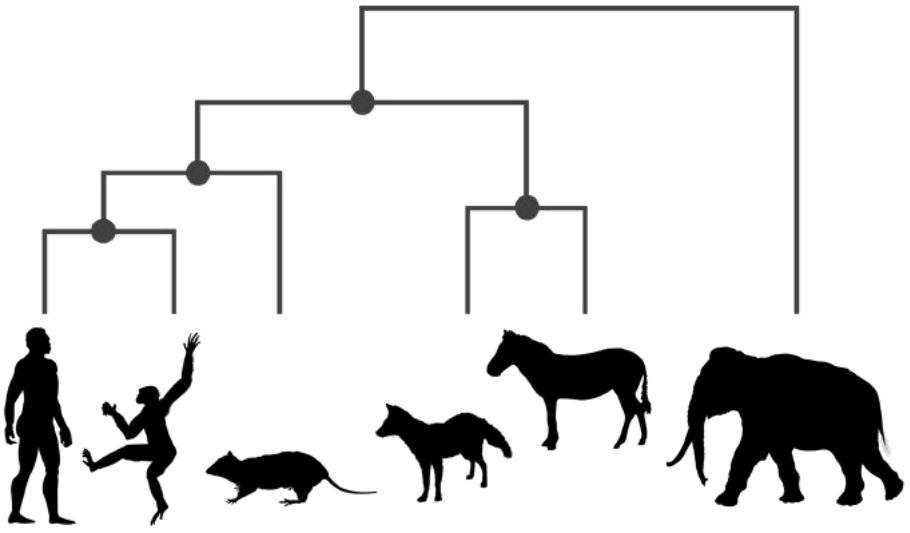
- Still need to optimize branch lengths → numerical method!
- Then, “simply” repeat for every possible tree topology to find the most likely tree :-)
- But: Number of possible trees grows over-exponentially with number of taxa! :-(

Phylogenetic Placement


$100M
$10M
$1M
$100K
$10K
$1K
$100
Moore's Law
Cost per genome
2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019

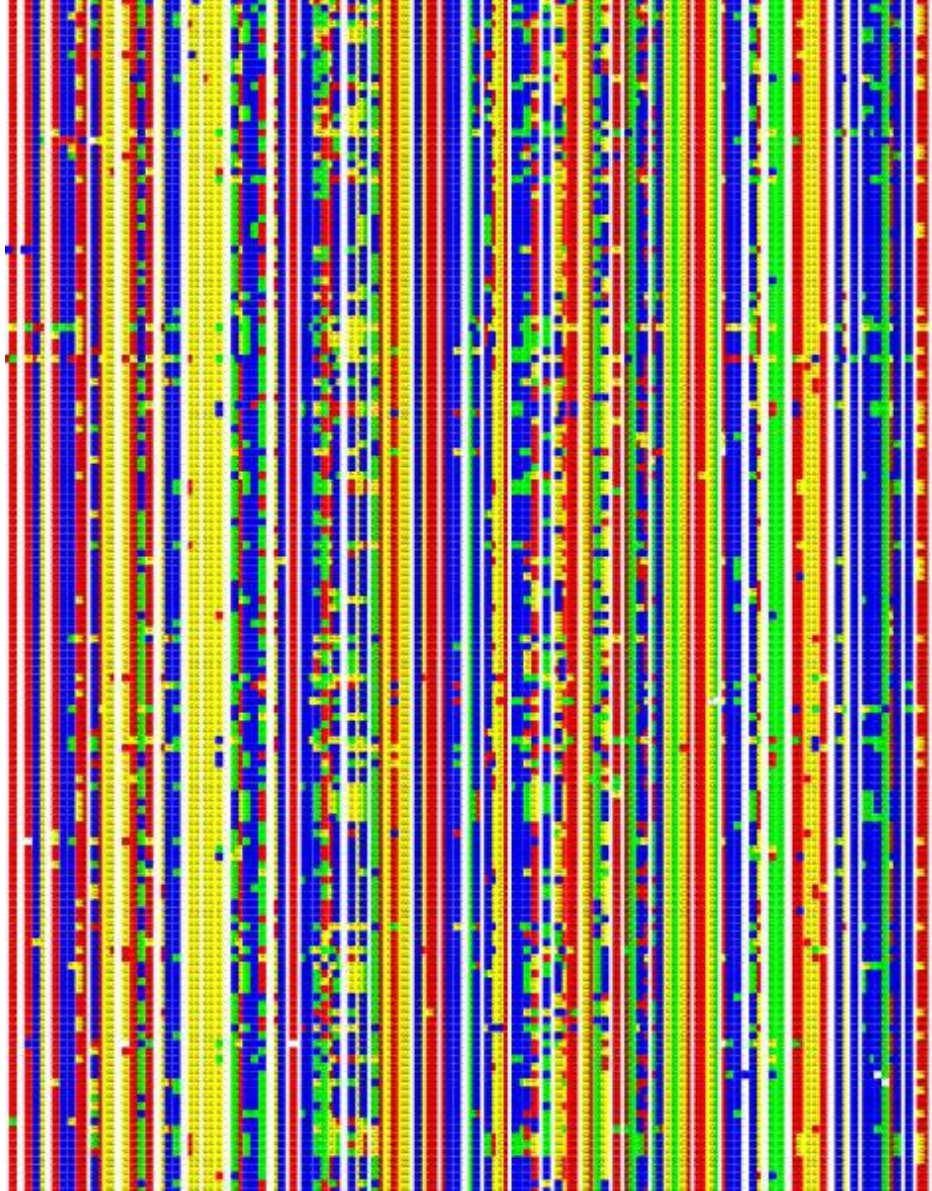
Multiple Sequence Alignment (MSA)

Human	C	A	A	A	T	C	C	A	C	A	T	A	C	A	A
Chimp	C	A	C	A	C	C	C	A	A	A	C	A	A	A	C
Mouse	C	C	T	A	C	C	A	A	C	T	C	C	C	A	T
Dog	C	A	C	A	T	C	C	A	A	A	C	G	A	A	C
Horse	C	A	C	A	T	G	C	A	C	G	C	G	C	A	C
Elephant	C	C	T	A	C	C	C	A	A	T	T	C	A	A	T

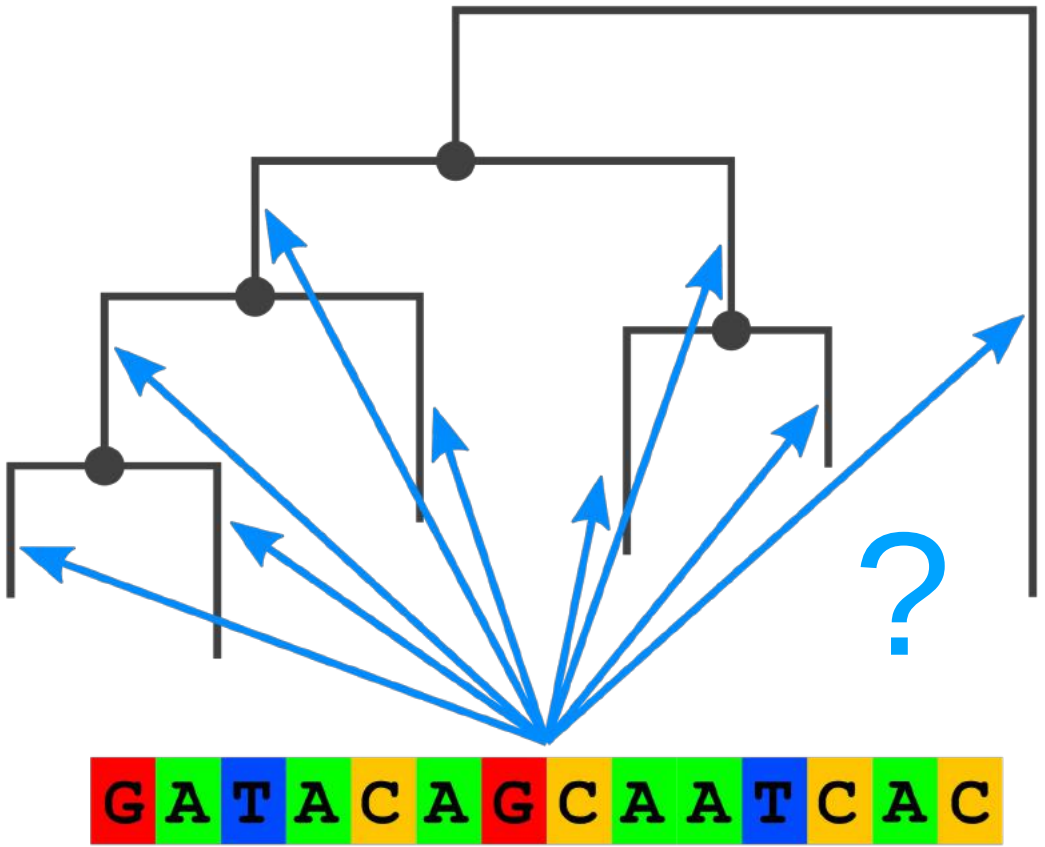


Multiple Sequence Alignment (MSA)

Human	C	A	A	A	T	C	C	A	C	A	T	A	C	A	A
Chimp	C	A	C	A	C	C	C	A	A	A	C	A	A	A	C
Mouse	C	C	T	A	C	C	A	A	C	T	C	C	C	A	T
Dog	C	A	C	A	T	C	C	A	A	A	C	G	A	A	C
Horse	C	A	C	A	T	G	C	A	C	G	C	G	C	A	C
Elephant	C	C	T	A	C	C	C	A	A	T	T	C	A	A	T



Phylogenetic Placement



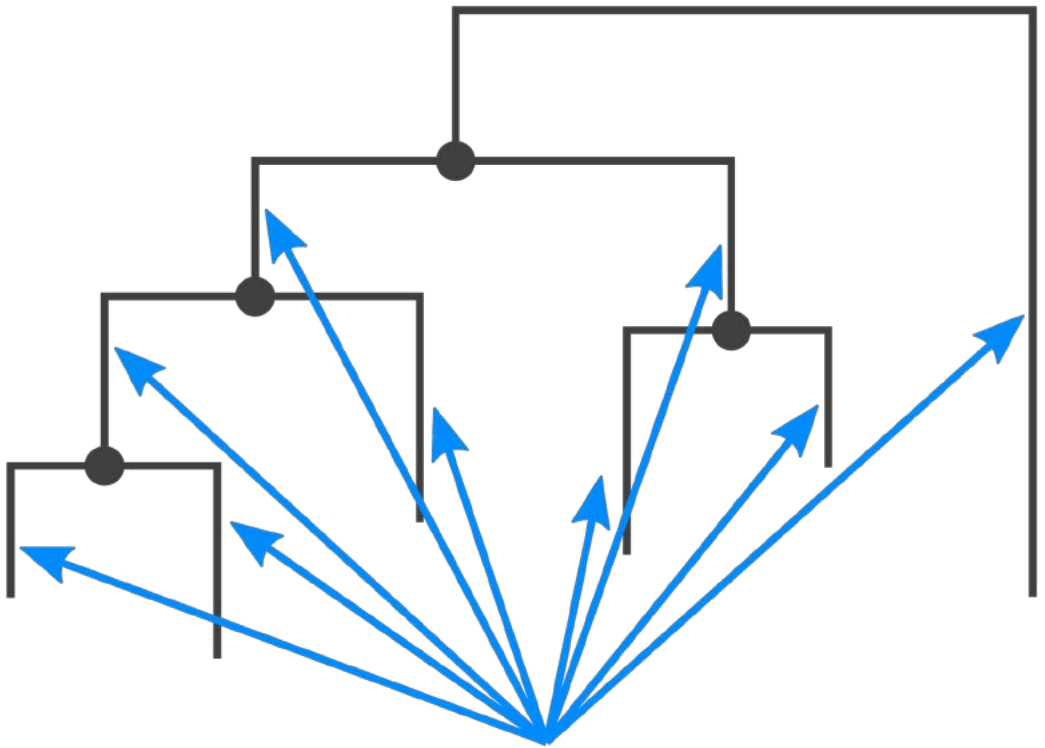
Phylogenetic Placement

- Given:
  - *Reference* tree and MSA
  - Set of *query sequences* (that we want to place on the tree)
- For each sequence:
  - Try out each branch as a potential *placement location*
  - Compute how likely this location is
- Repeat for all sequences → mapping from sequences to branches of the reference tree
- Tree is never changed, always stays fixed

Aligning to the Reference MSA

- However, typically, the query sequences are reads from a sequencing machine
- Have to align them to the given MSA first
- There are dedicated tools for aligning a set of new sequences to a given reference alignment:
  - hmmalign (part of hmmer): Uses a Markov model
  - PaPaRa: Uses reference tree to limit the number of sequences from the MSA that have to be considered
- There are also alignment-free placement methods, e.g., based on k-mers (not covered today)

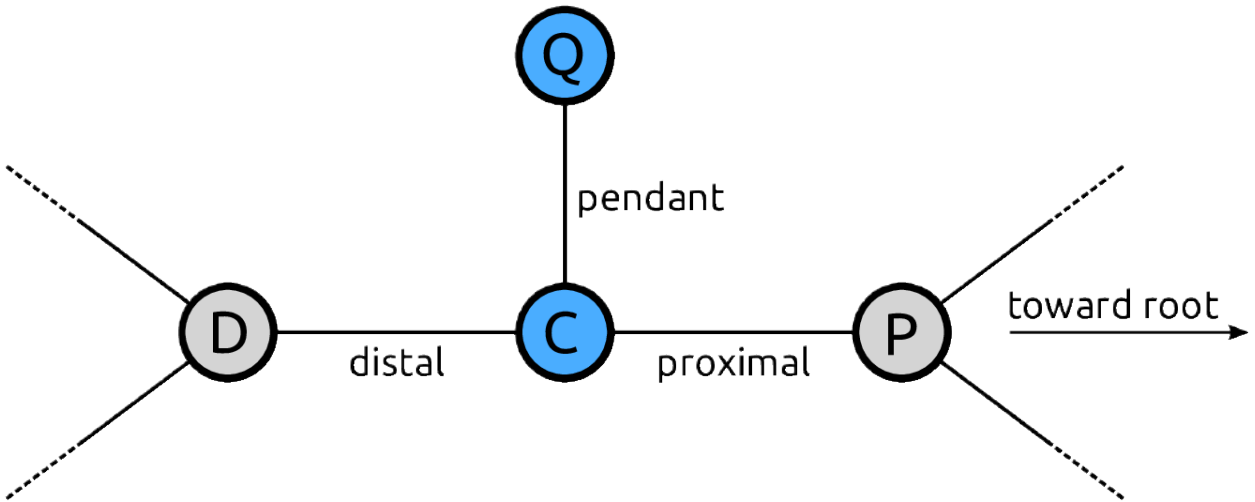
Phylogenetic Placement



Single Sequence:

Likelihood Computation

- For a single sequence on a single branch:
  - Pretend that this is actually a new tip node of the tree



  - Compute likelihood as before

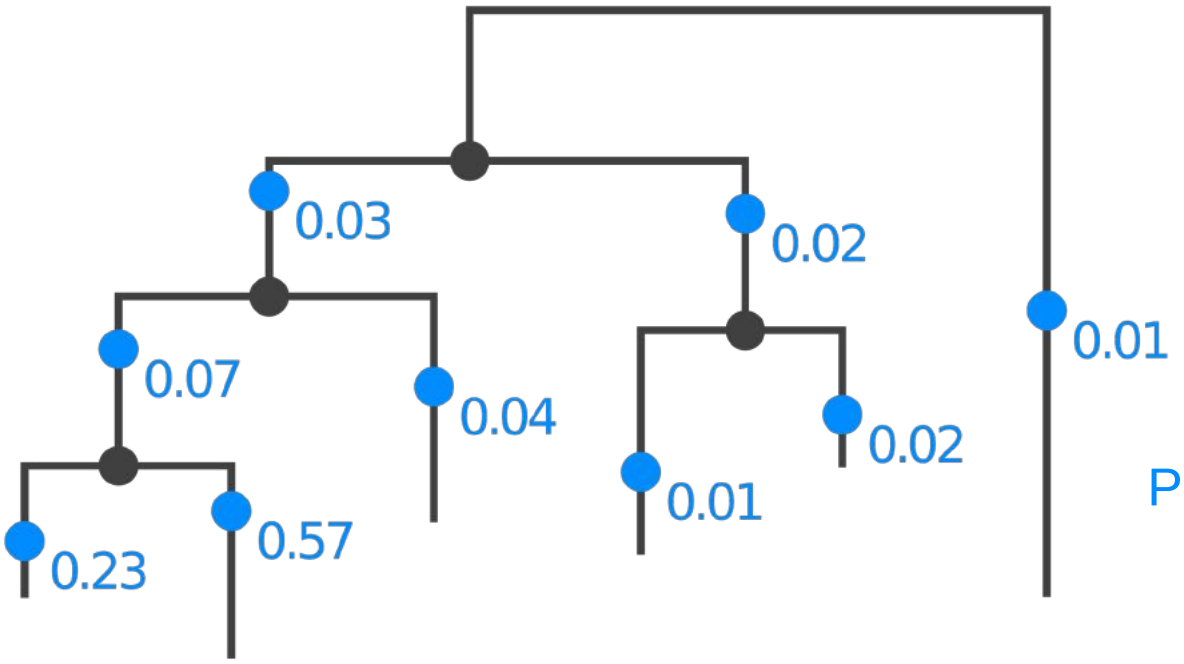
Likelihood Computation

- Repeat this for all branches of the tree
- Then, compute the *likelihood weight ratio* for each branch q:

$$\mathrm{LWR}(q) = \frac{\mathcal{L}(q)}{\sum_{i \in T} \mathcal{L}(i)}$$

- For a given query sequence, the sum of all LWRs over all branches is 1
- Can be interpreted as the probability of the sequence to be placed on that branch

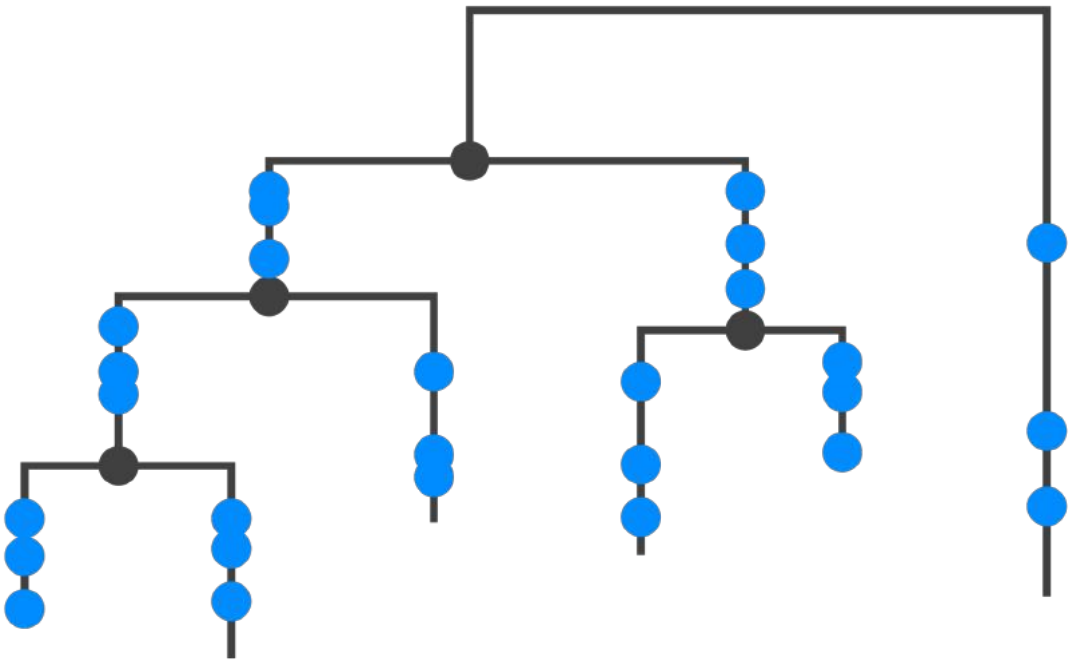
Phylogenetic Placement



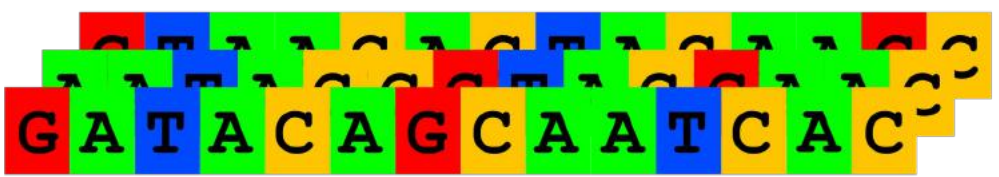
Placement Masses
≙
Probabilities

Single Sequence:

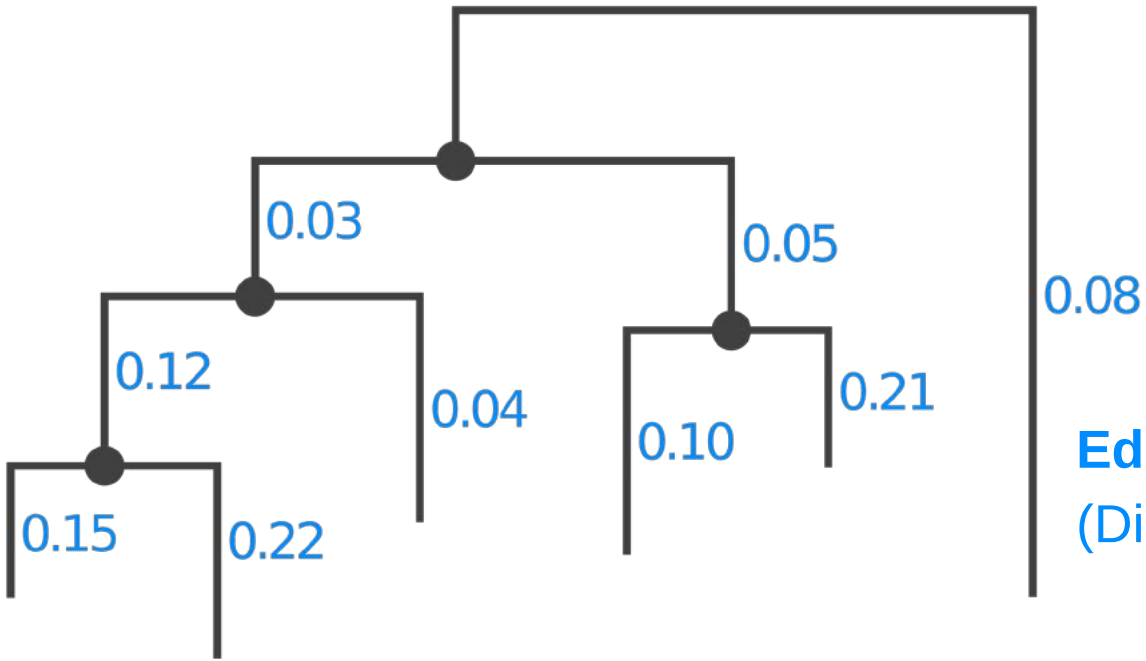
Phylogenetic Placement



Whole Sample:

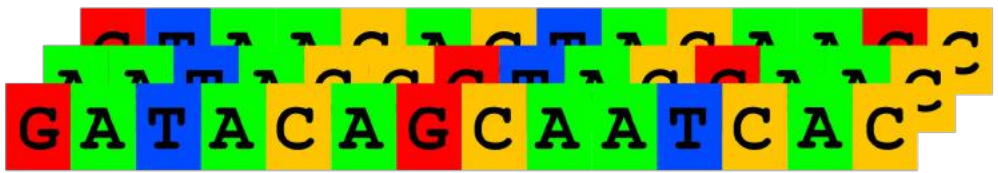


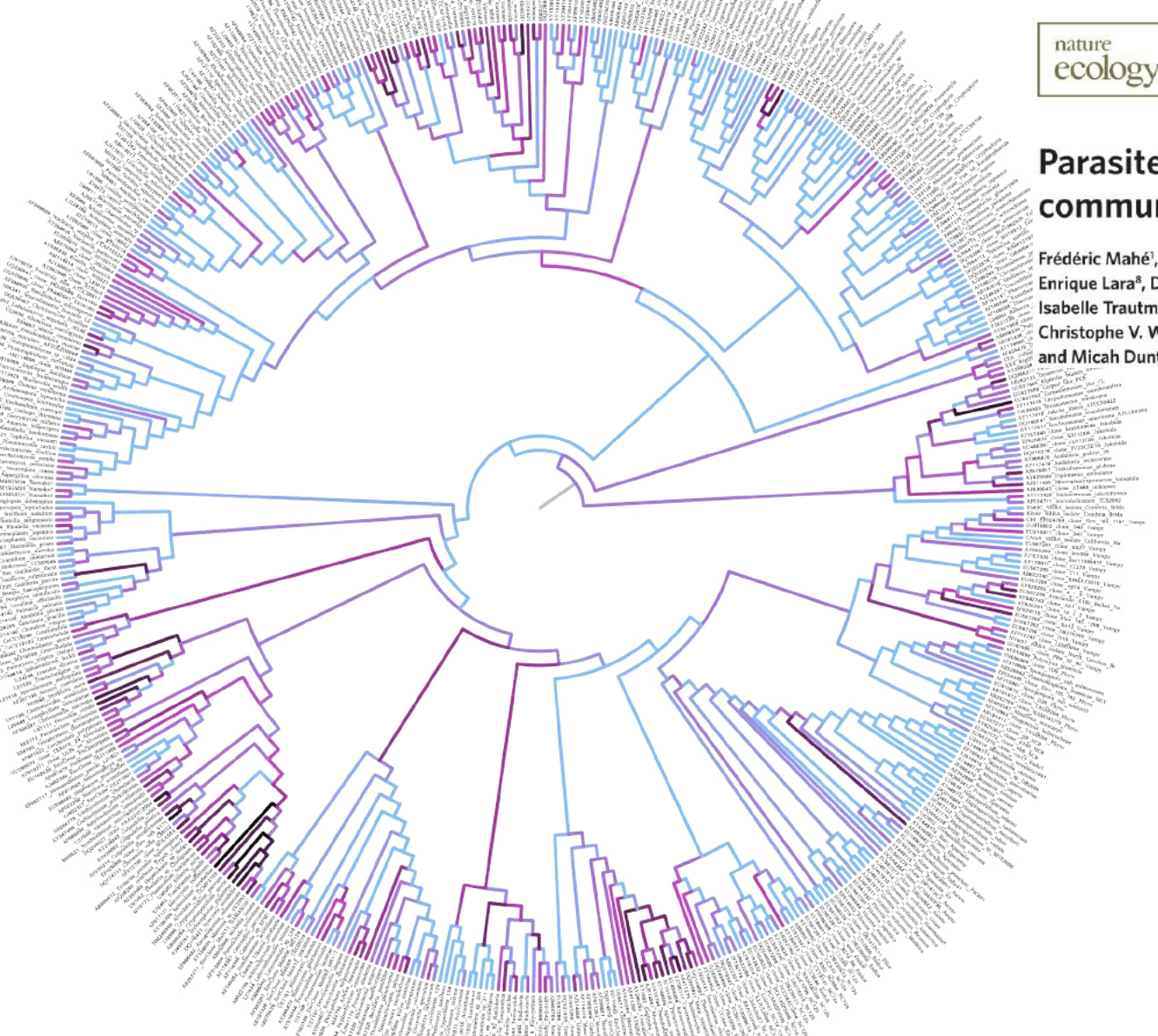
Phylogenetic Placement



Edge Masses
(Distribution on the Tree)

Whole Sample:





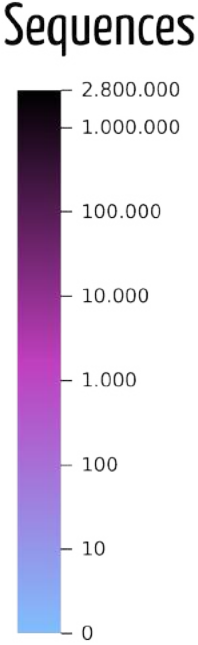
nature ecology & evolution

ARTICLES

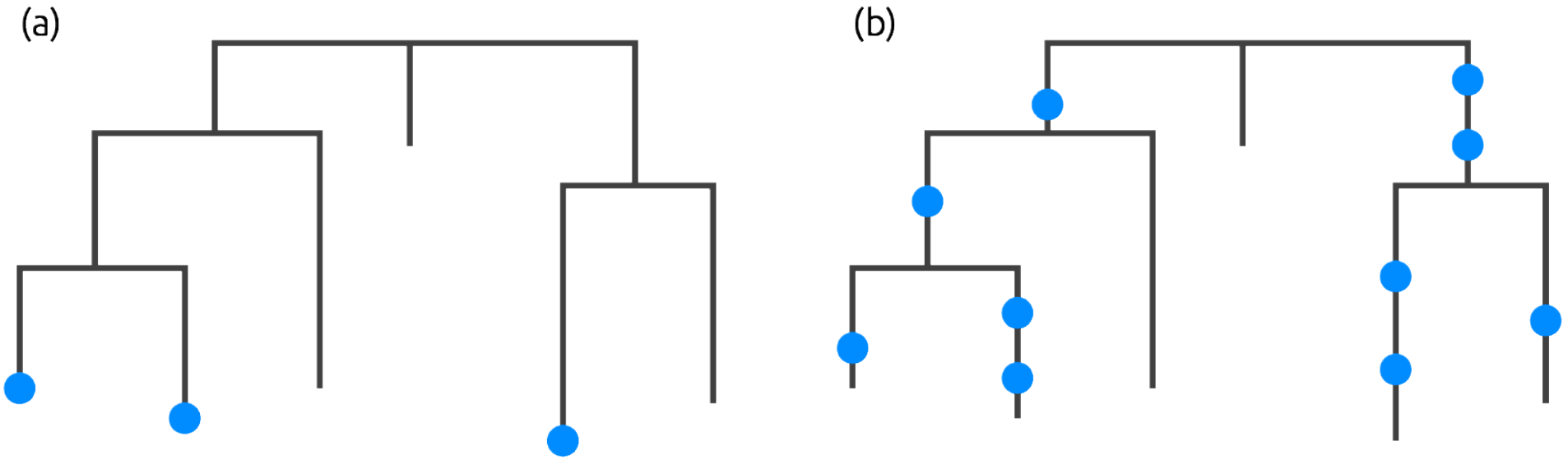
PUBLISHED: 20 MARCH 2017 | VOLUME: 1 | ARTICLE NUMBER: 0091

Parasites dominate hyperdiverse soil protist communities in Neotropical rainforests

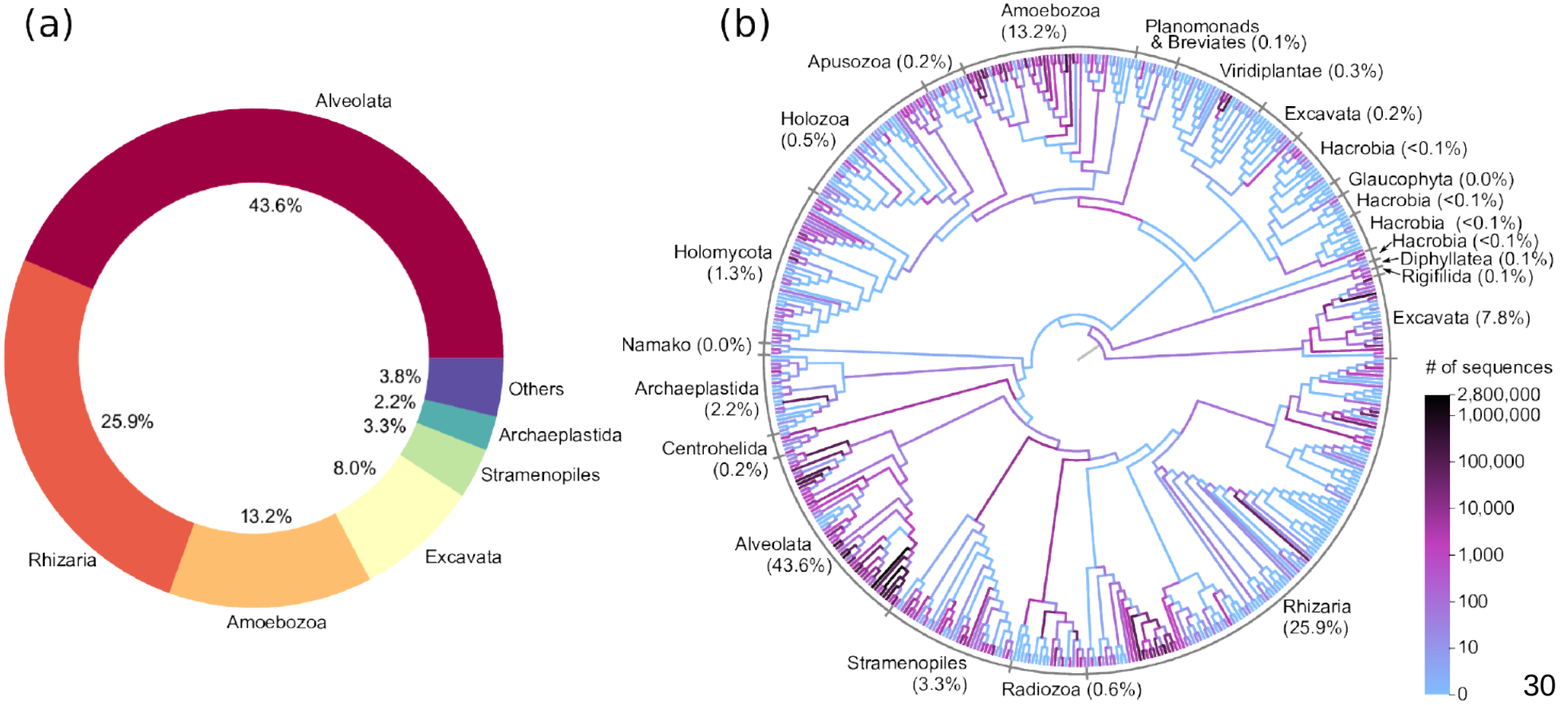
Frédéric Mahé[1], Colomban de Vargas[2,3], David Bass[4,5], Lucas Czech[6], Alexandros Stamatakis[6,7], Enrique Lara[8], David Singer[8], Jordan Mayor[9], John Bunge[10], Sarah Sernaker[11], Tobias Siemensmeyer[1], Isabelle Trautmann[1], Sarah Romac[2,3], Cédric Berney[2,3], Alexey Kozlov[6], Edward A. D. Mitchell[8,12], Christophe V. W. Seppey[8], Elianne Egge[13], Guillaume Lentendu[1], Rainer Wirth[14], Gabriel Trueba[15] and Micah Dunthorn[1*]



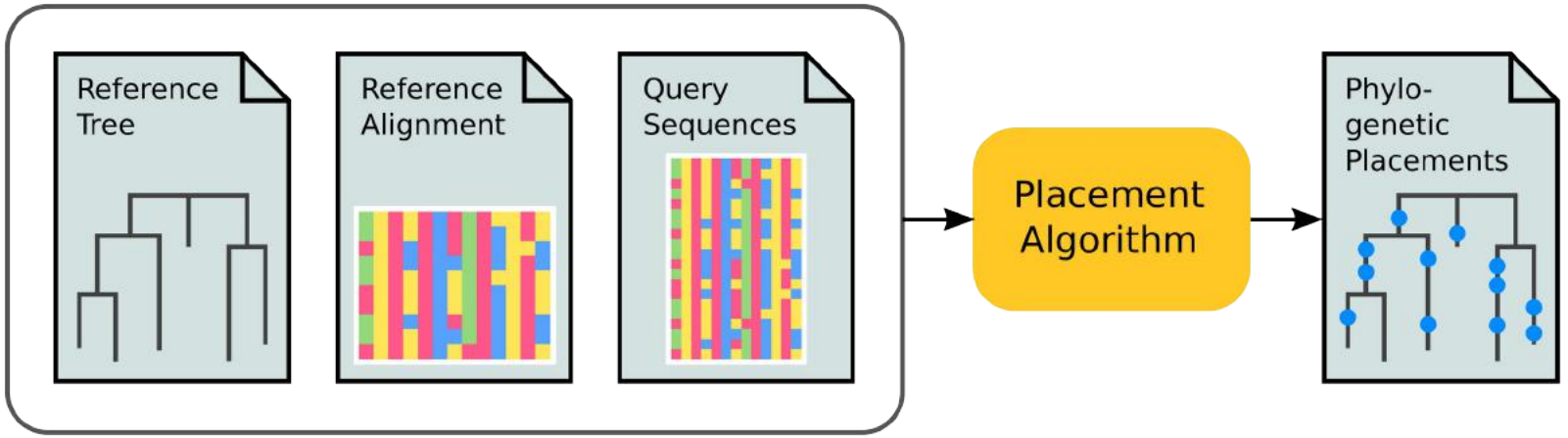
BLAST / vserach vs. Phylogenetic Placements



Abundances vs. Phylogenetic Placements



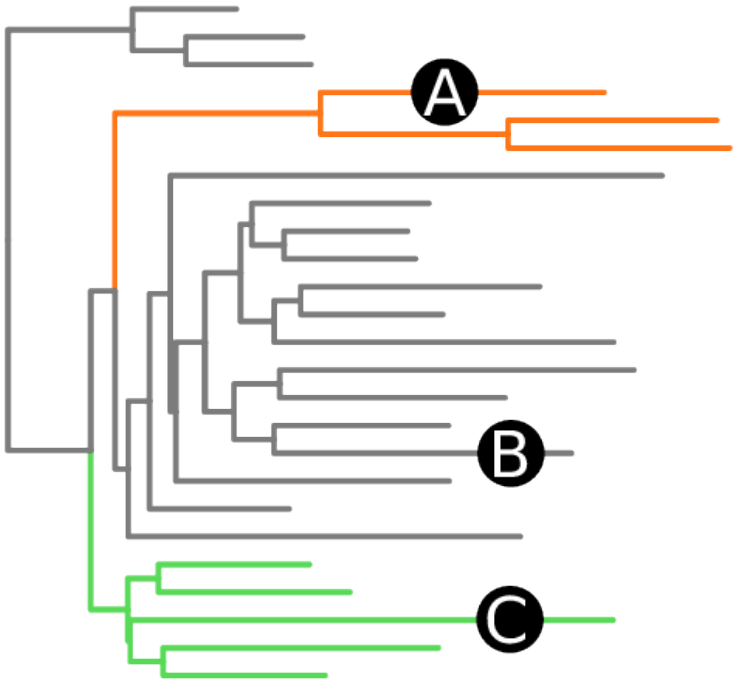
Phylogenetic Placement Pipeline



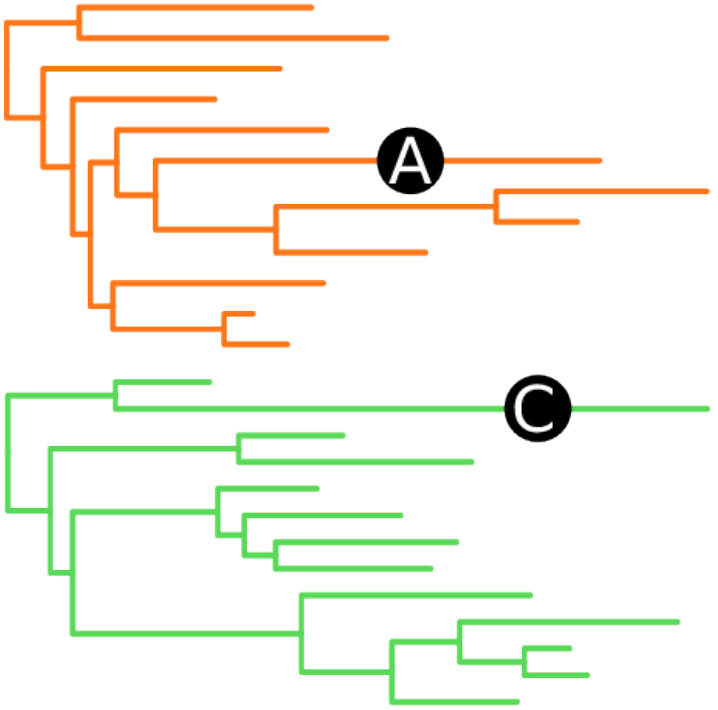
- Input:
  - Reference tree (newick)
  - Reference alignment (fasta or phylip)
  - Query sequences (fasta)
- Output:
  - Placements (jplace)

Multilevel Placement Pipeline

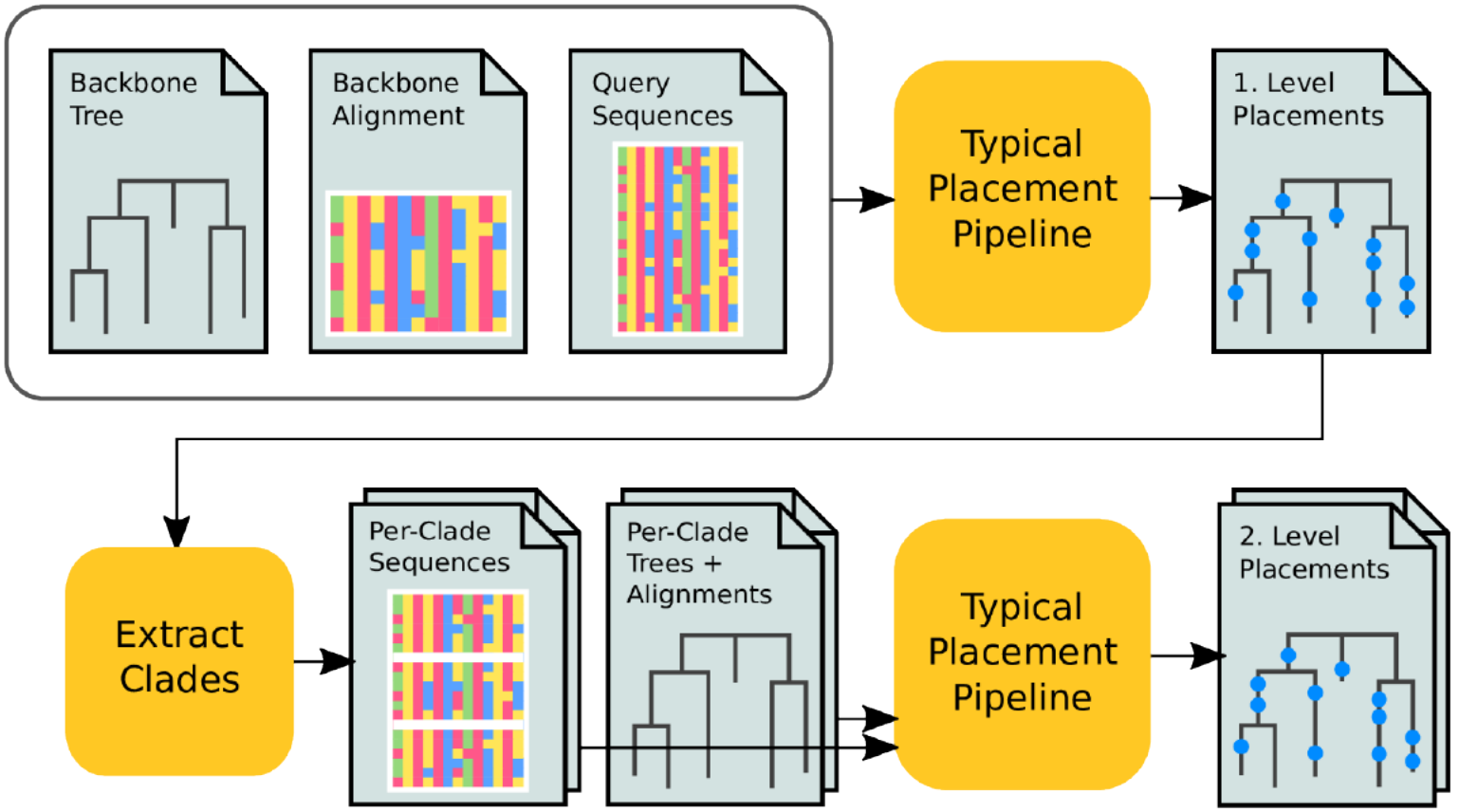
Backbone Tree (first level)



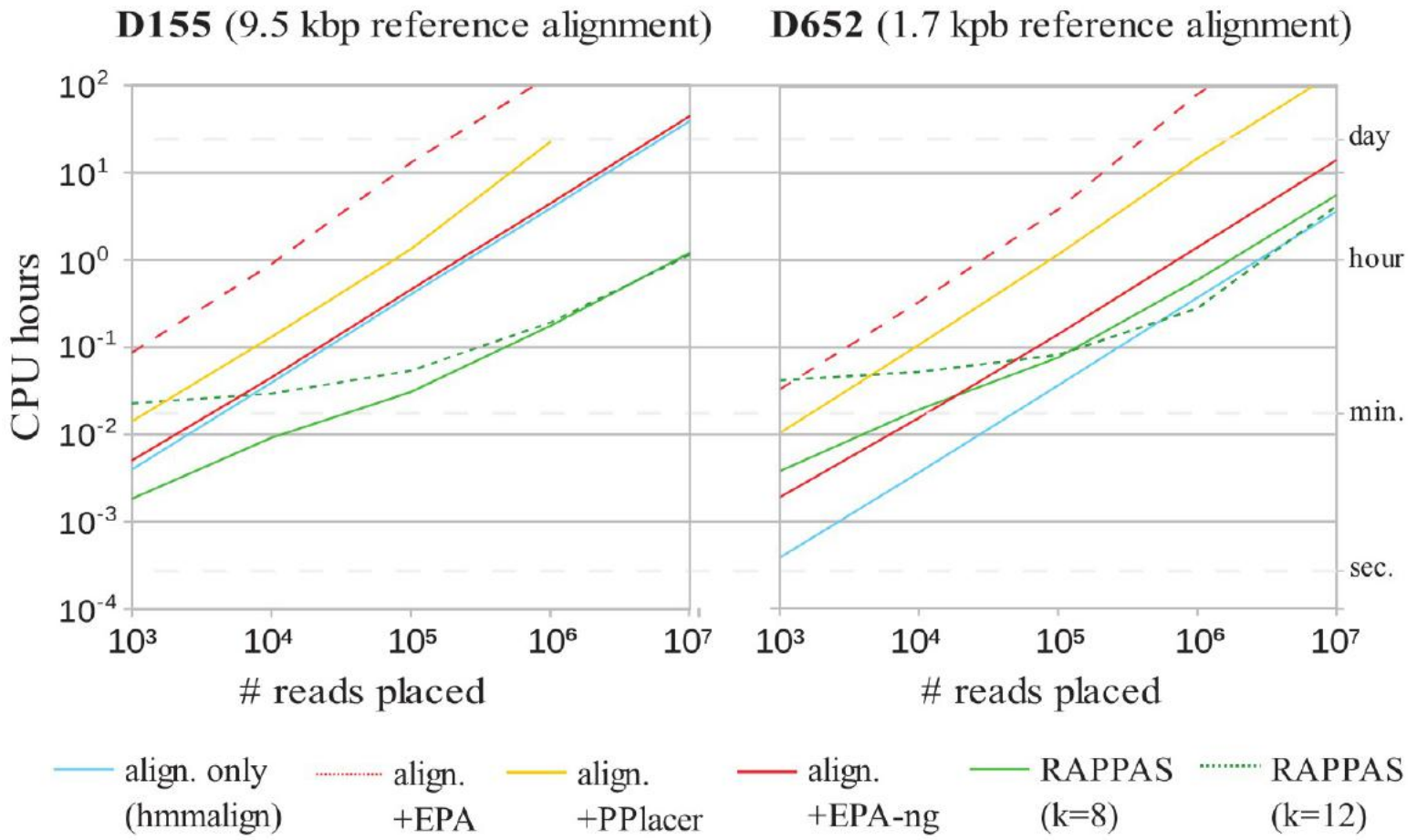
Clade Trees (second level)



Multilevel Placement Pipeline



Runtime Comparison

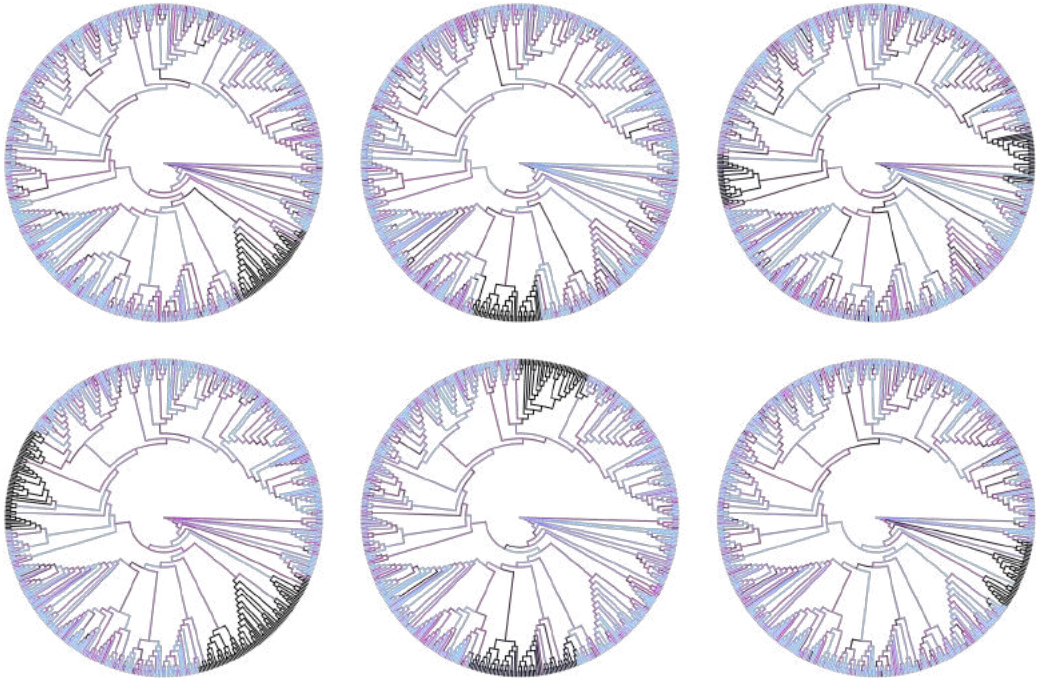


Placement Analysis and Visualization

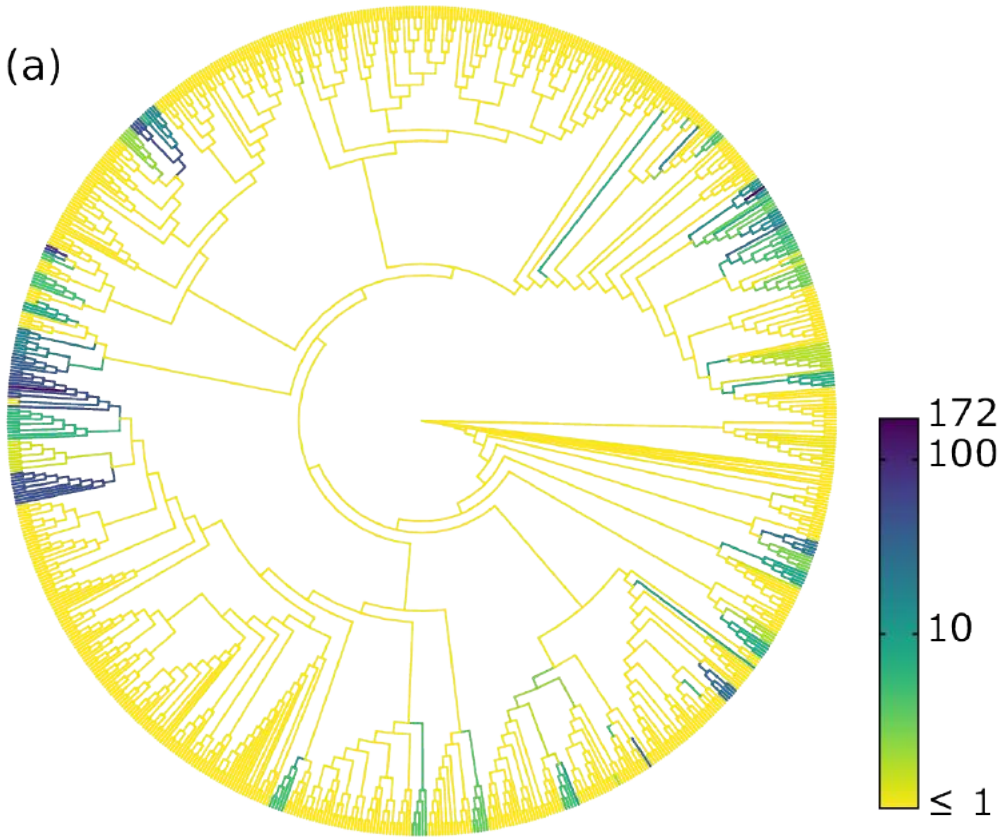
Placement of Multiple Samples

- Different people (human microbiome)
- Multiple locations in the forest / ocean / ...
- Points in time
- …

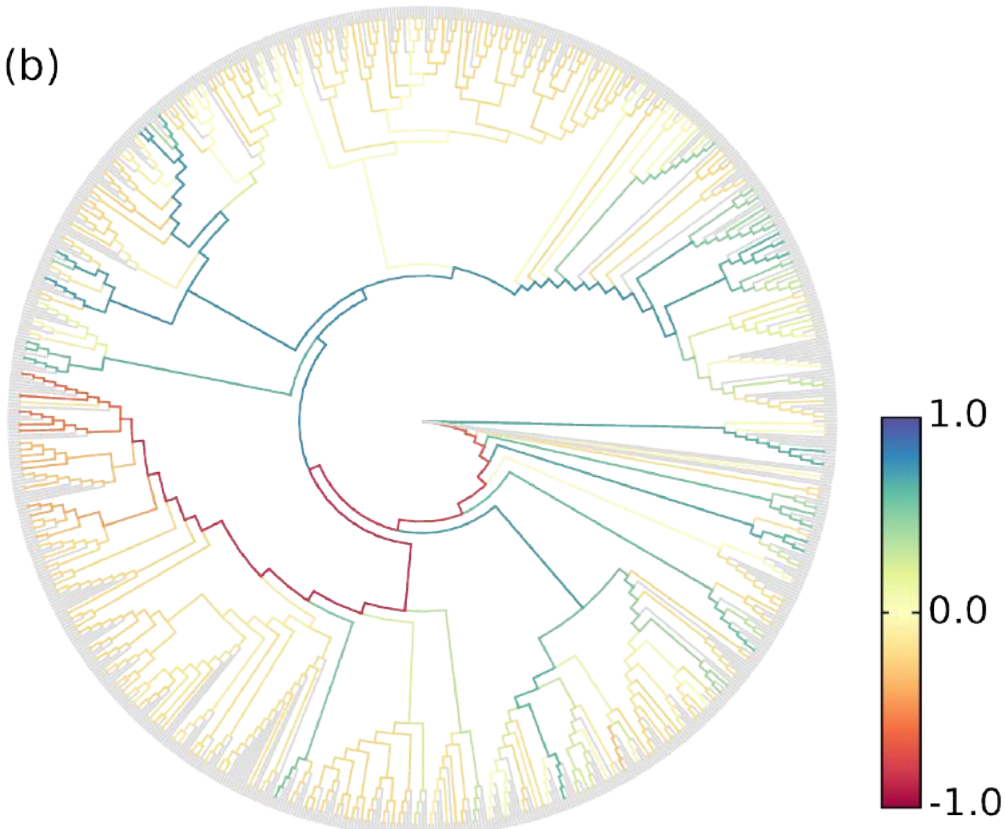
- Typically: Meta-data per sample
  - pH value
  - Temperature
  - ...



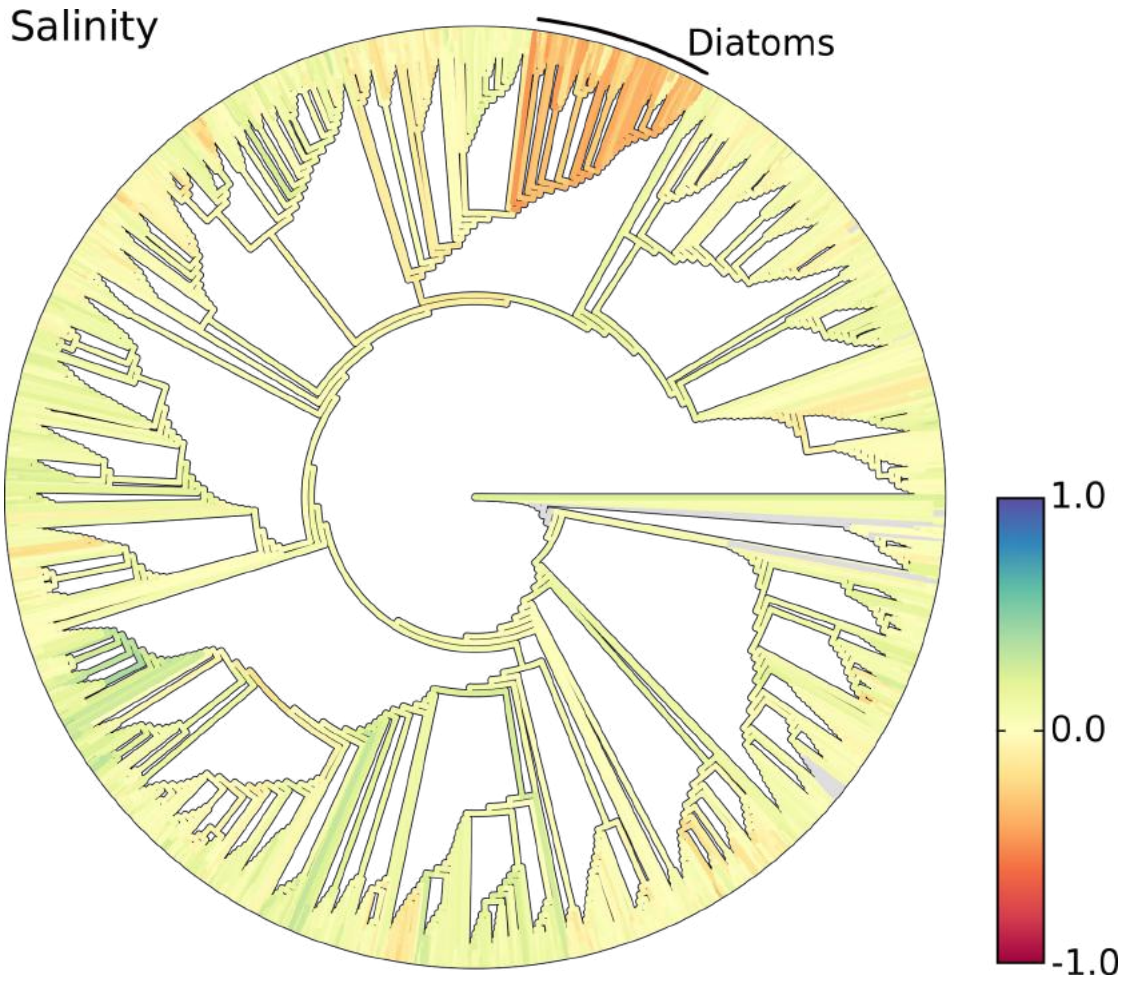
Edge Dispersion



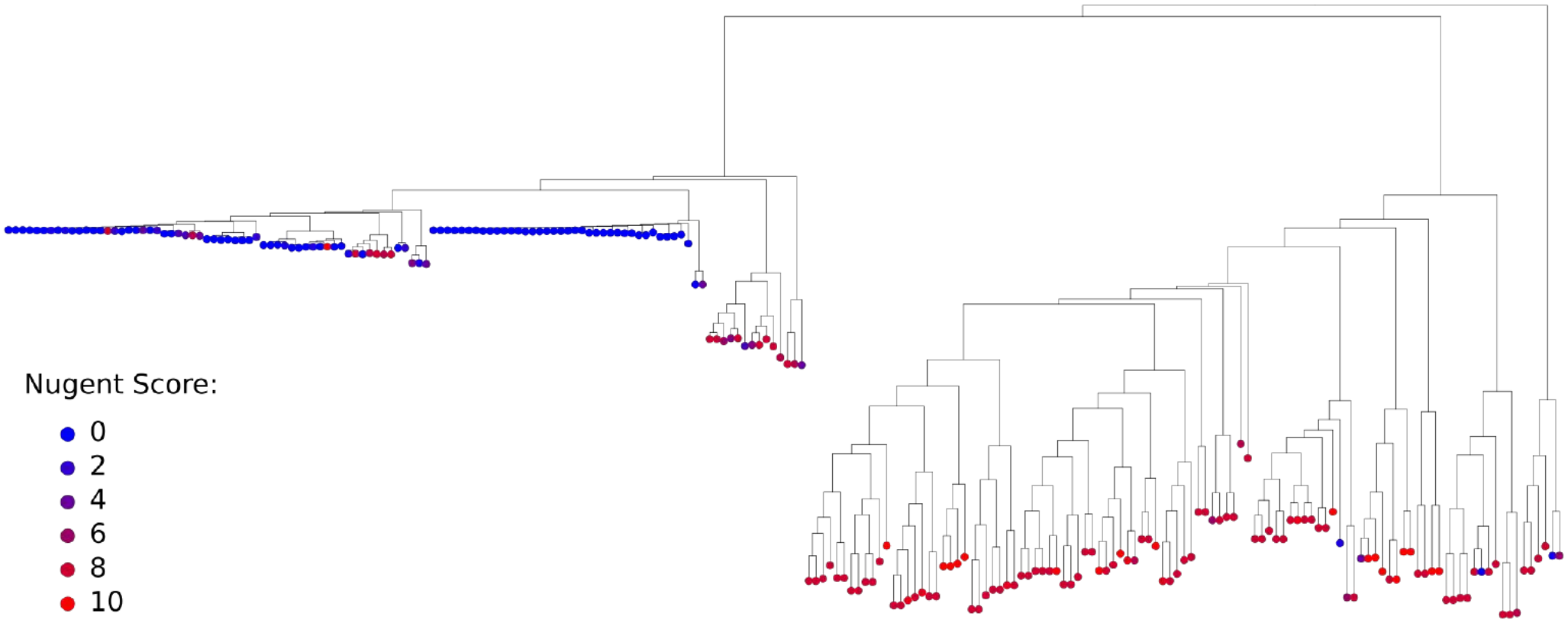
Edge Correlation



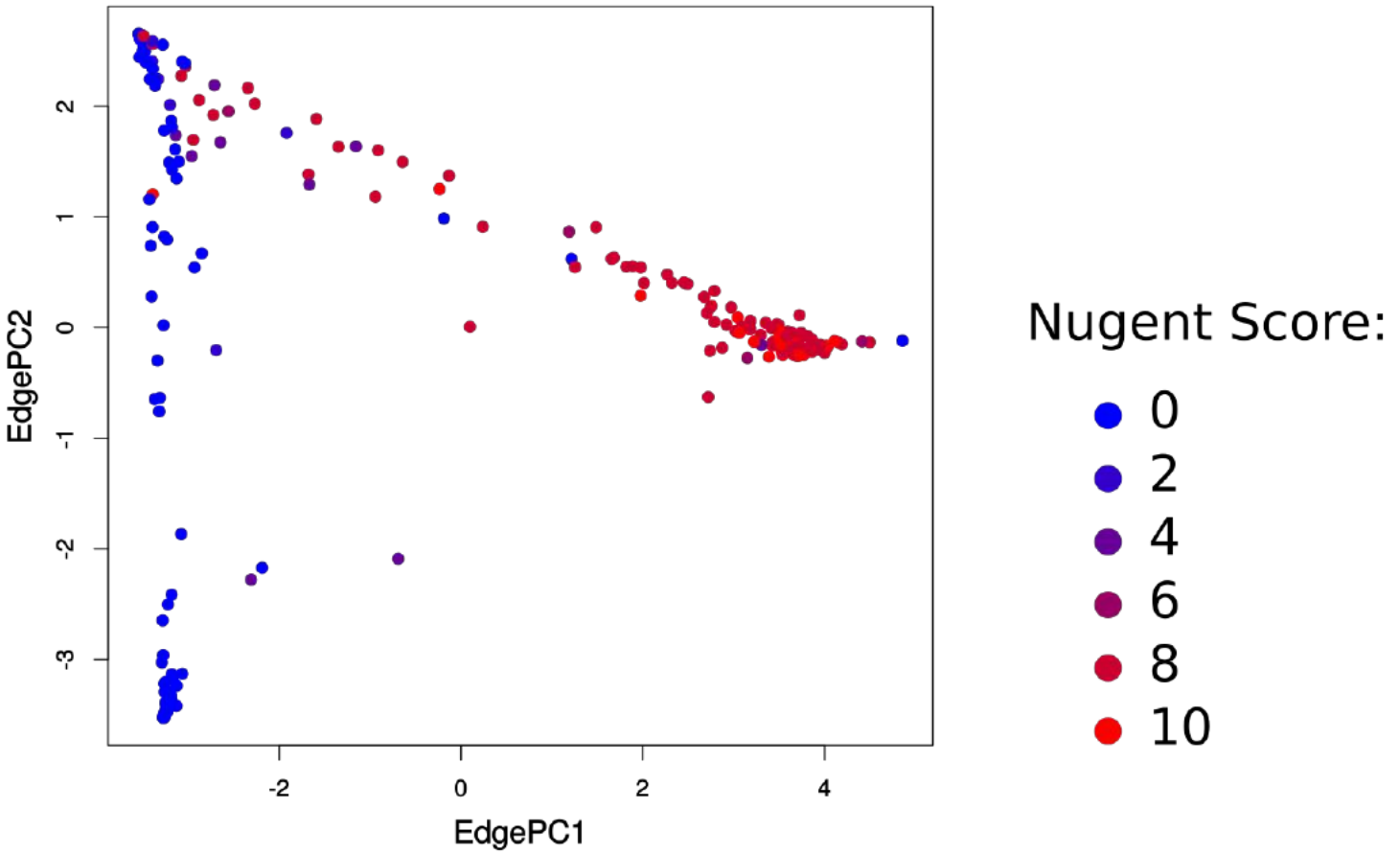
Edge Correlation



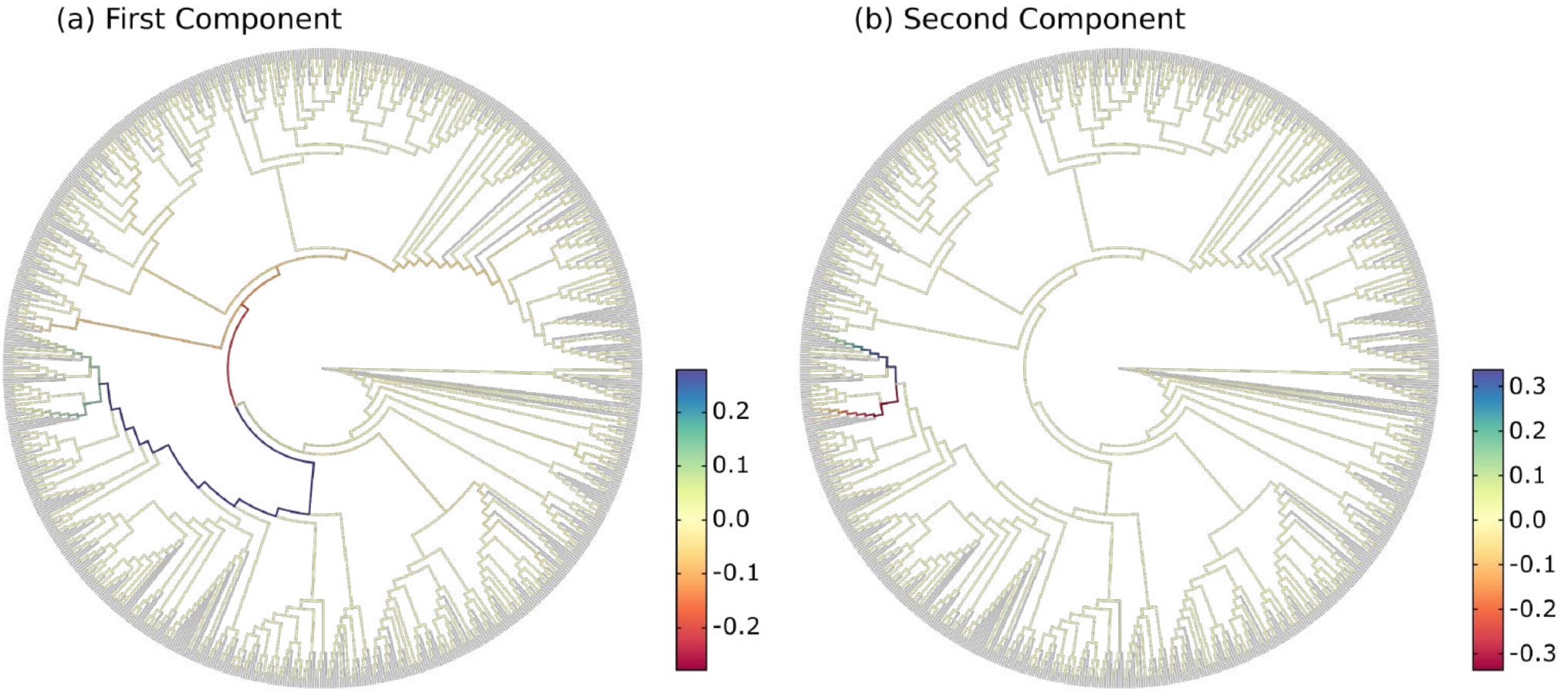
Squash Clustering



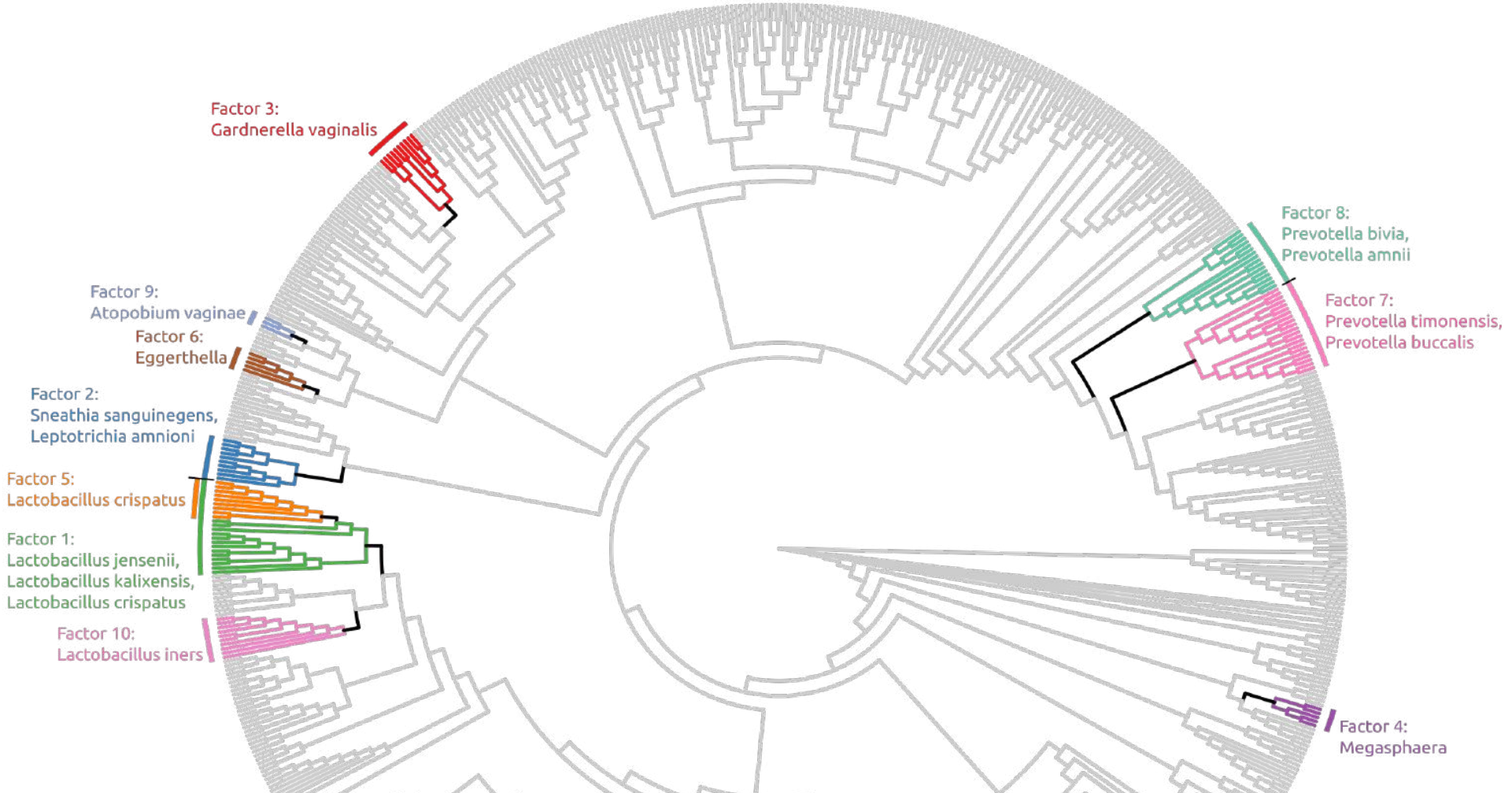
Edge PCA



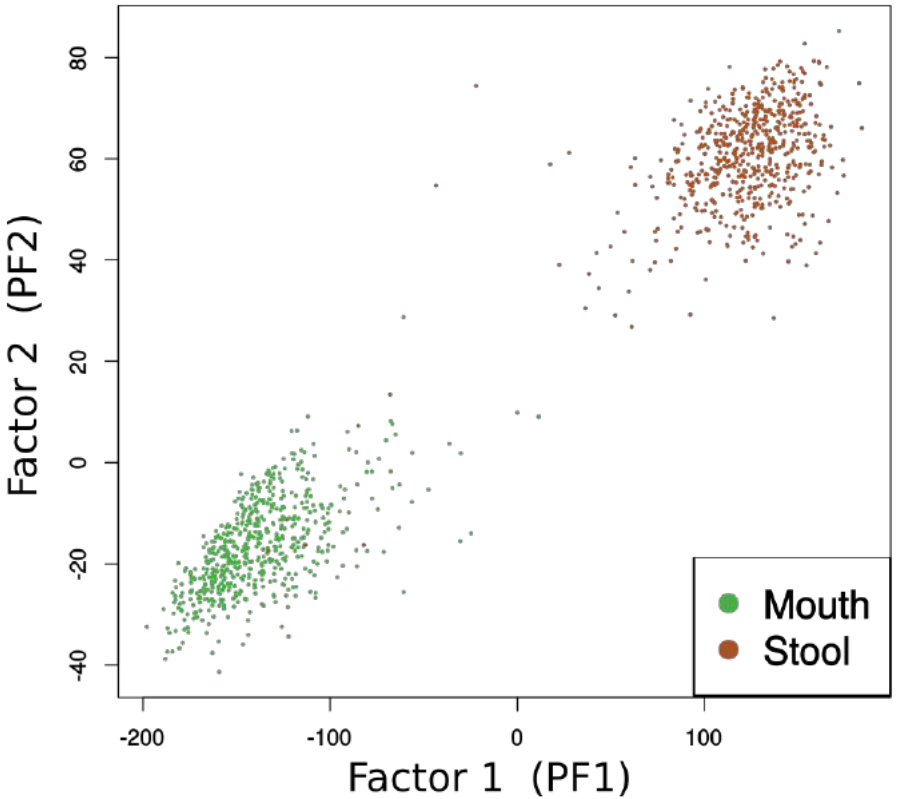
Edge PCA

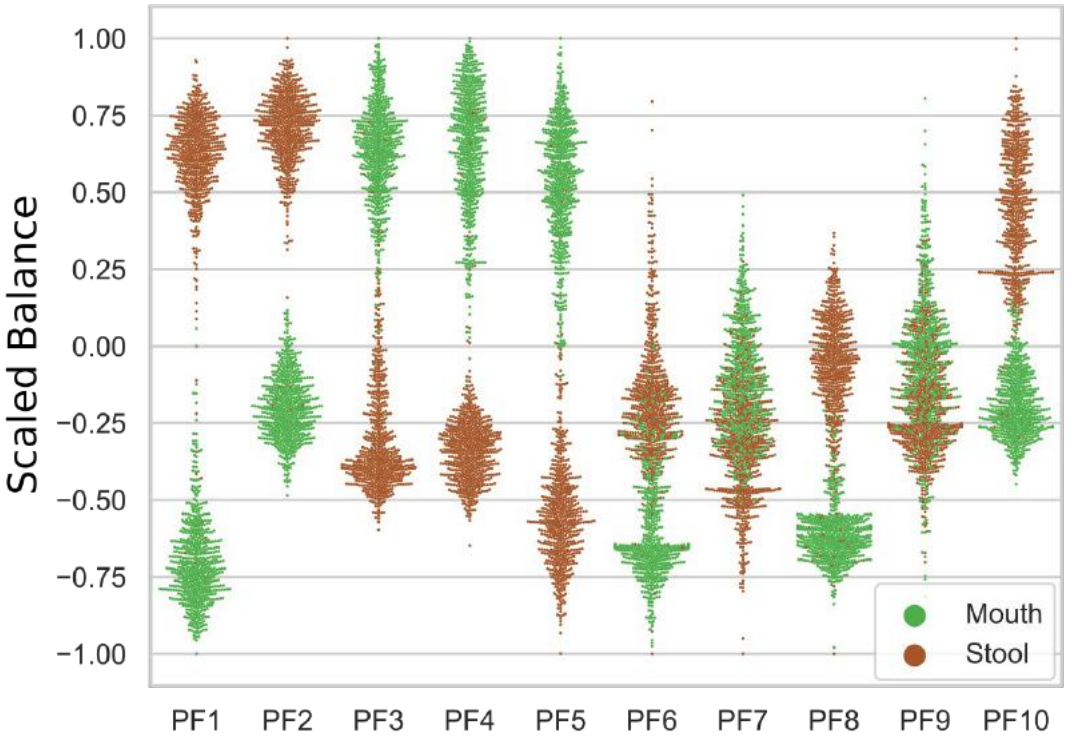

(a) First Component
0.2
0.1
0.0
-0.1
-0.2
(b) Second Component
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3

Placement-Factorization



Placement-Factorization





Thank you!

Time for your questions

Appendix

References

1. L. Czech, and A. Stamatakis, **“Scalable methods for analyzing and visualizing phylogenetic placement of metagenomic samples,”** PLOS ONE, 14(5), e0217050, 2019.
2. L. Czech, P. Barbera, and A. Stamatakis, **“Methods for Automatic Reference Trees and Multilevel Phylogenetic Placement”,** Bioinformatics, 299792. 2018.
3. L. Czech, P. Barbera, and A. Stamatakis, **“Genesis and Gappa: Processing, Analyzing and Visualizing Phylogenetic (Placement) Data”,** BioRxiv, 2019.
4. F. Mahé, C. de Vargas, D. Bass, L. Czech, A. Stamatakis, E. Lara, M. Dunthorn, **“Parasites dominate hyperdiverse soil protist communities in Neotropical rainforests,”** Nature Ecology & Evolution, 1(4), 0091, 2017.
5. S. Berger, D. Krompass, and A. Stamatakis, **“Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood,”** Syst. Biol., vol. 60, no. 3, pp. 291–302, 2011.
6. F. A. Matsen, R. B. Kodner, and E. V. Armbrust, **“pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree,”** BMC Bioinformatics, vol. 11, no. 1, p. 538, 2010.
7. P. Barbera, A. Kozlov, L. Czech, B. Morel, D. Darriba, T. Flouri, and A. Stamatakis, “**EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences,”** Systematic Biology, 68(2), 365–369, 2018.
8. F. A. Matsen and S. N. Evans, **“Edge principal components and squash clustering: using the special structure of phylogenetic placement data for sample comparison.,”** PLoS One, vol. 8, no. 3, pp. 1–17, 2011.
9. J. Silverman et al, **"A phylogenetic transform enhances analysis of compositional microbiota data,"** eLife, 6, e21887, 2017.
10. A. Washburne et al., **"Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets,"** PeerJ, 5, e2969, 2017.
11. S. N. Evans and F. A. Matsen, **“The phylogenetic Kantorovich-Rubinstein metric for environmental sequence samples,”** J. R. Stat. Soc. Ser. B Stat. Methodol., vol. 74, pp. 569–592, 2012.
12. S. Srinivasan, N. G. Hoffman, M. T. Morgan, F. A. Matsen, T. L. Fiedler, R. W. Hall, F. J. Ross, C. O. McCoy, R. Bumgarner, J. M. Marrazzo, and D. N. Fredricks, **“Bacterial communities in women with bacterial vaginosis: High resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria,”** PLoS One, vol. 7, no. 6, p. e37818, 2012.
13. B. A. Methé et al., **“A framework for human microbiome research,”** Nature, 486(7402):215–221, 2012.
14. C. Huttenhower et al., **“Structure, function and diversity of the healthy human microbiome,”** Nature, 486(7402):207–214, 2012.

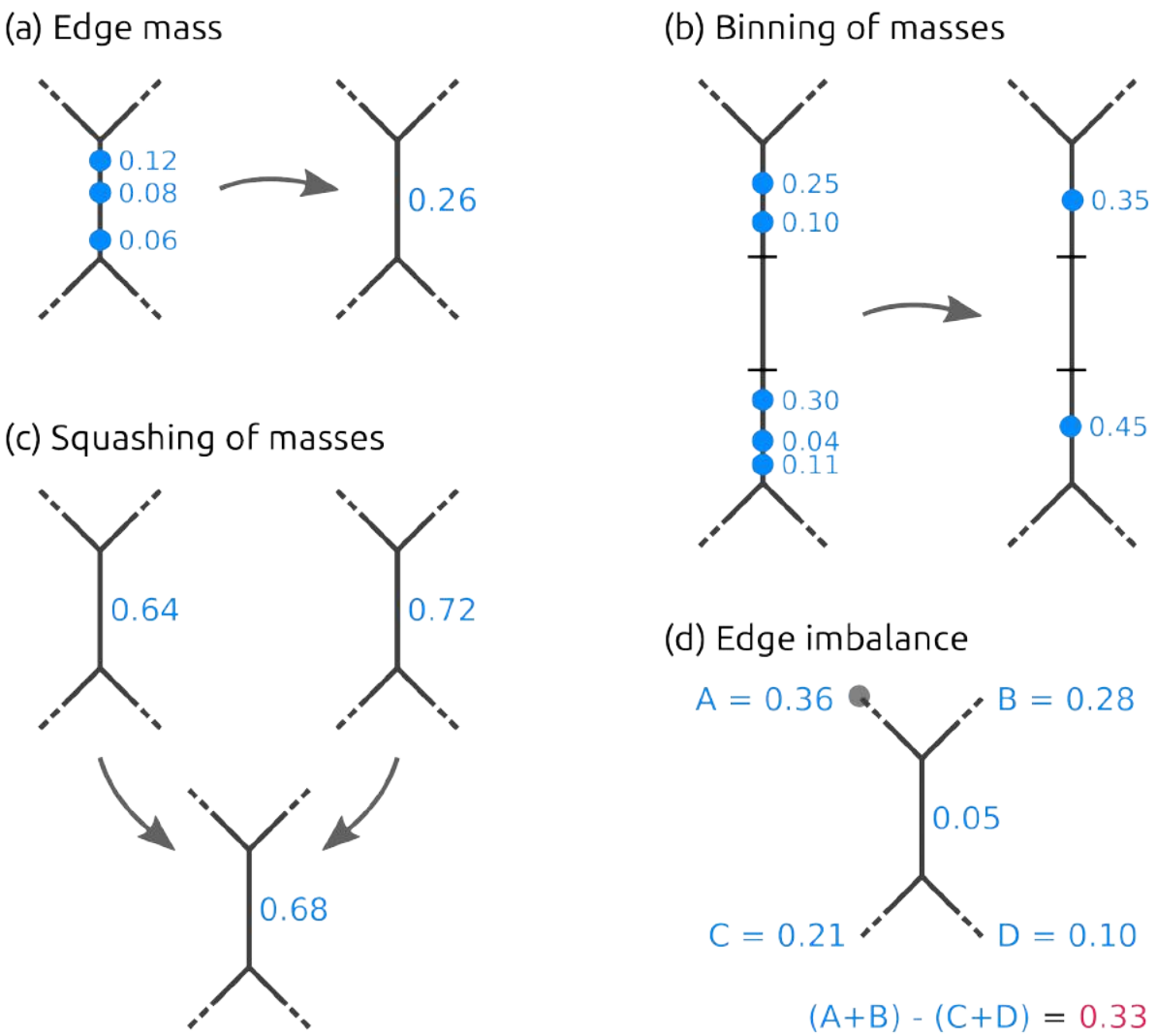
Implementation

genesis

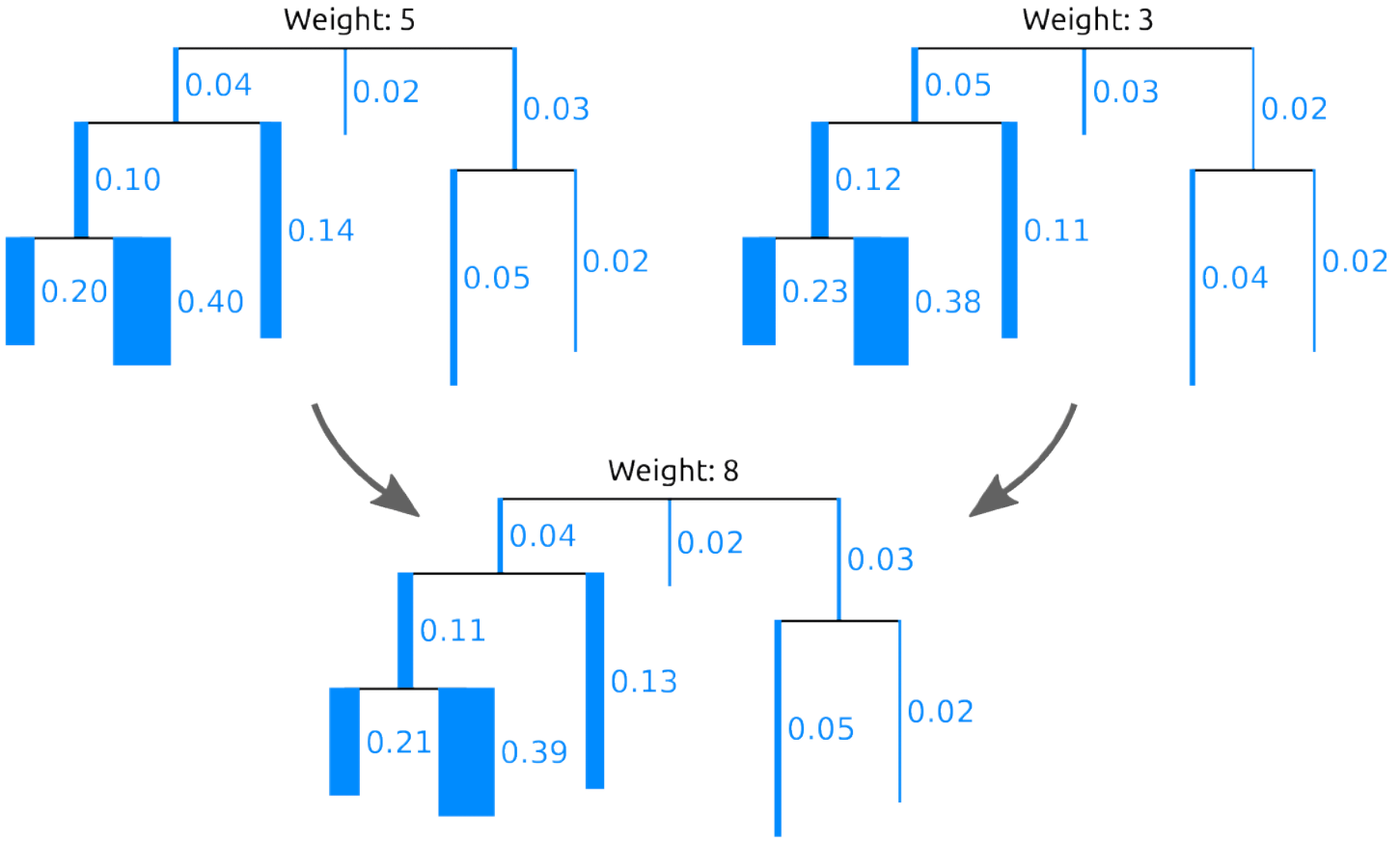
- C++ Library
- ~50k Lines of Code
- Used in many projects already
- Also re-implements existing methods

gappa

- Command Line Tool
- ~10k Lines of Code
- 5600 downloads on conda

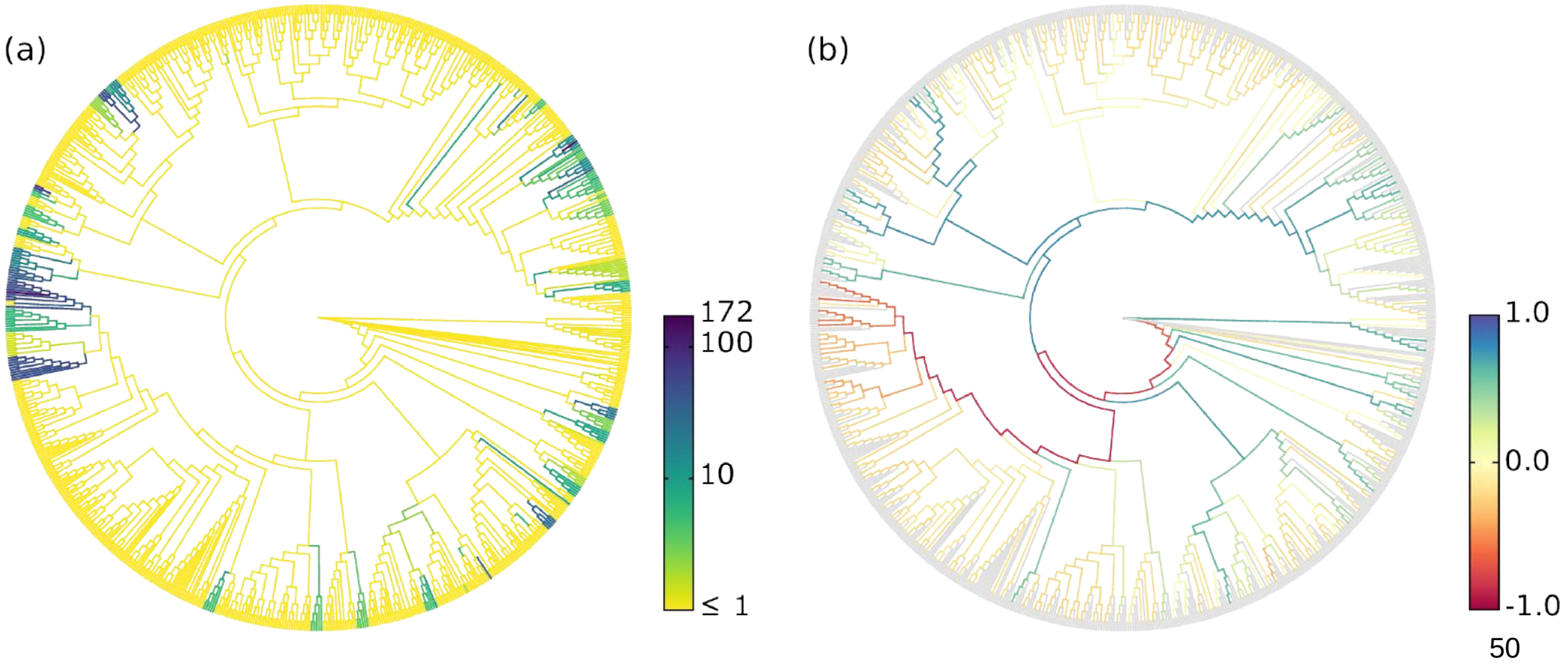

(a) Edge mass
0.12
0.08
0.06
0.26
(b) Binning of masses
0.25
0.10
0.30
0.04
0.11
0.35
0.45
(c) Squashing of masses
0.64
0.72
0.68
(d) Edge imbalance
A = 0.36
B = 0.28
0.05
C = 0.21
D = 0.10
(A+B) - (C+D) = 0.33

Squashing of Edge Masses



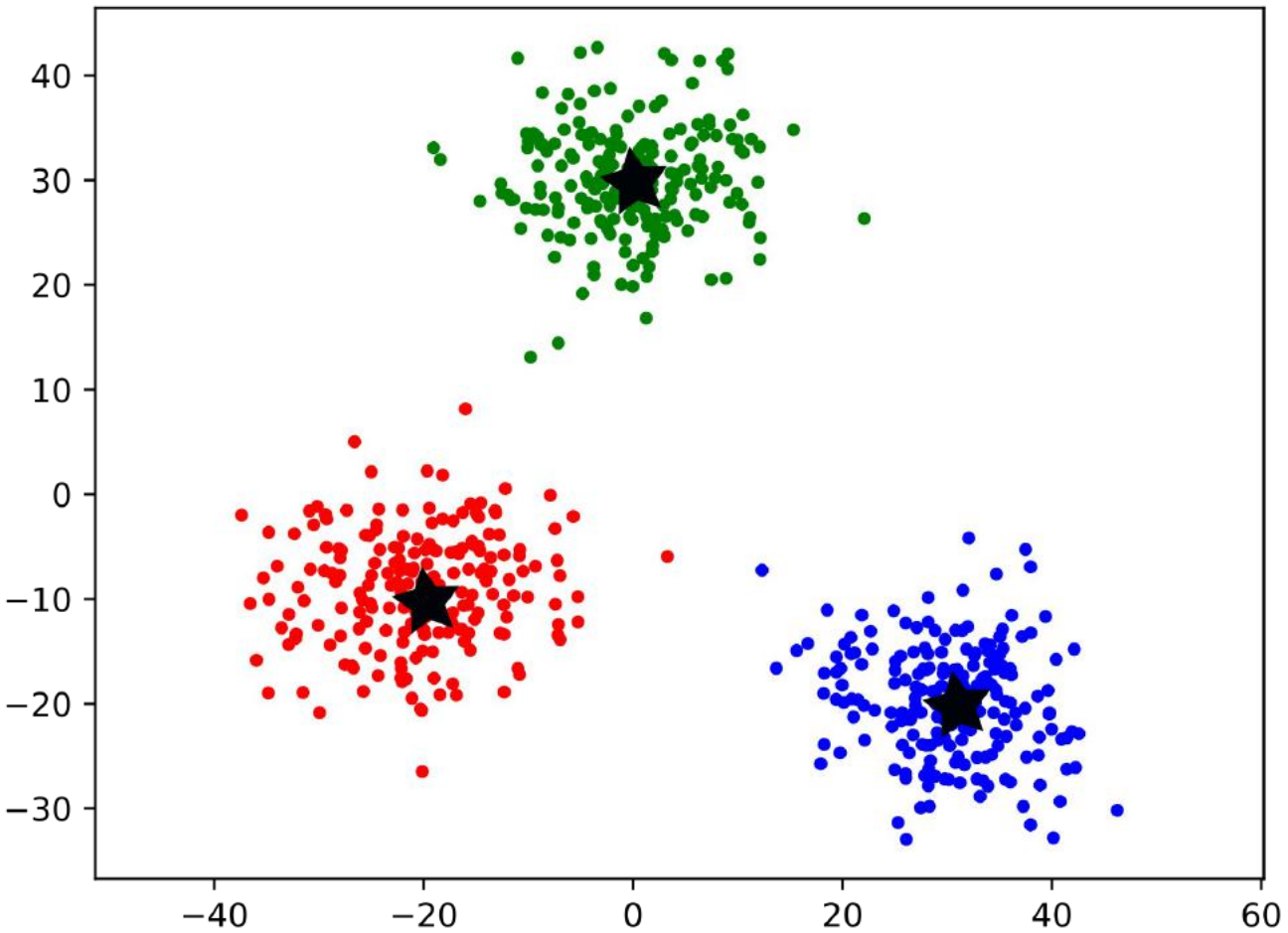
Edge Dispersion

Edge Correlation



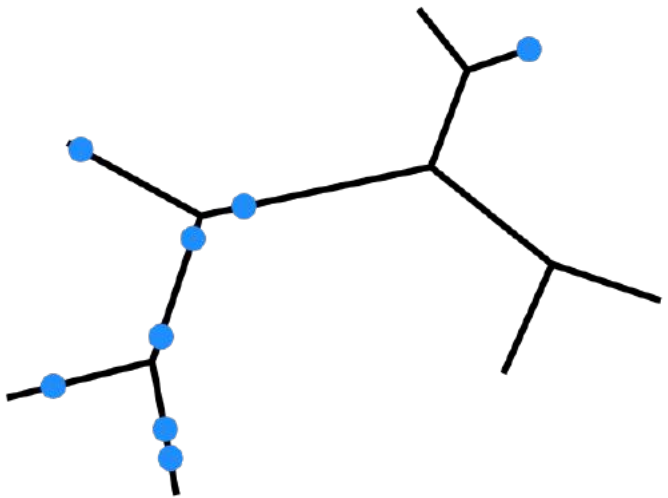
Clustering

K-means Clustering

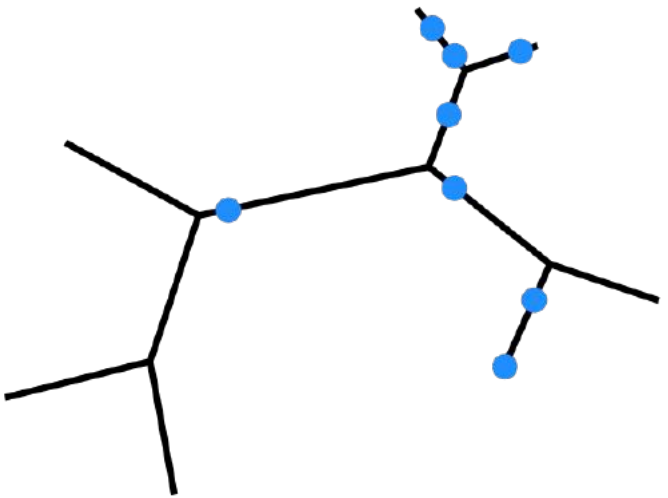


Earth Mover's Distance

(between masses on the reference tree)



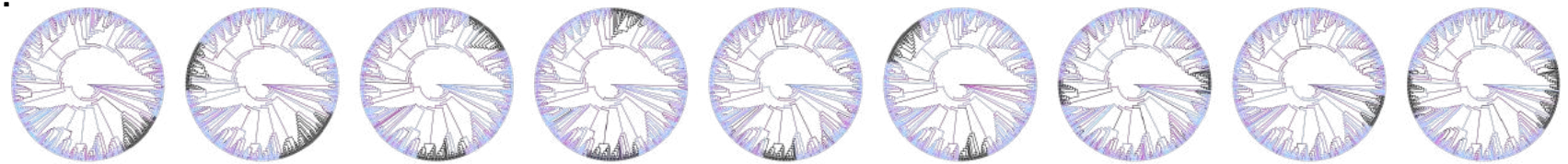
Sample A



Sample B

Phylogenetic K-means Clustering

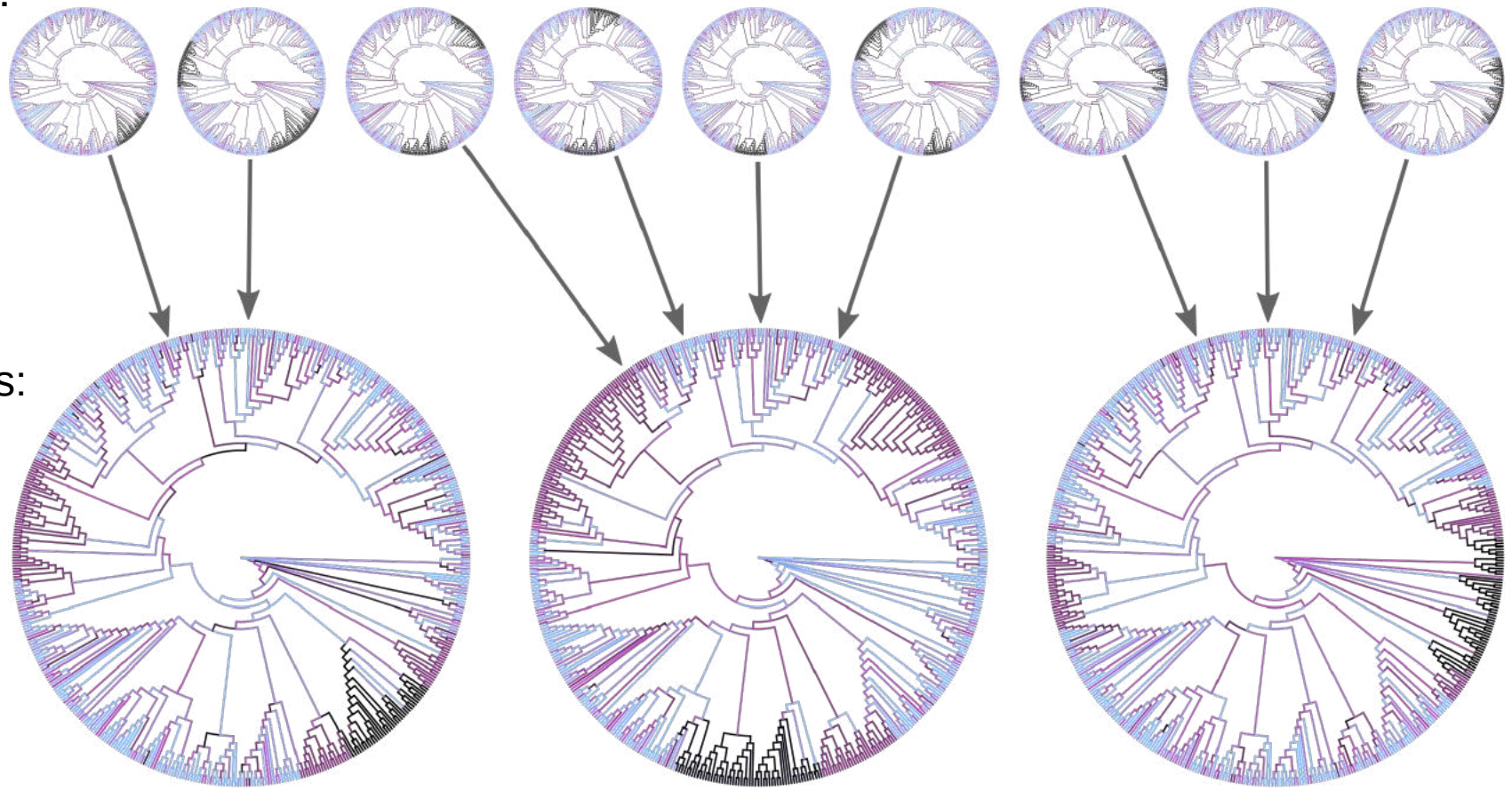
Samples:



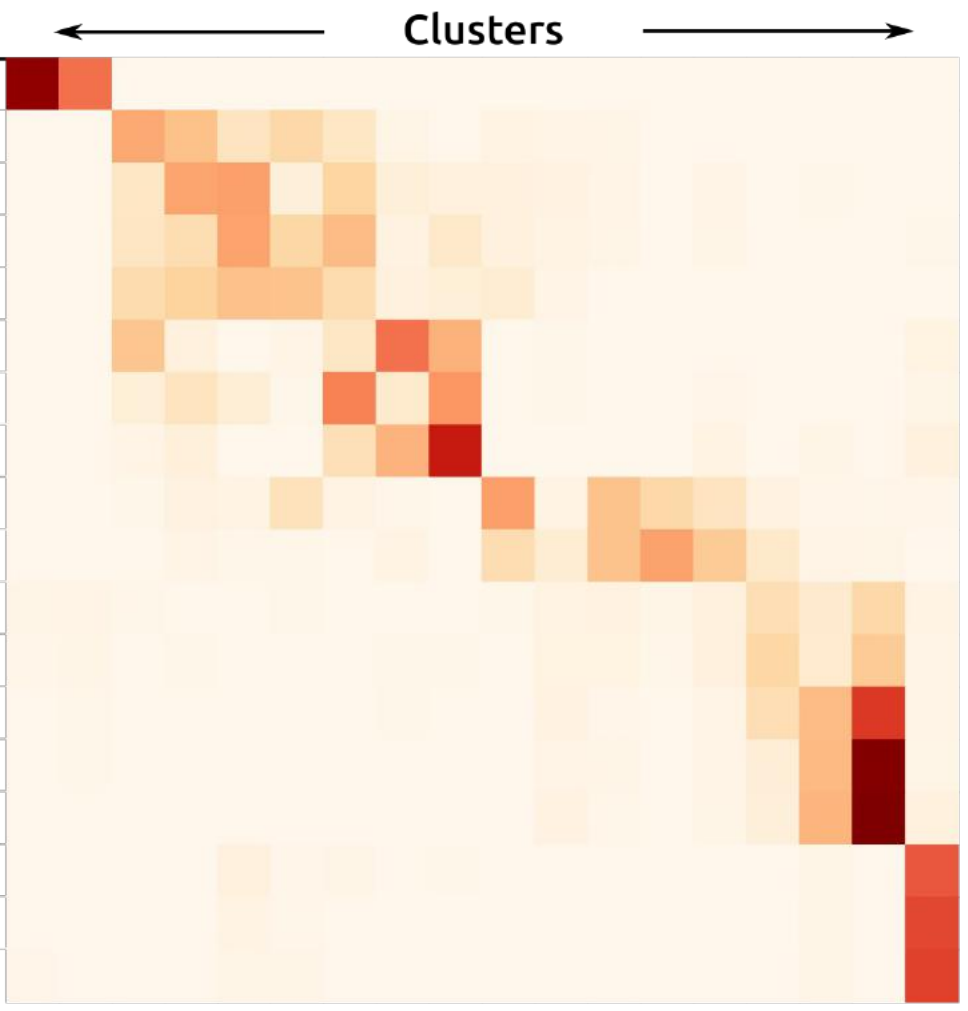
Phylogenetic K-means Clustering

Samples:

Centroids:



Body Site	Region
Stool	Stool
Saliva	Saliva
Tongue Dorsum	Mouth (back)
Throat	Mouth (back)
Palatine Tonsils	Mouth (back)
Attached Keratinized Gingiva	Mouth (front)
Hard Palate	Mouth (front)
Buccal Mucosa	Mouth (front)
Subgingival Plaque	Plaque
Supragingival Plaque	Plaque
Left Antecubital Fossa	Arm
Right Antecubital Fossa	Arm
Anterior Nares	Nose
Left Retroauricular Crease	Ear
Right Retroauricular Crease	Ear
Vaginal Introitus	Vagina
Mid Vagina	Vagina
Posterior Fornix	Vagina

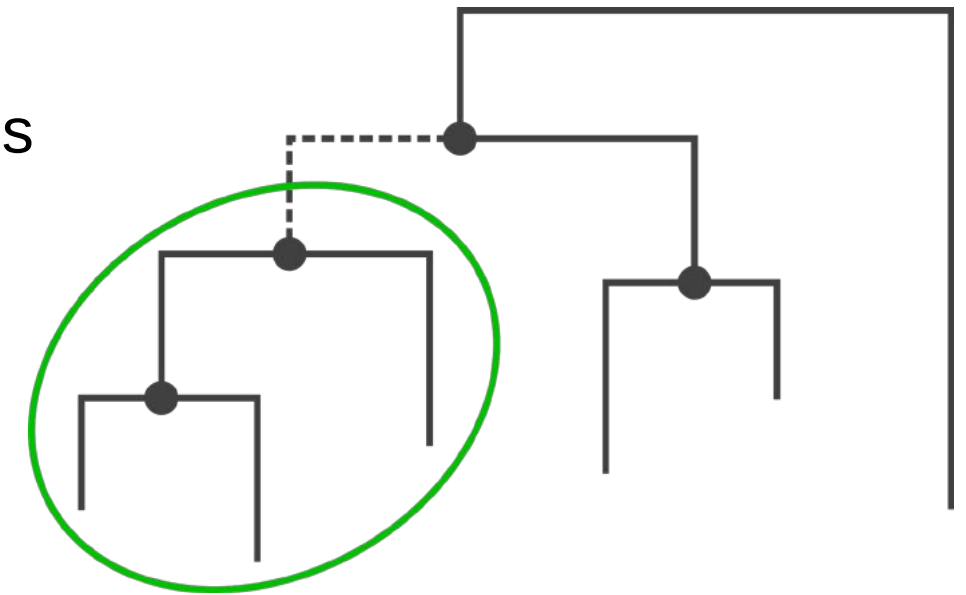


Balances

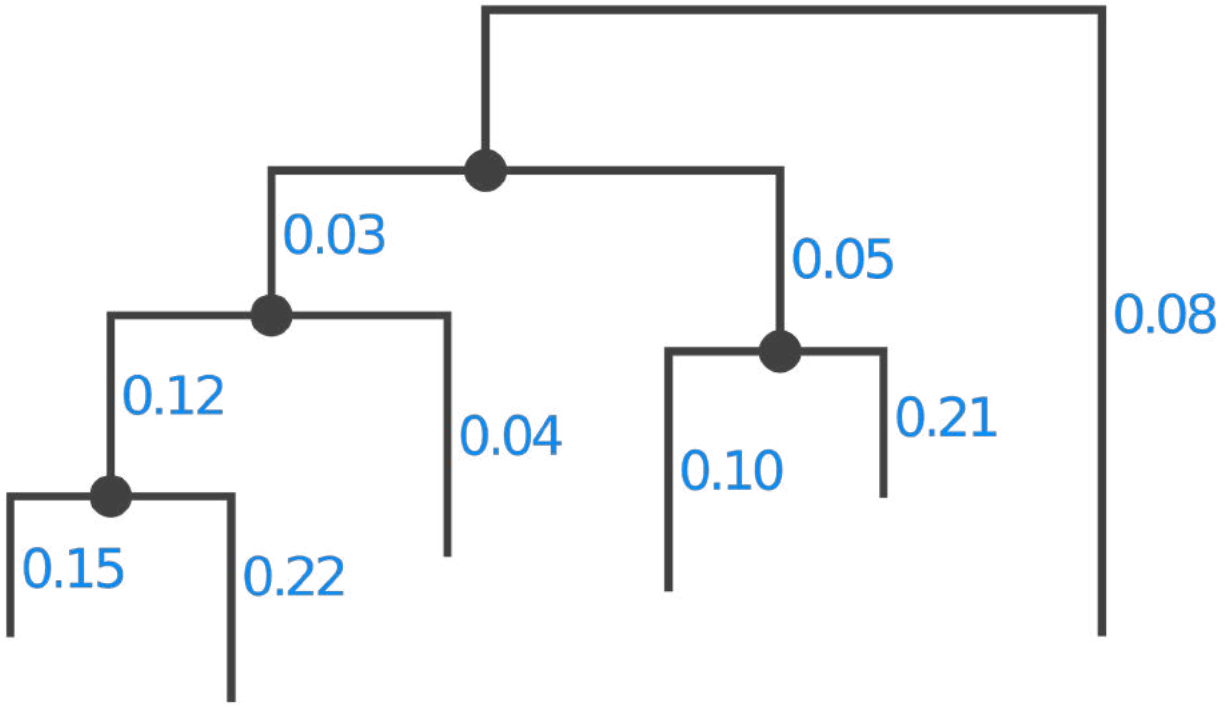
Motivation

- Often, groups of organisms are important biologically
- Hence, look at subtrees instead of single edges
- Need a transformation of the data!

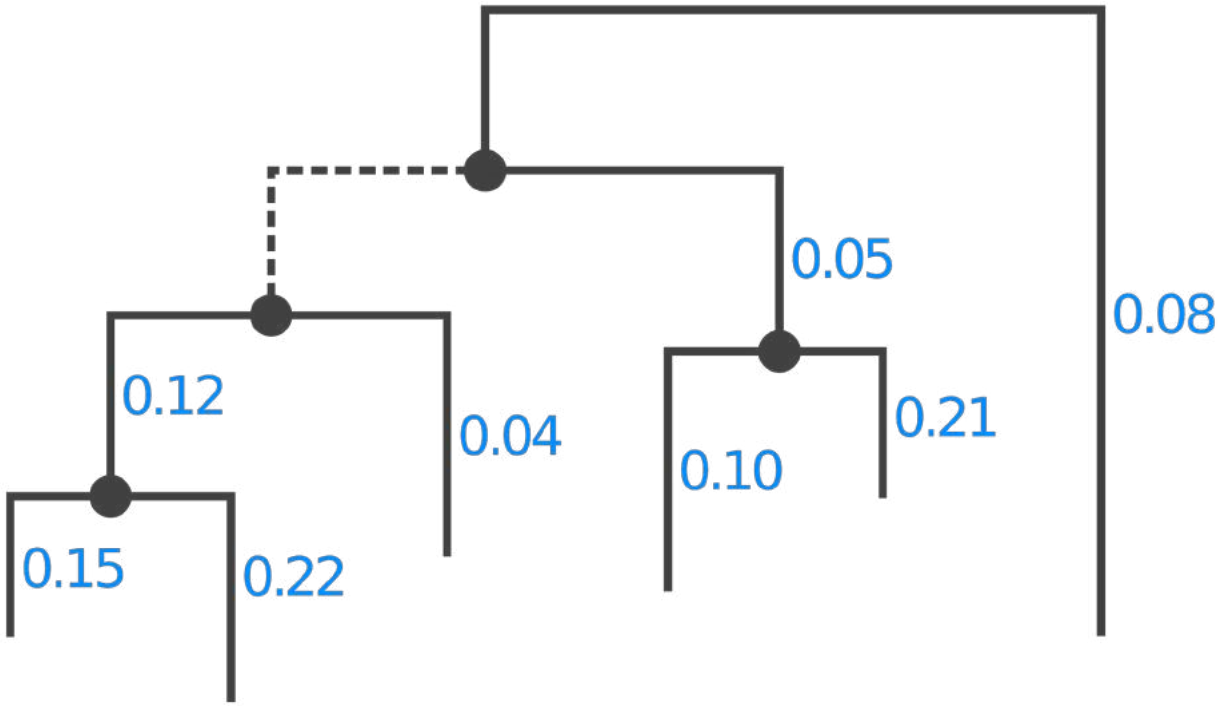
⟶ Adaptation of Balances to Placements



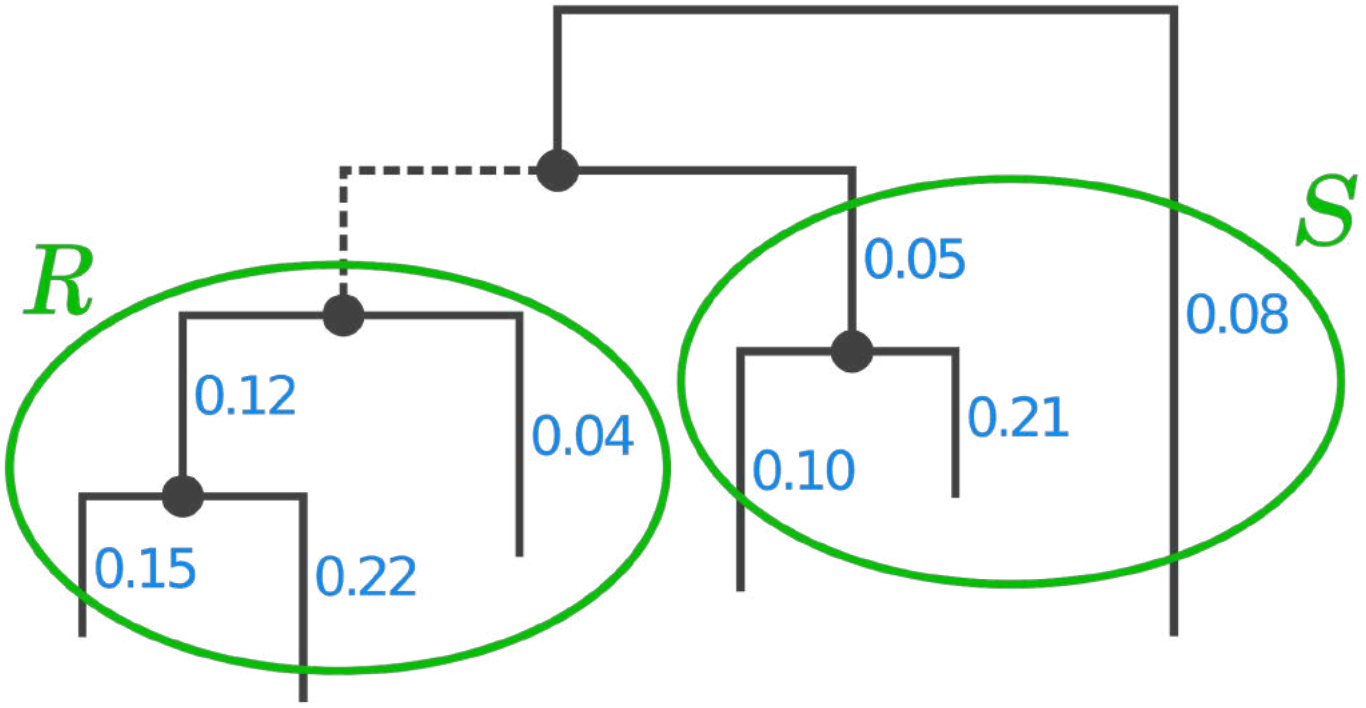
Edge Masses for one Sample



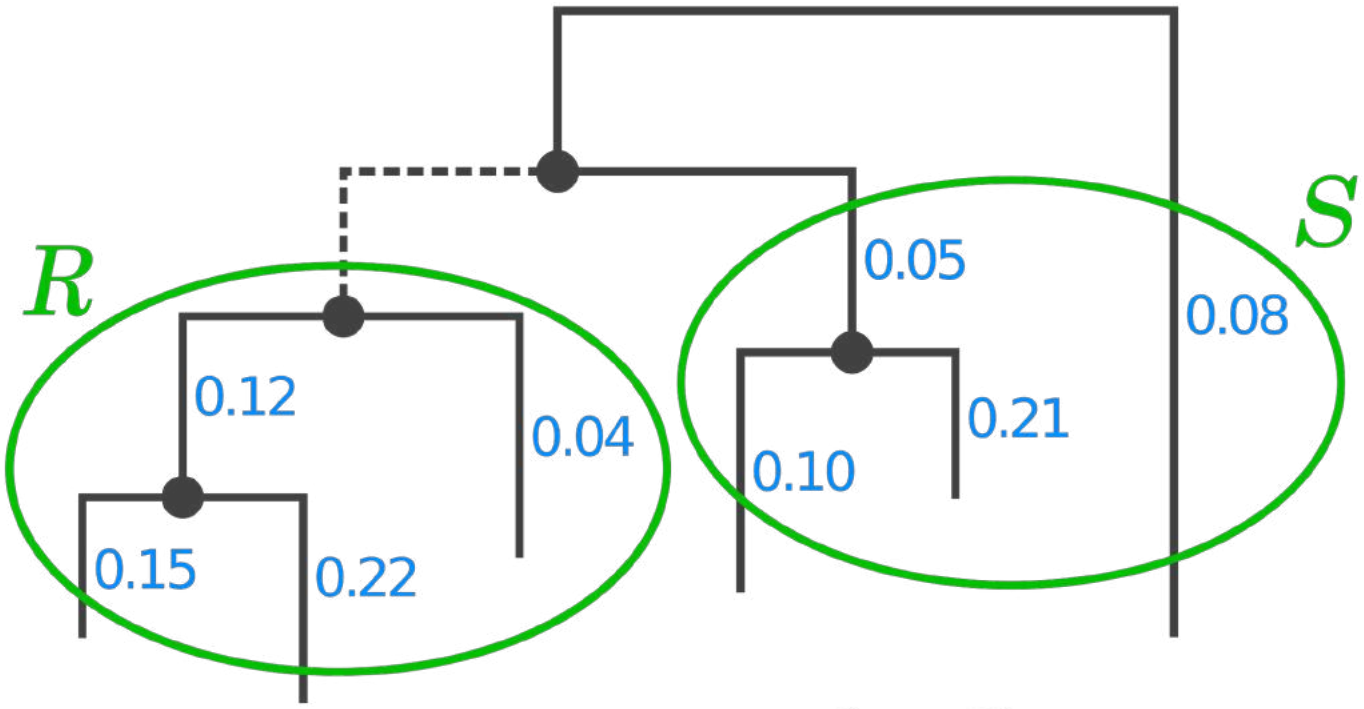
Subtrees induced by an Edge



Summarize the Subtrees

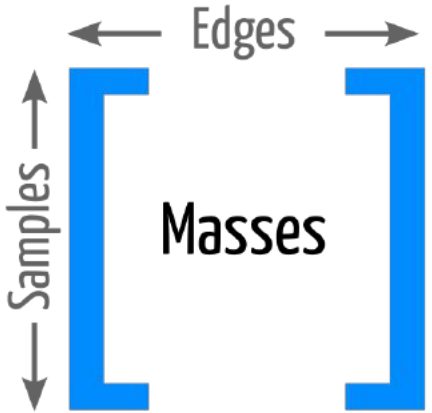


Balance across the Edge



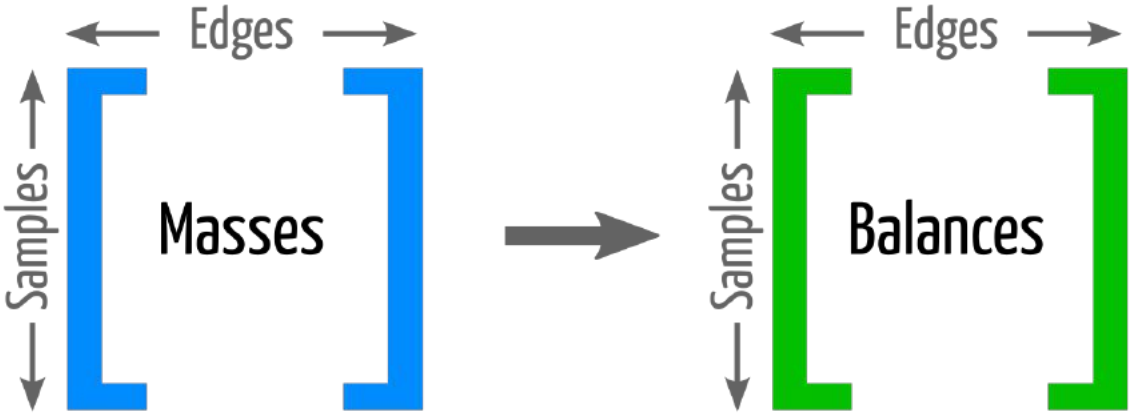
$$\text{balance}(R, S) = \lambda \cdot \log \frac{\text{gm}(R)}{\text{gm}(S)}$$

Masses for all Samples and Edges

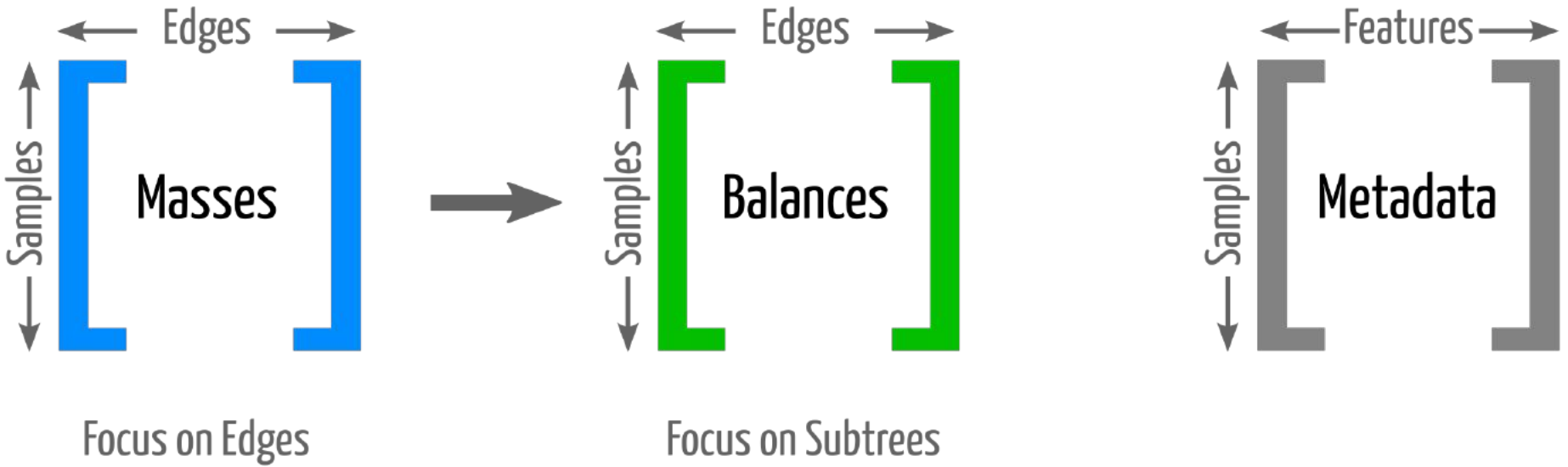


Focus on Edges

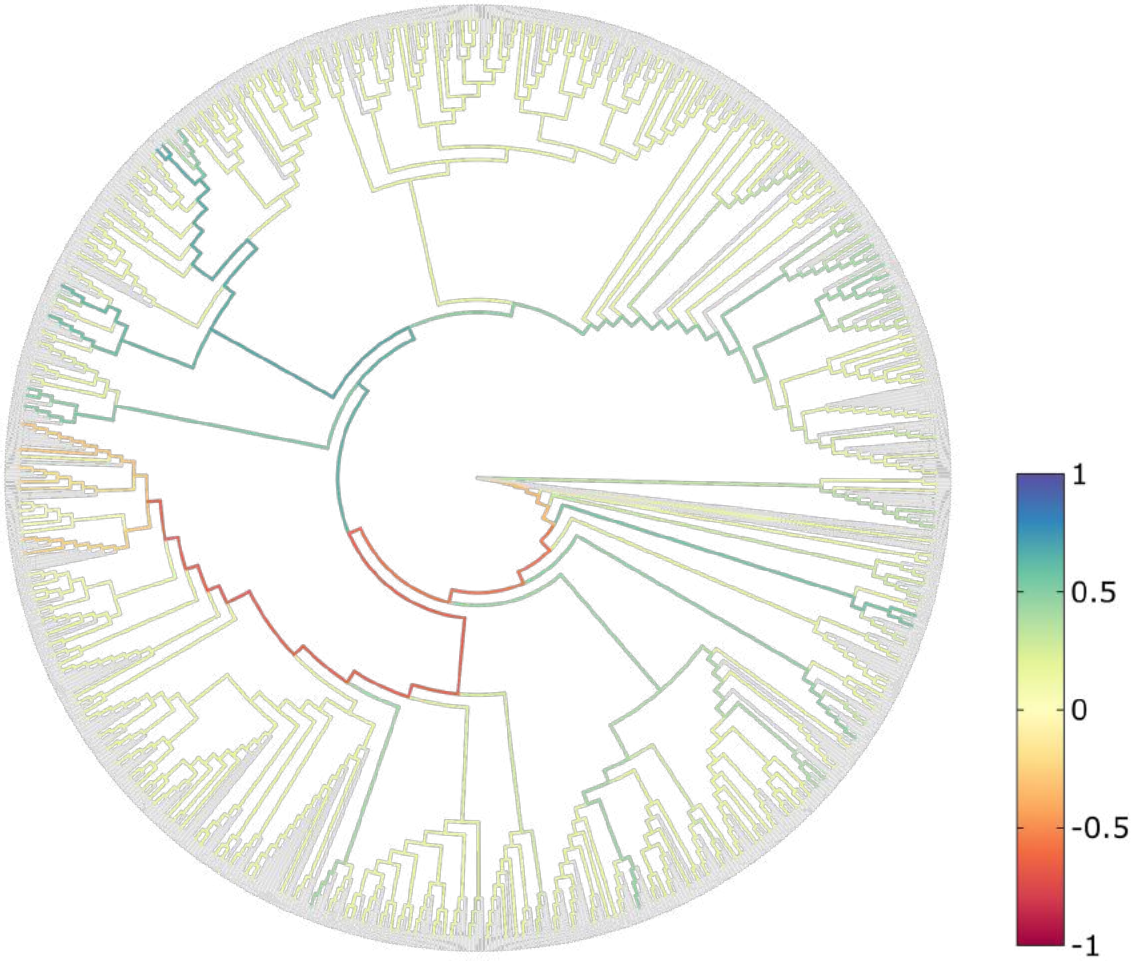
Transformation into Balances



Take Metadata into Account



Correlation between Balances and Metadata

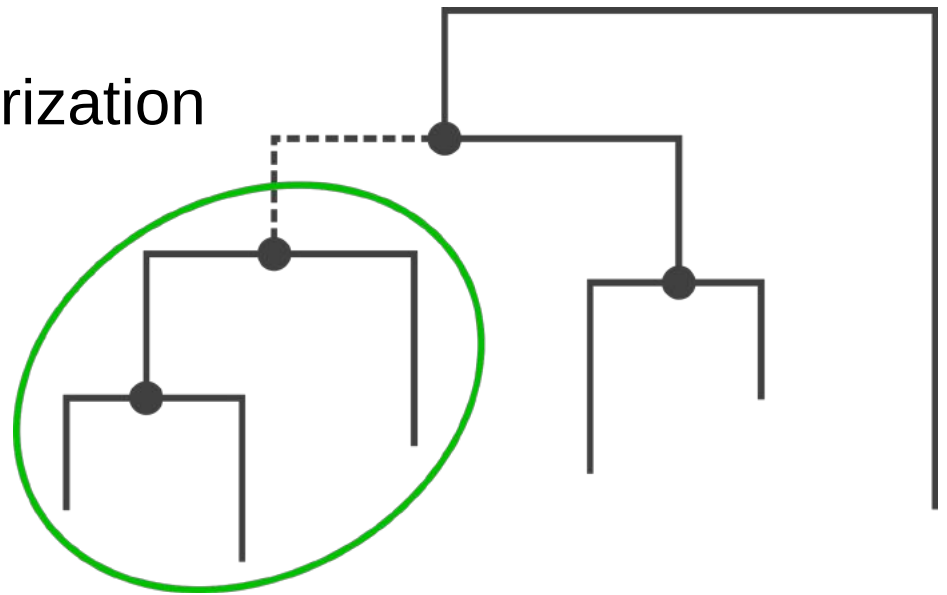


Placement-Factorization

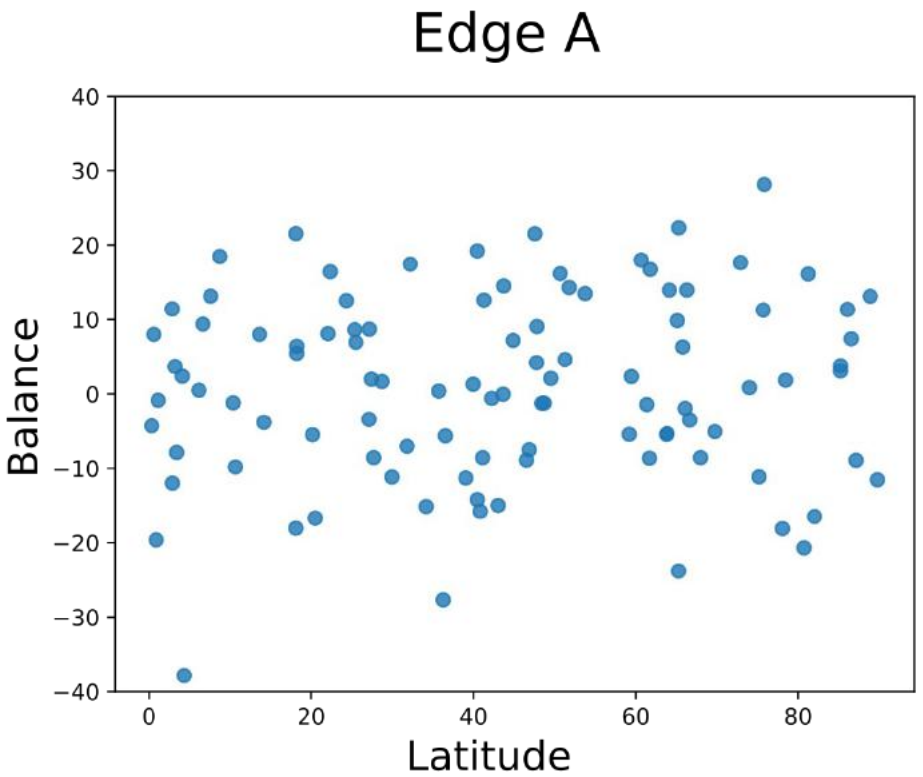
Motivation

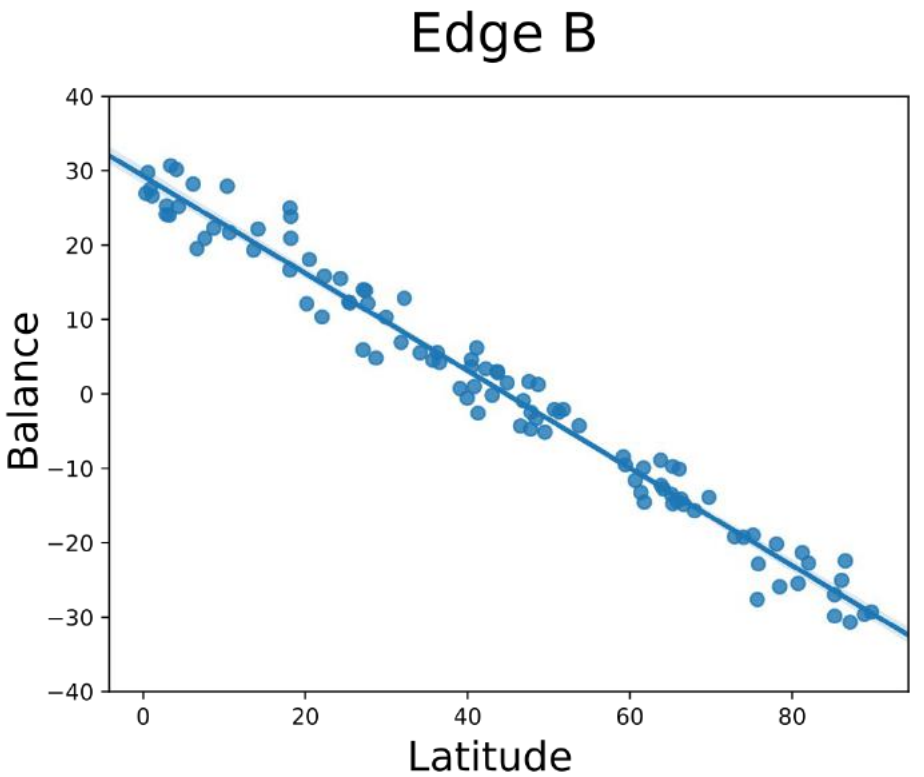
- Balances indicate which groups of organisms are abundant
- Use Balances to find edges / subtrees where abundances change with some metadata feature

⟶ Adaptation of PhyloFactorization to Placements

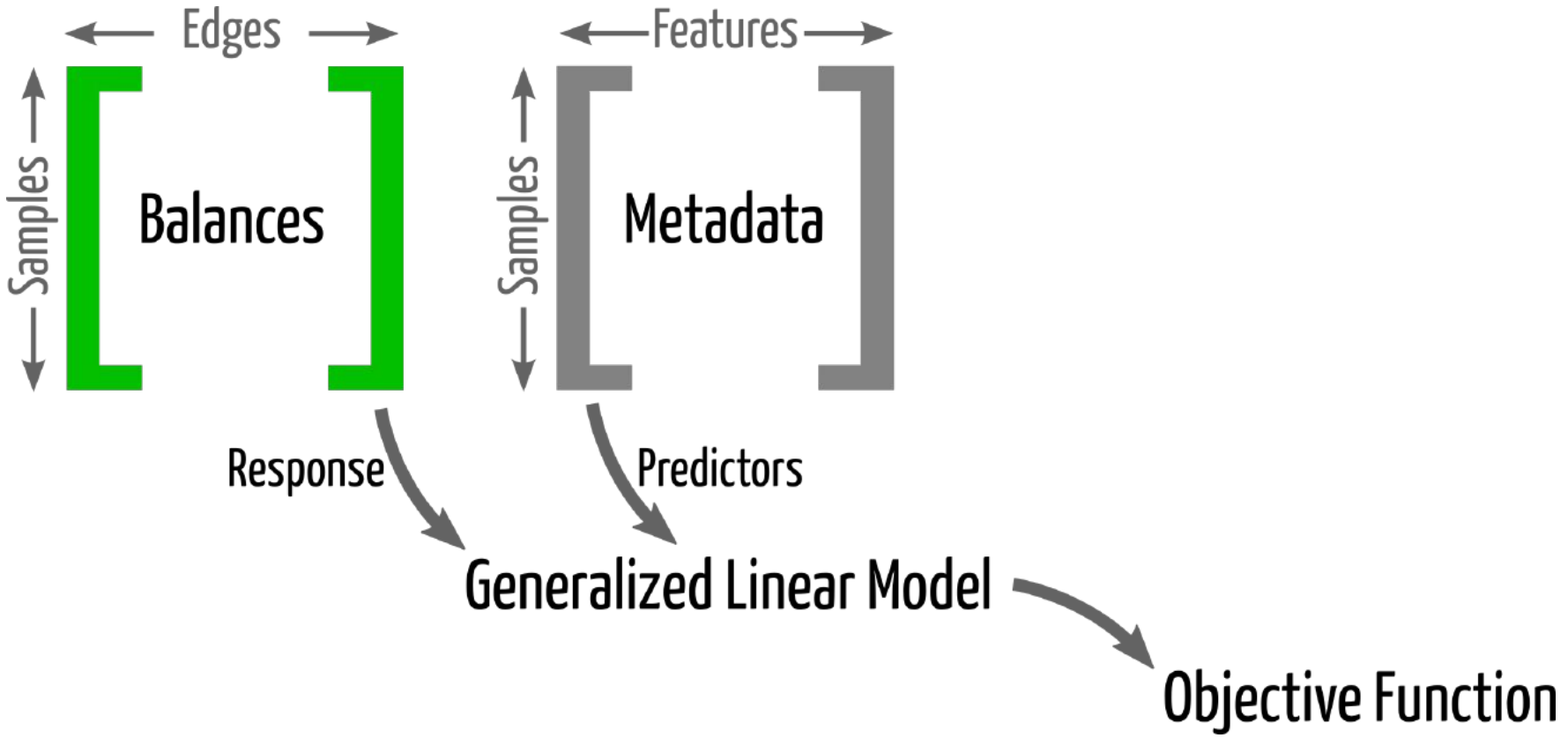


Balances vs. Metadata

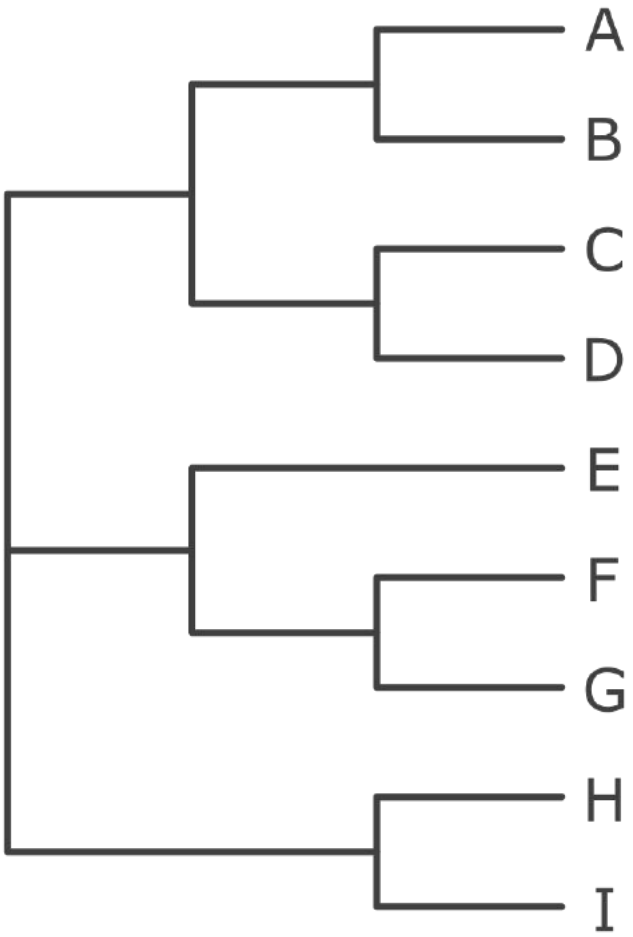




Balances vs. Metadata

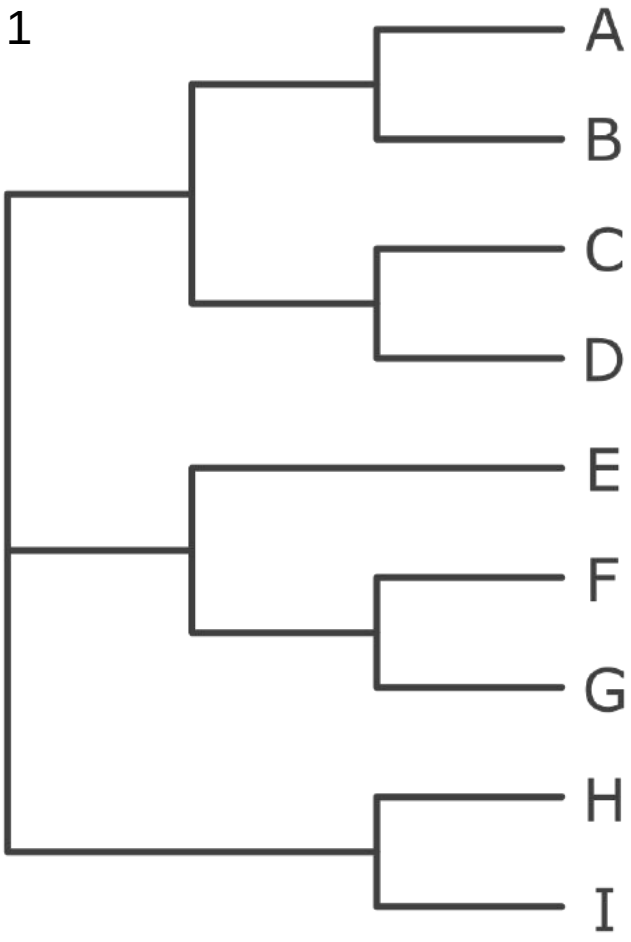


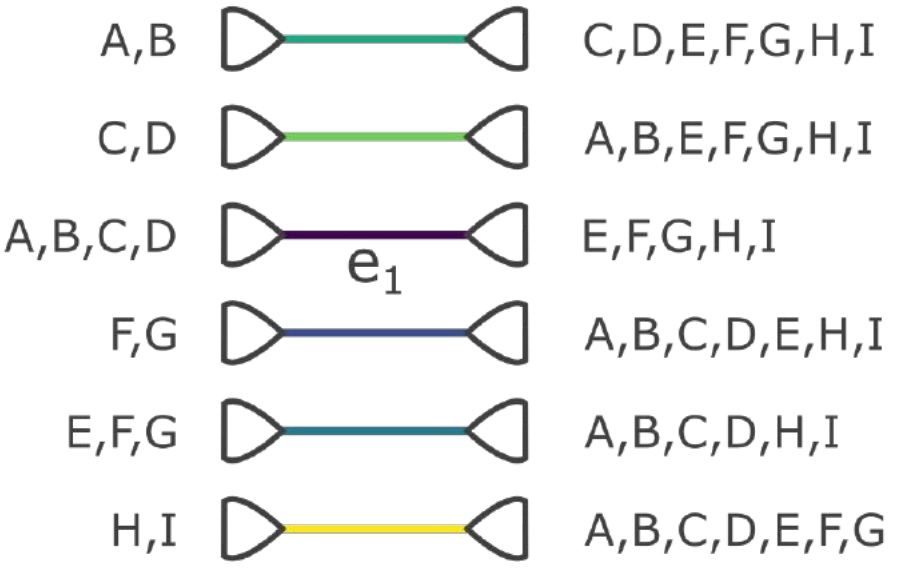
Placement-Factorization

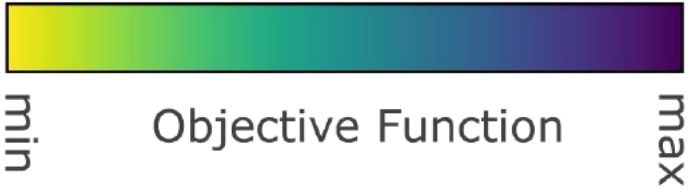


Placement-Factorization

Factor 1

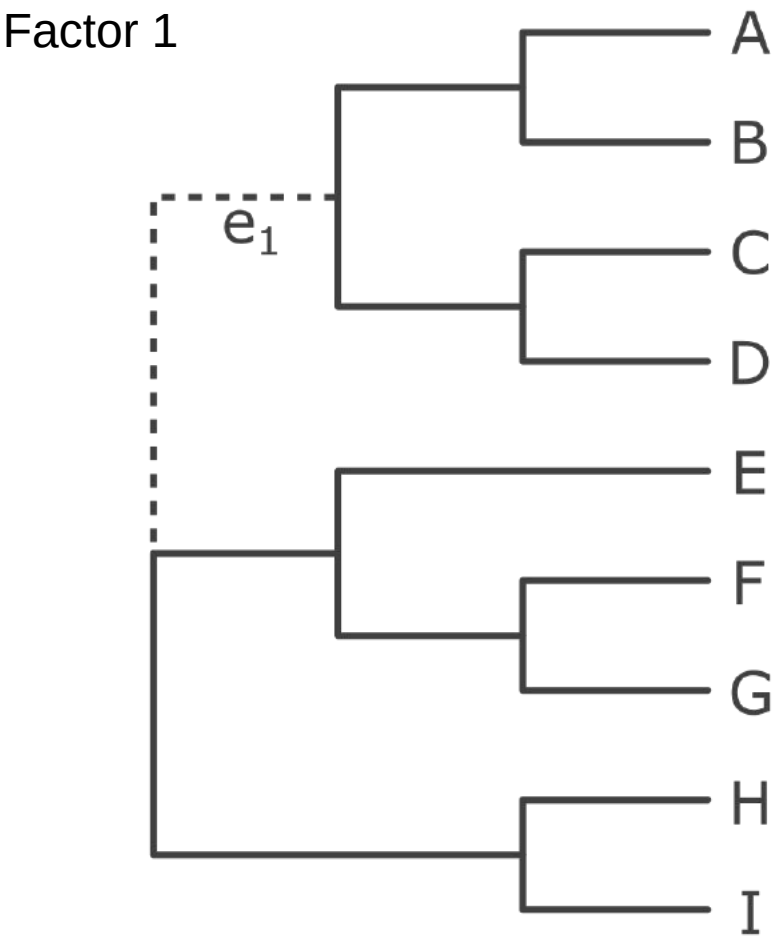


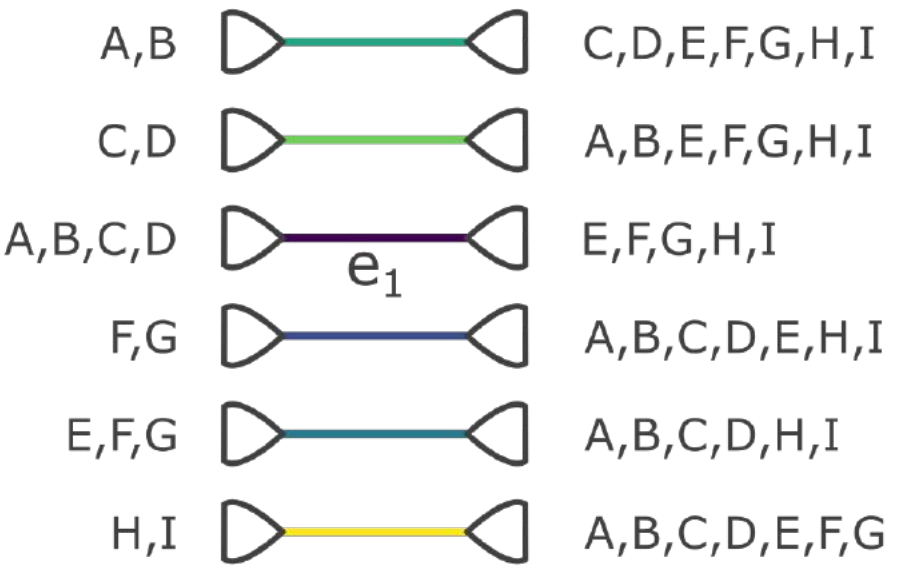


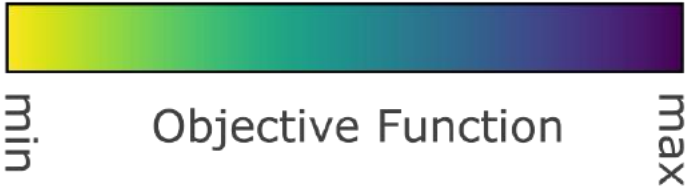


Placement-Factorization

Factor 1

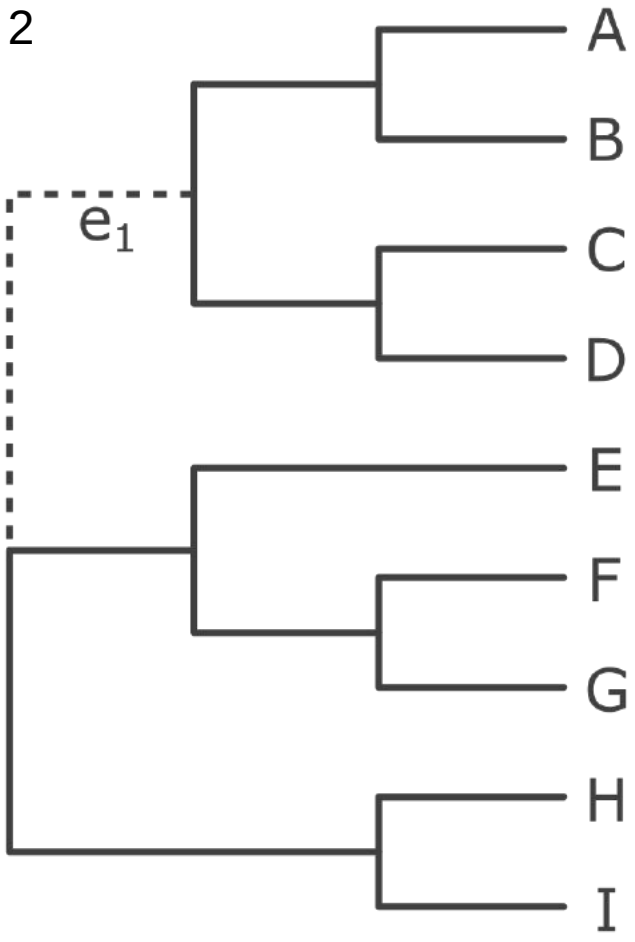


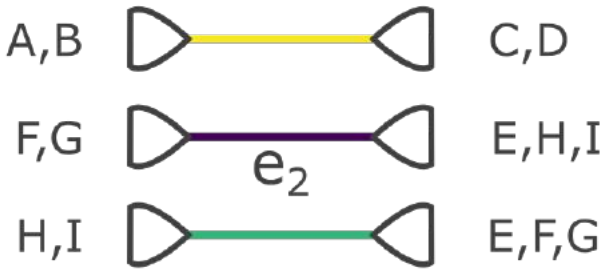


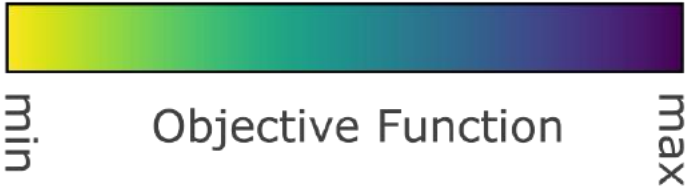


Placement-Factorization

Factor 2

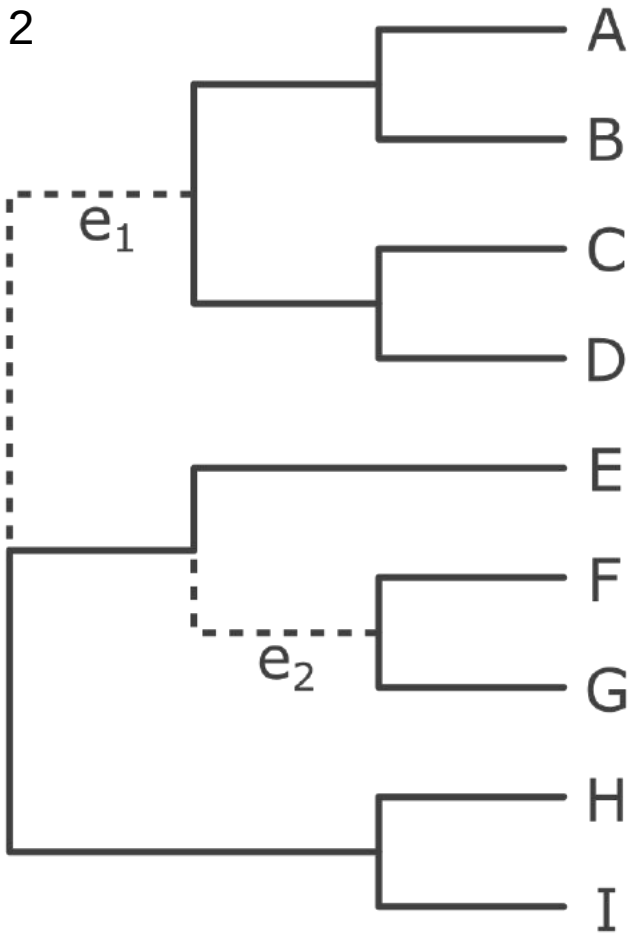


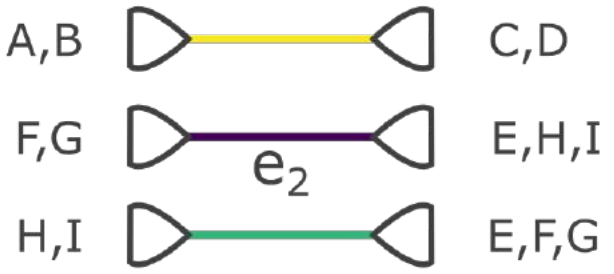


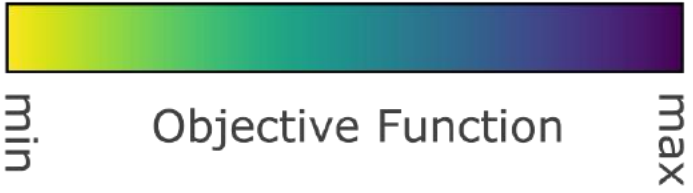


Placement-Factorization

Factor 2

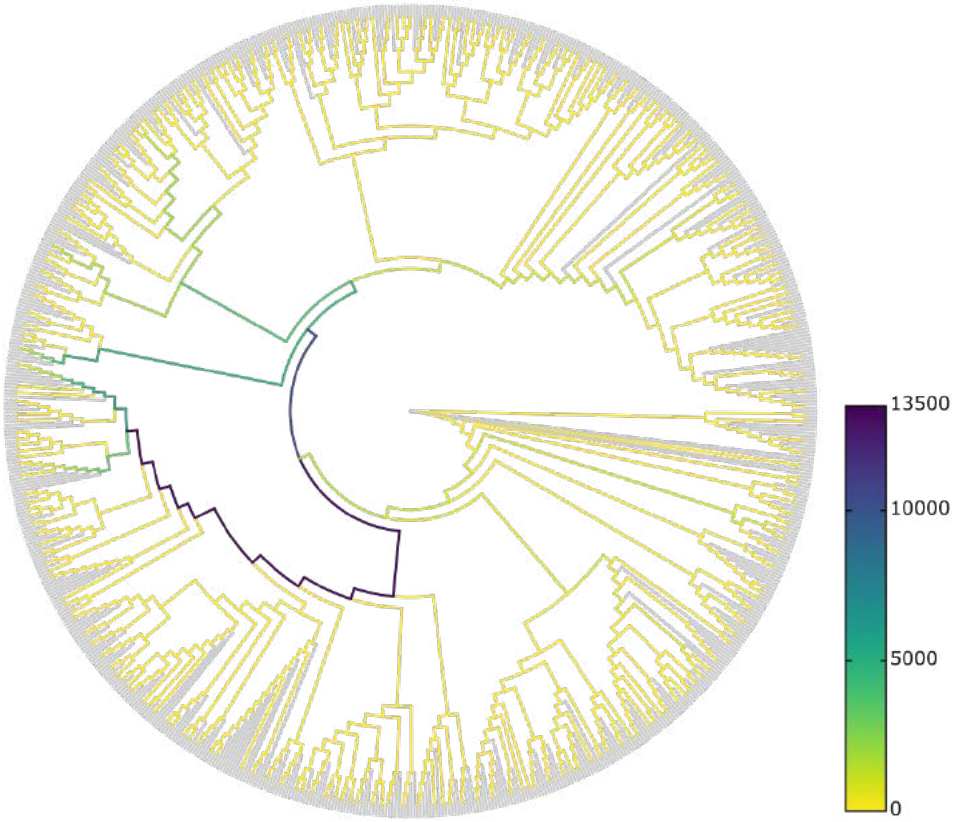




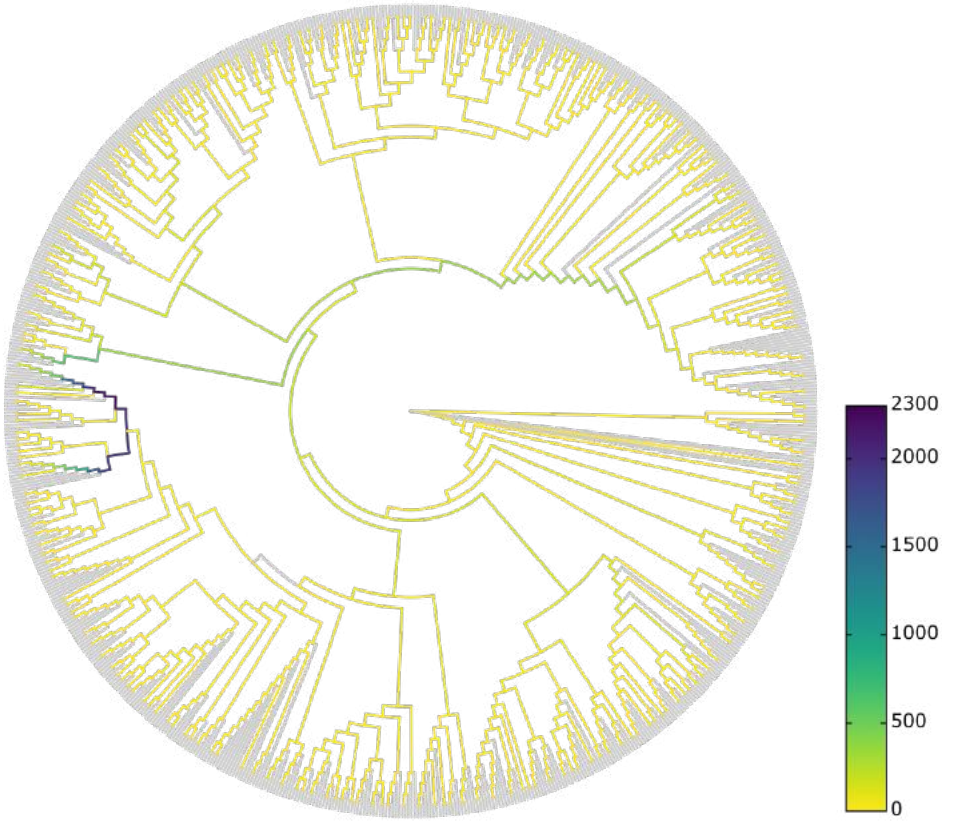


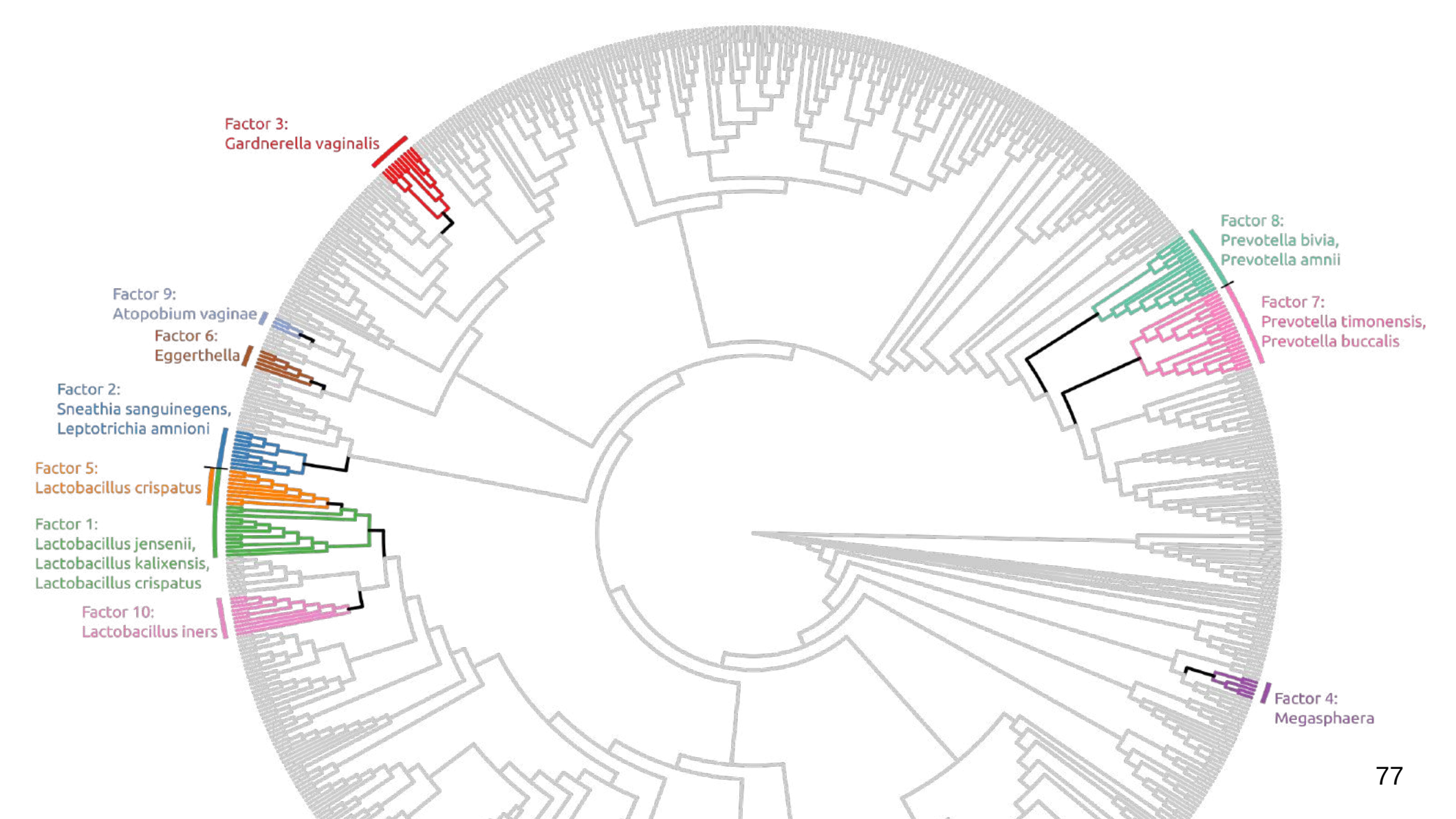
Values of the Objective Function

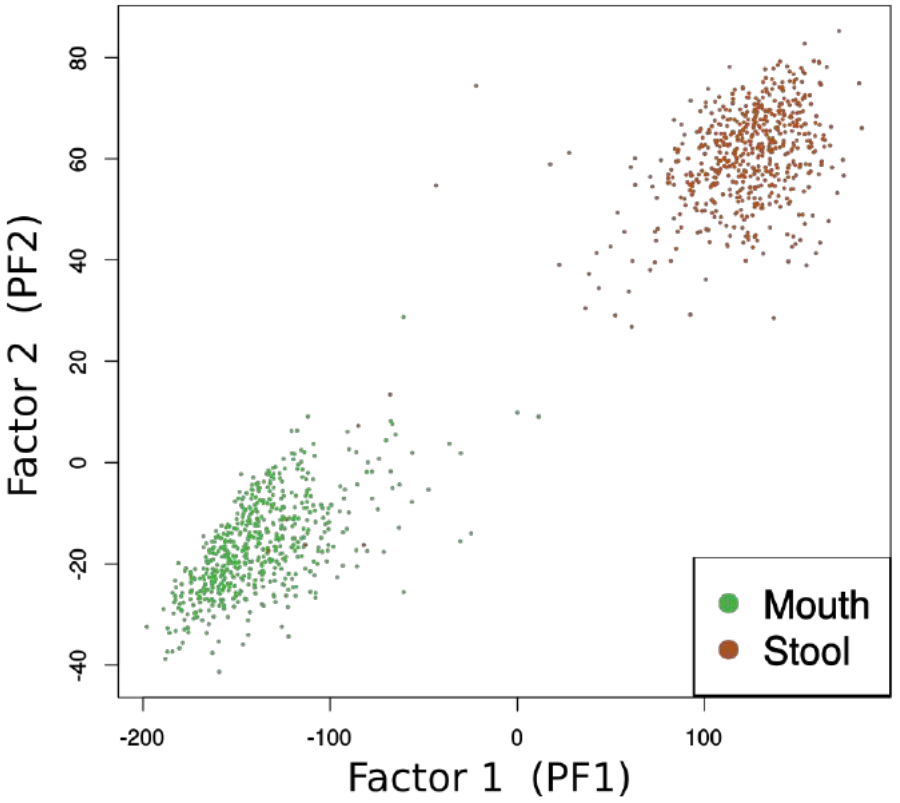
(a) Factor 1 (First Iteration)

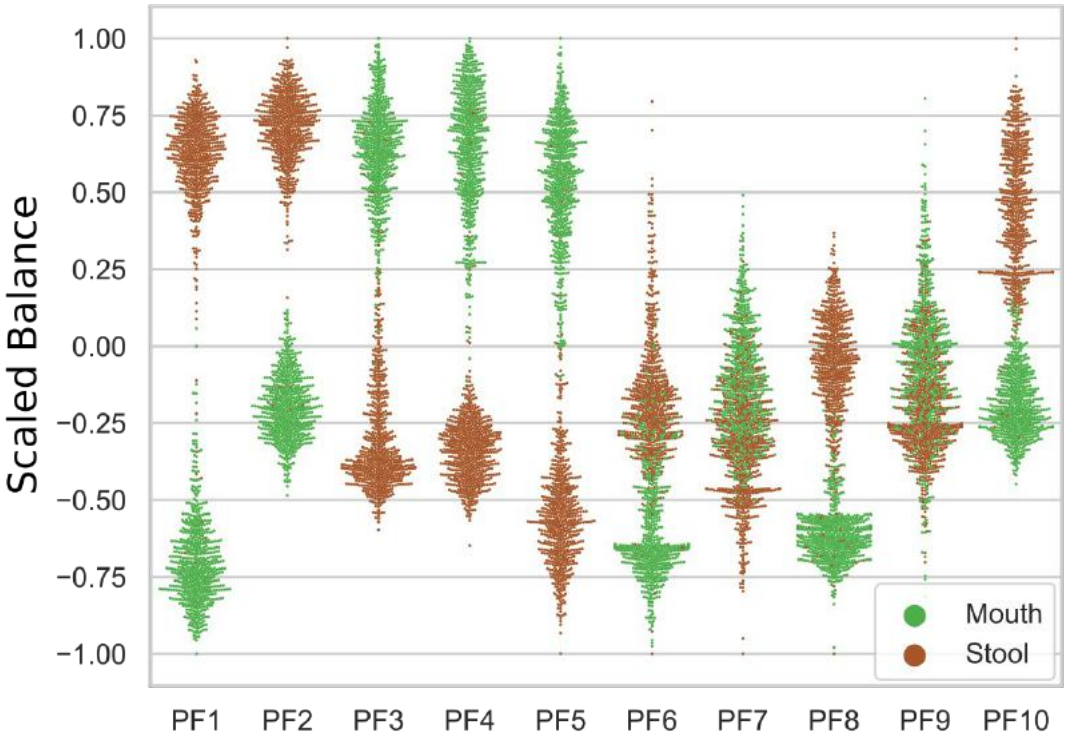


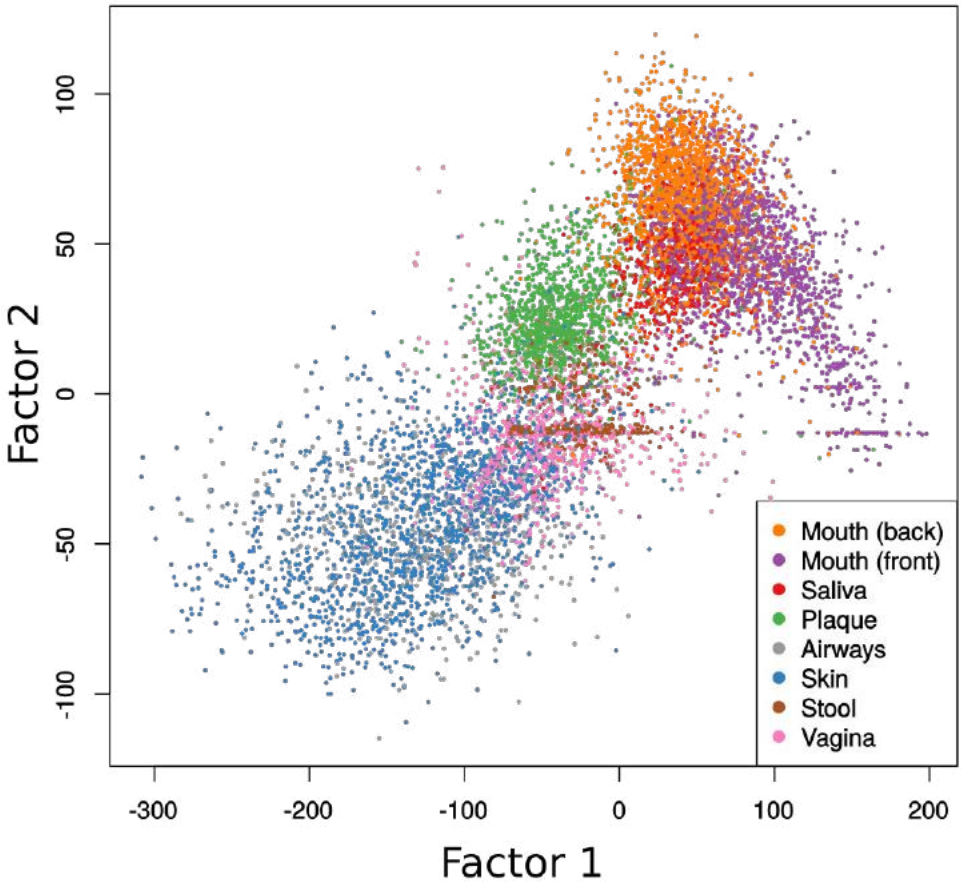
(b) Factor 2 (Second Iteration)

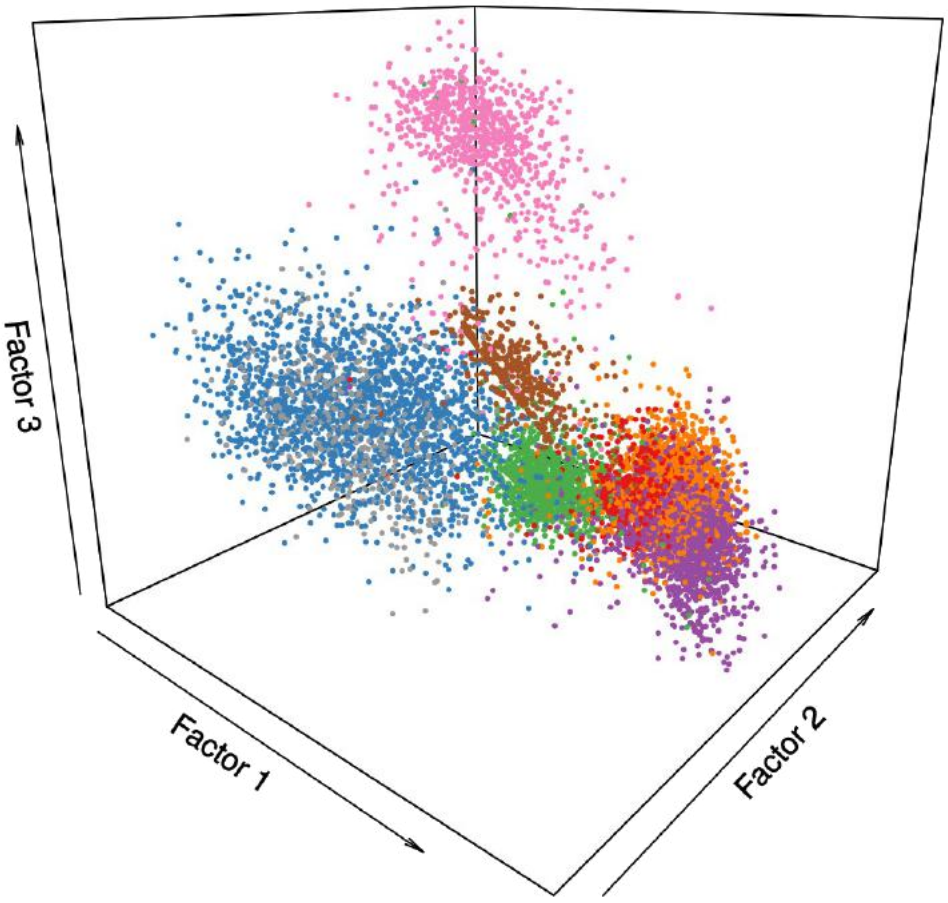



Factor 3:
Gardnerella vaginalis
Factor 8:
Prevotella bivia,
Prevotella amnii
Factor 9:
Atopobium vaginae
Factor 7:
Prevotella timonensis,
Prevotella buccalis
Factor 6:
Eggerthella
Factor 2:
Sneathia sanguinegens,
Leptotrichia amnioni
Factor 5:
Lactobacillus crispatus
Factor 1:
Lactobacillus jensenii,
Lactobacillus kalixensis,
Lactobacillus crispatus
Factor 10:
Lactobacillus iners
Factor 4:
Megasphaera


Factor 2 (PF2)
80
60
40
20
0
-20
-40
-200
-100
0
100
Factor 1 (PF1)
Mouth
Stool

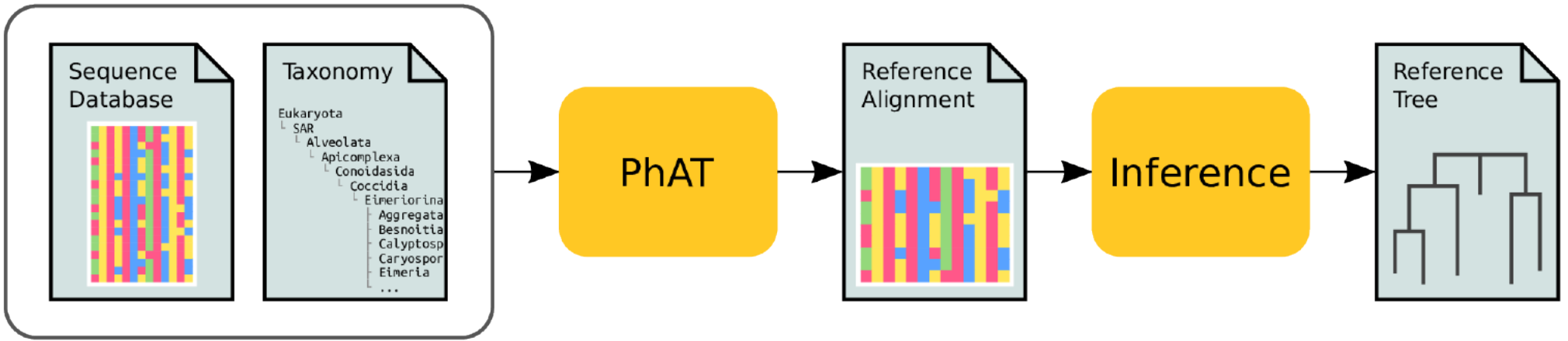

Scaled Balance
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
PF1
PF2
PF3
PF4
PF5
PF6
PF7
PF8
PF9
PF10
Mouth
Stool


Factor 2
100
50
0
-50
-100
-300
-200
-100
0
100
200
Factor 1
Mouth (back)
Mouth (front)
Saliva
Plaque
Airways
Skin
Stool
Vagina

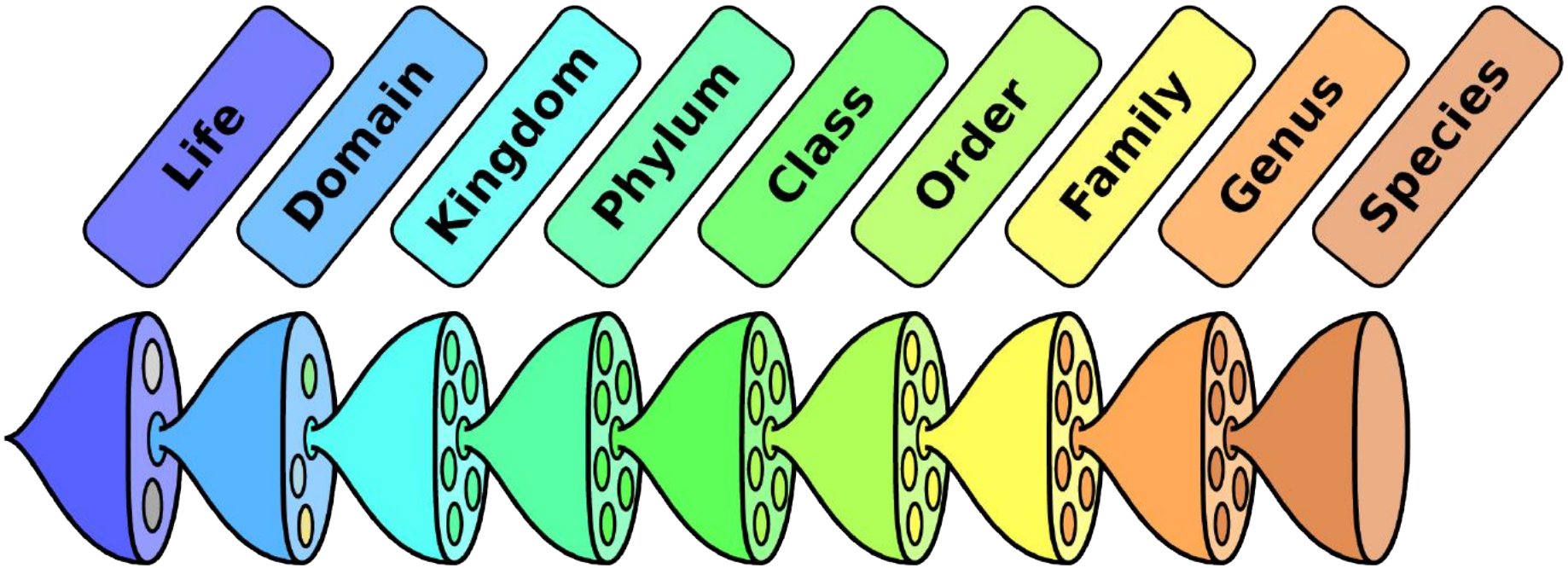

Factor 3
Factor 1
Factor 2

PhAT

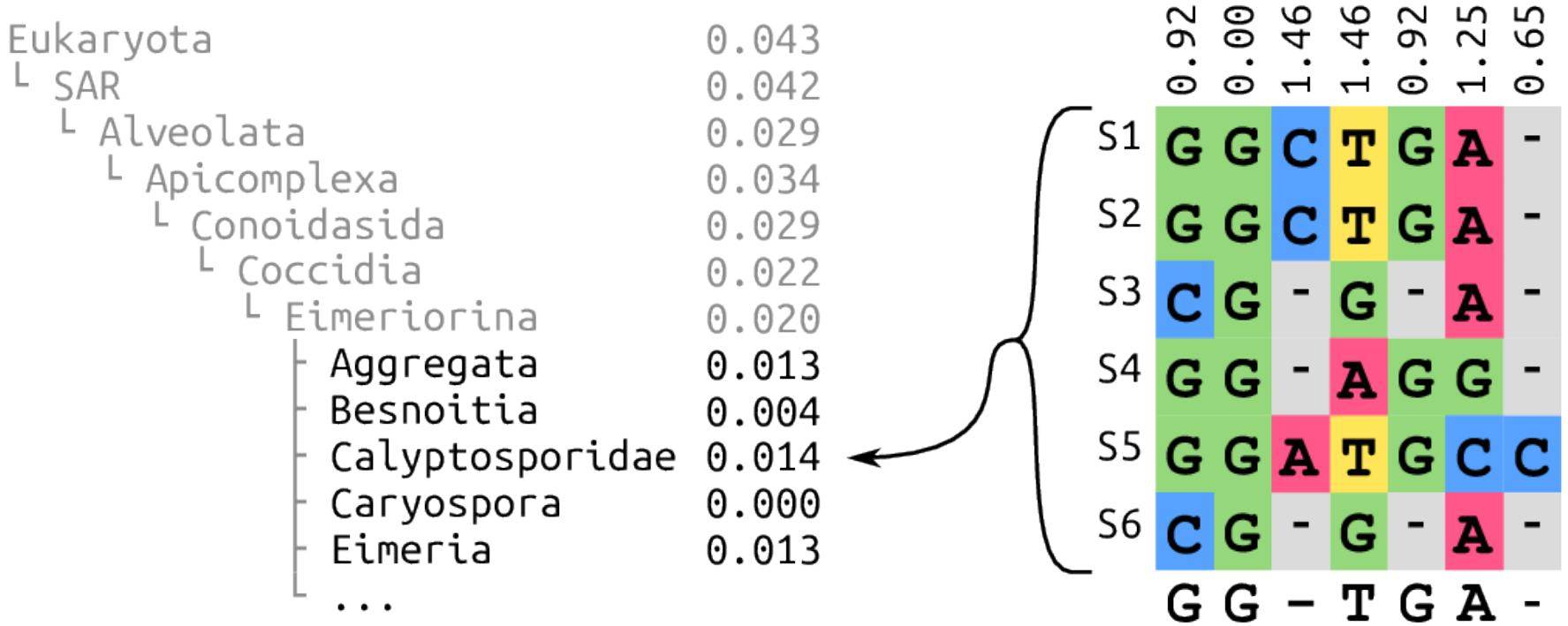
Phylogenetic Automatic Reference Trees



Taxonomy

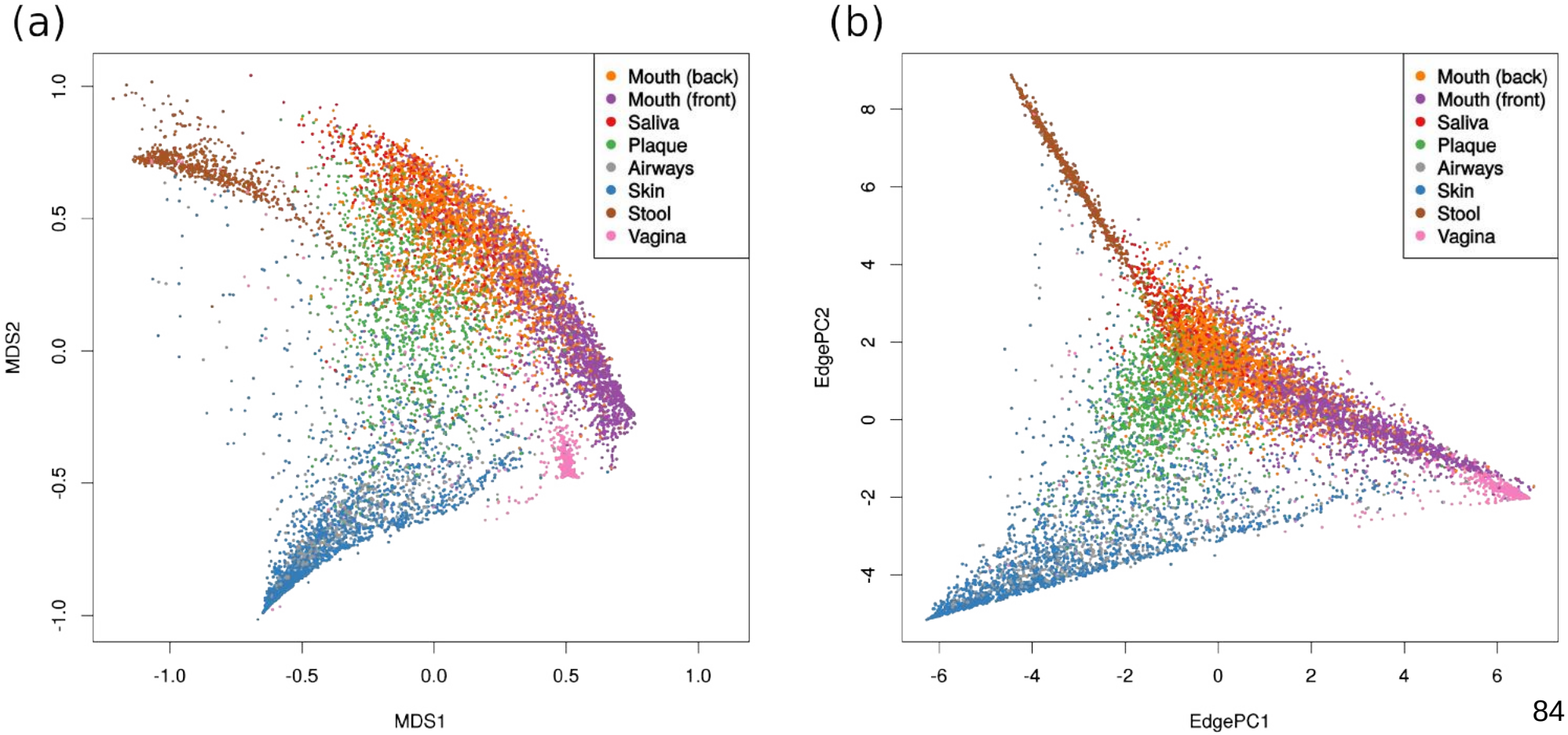

Life
Domain
Kingdom
Phylum
Class
Order
Family
Genus
Species

Phylogenetic Automatic Reference Trees



$$H_i = -\sum_c f_{c,i} \cdot \log f_{c,i}$$

Phylogenetic Automatic Reference Trees



Pipelines

Preprocessing Pipeline