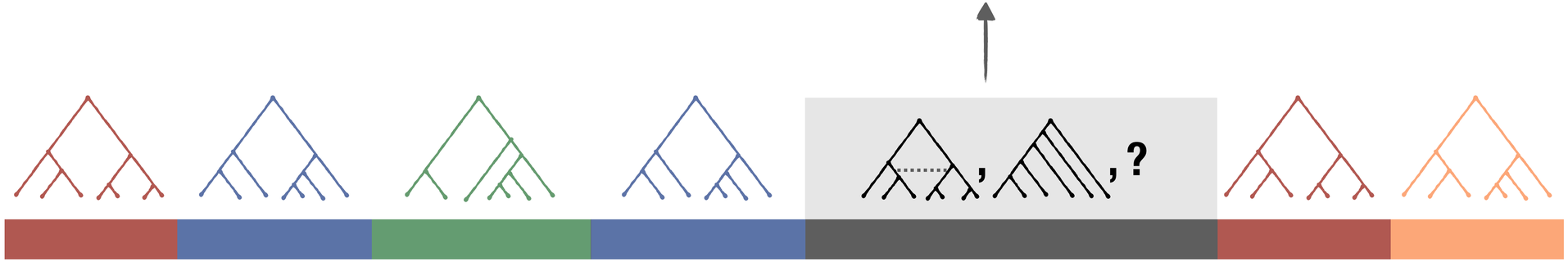
Mapping Strongly Discordant Regions on the Genome Using Hidden Markov Models

Ali Osman Berk Şapcı, Shayesteh Arasti, and Siavash Mirarab
UC San Diego

Coalescence with recombination versus non-ILS discordance

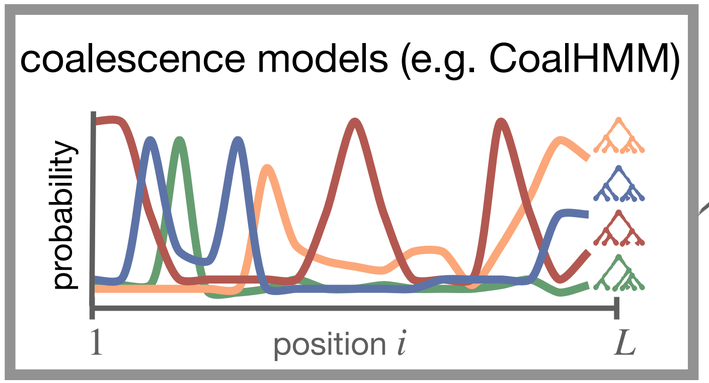
Expected discordance:
incomplete lineage sorting (ILS) & recombination

Strong non-ILS discordance:
biological (e.g., hybridization) & artifacts (e.g., errors)



Chromosome

Can we detect outlier regions using the posterior probabilities across the loci?



Expensive to compute!

	1	2	…	i	…	L-1	L
	0.03	0.03	…	0.09	…	0.26	0.22
	0.04	0.05	…	0.01	…	0.11	0.12
	…	…	…	…	…	…	…
	0.23	0.24	…	0.11	…	0.09	0.10
	0.06	0.06	…	0.32	…	0.03	0.04

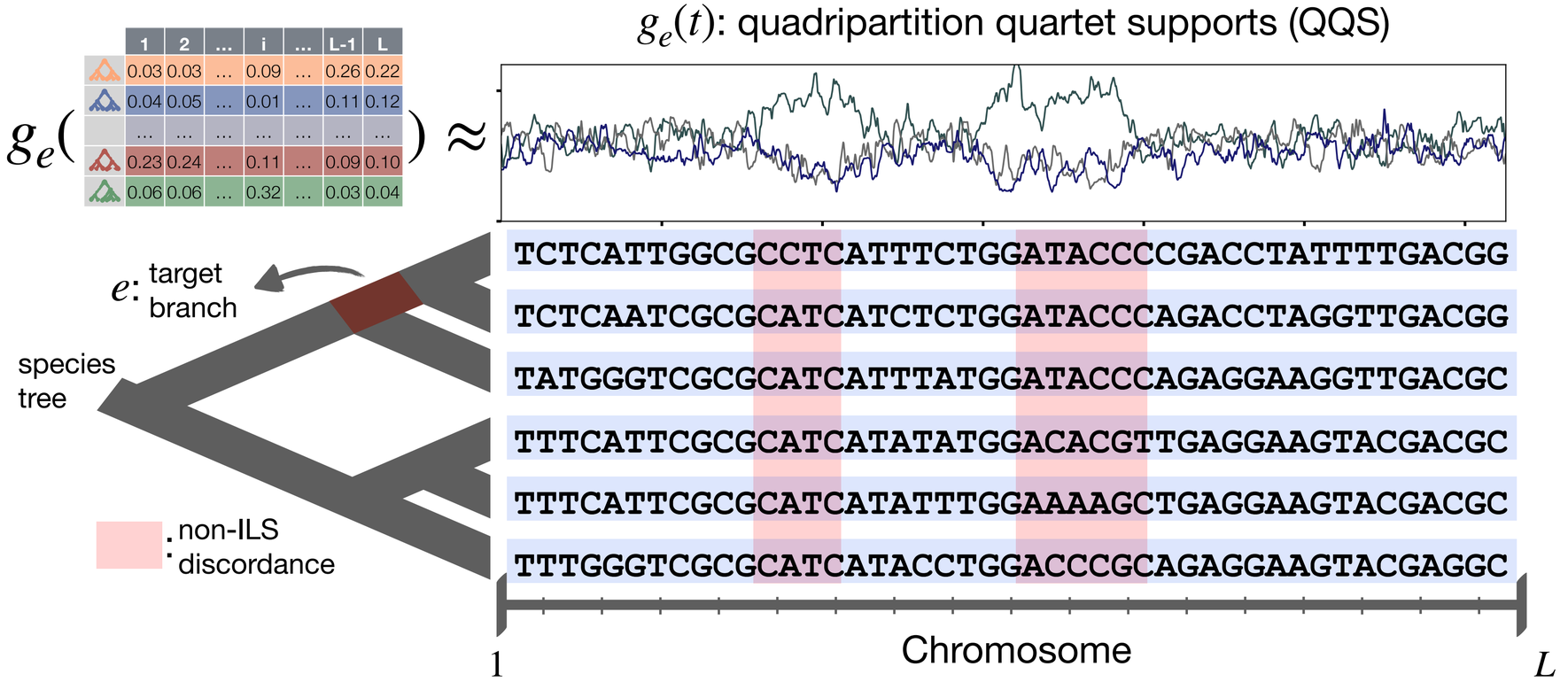
posterior probabilities

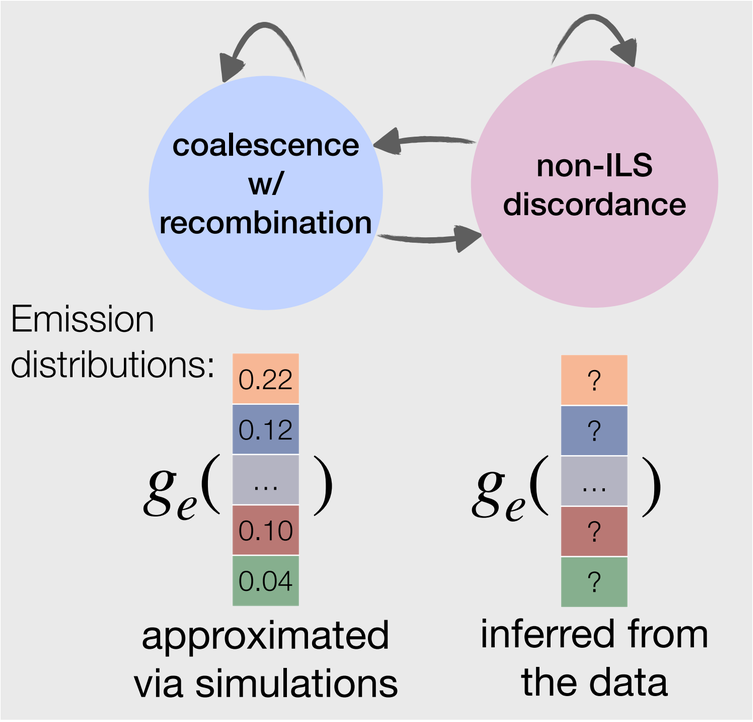
0.22
0.12
…
0.10
0.04

stationary distribution

An HMM w/ locus tree summary statistics as emissions

Given a species tree and gene trees sampled across a sequence:
- QQS values as emissions—*fast*
- approximating the emission distribution via simulations
- informative priors on transitions to deal with noise





Investigating Overfitting in Maximum Likelihood Phylogenetic Inference: A Systematic Approach

Anastasis Togkousidis

IMSI Workshop, August 2025

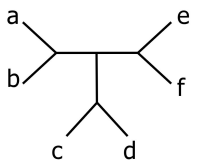
Problem

Input MSA

```
a: A G - C T T A
b: A C C C T T A
c: A G C C T T A
d: A G C C - T T
e: T C C T - T T
f: A G C C T T -
```

→ **ML tree inference** →

Binary tree



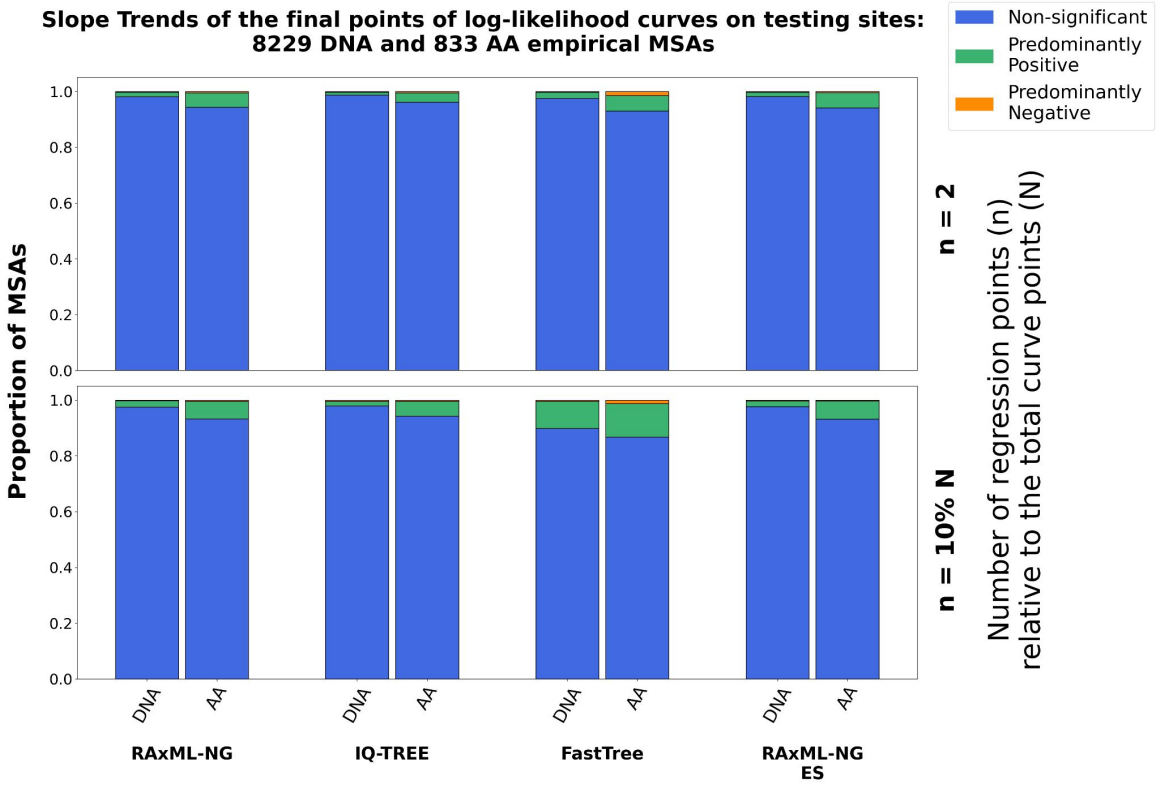
Notes:
1. **NP-hard** problem; tools deploy **iterative heuristics**
2. **MSAs** are **noisy**; risk of **overfitting**

Questions we ask

Do ML tree inference tools systematically overfit? ❌

Does holdout-based Early Stopping perform well in ML tree inference? ❌

Preliminary Results



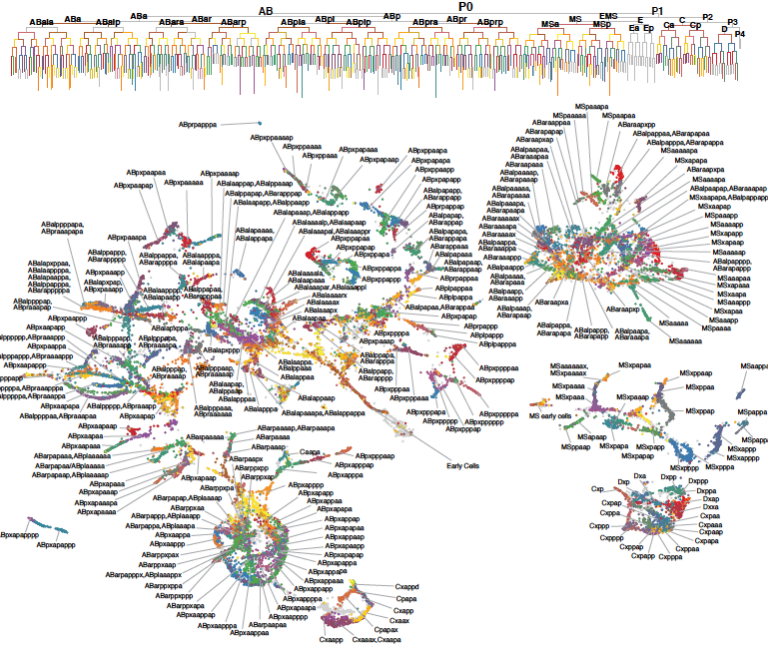
There is no systematic overfitting

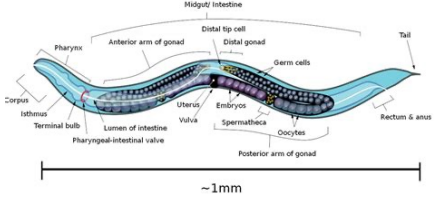
FastTree has higher fraction of **positive trends**

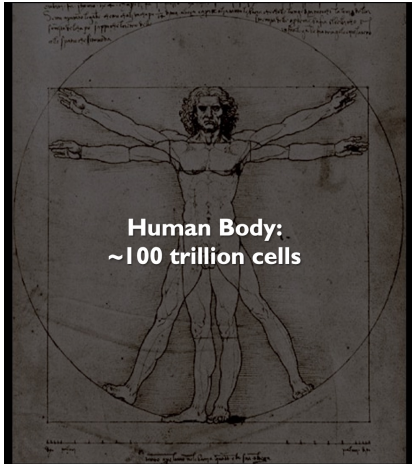
Reconstructing Cell Lineage Trees with Metric Learning

Da Kuang, Guanwen Qiu, Junhyong Kim
University of Pennsylvania
IMSI Workshop 2025

All cells in a metazoan originates from a single fertilized embryo by replication and differentiation.







Can we infer cell lineage trees from molecular phenotype data (e.g., transcriptome)?

#3

Challenge

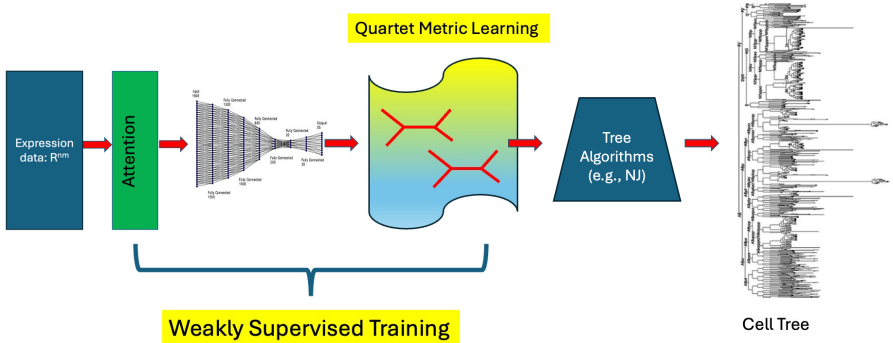
Fully supervised learning is impractical since cell history is unavailable.

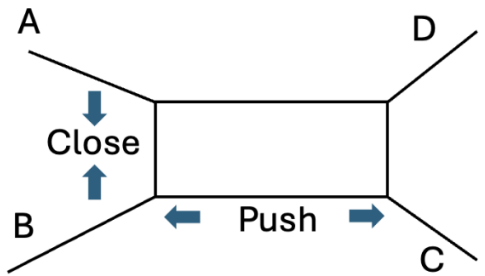
Opportunity

We found **Quartets can be learned with minimal supervision**.

Our Approach

Formulate lineage reconstruction as **metric learning** to generalize quartet graph structures.





$$\min_{f,T} \sum_{q \in \mathcal{Q}} \mathcal{L}(D(f(x_q)), D_T(x_q)) + \lambda \Omega(f)$$

UC San Diego

SPrUCE: Estimating Population-Level Nucleotide Diversity from UCEs

Contemporary Challenges in Large-Scale Sequence Alignments and Phylogenies

Daira Melendez, Ali Osman Berk Sapci, Vineet Bafna, Siavash Mirarab

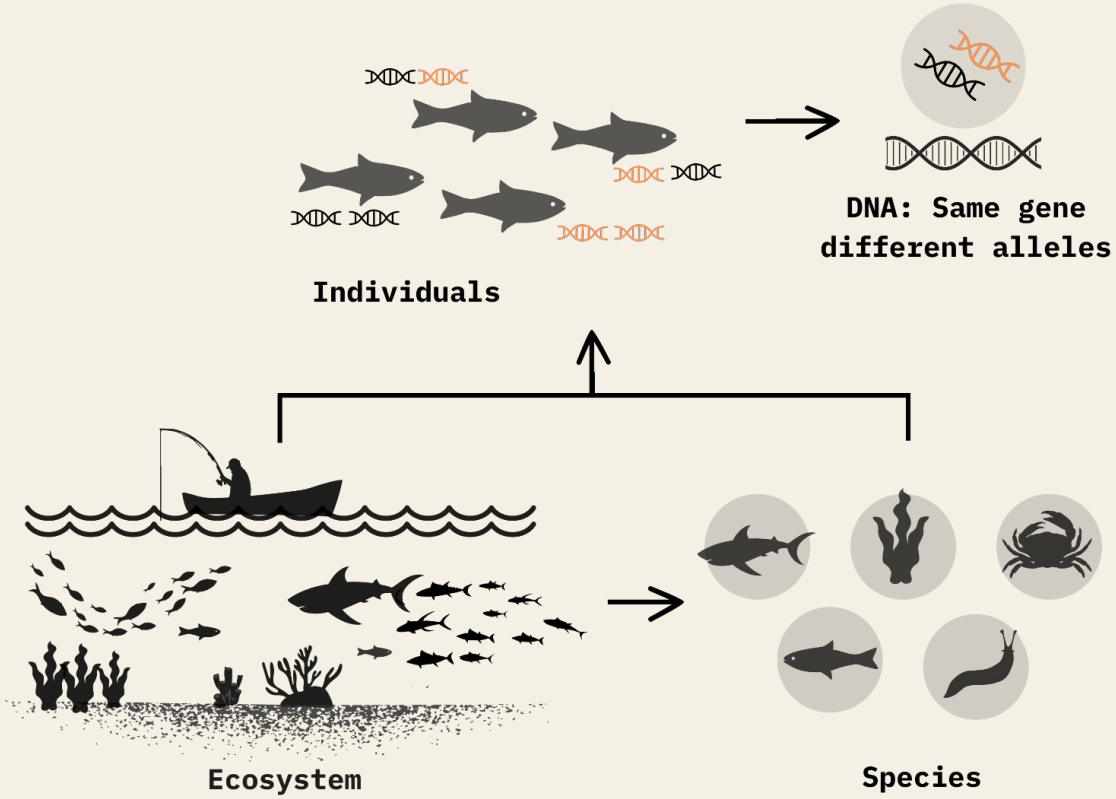
August 11, 2025

ALFRED P. SLOAN
FOUNDATION

Biodiversity Monitoring

Importance:

- Ecosystem stability



Traditional/Current Approaches:

- Targeted sequencing (16S, mtDNA)
- Whole genome sequencing
- Genome skimming (.5 - 4x coverage)

Challenges:

- Biomarkers: bias
- WGS: costly
  - Large alignments/computationally expensive
- Skimming: resolution depending on coverage

Proposed Solution:

- Biomarkers (cost-effective, scalable)
  - **Ultraconserved elements of DNA (UCEs)**
- Model data for population-level diversity estimates

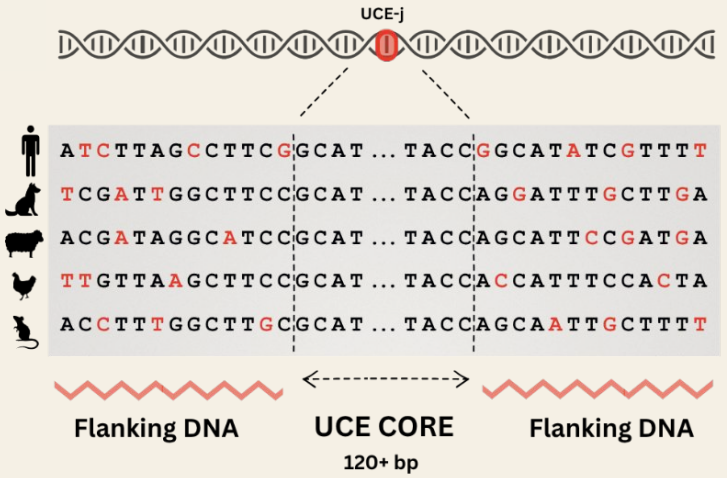
Population Genetics with UCEs

UCEs are **highly conserved** regions of DNA

- ~100-300 bp in length
- >97% seq. identity across species

For **population-level** genetics:

- **Flanking DNA**
  - surrounding UCE core
  - Variable and informative



Estimating nucleotide diversity

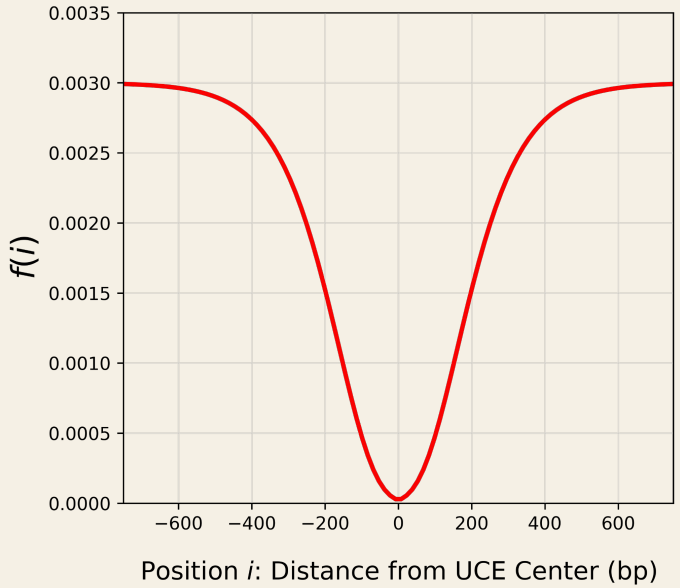
- Tajima's π : pairwise differences

$$\pi = \frac{1}{L}\sum_{i=1}^{L}\frac{s_i(n_i - s_i)}{\binom{n_i}{2}}$$

★ **π is biased near conserved core**

Gompertz Function:

$$f\,(i \mid \theta, \beta, \gamma) = \theta\, e^{-\beta\, e^{-\gamma |i|}}$$

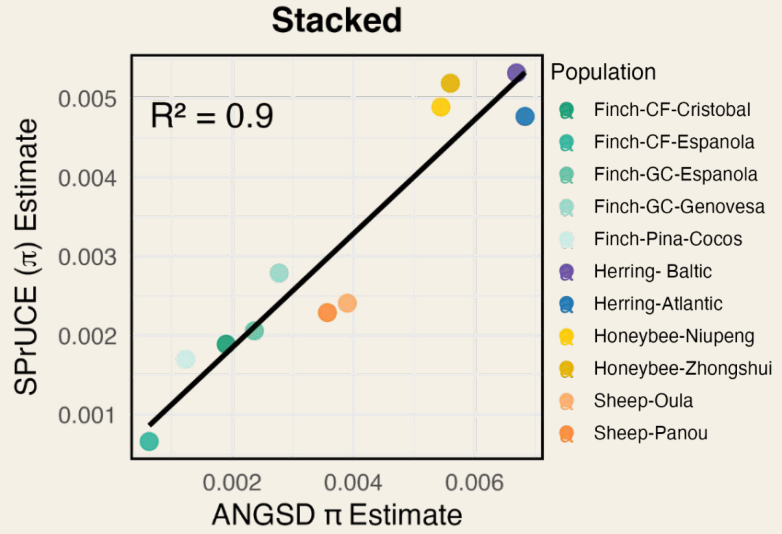


SPrUCE: Sigmoid Pi requiring UCEs

- Estimates Tajima's π
- Models diversity across flanking region using Gompertz function
- Optimizes by minimizing squared error (argmin)

Two methods: Stacked & Concatenated

$$\arg\min_{\theta,\beta,\gamma}\sum_{i=-l}^{l}\left(f(i;\theta,\beta,\gamma) - \frac{1}{k}\sum_{j=1}^{k}\pi_{i,j}\right)^2$$

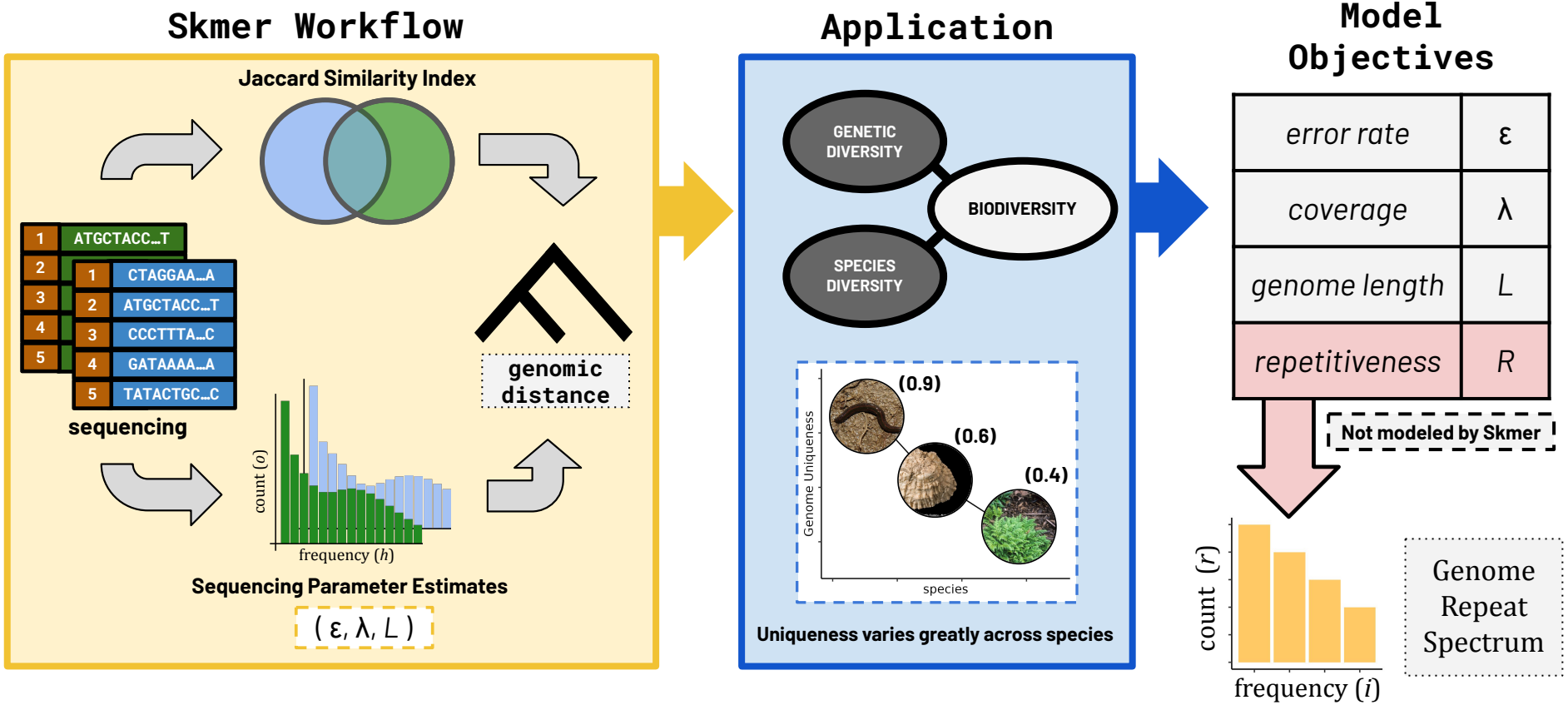


ReSkmer:
Modeling Repeats Improves Low-Coverage, Alignment-Free Estimates of Genomic Distance

Eduardo Charvel, Isaac Thomas, Homere J Alves Monteiro,

Glenn Dunshea, Vineet Bafna, Siavash Mirarab

Background: Accurate and Efficient Distance Estimation



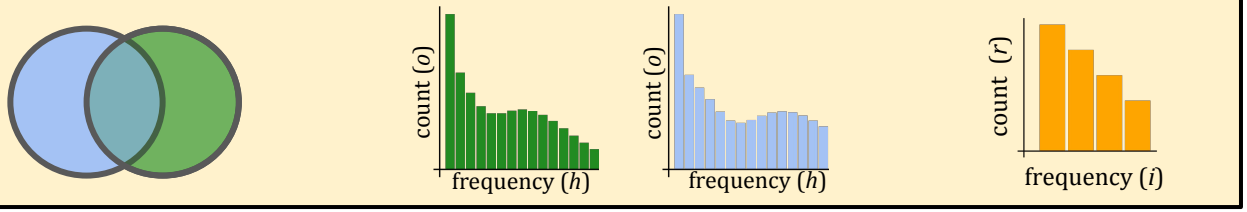
Methods: New Repeat-Aware Distance Estimator

ReSkmer
(models k-mer repeats)

`k-mer set intersection (I)`

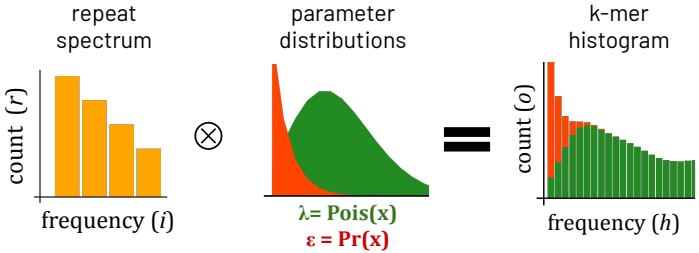
`observed k-mer histograms (O)`

`genome repeat spectrum (R)`



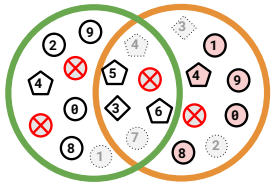
1. **Ordinary Least Squares**

estimate sequencing parameters using ***R*** and ***O***



2. **Repeat-Aware Intersection Model**

estimate distances from ***I*** using ***R*** and ***parameters***



I_0 intersection due to homology

I_1 intersection due to error

TIPP3: Improved abundance profiling in metagenomics

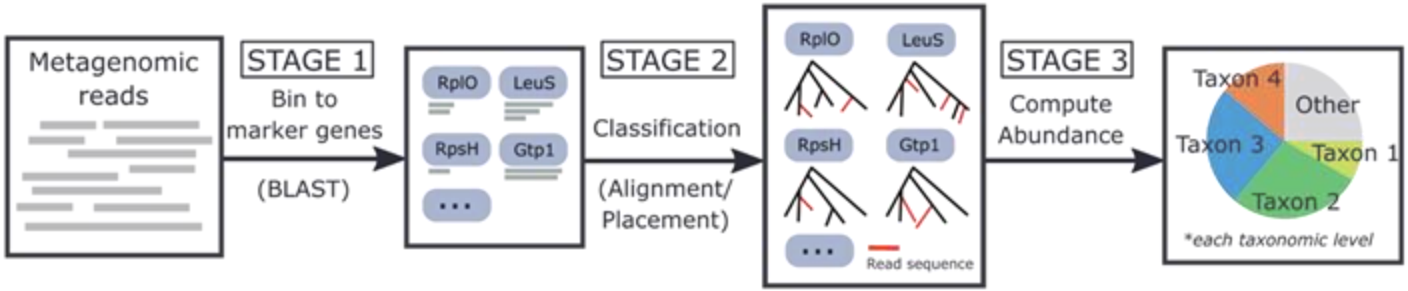
Presented by: Eleanor Wedell

C. Shen, E. Wedell, M. Pop, and T. Warnow, “TIPP3 and TIPP3-fast: Improved abundance profiling in metagenomics,” PLOS Computational Biology, vol. 21, pp. 1–21, 04 2025.

Goal of TIPP3: More accurate abundance profiling

Abundance Profiling – the distribution of species, genera, etc. in a metagenomic sample (*e.g. 20% Species A, 40% species B, … , for each taxonomic level*)

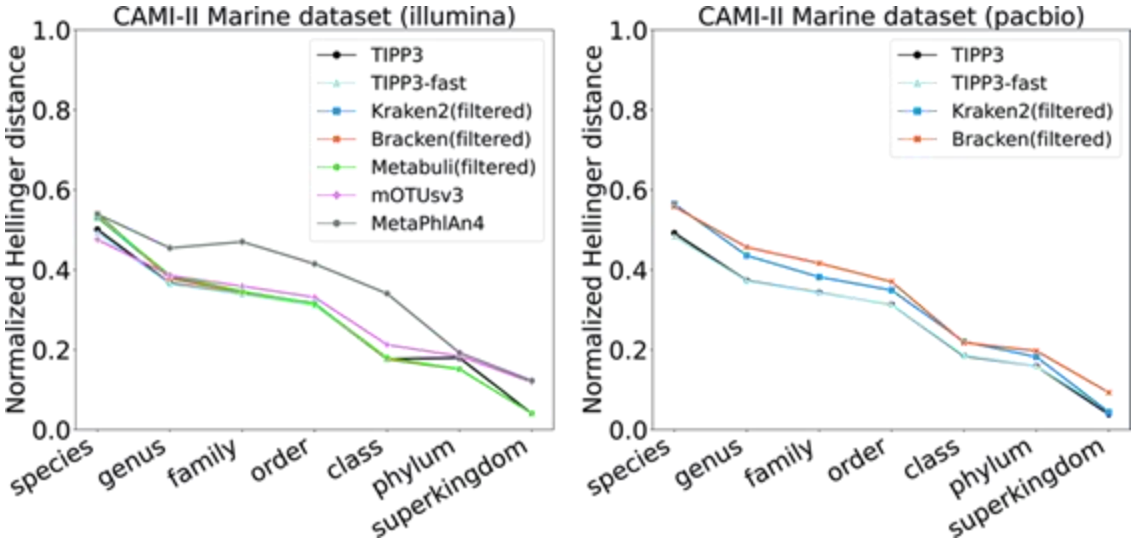
TIPP and TIPP2 (abundance profilers) both use a similar pipeline:



TIPP3 improves on TIPP2:

- Using BSCAMPP to do phylogenetic placement into large taxonomies with ~50,000 sequences
- Improved read alignment techniques (e.g., WITCH)
- Improved marker gene selection

Brief Preview of Results: CAMI-II Marine Dataset



“(filtered)” next to a method means the method was restricted to input reads mapped to marker genes by TIPP3.

Our study shows:

- Filtering improves accuracy for Bracken, Kraken2 and Metabuli
- TIPP3 is more accurate, especially when reads
  - are from novel genomes,
  - contain sequencing error

Machine learning enables alignment-free distance calculation and phylogenetic placement using k-mer frequencies

Eleonora Rachtman, Yueyu Jiang, Siavash Mirarab

08/11/25

UC San Diego

Distance based phylogenetics

Sequences

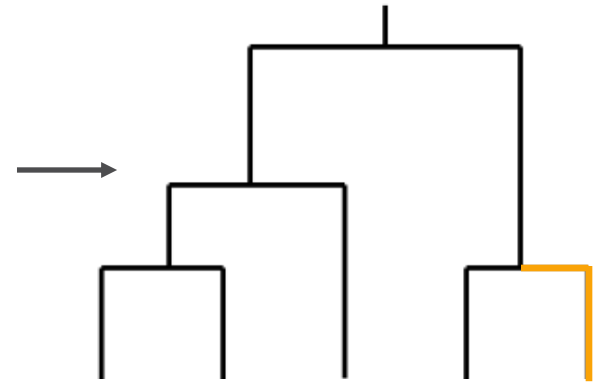
H	A	T	G	C	C	A	A	G
C	A	T	G	C	C	A	A	G
G	A	T	G	C	T	A	A	G
B	A	C	G	T	T	A	C	G
S	A	C	G	T	T	G	C	G

$d = 1/8$

Distance matrix

	H	C	G	B	S
H	0	0.0	0.2	1.1	1.5
C		0	0.2	1.1	1.5
G			0	0.7	1.1
B				0	0.2
S					0

Phylogenetic tree

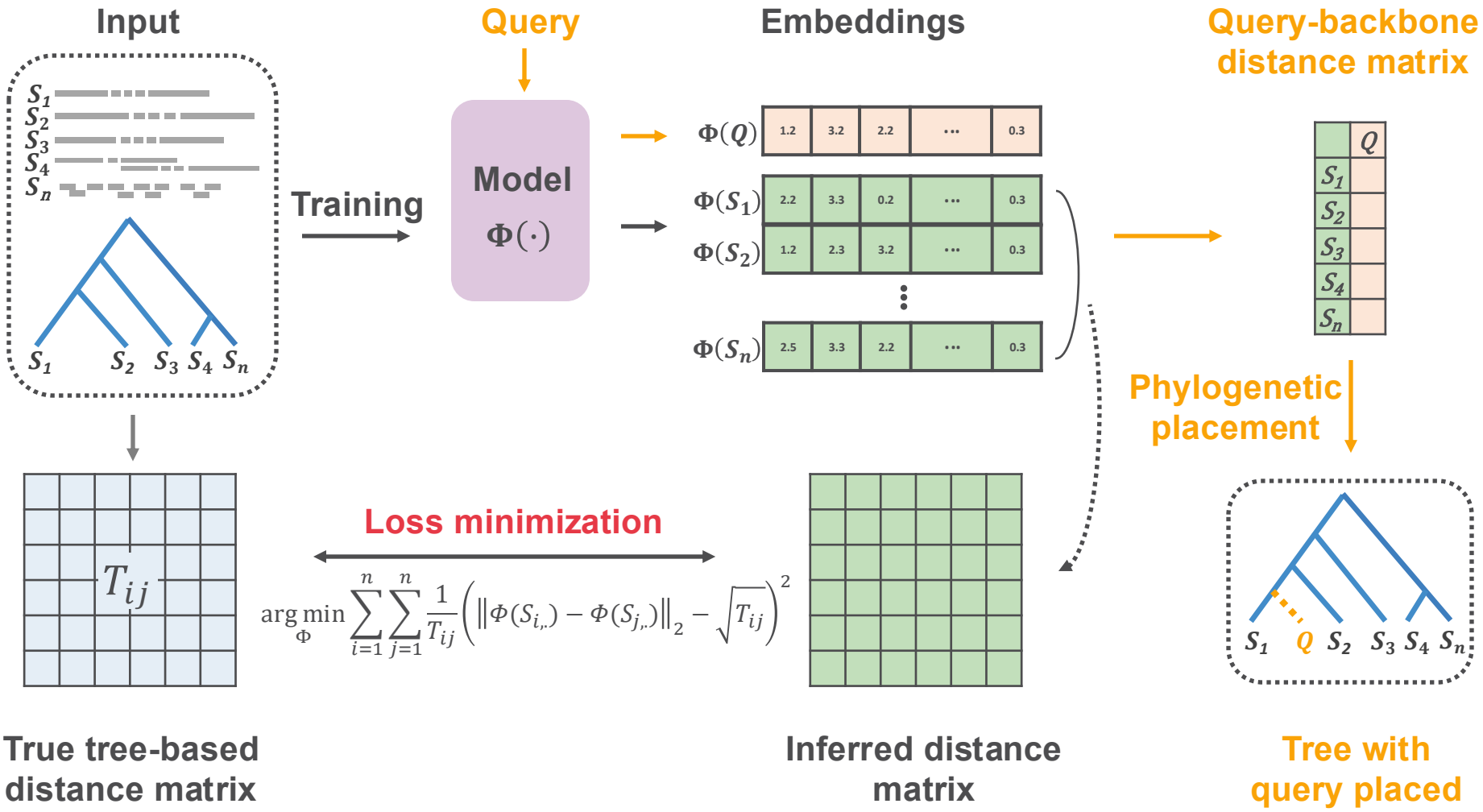


Jukes-Cantor model

$$\text{Distance} = -\frac{3}{4} ln(1 - \frac{4}{3} d)$$

- **Can we learn a distance function without explicitly modeling mechanistic parameters?**
- **Can workflow remain alignment-free?**

Phylogenetic placement using machine learning



https://github.com/noraracht/kf2vec

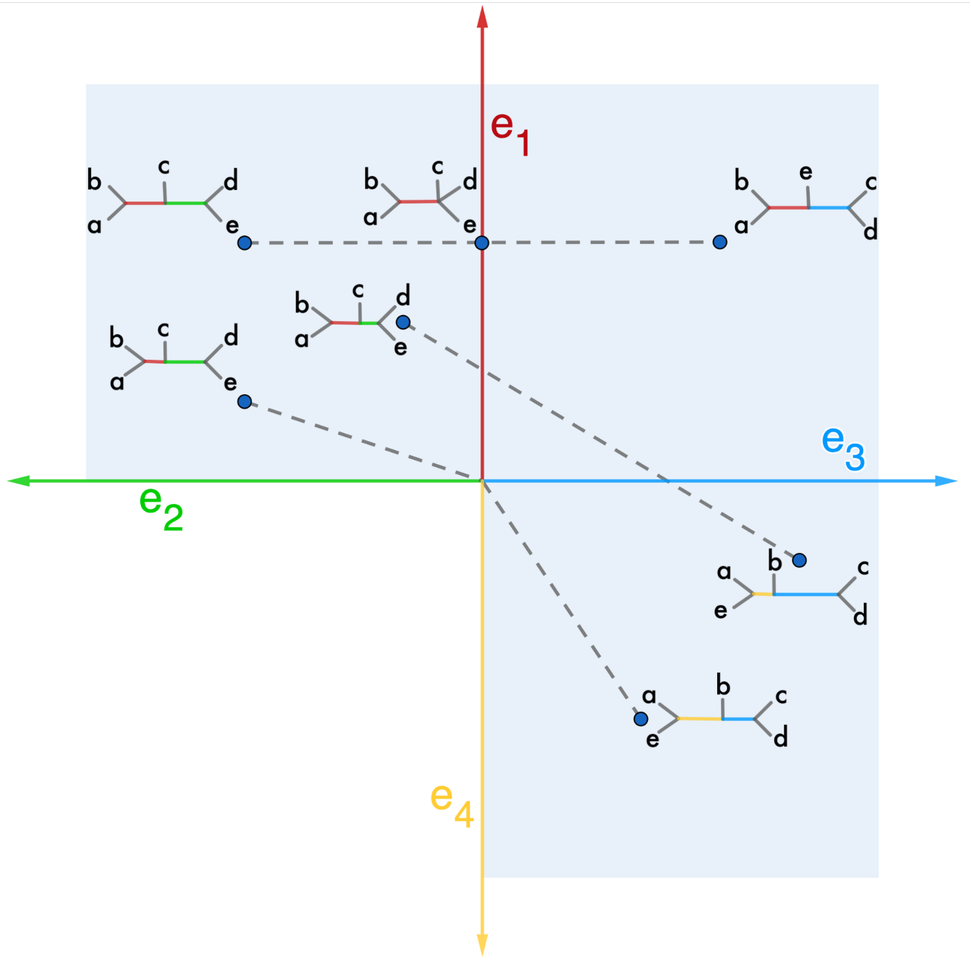
Preprint DOI: 10.22541/au.174539883.30231226/v1

Towering Tree Space: a metric between trees with differing leaf sets

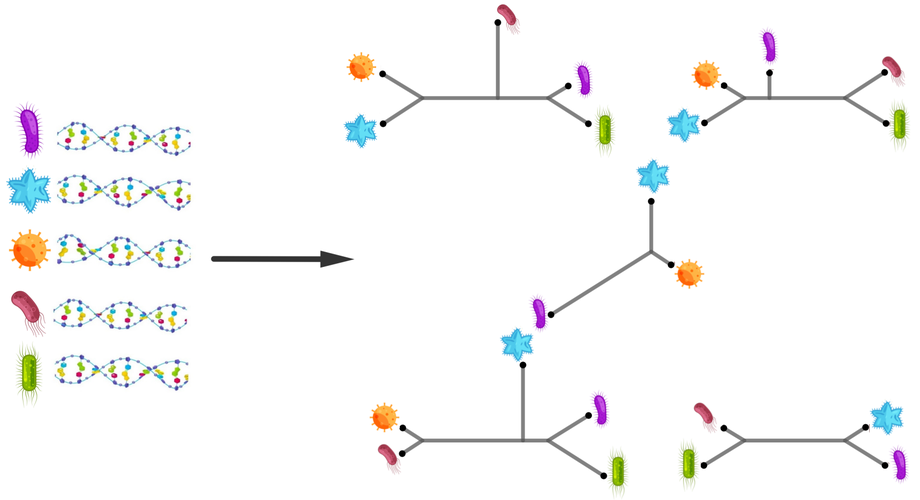
MARÍA VALDEZ-CABRERA, PHD
BIOSTATISTICS, UNIVERSITY OF WASHINGTON

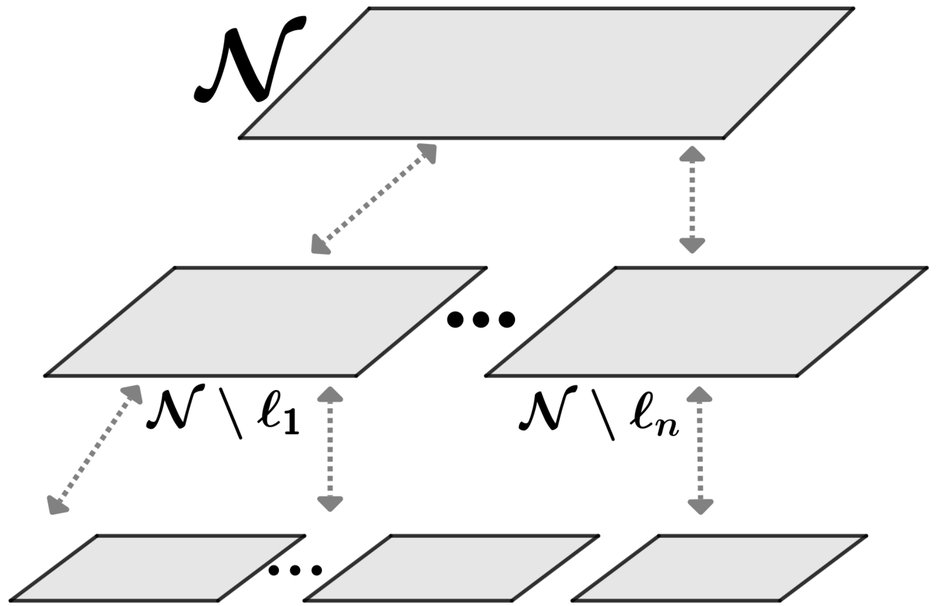
Contemporary Challenges In Large-Scale Alignments and Phylogenies
August 11th - 14th, 2025

SET-UP AND CHALLENGE



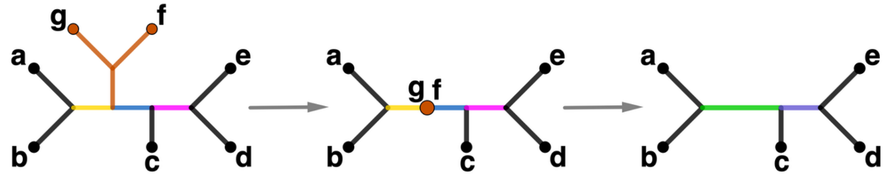
$$\sqrt{\sum_{i=1}^{k}\left(||A_i||_2+||B_i||_2\right)^2+\sum_{s\in C}\left(|s|_1-|s|_2\right)^2}$$



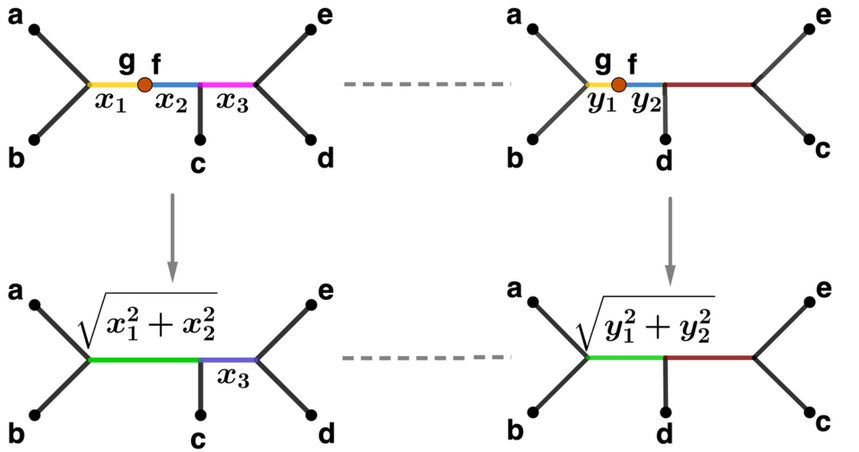


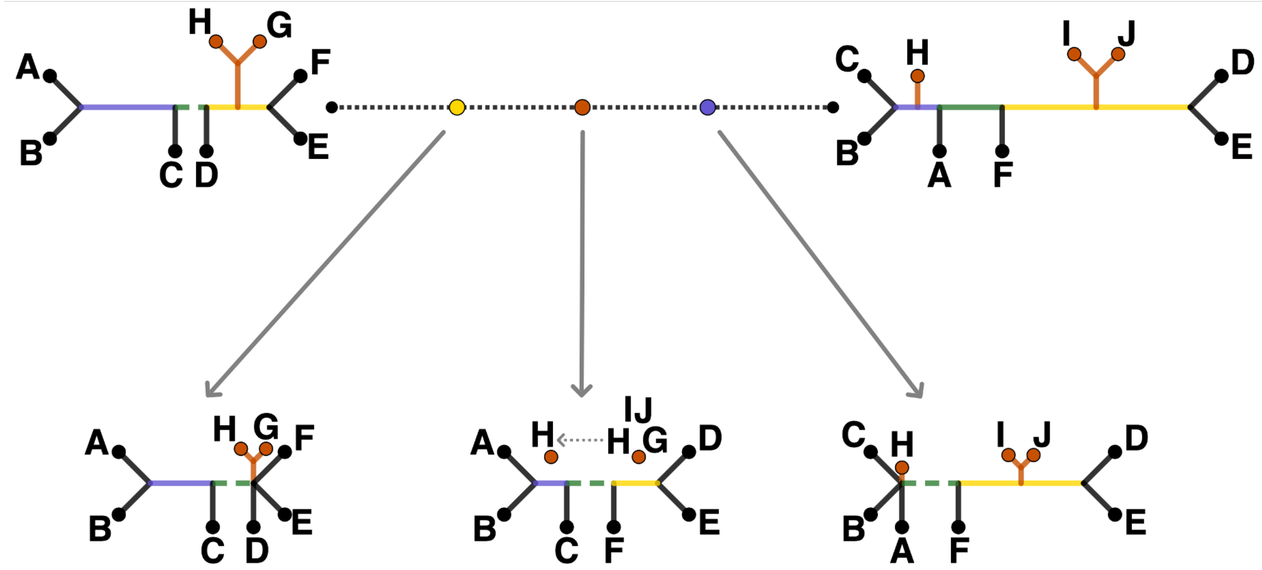
NEW METRIC: ✨TOWERING SPACE✨

Perform leaf prunings to transition



Merge edge lengths via L^2-norm





$$\sqrt{(||\bullet||_2 + ||\bullet||_2)^2 + (||\bullet||_2 + ||\bullet||_2)^2 + (||\bullet||_2 + ||\bullet||_2)^2 + (\bullet - \bullet)^2}$$

Theoretical and Empirical Performance of Pseudo-likelihood-based Bayesian Inference of Species Trees under the Multispecies Coalescent

Nicolae Sapoval

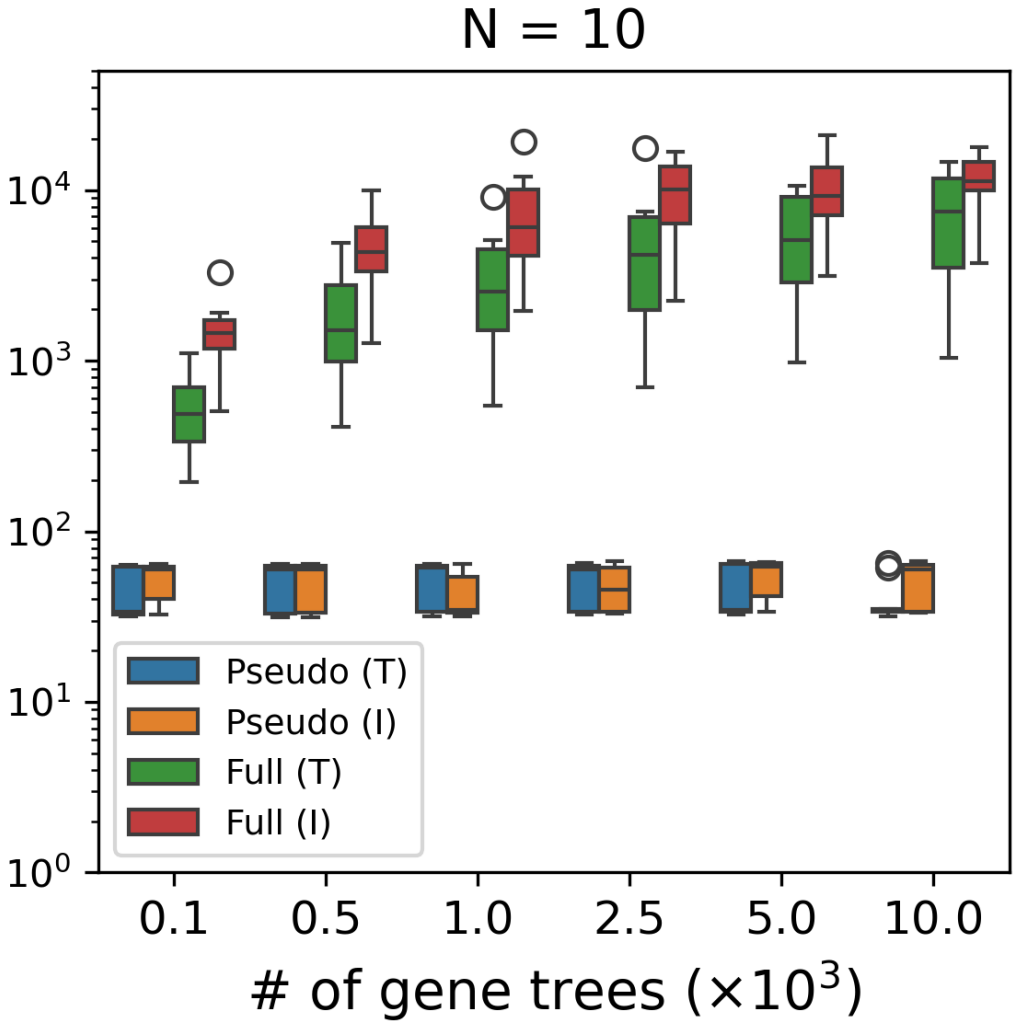
Rice University

August 11, 2025

Pseudo-likelihood is scalable

- Pseudo-likelihood: count rooted triplets from gene trees and use multinomial distribution arising from MSC for 3 taxa

- Easy to compute
- Statistically consistent MLE of species tree topology under MSC



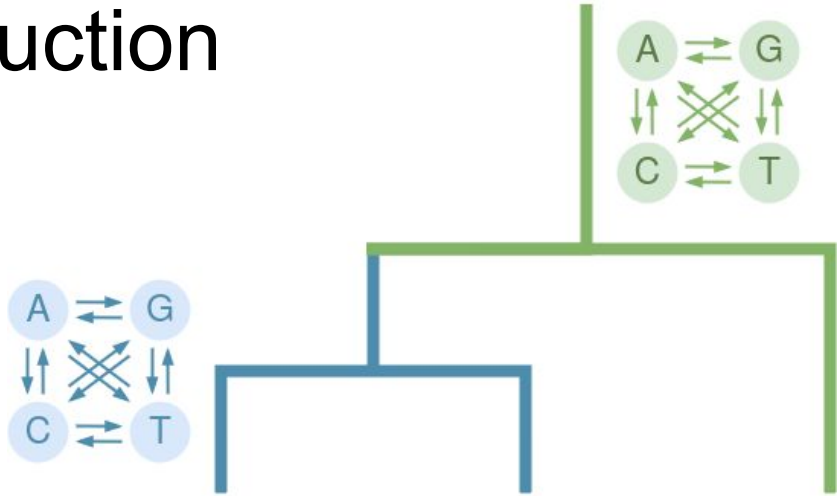
Can we do Bayesian inference with PL?

- Refined convergence bound for maximum pseudo-likelihood inference of the species tree topology
- Established a Bernstein-von Mises result under model misspecification for branch length estimation
- Conducted empirical evaluation of the described phenomena and assessed practical utility of Bayesian inference under pseudo-likelihood

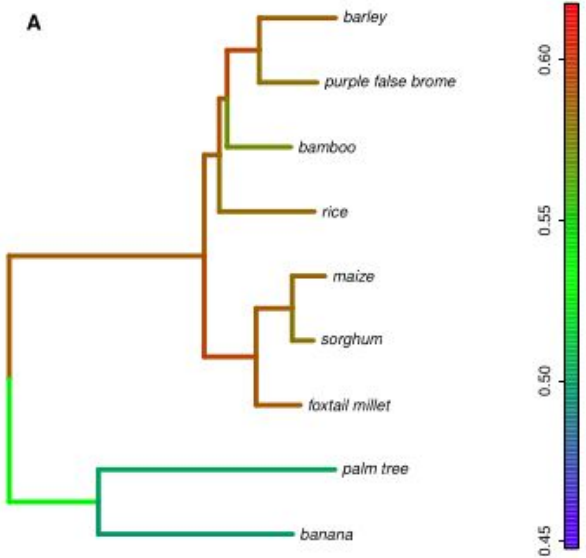
- Come to poster #9 to learn more about Bayesian inference with PL

A Performance Study of Statistical Methods for Phylogenetic Reconstruction Under Branch-Variable Substitution Models

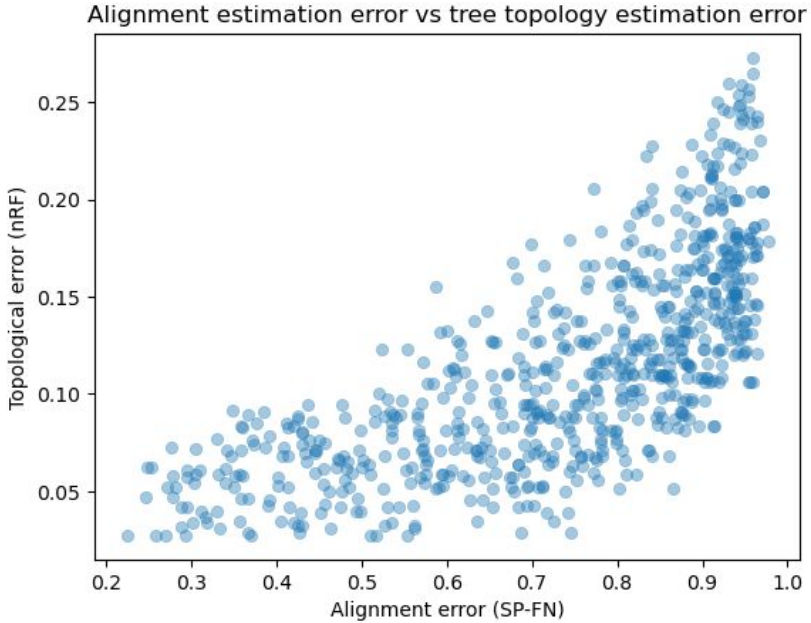
Rei Doko
Michigan State University



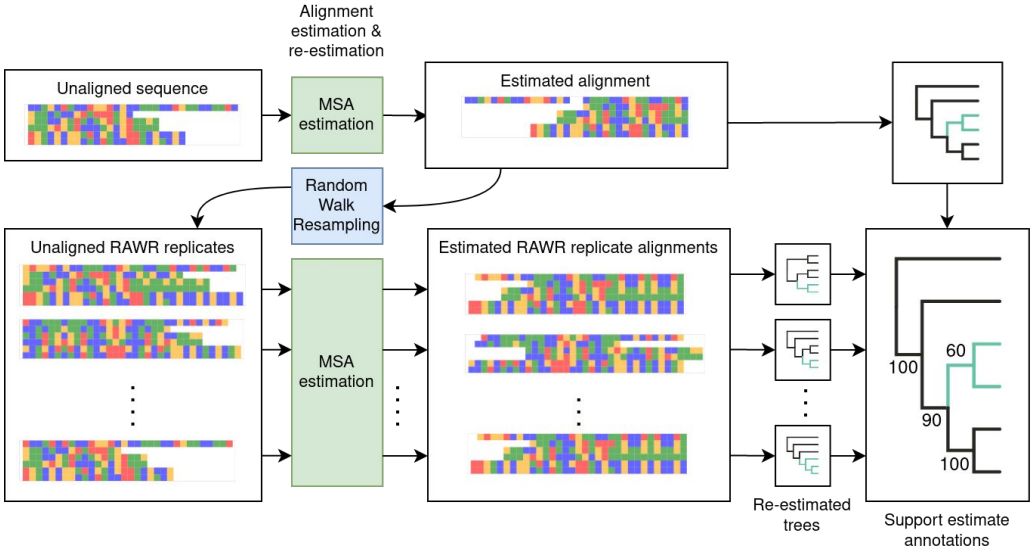
Complex sequence evolution & alignment quality



Reproduced from Clément et al., “The bimodal distribution of genic GC content is ancestral to Monocot species”, GBE 2015



MSA aware confidence for branch-variable model MLE



Model condition	ROC-AUC	
	Bootstrap	NoHTS
10.A	0.875	0.957
10.C	0.748	0.926
10.E	0.791	0.922

Data & scripts: https://gitlab.msu.edu/liulab/nonhomogeneous-substitution-model-study-data-scripts
Website: https://www.cse.msu.edu/~dokorei/

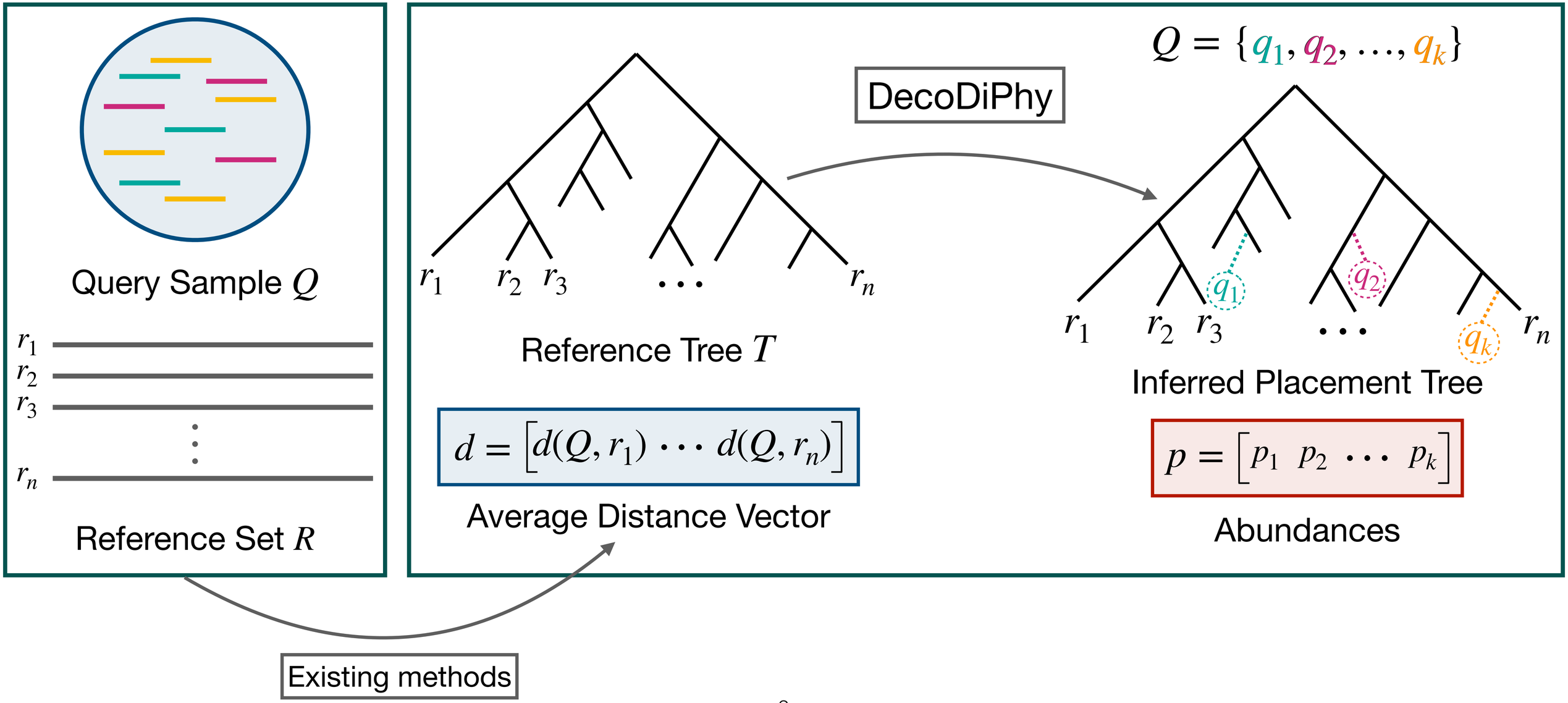
UC San Diego

Deconvolving Phylogenetic Distance Mixtures

Shayesteh Arasti, Ali Osman Berk Şapcı, Mohammed El-Kebir, Siavash Mirarab

IMSI 2025

Metagenomic Profiling

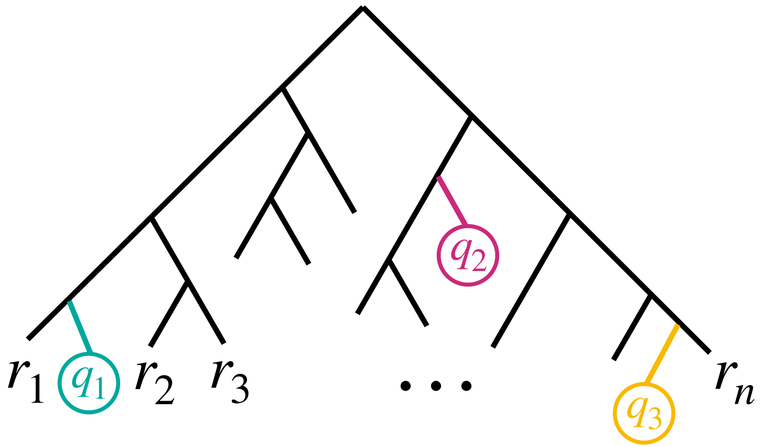


DecoDiPhy

Average Distance Vector

Abundances

$$d = \begin{bmatrix} d(Q, r_1) \\ d(Q, r_2) \\ \vdots \\ d(Q, r_n) \end{bmatrix} = \begin{bmatrix} d(q_1, r_1) & d(q_2, r_1) & \cdots & d(q_k, r_1) \\ d(q_1, r_2) & & & \\ \vdots & & \ddots & \vdots \\ & & & \\ d(q_1, r_n) & d(q_2, r_n) & \cdots & d(q_k, r_n) \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_k \end{bmatrix}$$



Phylogenetic Tree T

$D_S \in R^{n \times k}$ $\quad p \in [0,1]^k$

$$d = D \cdot A \cdot p + C \cdot L \cdot A \cdot (x \circ p) + \bar{y} \cdot \mathbf{1}_n$$