Phylogenetic Placement: Computation, Analysis, and Visualization

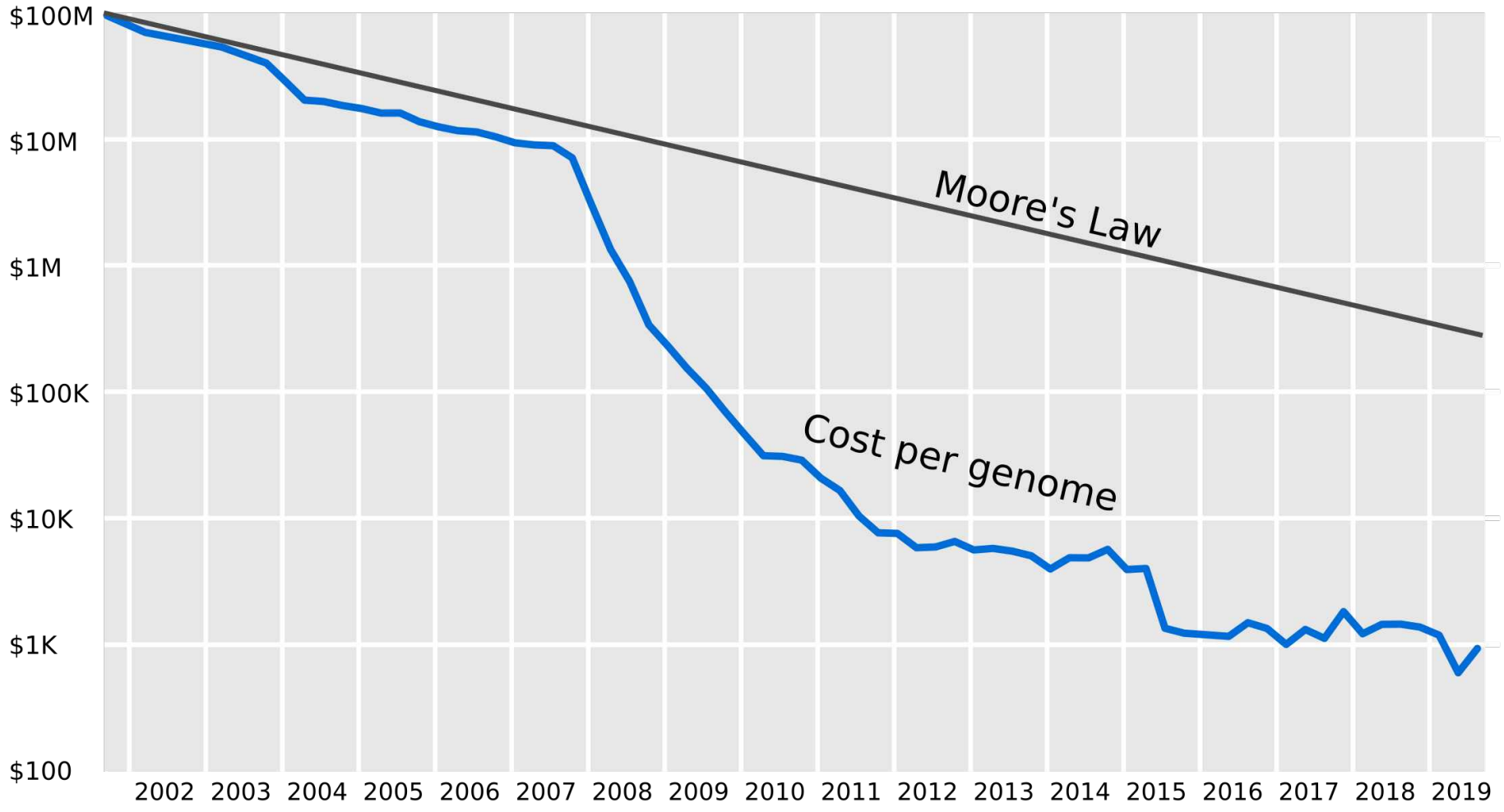
Lucas Czech
Carnegie Institution for Science
Stanford, USA

2023-04-20
Guest Lecture
University of Oslo

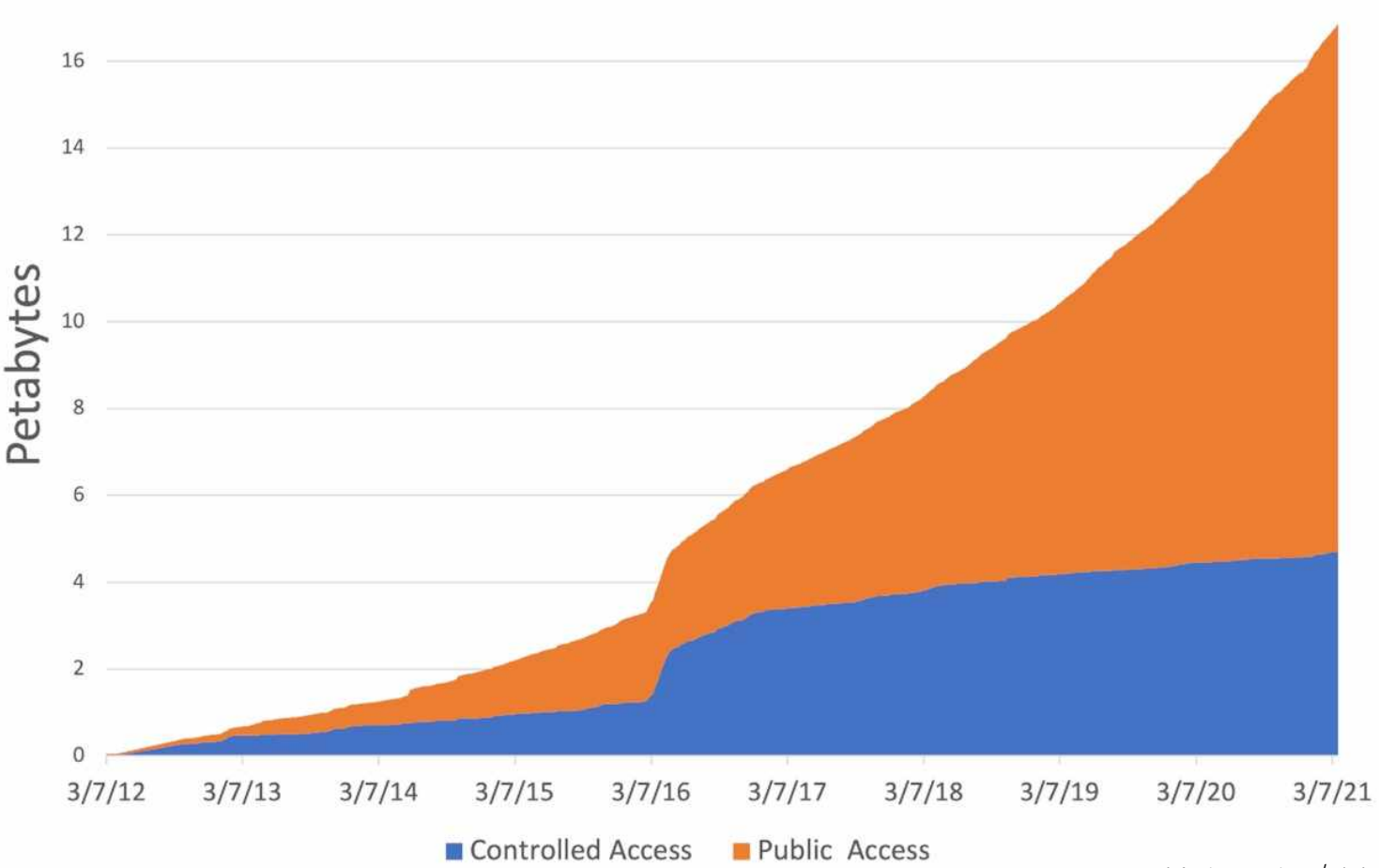
Overview

- Motivation
- Phylogenetic Tree Inference
- Phylogenetic Placement
- Placement Analysis and Visualization

Motivation

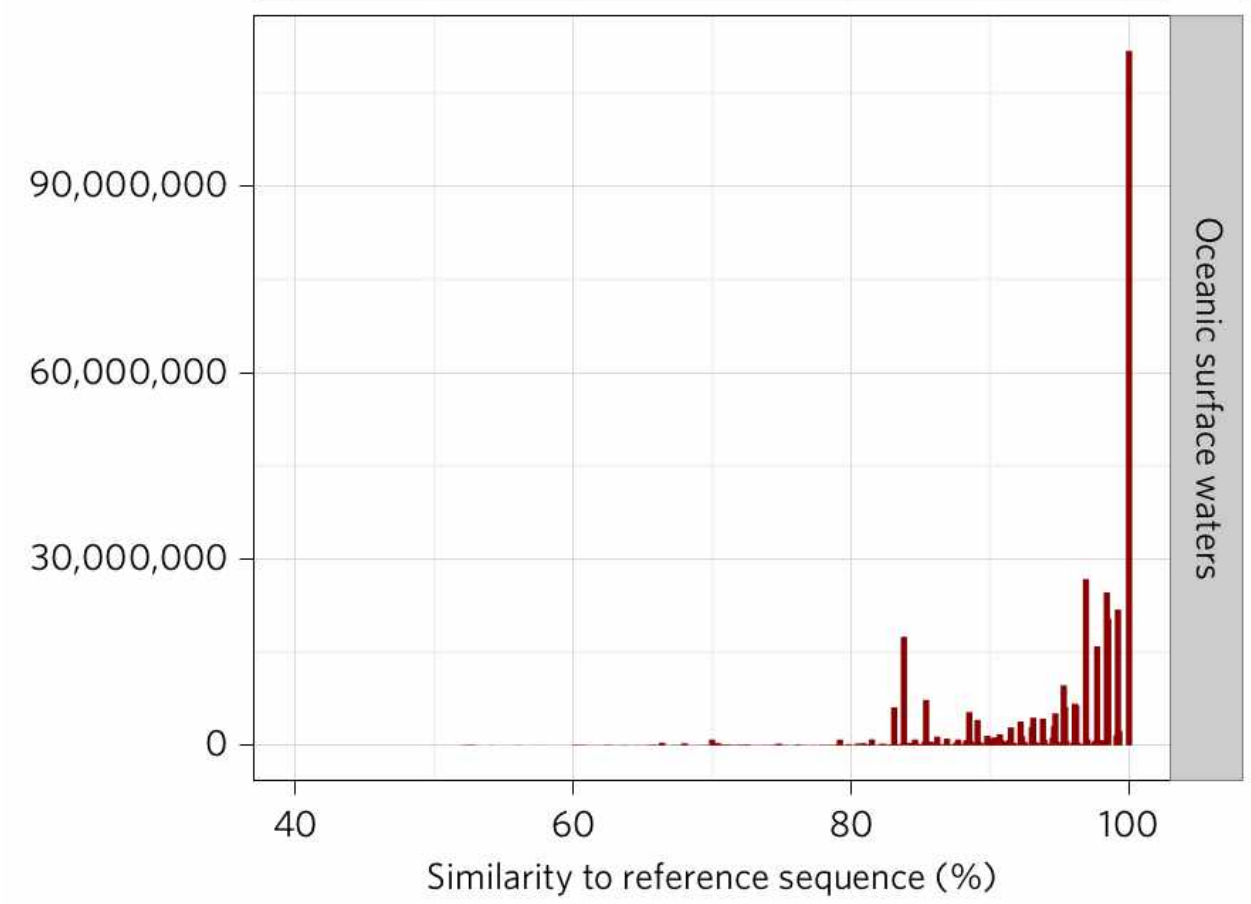

$100M
$10M
$1M
$100K
$10K
$1K
$100
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
Moore's Law
Cost per genome

Sequence Database Growth (SRA)



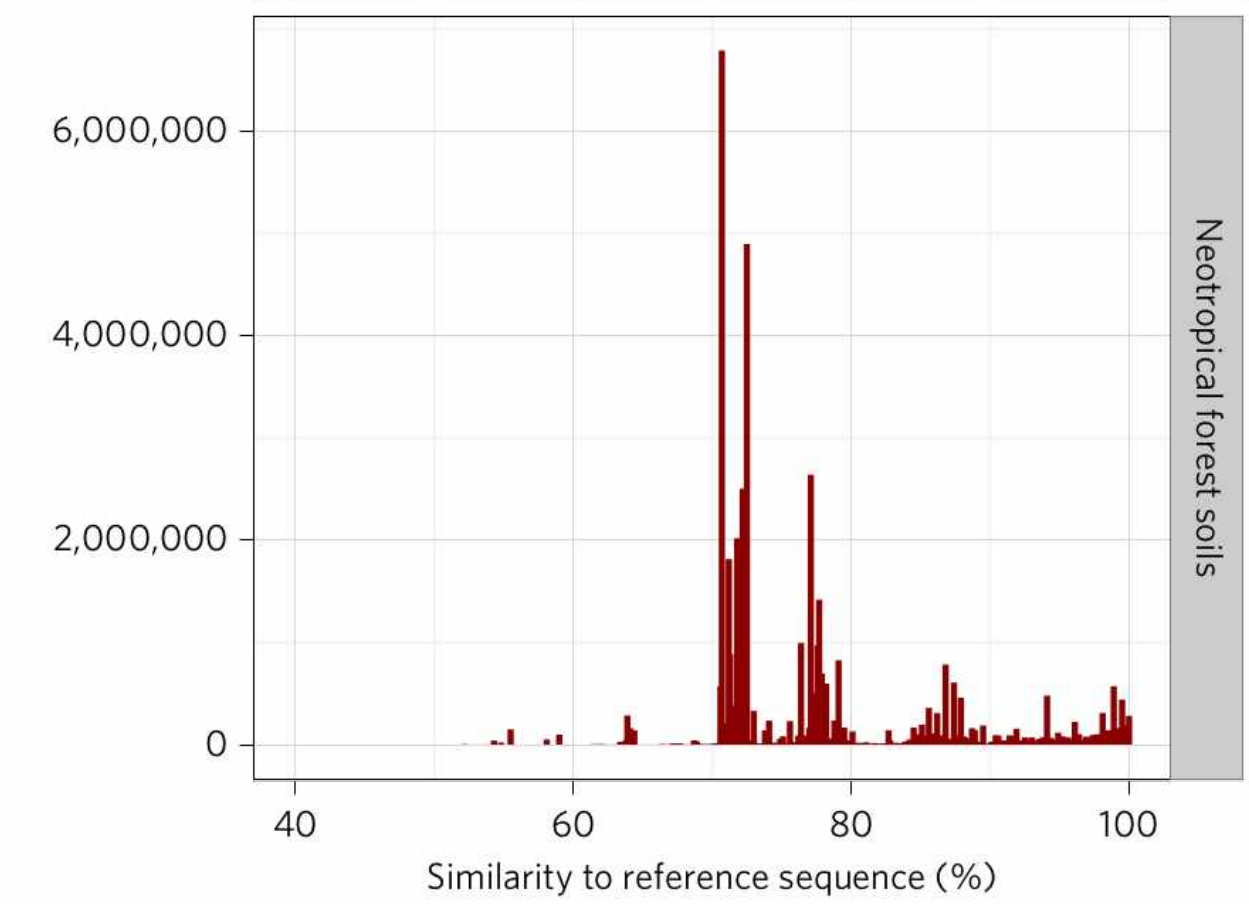
Katz *et al*, Nuc. Ac. Res., 2021

(Almost) complete reference database



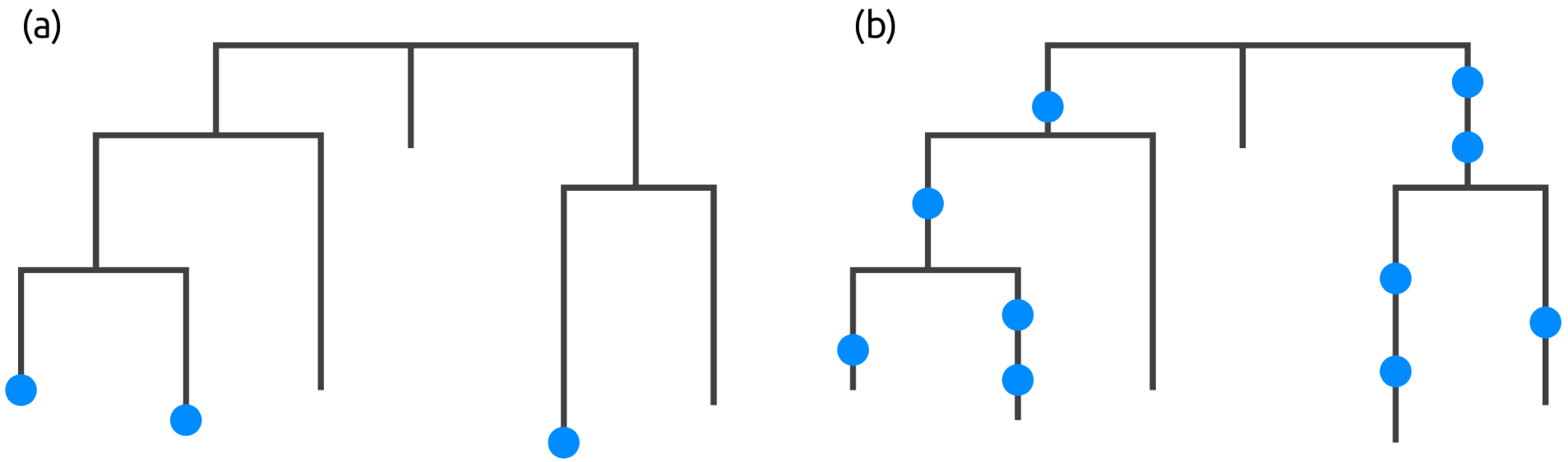
Mahé *et al*, Nature E&E, 2017

Incomplete reference database



Mahé *et al*, Nature E&E, 2017

BLAST / vsearch vs. Phylogenetic Placements



Research Questions

- “Who lives there?” - Microbial composition of samples
- How do samples differ from each other?
- Which environmental factors drive these differences?

Some Applications

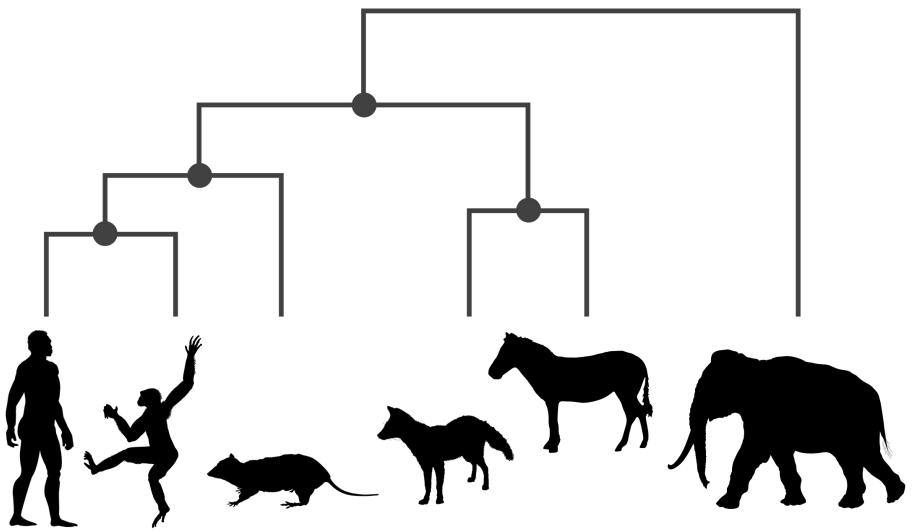
- Data cleaning and retention (Mahé et al., 2017)
- Inference of new clades (Dunthorn et al., 2014; Bass et al., 2018)
- Estimation of ecological profiles (Keck et al., 2018)
- Identification of low-coverage genomes of viral strains (Mühlemann et al., 2020)
- Phylogenetic analysis of viruses such as SARS-CoV-2 (Morel et al., 2020; Turakhia et al., 2021)
- Clinical studies of microbial diseases (Srinivasan et al., 2012)

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 phylogenetic tree:
maximize the likelihood of producing 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!

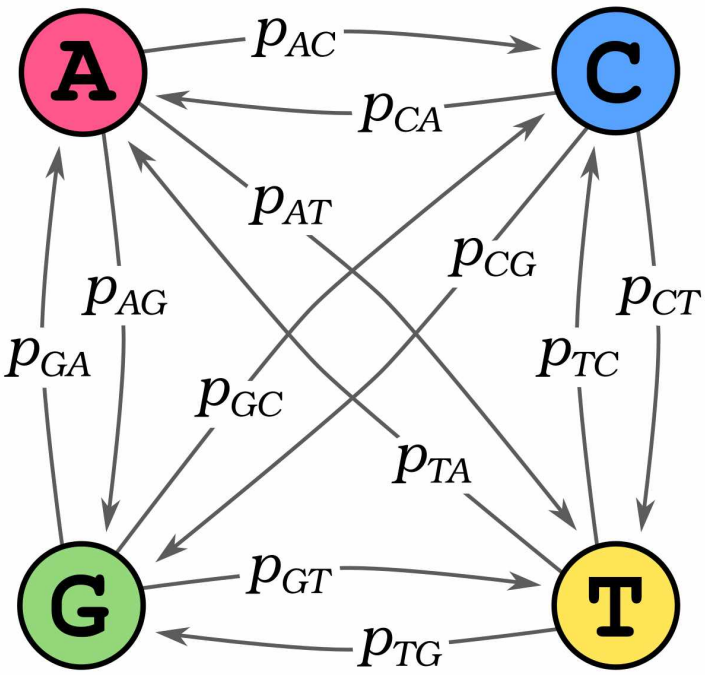
- Many heuristics and methods developed over the years
  - Greedy hill-climbing from a (reasonable) starting tree
  - Felsenstein Pruning Algorithm
  - ...

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 loci (columns of the MSA), we only consider mutations (no insertions or deletions)
  - 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

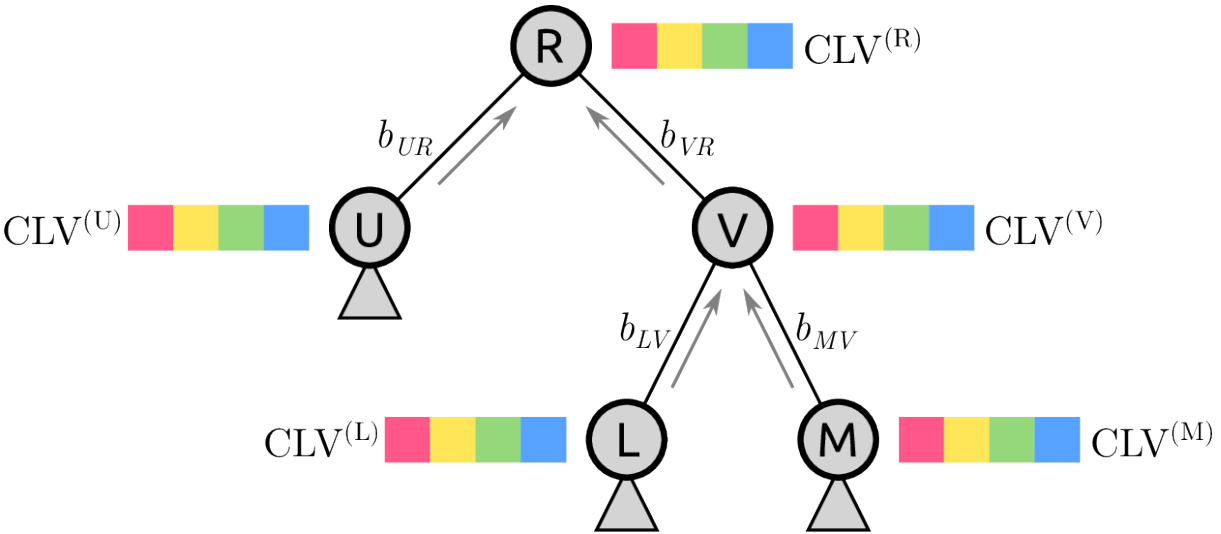
  → Sum over probabilities of every possible states

  → 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
- 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

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

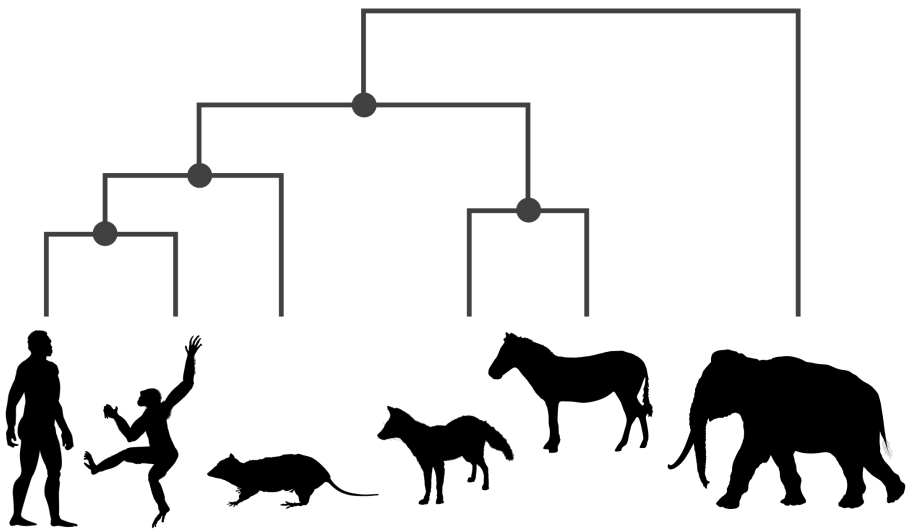
- Still need to optimize branch lengths → numerical method!
- Then, “simply” repeat for every possible tree topology to find the most likely tree :-)
- But: Number of possible trees grows over-exponentially with number of taxa! :-(

Phylogenetic Placement

MSA and Phylogenetic Tree

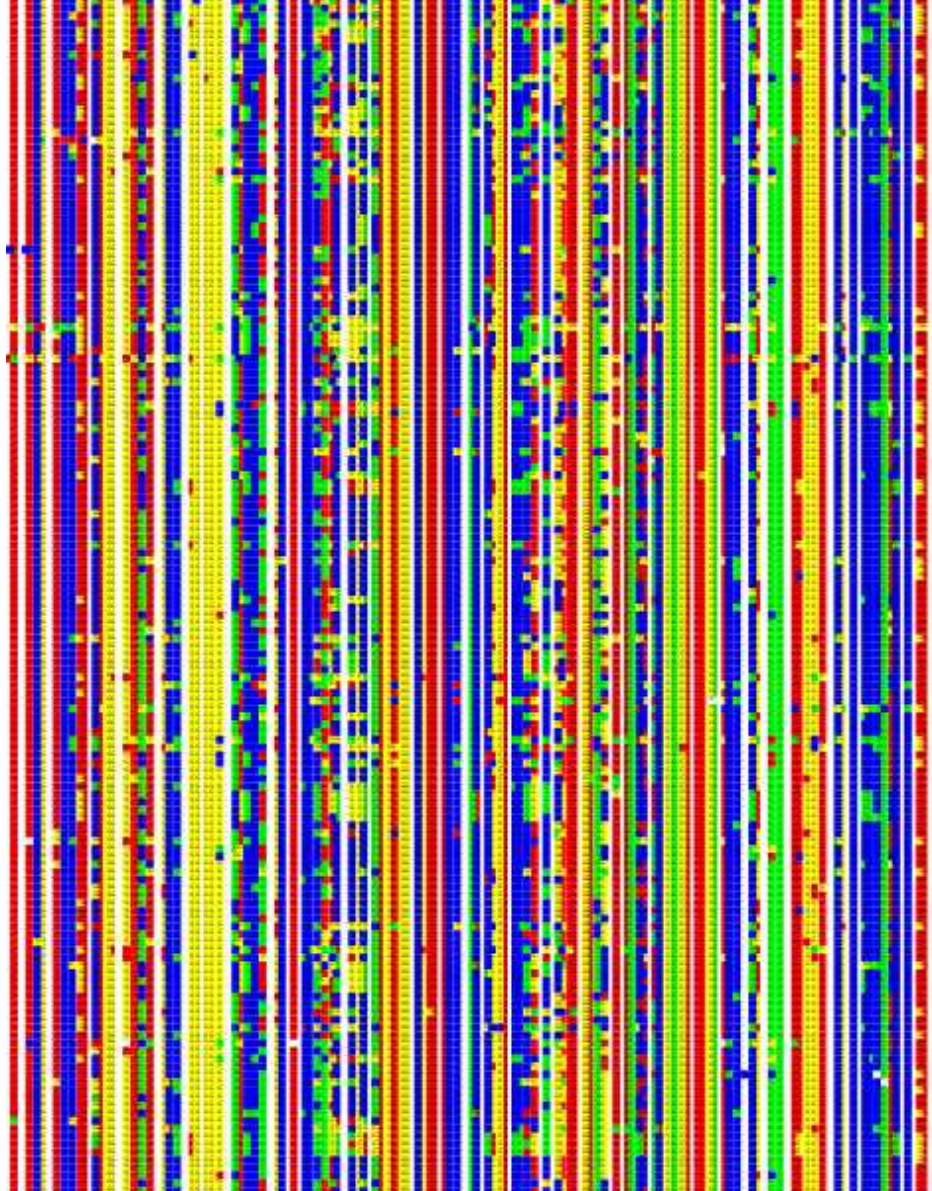
Multiple Sequence Alignment (MSA)

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

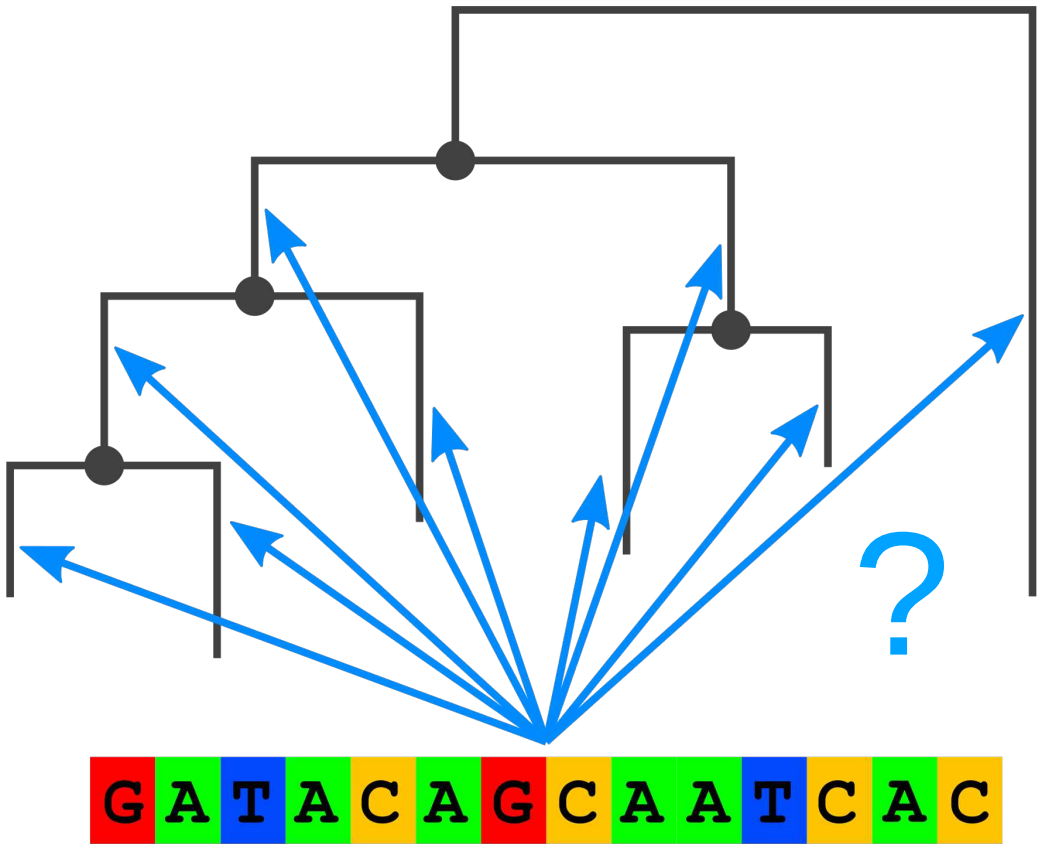


Multiple Sequence Alignment (MSA)

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



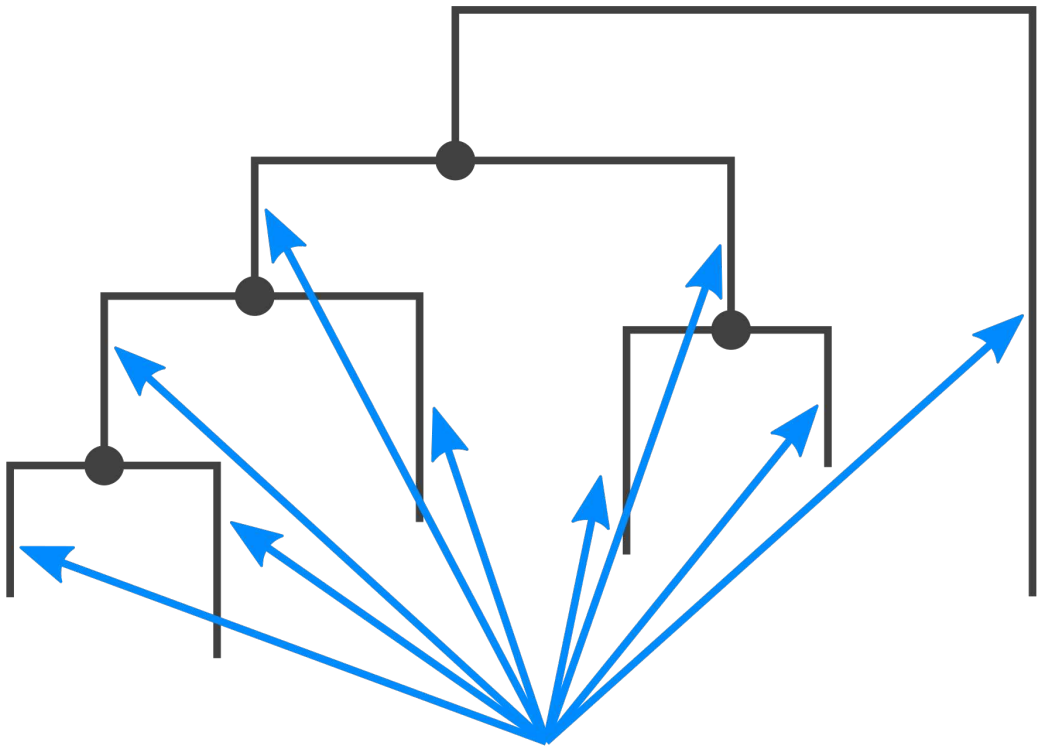
Phylogenetic Placement

- Given:
  - *Reference tree* and MSA
  - Set of *query sequences*
- 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

- Typically: query sequences are reads from a sequencing machine
- Have to align them to the given MSA first
- Dedicated tools for aligning queries to a given MSA:
  - 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

Phylogenetic Placement

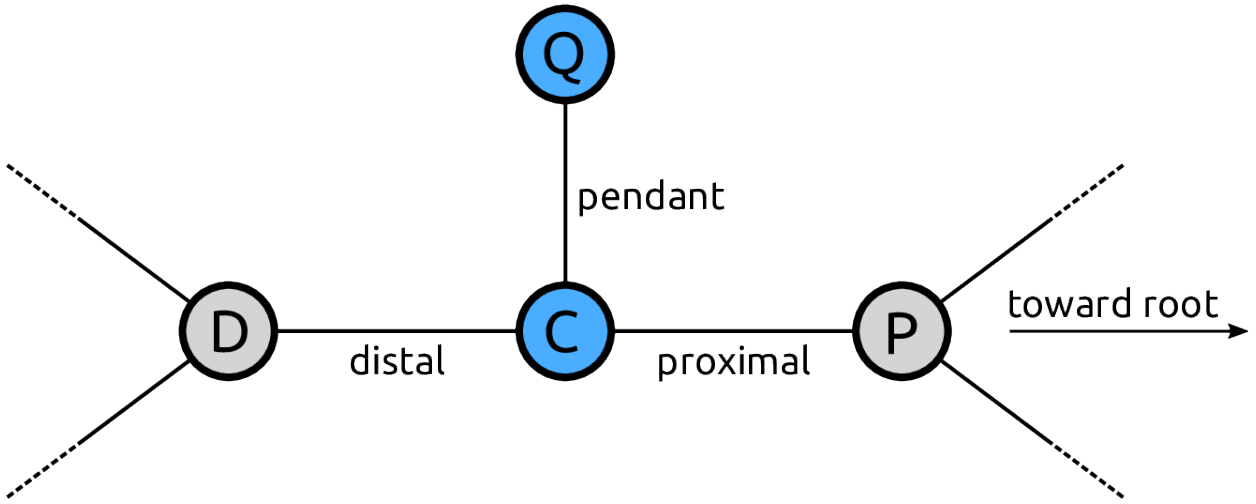


Single Sequence:

Likelihood Computation

For a single sequence on a single branch:

- Pretend that this is actually a new tip node of the tree



- Compute likelihood (or some other score)

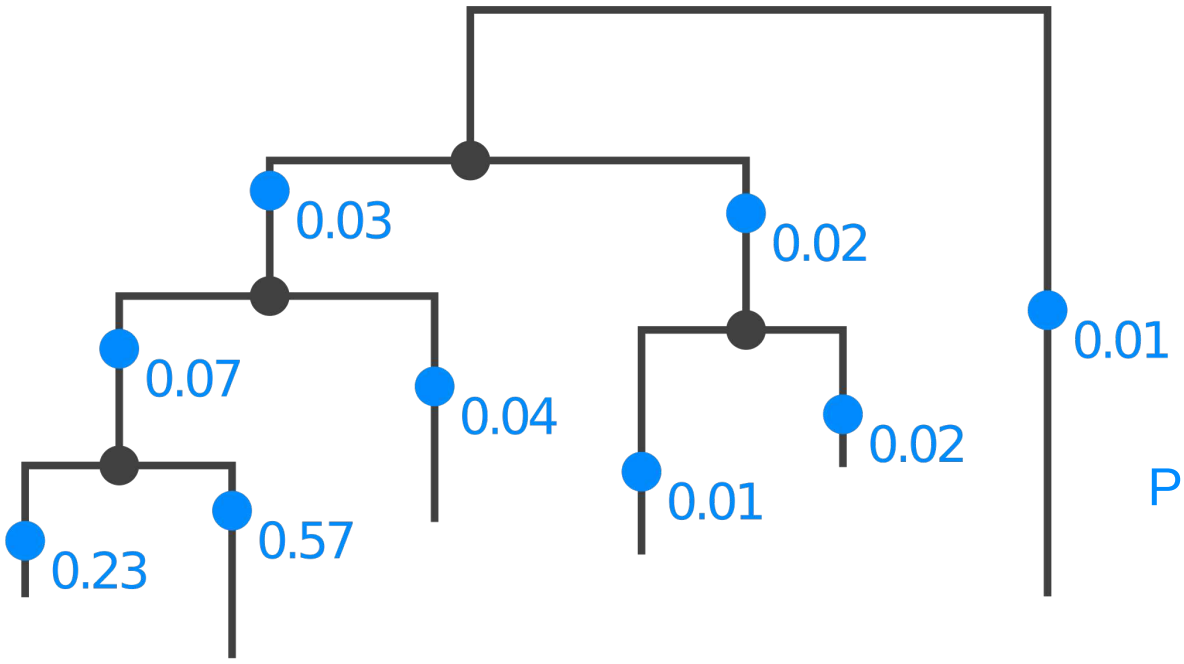
Likelihood Computation

- 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

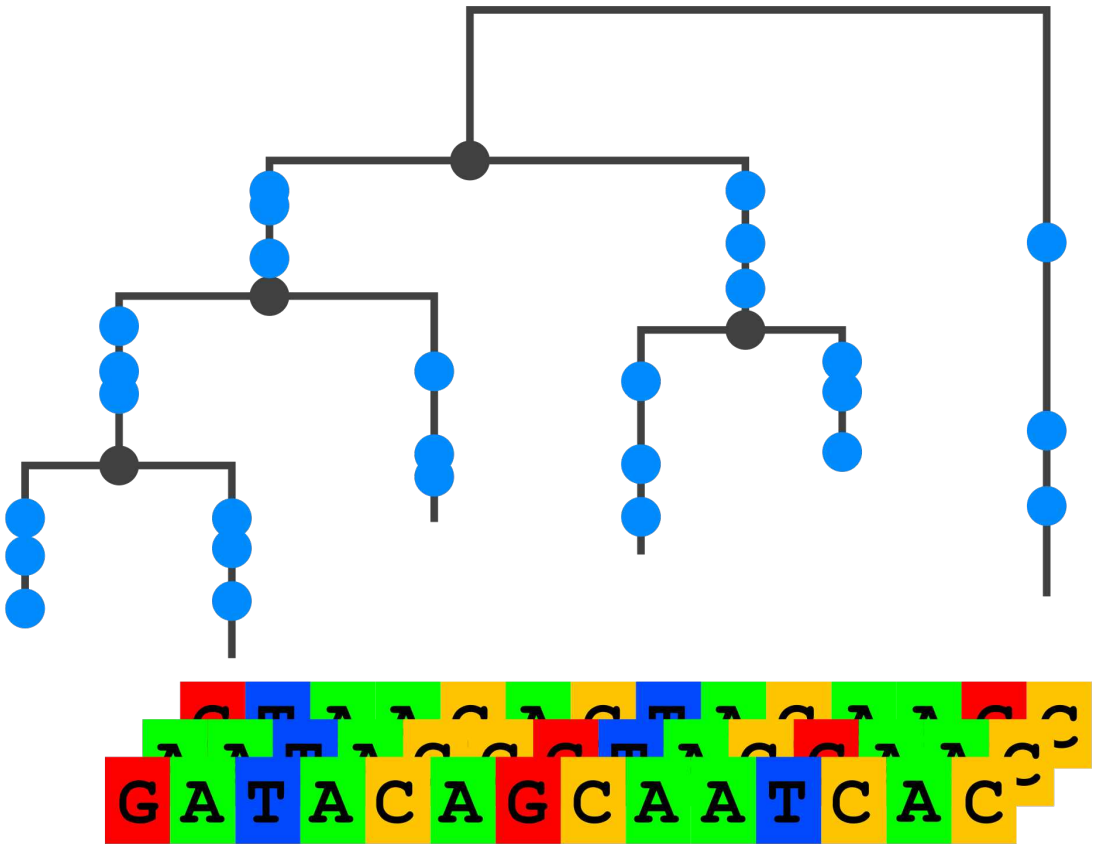
Phylogenetic Placement



Placement Masses
≙
Probabilities

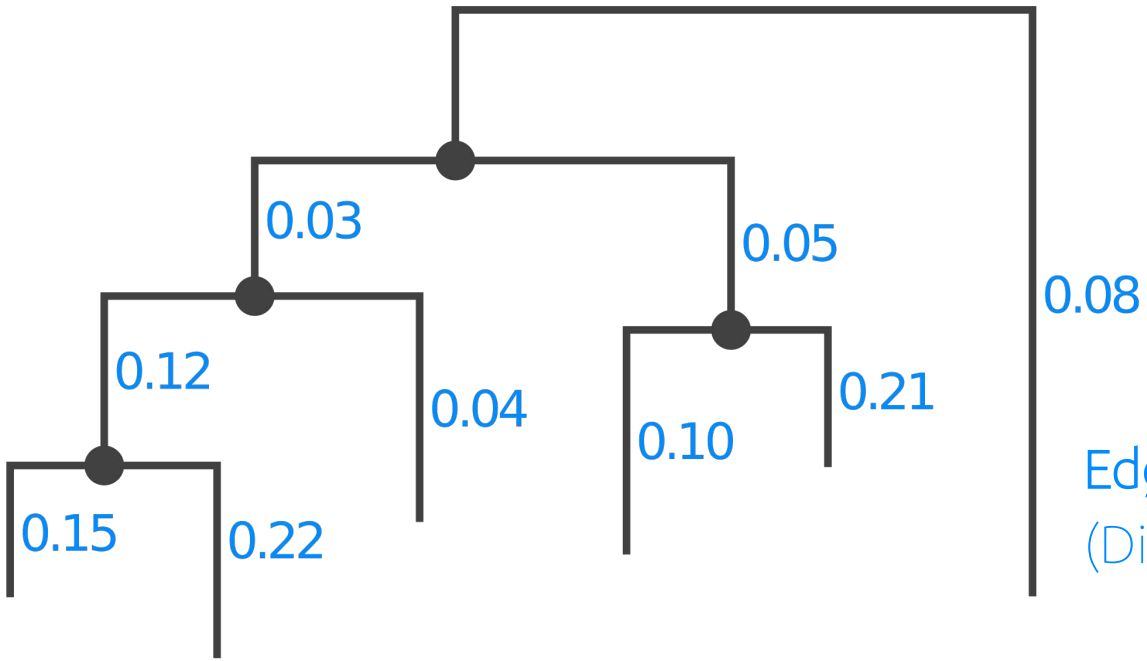
Single Sequence:

Phylogenetic Placement



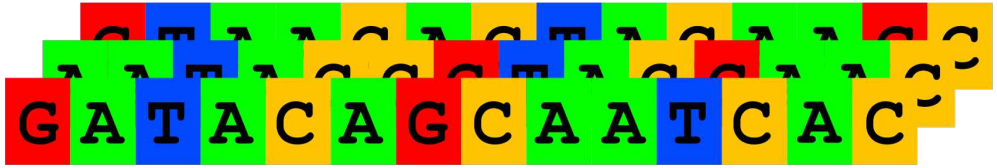
Whole Sample:

Phylogenetic Placement

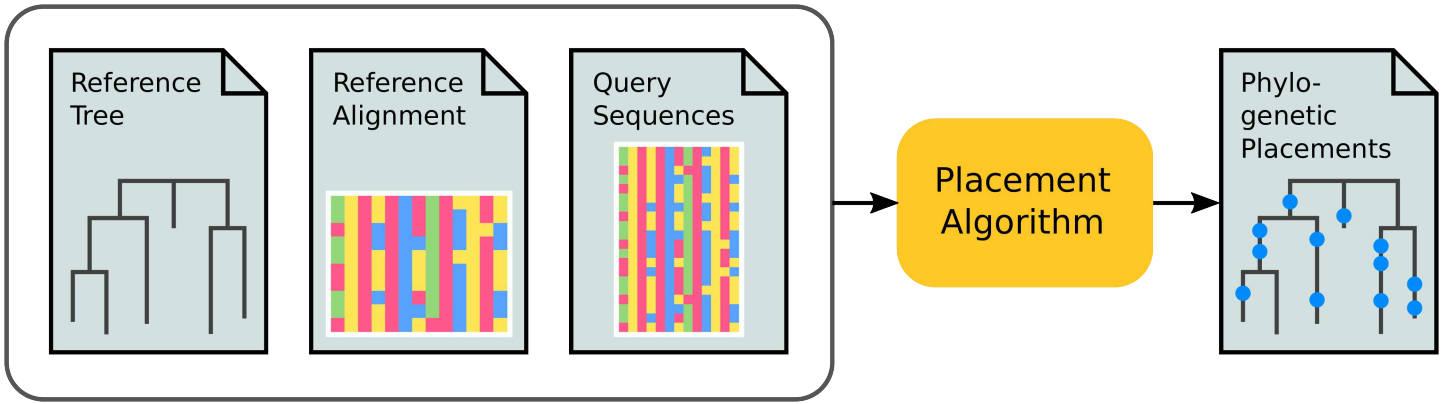


Edge Masses
(Distribution on the Tree)

Whole Sample:



Phylogenetic Placement Pipeline



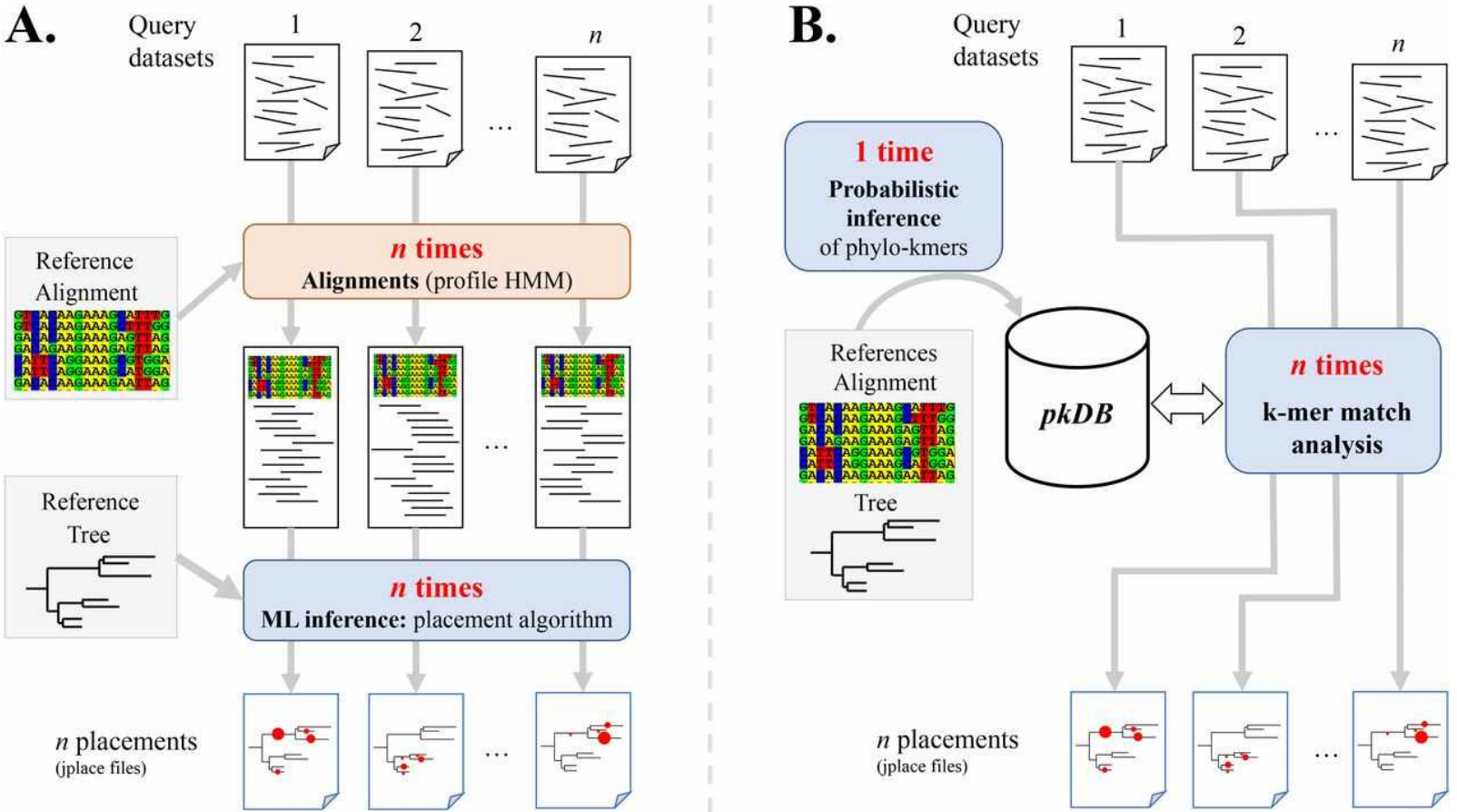
Input:

- Reference tree (newick)
- Reference alignment (fasta or phylip)
- Query sequences (fasta)

Output:

- Placements (jplace)

Alignment-free (k-mer based) placement

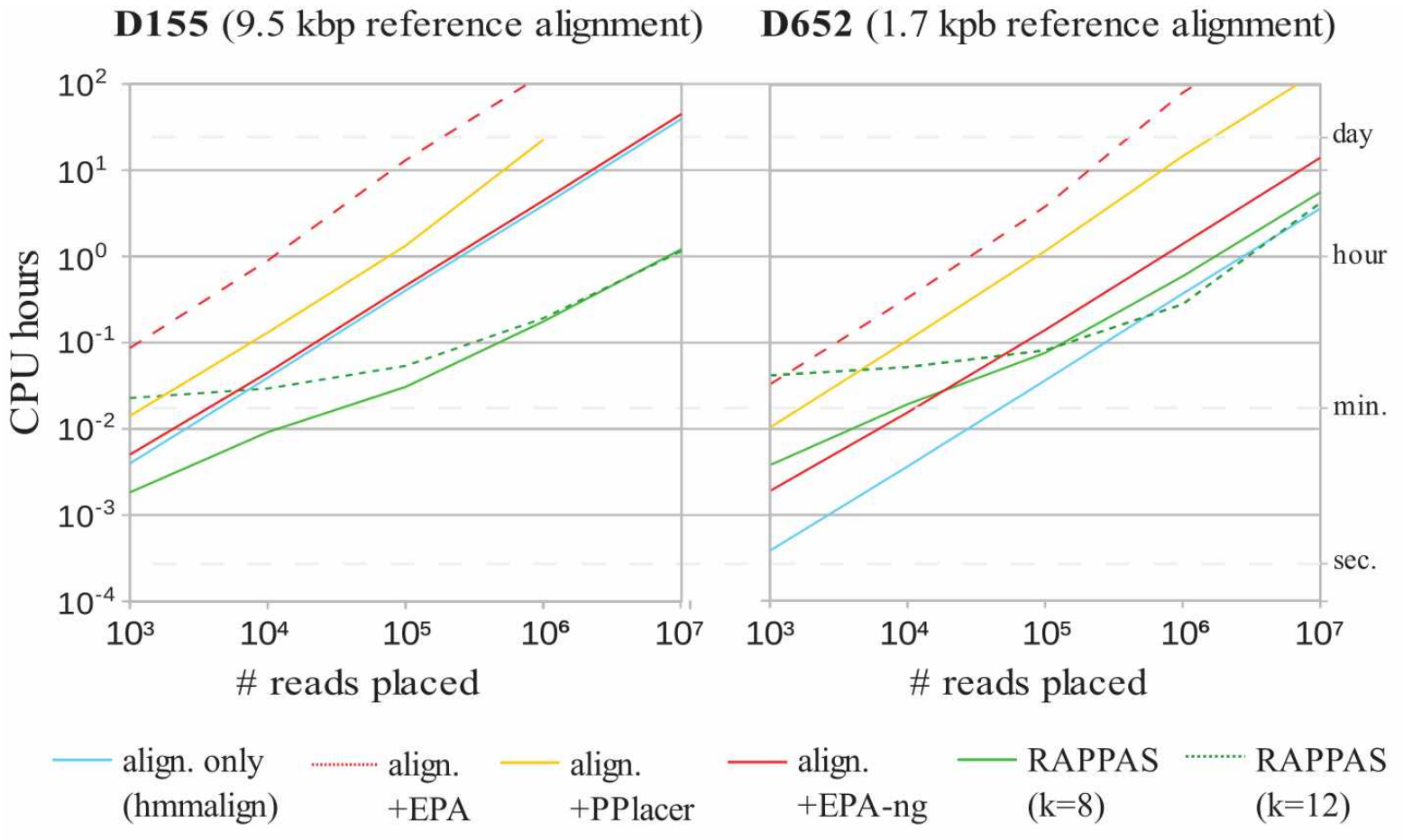


Linard *et al*, Bioinformatics, 2019

General Purpose Placement Methods

Placement Tool	Alignment	Multiple	Uncertainty	Branch Lengths
PPLACER	yes	yes	yes	yes
RAXML-EPA	yes	yes	yes	yes
EPA-NG	yes	yes	yes	yes
RAPPAS	no	yes	yes	no
APPLES	no	no	no	yes
APP-SPAM	no	no	no	yes

Runtime Comparison



Placement Analysis and Visualization

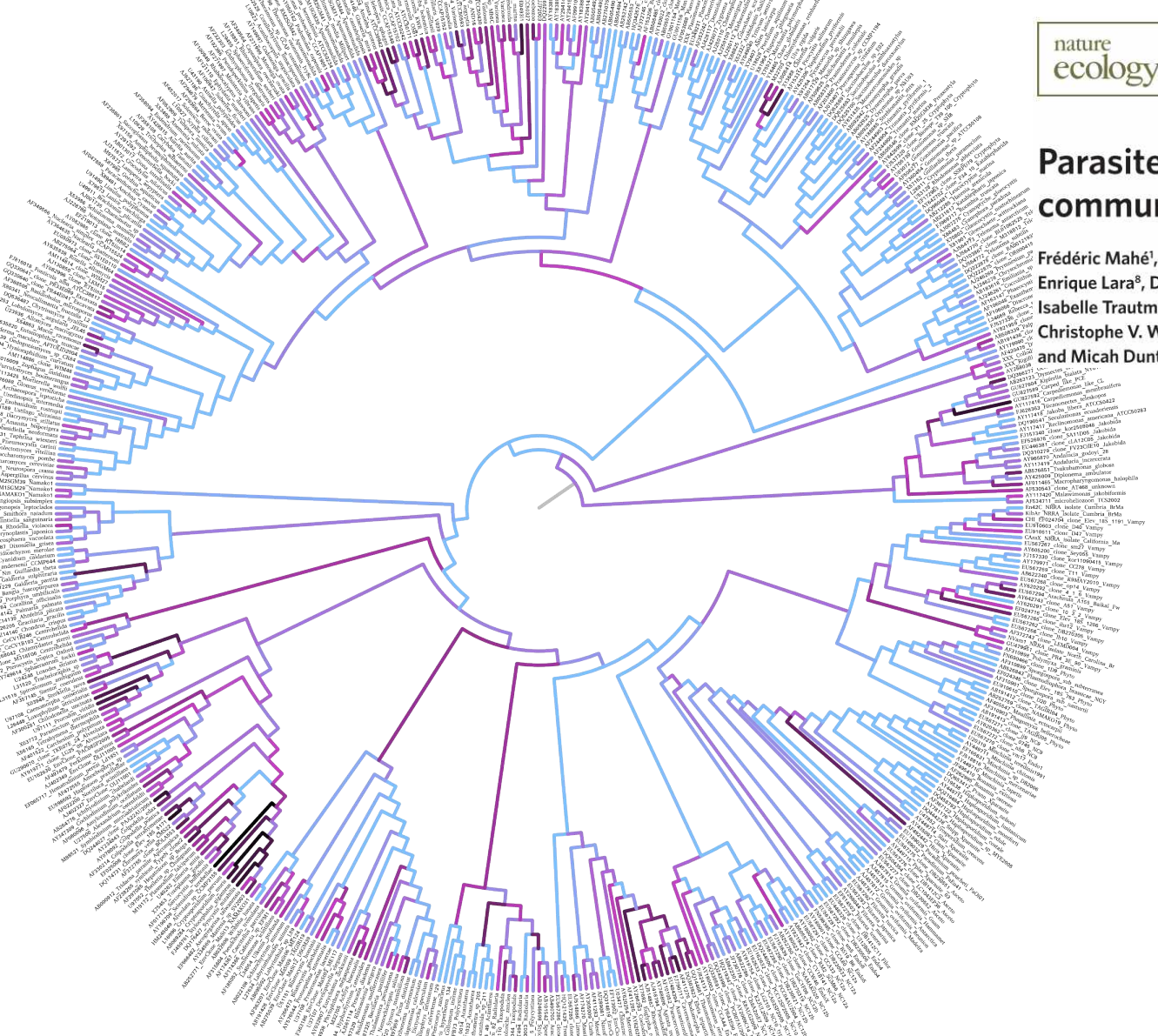
Placement Analysis

Two interpretations:

- Assignment (or mapping) of sequences to branches
- Distribution of sequences across the tree

Typical types of analysis:

- Examine a single sample
- Relate multiple samples to each other
- Relate samples to environmental factors / variables



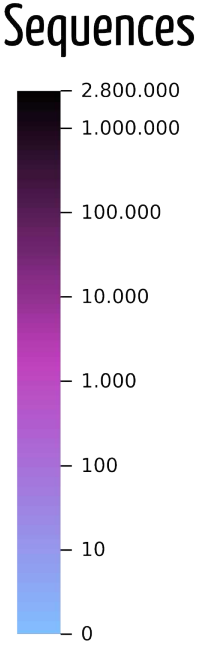
nature ecology & evolution

ARTICLES

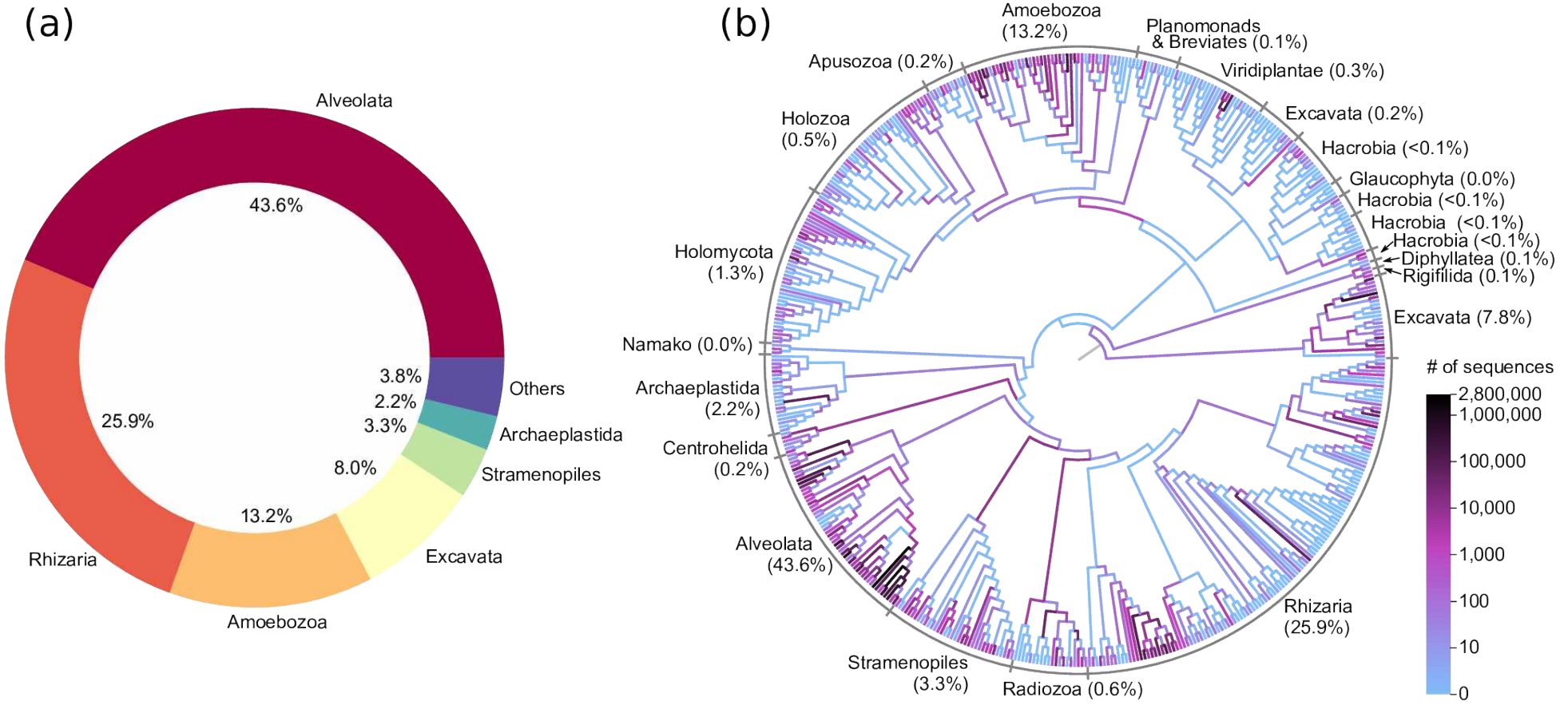
PUBLISHED: 20 MARCH 2017 | VOLUME: 1 | ARTICLE NUMBER: 0091

Parasites dominate hyperdiverse soil protist communities in Neotropical rainforests

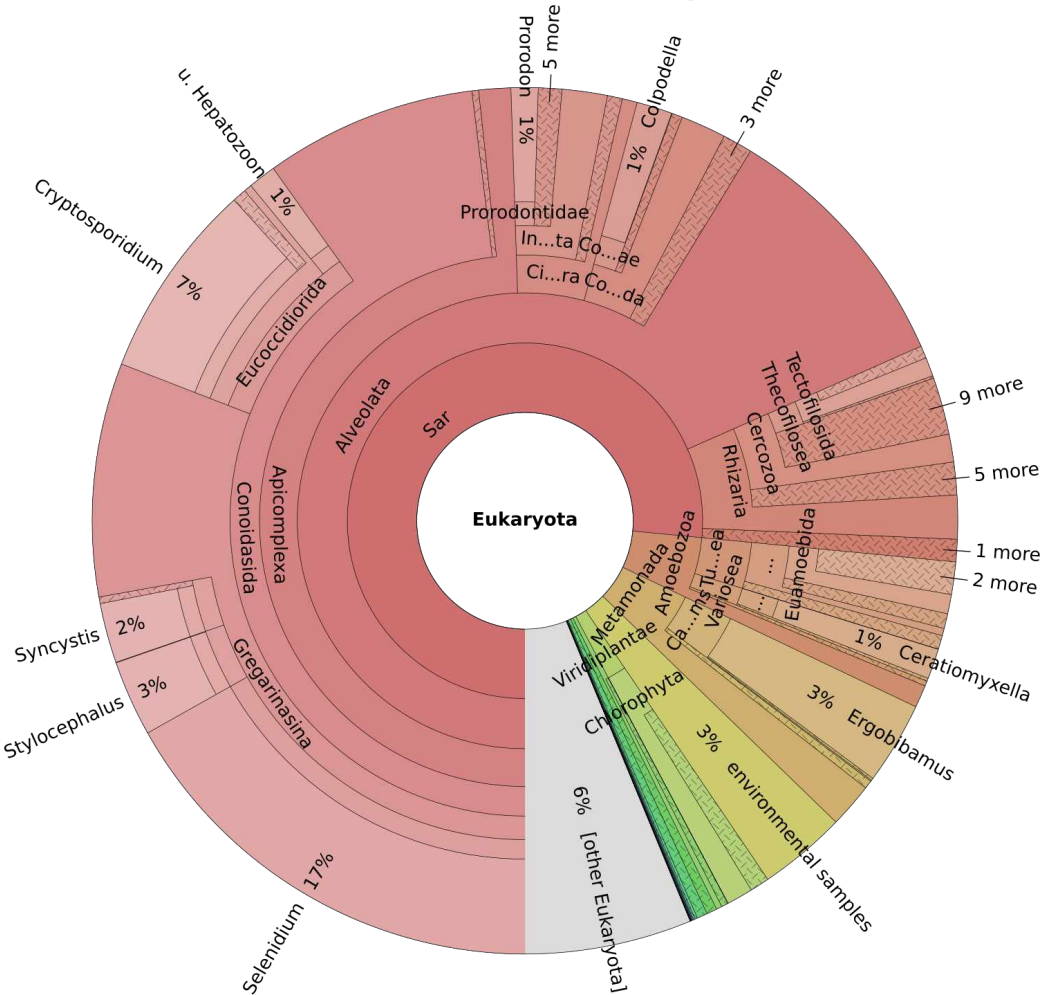
Frédéric Mahé[1], Colomban de Vargas[2,3], David Bass[4,5], Lucas Czech[6], Alexandros Stamatakis[6,7], Enrique Lara[8], David Singer[8], Jordan Mayor[9], John Bunge[10], Sarah Sernaker[11], Tobias Siemensmeyer[1], Isabelle Trautmann[1], Sarah Romac[2,3], Cédric Berney[2,3], Alexey Kozlov[6], Edward 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]*



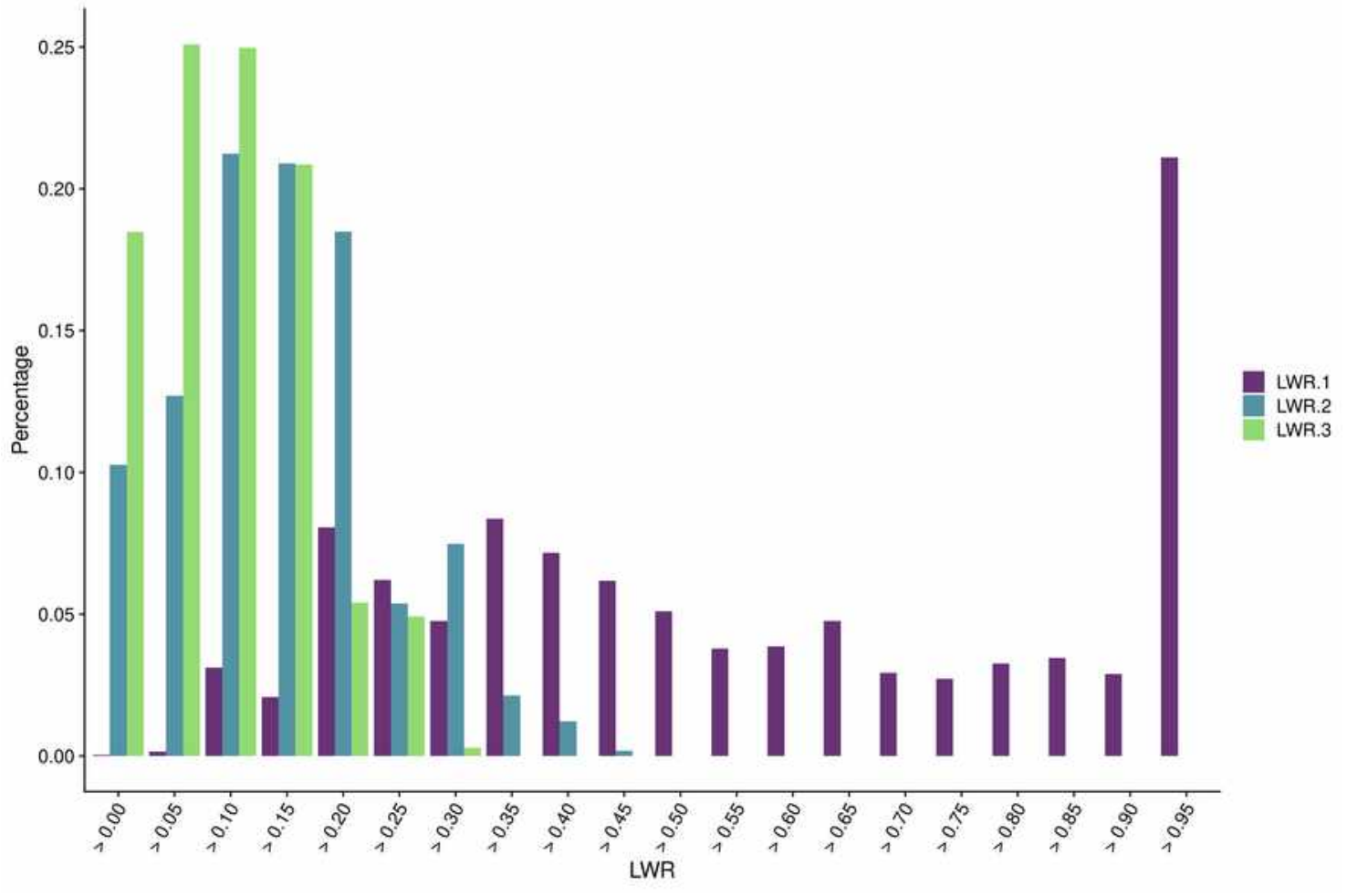
Abundances vs. Phylogenetic Placements



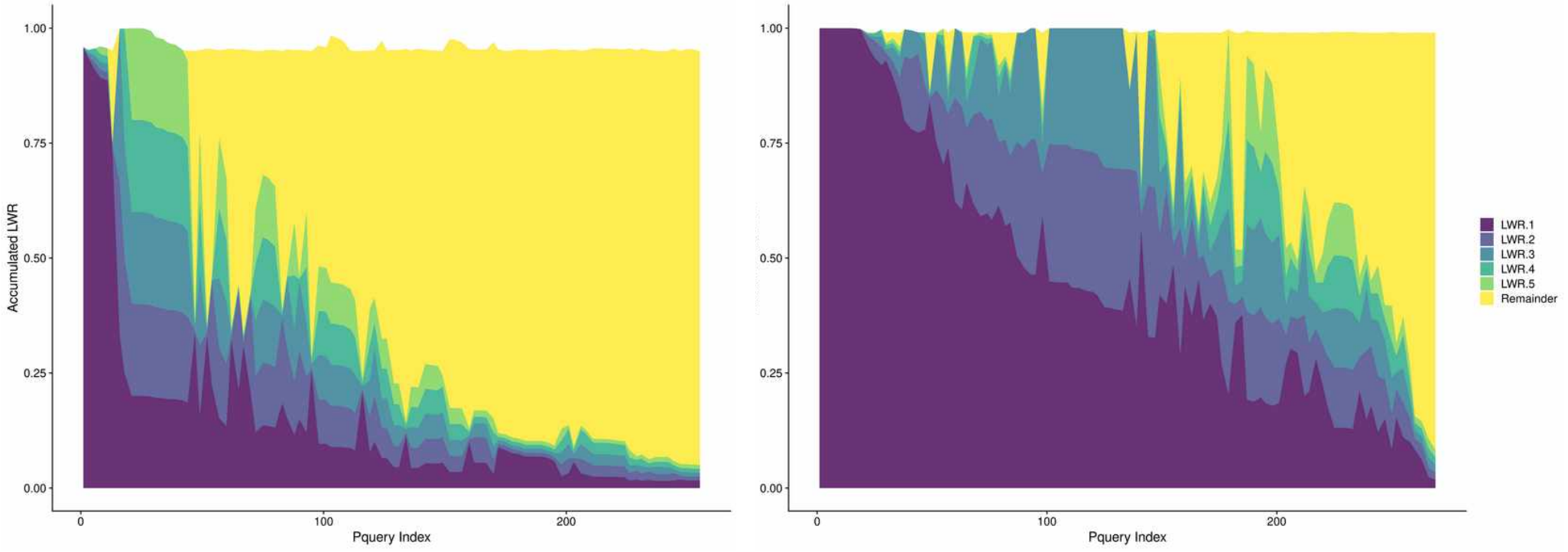
Taxonomic Assignment



LWR Distribution

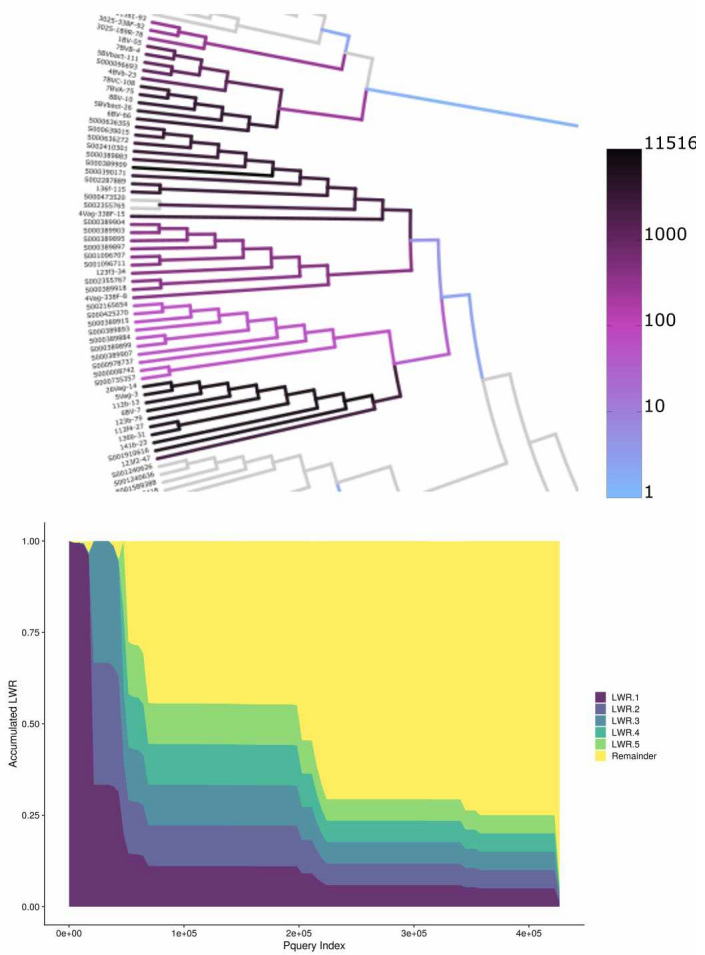


LWR Distribution

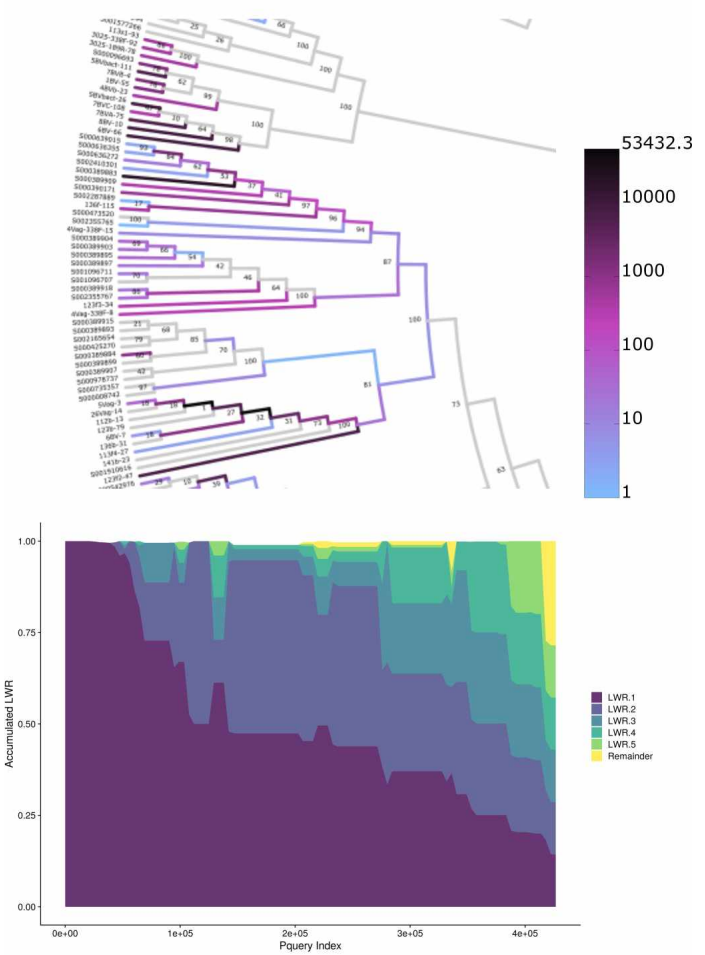


Comparing ML vs k-mer based placement

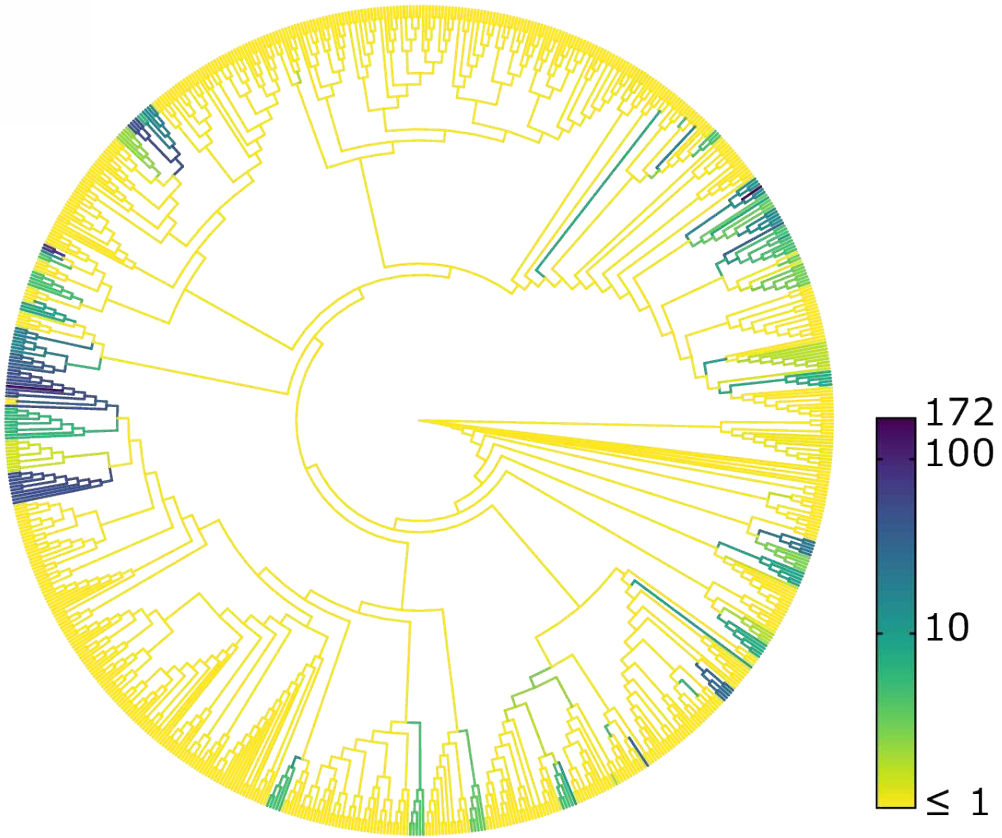
ML



k-mer



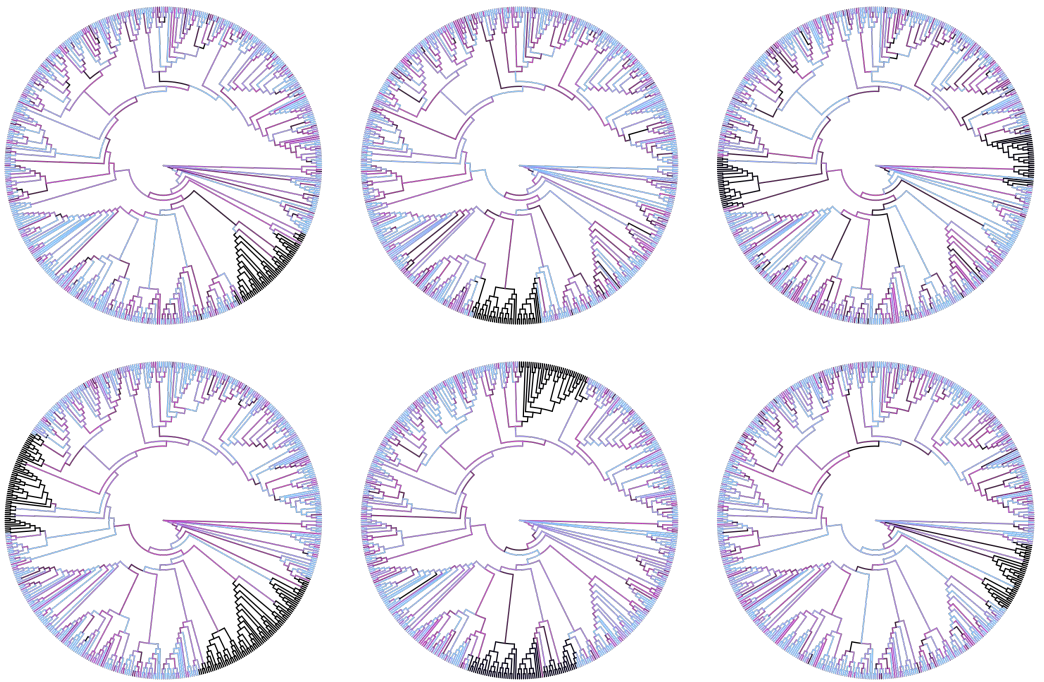
Edge Dispersion



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
  - …

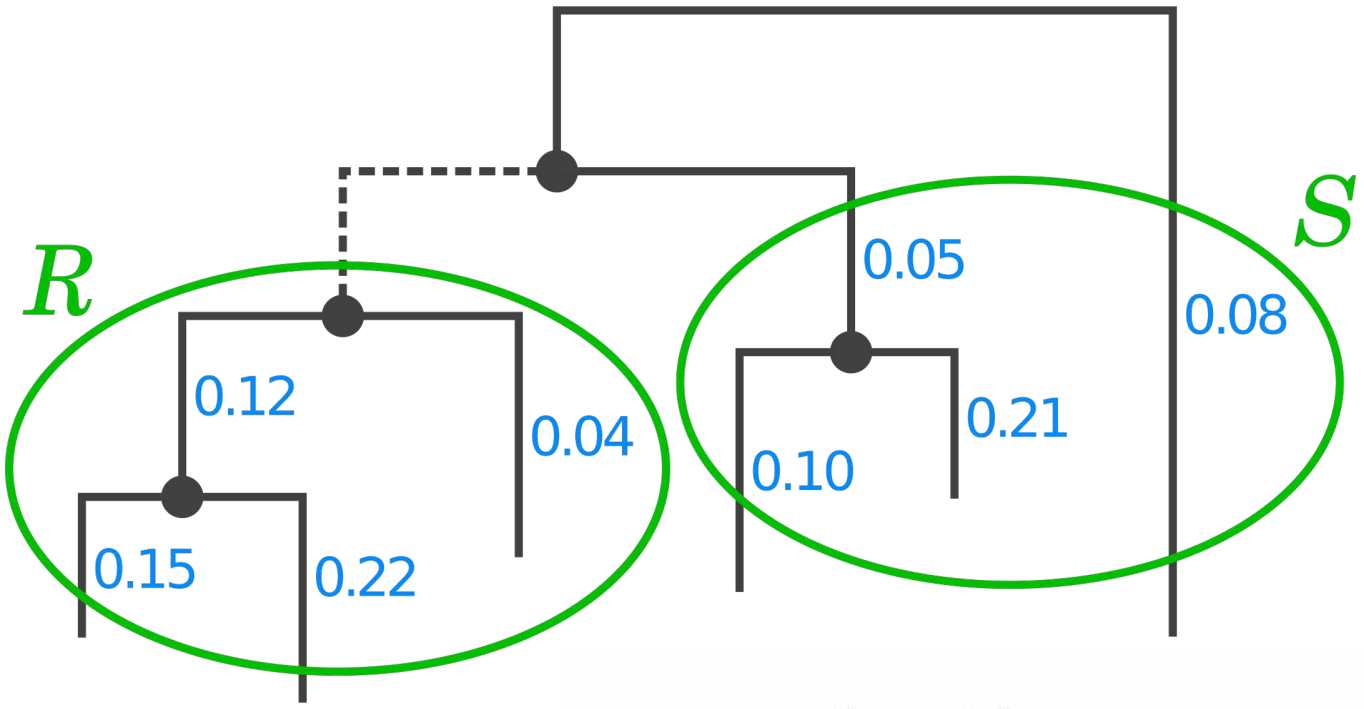


Caveat

- Metagenomic data is compositional!
- This has statistical implications:
  Sequence abundances cannot be interpreted absolutely
- Same for phylogenetic placements

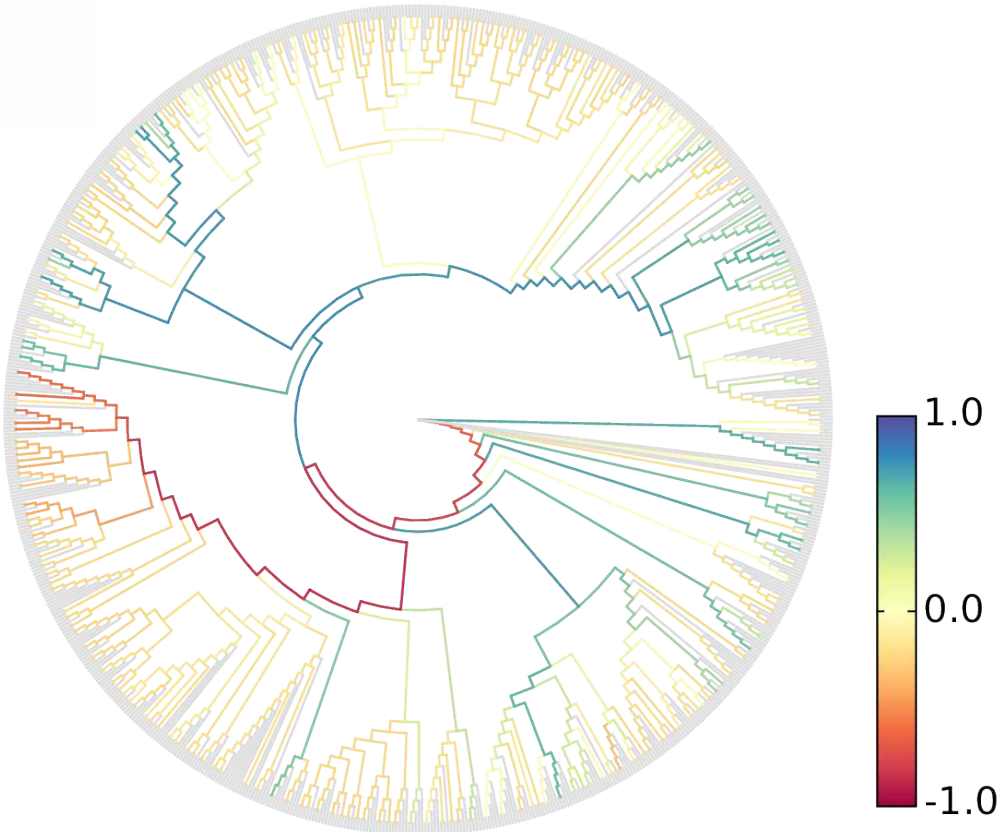
  → transform the Edge Masses

Edge Balance

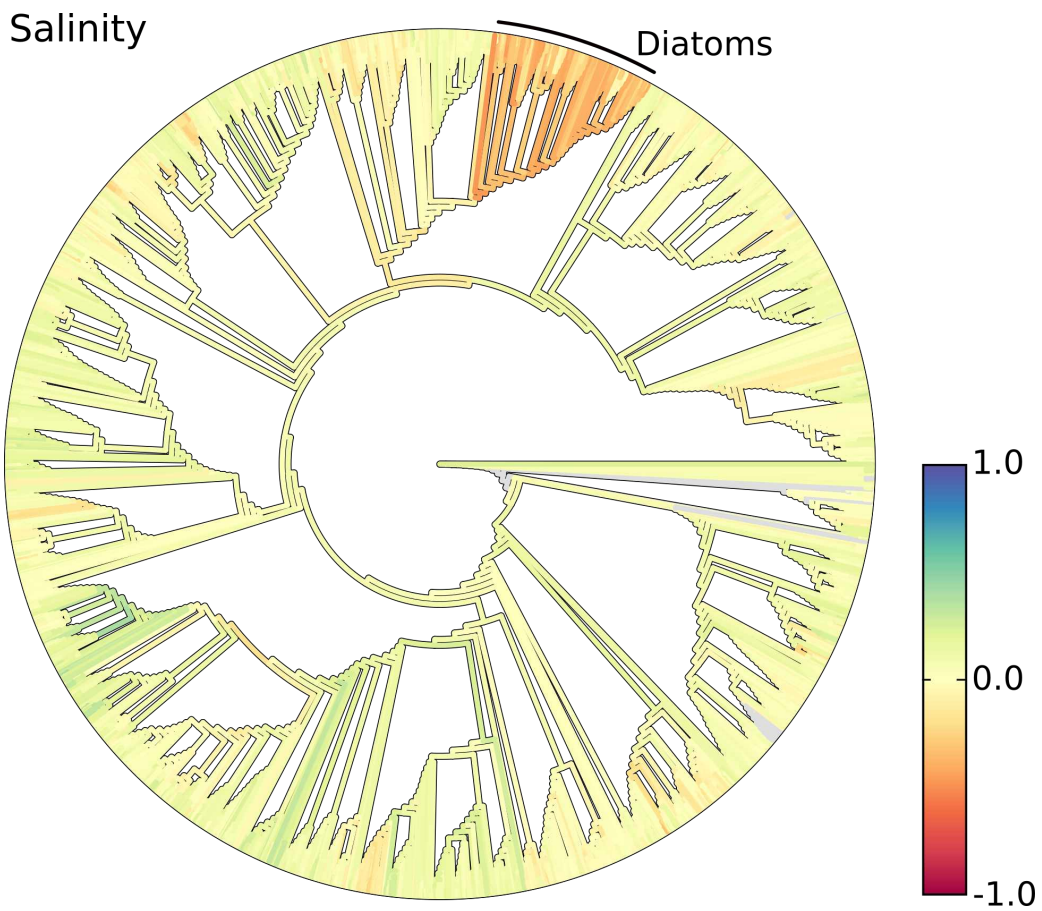


$$\text{balance}(R, S) = \lambda \cdot \log \frac{\text{gm}(R)}{\text{gm}(S)}$$

Edge Correlation



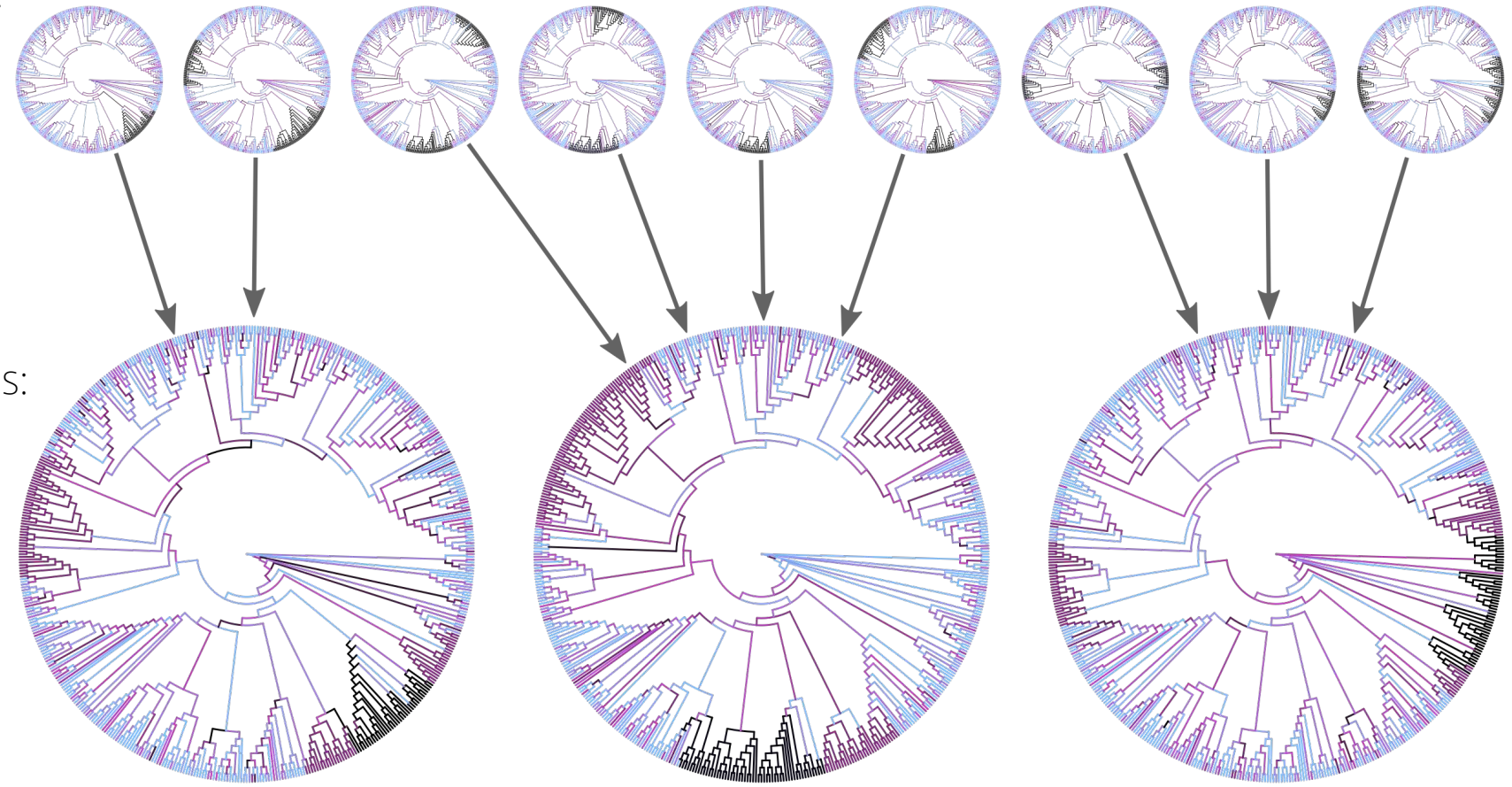
Edge Correlation



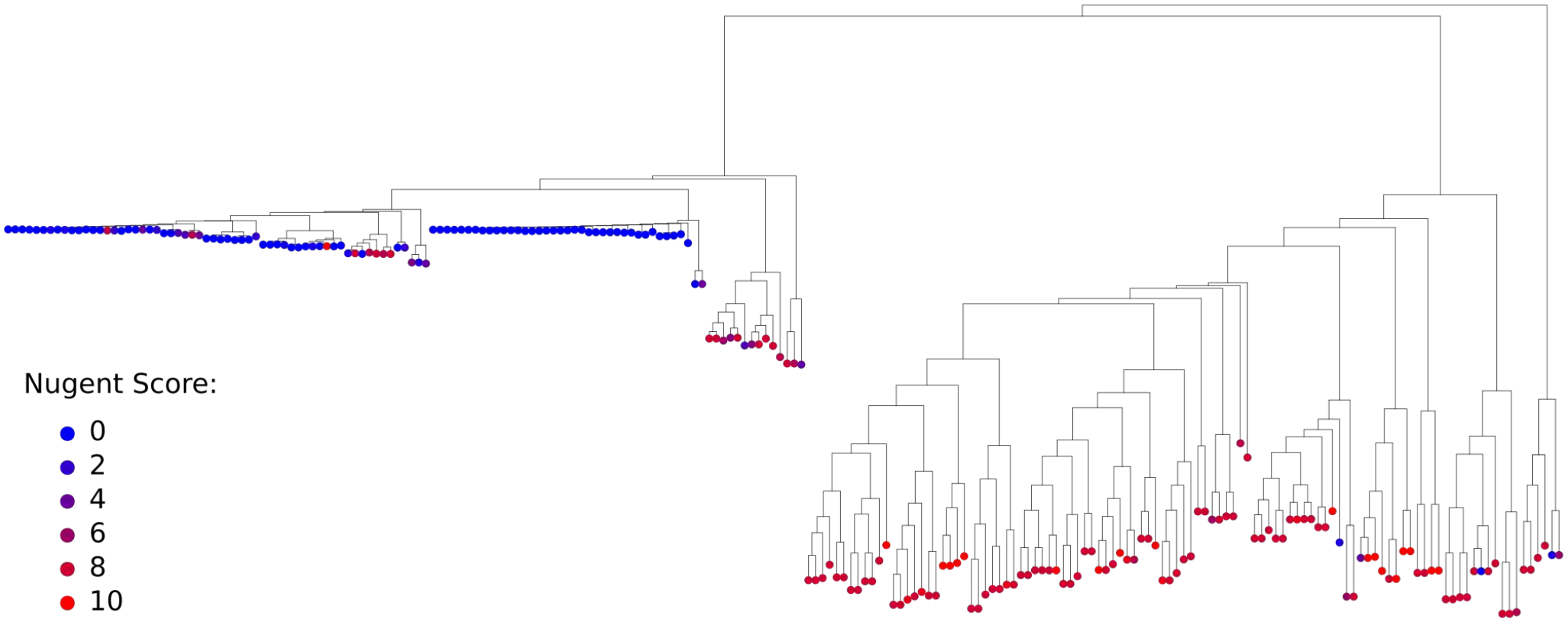
Phylogenetic K-means Clustering

Samples:

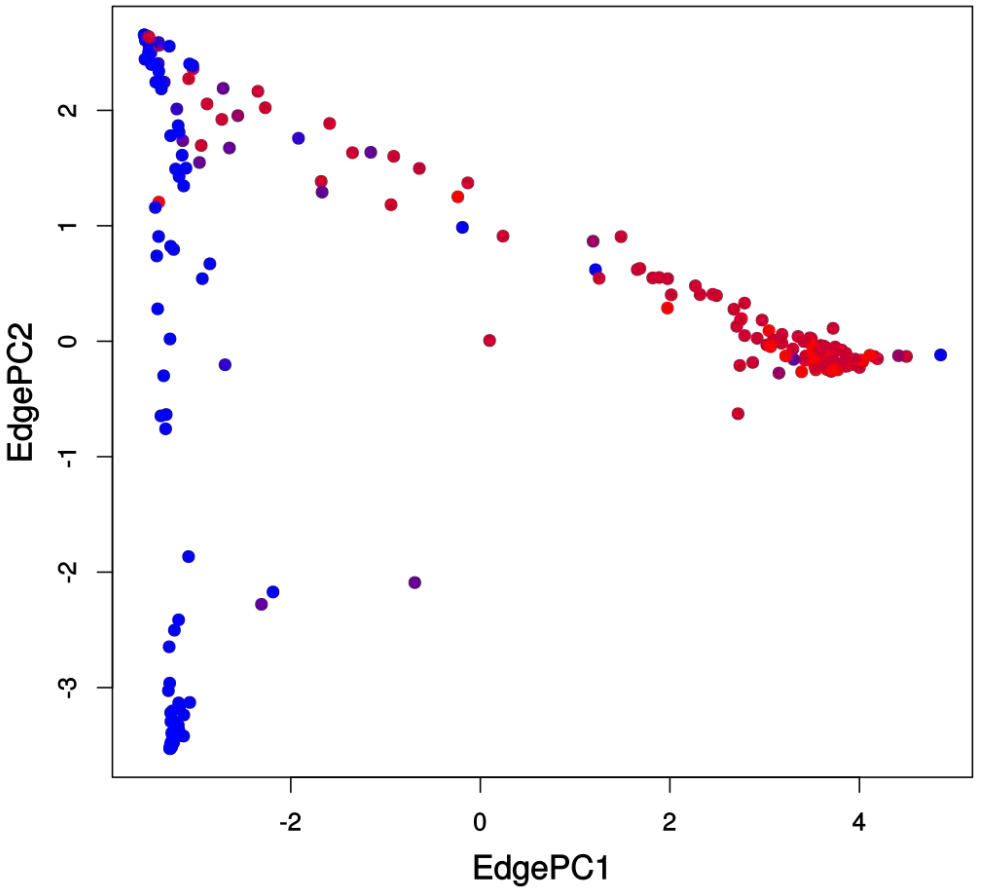
Centroids:



Squash Clustering



Edge PCA

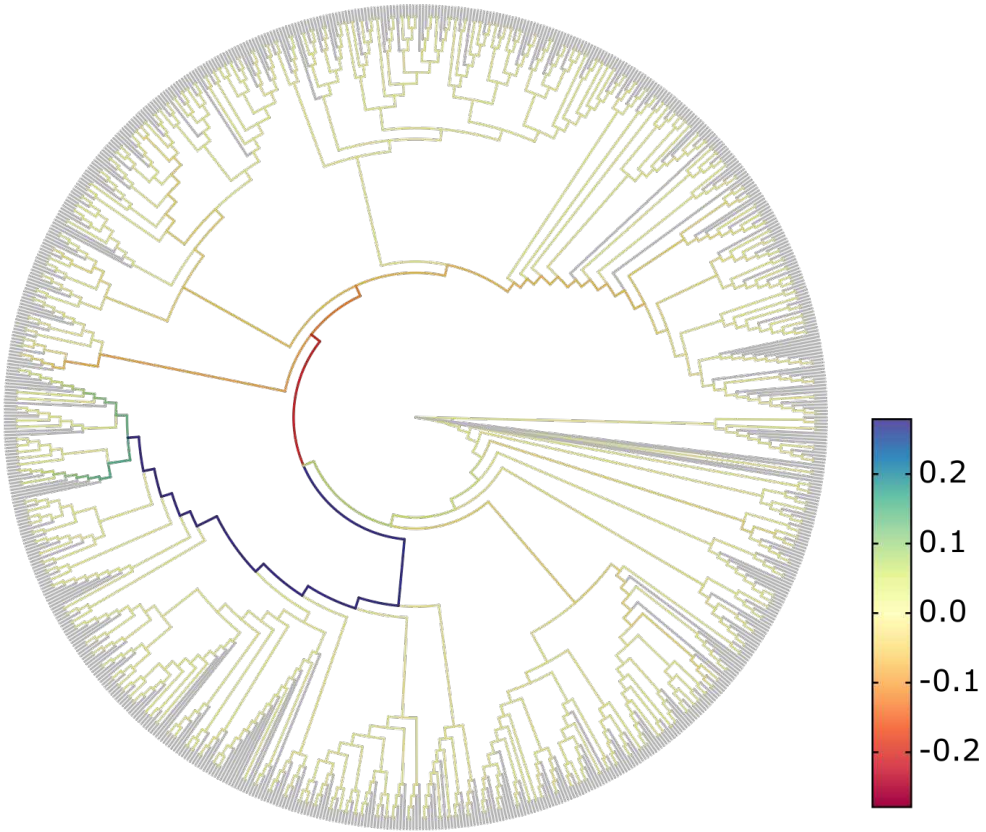


Nugent Score:

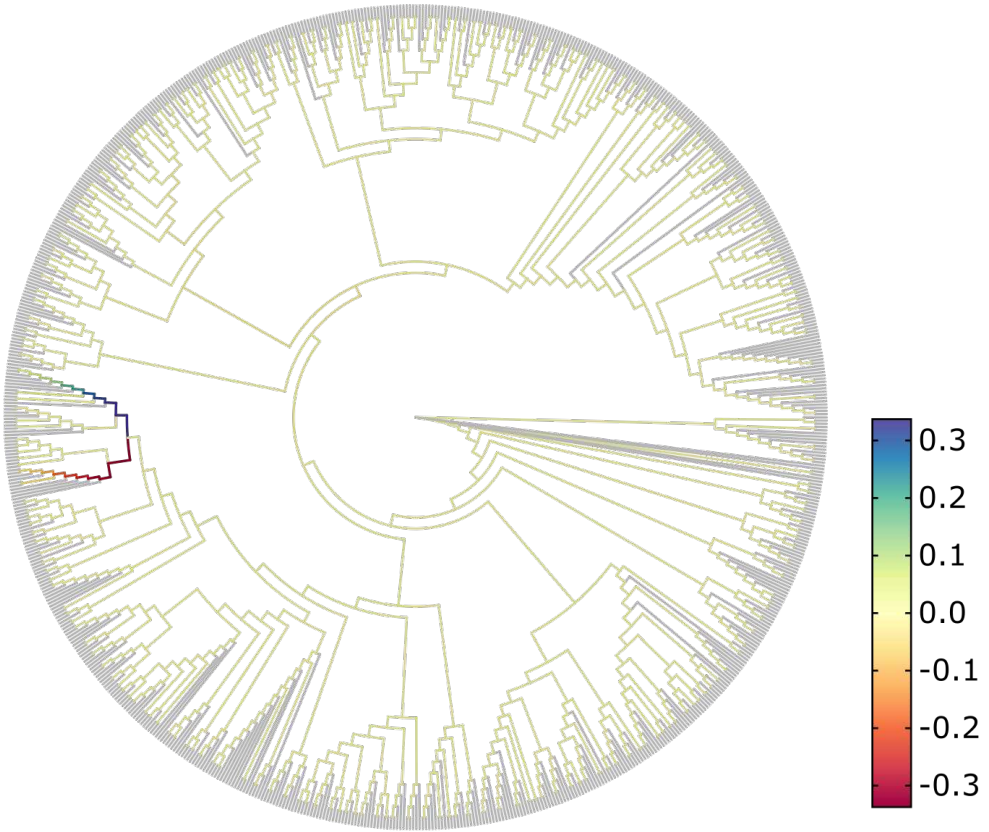
- 0
- 2
- 4
- 6
- 8
- 10

Edge PCA

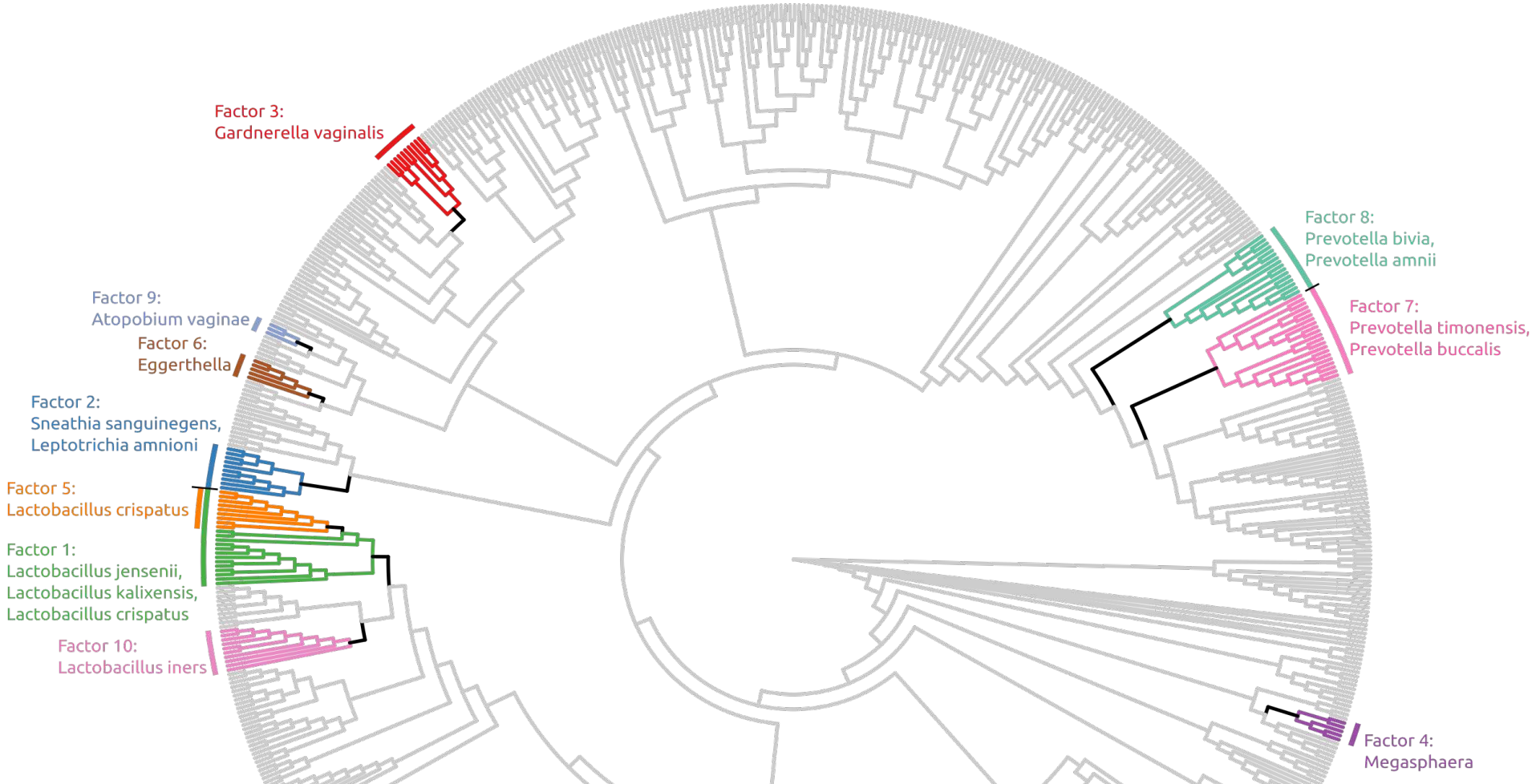
(a) First Component



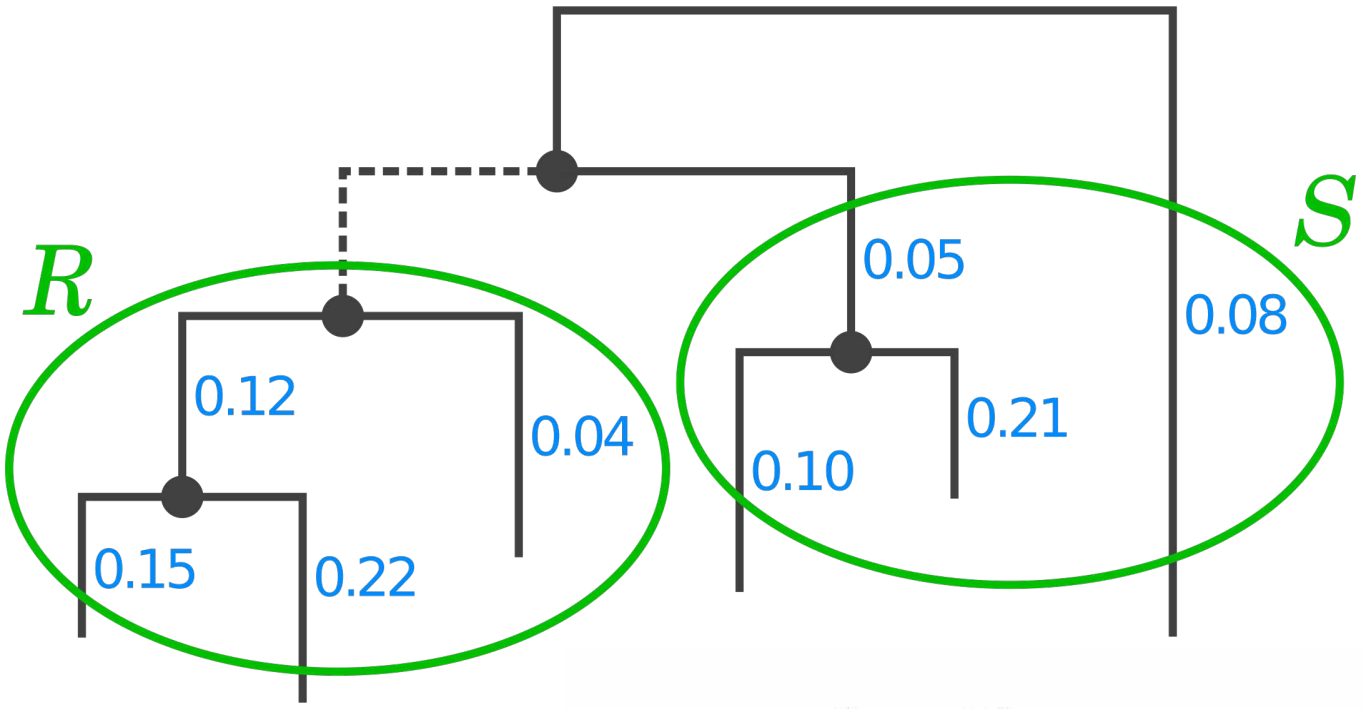
(b) Second Component



Placement-Factorization

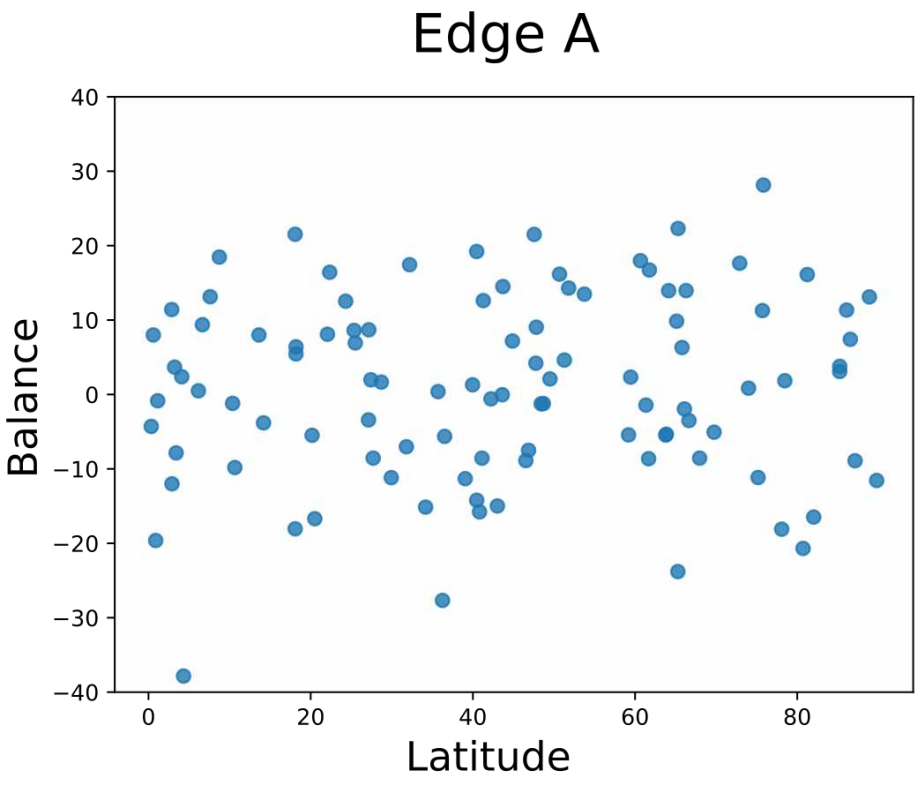


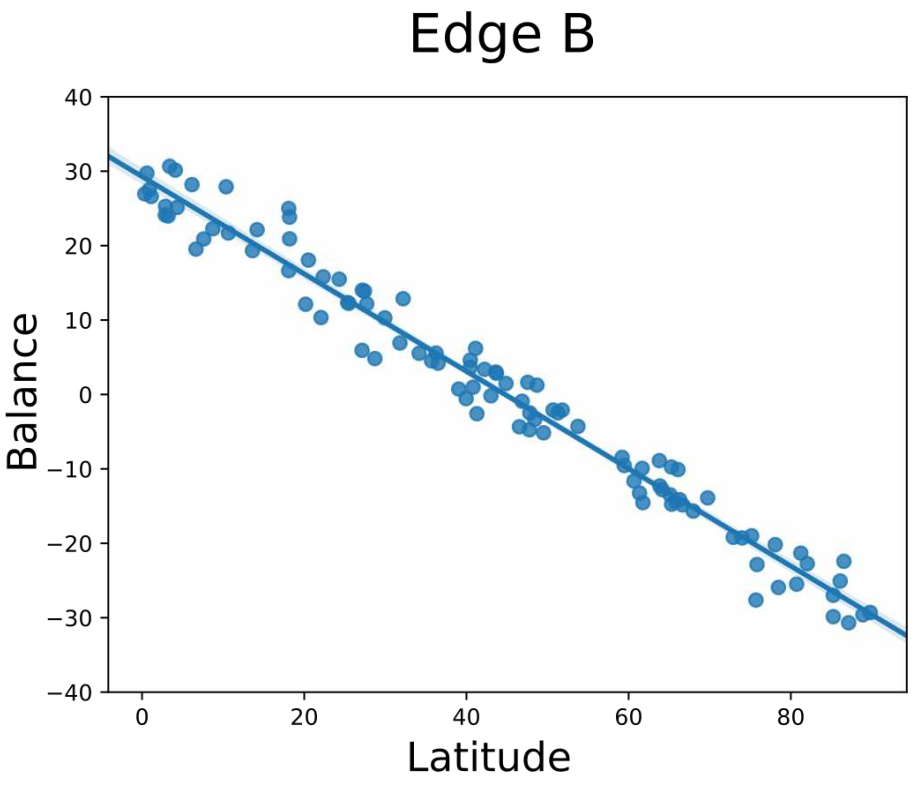
Edge Balance



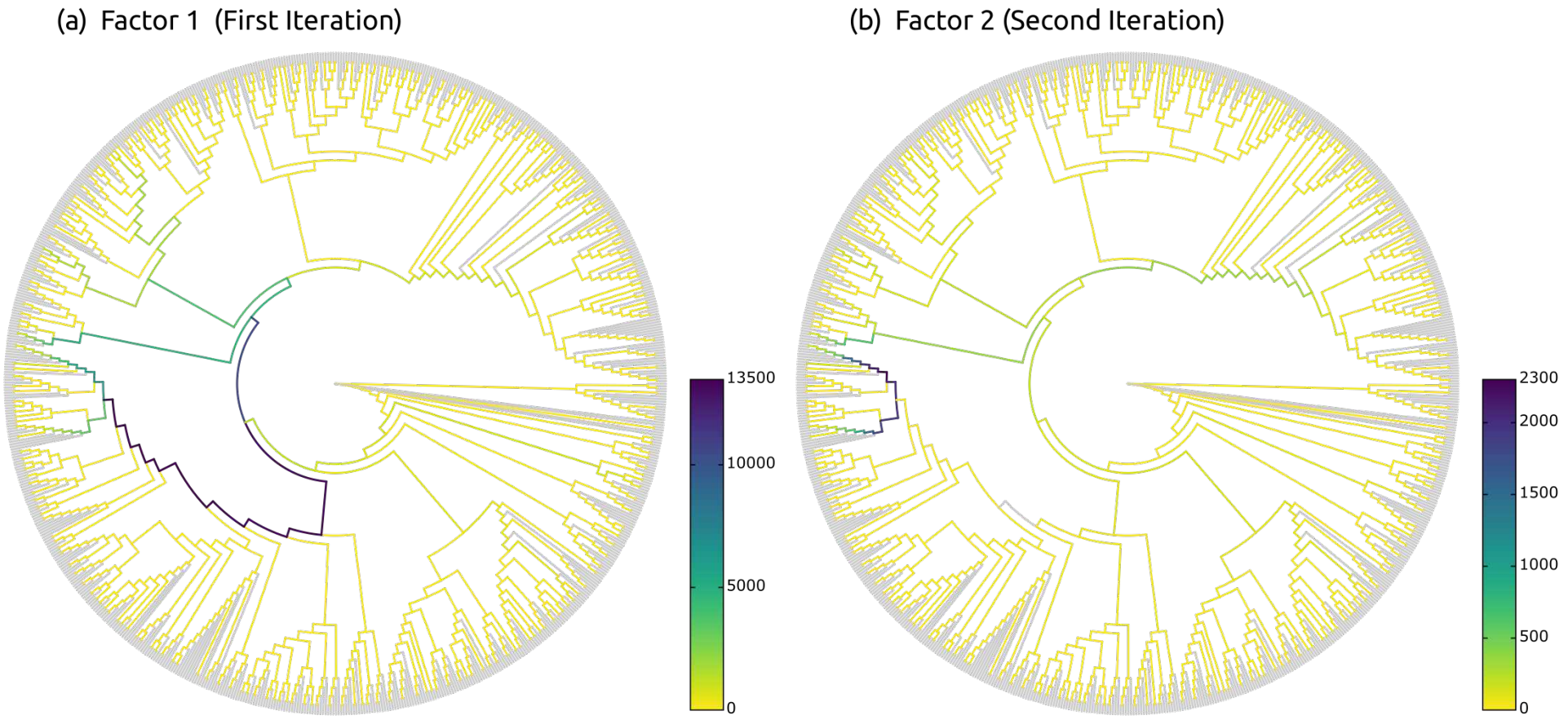
$$\text{balance}(R, S) = \lambda \cdot \log \frac{\text{gm}(R)}{\text{gm}(S)}$$

Balances vs. Metadata

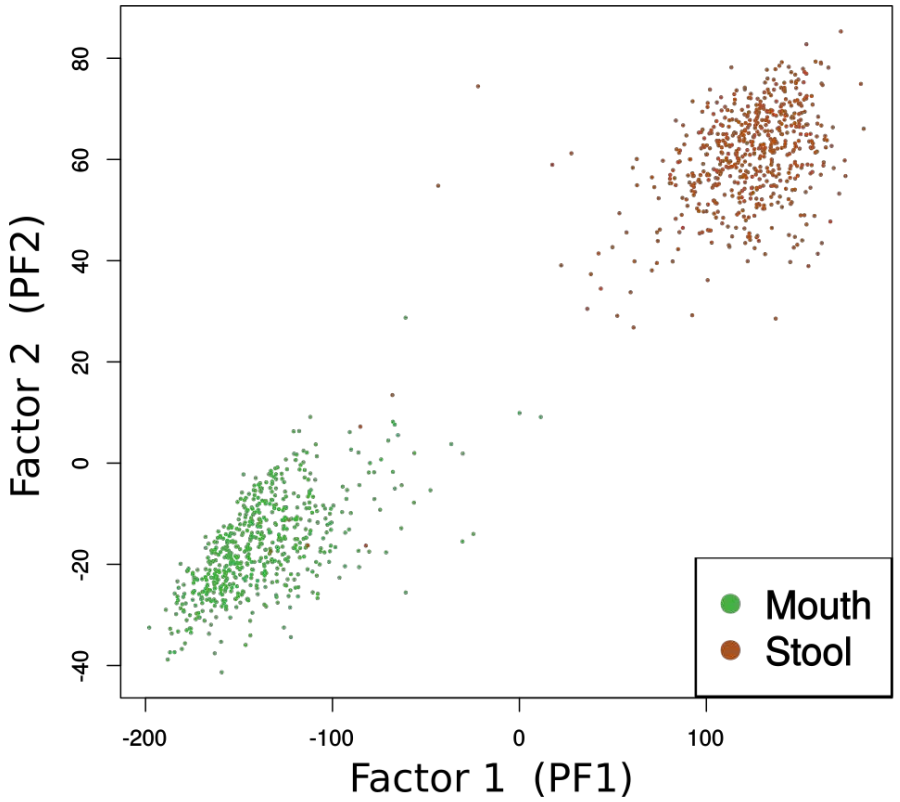


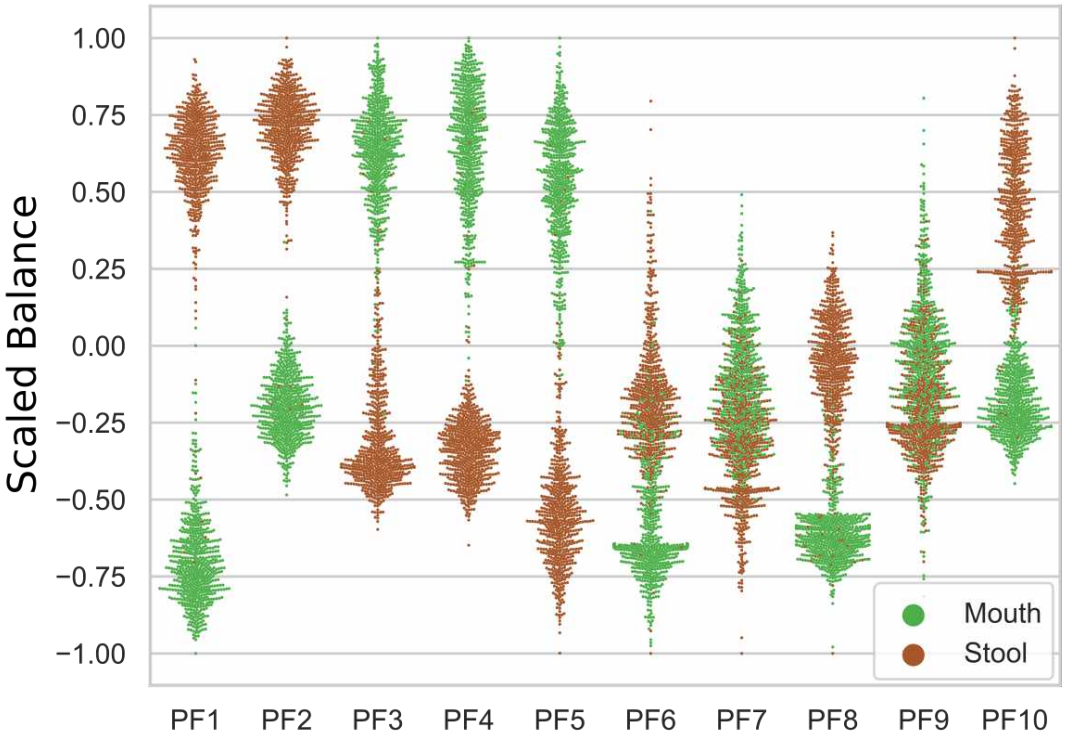


Placement-Factorization

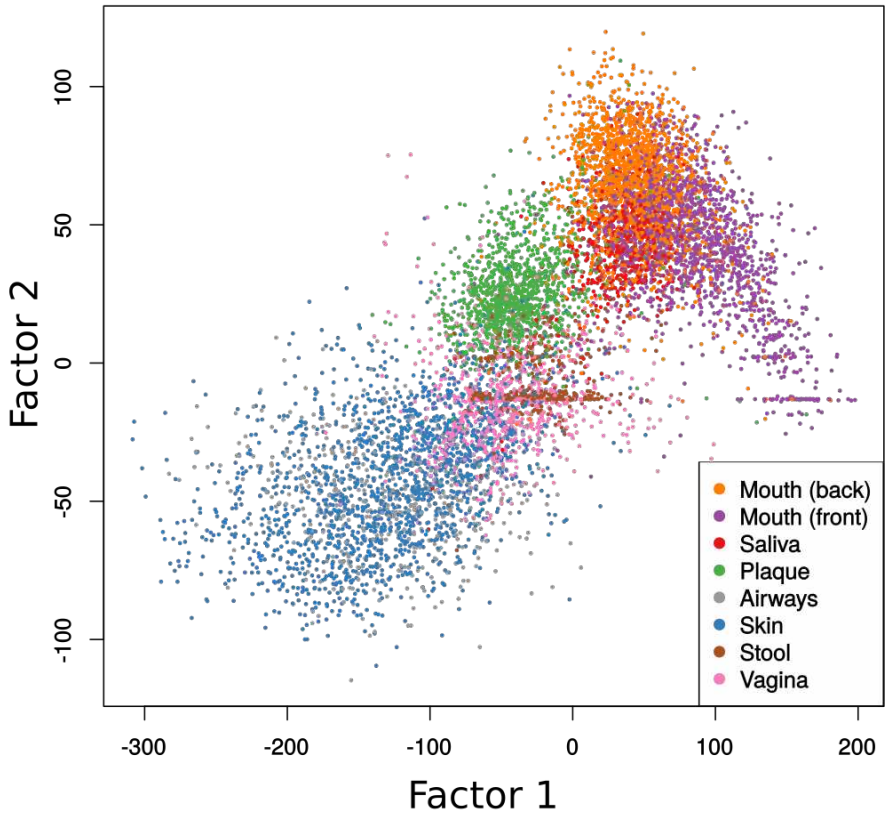


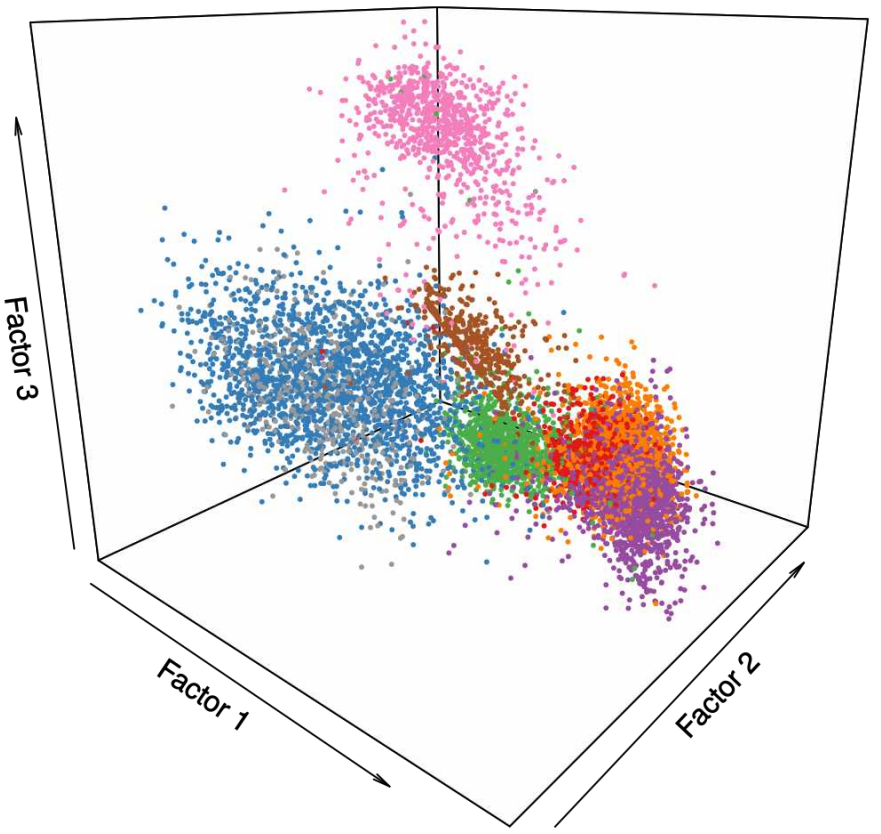
Placement-Factorization



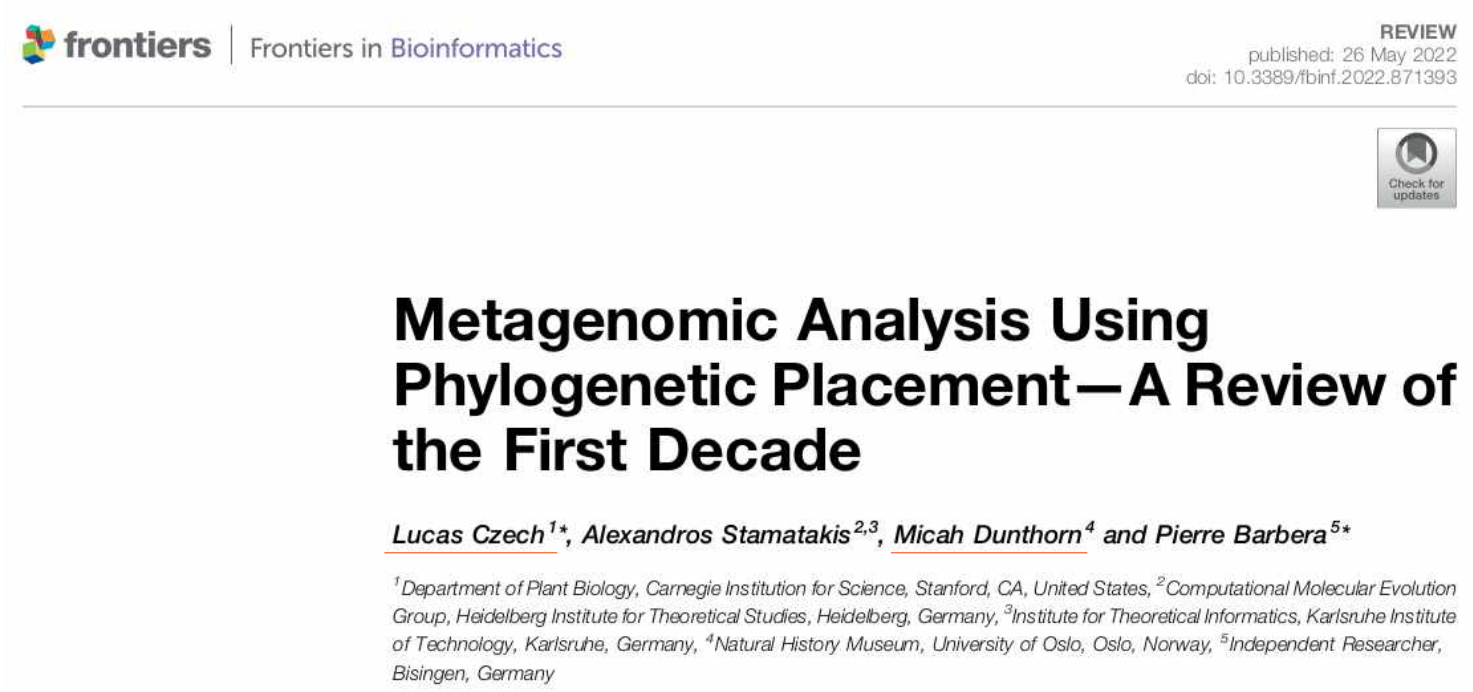


Placement-Factorization





Into the Placement-verse



frontiers | Frontiers in Bioinformatics

REVIEW
published: 26 May 2022
doi: 10.3389/fbinf.2022.871393

Check for updates

Metagenomic Analysis Using Phylogenetic Placement—A Review of the First Decade

Lucas Czech[1]*, *Alexandros Stamatakis*[2,3], *Micah Dunthorn*[4] and *Pierre Barbera*[5]*

[1]Department of Plant Biology, Carnegie Institution for Science, Stanford, CA, United States, [2]Computational Molecular Evolution Group, Heidelberg Institute for Theoretical Studies, Heidelberg, Germany, [3]Institute for Theoretical Informatics, Karlsruhe Institute of Technology, Karlsruhe, Germany, [4]Natural History Museum, University of Oslo, Oslo, Norway, [5]Independent Researcher, Bisingen, Germany

doi: 10.3389/fbinf.2022.871393

Thank you!
Time for your questions