Deconvolving Phylogenetic Distance Mixtures

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

Metagenomic profiling (taxonomic)

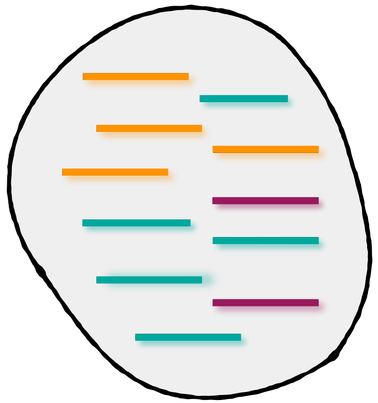
Identifying the taxa in a query sample $\mathscr{X}$ with respect to a reference set $\mathscr{R}$

non-trivial: lack of reference & uncertainity, relationships between references

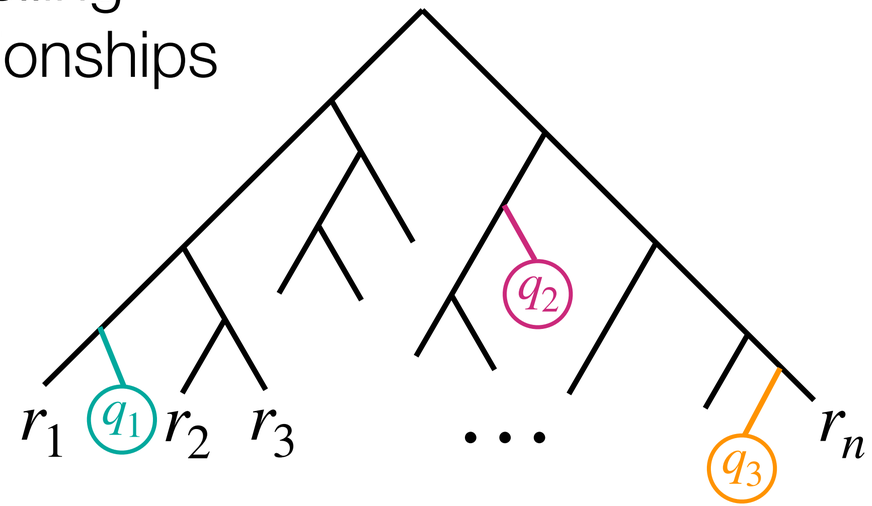
Metagenomic profiling (phylogenetic)

Phylogenetic placement all query taxa on a reference tree R

challenge: modelling evolutionary relationships



Query Sample $\mathcal{X}$



Phylogenetic Reference Tree R

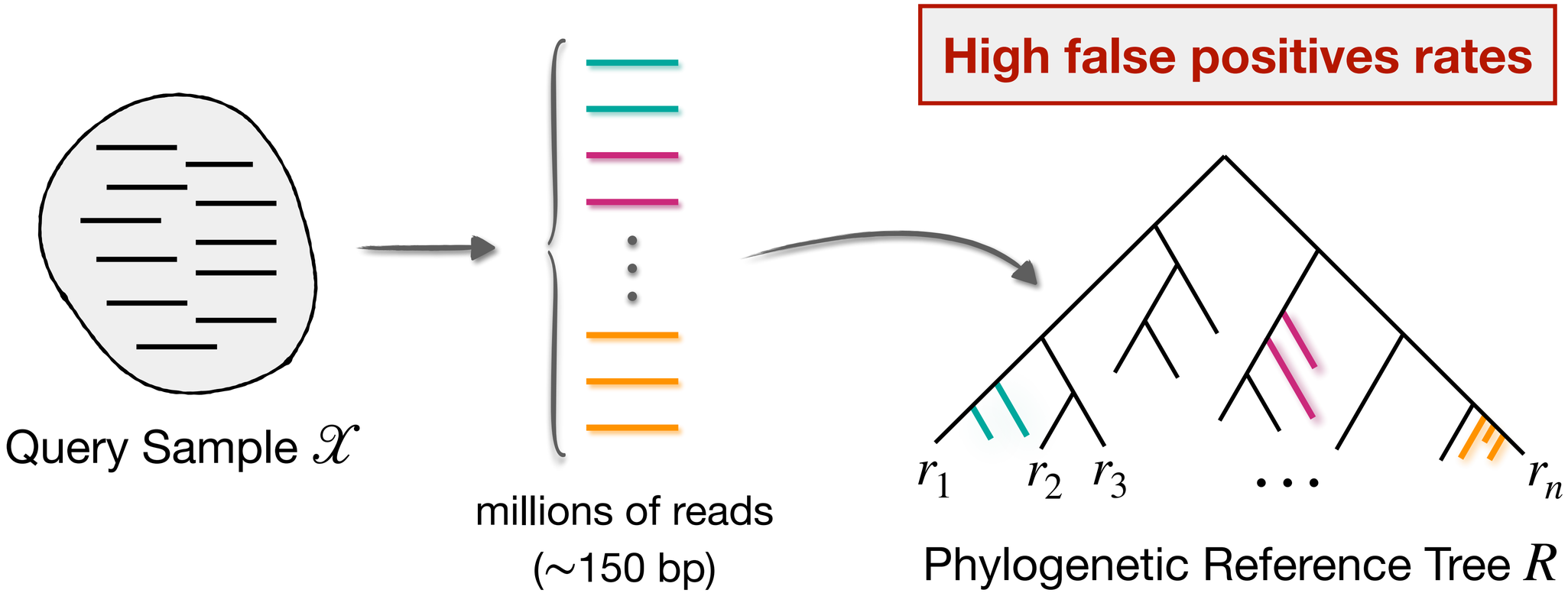
$$\begin{matrix} q_1 \\ q_2 \\ q_3 \end{matrix} \begin{bmatrix} 0.3 \\ 0.4 \\ 0.3 \end{bmatrix}$$

Abundances $\mathbf{p}$

Traditionally limited to marker genes…

Read-level metagenomic identification

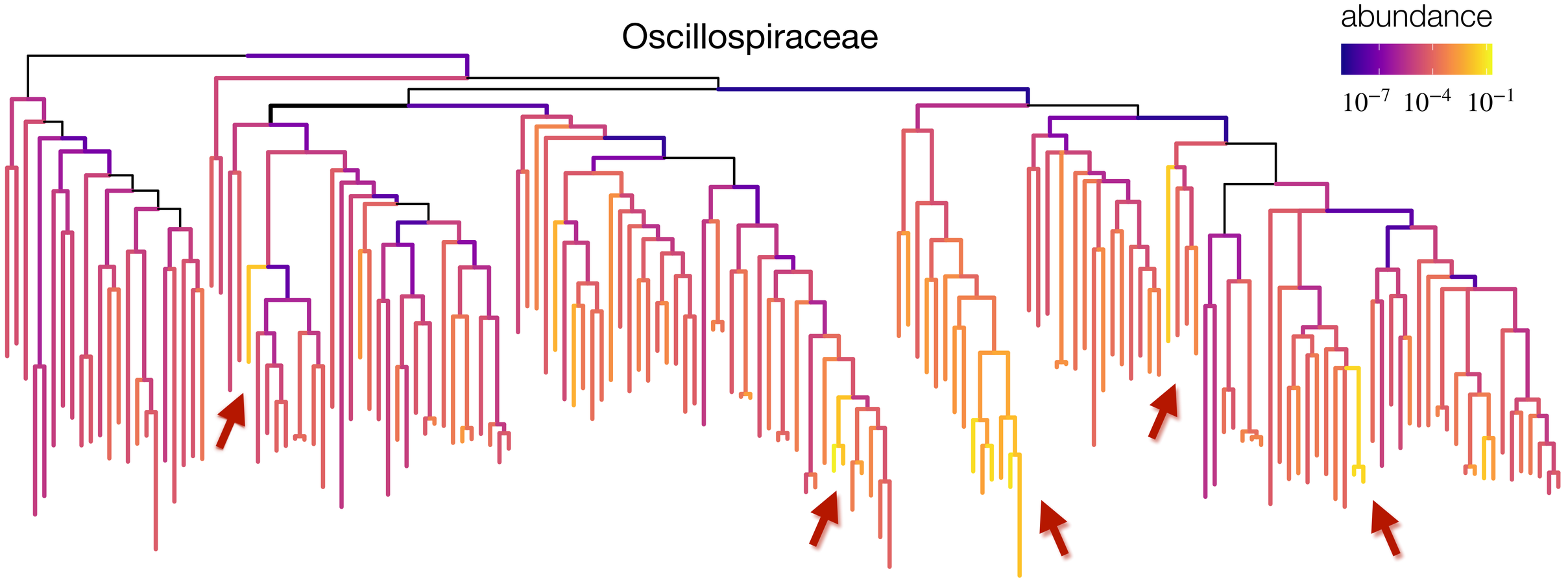
- Phylogenetic placement of individual reads: **krepp** [RECOMB 2025]
- Taxonomic classification (e.g., Kraken2) & read mapping are widely used



Placement of genome-wide reads (krepp)

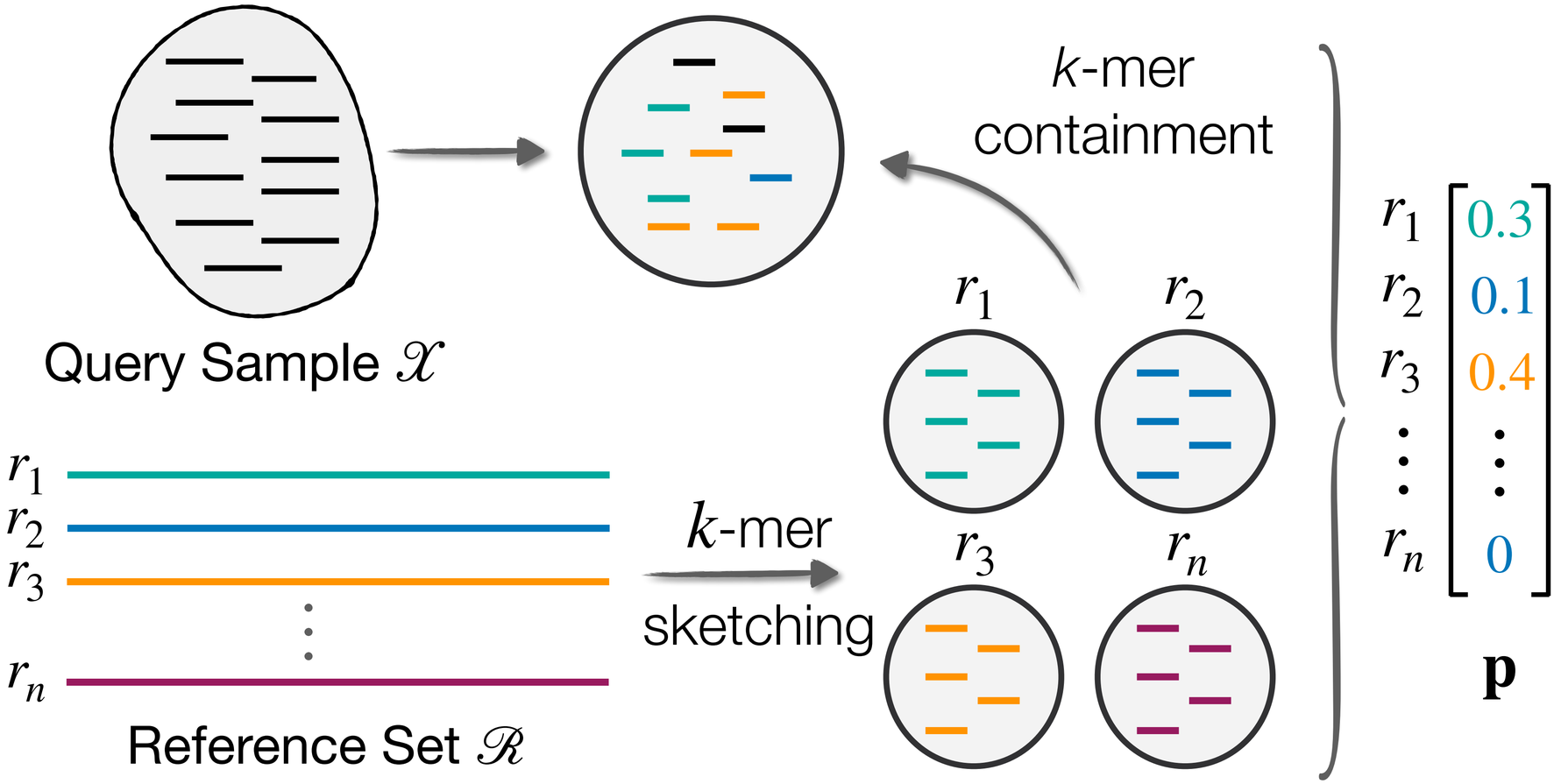
An inflammatory bowel disease (IBD) sample from Franzosa et al. 2019

Almost all branches have non-zero abundances → methodological artifacts

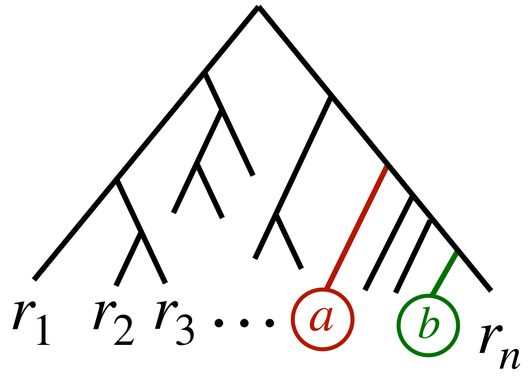


Joint abundance estimation

- Jointly estimating abundances and incorporating query dependencies
- Sketching-based: sourmash (Irber et al. 2022) & sylph (Shaw and Yu 2024)



Struggle to identify novel taxa



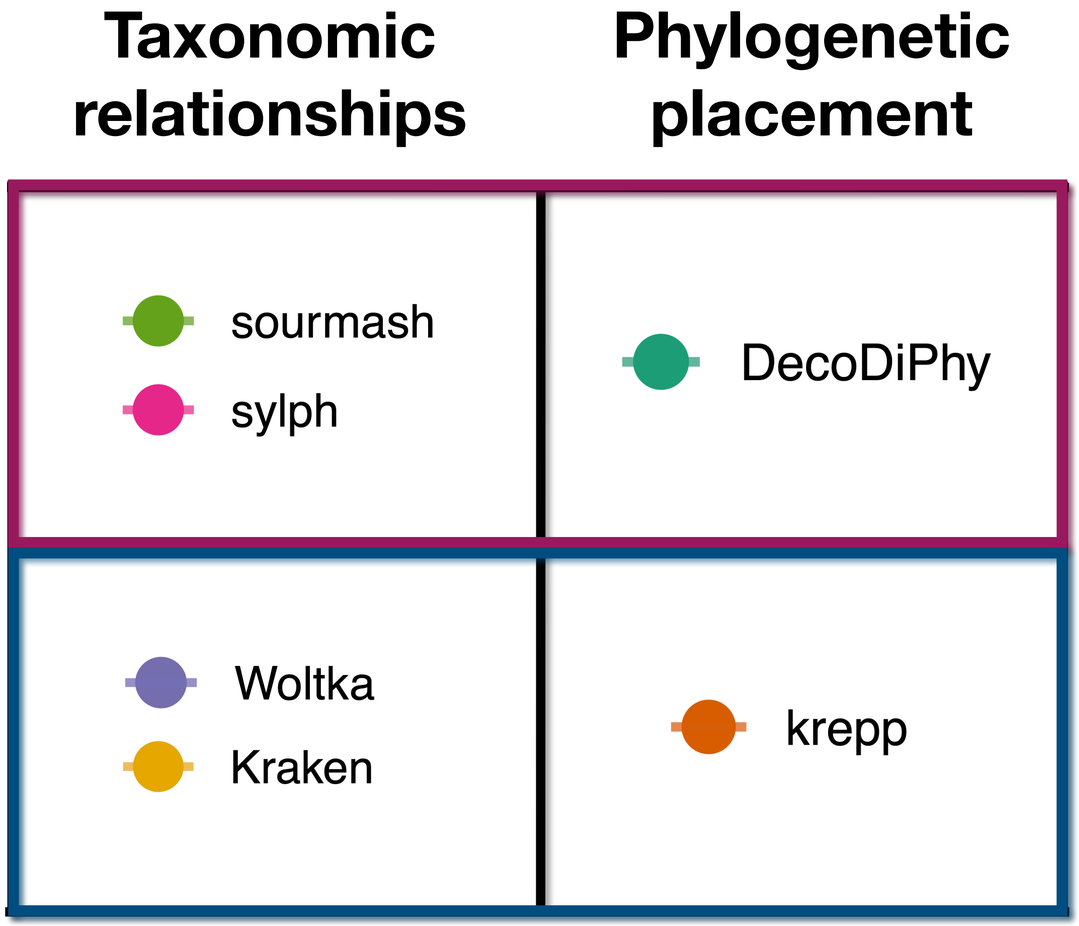
Spectrum of metagenomic analysis

fewer features
& consolidated

Joint analysis

Read-level analysis

noisy & with FPs

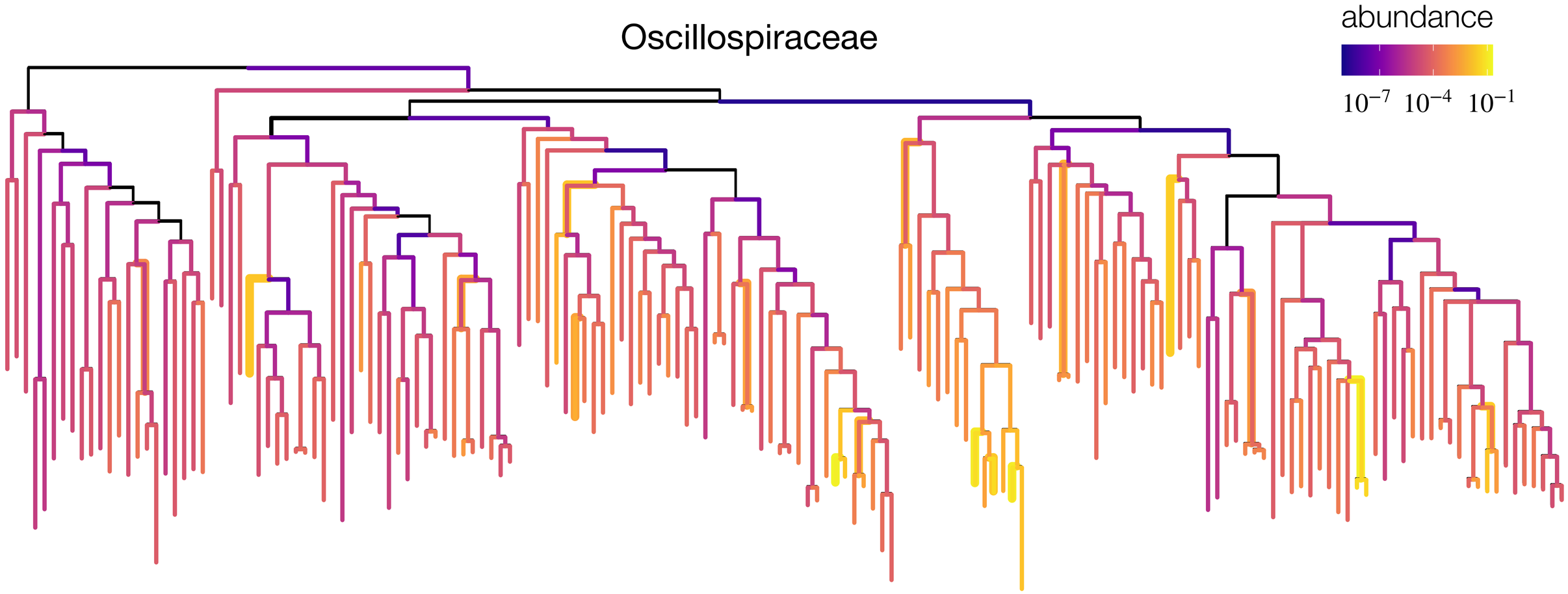


…many other methods…

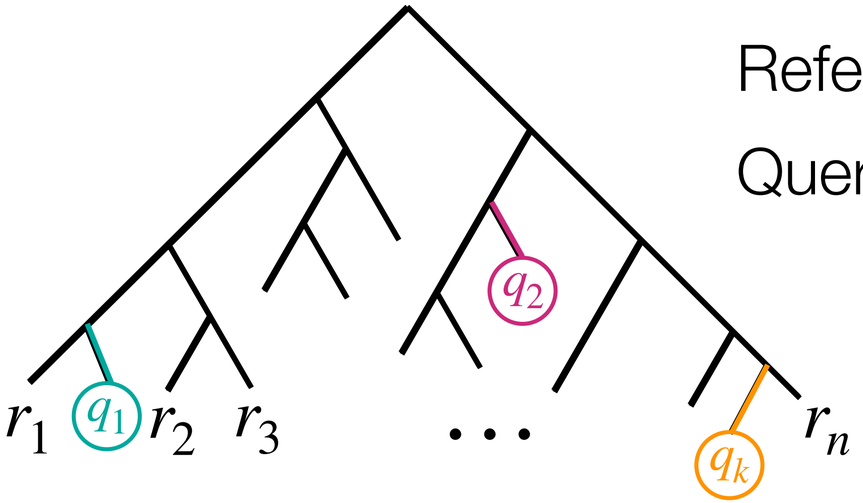
DecoDiPhy: consolidating the signal & interpretability

An inflammatory bowel disease (IBD) sample from Franzosa et al. 2019

Deconvolution on a reference tree with 16,000 leaves



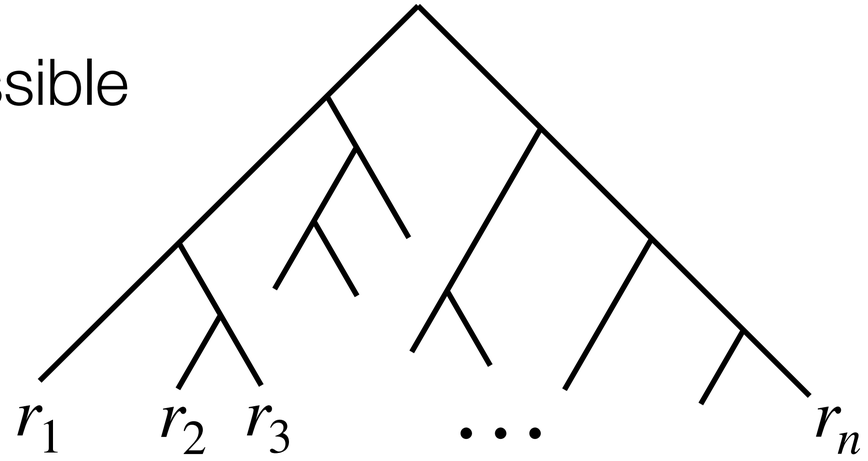
Formulating the problem



A true tree F
(latent)

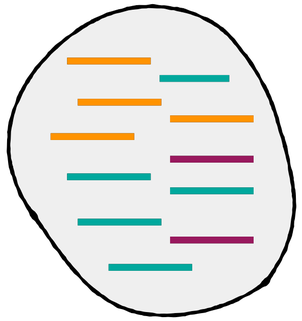
References $r_i \in \mathcal{R}$ are accessible

Queries $q_i \in \mathcal{Q}$ are unknown



Reference tree R

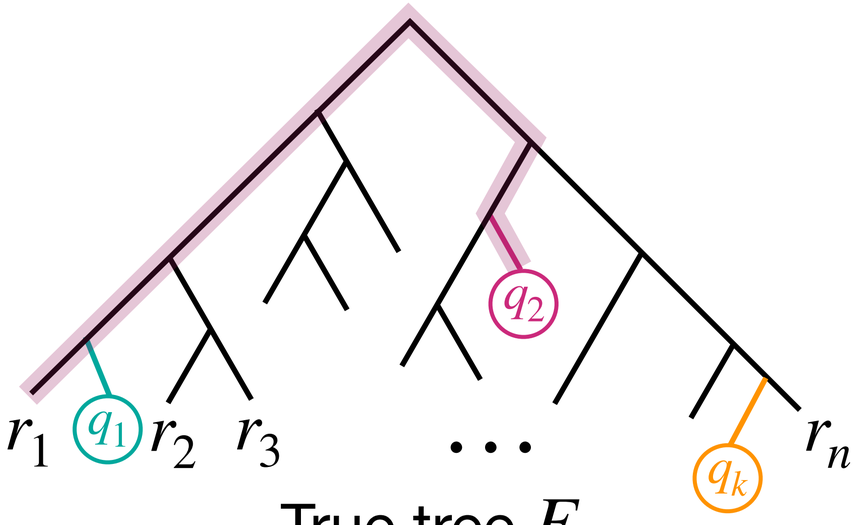
A set of reads $\mathcal{X}$:



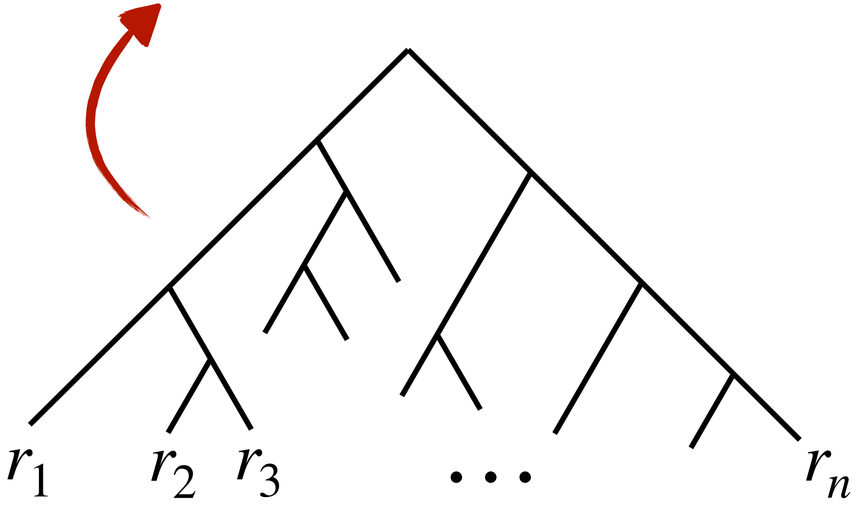
Each $x \in \mathcal{X}$ belongs to an unknown $q \in \mathcal{Q}$

p_q: abundance of q

Phylogenetic Distance Deconvolution (PDD)



True tree F



Reference tree R

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

$D(F) \in \mathbb{R}^{n \times k}$ $\quad \mathbf{p} \in [0,1]^k$ $\quad \mathbf{d} \in \mathbb{R}^n$

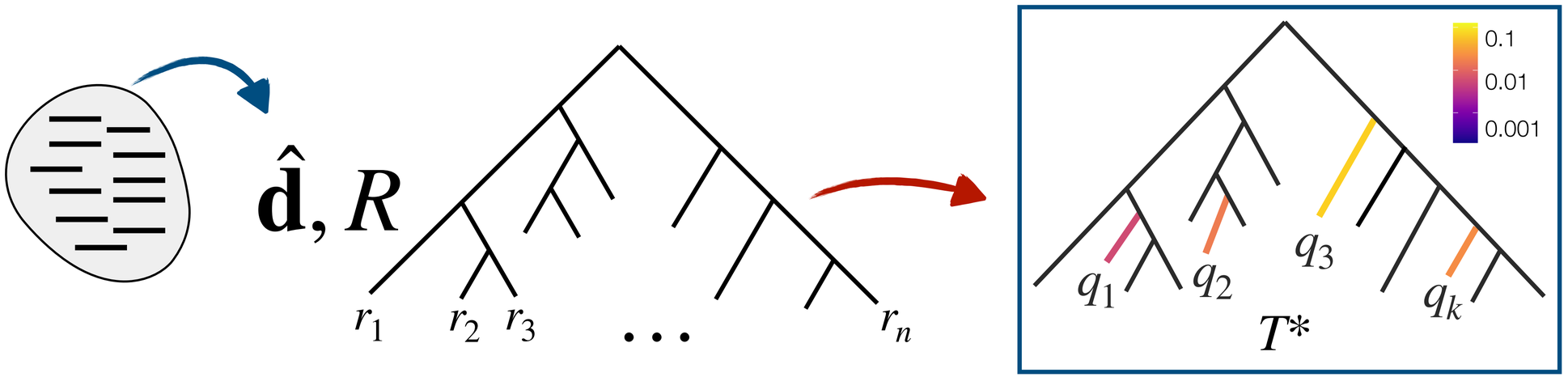
Core problem:

Given R and estimated mixture distances $\hat{\mathbf{d}}$,

find an F that induces R & proportions $\mathbf{p}$

DecoDiPhy: Solving the PDD optimization

Given a reference tree & an distance mixture, find multi-placements:



Placements
$q_1, q_2, \ldots, q_k$
with abundances
$p_1, p_2, \ldots, p_k$

$$T^*, \mathbf{p}^* = \text{argmax}_{T \in \mathcal{T}(R,k),\ \mathbf{p}} \|\hat{\mathbf{d}} - D(T) \cdot \mathbf{p}\|_2^2$$

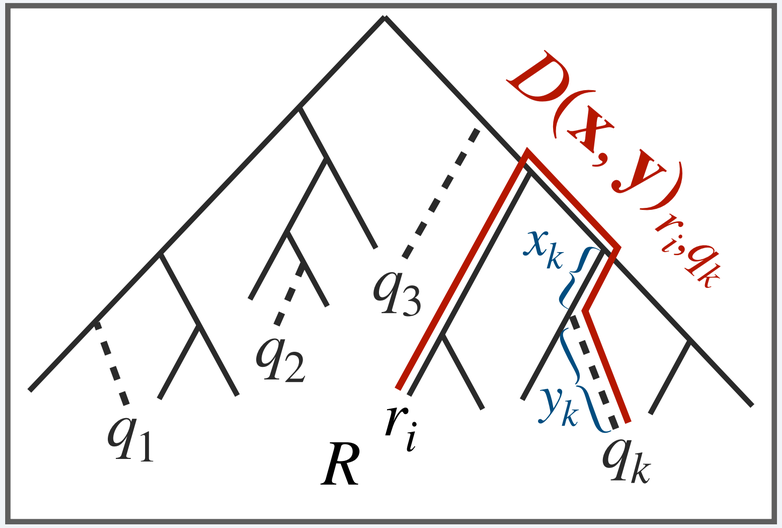
$\mathcal{T}(R,k)$ = Space of all trees resulting from k placement on R $\rightarrow O(n^k)$

Fixing the discrete k-placements & optimizing the continuos variables

$\mathcal{T}(R, k)$ and k are **combinatorial:** $O(n^K)$ for bounded $k \leq K$

However, if we **fix** k **and edges** $e_1, \ldots, e_k$ on which $q_1, \ldots, q_k$ lie

Small-Problem $\mathrm{SP}(\mathbf{q})$: a simple **convex** objective for k-placements



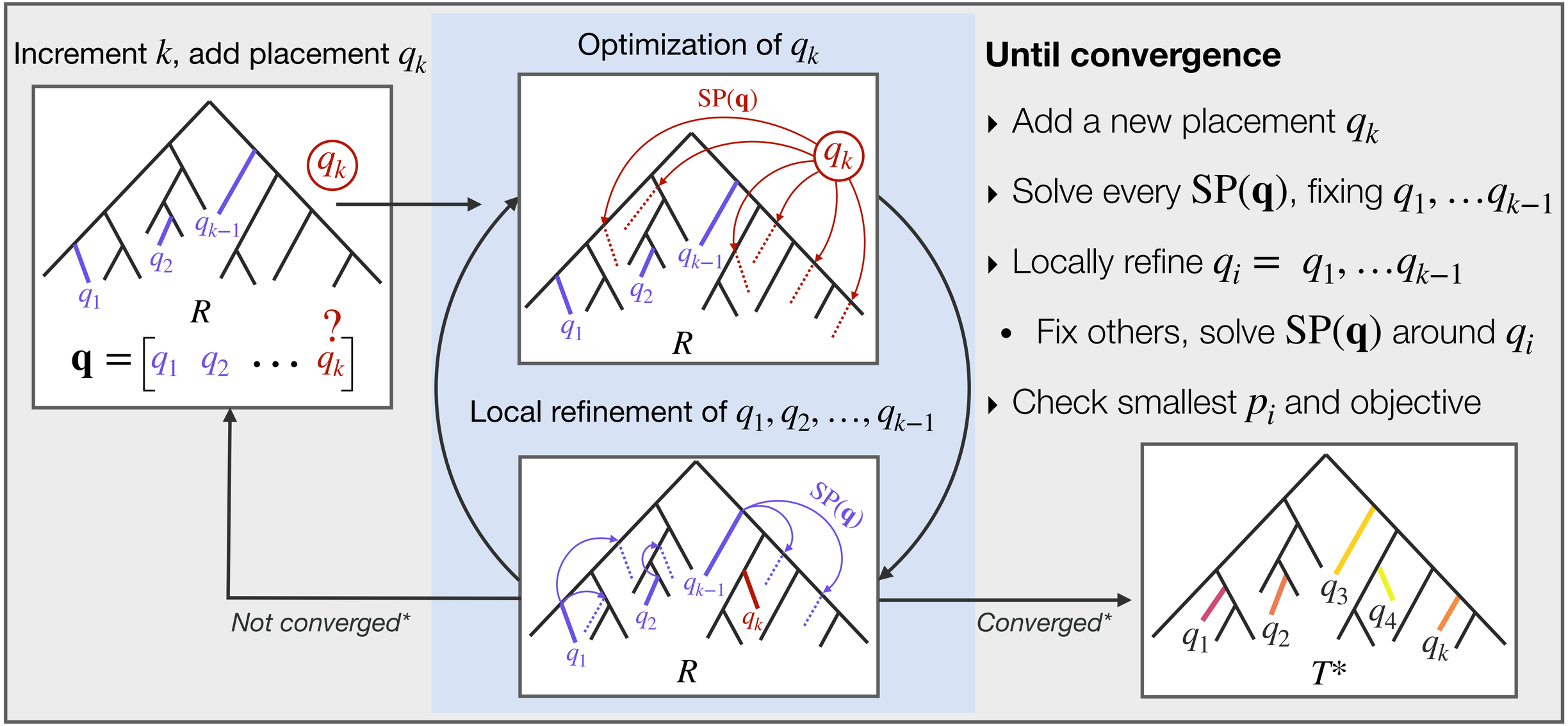
$\mathbf{p}$: abundances of $\mathbf{q}$

x_i: exact placement of q_i on e_i

$\bar{\mathbf{y}}$: avg. terminal branch lengths for $\mathbf{q}$

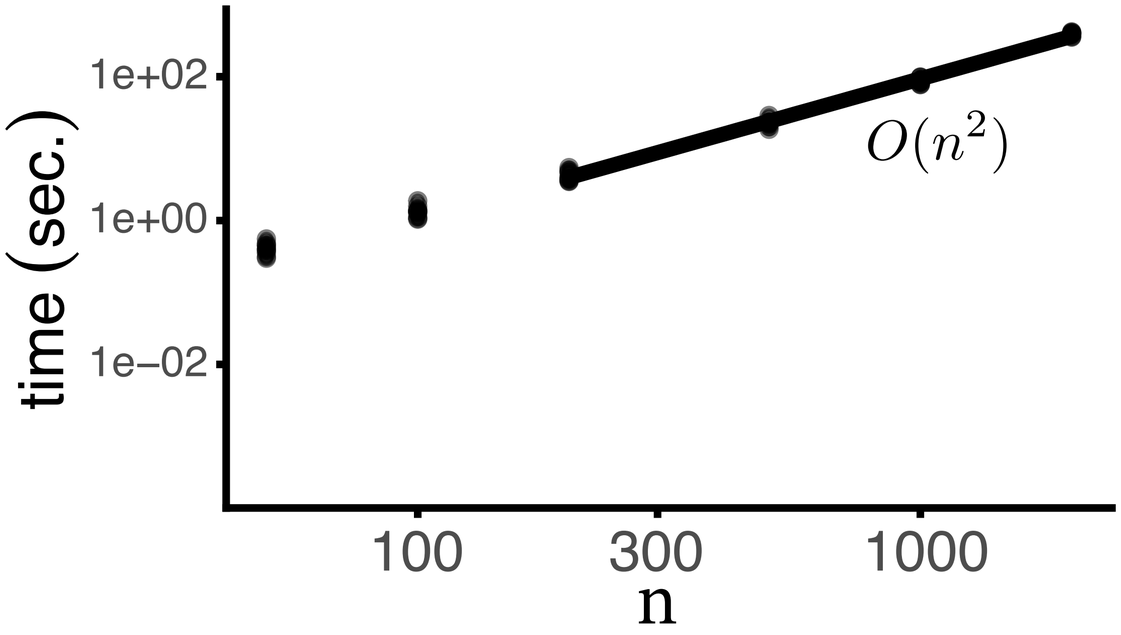
$$\mathrm{argmin}_{\mathbf{p},\mathbf{x},\bar{\mathbf{y}}} \| D(\mathbf{x}, \bar{\mathbf{y}}) \cdot \mathbf{p} - \hat{\mathbf{d}} \|_2$$

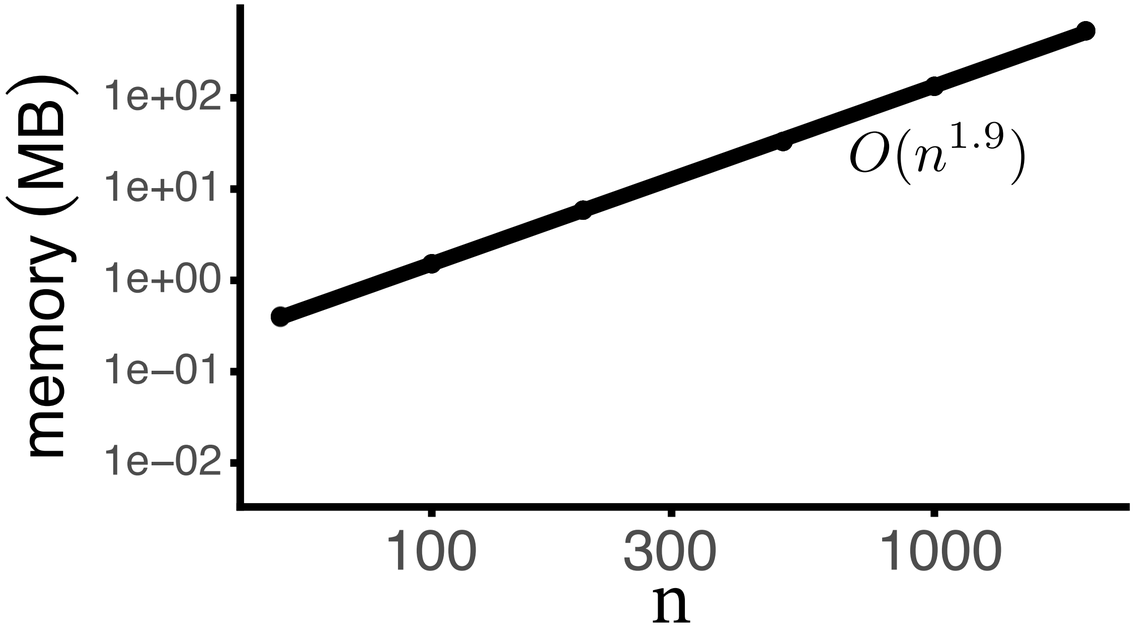
DecoDiPhy: a heuristic search with local refinements



Scalability

Total complexity: $O(nk^2) \approx O(n^2)$ when $k \leq \sqrt{n}$ instead of $O(n^K)$

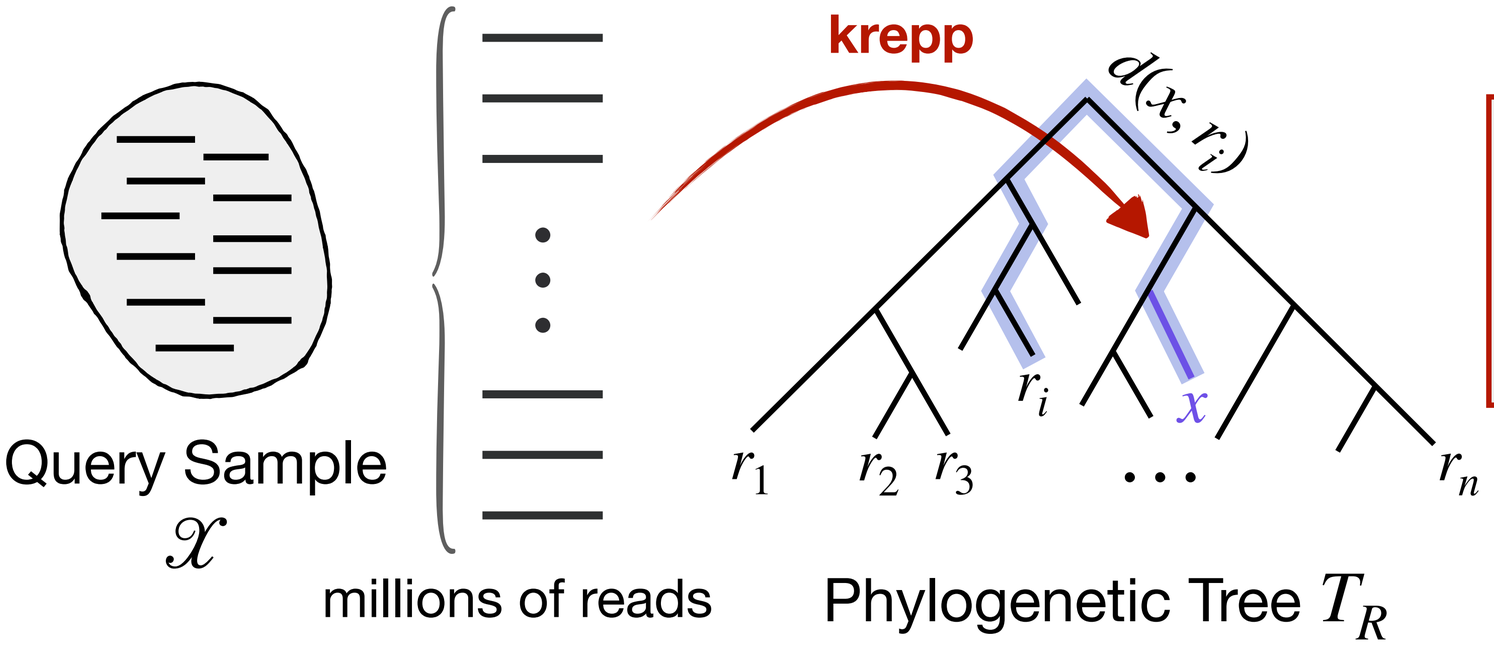




Estimating a mixture distance vector $\hat{\mathbf{d}}$

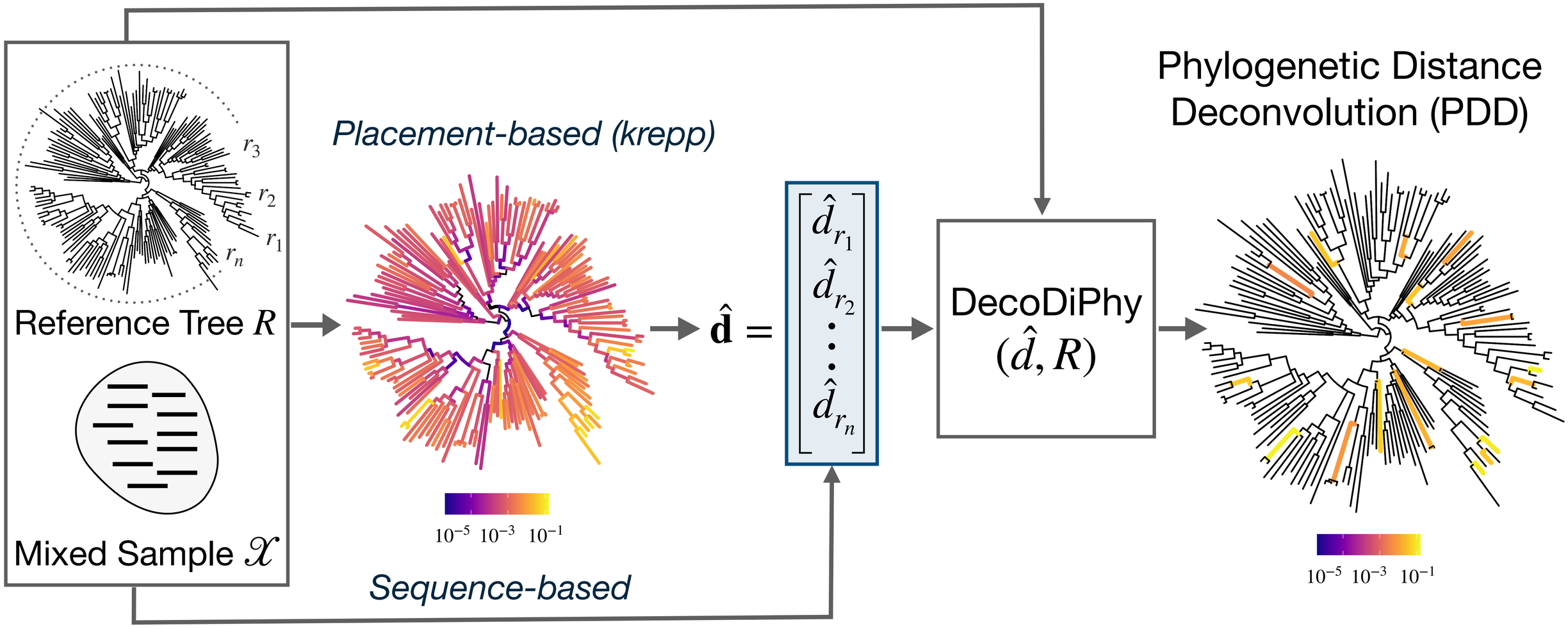
Calculate distance **per read & then average** (sequence or branch length distances)

We place reads using krepp: high placement rate, **distances in branch length units**



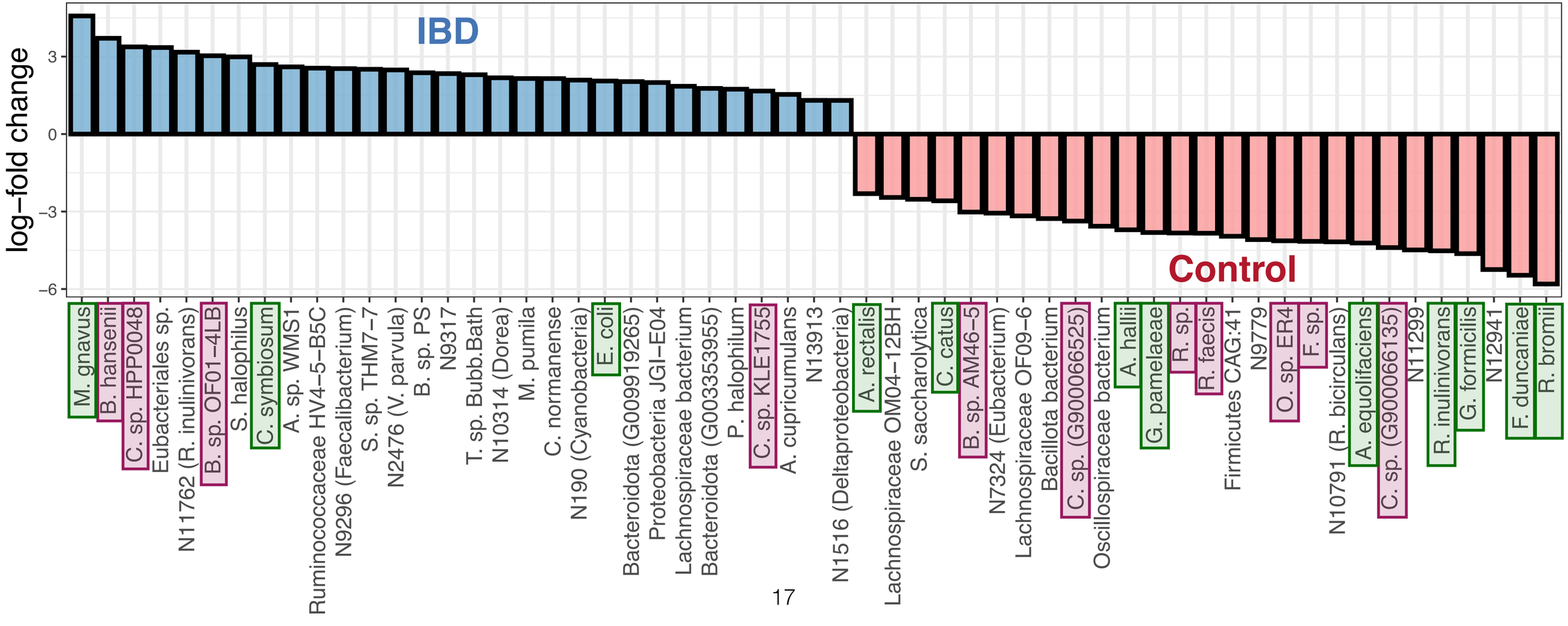
$$\hat{d}(\mathcal{X}, r_i) = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} d(x, r_i)$$

consolidating the signal from many *potentially* FP features
(branches with non-zero abundances)



Differentially abundant features identified by DecoDiPhy

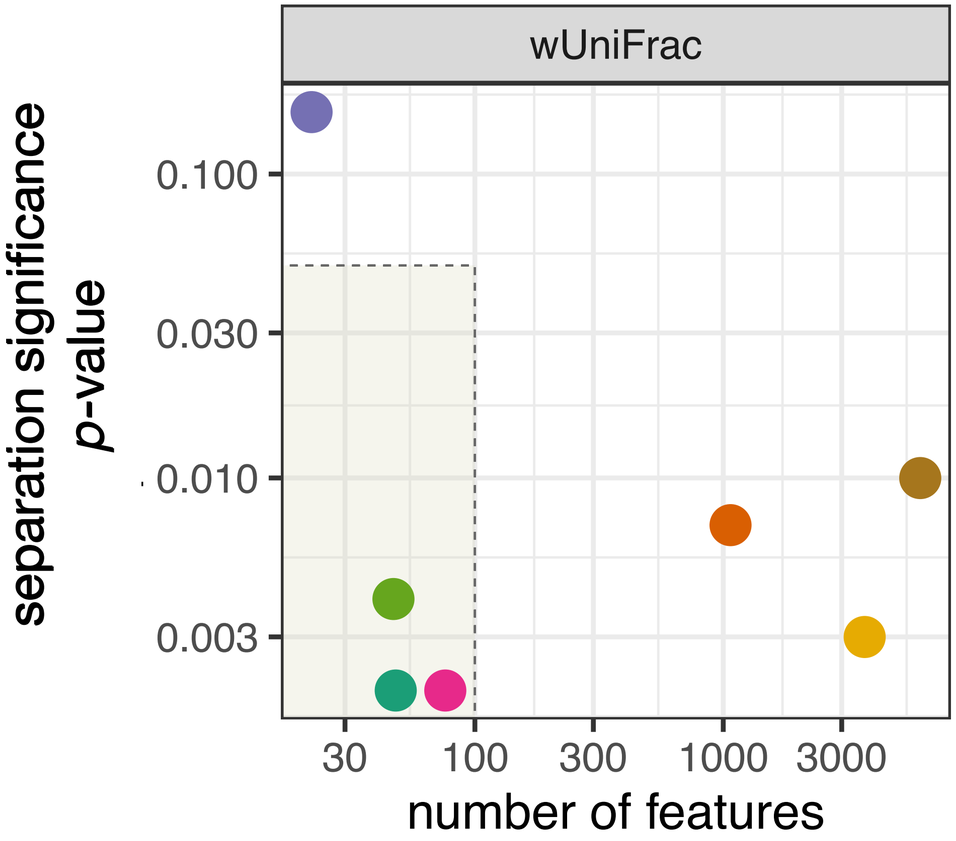
- **>1000 (krepp) vs. 57 (DecoDiPhy)** differentially abundant features (ANCOM-BC test)
- 12 features: **known associations** in the IBD literature: not all detected by other methods!
- 11 features: **not documented but also found by other tools** (sylph and/or sourmash)
- Rest: internal placements and others with weak or no connection



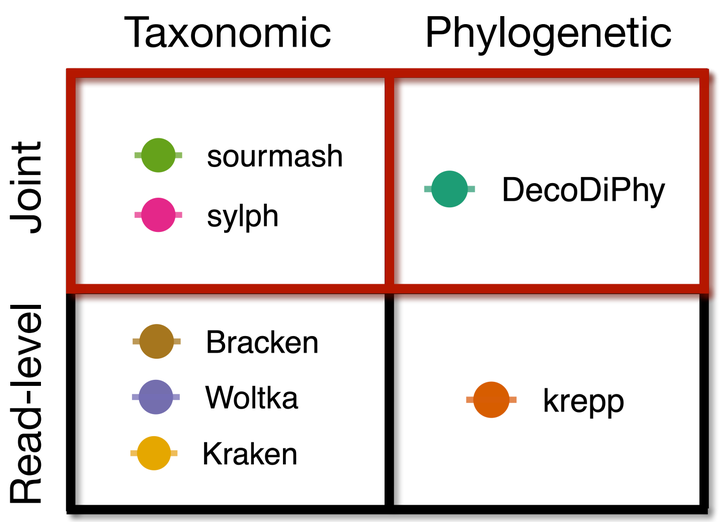
DecoDiPhy features improves IBD vs healthy separation

- Comparing samples using weighted UniFrac distances **(abundances & phylogeny)**
- **Desired:** p-value ≤ 0.05 and a (arbitrarty) interpretable # of features

Franzosa et al. 2019
Human gut samples
(164 IBD & 56 Control)

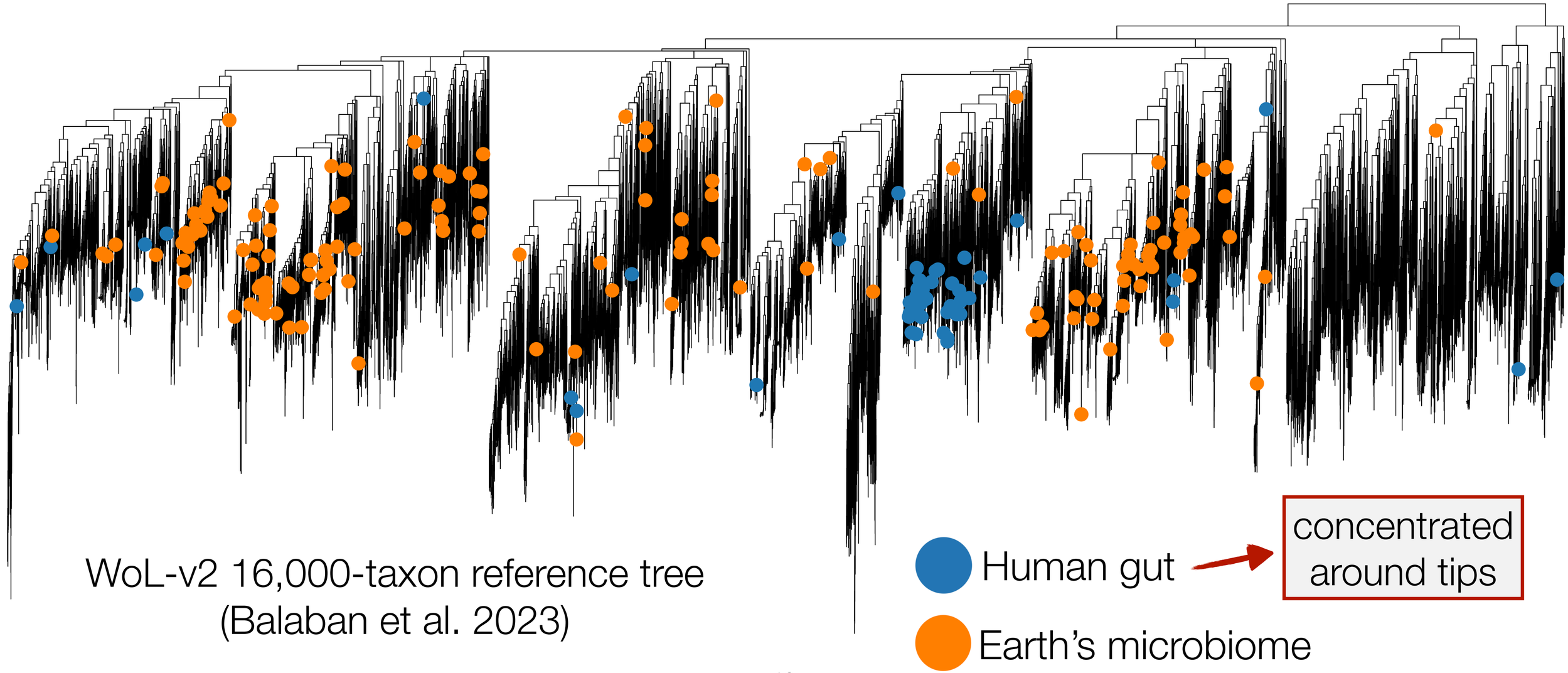


works well!



Impact of novel taxa from less-studied environments

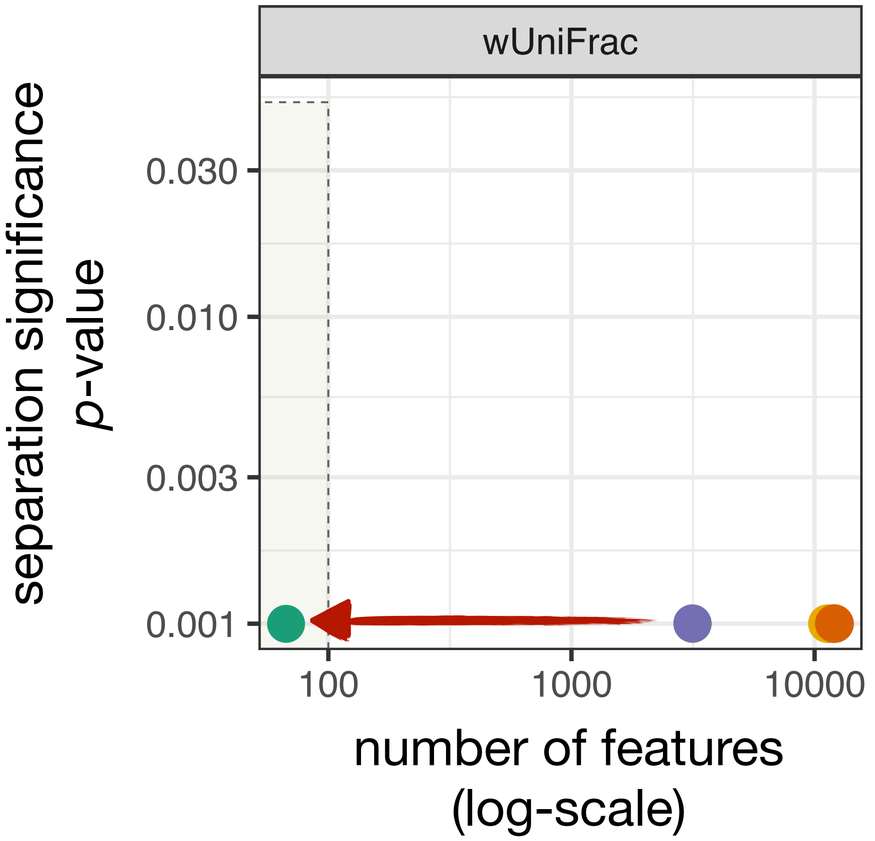
210 samples from the Earth Microbiome Project (Shaffer et al. 2022)

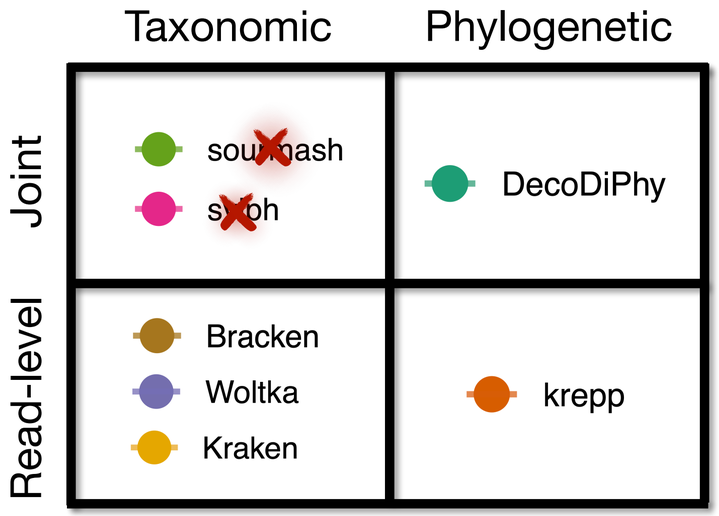


DecoDiPhy performs the best in saline vs non-saline separation

- Comparing saline versus non-saline samples using weighted UniFrac
- Methods based on *k*-mer containment fails to detect taxa for many samples

Shaffer et al. 2022
Earth's microbiome samples
(106 saline, 104 non-saline)

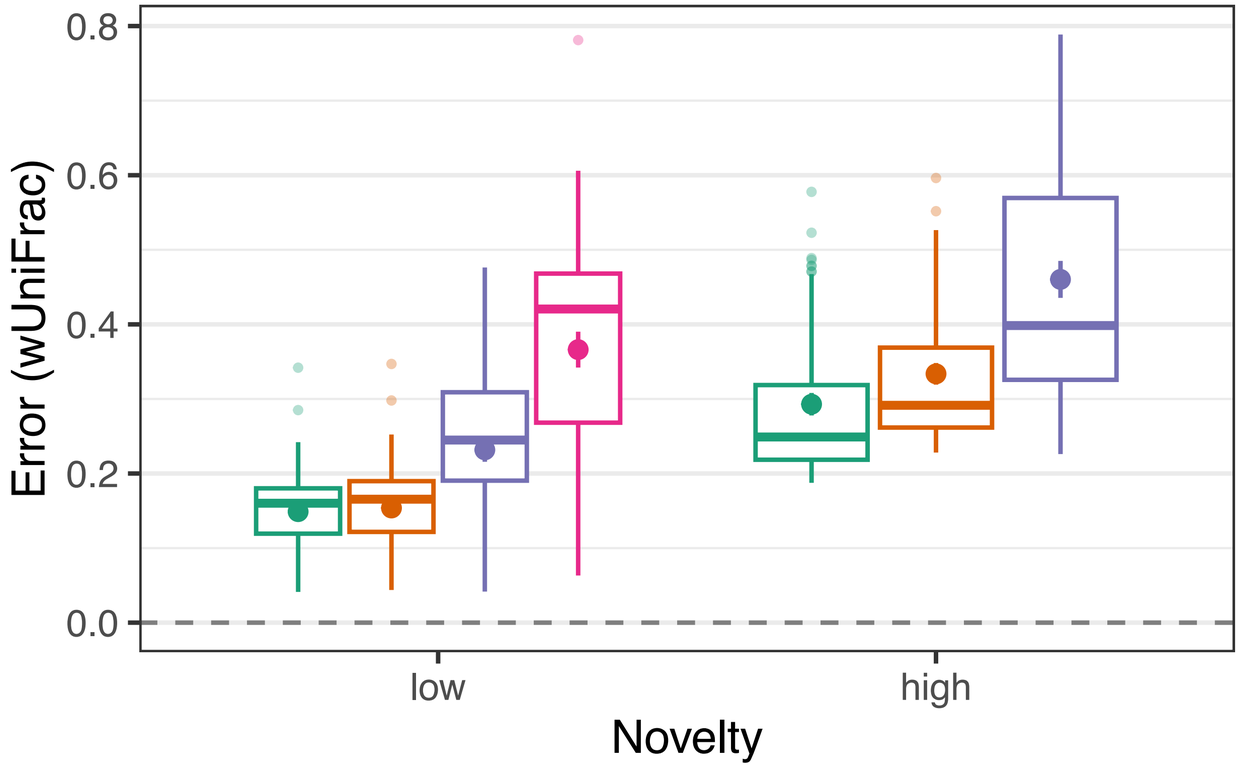


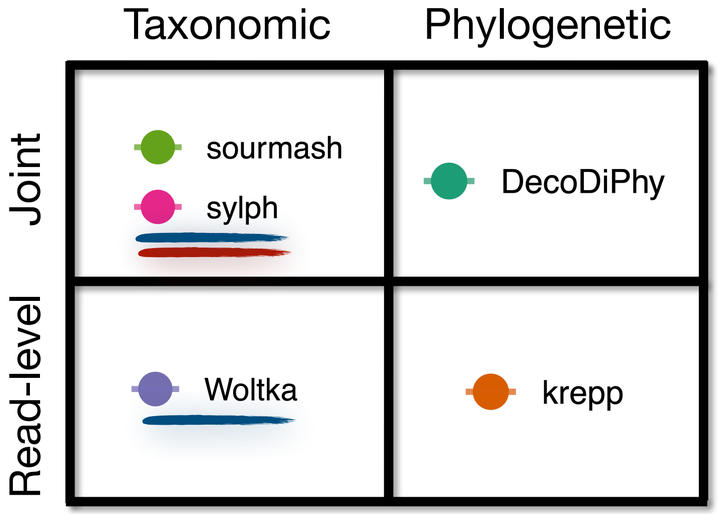


Simulations with ground truth

- Simulations mimicking on microbiome samples from human body sites (Zhu et al. 2022)
- Controlling the query novelty in leave-out experiments

DecoDiPhy outperforms other methods, especially for high novelty





placements only on leaves

does not extend to novel taxa

Summary

DecoDiPhy

- a framework designed to deconvolve phylogenetic distance mixtures
- solves distance-based objective through phylogenetic multi-placement

$$\text{argmax}_{T \in \mathcal{T}(R,k),\ p \in \Delta^{k-1}} \| \hat{d} - D(T) \cdot p \|_2^2$$

- increases strength and interpretability in metagenomic analysis
- extends to novel queries from less-studied enviroments

Thank you!

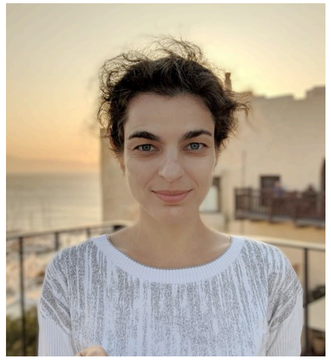
Funding

RECOMB Travel Fellowship
NSF & ISCB

software

preprint

Shayesteh Arasti



Eleonora Rachtman

Mohammed El-Kebir

Siavash Mirarab

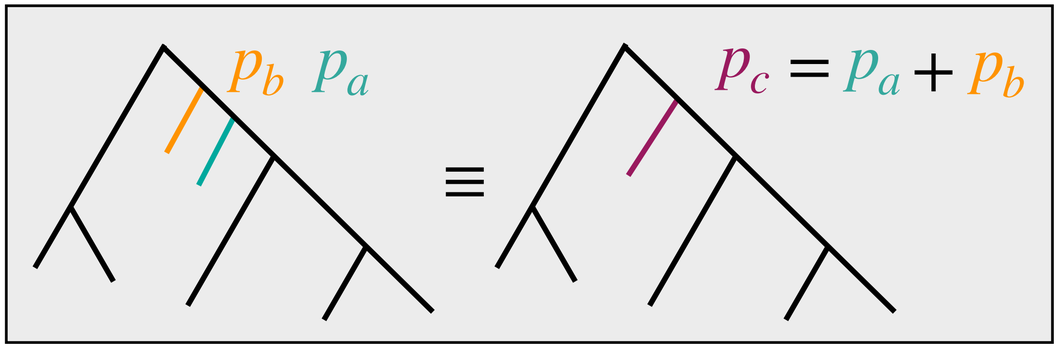
Extra Slides

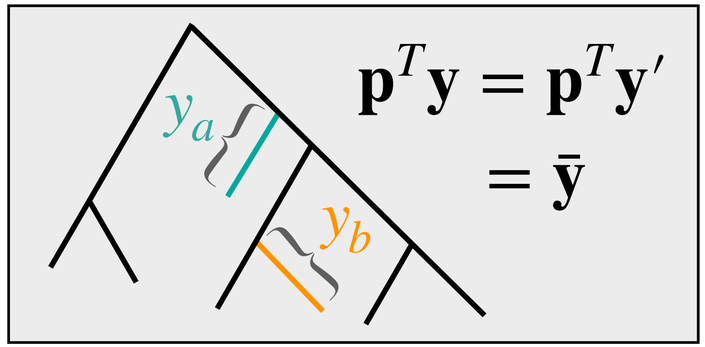
Many questions to be addressed…

- How to obtain a mixture distance from a sample $\mathcal{X}$?
- How many queries (k) are there?

- Is PDD identifiable? If (F, p) satisfies $\mathbf{d} = D(F) \cdot \mathbf{p}$, there is no other $(F', \mathbf{p}')$.

Not generally…





The rest of the unidentifable cases seems to be extreme for realistic samples

Comparison of differentially abundant features across methods