Phylogenetic Placement: Computation, Analysis, and Visualization

Lucas Czech

2025-04-10
Guest Lecture
University of Oslo

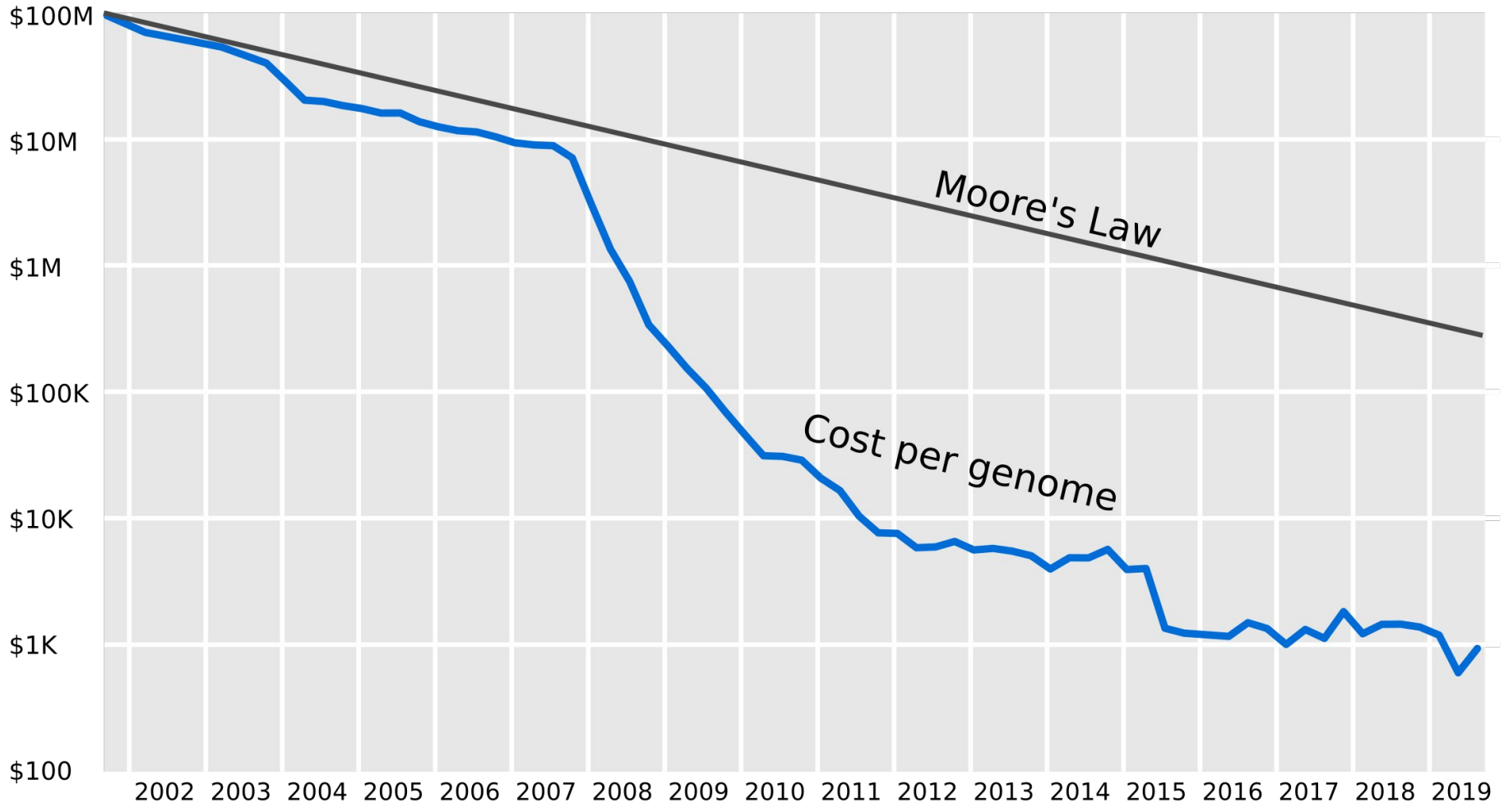
Agenda

- Motivation
- Phylogenetic Tree Inference

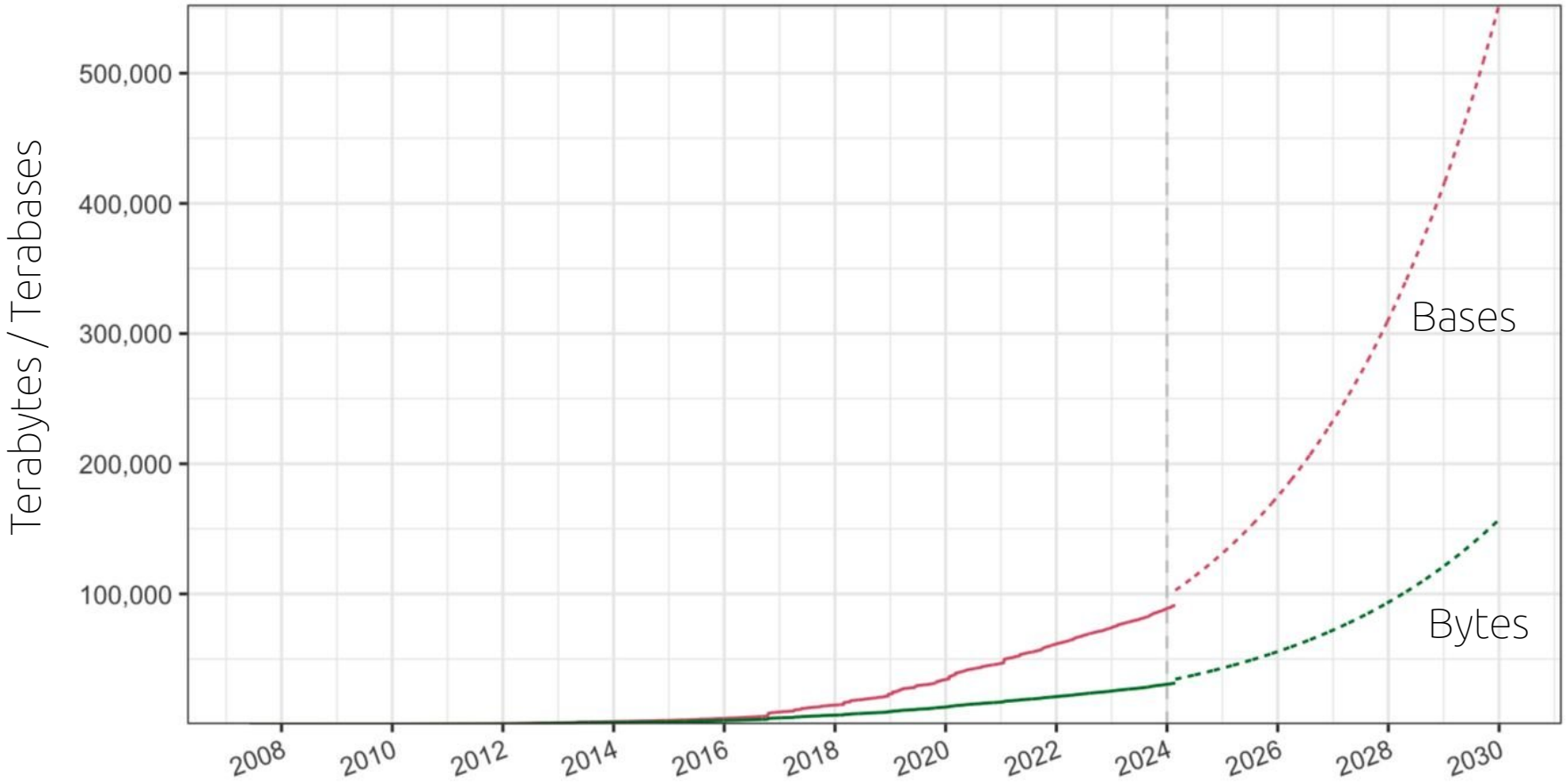
  – short break –

- 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 Read Archive data and prediction (exponential)



https://doi.org/10.1101/2024.06.21.599698

Metagenomics.

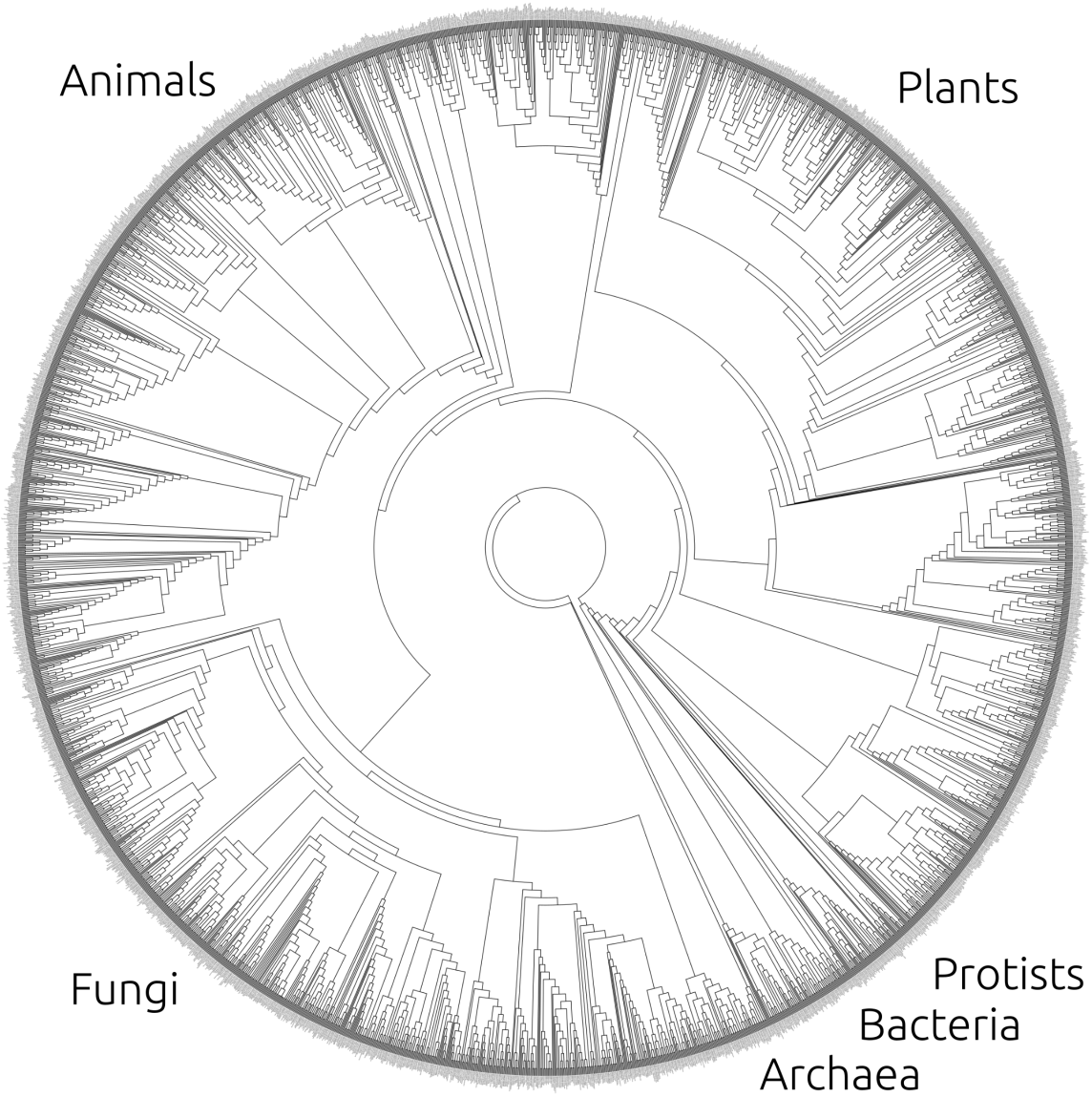
The study of genetic material recovered directly from environmental (or clinical) samples.

Human Microbiome

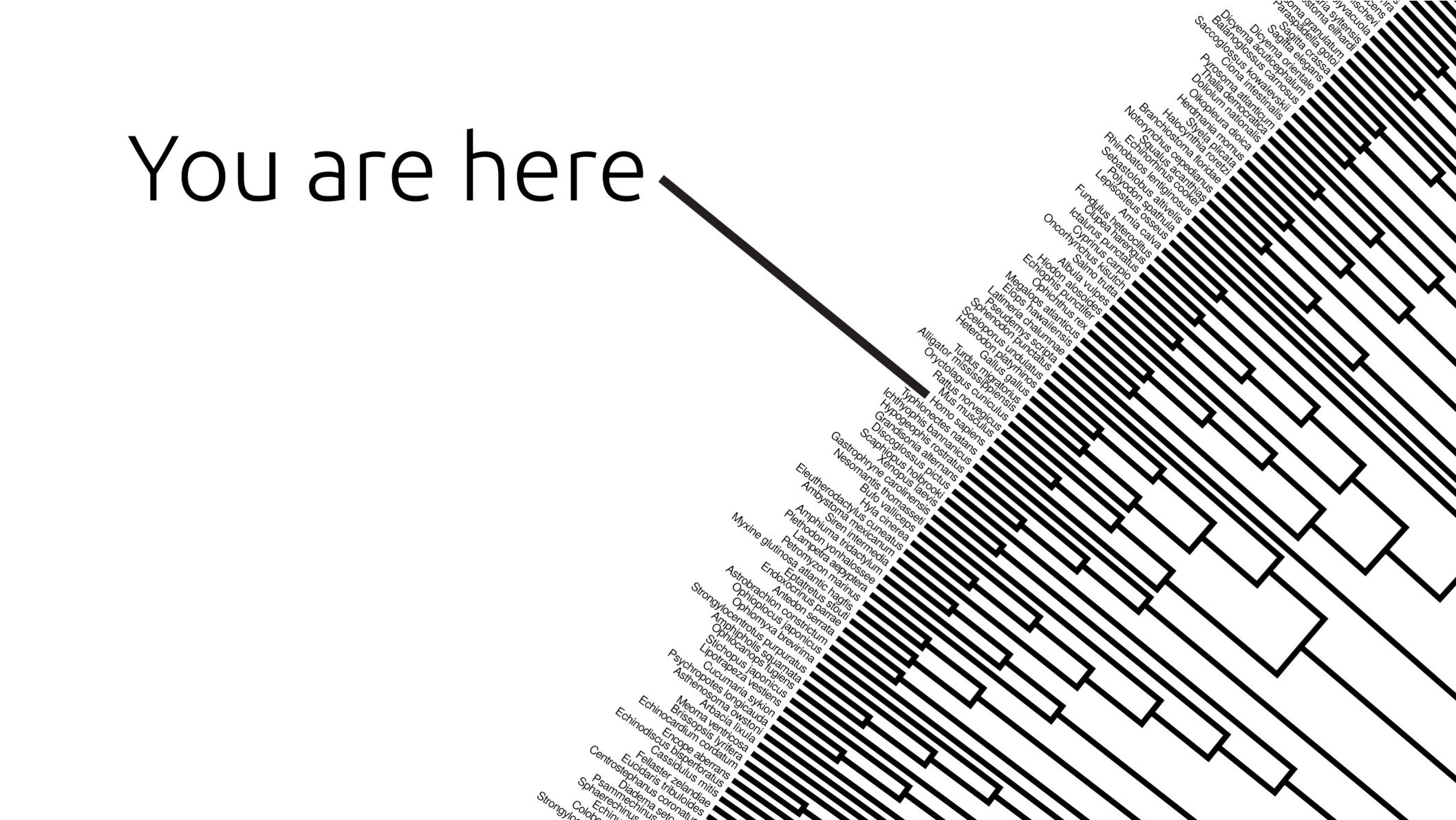
Oceanic Micro-Organisms

Forest Soils

Ancient Ecosystems


Animals
Plants
Fungi
Protists
Bacteria
Archaea

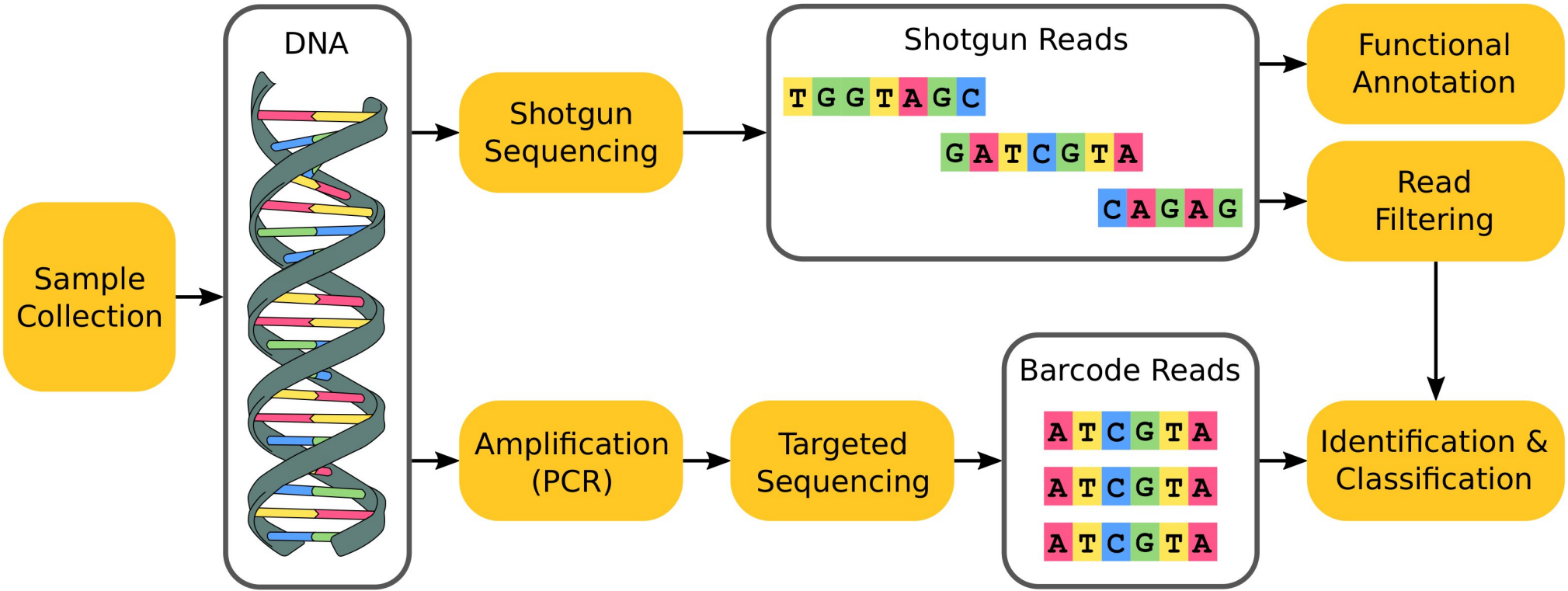
You are here



Typical questions in meta-genomics research

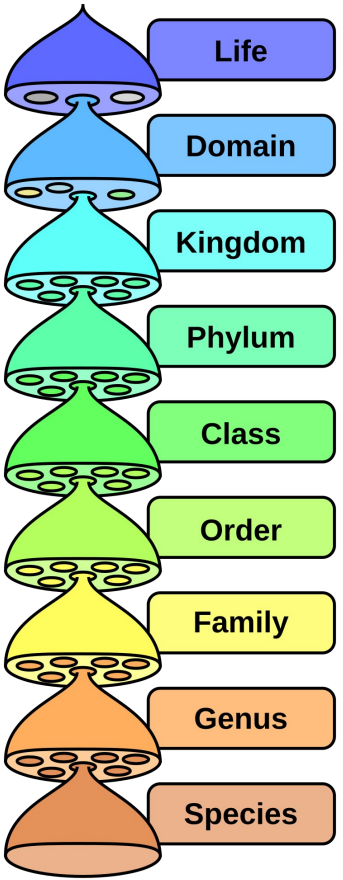
- “Who lives there?” — Genetic composition of samples
- How do samples differ from each other?
- Which environmental factors drive these differences?

Typical meta-genomic pipelines



Common approaches to meta-genomic and eDNA data

- Taxonomic classification
- Taxonomic profiling
- Phylogenetic placement



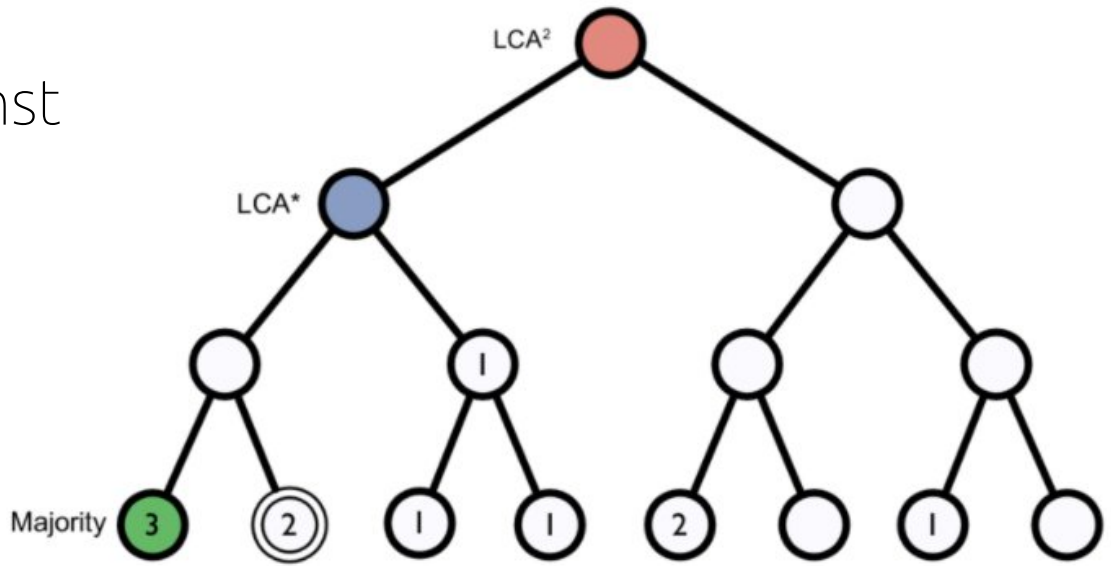
Taxonomic classification

Goals

- Directly assign reads to known taxonomic groups
- Compare sequences against reference databases

Typical tools

- Kraken
- Centrifuge



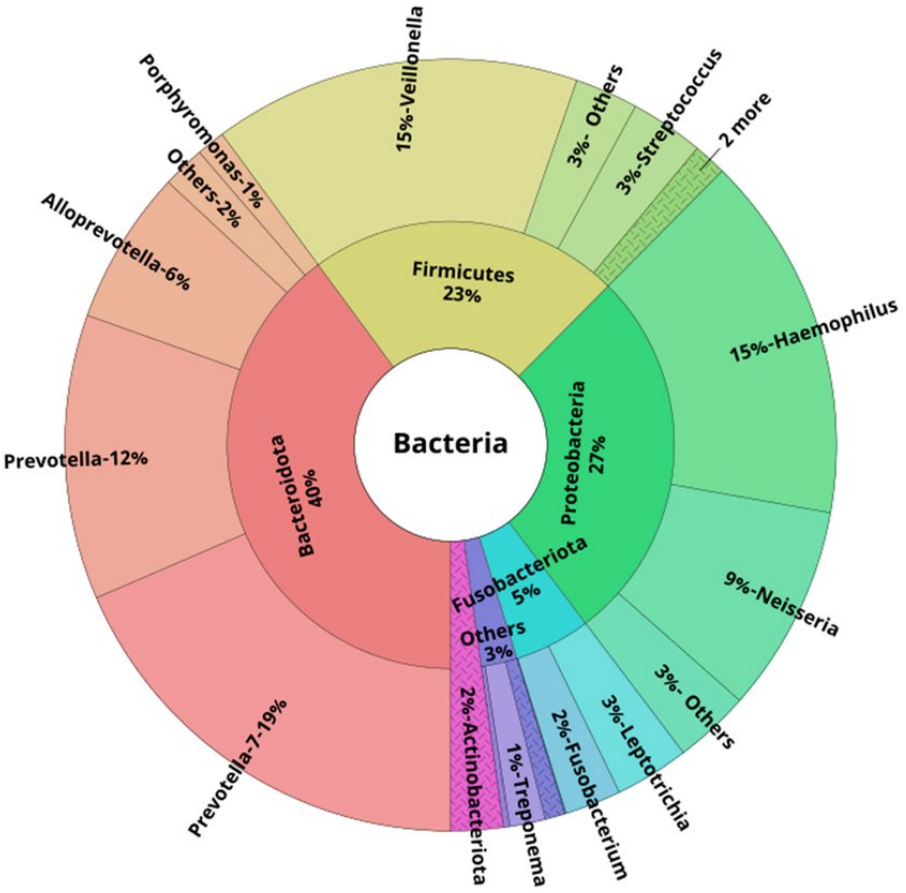
Taxonomic profiling

Goals

- Estimate relative abundances of organisms in a sample
- Use conserved marker genes instead of all reads

Typical tools

- MetaPhlAn
- mOTUs



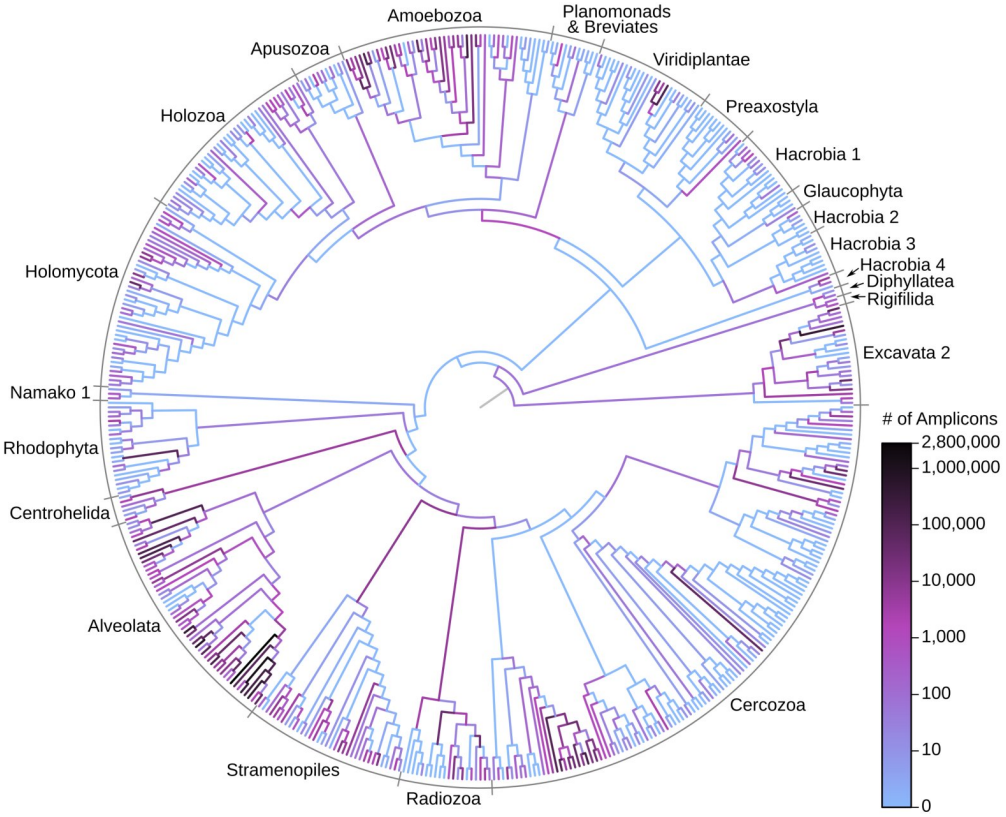
Phylogenetic placement

Goals

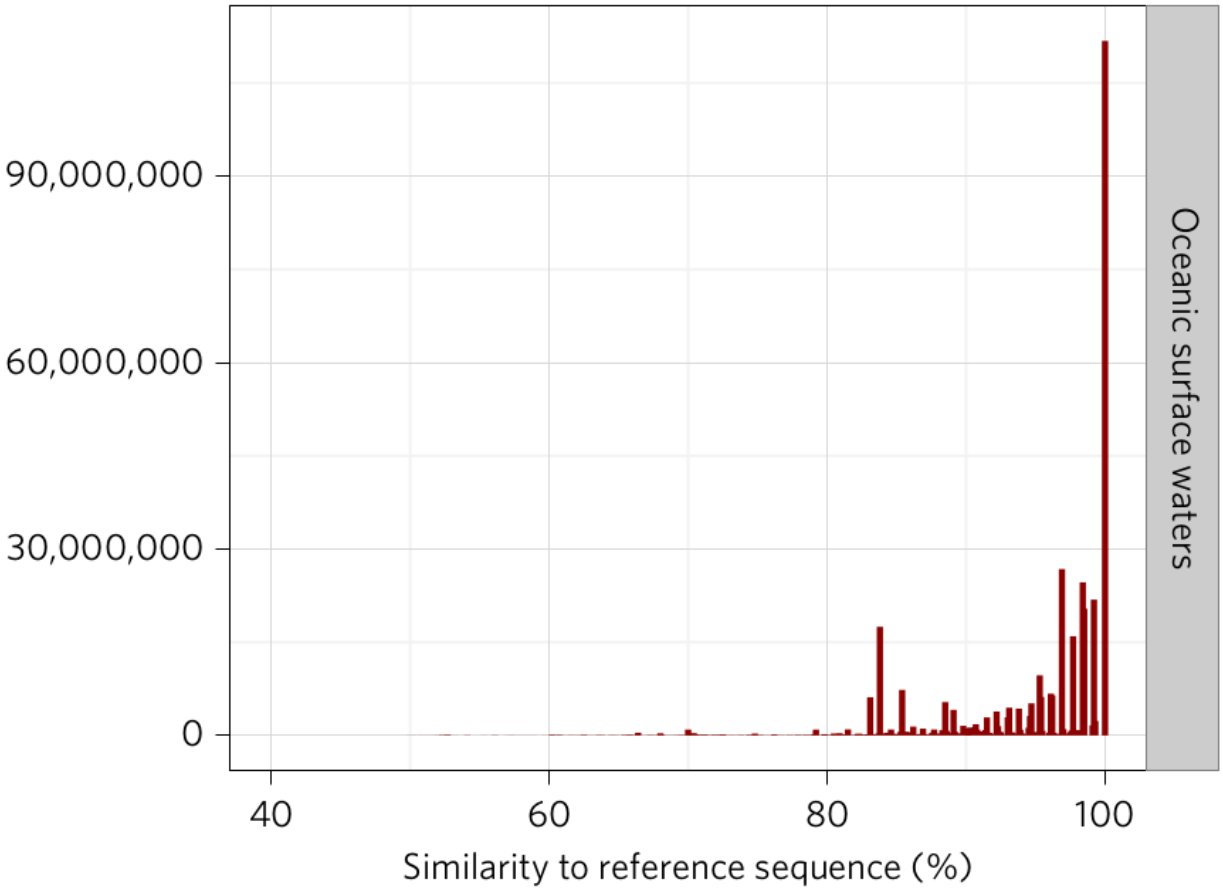
- Place reads into a given phylogenetic reference tree
- Adds evolutionary context to the reads

Typical tools

- EPA-ng
- pplacer

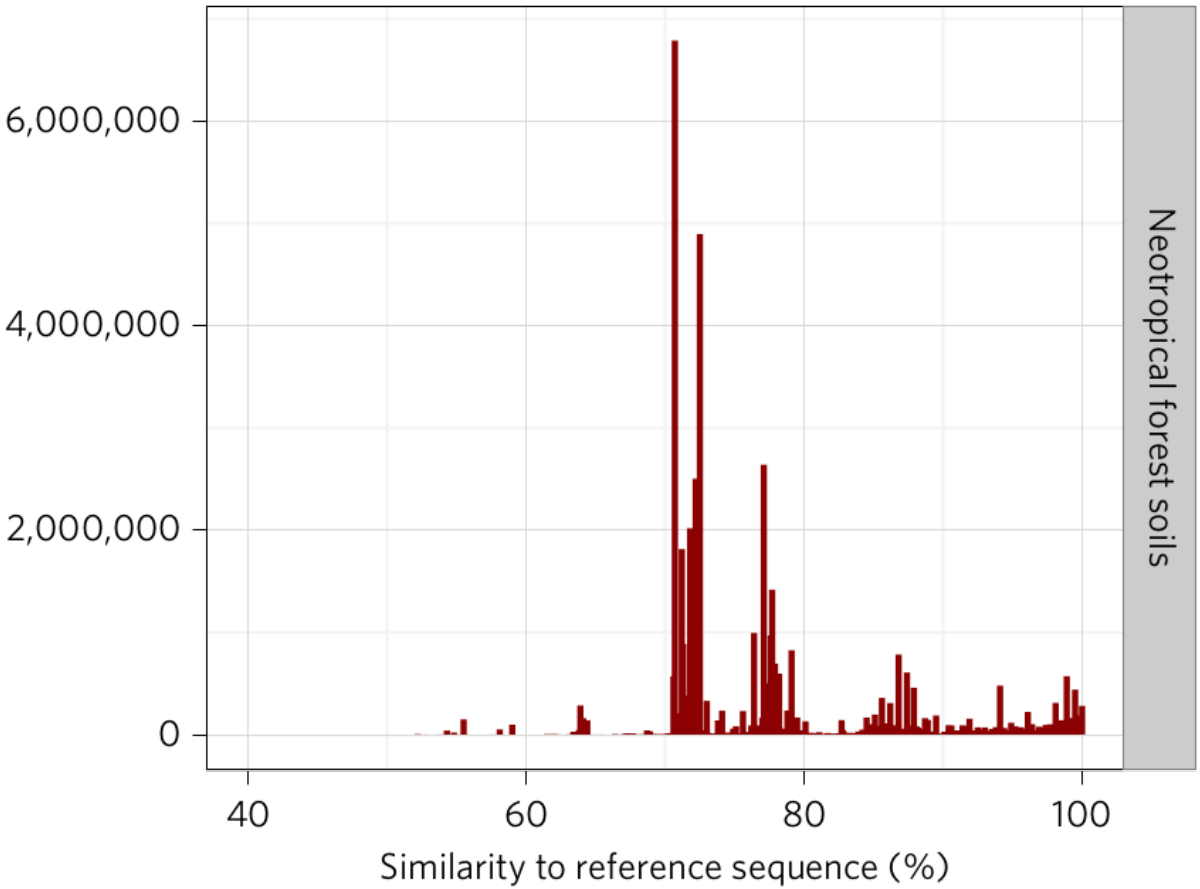


(Almost) complete reference database



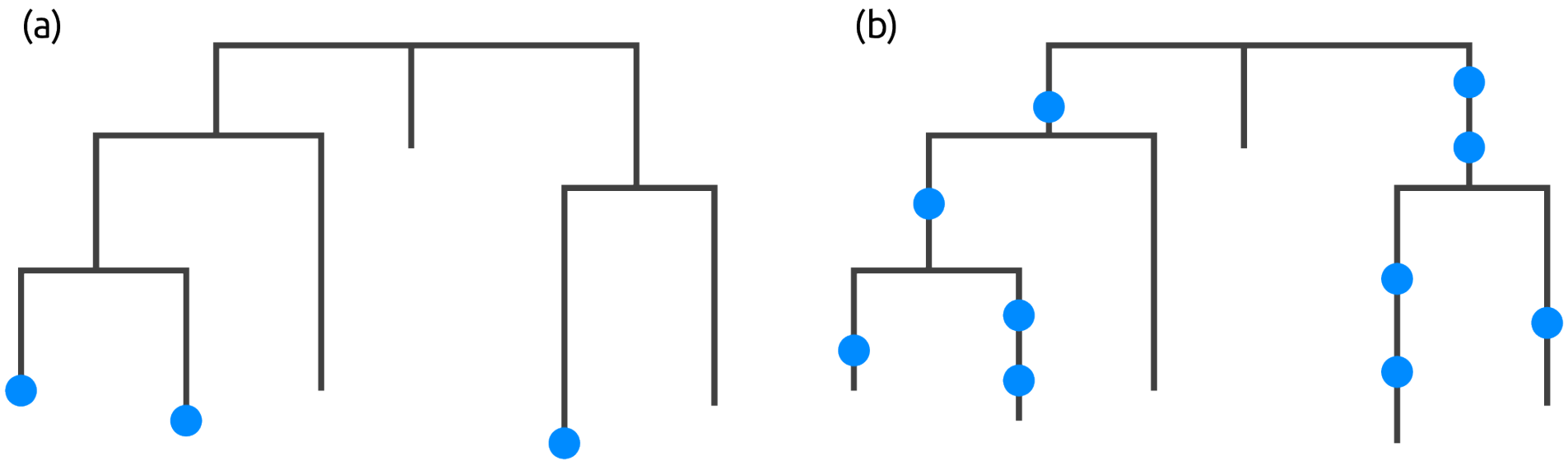
Mahé *et al*, Nature E&E, 2017

Incomplete reference database



Mahé *et al*, Nature E&E, 2017

BLAST / vsearch vs. Phylogenetic Placements



Some applications

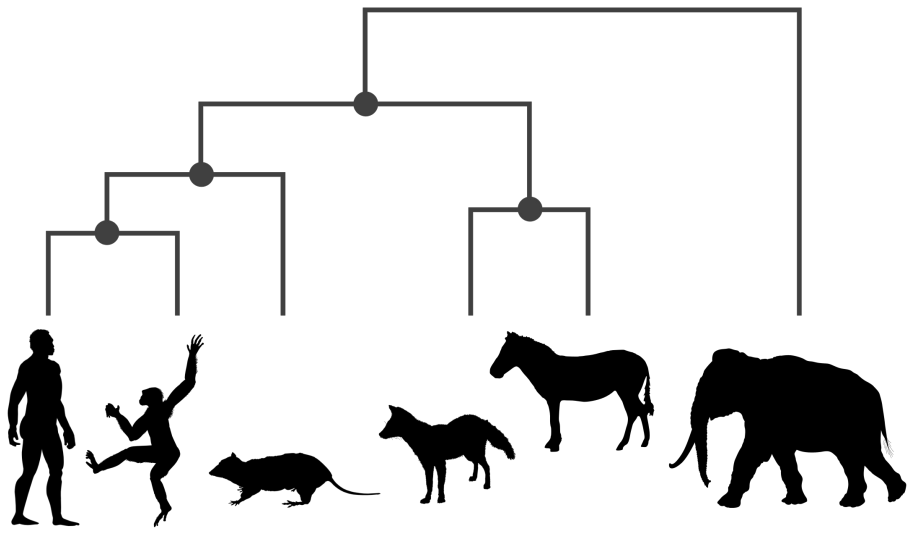
- Data cleaning and retention (Mahé et al., 2017)
- Inference of new clades (Dunthorn et al., 2014; Bass et al., 2018)
- Estimation of ecological profiles (Keck et al., 2018)
- Identification of low-coverage genomes of viral strains (Mühlemann et al., 2020)
- Phylogenetic analysis of viruses such as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

Phylogenetic Tree Inference

MSA and Phylogenetic Tree

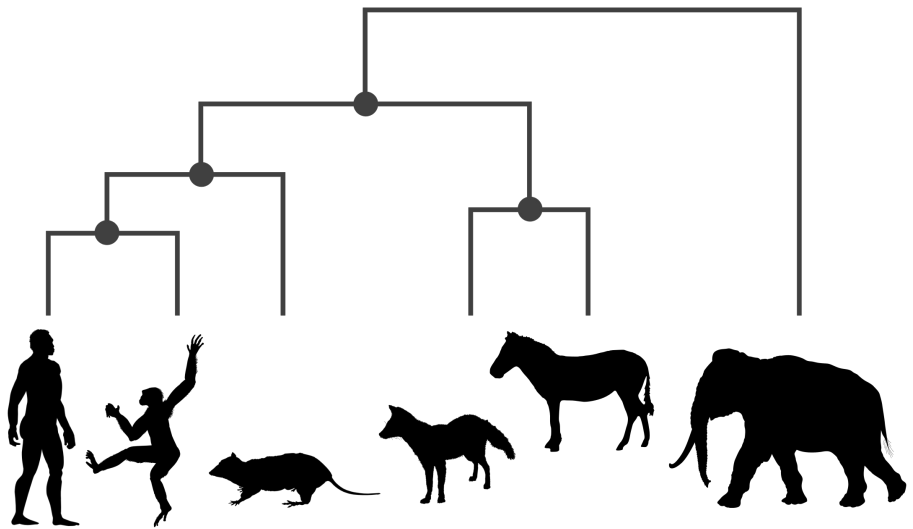
Multiple Sequence Alignment (MSA)

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



Tree inference methods

- Distance-based methods
  - UPGMA
  - Neighbor-Joining
- Character-based methods
  - Maximum Parsimony
  - Maximum Likelihood ← today
  - Bayesian Inference
- Other approaches
  - Coalescent-based methods



Maximum Likelihood Estimation

Find the values of the model parameters
that maximize the likelihood function
over the parameter space

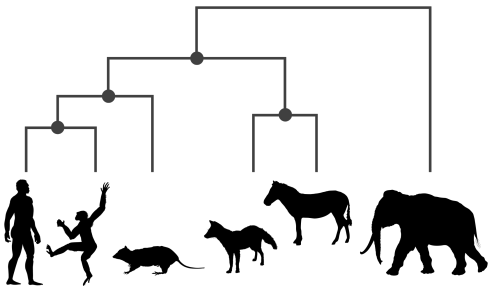
$$\hat{\theta} = \underset{\theta \in \Theta}{\arg\max}\, \mathcal{L}_n(\mathbf{y} \mid \theta)$$

Under which parameters is it most likely that we observe our data?!

Maximum Likelihood Tree Inference

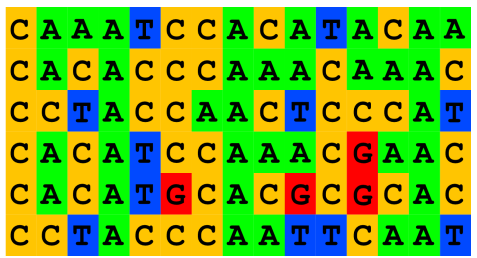
Find phylogenetic tree:
maximize the likelihood of producing the given MSA

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



with

- T: Tree
- b: Branch lengths
- M: Model of evolution
- θ: Model parameters

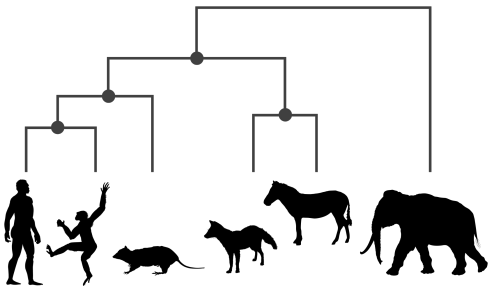


Note that this is the reverse of the intuitive direction!

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
  - ...

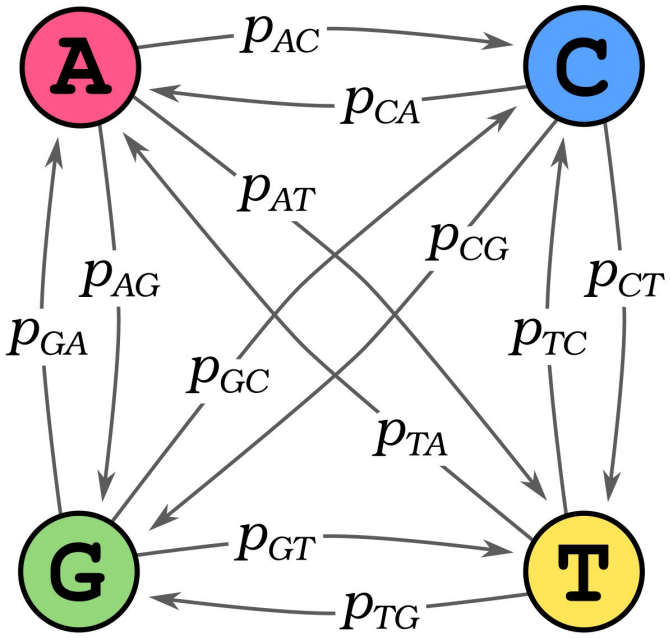
Model of Nucleotide Substitution

- Assumption: Columns/sites of the MSA evolved independently!
- Assumption: (Evolutionary) time is reversible!

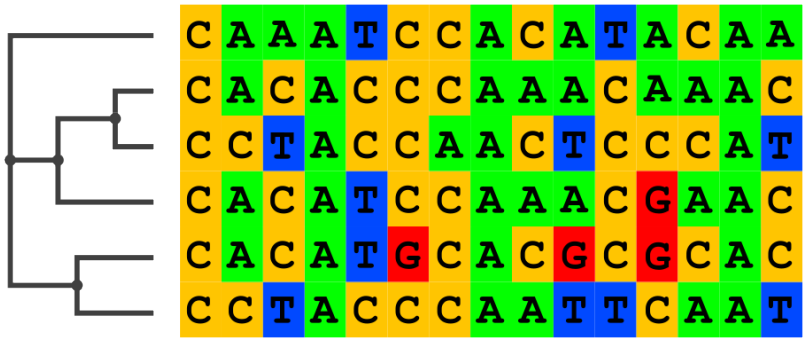
- How did the sequences evolve?
  - Need a model to estimate of the evolutionary distance between sequences
  - As we assume homologous loci (columns of the MSA), we only consider mutations (no insertions or deletions)
  - We use a continuous-time Markov chain (MC) model

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



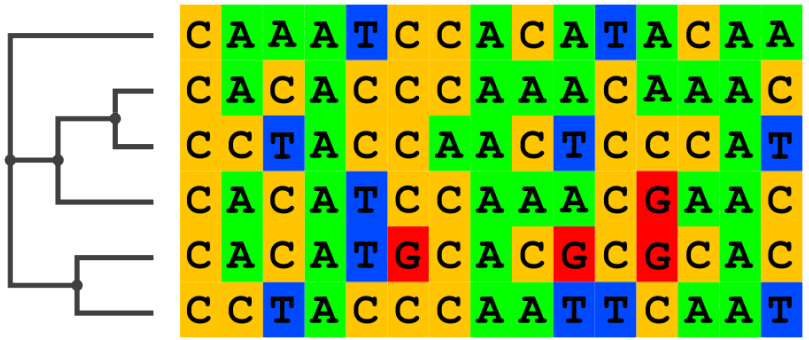
Likelihood Computation



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

Likelihood Computation

- 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)

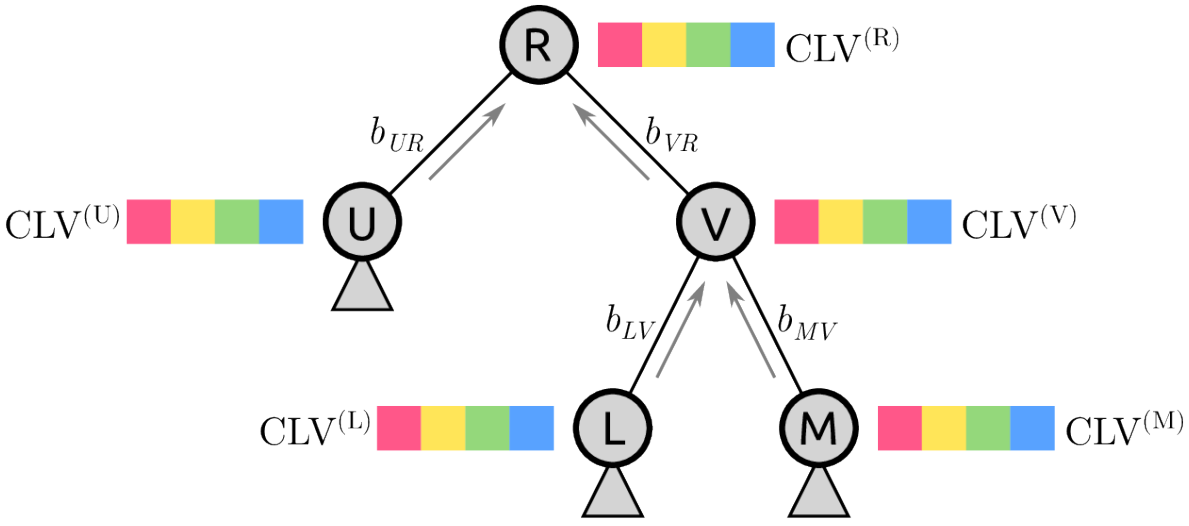
The CLV "summarizes" the subtree below its node:

- 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

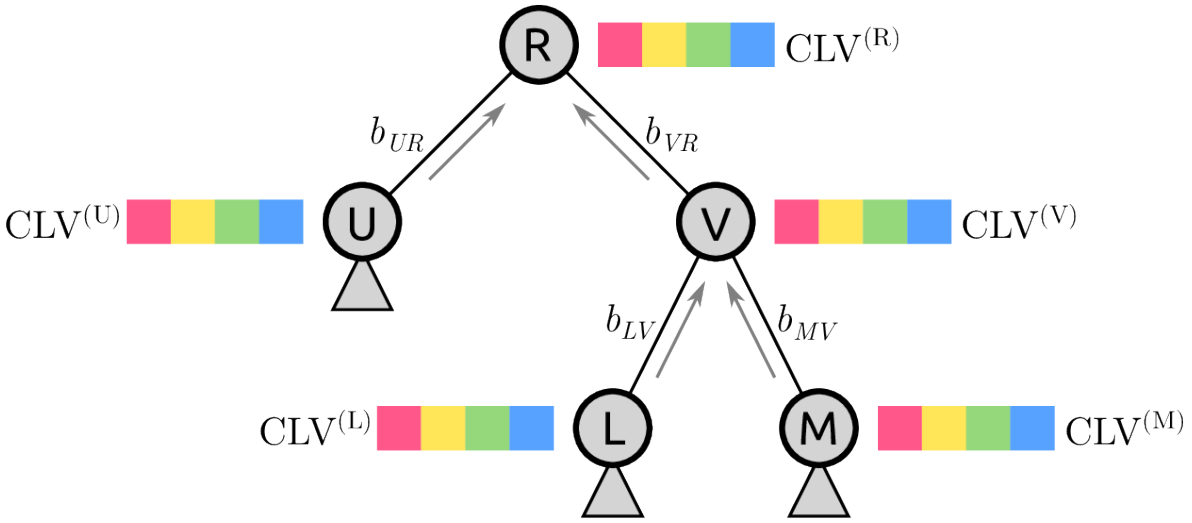
Felsenstein Pruning Algorithm

- 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



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

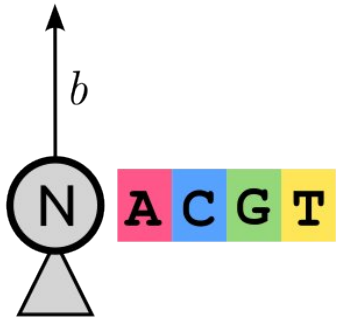
$$\mathrm{CLV}_{s,c}^{(V)} = \left(\sum_{j \in N} p_{cj}(r \cdot b_{LV}) \cdot \mathrm{CLV}_{s,j}^{(L)} \right) \left(\sum_{k \in N} p_{ck}(r \cdot b_{MV}) \cdot \mathrm{CLV}_{s,k}^{(M)} \right)$$

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! :-(

Agenda

- Motivation
- Phylogenetic Tree Inference

  – short break –

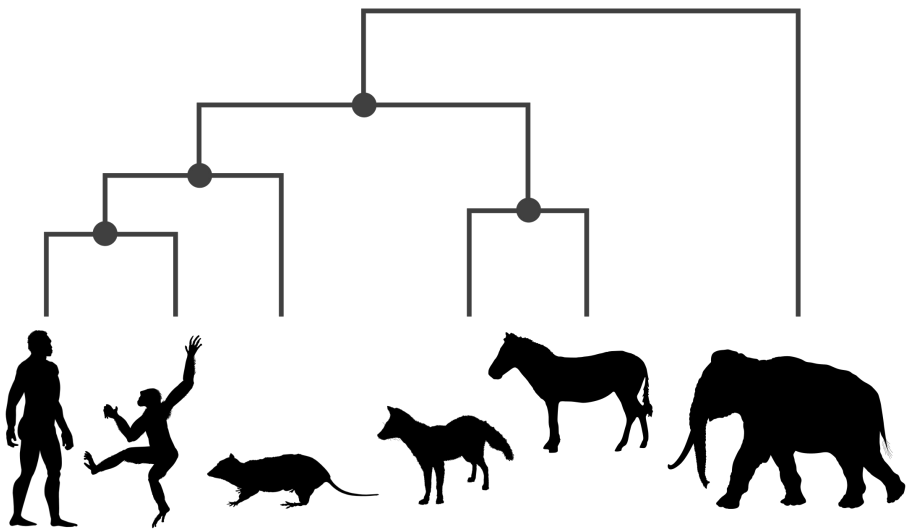
- Phylogenetic Placement
- Placement Analysis and Visualization

Phylogenetic Placement

MSA and Phylogenetic Tree

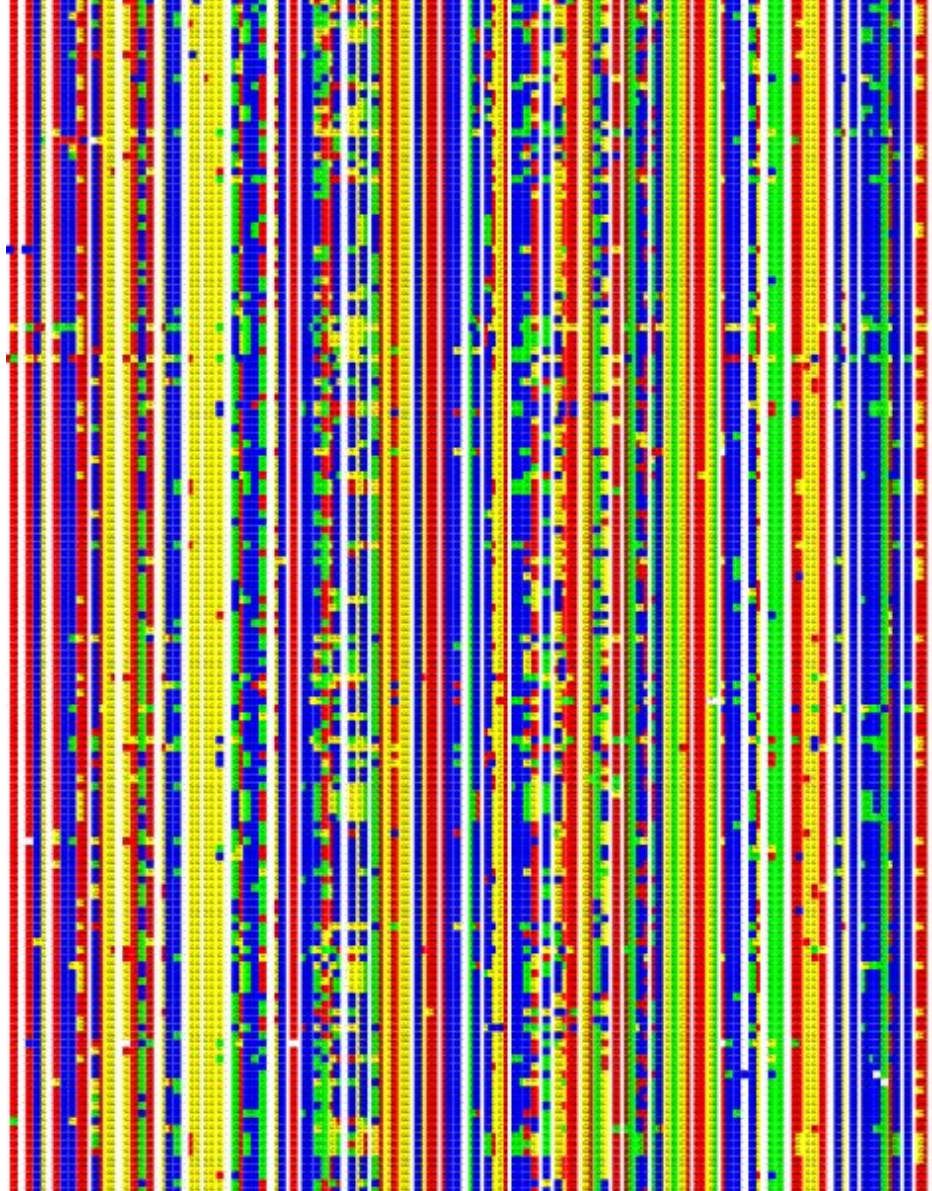
Multiple Sequence Alignment (MSA)

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

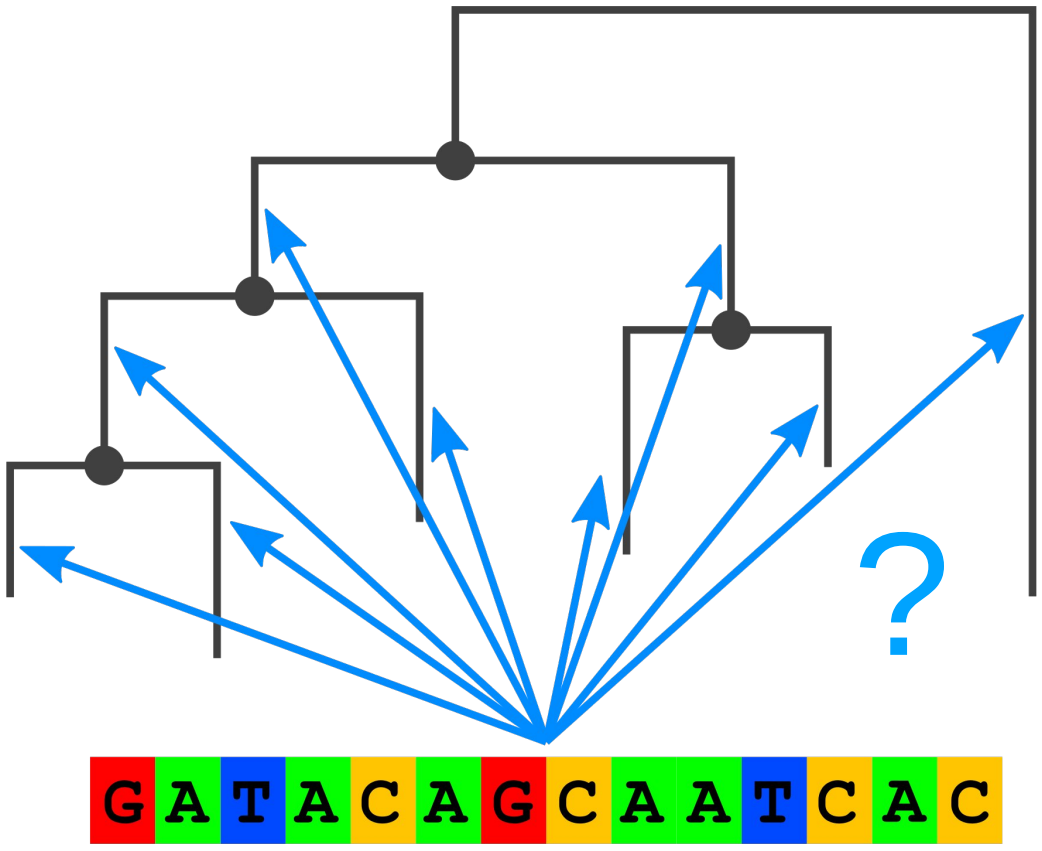


Multiple Sequence Alignment (MSA)

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

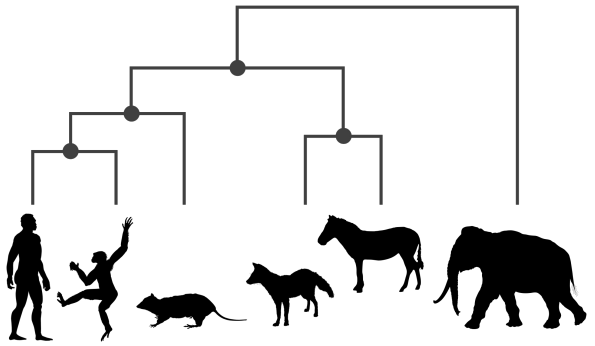


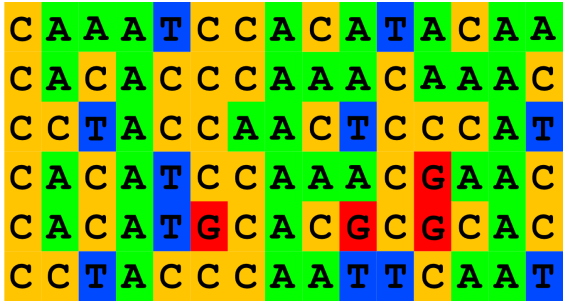
Phylogenetic Placement



Phylogenetic Placement

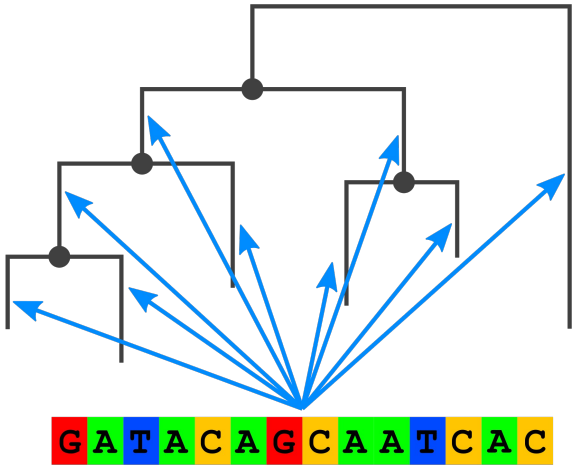
- Given:
  - *Reference tree* and MSA
  - Set of *query sequences*
- Typically, we target conserved marker regions
- Queries can be amplicons, OTUs, etc





Phylogenetic Placement

- For each sequence:
  - Try out each branch as a potential *placement location*
  - Compute how likely this location is



- Repeat for all sequences

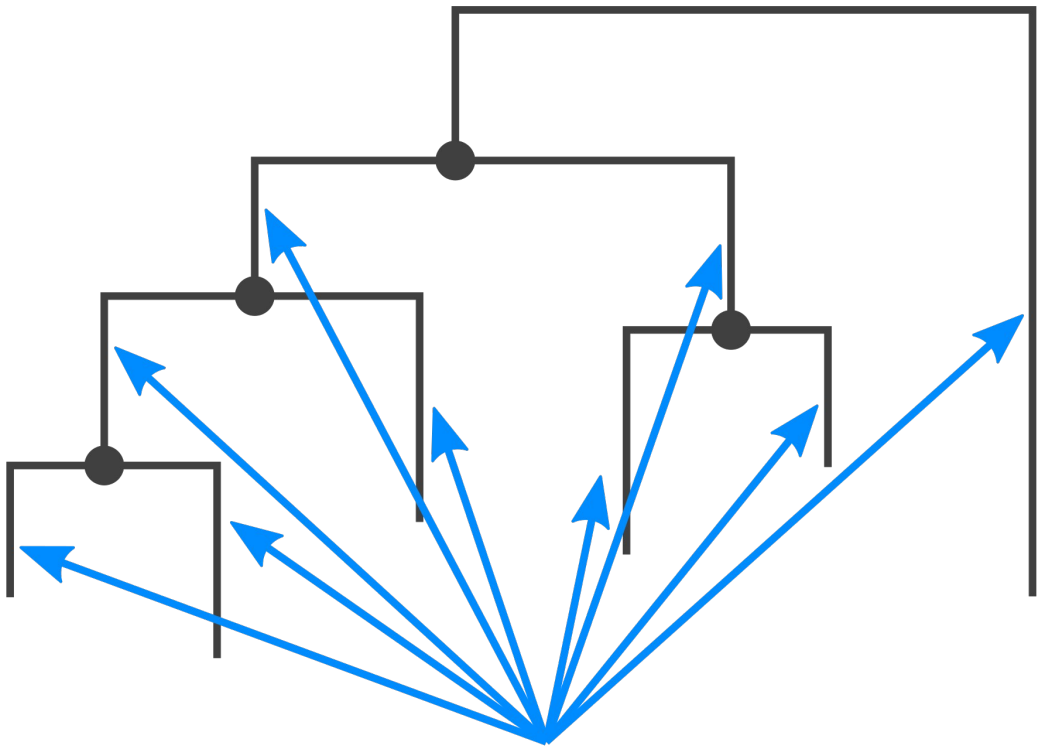
  → mapping from sequences to branches of the reference tree

- 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

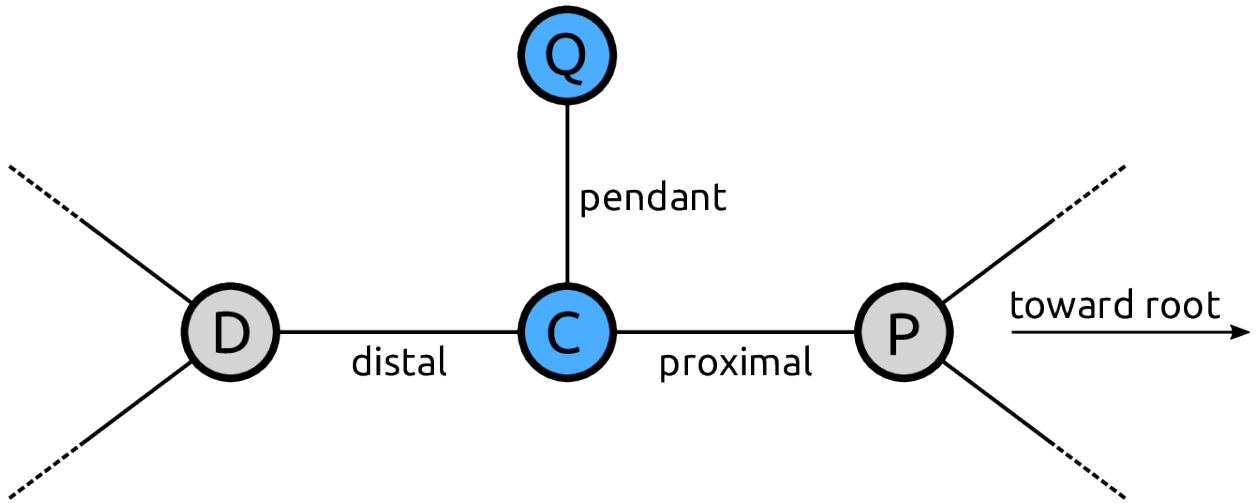


Single Sequence:

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)

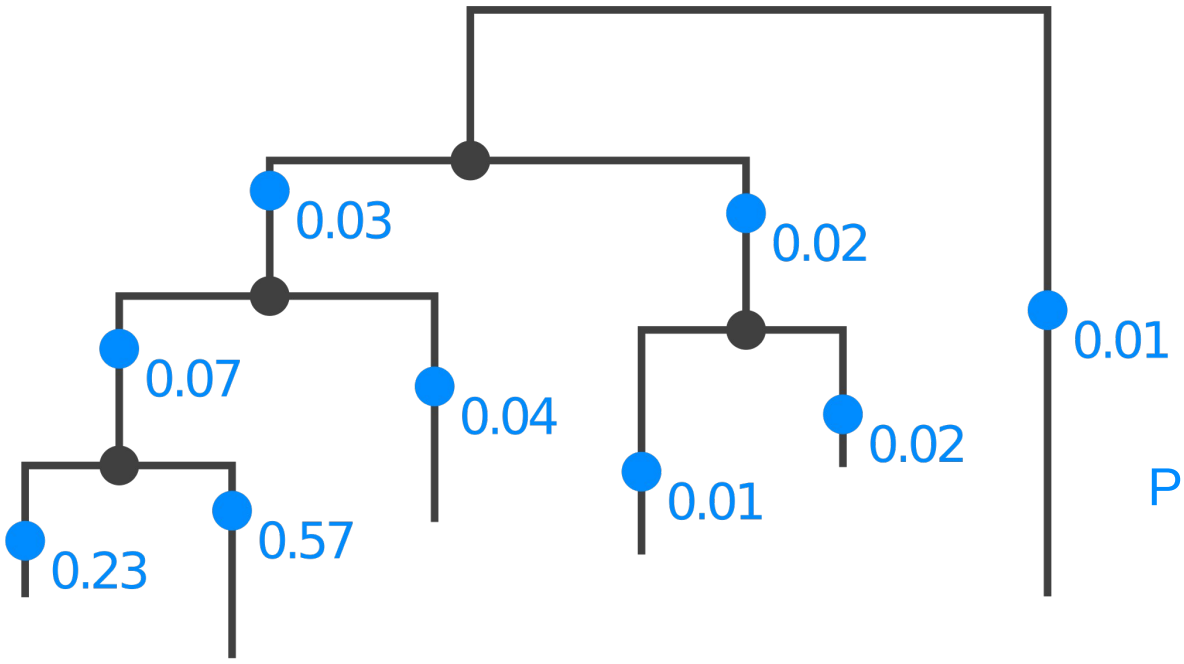
Likelihood Computation

- 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

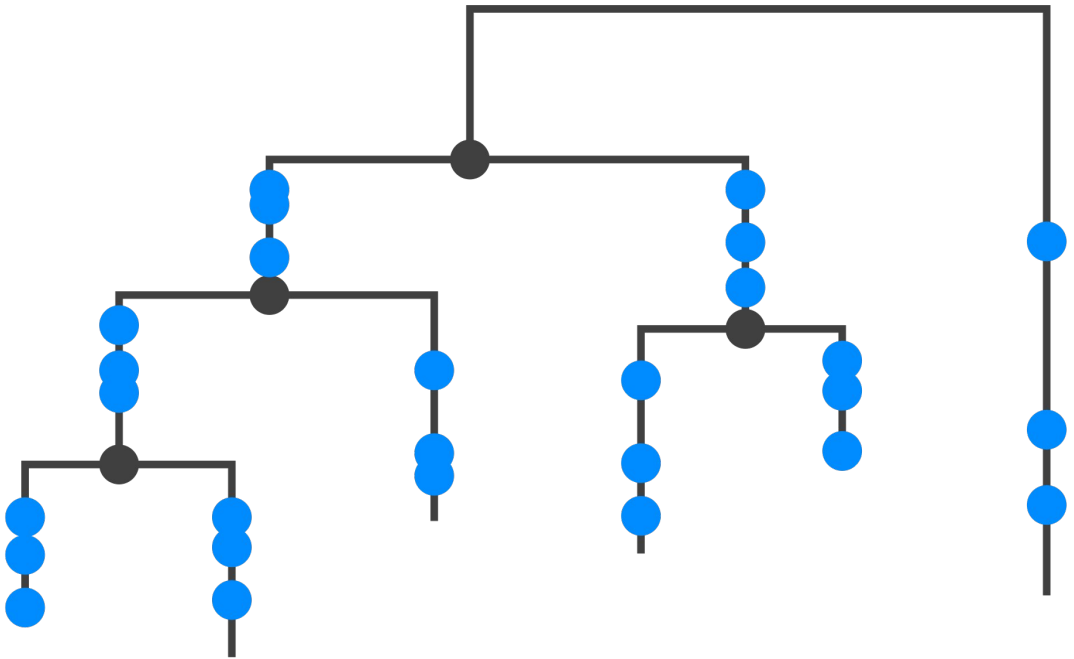
Phylogenetic Placement



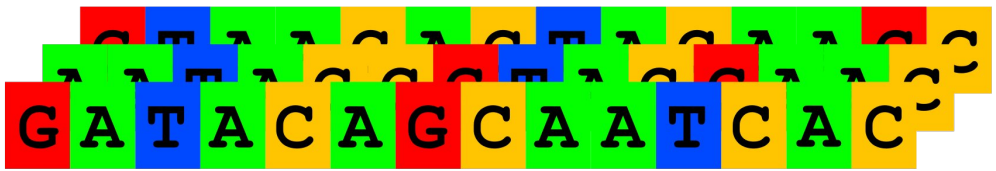
Placement Masses
≙
Probabilities

Single Sequence:

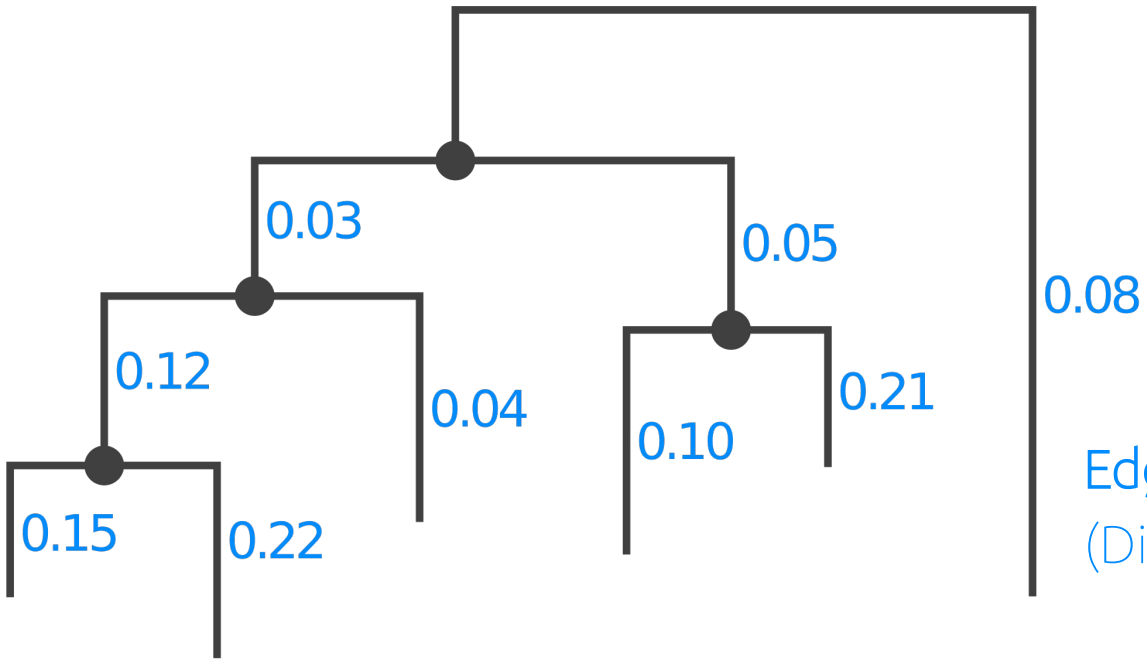
Phylogenetic Placement



Whole Sample:

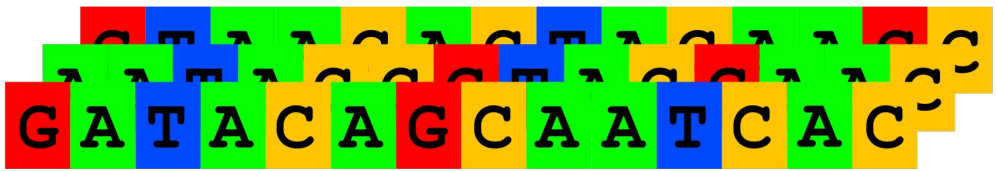


Phylogenetic Placement

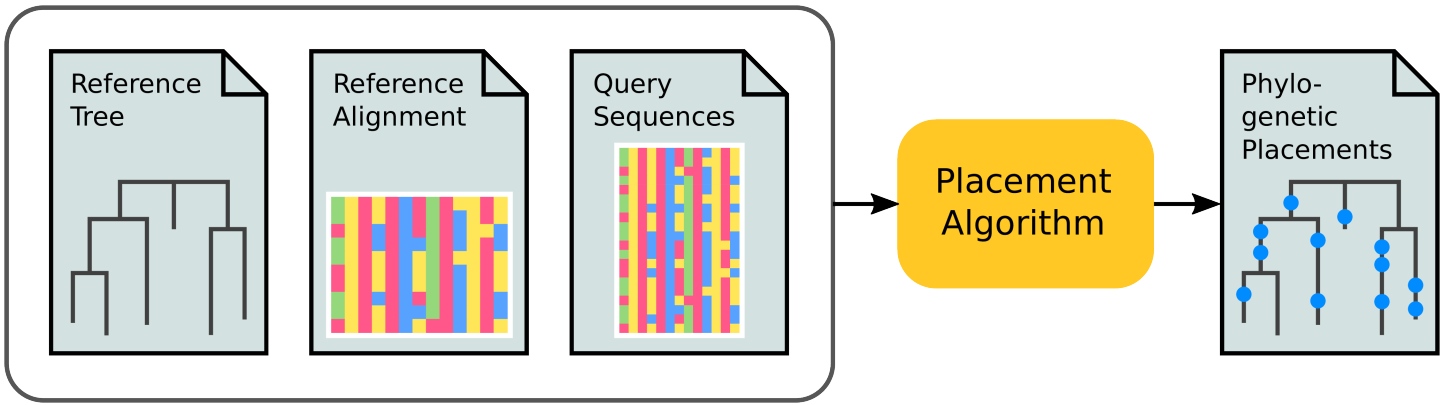


Edge Masses
(Distribution on the Tree)

Whole Sample:



Phylogenetic Placement Pipeline



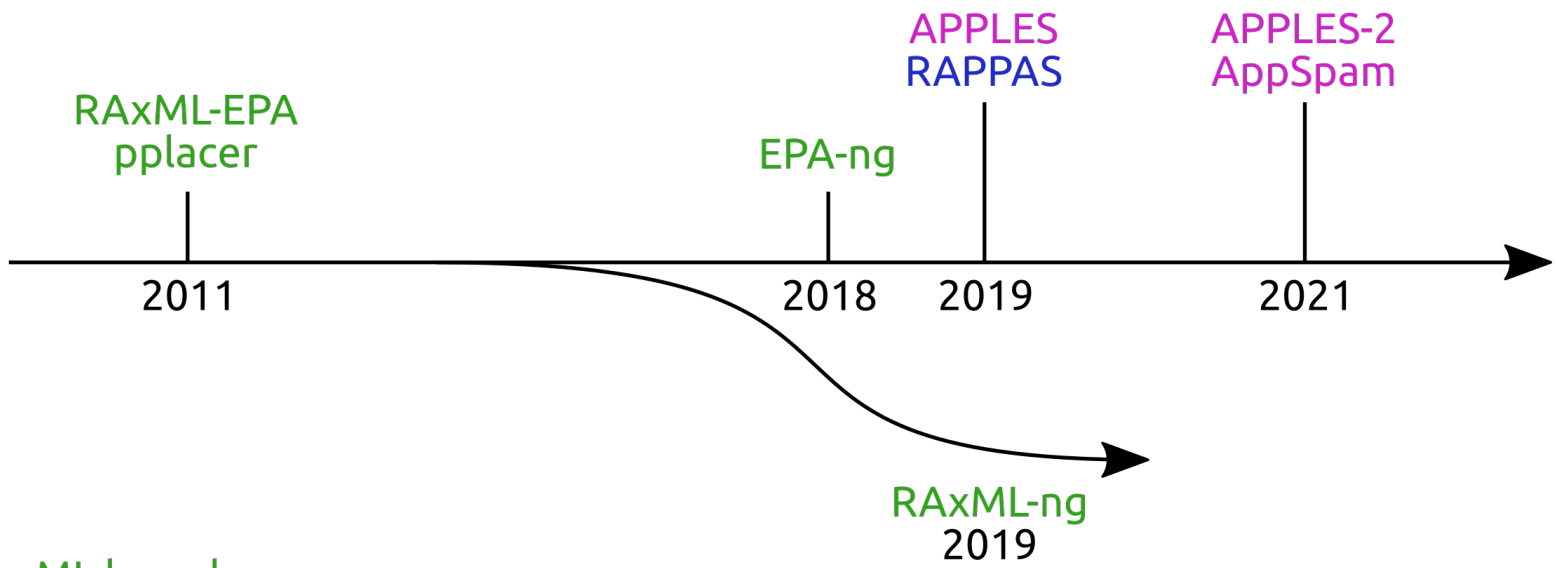
Input:

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

Output:

- Placements (jplace)

New: alignment-free (k-mer based) placement



Czech *et al*, Frontiers in Bioinformatics, 2022

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

Czech *et al*, Frontiers in Bioinformatics, 2022

Placement Analysis and Visualization

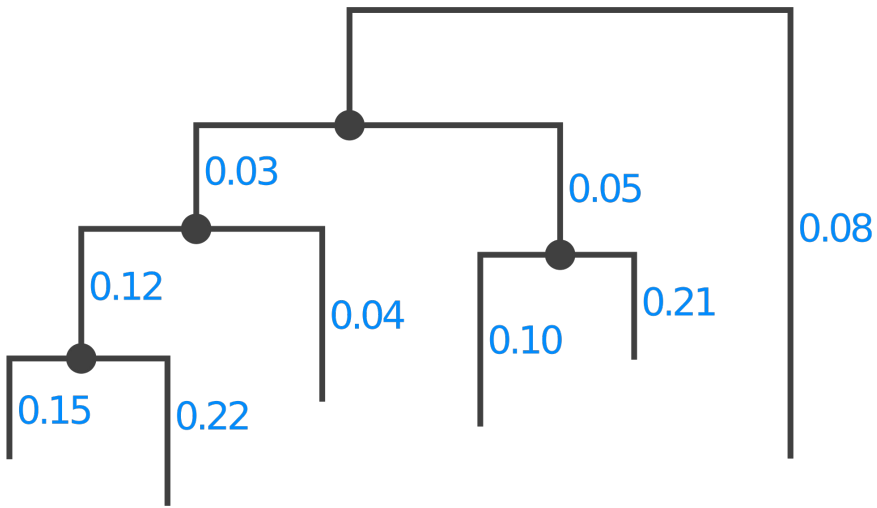
Placement Analysis

Two interpretations:

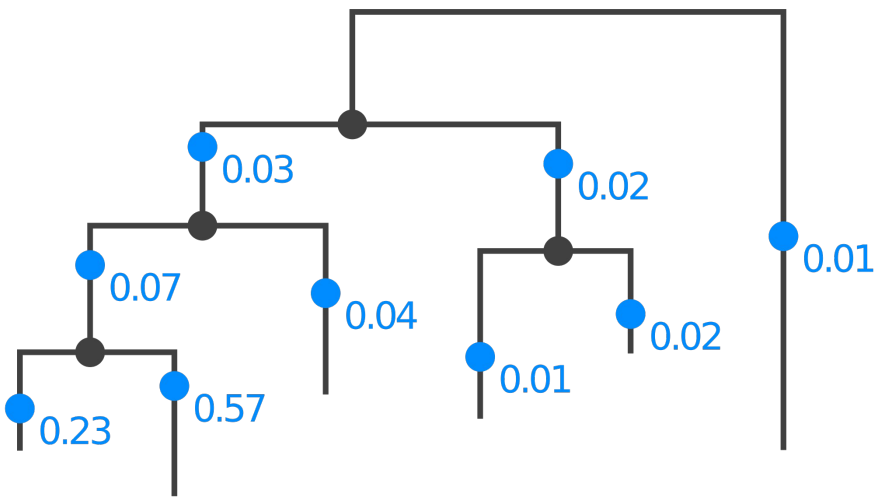
- Assignment (or mapping) of sequences to branches
- Distribution of sequences across the tree → today

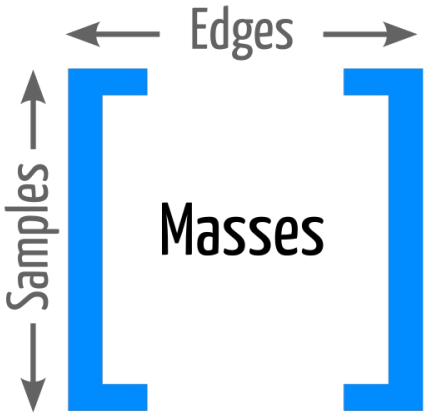
Typical types of analysis:

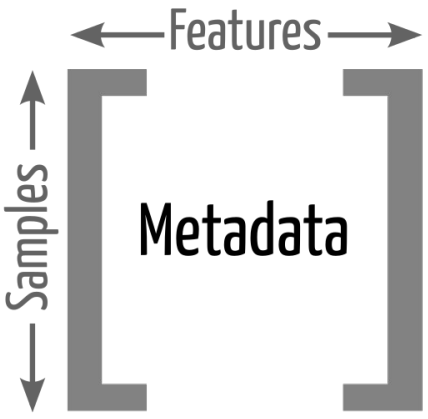
- Examine a single sample
- Relate multiple samples to each other
- Relate samples to environmental factors / variables



Masses for all Samples and Edges







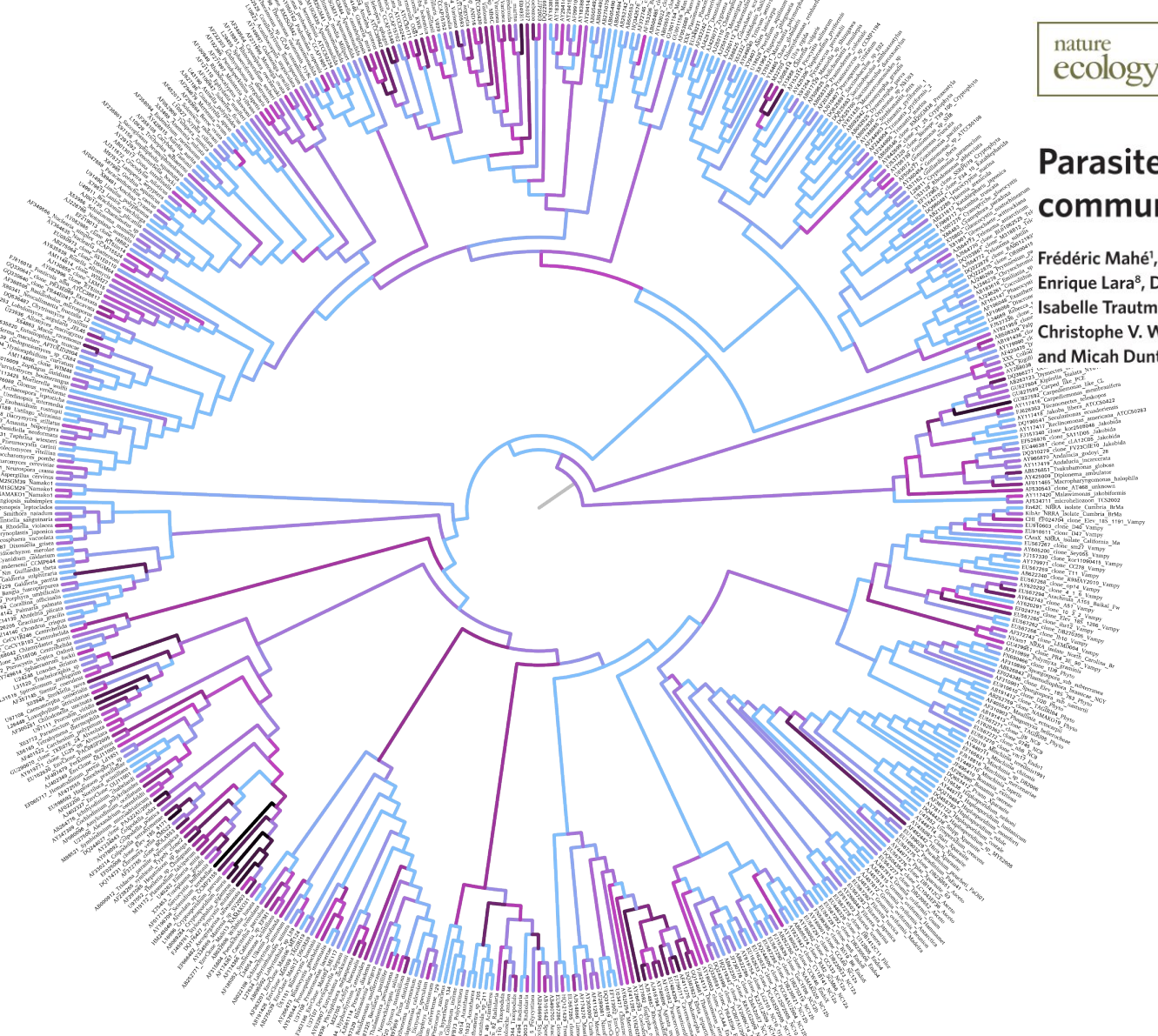
nature ecology & evolution

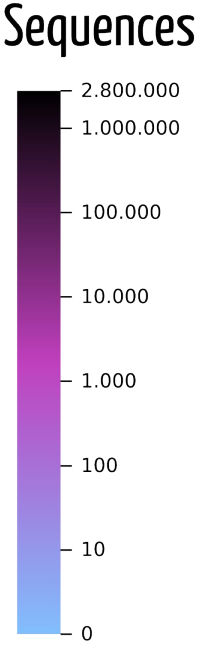
ARTICLES

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

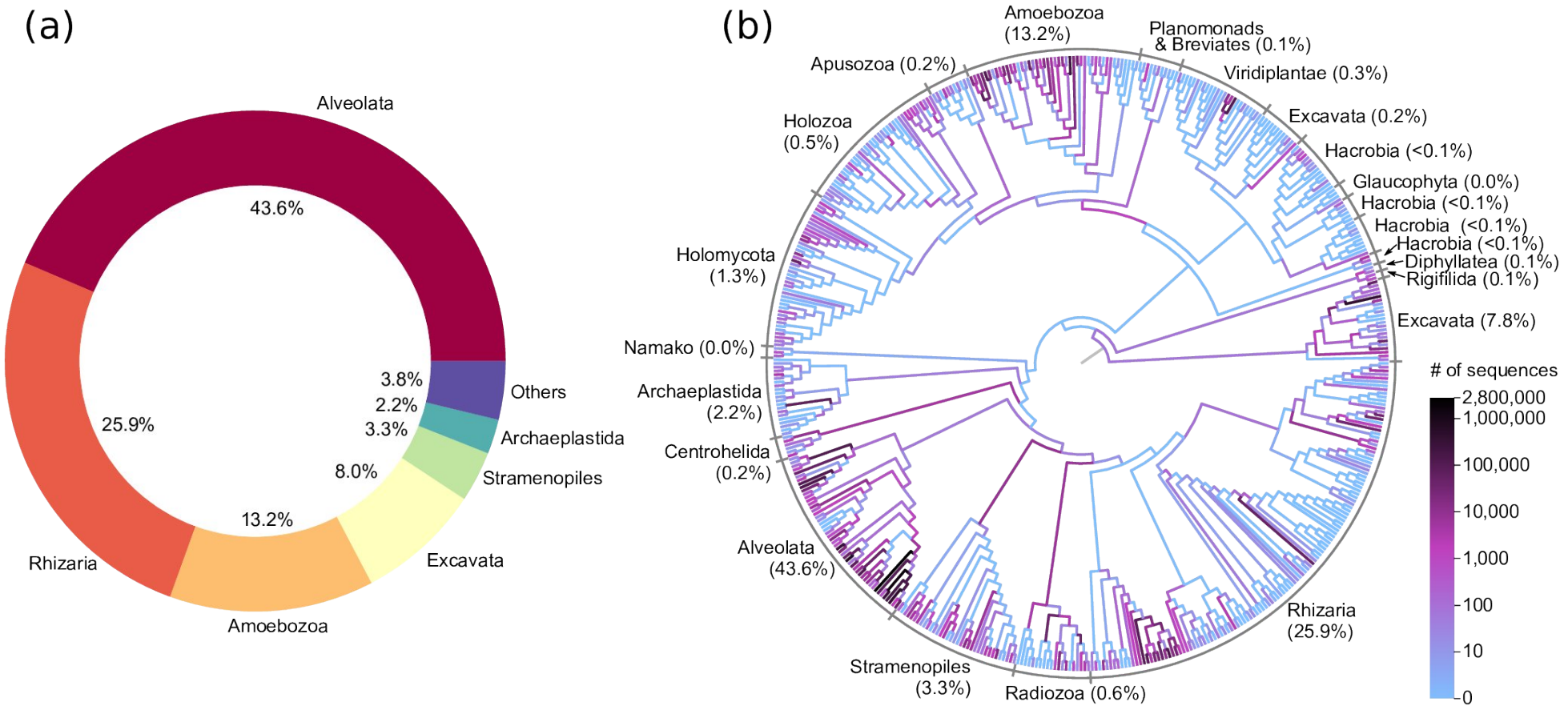
Parasites dominate hyperdiverse soil protist communities in Neotropical rainforests

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]*

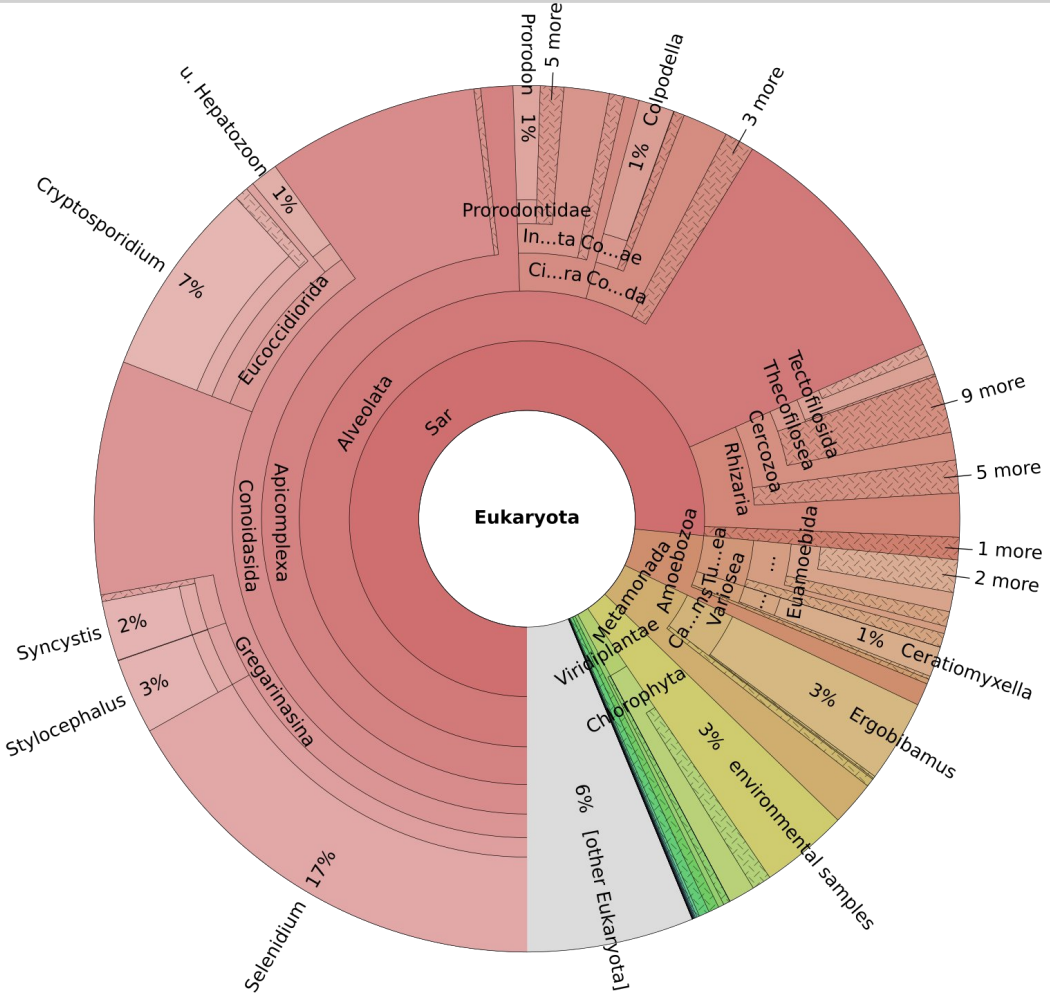




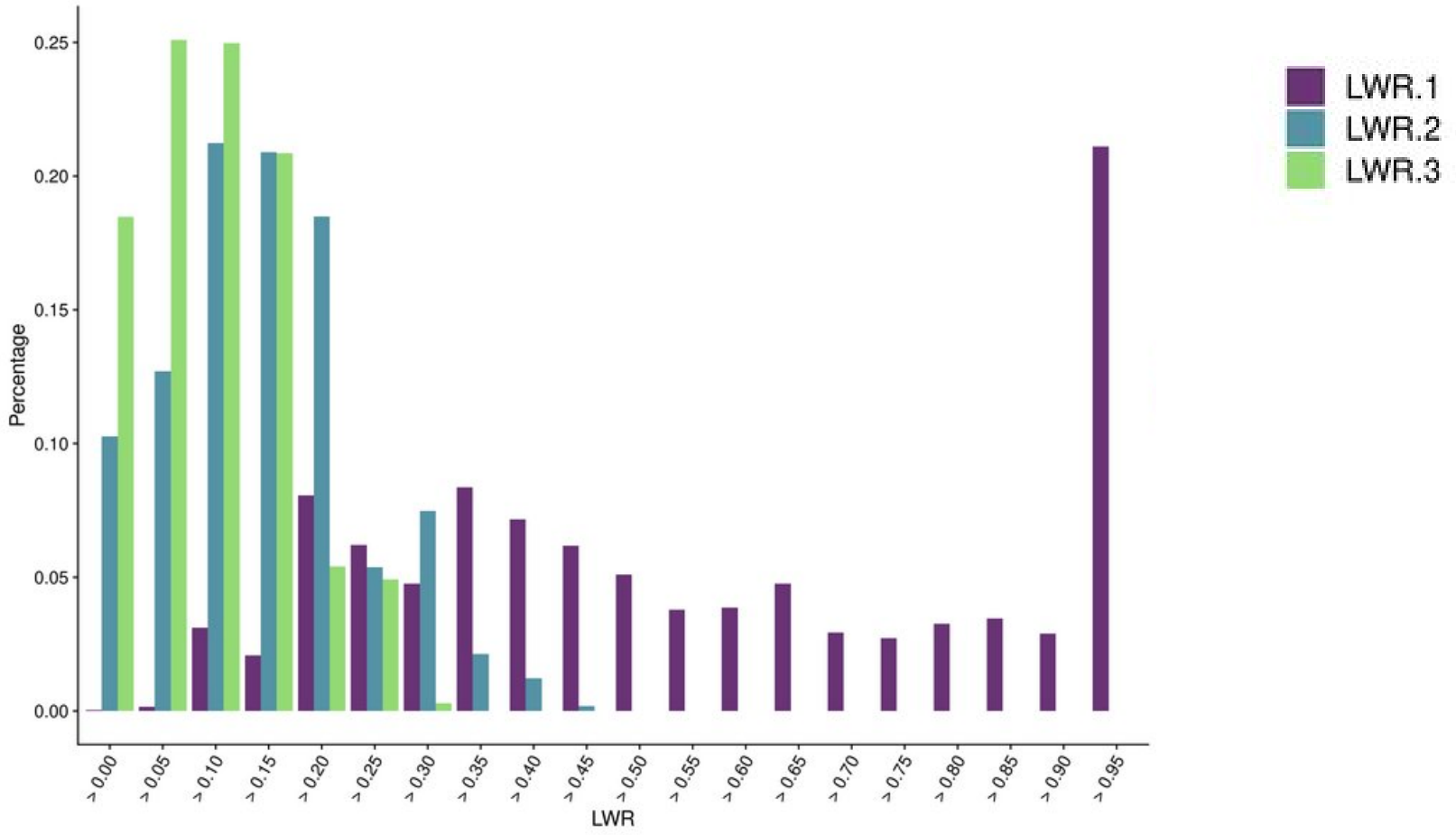
Abundances vs. Phylogenetic Placements



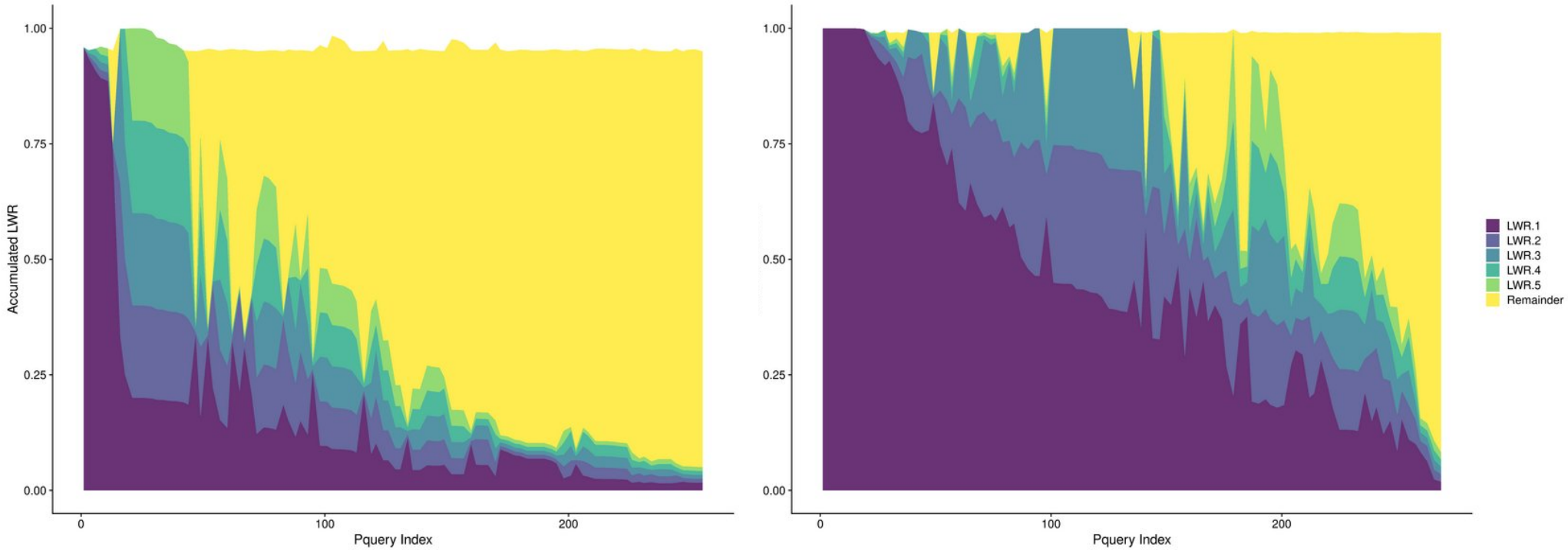
Taxonomic Assignment



LWR Distribution



LWR Distribution

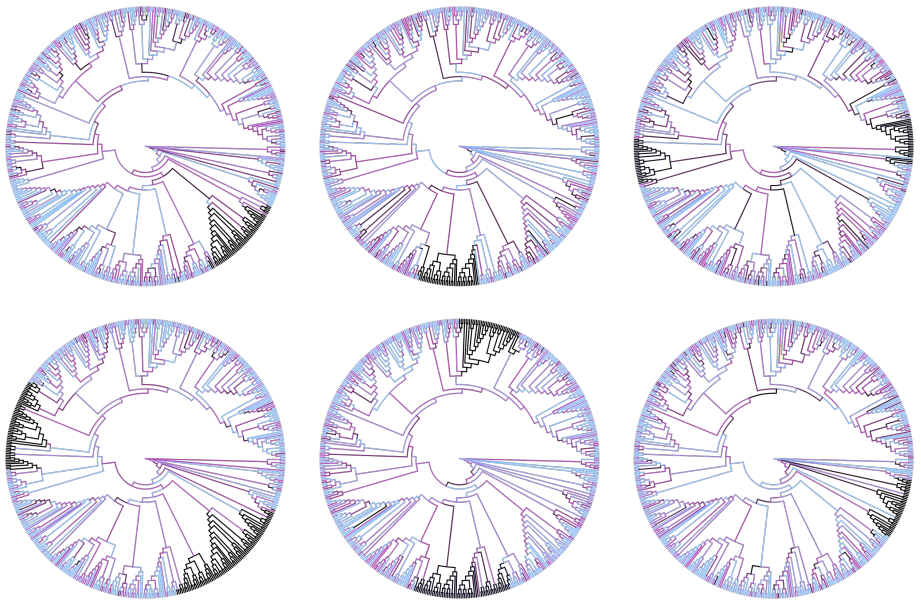


Placement of Multiple Samples

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

Typically: Meta-data per sample

- pH value
- Temperature
- …



Phylogenetic Placement as applied to a disease

OPEN ACCESS Freely available online

PLoS one

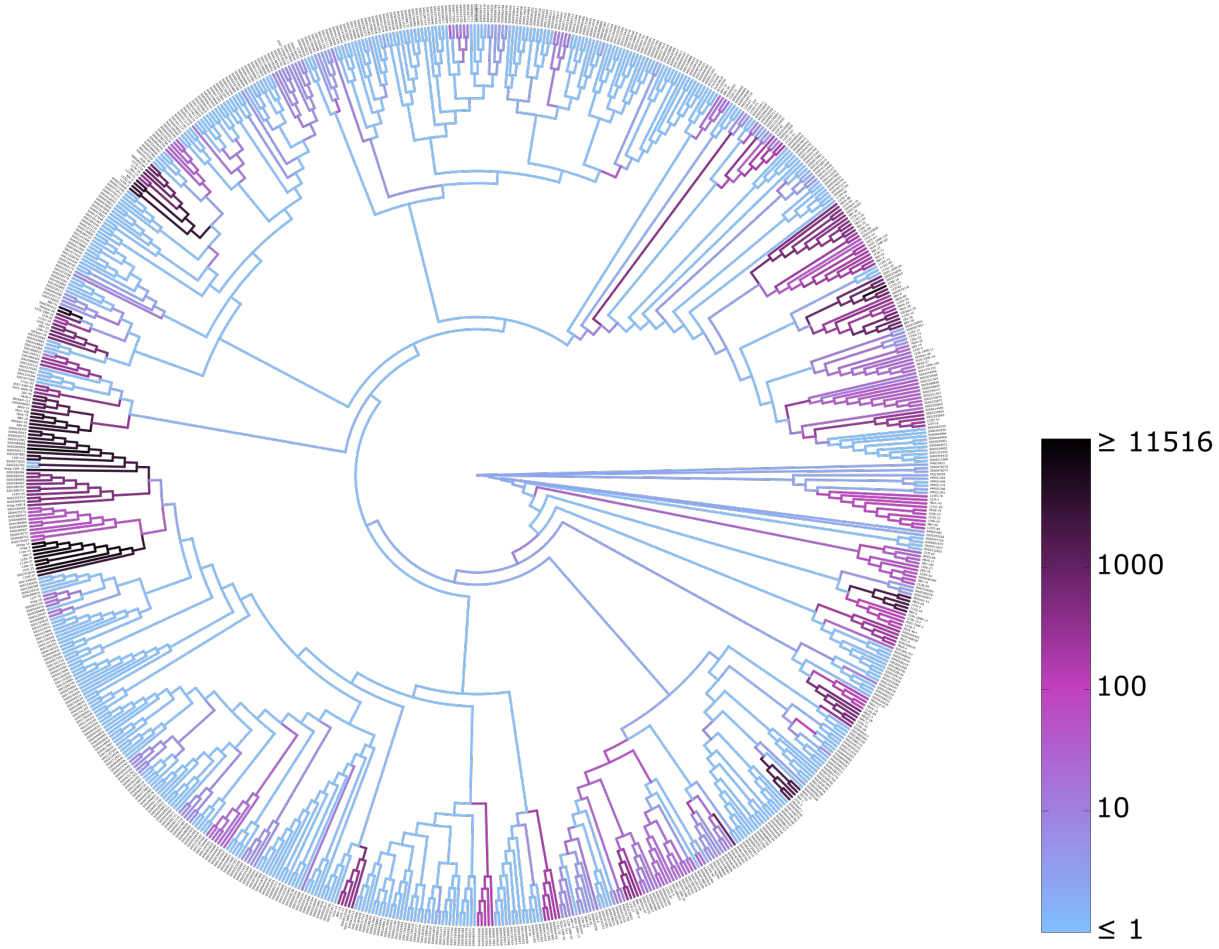
Bacterial Communities in Women with Bacterial Vaginosis: High Resolution Phylogenetic Analyses Reveal Relationships of Microbiota to Clinical Criteria

Sujatha Srinivasan[1]*, Noah G. Hoffman[2], Martin T. Morgan[3], Frederick A. Matsen[3], Tina L. Fiedler[1], Robert W. Hall[4], Frederick J. Ross[3], Connor O. McCoy[3], Roger Bumgarner[4], Jeanne M. Marrazzo[5], David N. Fredricks[1,4,5]*

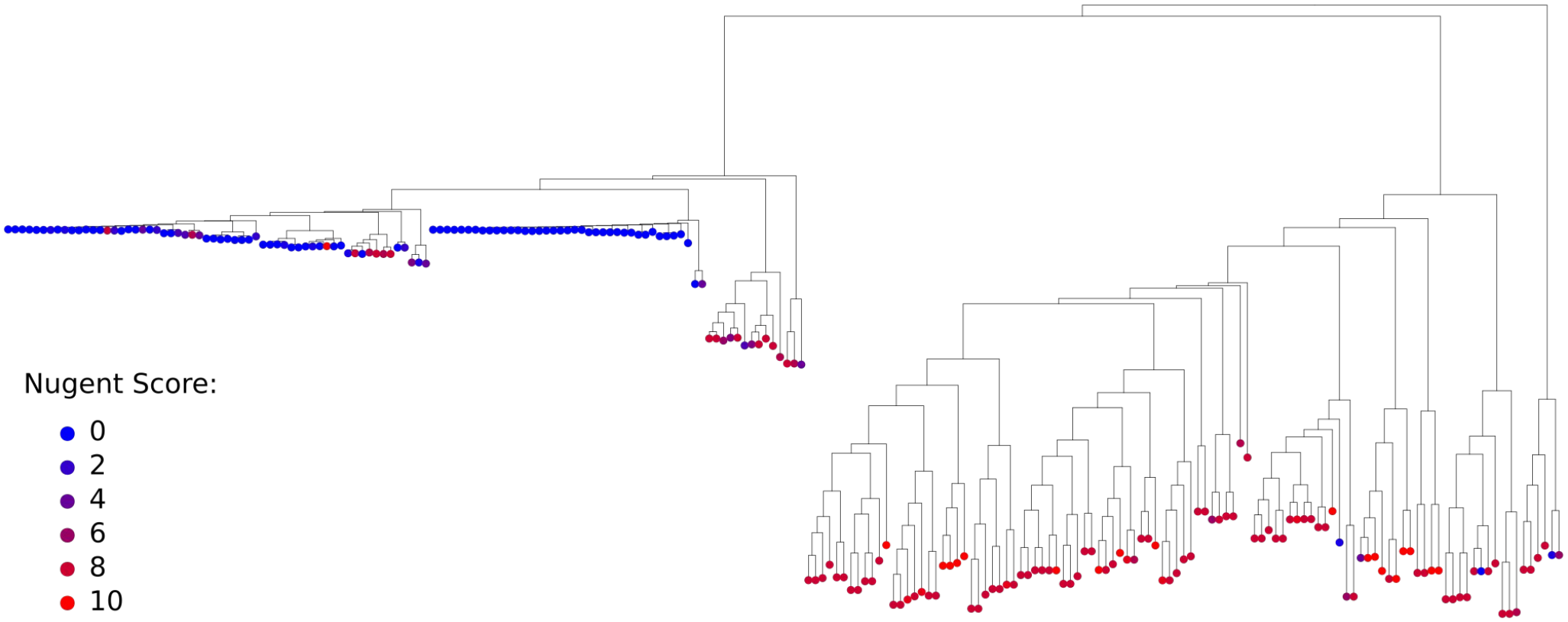
1 Vaccine & Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America, **2** Department of Laboratory Medicine, University of Washington, Seattle, Washington, United States of America, **3** Public Health Science Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America, **4** Department of Microbiology, University of Washington, Seattle, Washington, United States of America, **5** Department of Medicine, University of Washington, Seattle, Washington, United States of America

Srinivasan *et al*, PLoS ONE, 2012

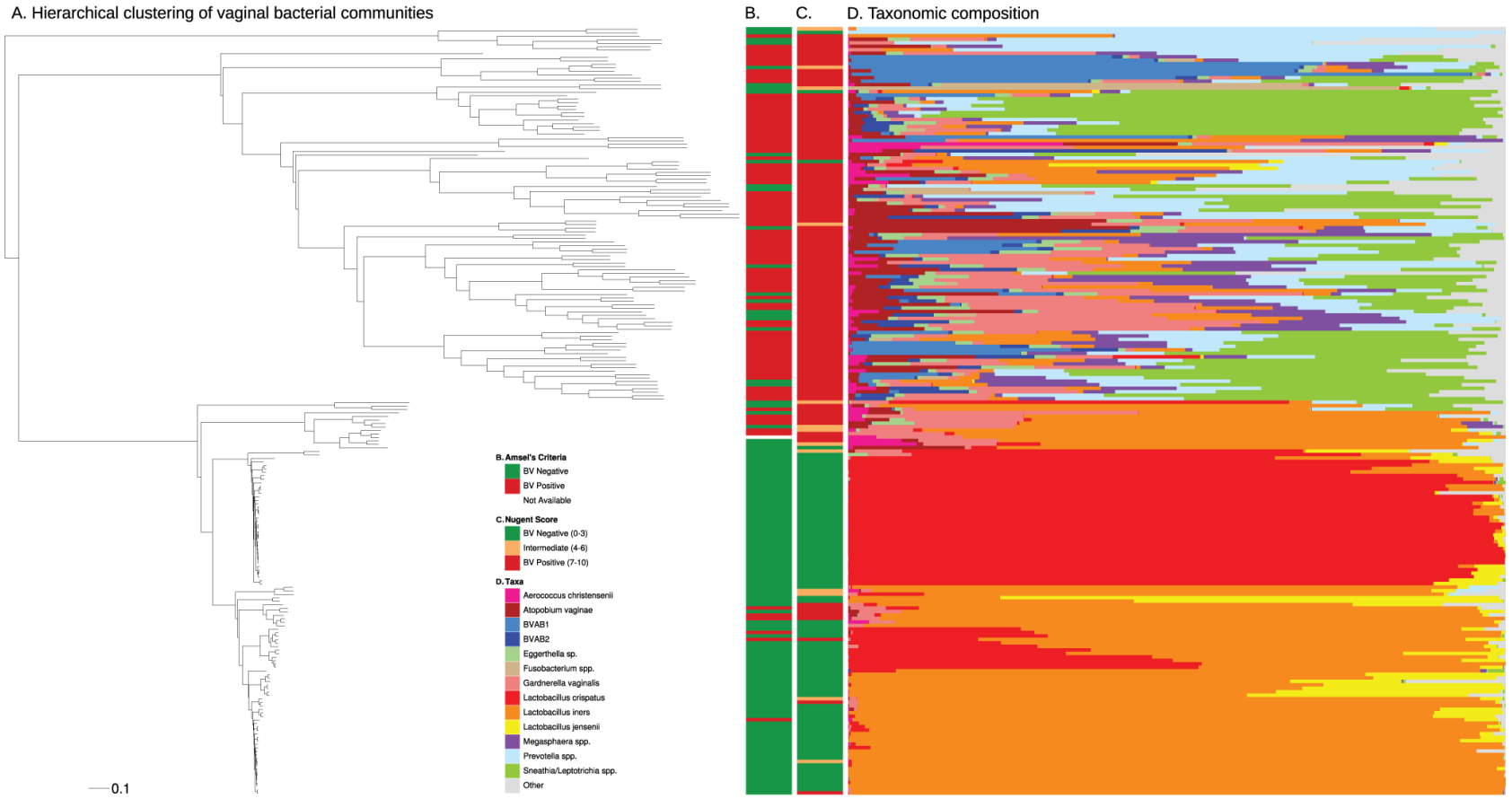
All 220 samples placed on a reference tree



Squash Clustering

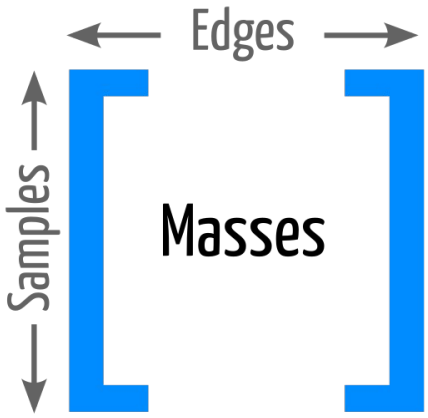


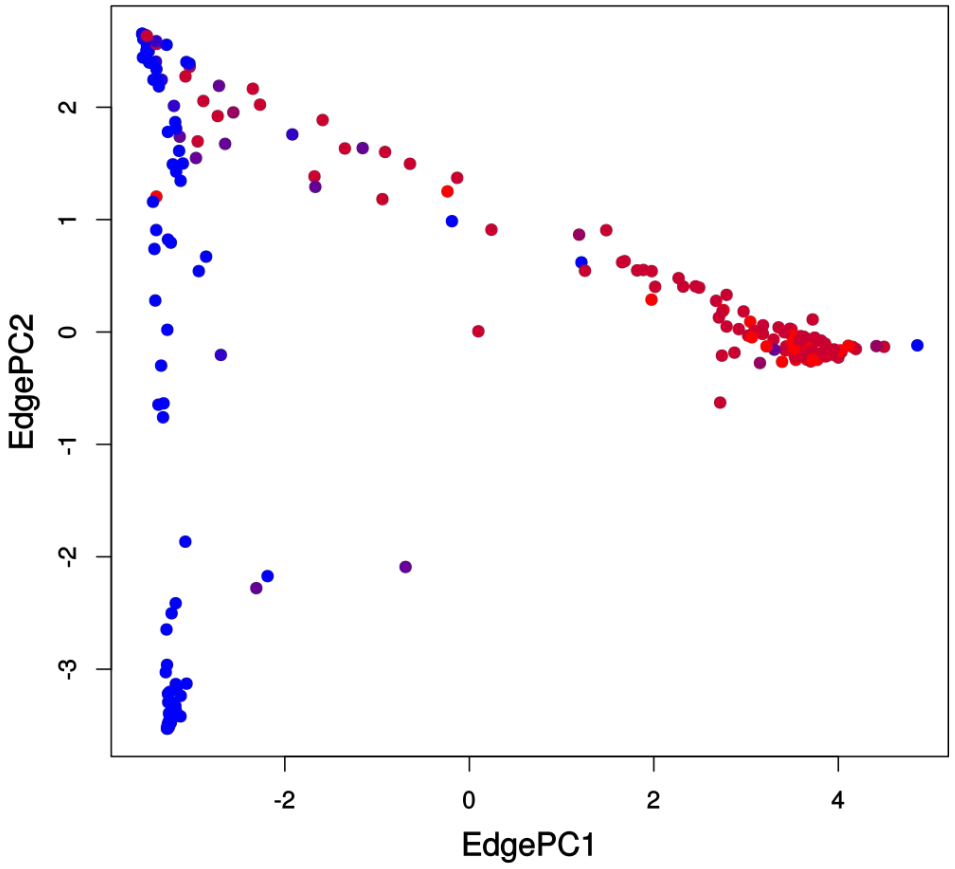
Samples cluster by disease status



Srinivasan *et al*, PLoS ONE, 2012

Edge PCA



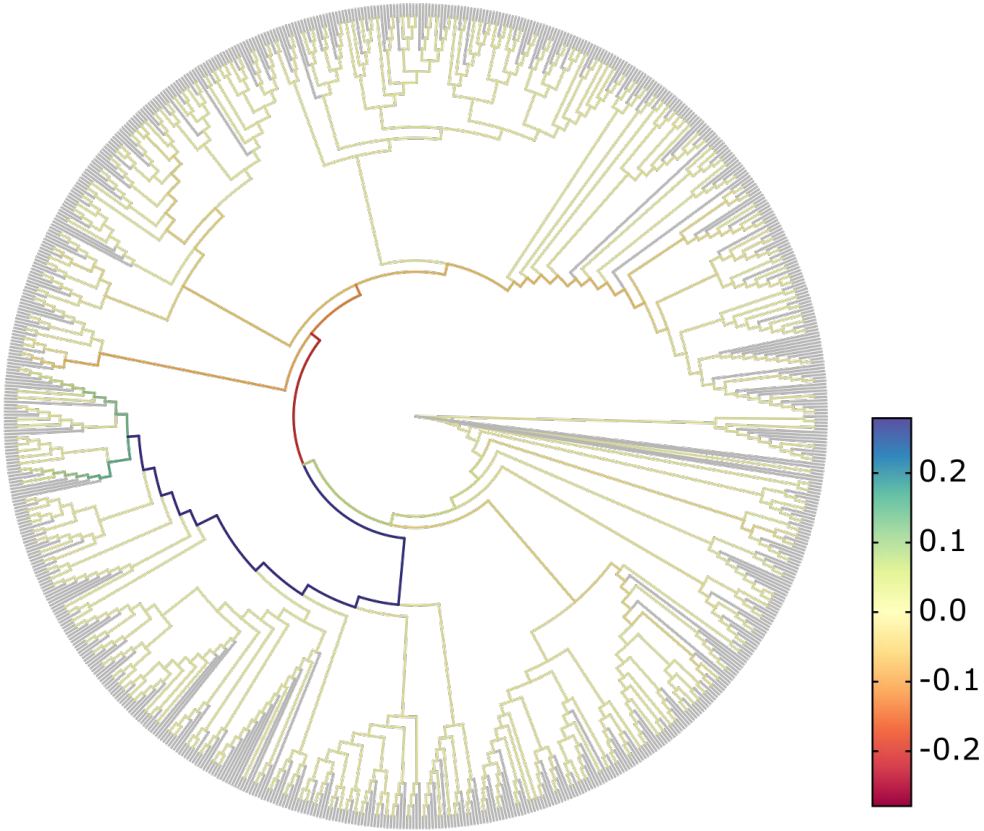


Nugent Score:

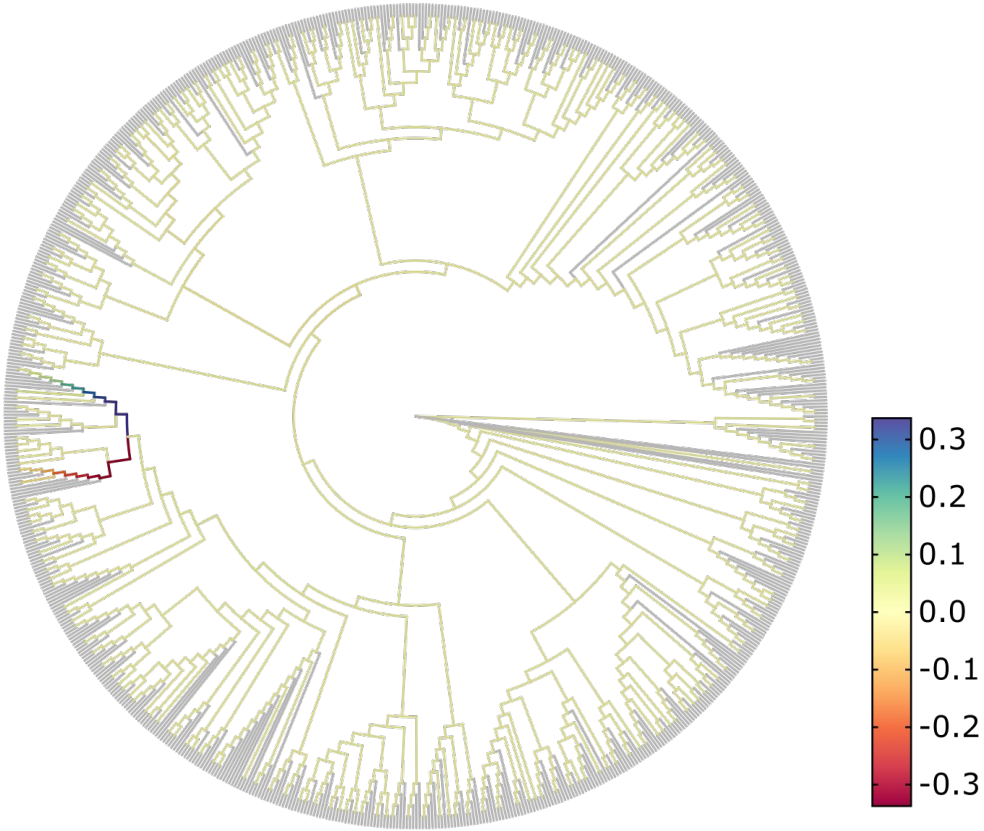
- 0
- 2
- 4
- 6
- 8
- 10

Edge PCA

(a) First Component



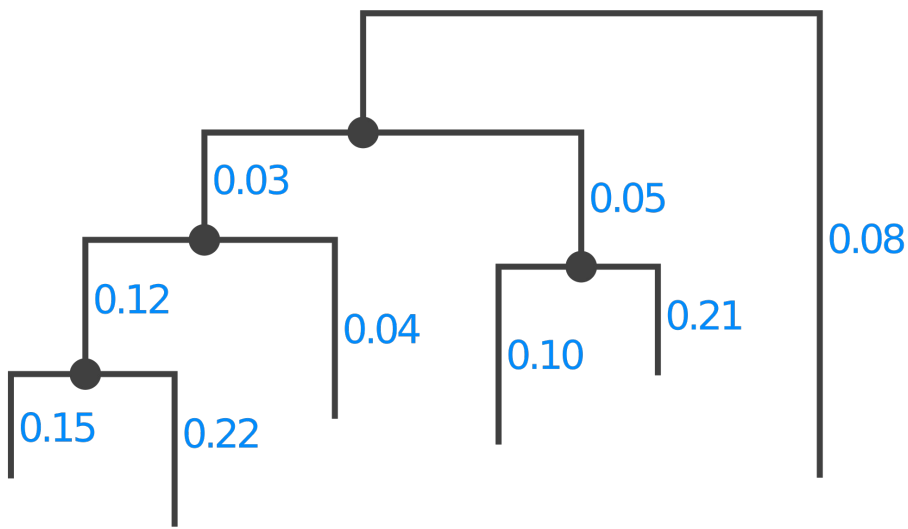
(b) Second Component



Caveat

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

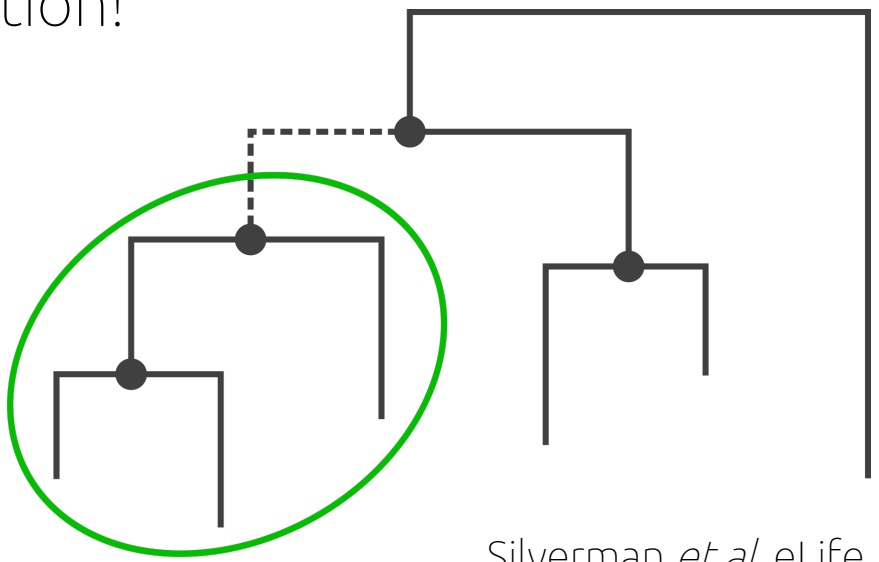
  → transform the Edge Masses



Adaptation of Balances to Placements

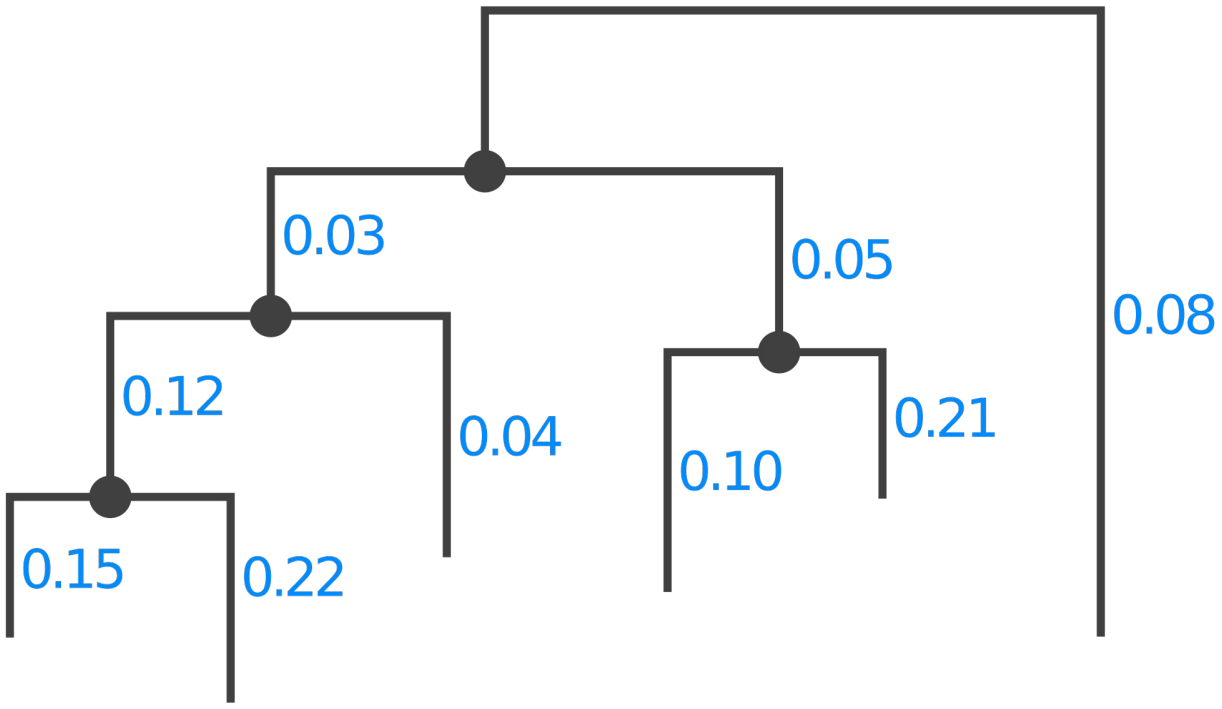
- Often, groups of organisms are important biologically
- We transform from per-branch data to per-clade data
- Hence, look at subtrees instead of single edges
- This yields our desired transformation!

→ Adaptation of Balances to Placements



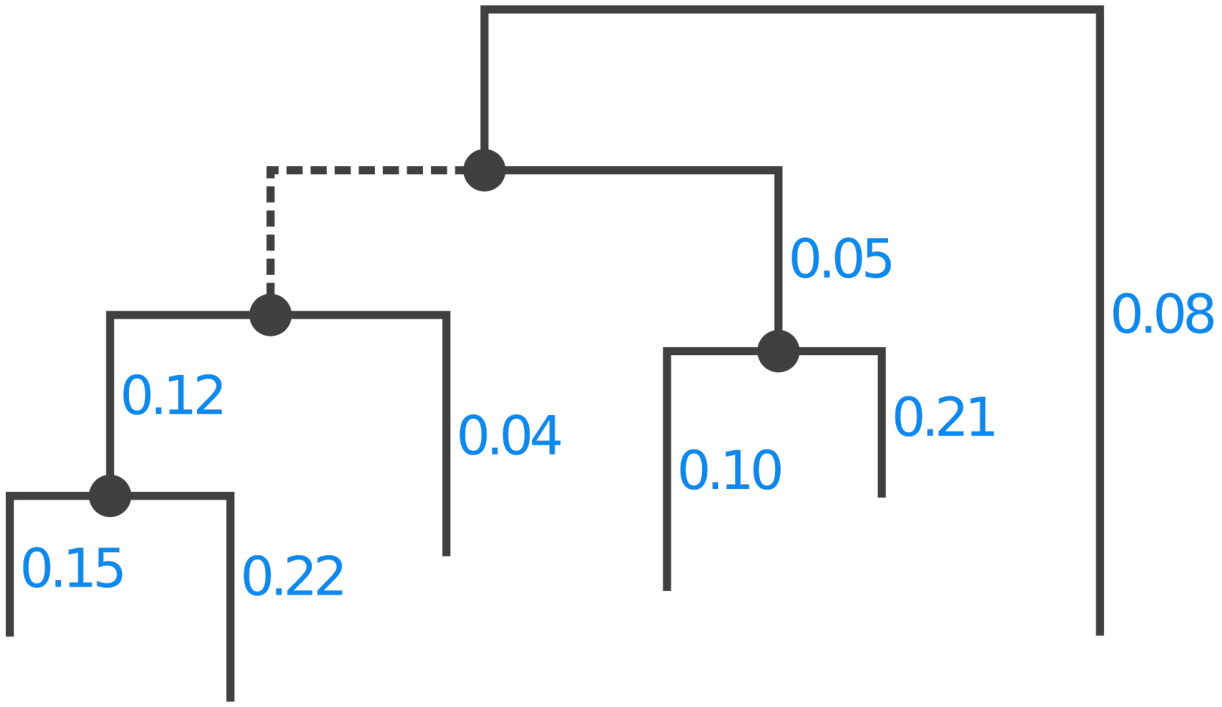
Silverman *et al*, eLife, 2017

Edge Masses for one Sample



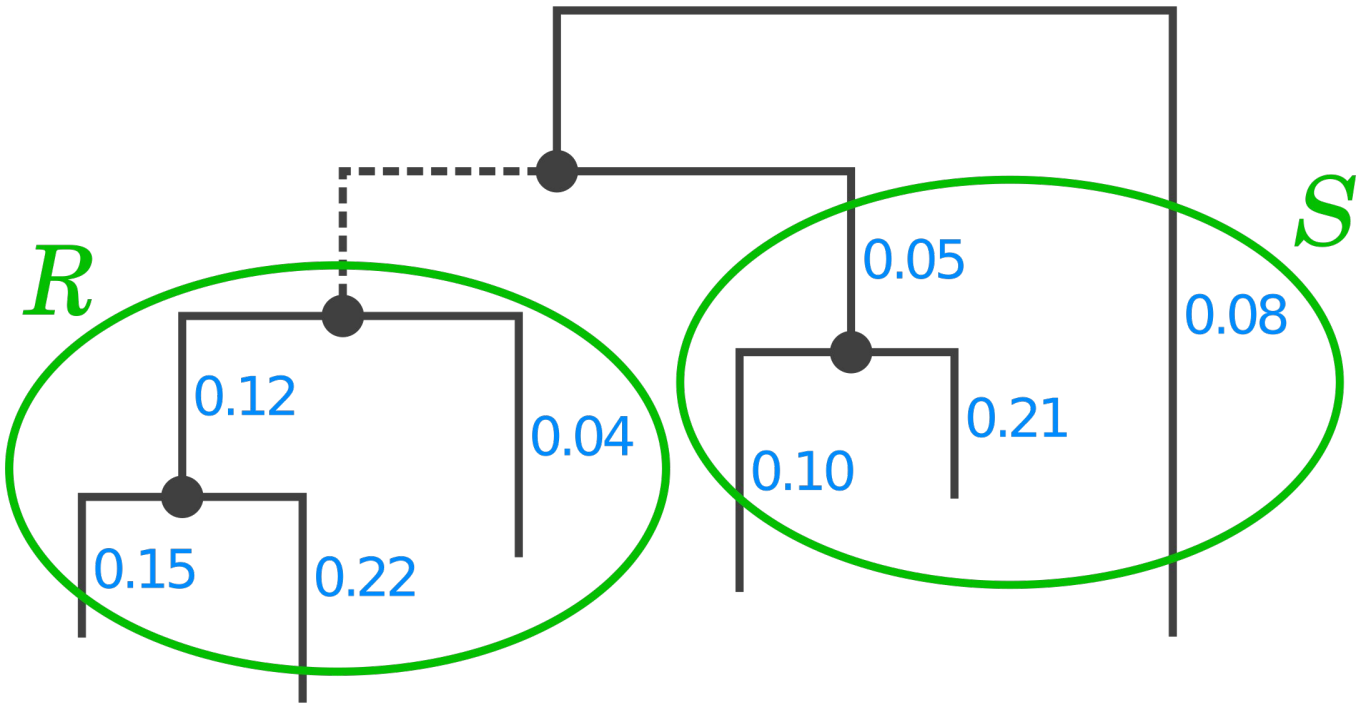
Silverman *et al*, eLife, 2017

Subtrees induced by an Edge



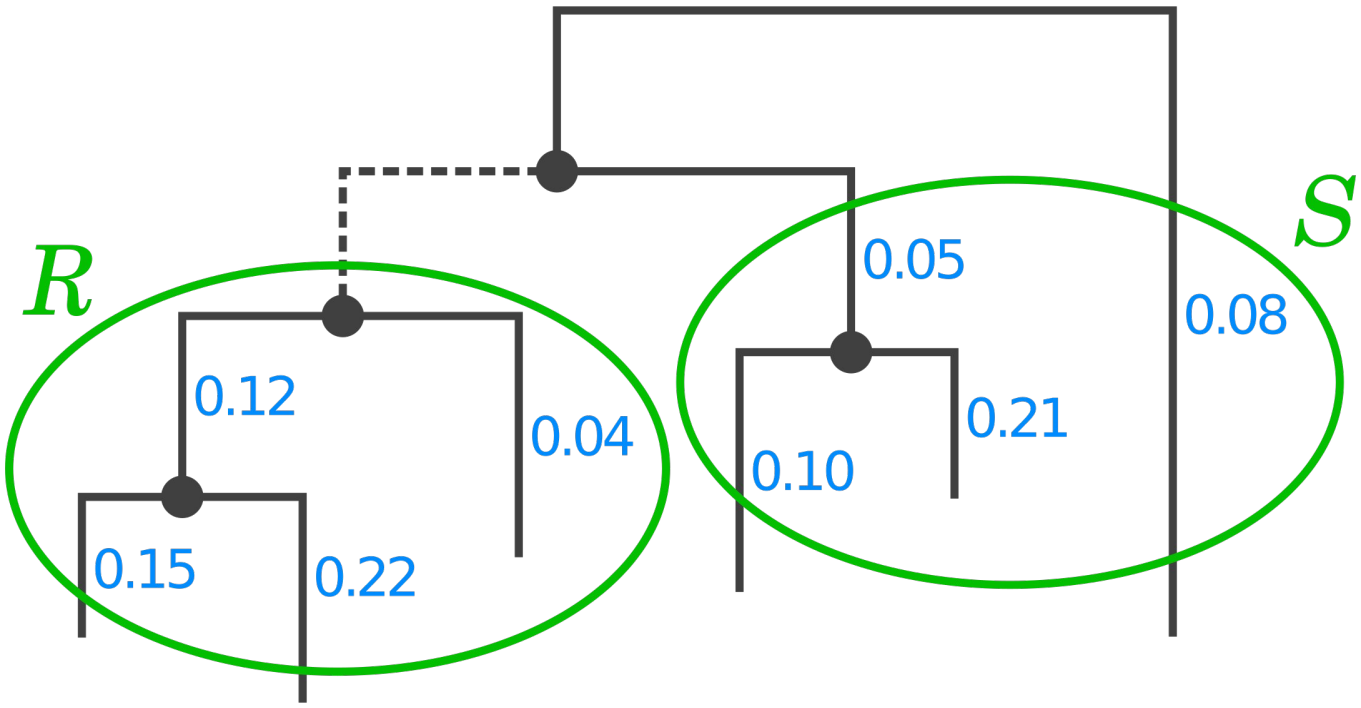
Silverman *et al*, eLife, 2017

Summarize the Subtrees



Silverman *et al*, eLife, 2017

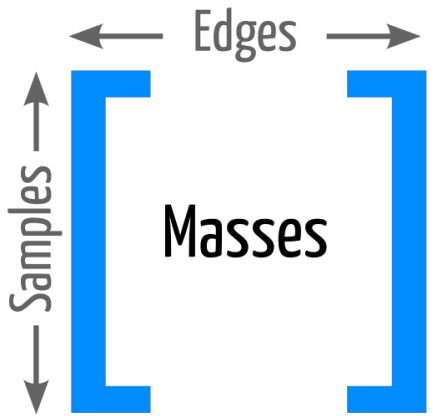
Balance across the Edge



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

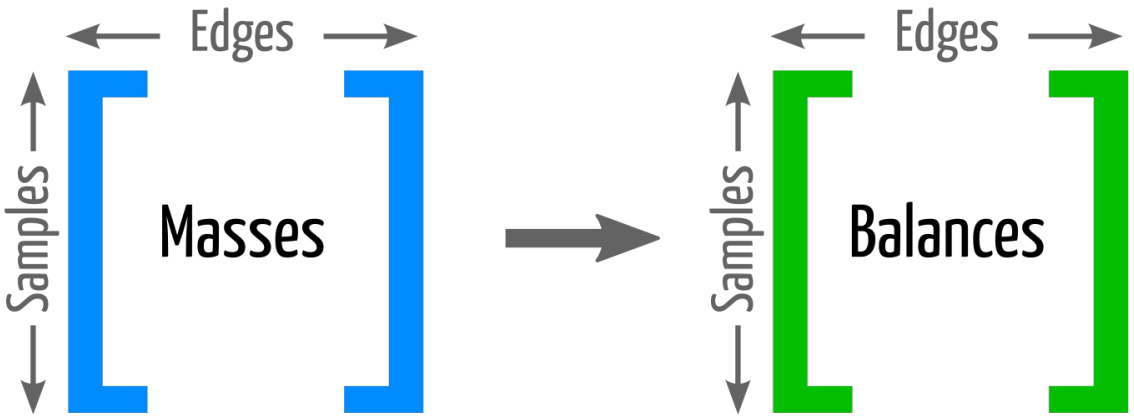
Silverman *et al*, eLife, 2017

Masses for all Samples and Edges



Focus on Edges

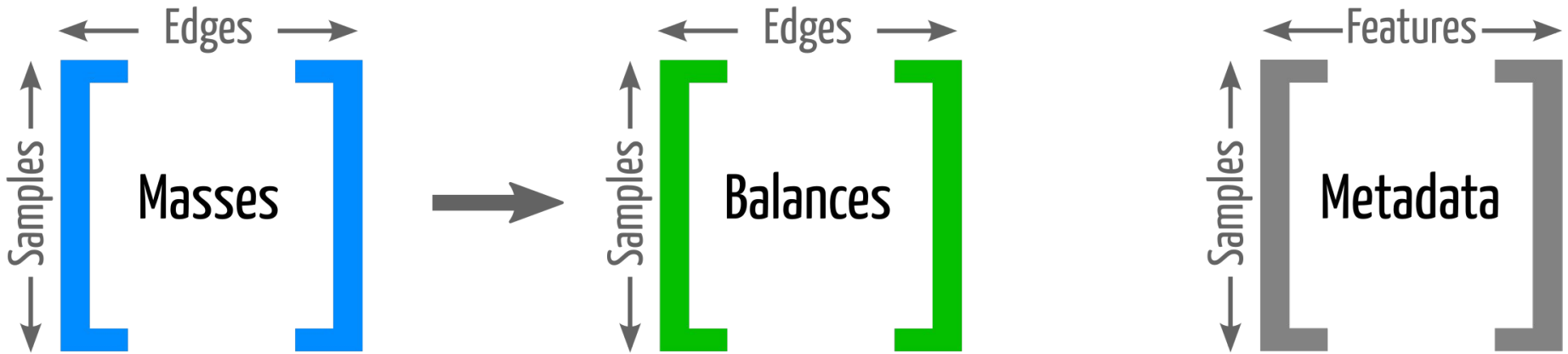
Transformation into Balances



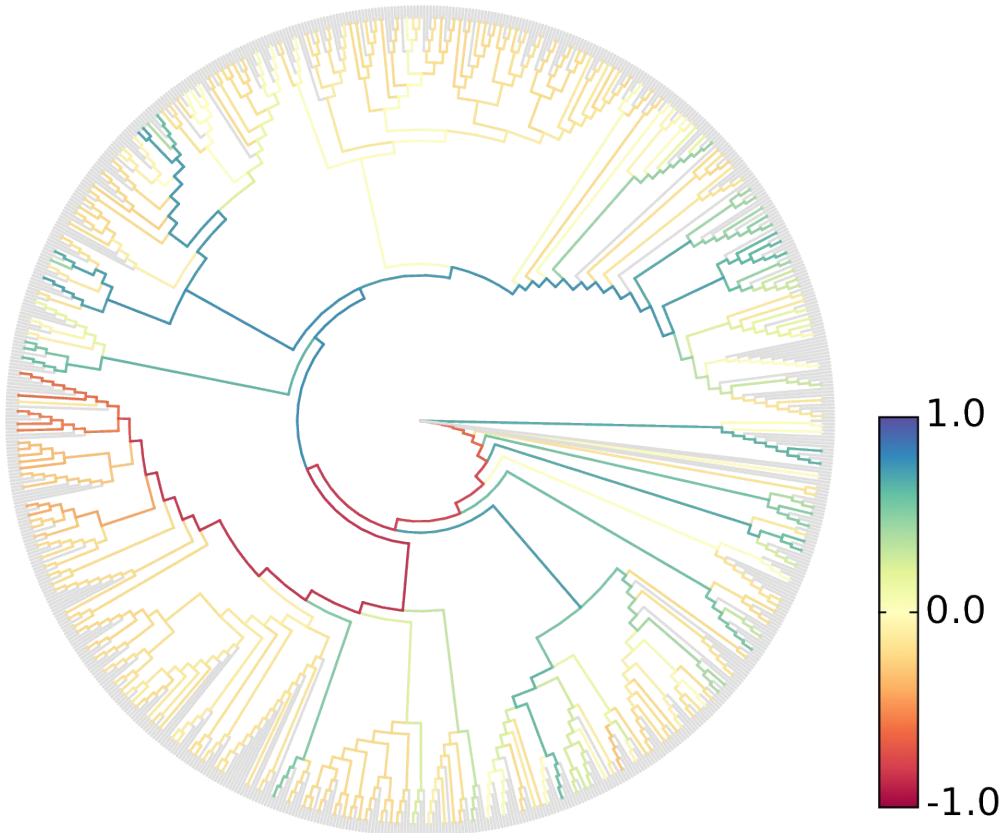
Focus on Edges

Focus on Subtrees

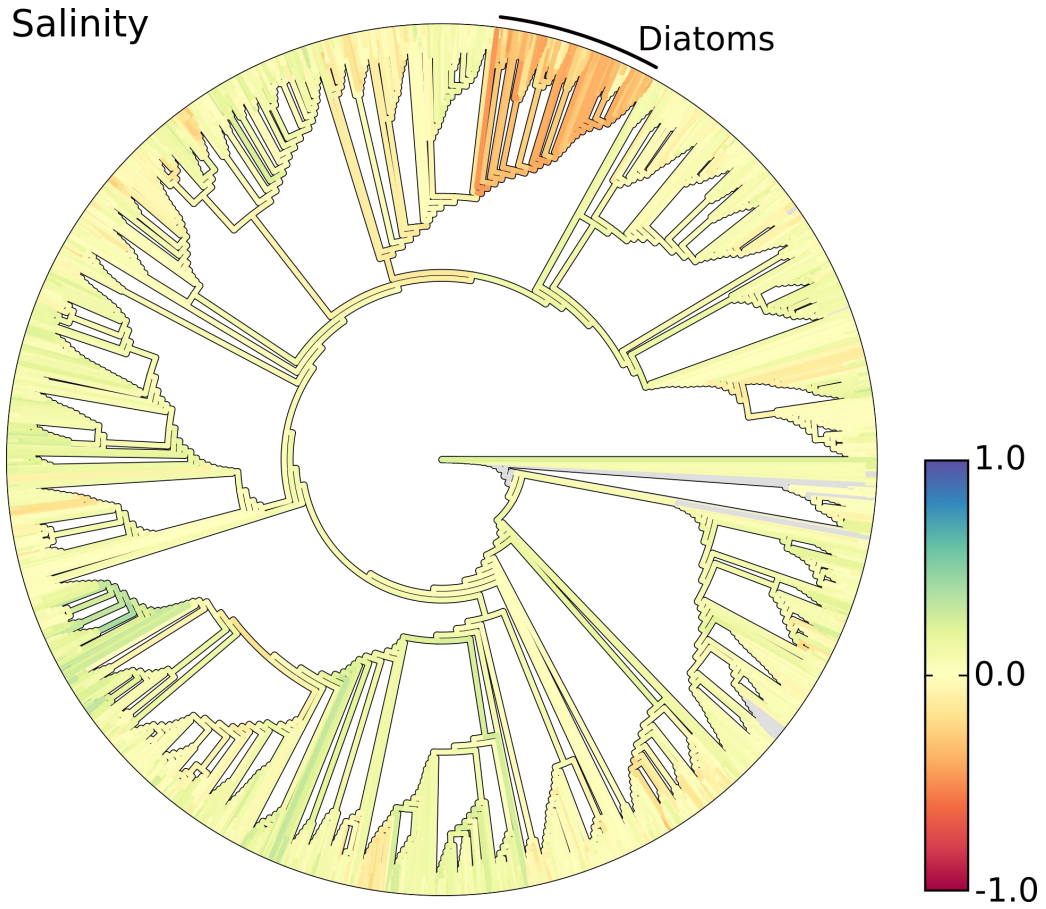
Take Metadata into Account



Edge Correlation between Balances and Metadata



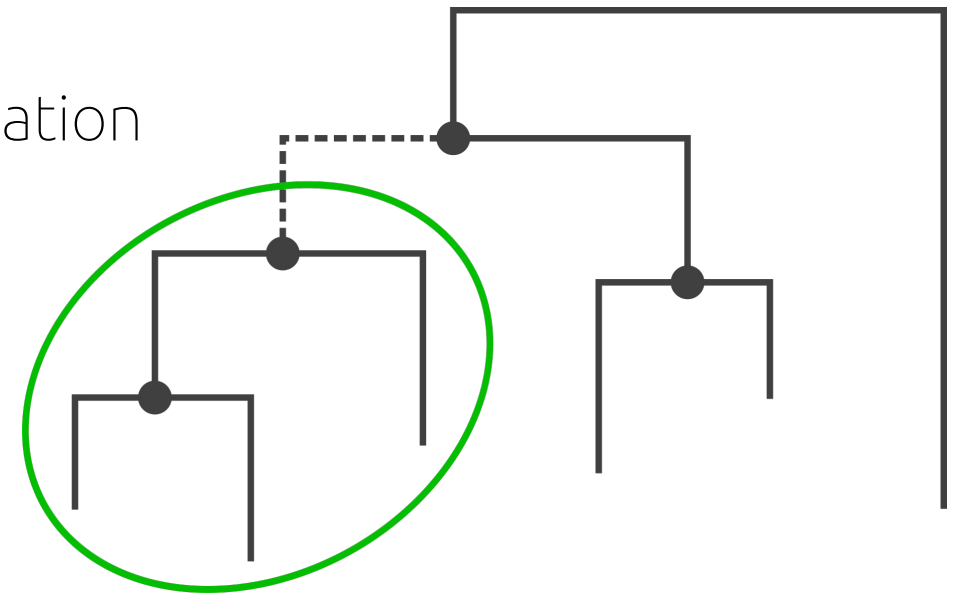
Edge Correlation between Balances and Metadata



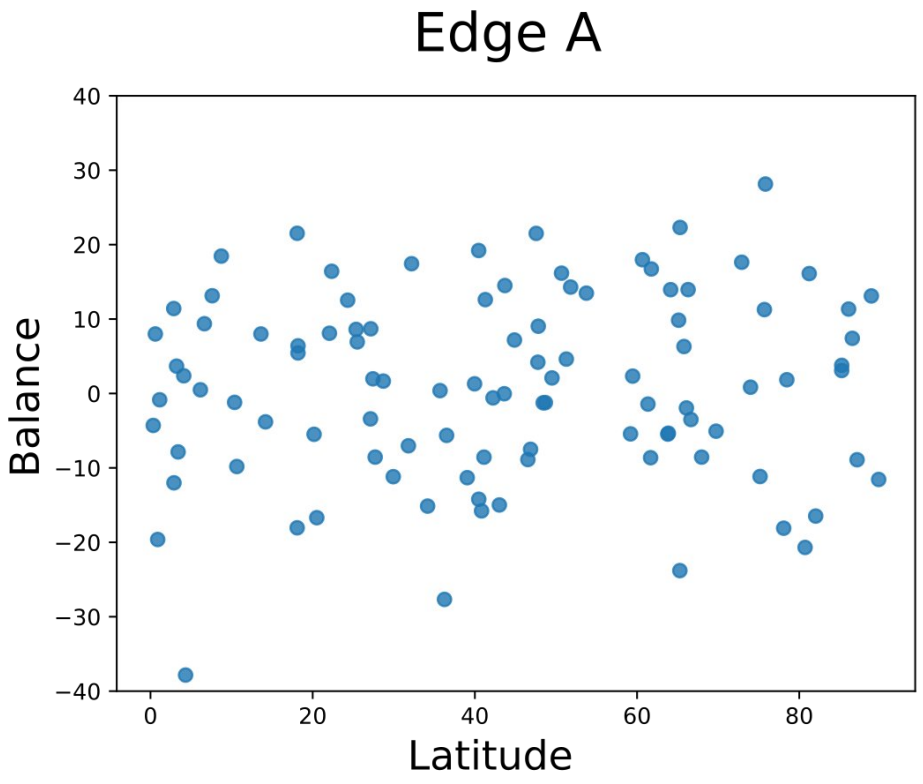
Placement-Factorization

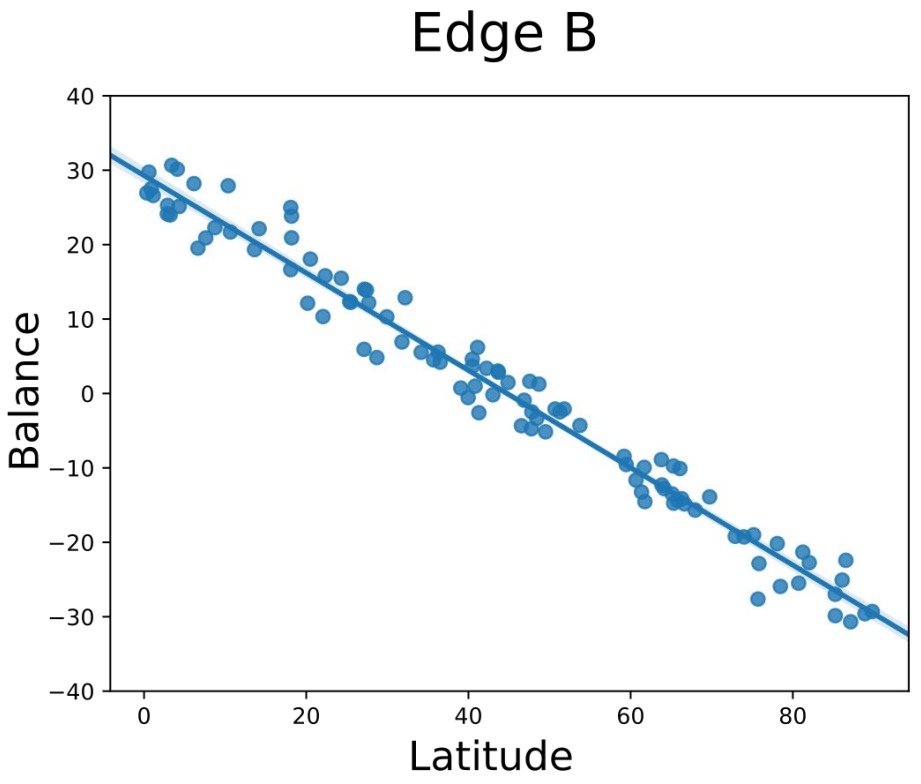
- Balances indicate differences in abundances across clades
- Use Balances to find edges / subtrees where abundances change with some metadata feature

→ Adaptation of PhyloFactorization to Placements

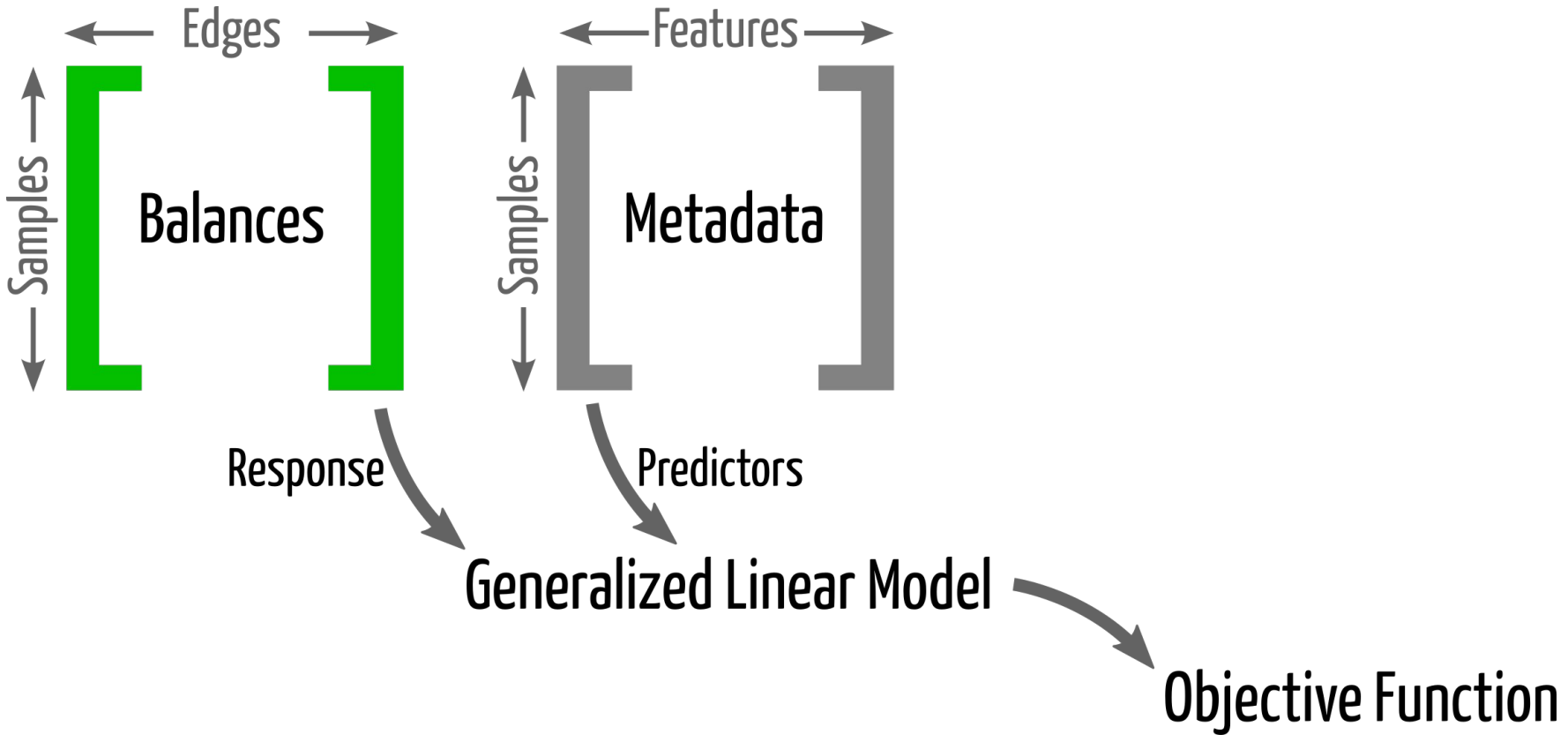


Balances vs. Metadata

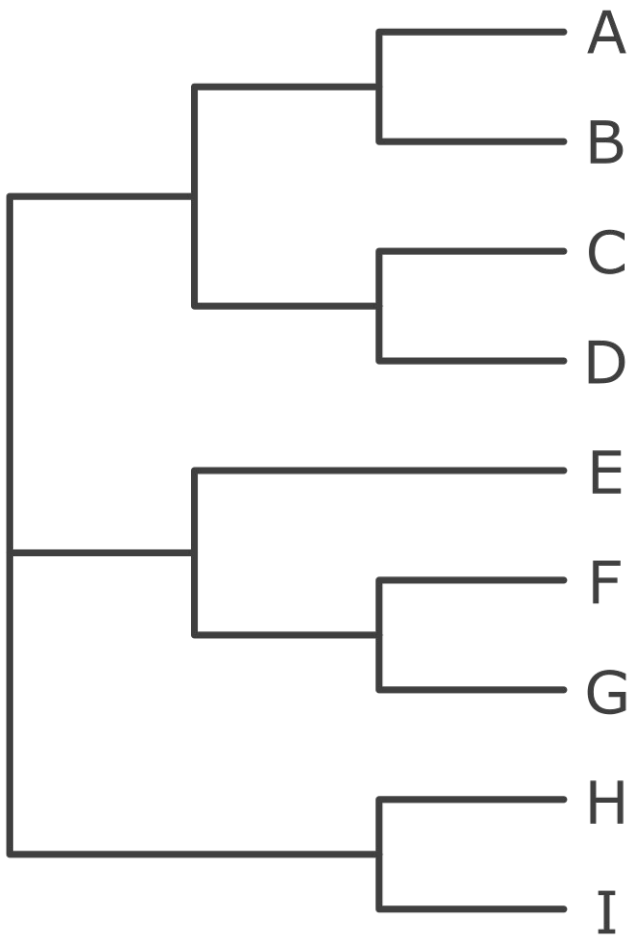




Balances vs. Metadata

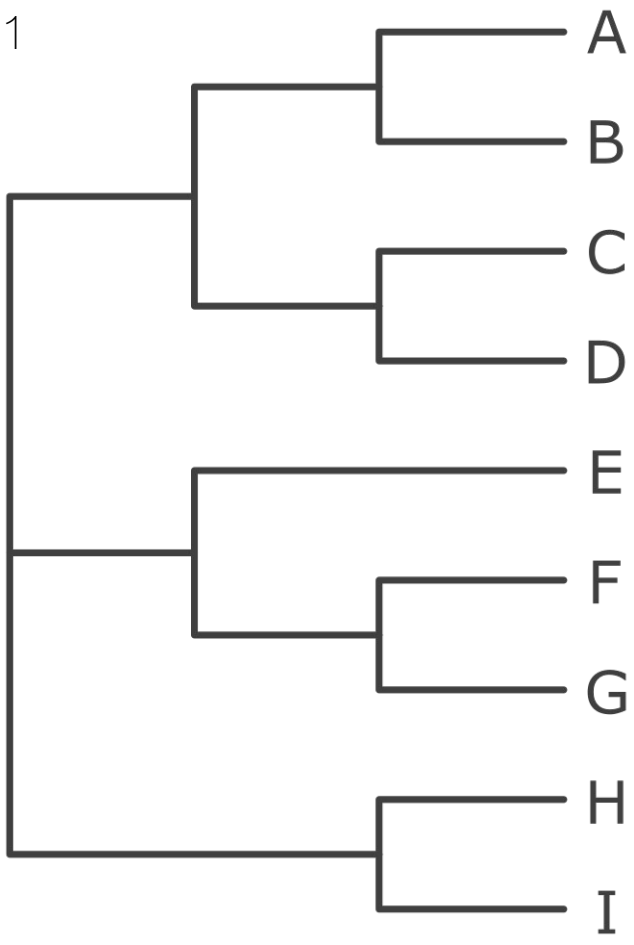


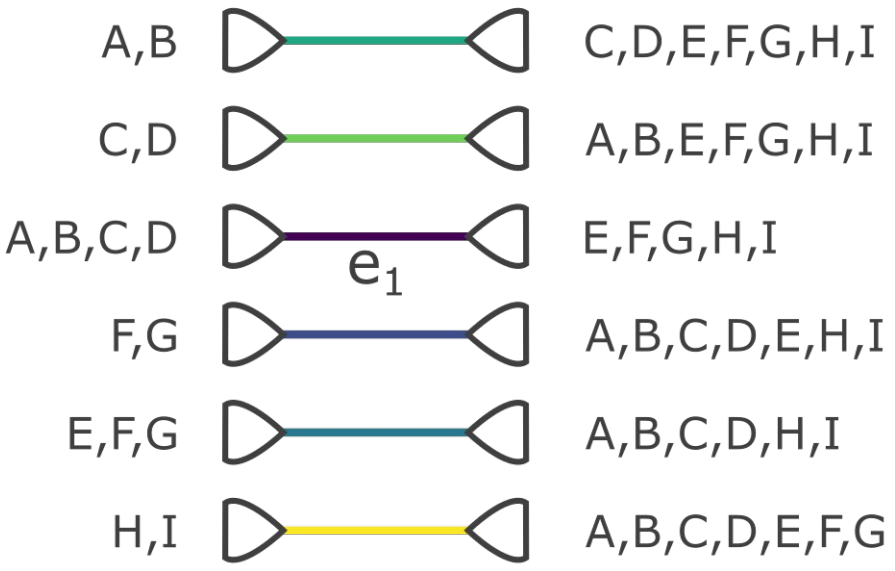
Placement-Factorization

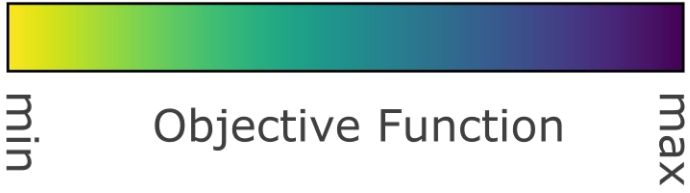


Placement-Factorization

Factor 1

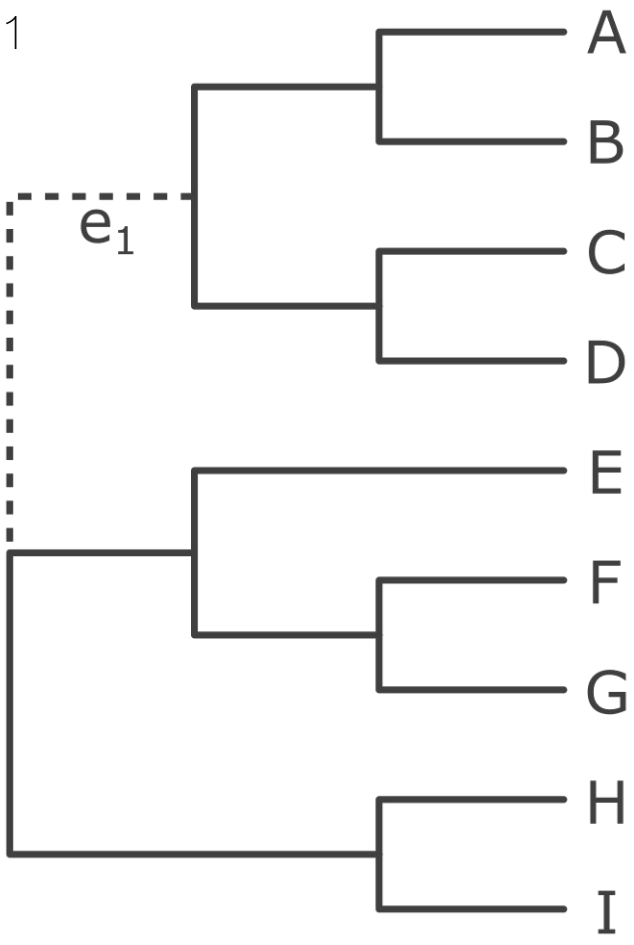


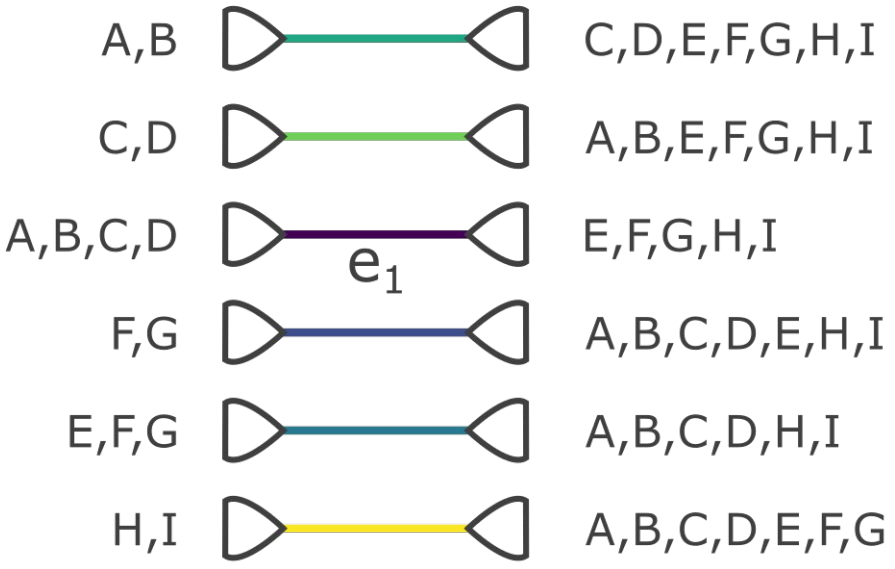


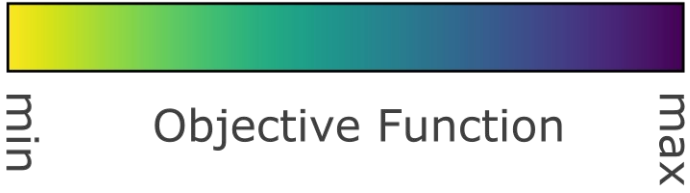


Placement-Factorization

Factor 1

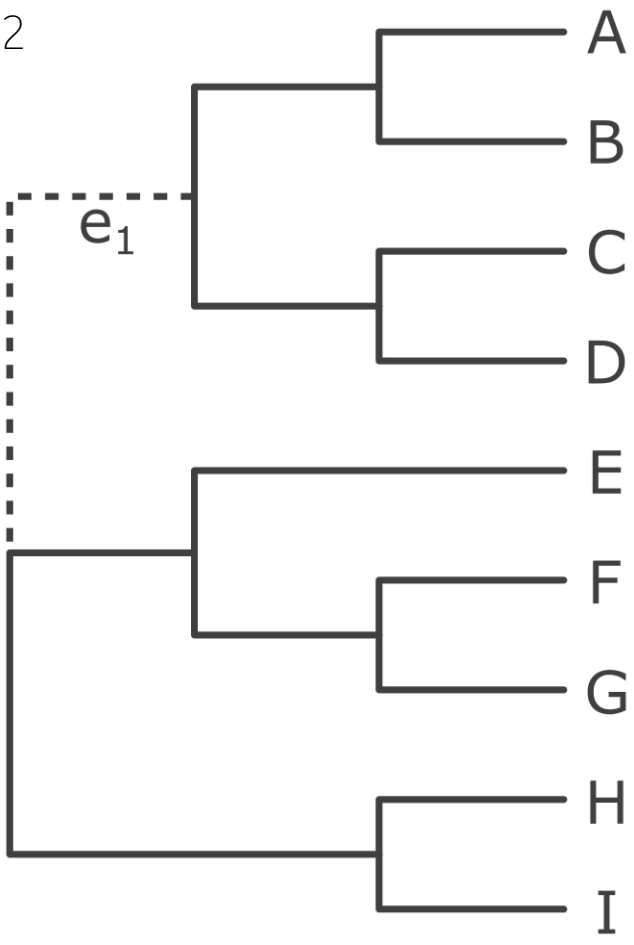


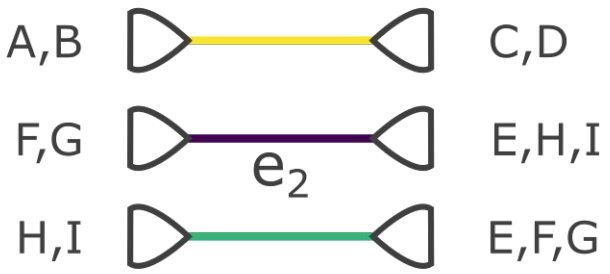


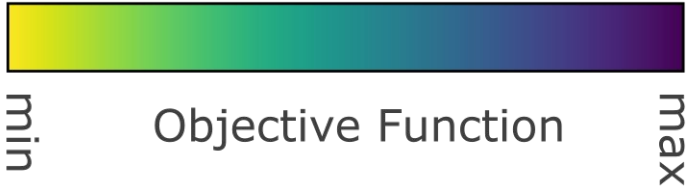


Placement-Factorization

Factor 2

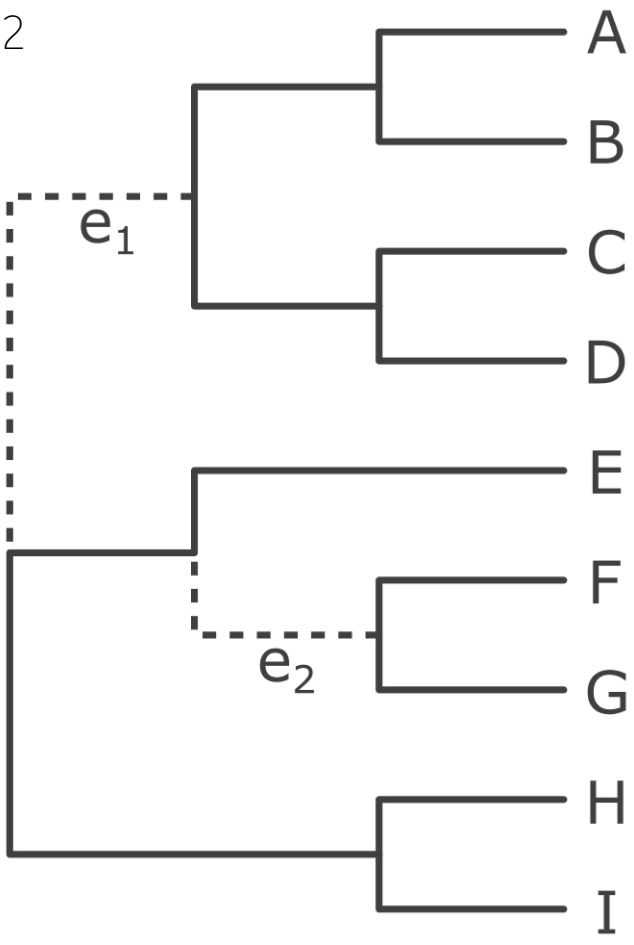


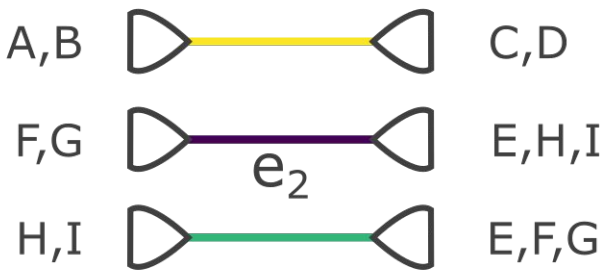


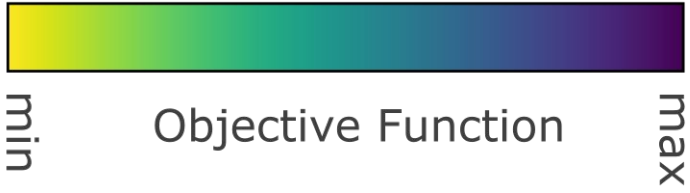


Placement-Factorization

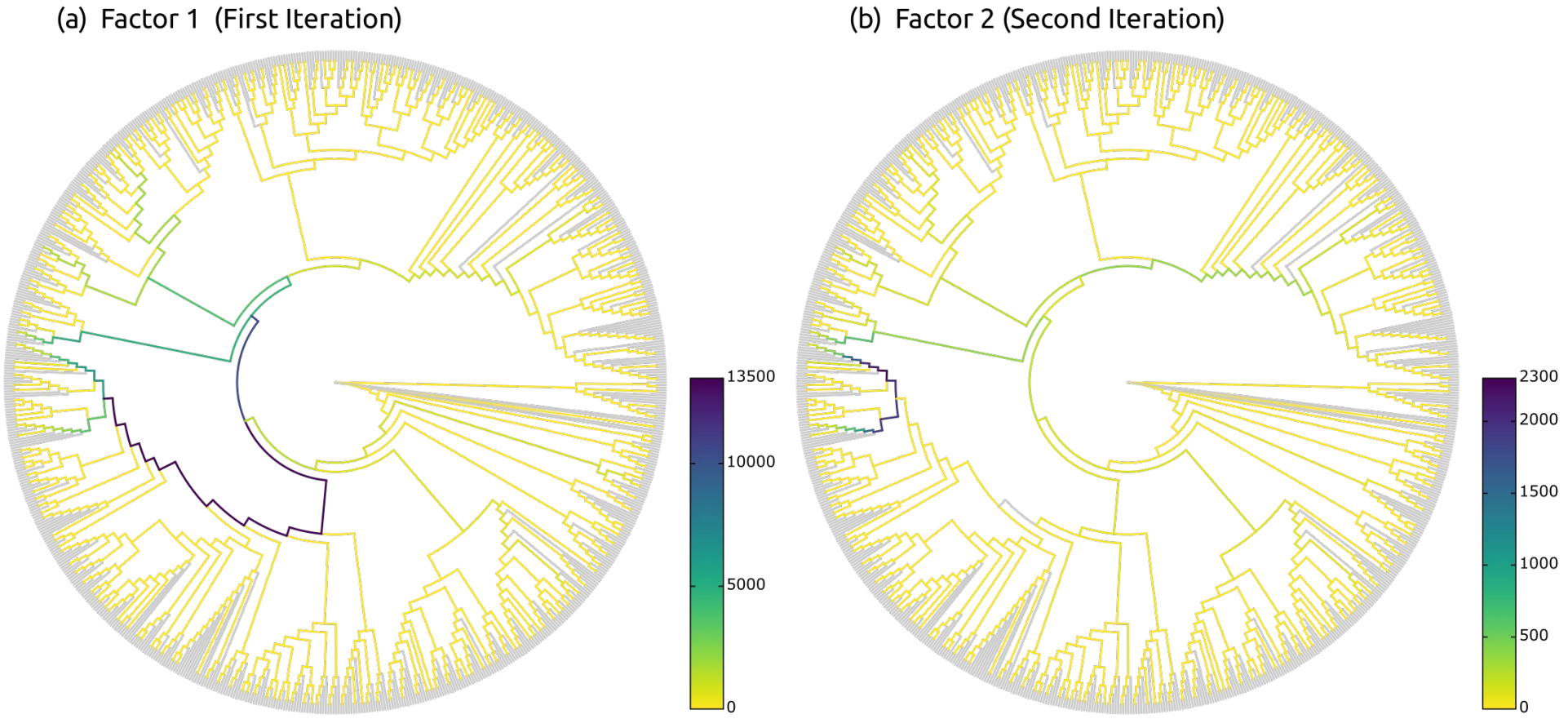
Factor 2



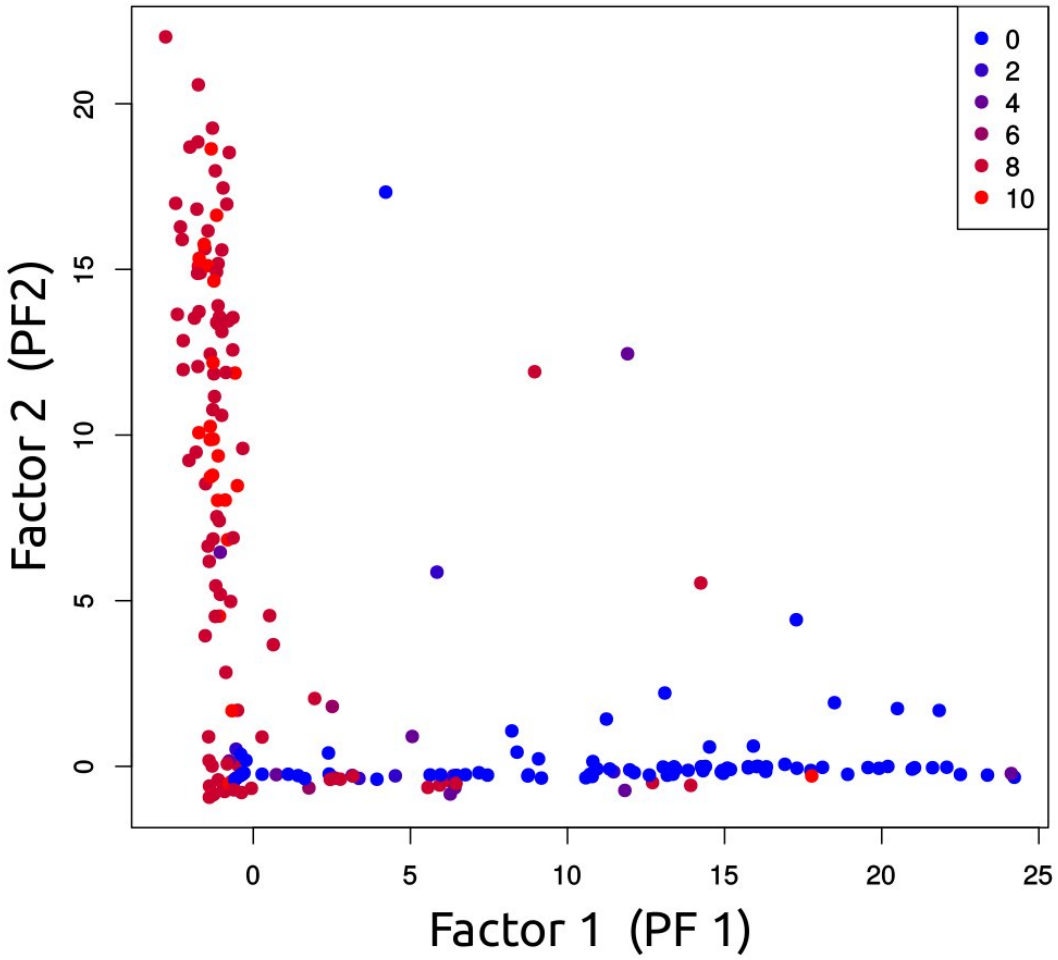




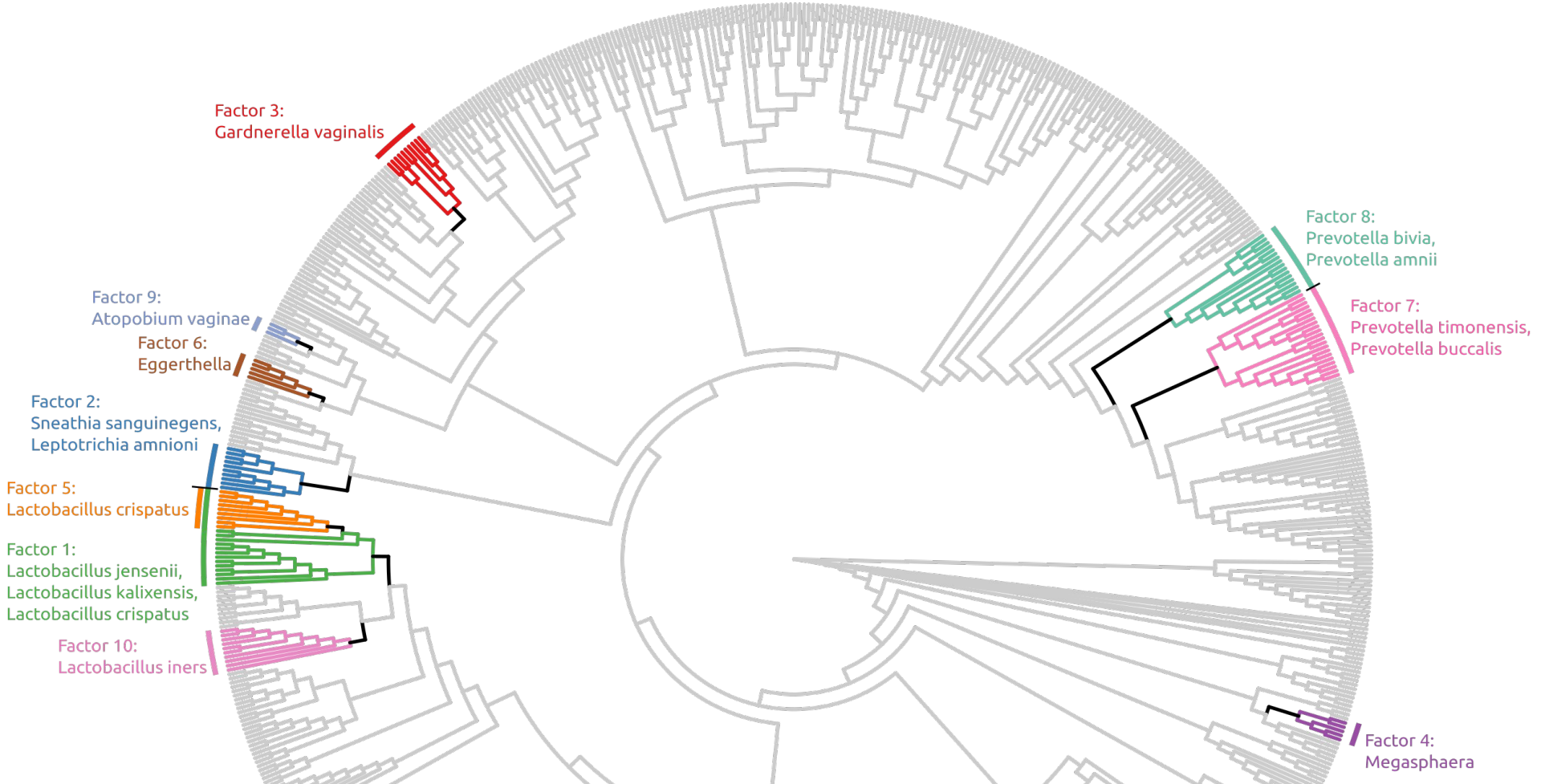
Placement-Factorization



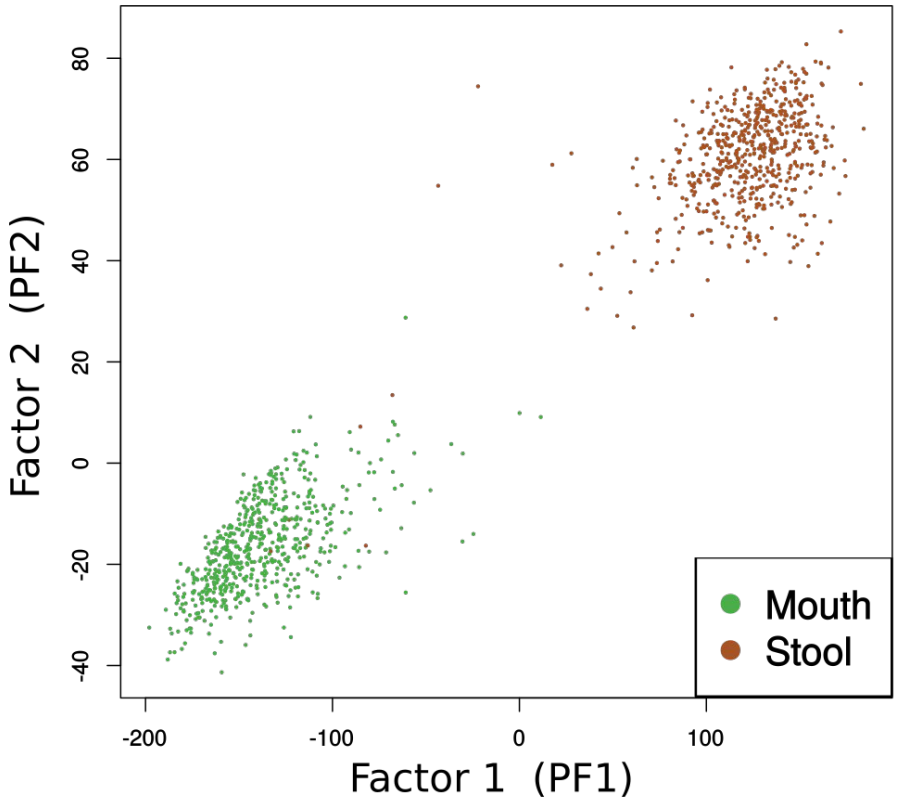
Factors are an ordination of samples

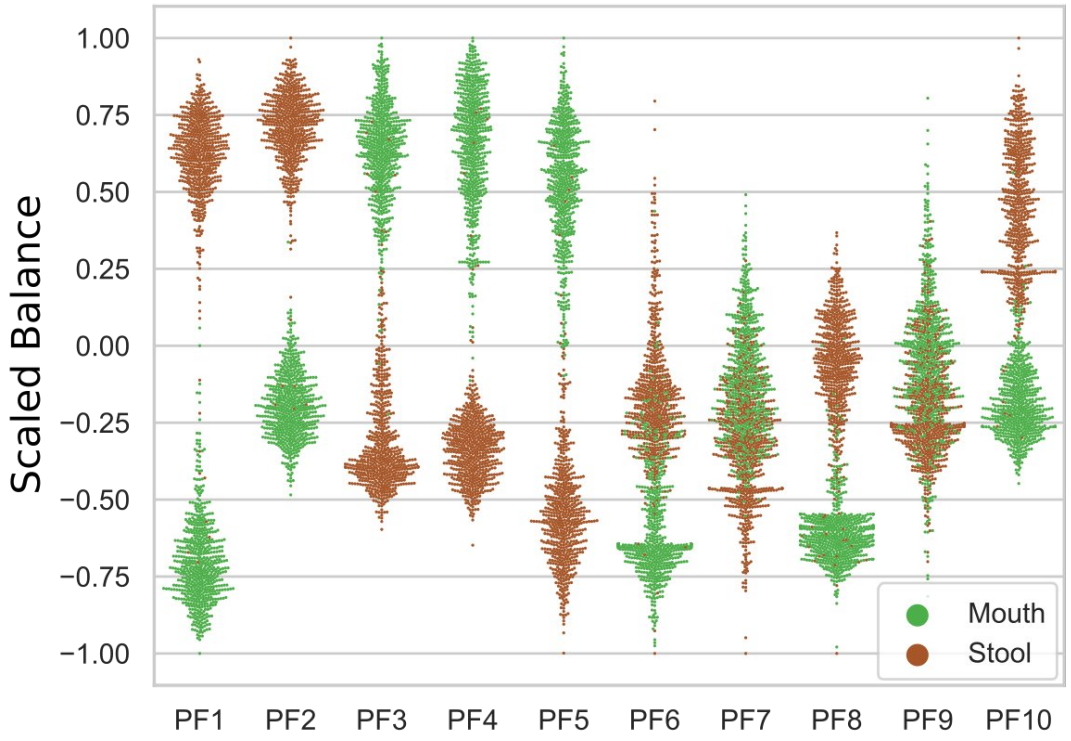


Placement-Factorization

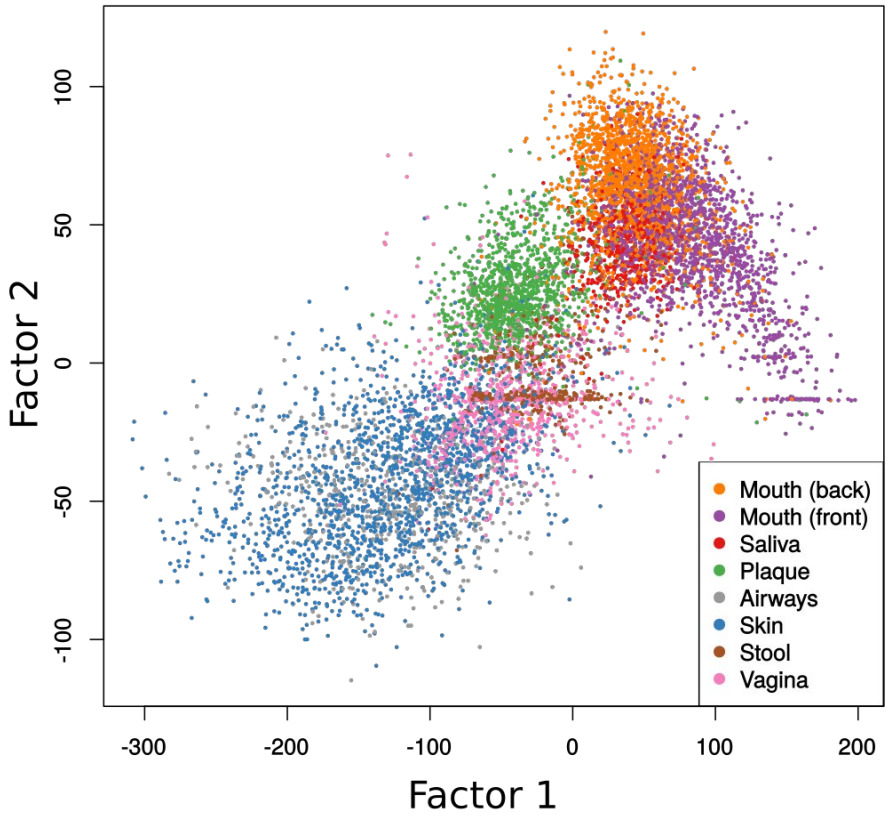


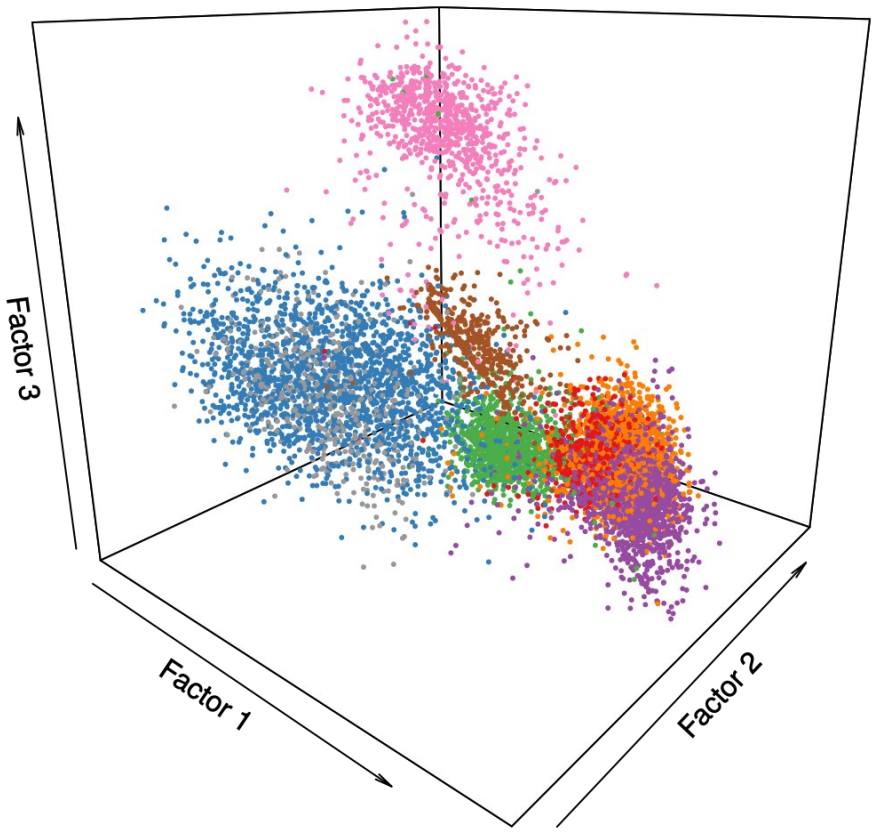
Placement-Factorization





Placement-Factorization





Into the Placement-verse

Syst. Biol. 68(2):365–369, 2019

DOI:10.1093/sysbio/syy054
Advance Access publication September 21, 2018

EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences

PIERRE BARBERA[1,*], ALEXEY M. KOZLOV[1], LUCAS CZECH[1], BENOIT MOREL[1], DIEGO DARRIBA[1,2], TOMÁŠ FLOURI[1,3], AND ALEXANDROS STAMATAKIS[1,4]

[1]*Heidelberg Institute for Theoretical Studies, Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany;*
[2]*Department of Computer Engineering, University of A Coruña, 15071 A Coruña, Spain;*
[3]*Department of Genetics, Evolution and Environment,University College London, Gower St., Bloomsbury, London WC1E 6BT, UK; and*
[4]*Karlsruhe Institute of Technology, Department of Informatics, Institute of Theoretical Informatics, Postfach 6980, 76128 Karlsruhe, Germany*
**Correspondence to be sent to: Heidelberg Institute for Theoretical Studies, Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany;*
E-mail: pierre.barbera@h-its.org.

Received 30 March 2018; reviews returned 21 August 2018; accepted 21 August 2018
Associate Editor: David Posada

doi: 10.1093/sysbio/syy054

gappa

Bioinformatics, 36(10), 2020, 3263–3265
doi: 10.1093/bioinformatics/btaa070
Advance Access Publication Date: 4 February 2020
Applications Note

OXFORD

Phylogenetics

Genesis and Gappa: processing, analyzing and visualizing phylogenetic (placement) data

Lucas Czech [1,*], **Pierre Barbera** [1] and **Alexandros Stamatakis** [1,2,*]

[1]Computational Molecular Evolution Group, Heidelberg Institute for Theoretical Studies, Heidelberg 69118, Germany and [2]Institute for Theoretical Informatics, Karlsruhe Institute of Technology, Karlsruhe 76131, Germany

doi: 10.1093/bioinformatics/btaa070

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