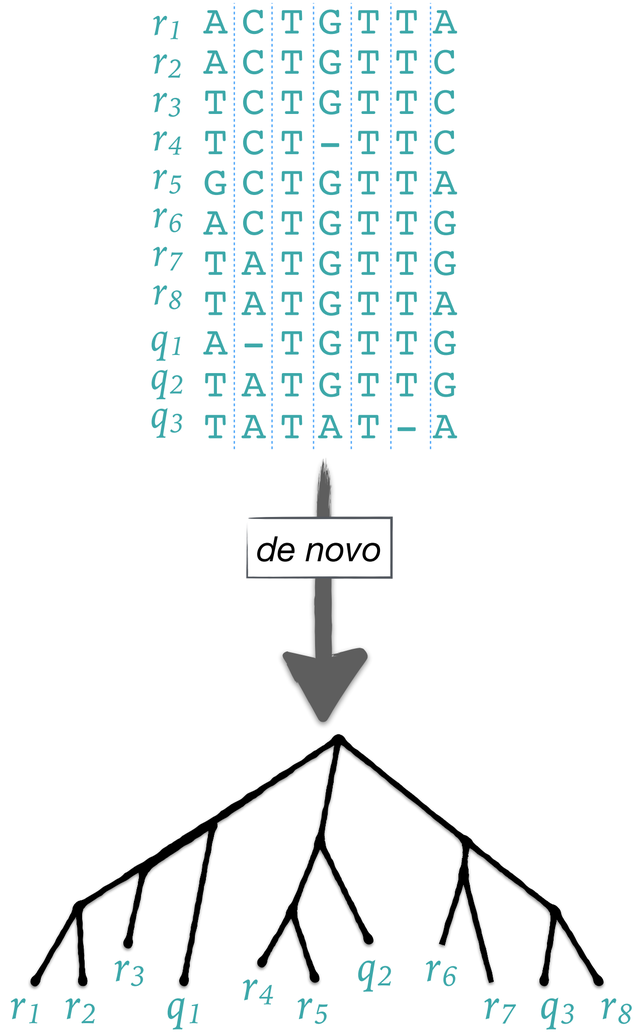
Distance calculation and phylogenetic placement using k-mers

Ali Osman Berk Şapcı & Siavash Mirarab
UC San Diego

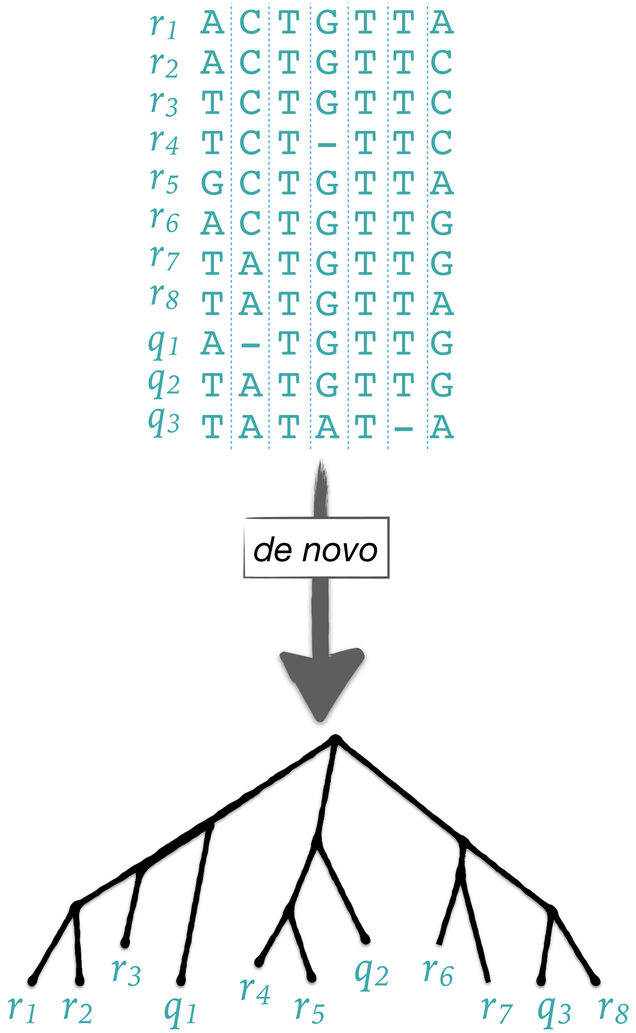
Ali Osman Berk Şapcı

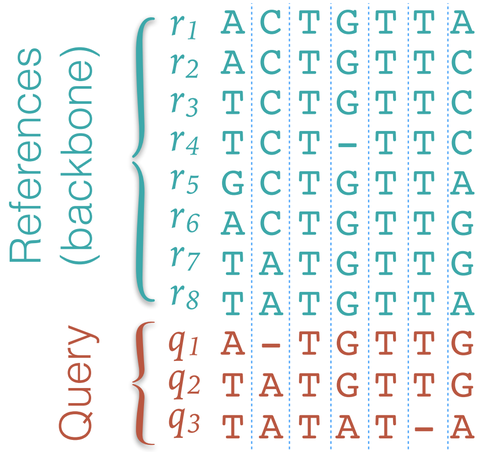
UC San Diego
Electrical and Computer Engineering
JACOBS SCHOOL OF ENGINEERING

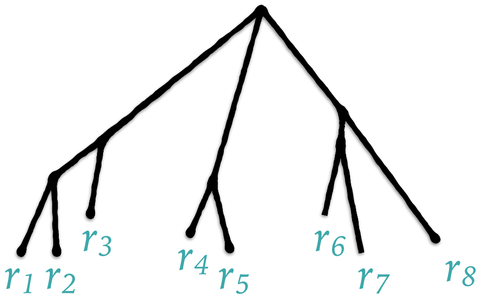
Phylogenetic placement (PP)



Phylogenetic placement (PP)



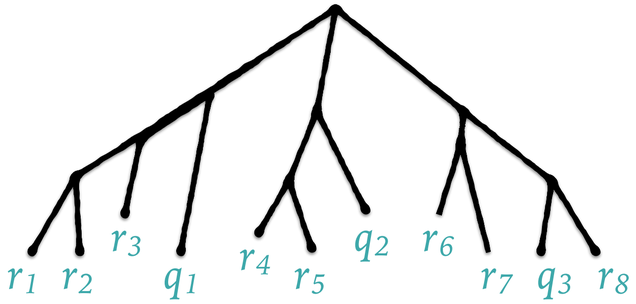




Phylogenetic placement (PP)

r_1	A	C	T	G	T	T	A
r_2	A	C	T	G	T	T	C
r_3	T	C	T	G	T	T	C
r_4	T	C	T	–	T	T	C
r_5	G	C	T	G	T	T	A
r_6	A	C	T	G	T	T	G
r_7	T	A	T	G	T	T	G
r_8	T	A	T	G	T	T	A
q_1	A	–	T	G	T	T	G
q_2	T	A	T	G	T	T	G
q_3	T	A	T	A	T	–	A

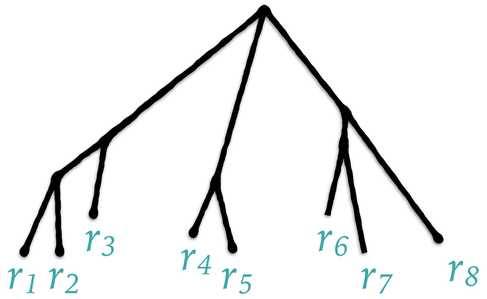
de novo



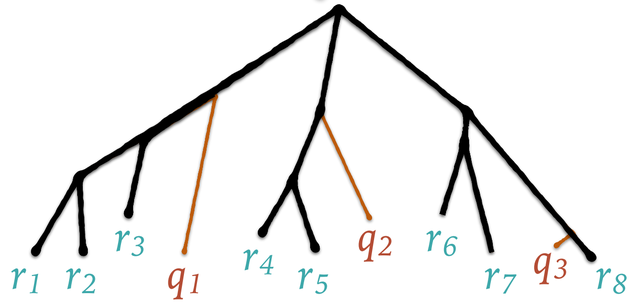
References (backbone)

Query

r_1	A	C	T	G	T	T	A
r_2	A	C	T	G	T	T	C
r_3	T	C	T	G	T	T	C
r_4	T	C	T	–	T	T	C
r_5	G	C	T	G	T	T	A
r_6	A	C	T	G	T	T	G
r_7	T	A	T	G	T	T	G
r_8	T	A	T	G	T	T	A
q_1	A	–	T	G	T	T	G
q_2	T	A	T	G	T	T	G
q_3	T	A	T	A	T	–	A



Placement



Benefits of Placement

- **Scalability**:
  - The backbone tree is fixed
  - Queries are independent
    - Linearly scales with more queries
    - Embarrassingly parallel
- **Error tolerance**: high-quality backbone sequences help situate short low-quality queries
  [Janssen et al, 2018, msystems]

Shortcoming

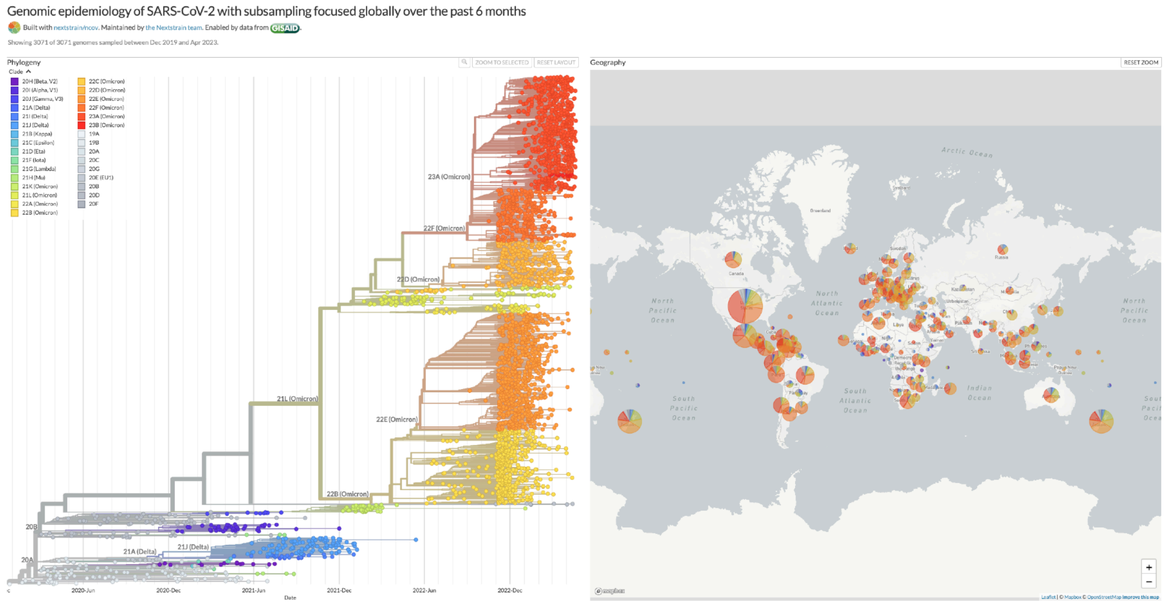
- Queries cannot refine backbone relationships
- Relationships between queries are not inferred

PP Methods

- Traditionally, for greedy *de novo* inference
  [Felsenstein, JME, 1981; Desper and Gascuel, JCB, 2002]
- Designed explicitly for PP:
  - Maximum likelihood — PPlacer [Matsen, et al., BMC Bioinformatics, 2010]
    — EPA(-ng) [Berger, et al., Sys Bio, 2011][Barbera, et al., Sys Bio, 2019]
  - Divide-and-conquer — SEPP [Mirarab, et al., PBC, 2013]
    — (B)SCAMPP [Wedell, et al., TCBB, 2023]
  - Parsimony — UShER [Turakhia, 2021]
  - Distance-based (least squares error) — APPLES(-II) [Balaban & Mirarab, Sys Bio, 2020]
    [Balaban, et al, 2022]

Application: molecular epidemic tracking

- Tracking progression of pandemics



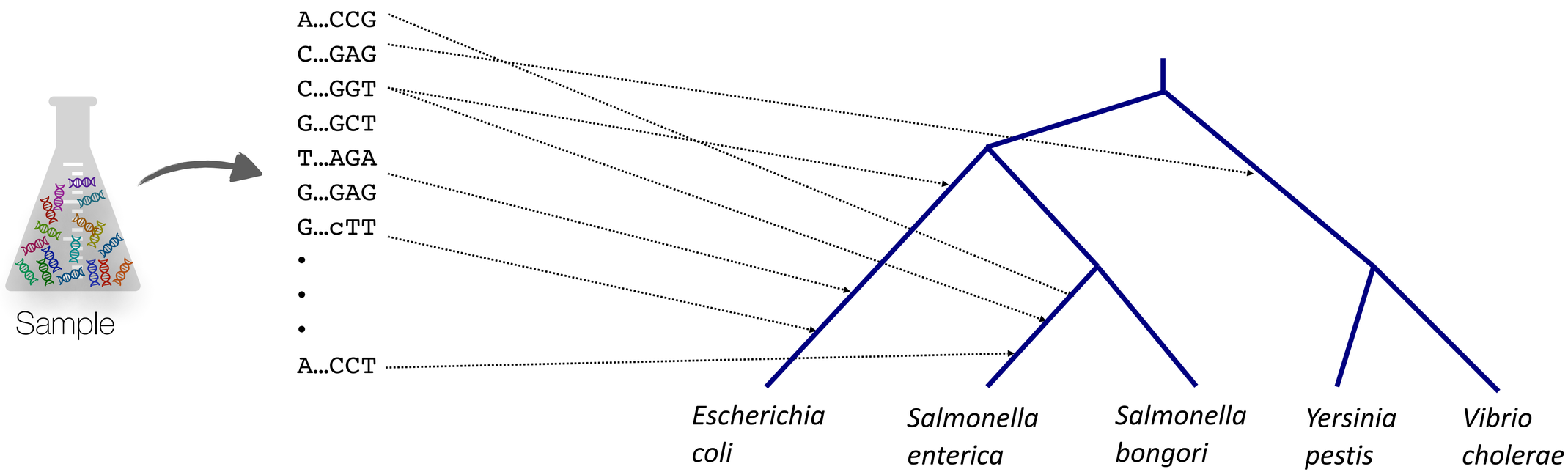
Article | Published: 10 May 2021

Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic

Yatish Turakhia, Bryan Thornlow, Angie S. Hinrichs, Nicola De Maio, Landen Gozashti, Robert Lanfear, David Haussler & Russell Corbett-Detig

Nature Genetics **53**, 809–816 (2021) | Cite this article

Application: microbiome identification

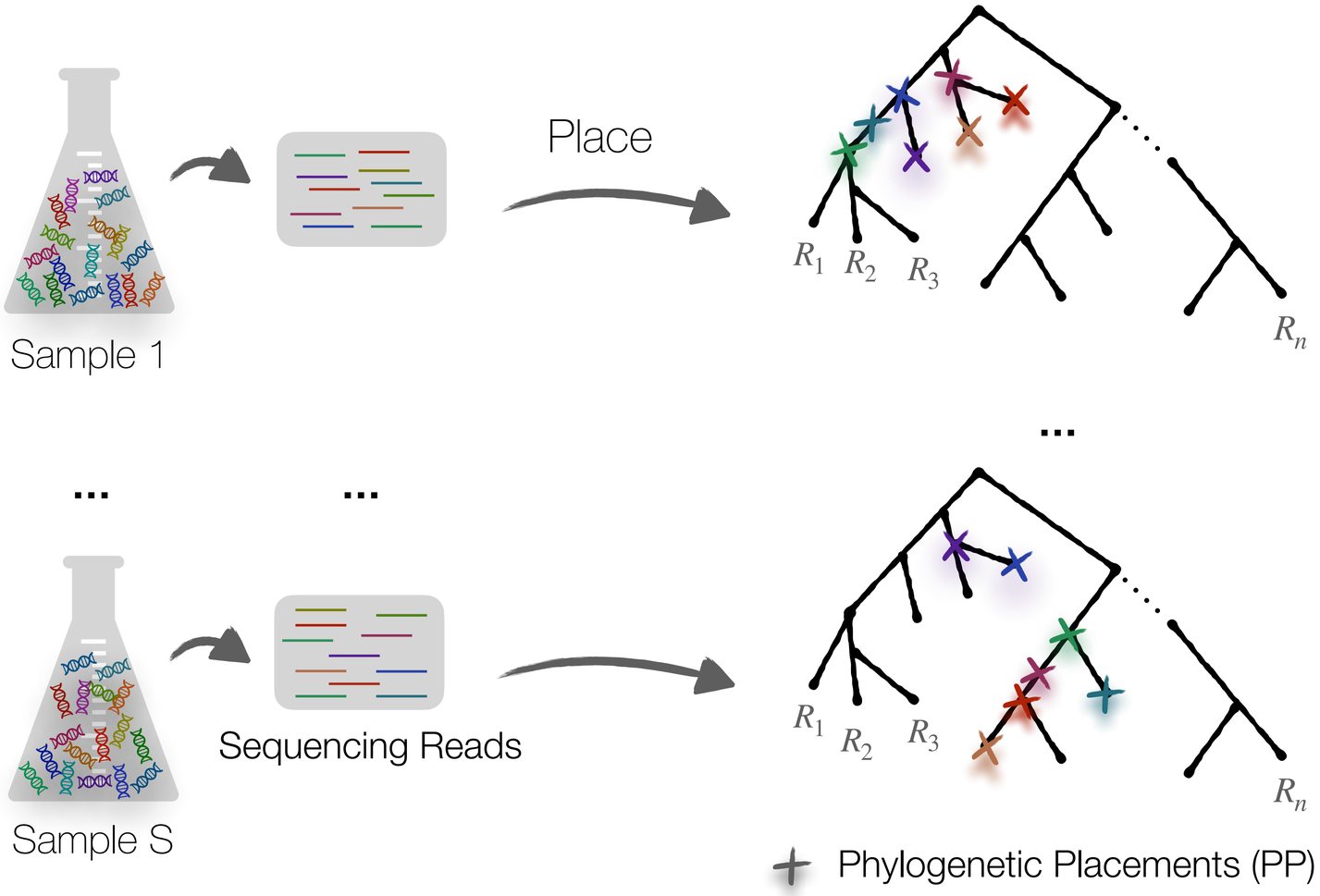


Unknown reads
metagenomic or amplicon

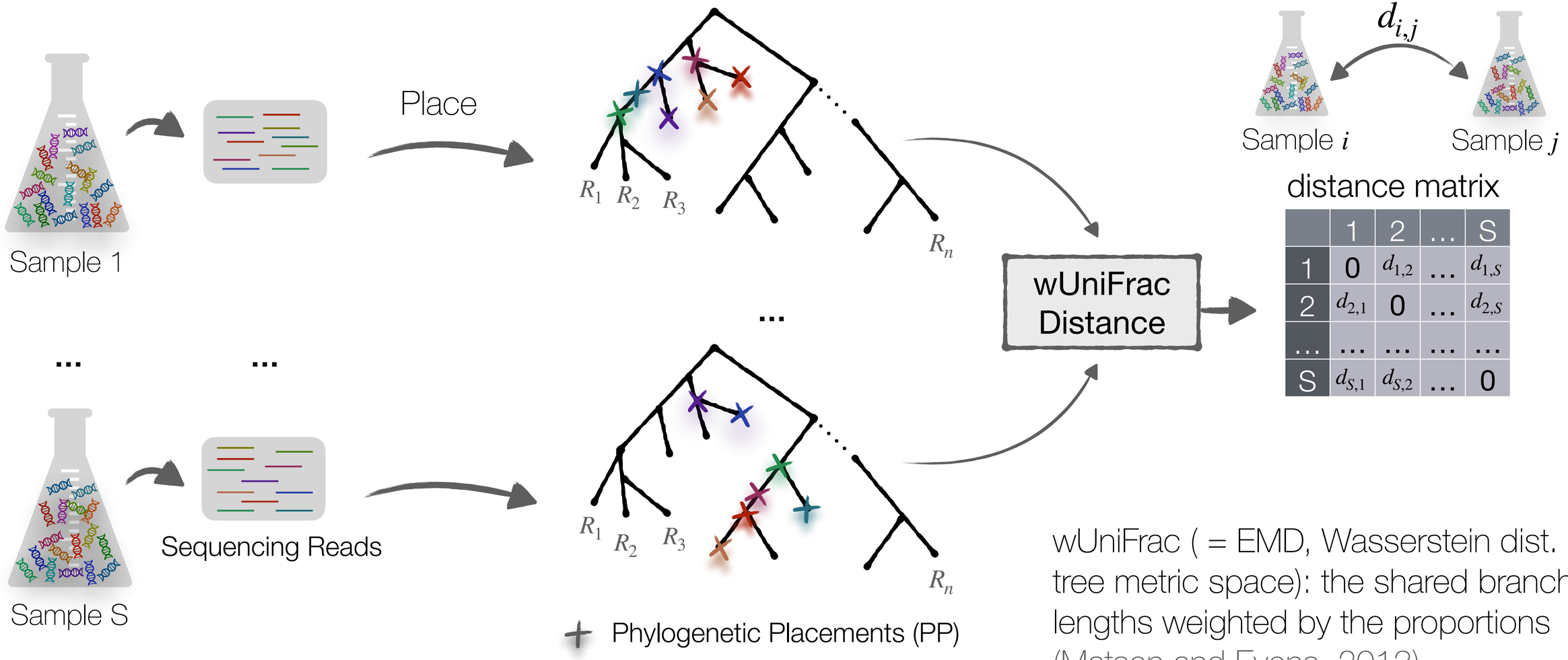
A **reference dataset** of known sequences
with an alignment and a tree

Place each read on a *reference tree* of known sequences to identify them

Comparing metagenomic samples using phylogenetic placement



Comparing metagenomic samples using phylogenetic placement



	1	2	...	S
1	0	$d_{1,2}$	...	$d_{1,S}$
2	$d_{2,1}$	0	...	$d_{2,S}$
...	...	...	...	...
S	$d_{S,1}$	$d_{S,2}$	...	0

wUniFrac (= EMD, Wasserstein dist. on tree metric space): the shared branch lengths weighted by the proportions (Matsen and Evens, 2013)

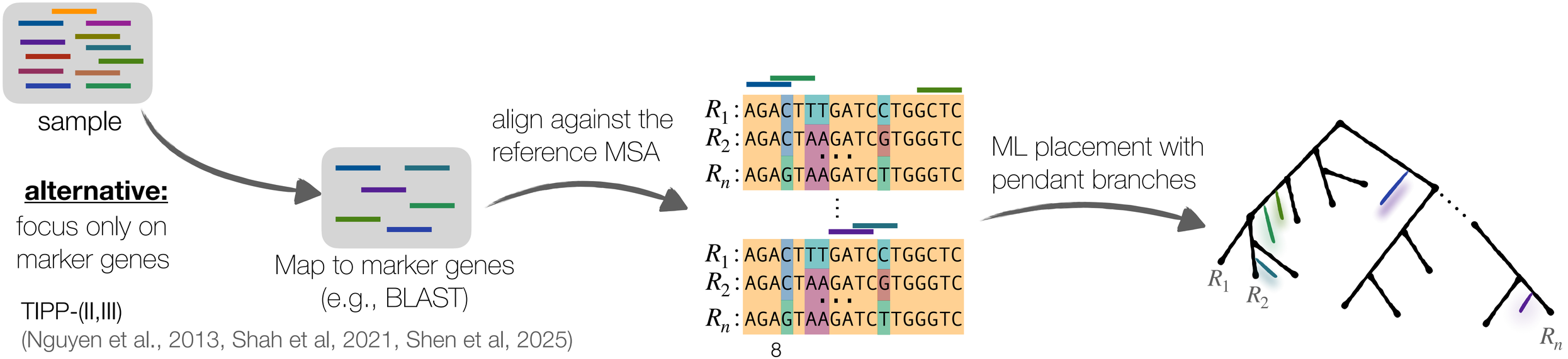
How well does phylogenetic placement of **all reads** from a sample on a (species) tree labelled with reference genomes work?

How well does phylogenetic placement of **all reads** from a sample on a (species) tree labelled with reference genomes work?

No (scalable) existing method was designed for this

How well does phylogenetic placement of **all reads** from a sample on a (species) tree labelled with reference genomes work?

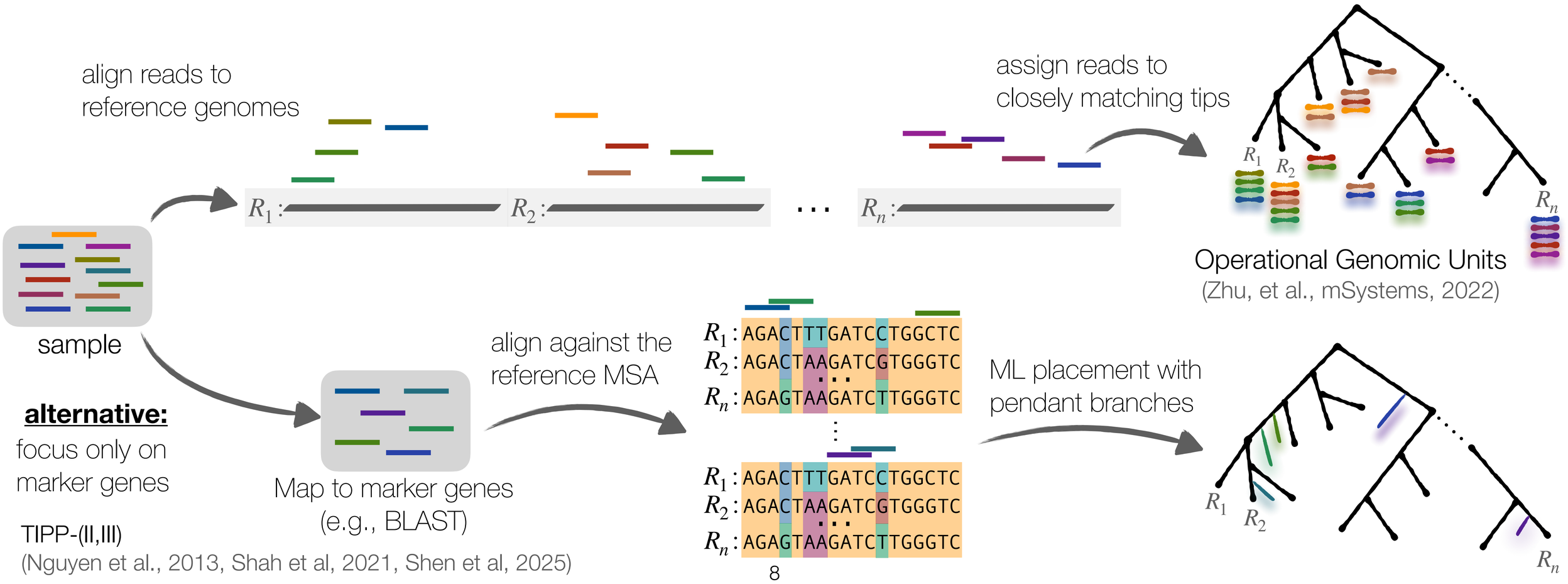
No (scalable) existing method was designed for this



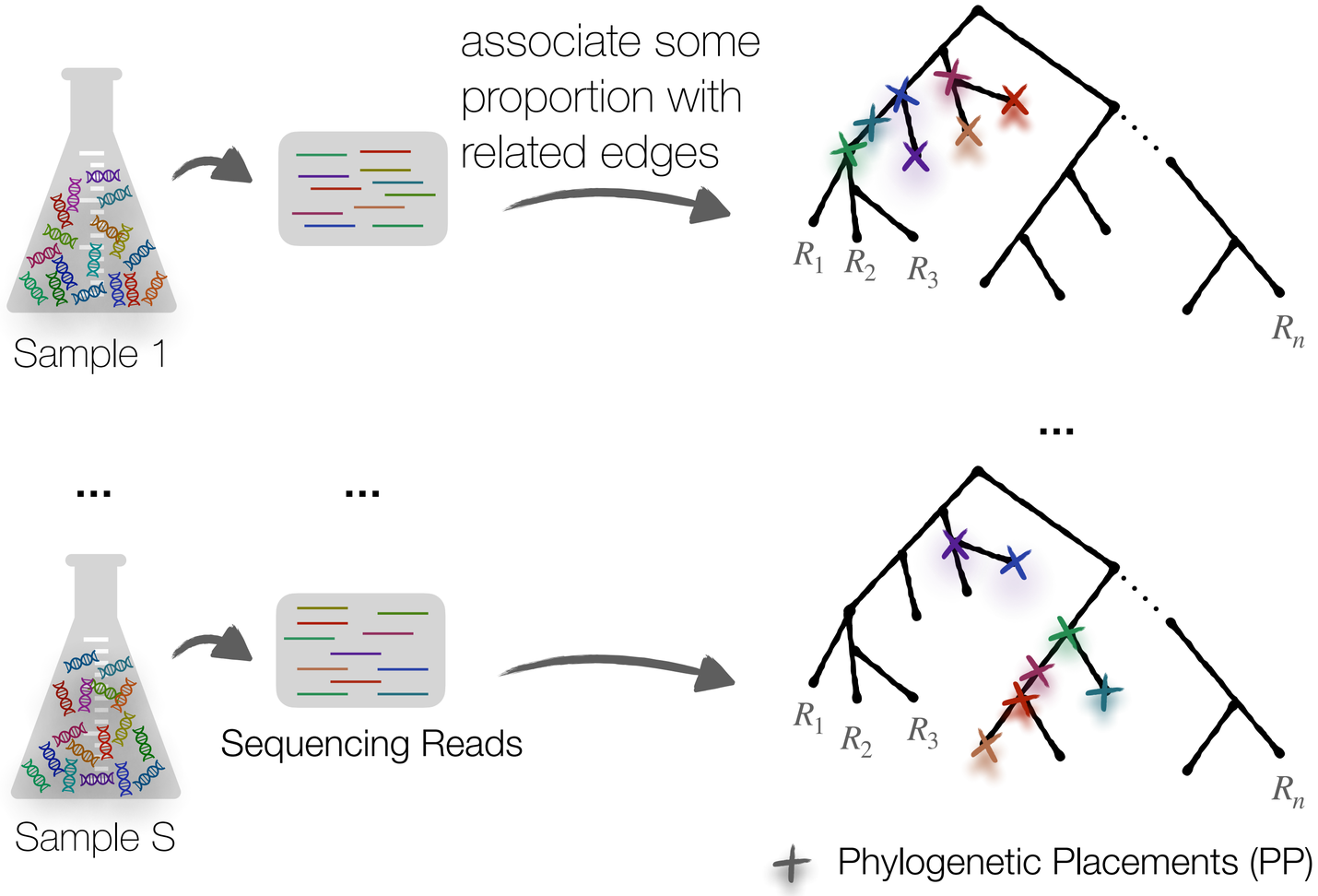
TIPP-(II,III)
(Nguyen et al., 2013, Shah et al, 2021, Shen et al, 2025)

How well does phylogenetic placement of **all reads** from a sample on a (species) tree labelled with reference genomes work?

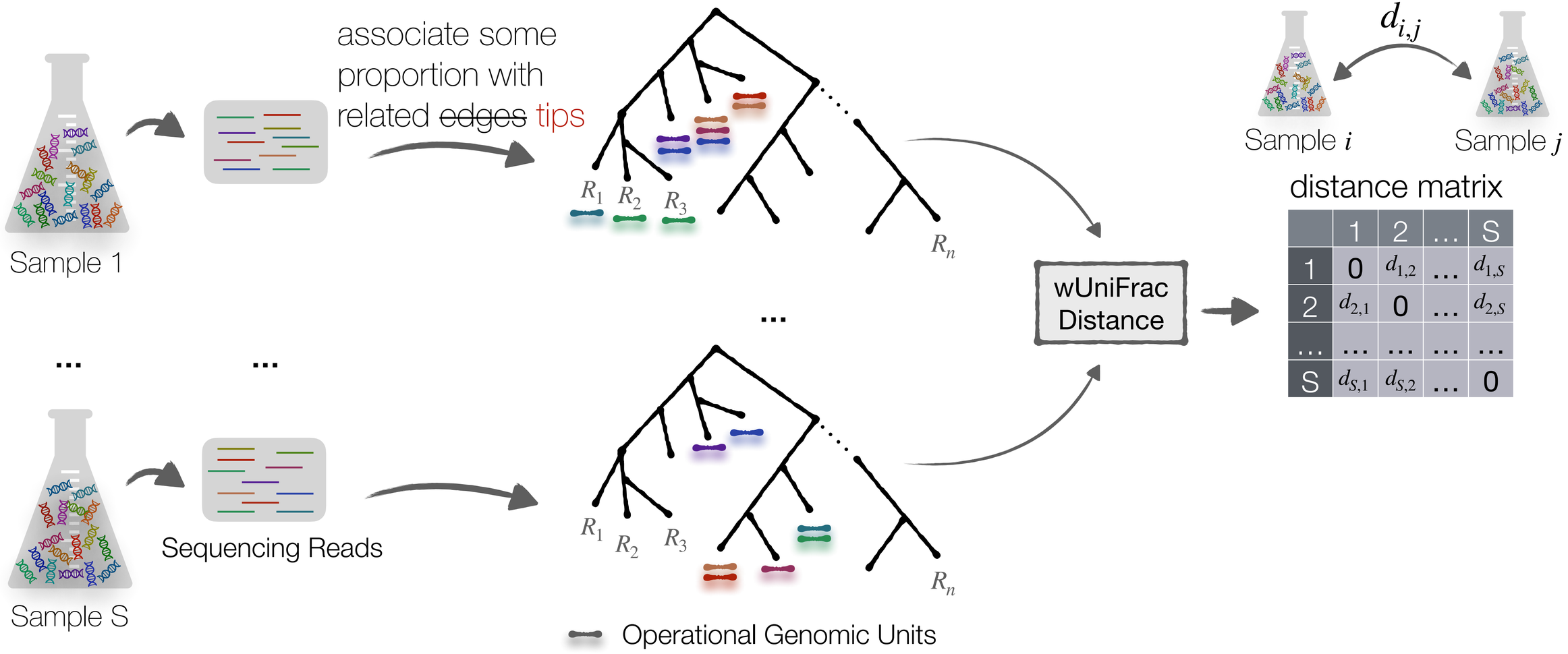
No (scalable) existing method was designed for this



Characterization of metagenomic samples on a phylogeny

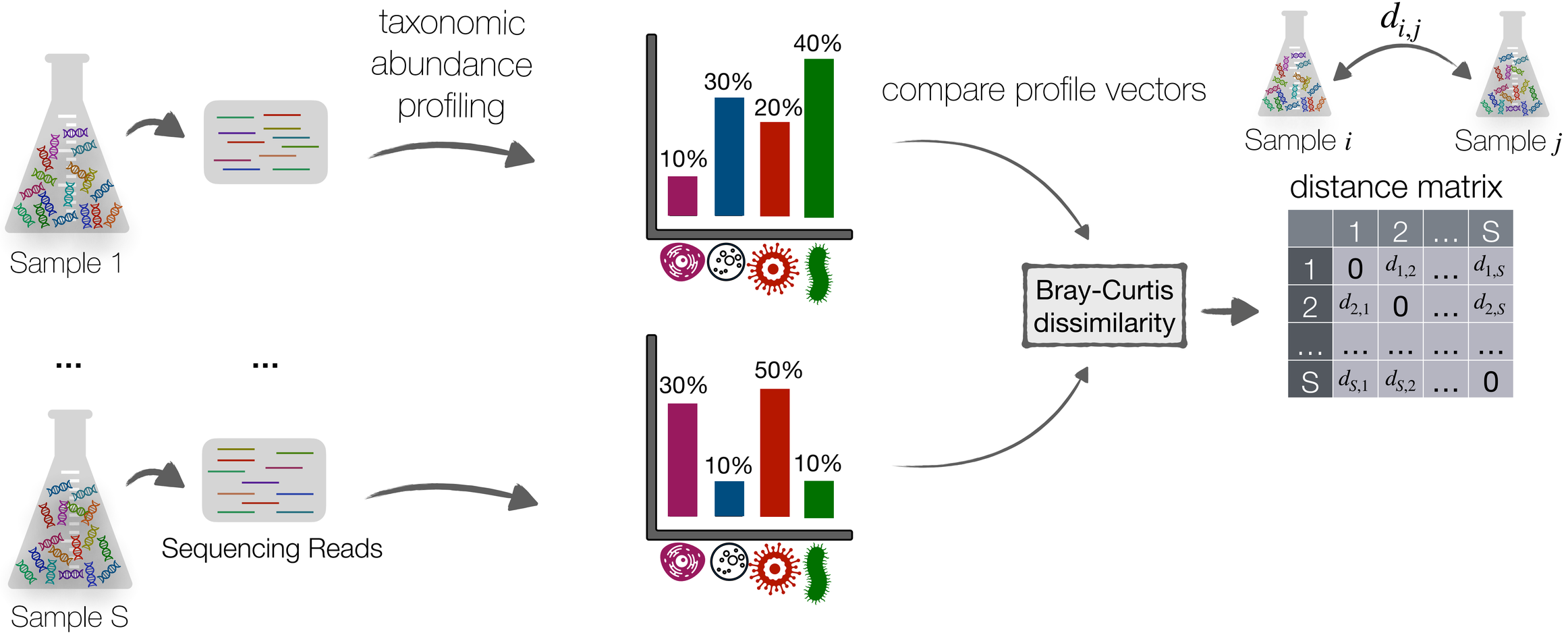


Characterization of metagenomic samples on a phylogeny

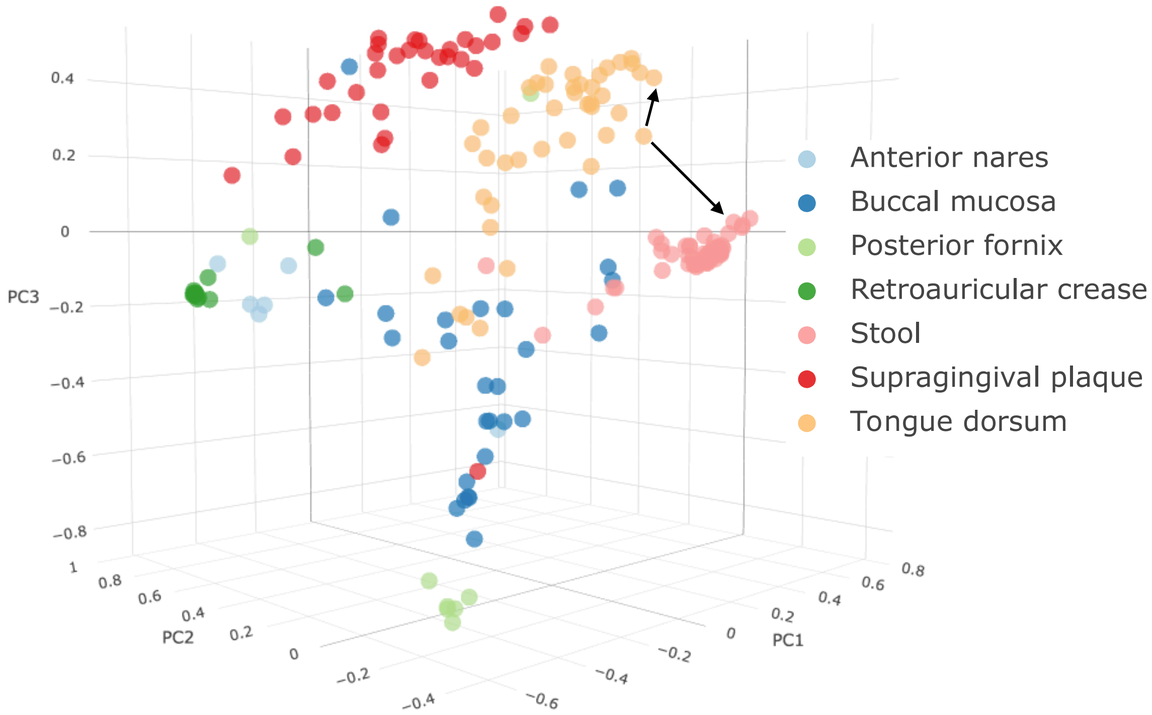


	1	2	...	S
1	0	$d_{1,2}$	...	$d_{1,S}$
2	$d_{2,1}$	0	...	$d_{2,S}$
...	...	...	...	...
S	$d_{S,1}$	$d_{S,2}$	...	0

Characterization of metagenomic samples on a phylogeny



Analyzing human microbiome with larger references



Phylogeny-Aware Analysis of Metagenome Community Ecology Based on Matched Reference Genomes while Bypassing Taxonomy

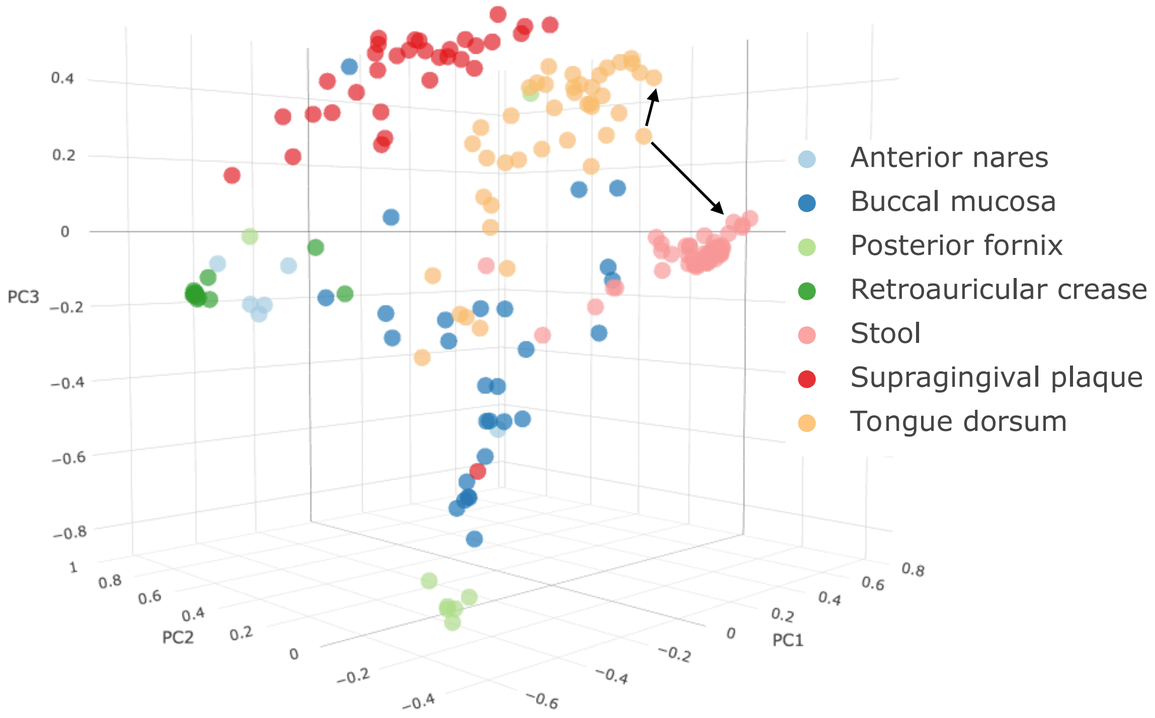
This article has been corrected. VIEW CORRECTION

Authors: Qiyun Zhu, Shi Huang, Antonio Gonzalez, Imran McGrath, Daniel McDonald, Niina Haiminen, George Armstrong, Yoshiki Vázquez-Baeza, Julian Yu, Justin Kuczynski, Gregory D. Sepich-Poore, Austin D. Swafford, Promi Das, Justin P. Shaffer, Franck Lejzerowicz, Pedro Belda-Ferre, Aki S. Havulinna, Guillaume Méric, Teemu Niiranen, Leo Lahti, Veikko Salomaa, Ho-Cheol Kim, Mohit Jain, Michael Inouye, Jack A. Gilbert, Rob Knight SHOW FEWER | AUTHORS INFO & AFFILIATIONS

Evaluation:

- samples often have categorical labels
- **pseudo *F* statistic**: compare within-group versus across-group distances

Analyzing human microbiome with larger references

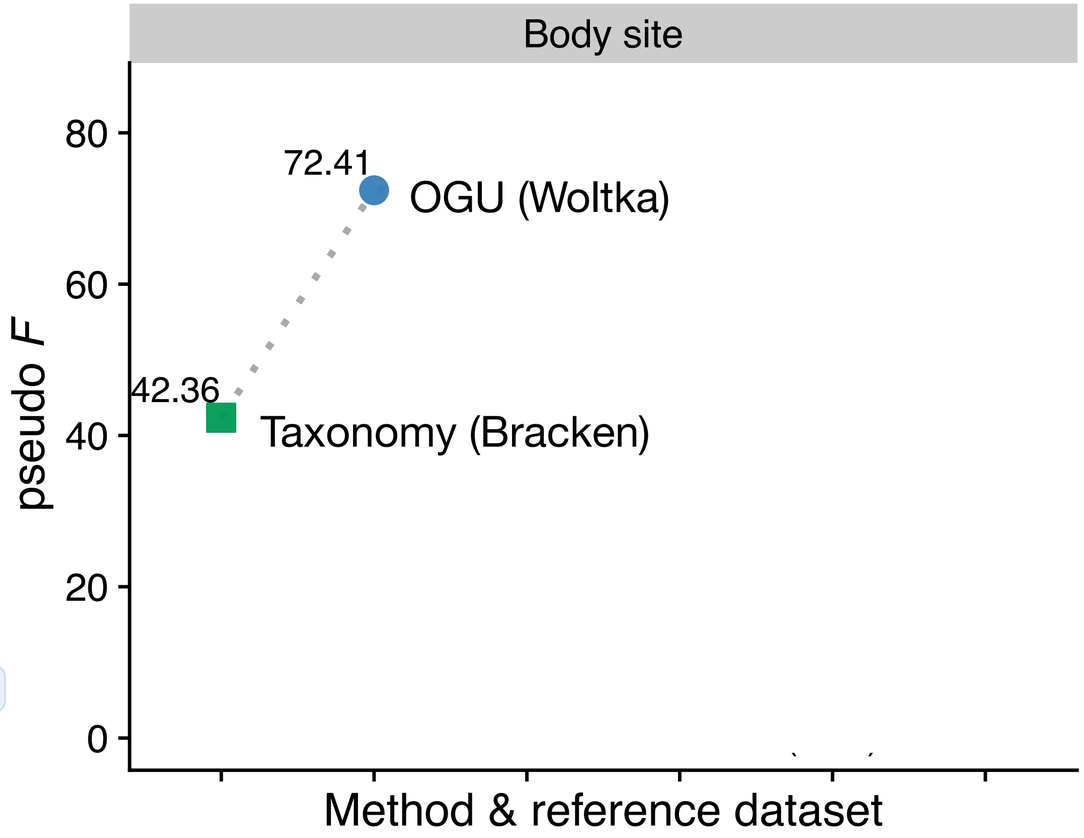


Phylogeny-Aware Analysis of Metagenome Community Ecology Based on Matched Reference Genomes while Bypassing Taxonomy

This article has been corrected. VIEW CORRECTION

Authors: Qiyun Zhu, Shi Huang, Antonio Gonzalez, Imran McGrath, Daniel McDonald, Niina Haiminen, George Armstrong, Yoshiki Vázquez-Baeza, Julian Yu, Justin Kuczynski, Gregory D. Sepich-Poore, Austin D. Swafford, Promi Das, Justin P. Shaffer, Franck Lejzerowicz, Pedro Belda-Ferre, Aki S. Havulinna, Guillaume Méric, Teemu Niiranen, Leo Lahti, Veikko Salomaa, Ho-Cheol Kim, Mohit Jain, Michael Inouye, Jack A. Gilbert, Rob Knight SHOW FEWER | AUTHORS INFO & AFFILIATIONS

Reference: Web of Life (v2)
16,000 microbial genomes



Perhaps the OGU approach
(alignment + assigning to all mapped tree tips)
is good enough?

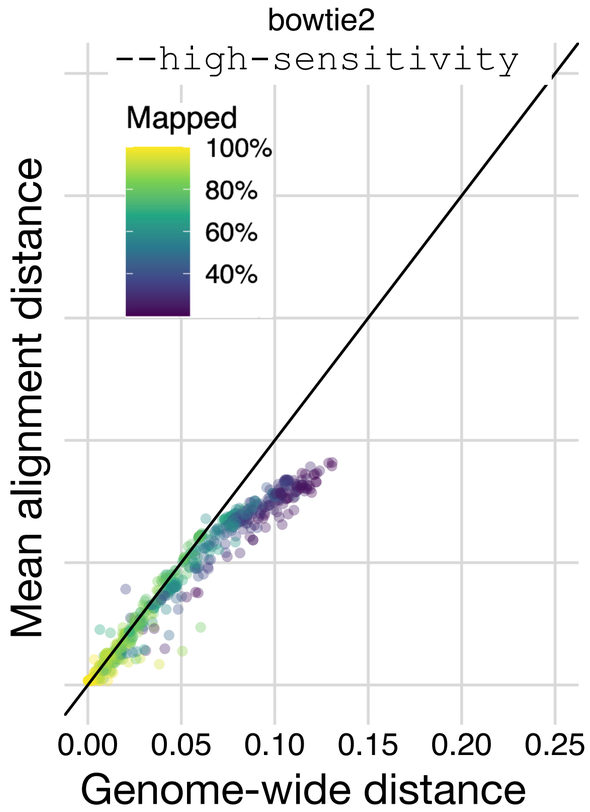
OGU: Challenges of aligning reads

align reads to reference genomes

(-) not scalable for large N, even with efficient indexes

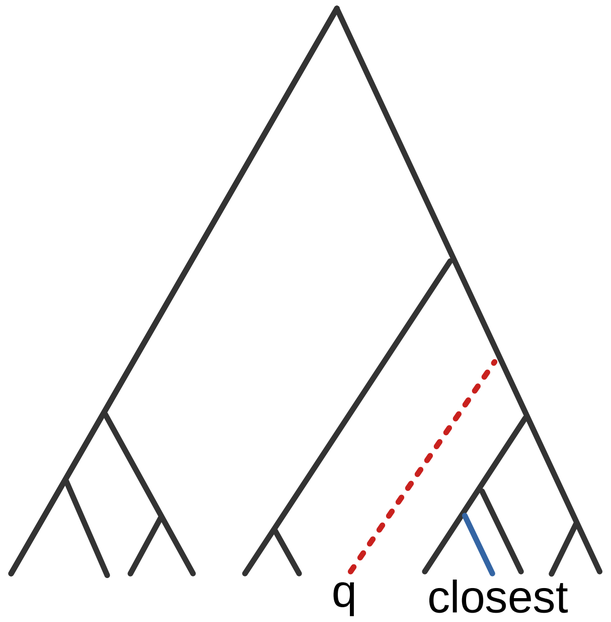
(-) not suitable for higher distances & novel sequences (>10%)

— Increasing the sensitivity by relaxing the alignment is costly



Closest match is not enough

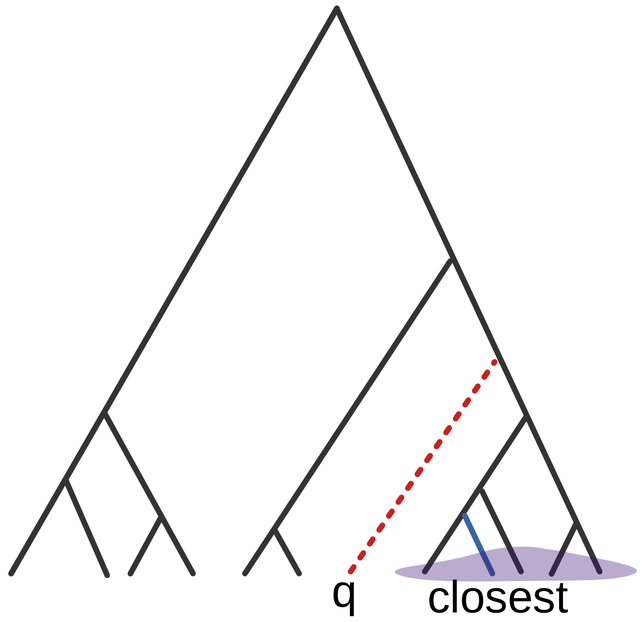
- Closest match may mislead
  - Assigning to multiple tips may not be sufficient.



[Balaban, et al, 2022]

Closest match is not enough

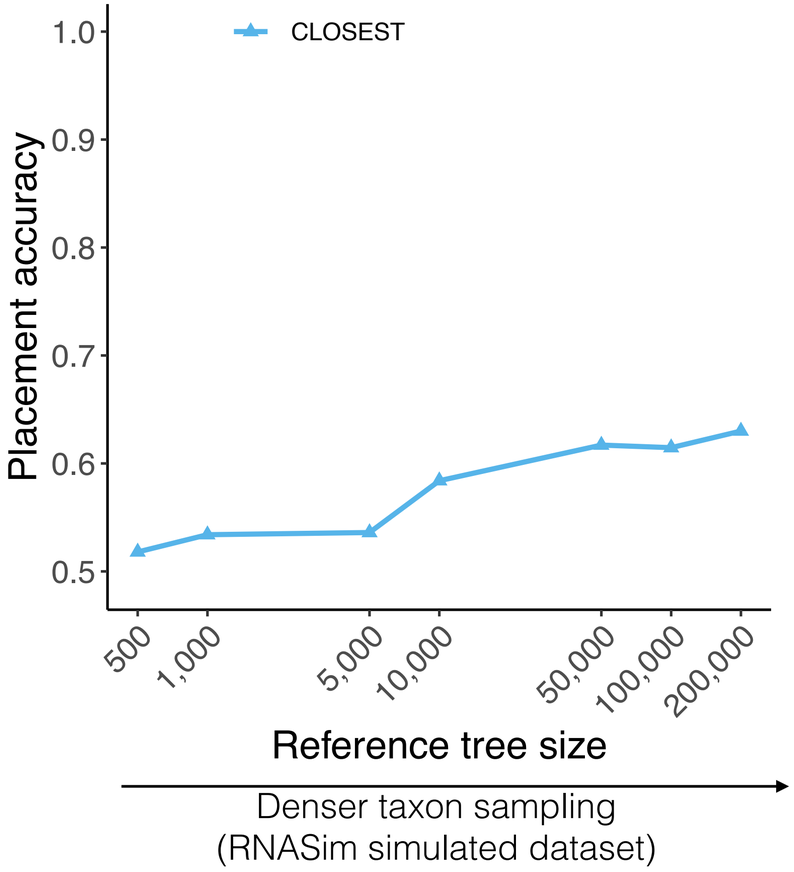
- Closest match may mislead
  - Assigning to multiple tips may not be sufficient.

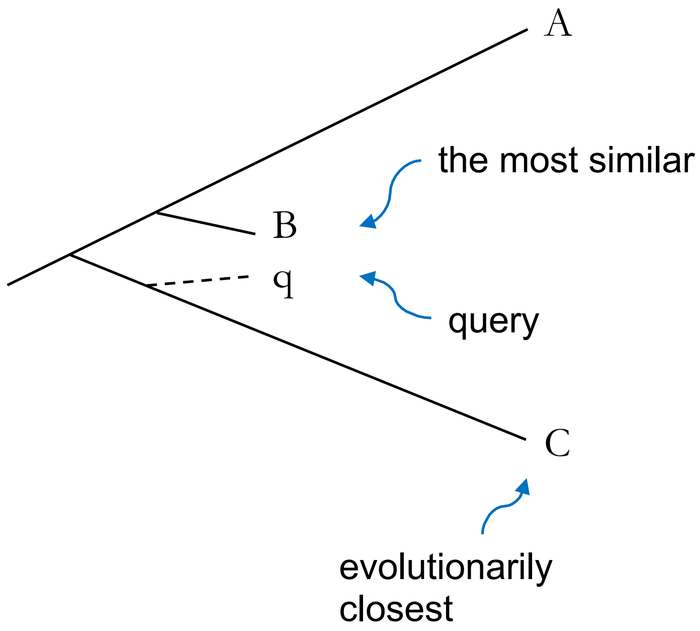


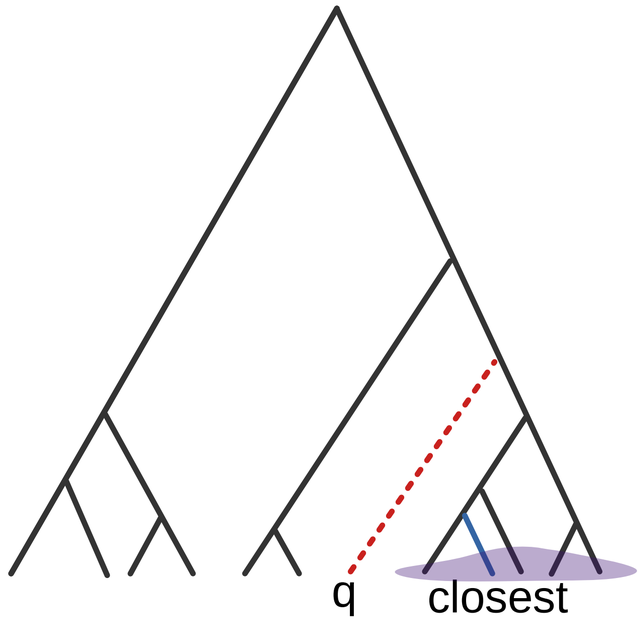
[Balaban, et al, 2022]

Closest match is not enough

- Closest match may mislead
  - Assigning to multiple tips may not be sufficient.
- Closest length ≠ Closest clade for non-ultrametric trees (e.g., uneven rates of evolution)

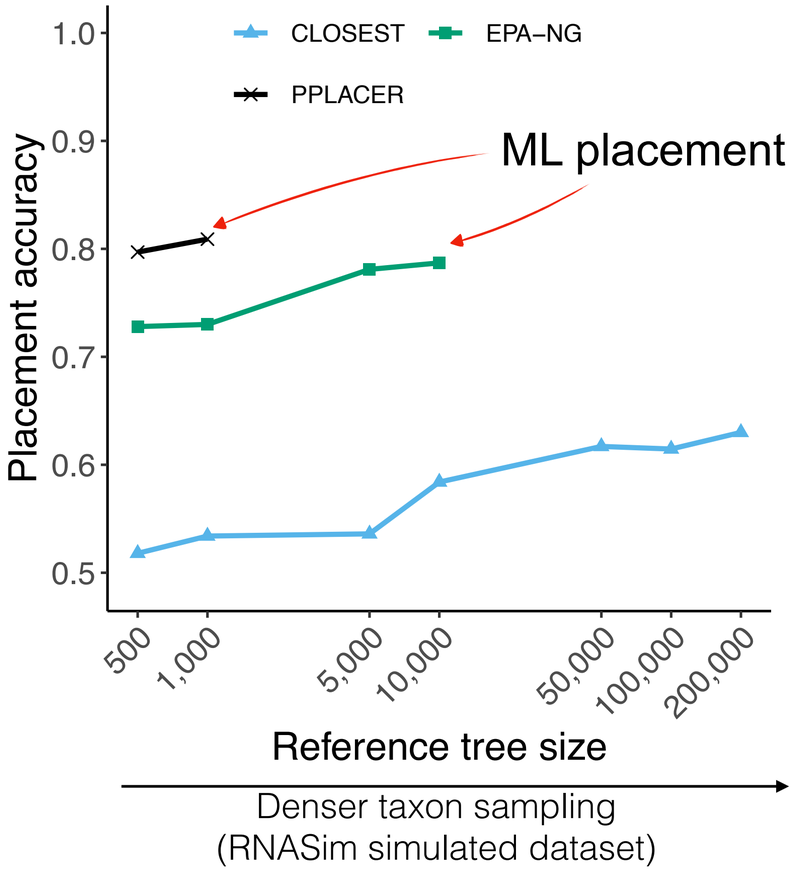


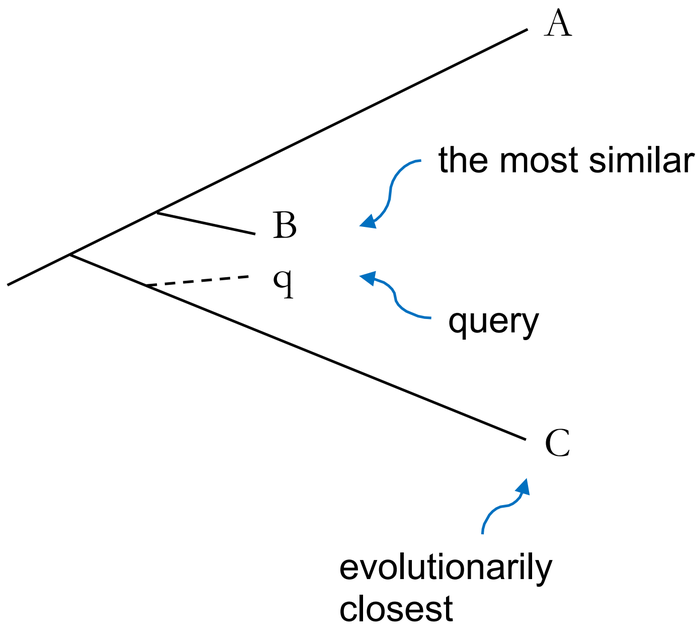


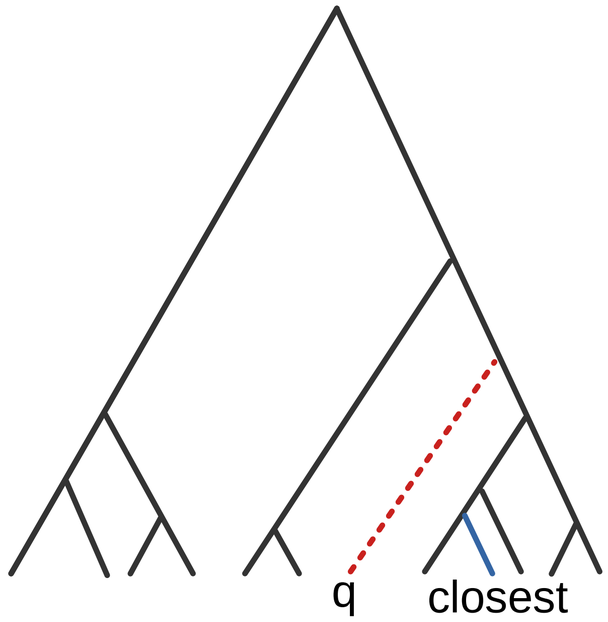


[Balaban, et al, 2022]

Phylogenetic placement can do better!







Perhaps the OGU approach is good enough?

Not in theory + Alignment is not scalable

Perhaps the OGU approach is good enough?

Not in theory + Alignment is not scalable

Is full placement of all reads on a species tree better?

No method is designed for this.

Perhaps the OGU approach is good enough?

Not in theory + Alignment is not scalable

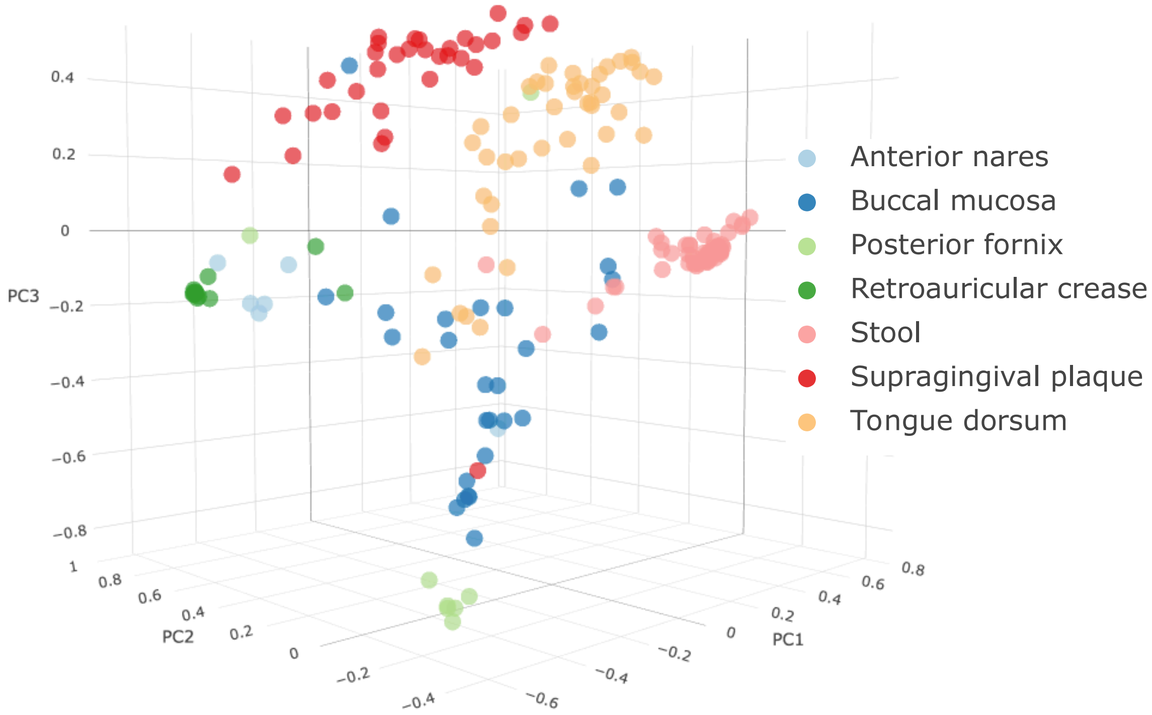
Is full placement of all reads on a species tree better?

No method is designed for this.

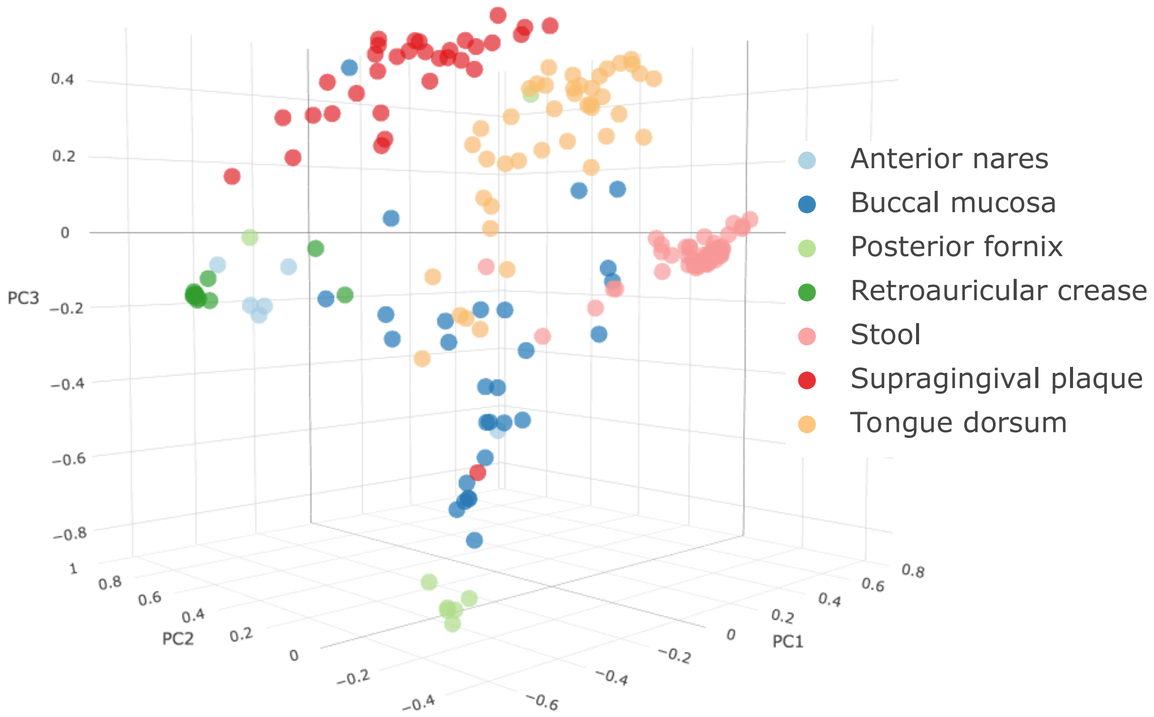
krepp: k-mer-based read phylogenetic placement

Place reads from **anywhere on a genome** onto an **ultra-large tree** of reference genomes **without alignment**

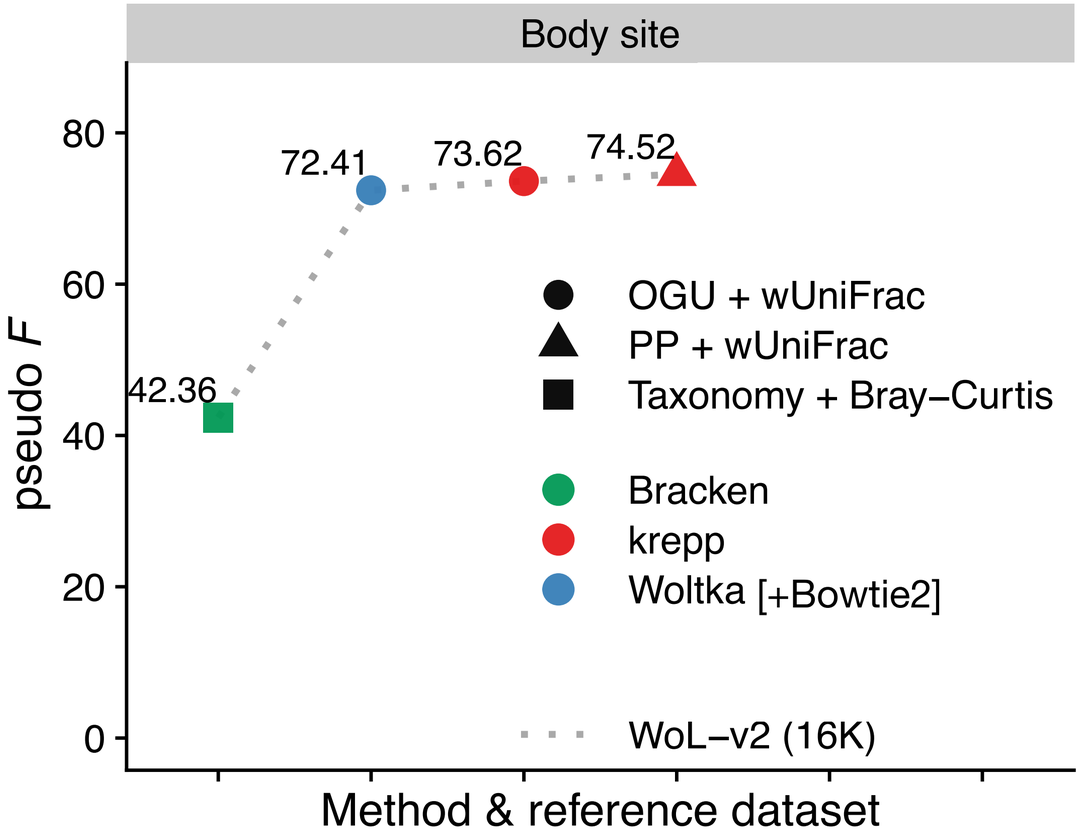
Analyzing human microbiome with larger references



Analyzing human microbiome with larger references

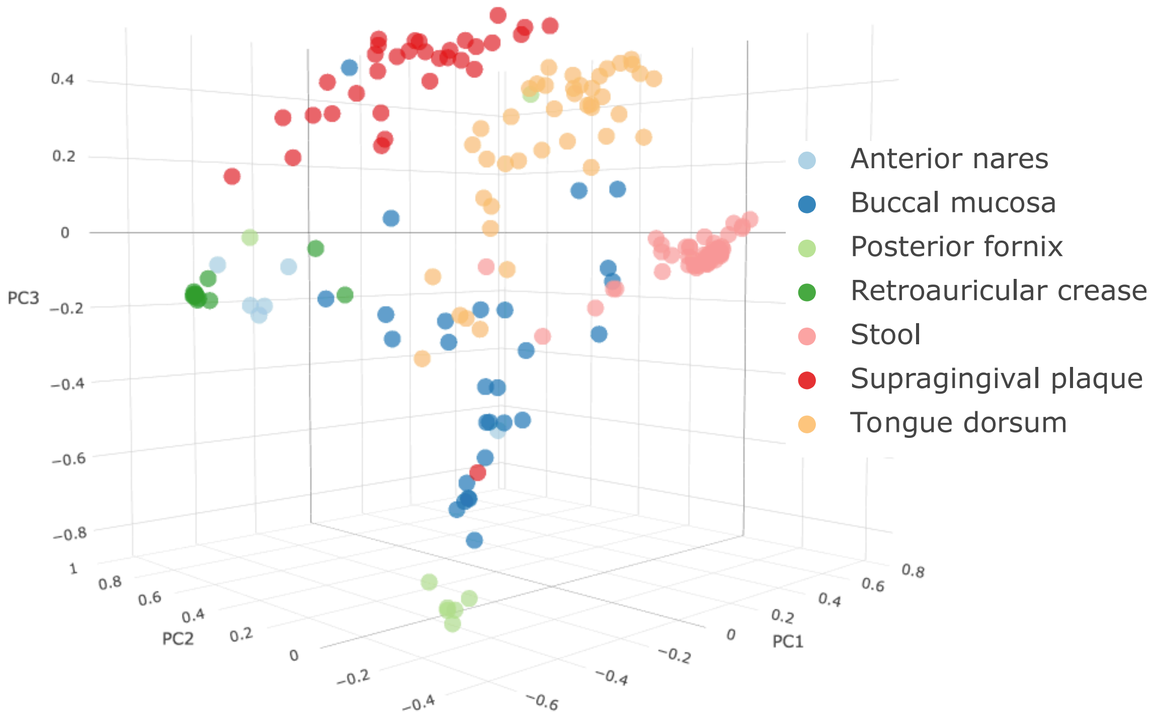


Reference: Web of Life (v2)
16,000 microbial genomes

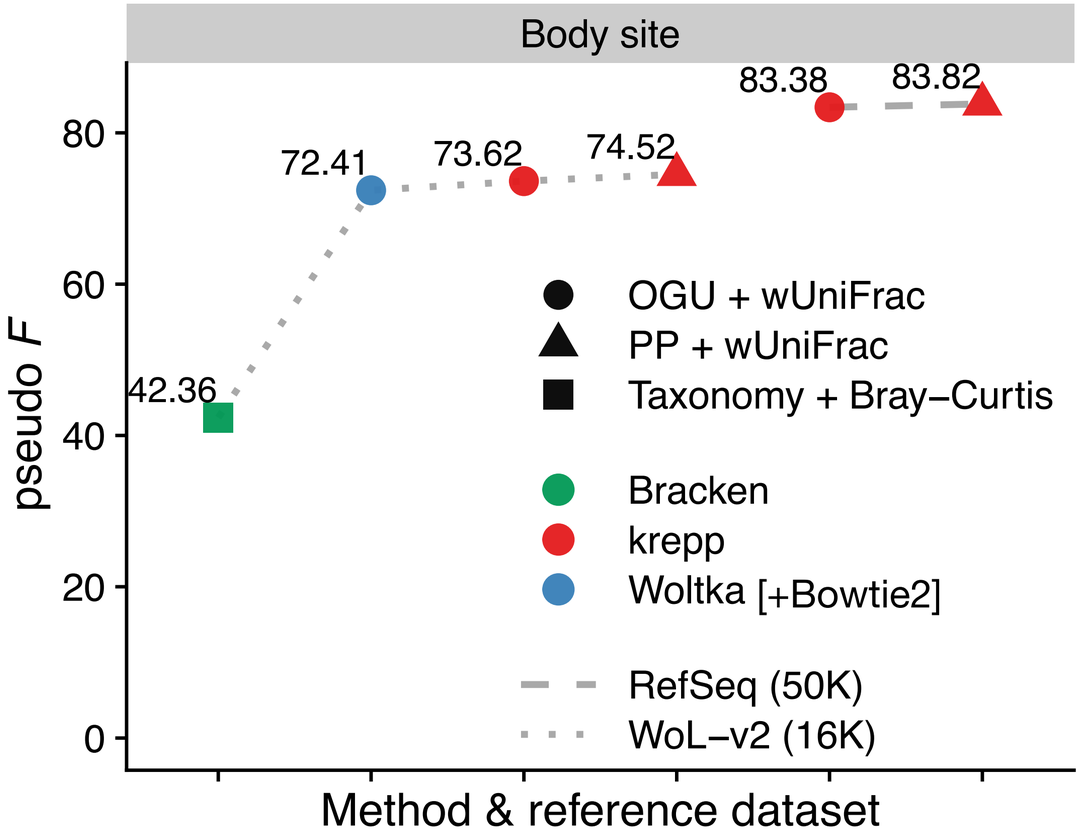


Analyzing human microbiome with larger references

Scaling to large references further **improves** separation of body sites.

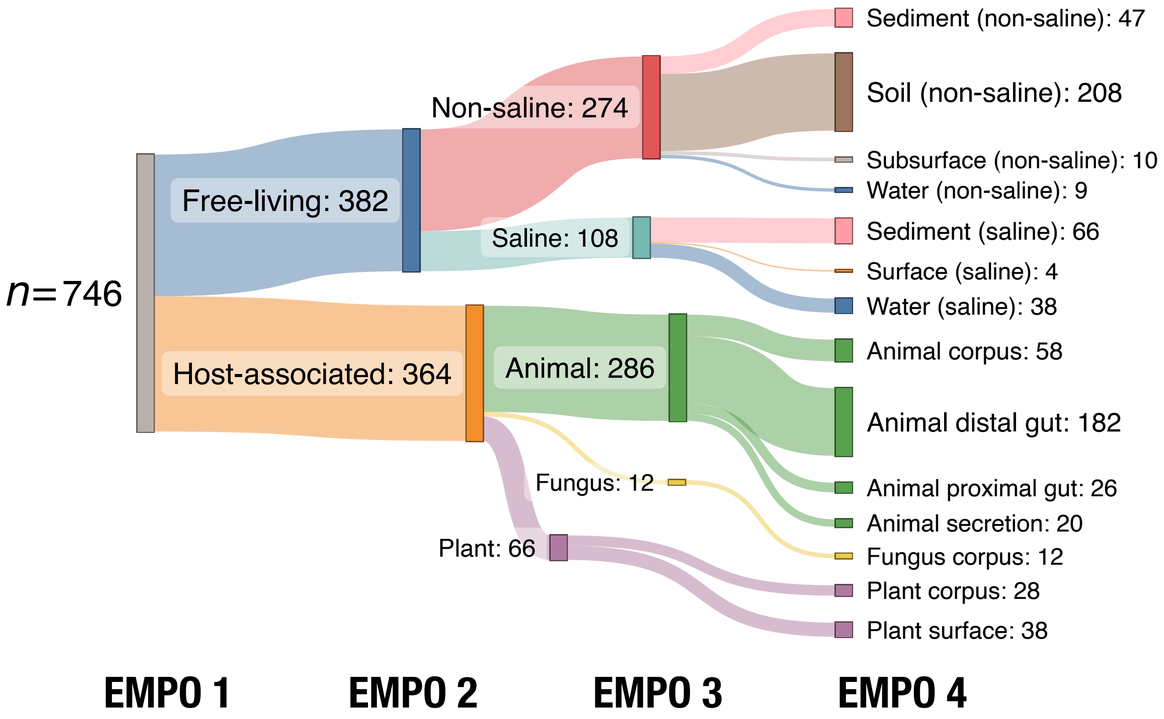


Reference: RefSeq subset
50,000 microbial genomes

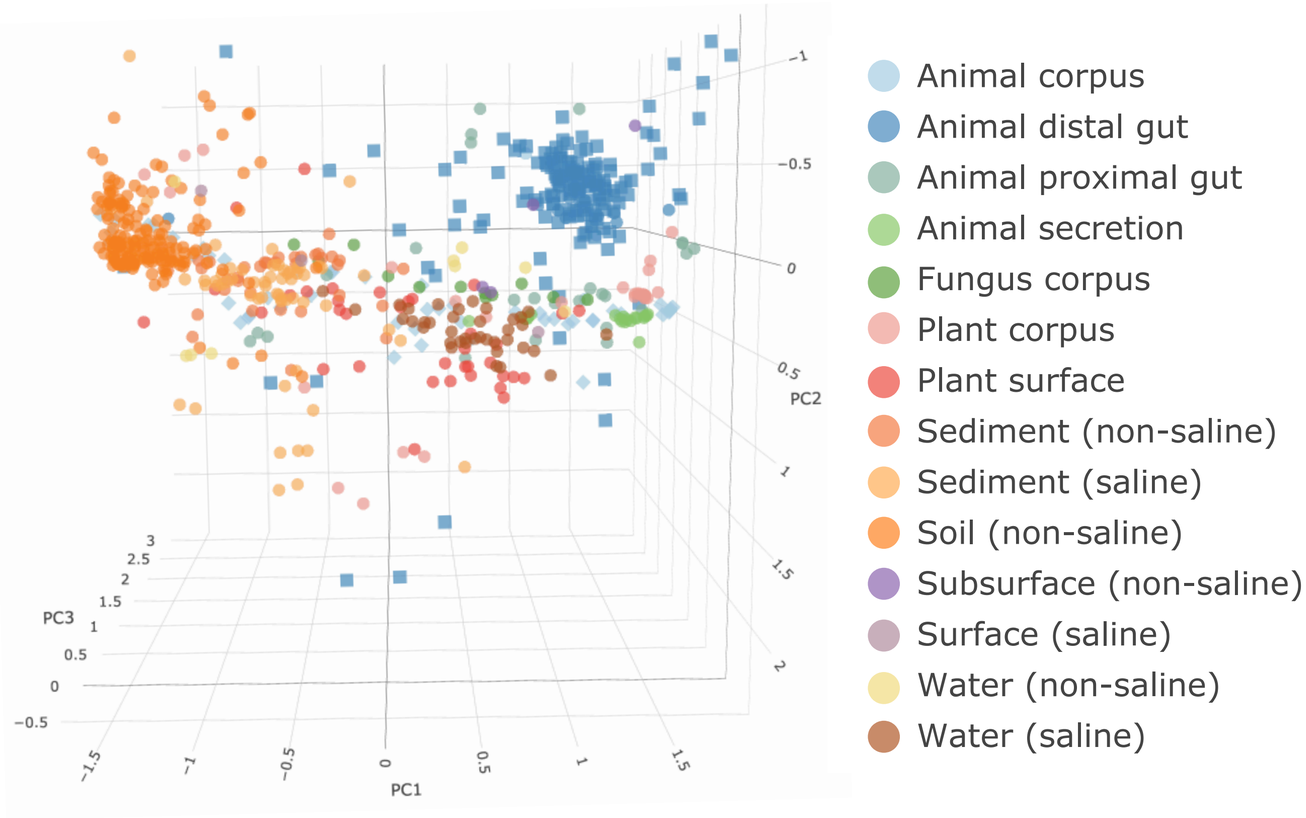


Better characterization of less-studied microbiome of earth

Hierarchical categorization
of earth microbiome samples

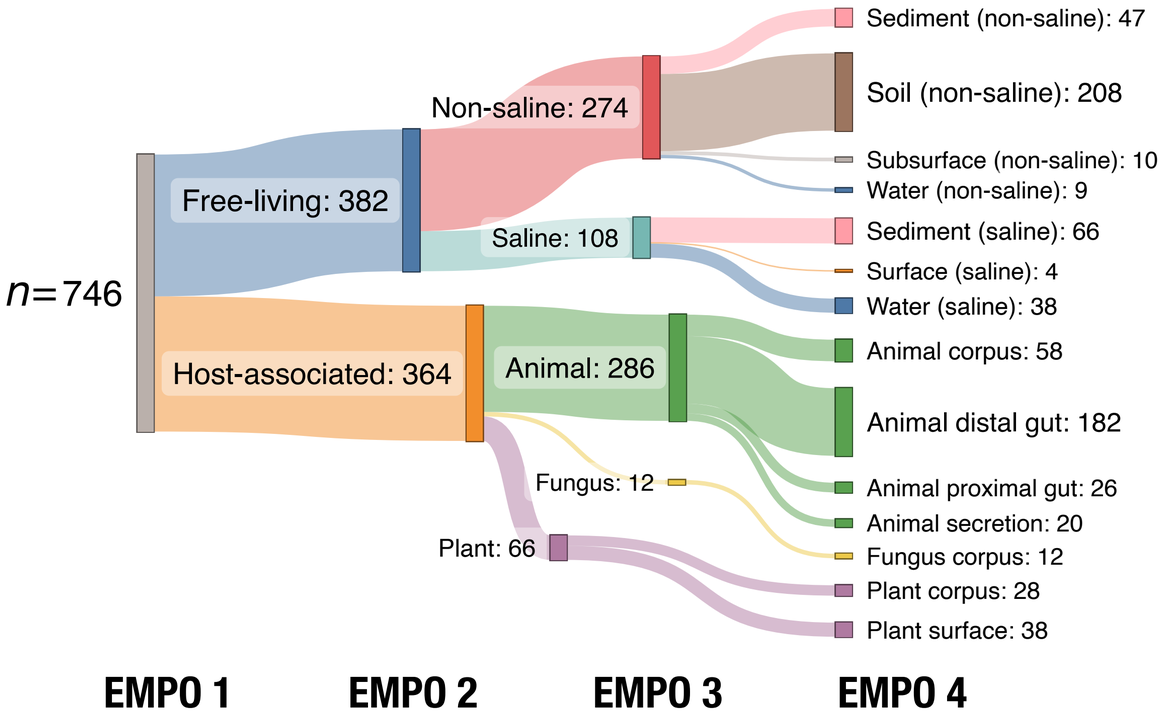


Reference: Web of Life (v1)
10,500 microbial genomes

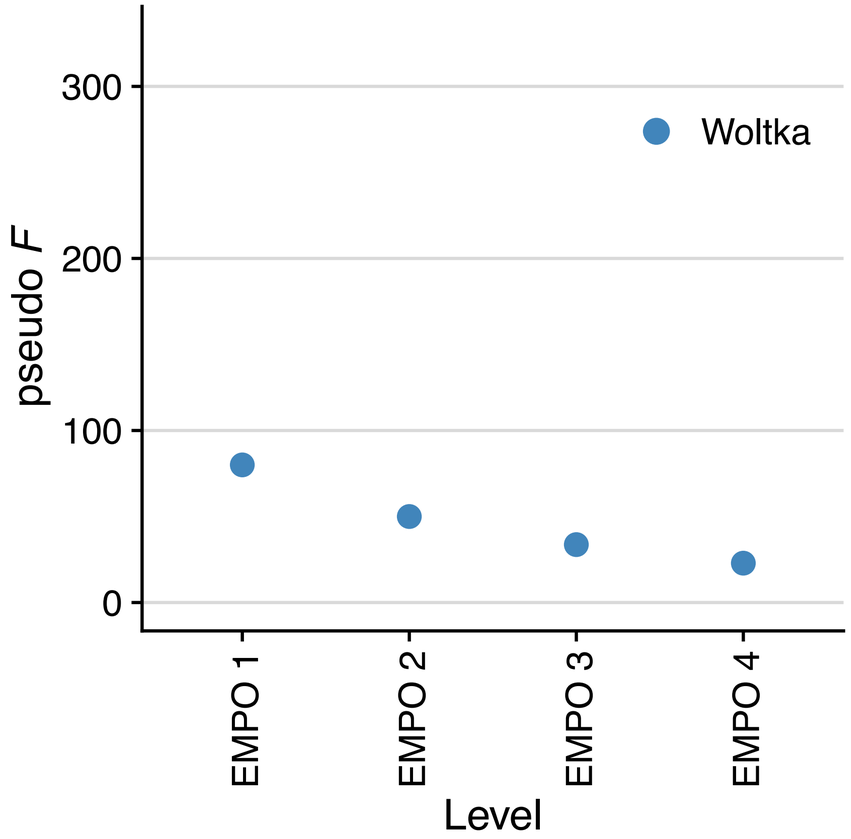


Better characterization of less-studied microbiome of earth

Hierarchical categorization
of earth microbiome samples

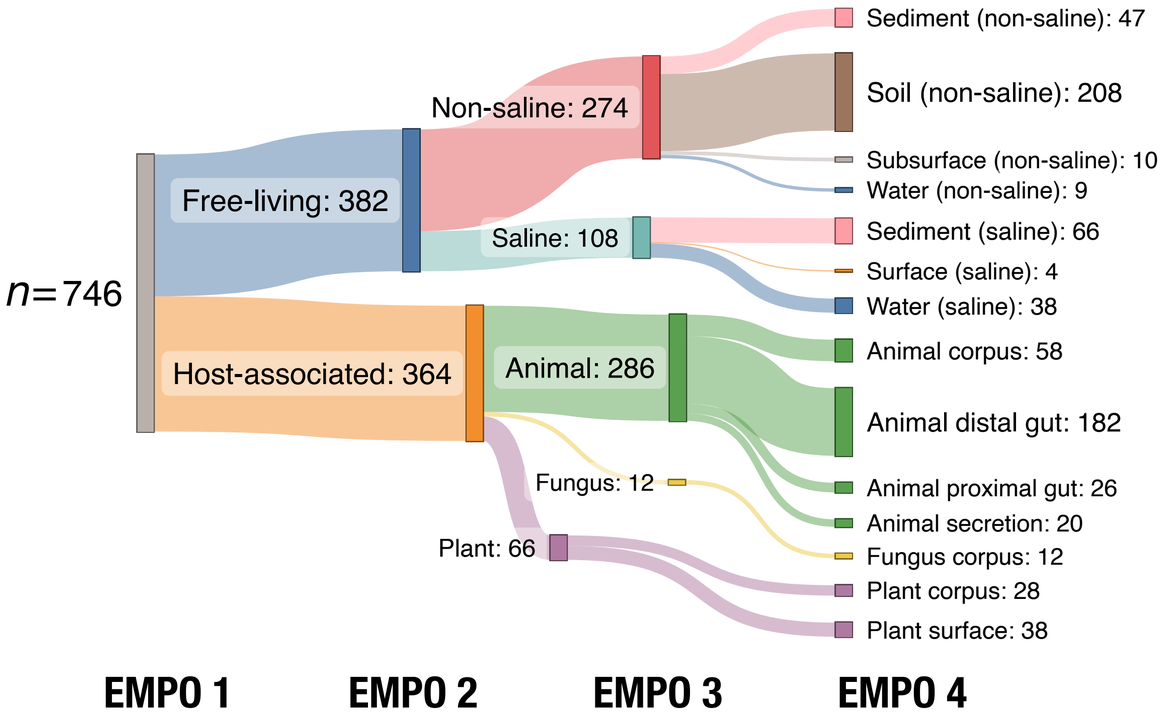


Reference: Web of Life (v1)
10,500 microbial genomes

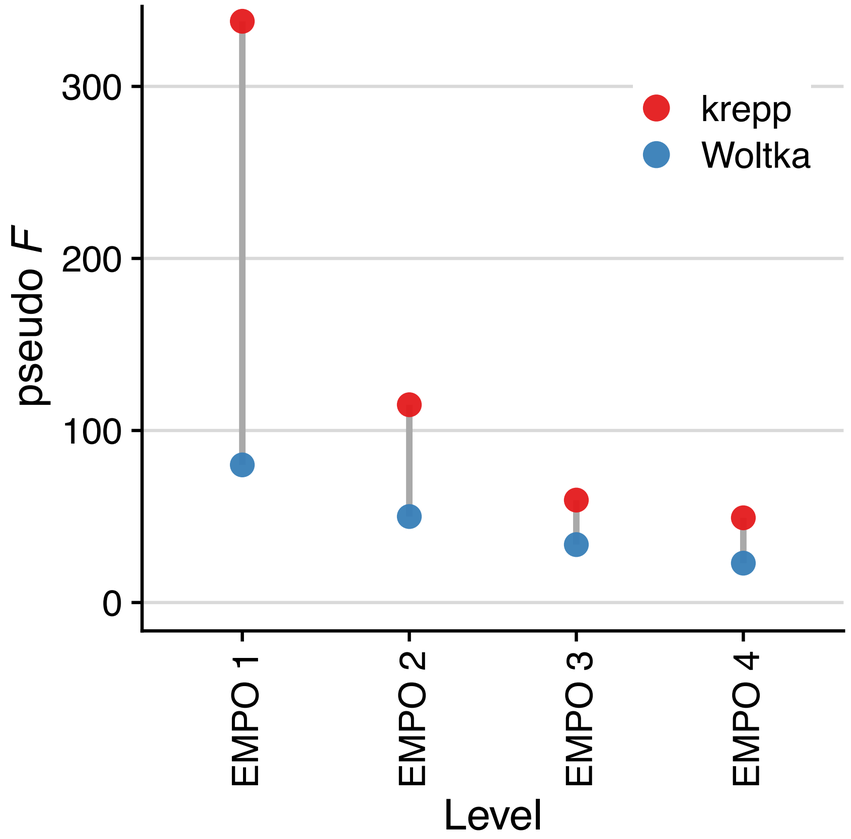


Better characterization of less-studied microbiome of earth

Hierarchical categorization
of earth microbiome samples



Reference: Web of Life (v1)
10,500 microbial genomes



Problem statement for krepp

Given:

- query sequence q
- set of references $\mathscr{R} = \{R_1, \ldots, R_N\}$
- a backbone phylogeny T

>q
CCTGCTA...

reference genomes

R_1: TCCCTGCTCA...
R_2: TCCCTGCTAA...
R_3: CCCCTGGCAG...
R_4: ATTATCTGAT...
...
R_N: CCCCAAACAA...

Problem statement for krepp

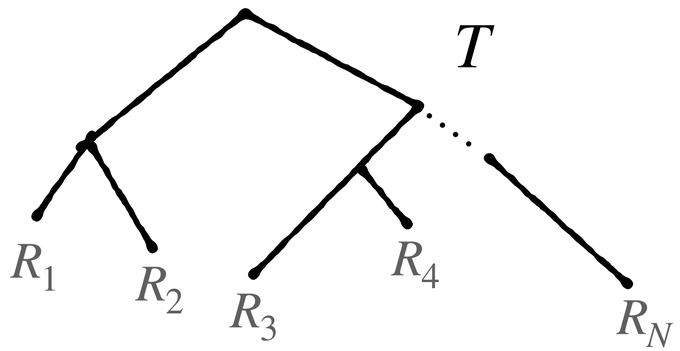
Given:

- query sequence q
- set of references $\mathscr{R} = \{R_1, \ldots, R_N\}$
- a backbone phylogeny T

>q
CCTGCTA...

reference genomes

R_1: TCCCTGCTCA...
R_2: TCCCTGCTAA...
R_3: CCCCTGGCAG...
R_4: ATTATCTGAT...
. . .
R_N: CCCCAAACAA...



Problem statement for krepp

Given:

- query sequence q
- set of references $\mathscr{R} = \{R_1, \ldots, R_N\}$
- a backbone phylogeny T

Interpretation of the distances:

$i)$ $d(q, R) = \dfrac{\text{\# of mismatches}}{\text{length of } q}$

CCTGCTA…
…AGTTATCCCTGCTCA…
x

$ii)$ $\mathbb{E}_Q[d(q, R)] \approx 1 - \text{ANI}(Q, R)$
where Q is the source genome of q

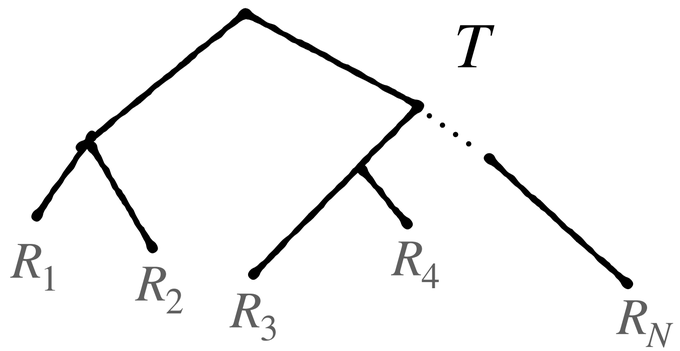
>q
CCTGCTA…

reference genomes

R_1: TCCCTGCTCA…
R_2: TCCCTGCTAA…
R_3: CCCCTGGCAG…
R_4: ATTATCTGAT…
. . .
R_N: CCCCAAACAA…

distance estimates

$\mathrm{d}(q, R_1) = 0.141$
$\mathrm{d}(q, R_2) = 0.001$
$\mathrm{d}(q, R_3) = 0.195$
$\mathrm{d}(q, R_4) = \mathrm{NA}$
. . .
$\mathrm{d}(q, R_N) = 0.244$



Problem statement for krepp

Given:

- query sequence q
- set of references $\mathscr{R} = \{R_1, \ldots, R_N\}$
- a backbone phylogeny T

Interpretation of the distances:

$i)$ $d(q, R) = \dfrac{\text{\# of mismatches}}{\text{length of } q}$

```
CCTGCTA...
...AGTTATCCCTGCTCA...
          x
```

$ii)$ $\mathbb{E}_Q[d(q, R)] \approx 1 - \text{ANI}(Q, R)$
where Q is the source genome of q

```
>q
CCTGCTA...
```

reference genomes

R_1: TCCCTGCTCA...
R_2: TCCCTGCTAA...
R_3: CCCCTGGCAG...
R_4: ATTATCTGAT...
...
R_N: CCCCAAACAA...

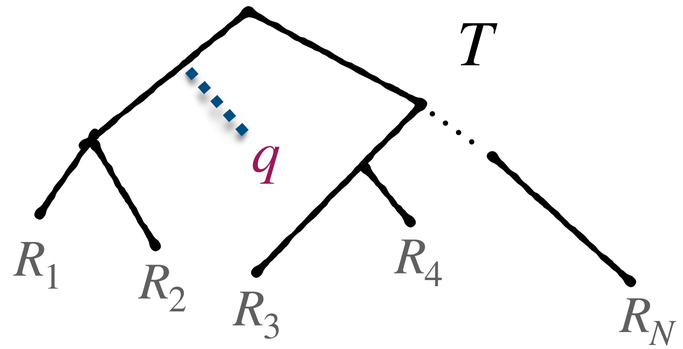
→

distance estimates

$\mathrm{d}(q, R_1) = 0.141$
$\mathrm{d}(q, R_2) = 0.001$
$\mathrm{d}(q, R_3) = 0.195$
$\mathrm{d}(q, R_4) = \mathrm{NA}$
...
$\mathrm{d}(q, R_N) = 0.244$

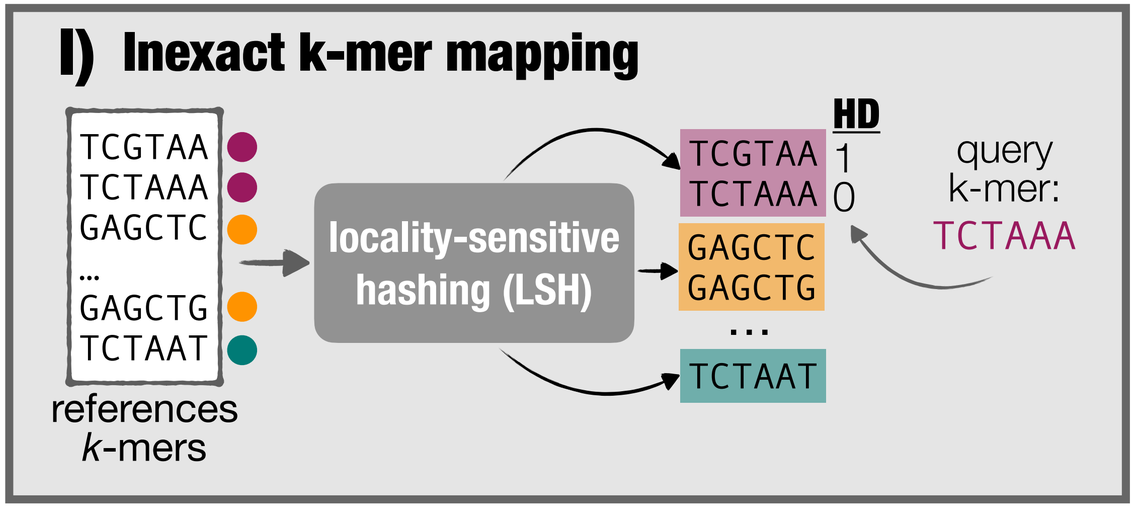
→

phylogenetic placement:

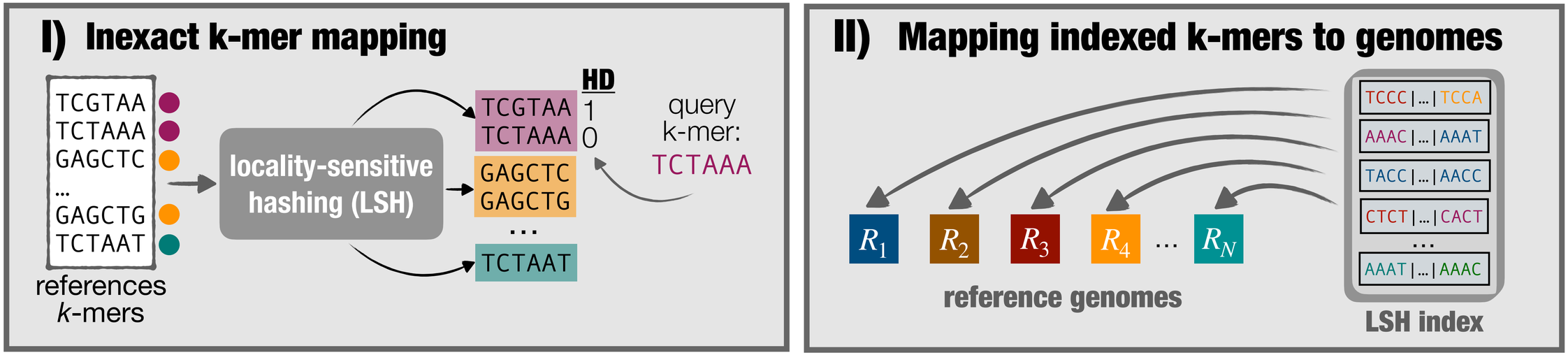


Four computational (sub)problems

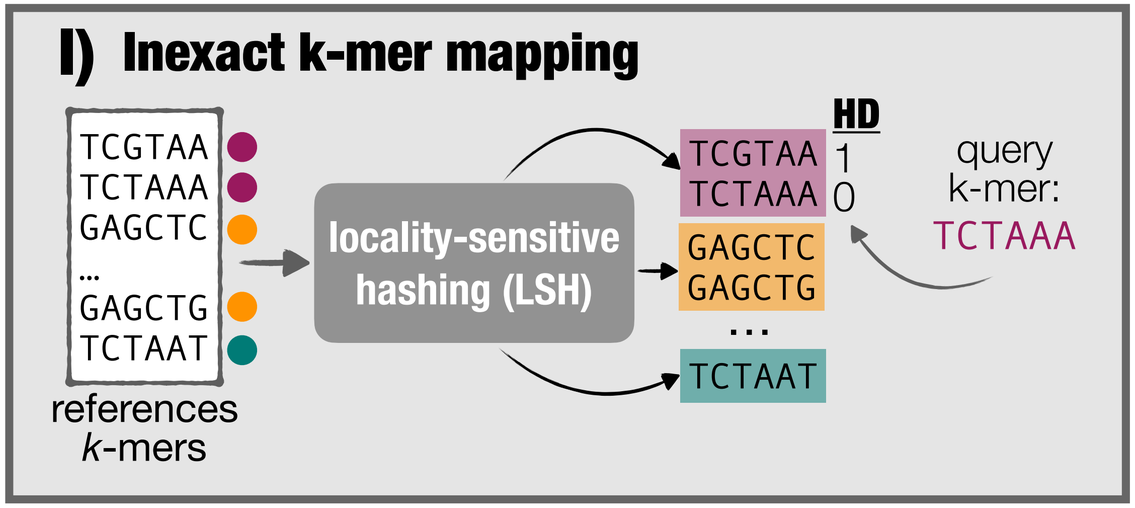
Four computational (sub)problems

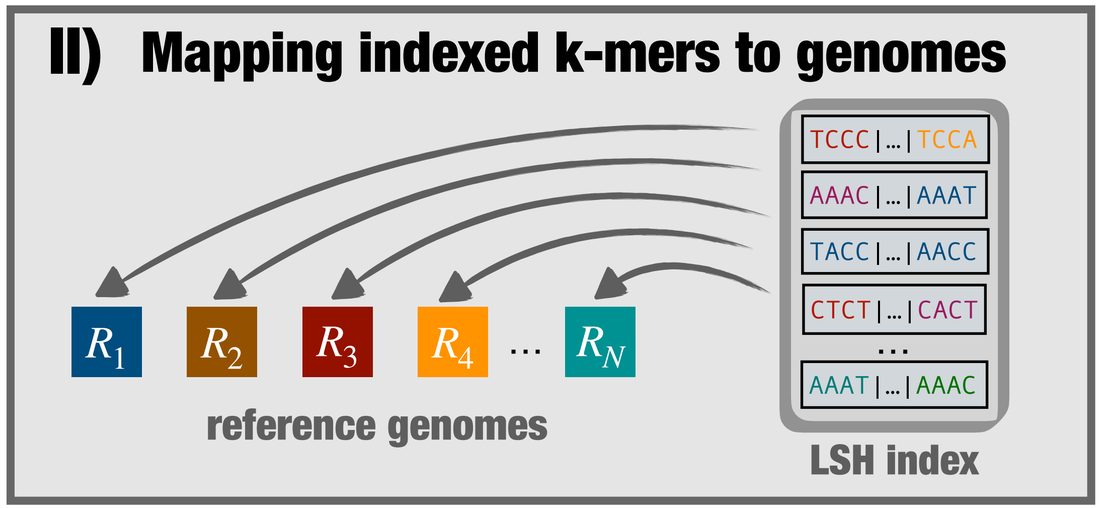


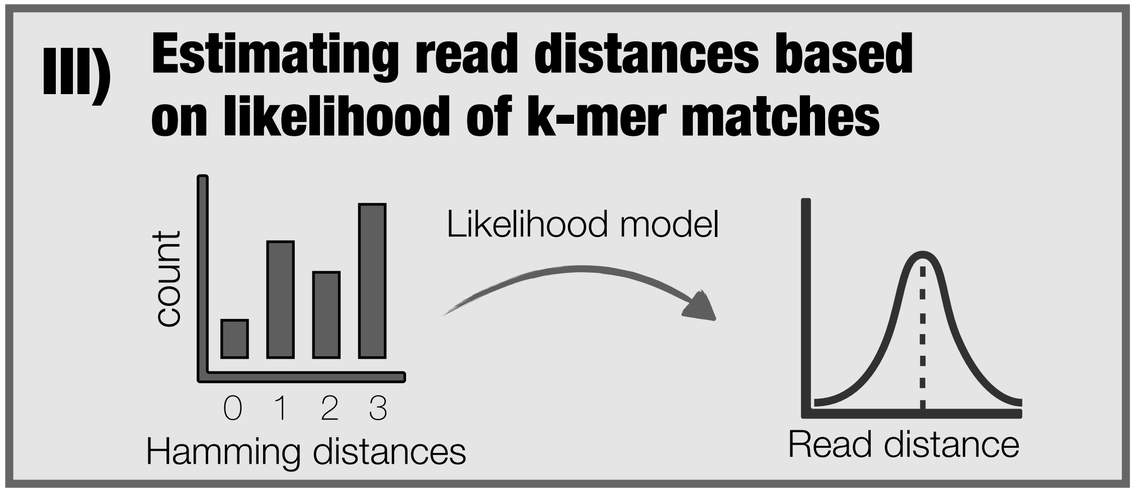
Four computational (sub)problems



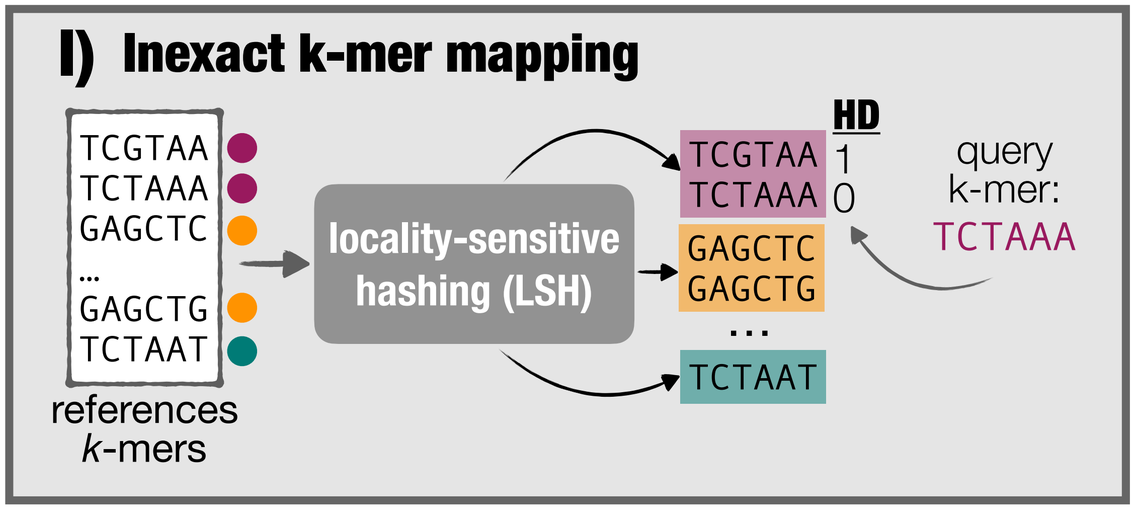
Four computational (sub)problems

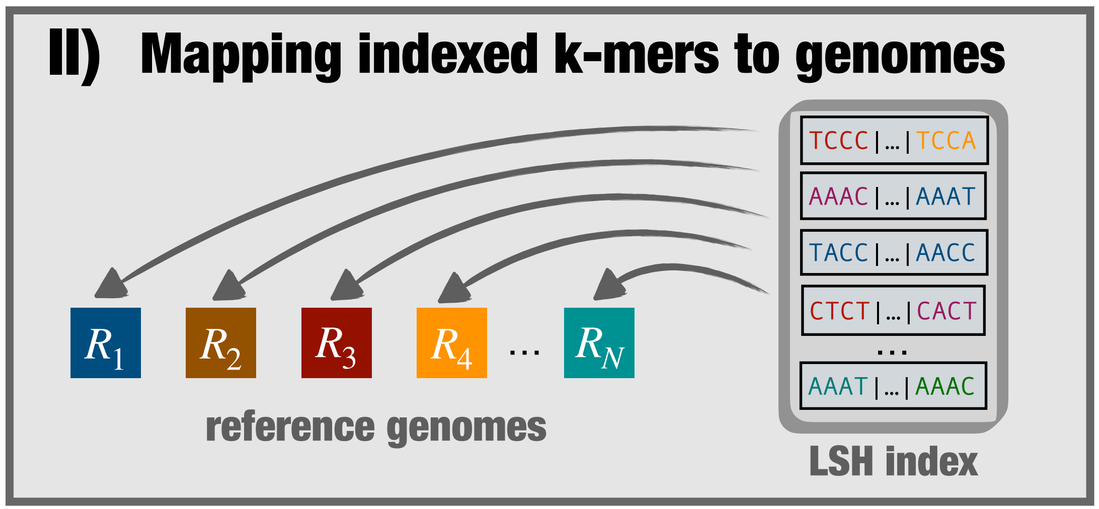


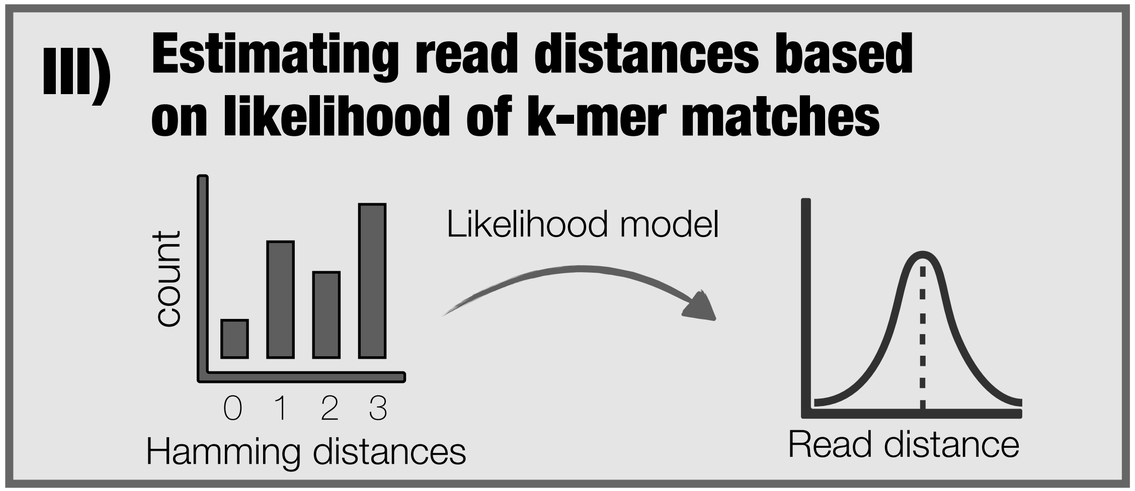


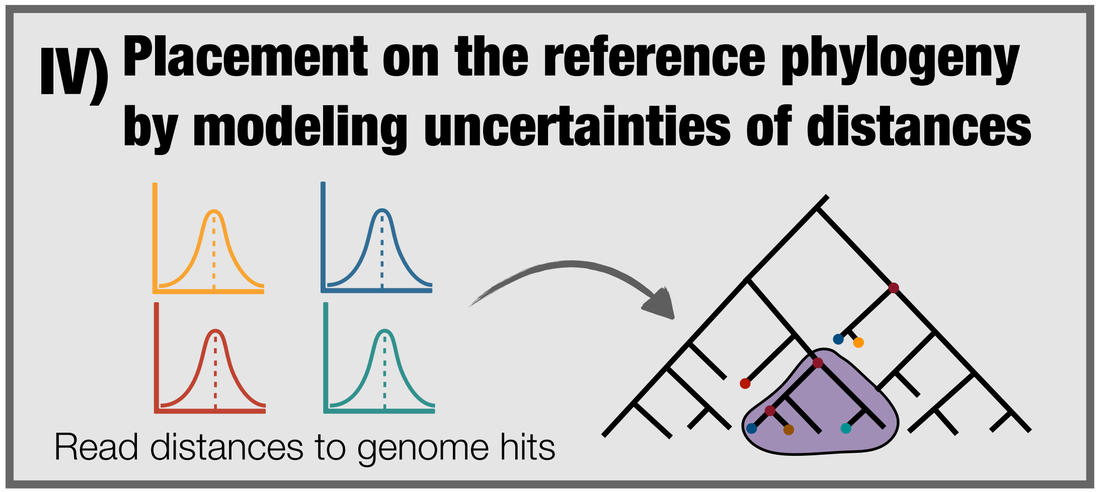


Four computational (sub)problems









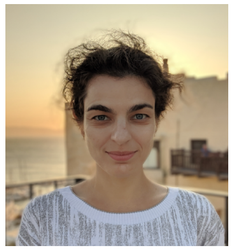
Problem 1: Find similar (long) k-mers

- Represent each reference genome as a set of k-mers: $R_i = \{x_1^{(i)} \ldots x_L^{(i)}\}$
- Create a table of all $\mathscr{R} = R_1 \cup \ldots \cup R_N$
- For each k-mer y of a query read:
  - Find all $x_j^{(i)} \in \mathscr{R}$:

    $$\text{Hamming distance}\,(y, x_i^{(j)}) \leq \delta$$

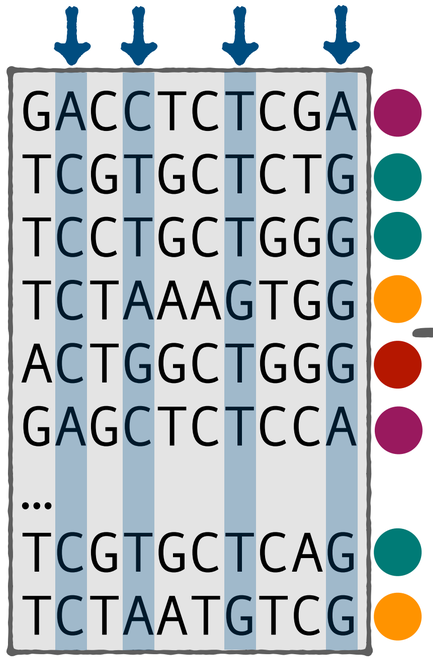
    for a fixed δ
    (e.g., $\delta = 4$ for $k = 31$)

Locality-Sensitive Hashing (LSH)

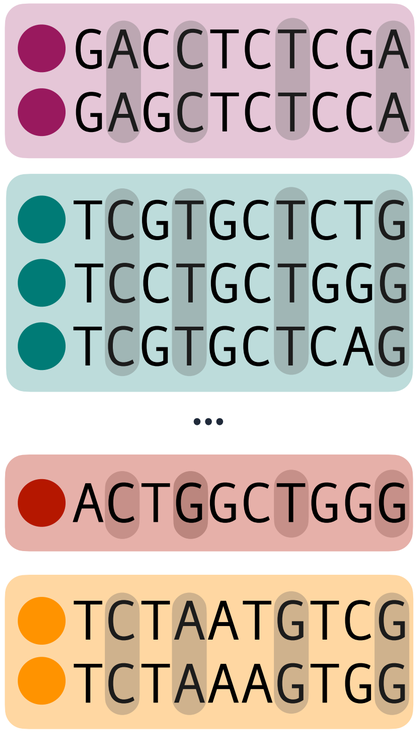


CONSULT
(Rachtman, et al., 2021)

Select h random but fixed positions (default h: 14, k: 29)



reference *k*-mers



4^h LSH buckets

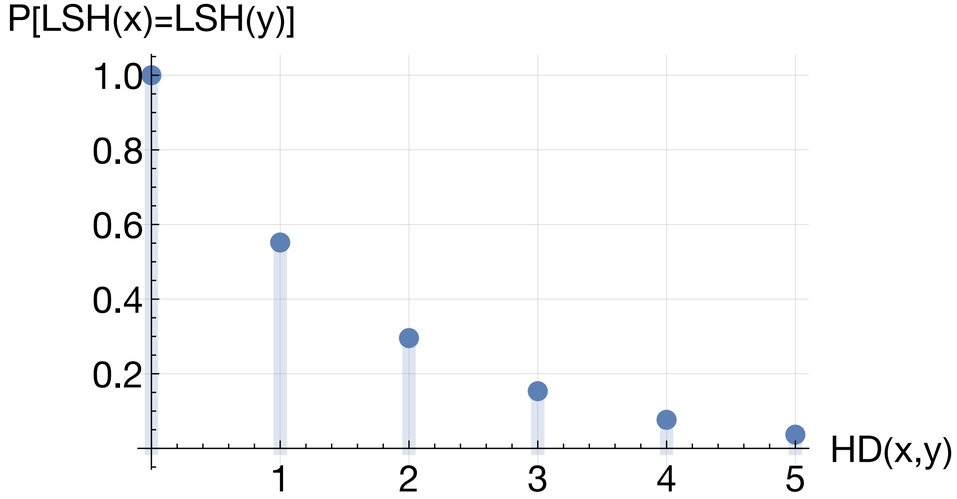
HD
4
1
4

miss at HD=2

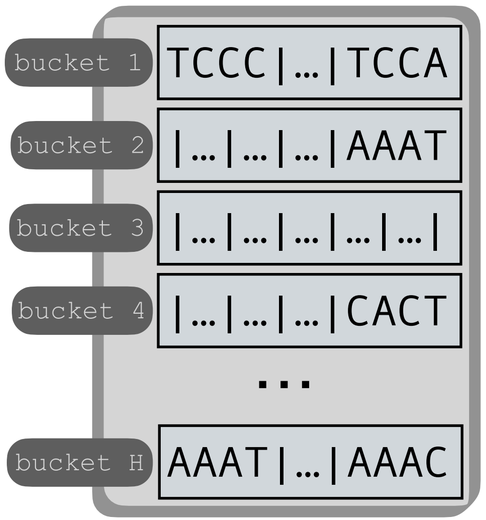
Given a query *k*-mer

ACCTGCTGGG

collides for HD=x with probability: $\frac{\binom{k-h}{x}}{\binom{k}{x}}$

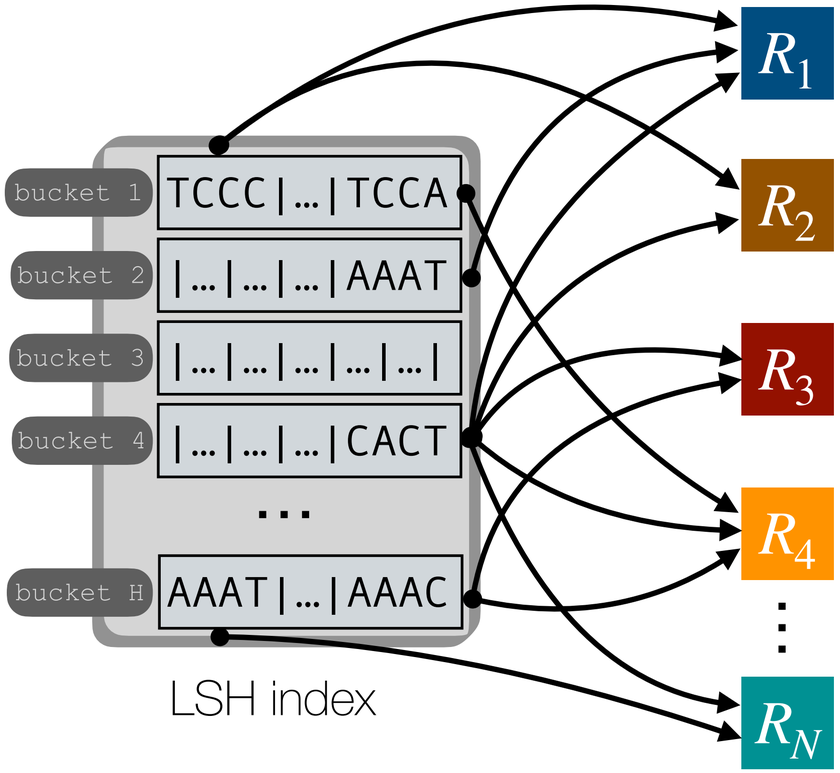


Problem 2: Mapping indexed k-mers to reference genomes



LSH index

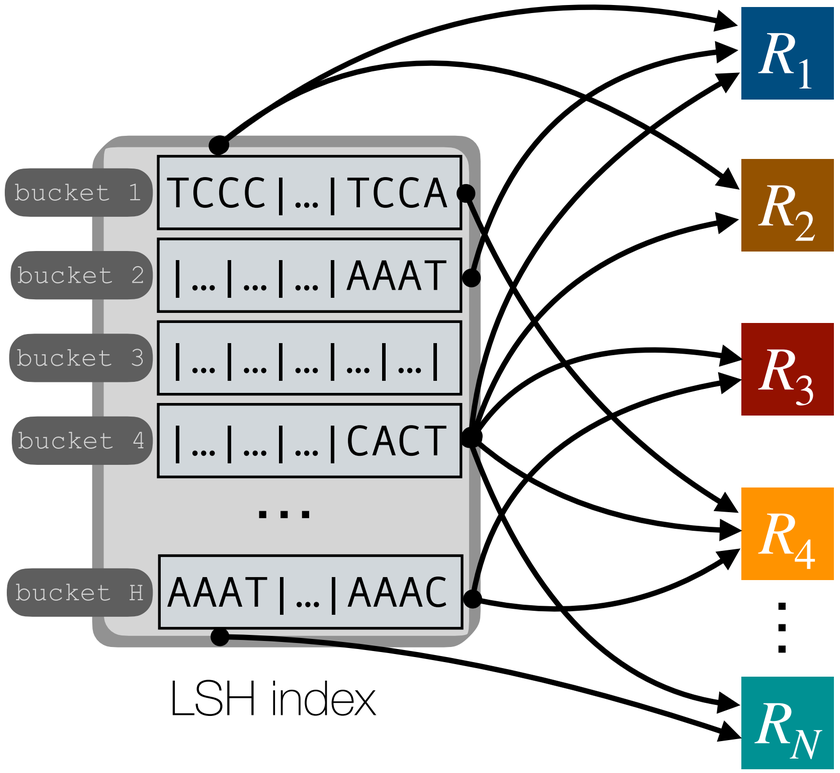
Problem 2: Mapping indexed k-mers to reference genomes

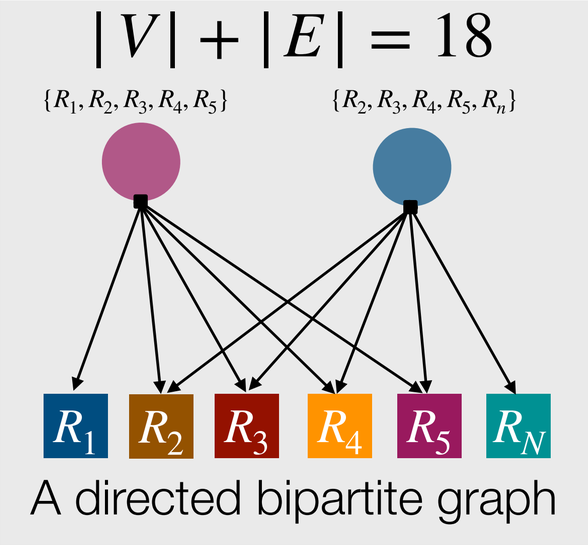


colored *k*-mer problem

color: a subset of references
(including singletons)

Problem 2: Mapping indexed k-mers to reference genomes

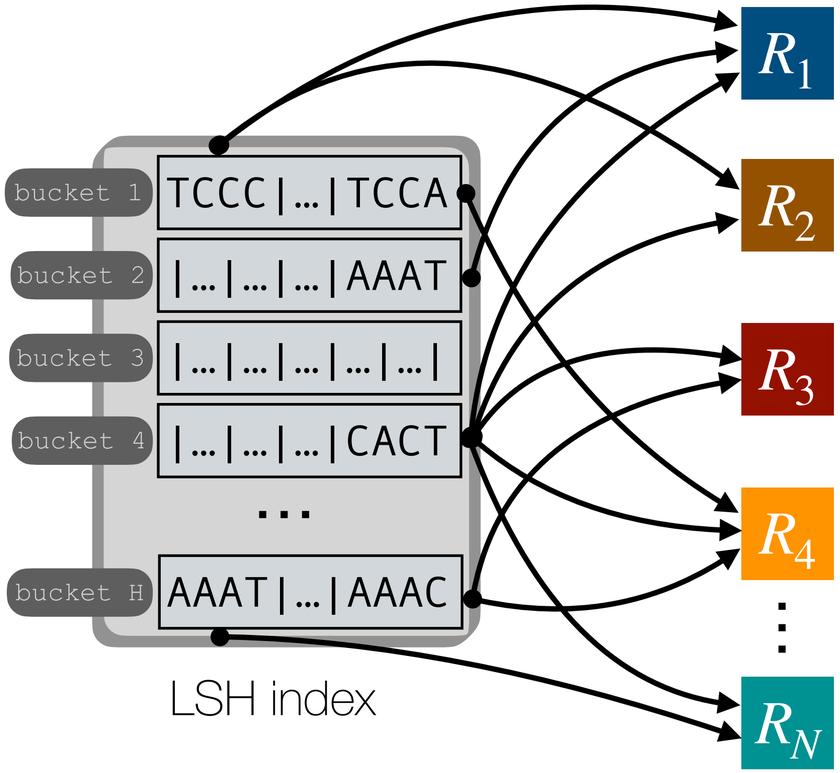




colored *k*-mer problem

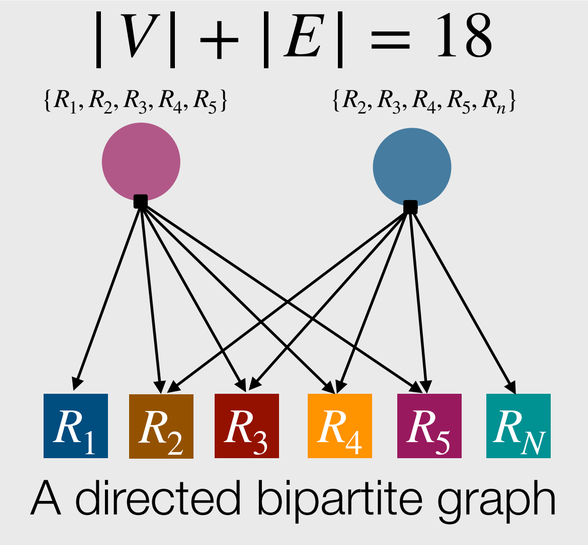
color: a subset of references (including singletons)

Problem 2: Mapping indexed k-mers to reference genomes



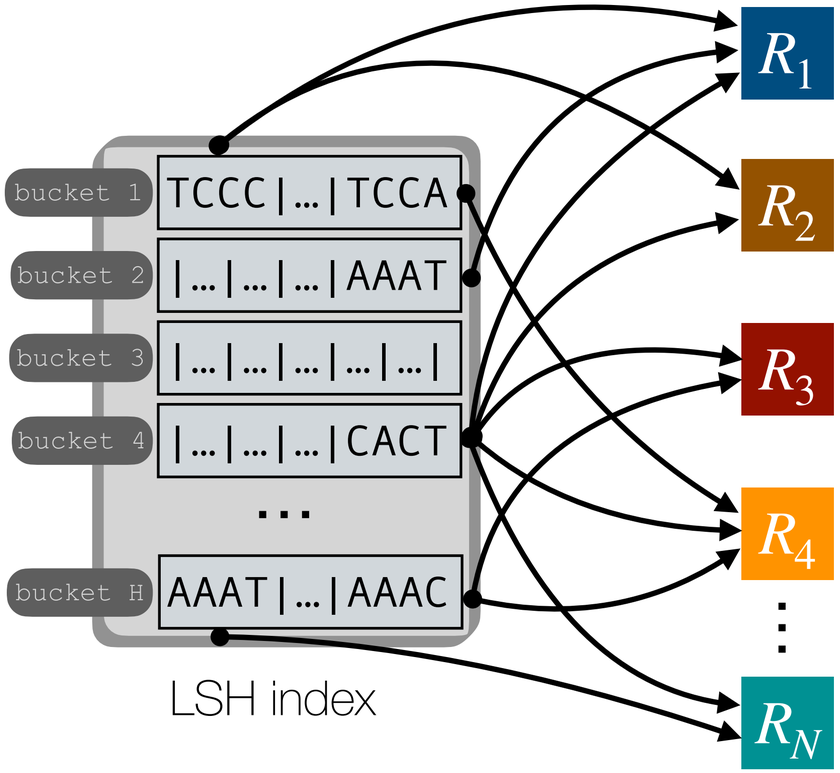
colored *k*-mer problem

color: a subset of references (including singletons)



Minimize $|V| + |E|$ **?**

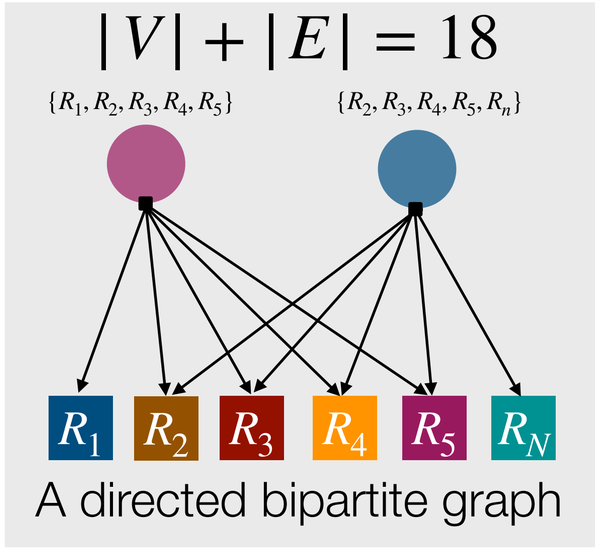
Problem 2: Mapping indexed k-mers to reference genomes



LSH index

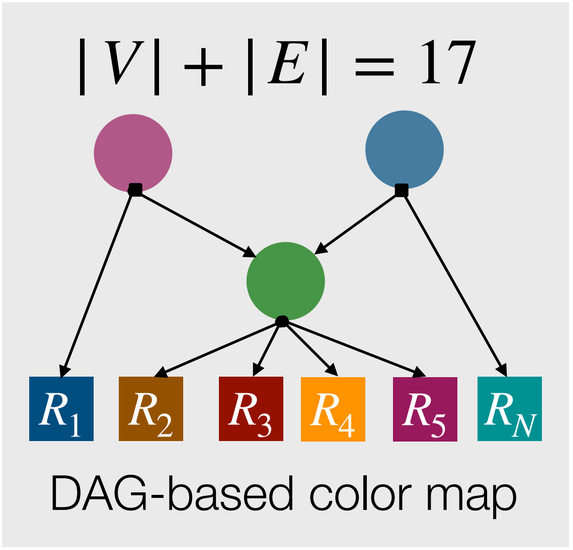
colored *k*-mer problem

color: a subset of references (including singletons)



A directed bipartite graph

Add extra colors



DAG-based color map

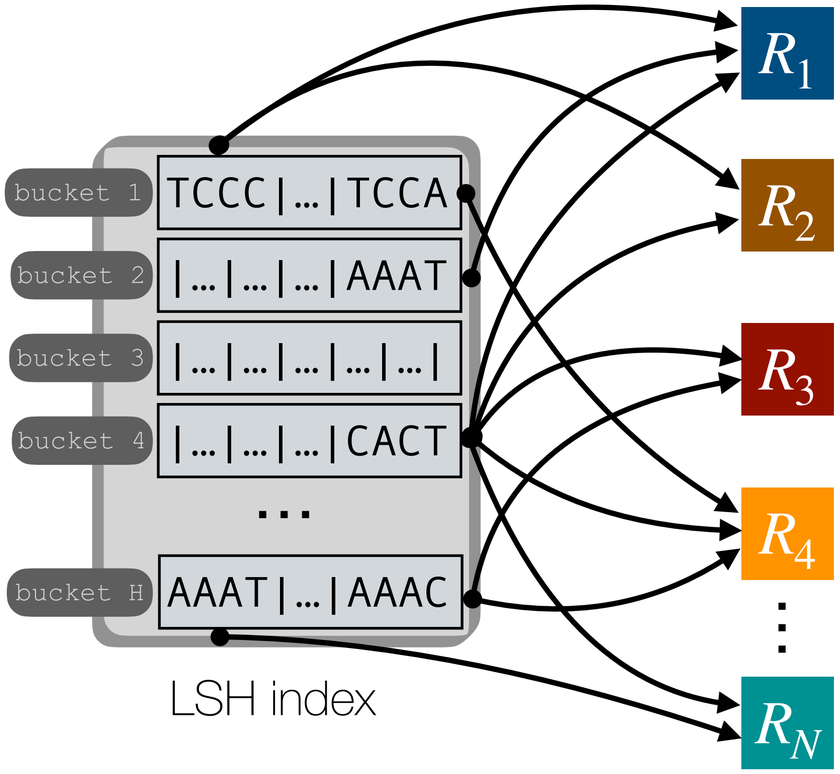
Minimize $|V| + |E|$?

i. add nodes for frequently shared sub-colors (similar to *meta-colors* from *Campanelli et al., 2024)*

ii. explain larger color w/ smaller existing colors

iii. follow edges to reconstruct colors

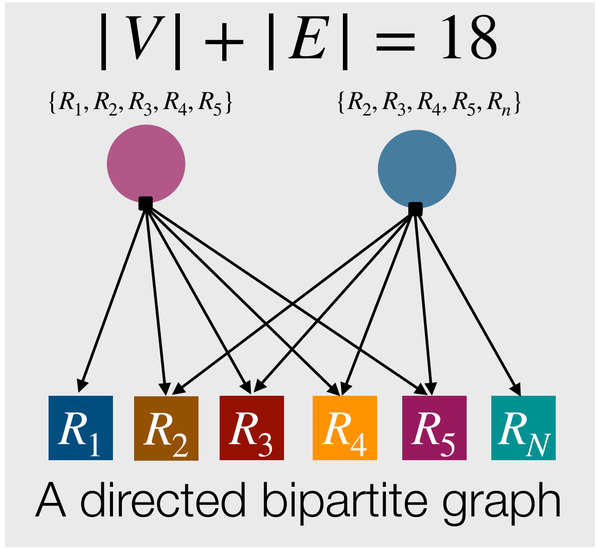
Problem 2: Mapping indexed k-mers to reference genomes



LSH index

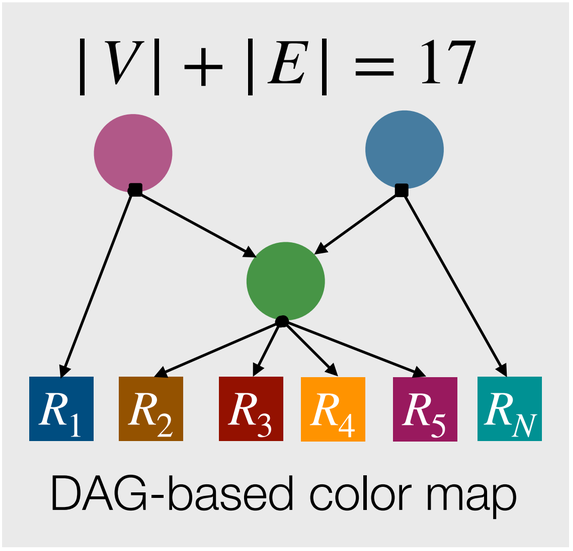
colored *k*-mer problem

color: a subset of references (including singletons)



A directed bipartite graph

Add extra colors



DAG-based color map

Minimize $|V| + |E|$?

i. add nodes for frequently shared sub-colors (similar to *meta-colors* from *Campanelli et al., 2024)*

ii. explain larger color w/ smaller existing colors

iii. follow edges to reconstruct colors

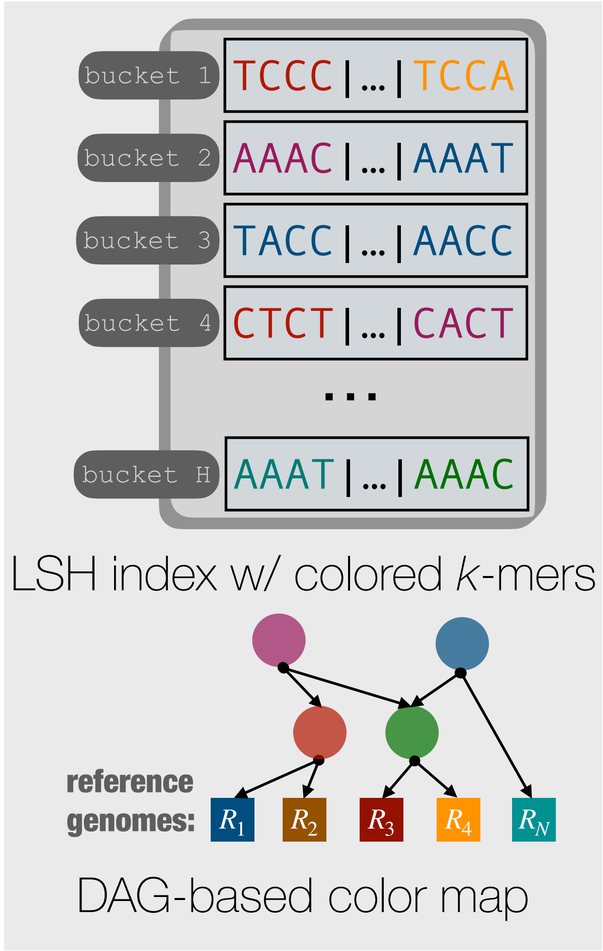
We use a phylogeny-guided heuristic to build a *multi-tree*.

Problem 3: Estimate distances from k-mer matches

>q
ATACCTAGGAGTACGGGAC

1: ATAC
2: TACC
3: ACCT
4: CCTA
5: CTAG
...
L-k+1: GGAC

search *k*-mer matches up to a HD threshold δ



keep the closest match as orthologous

sparse table

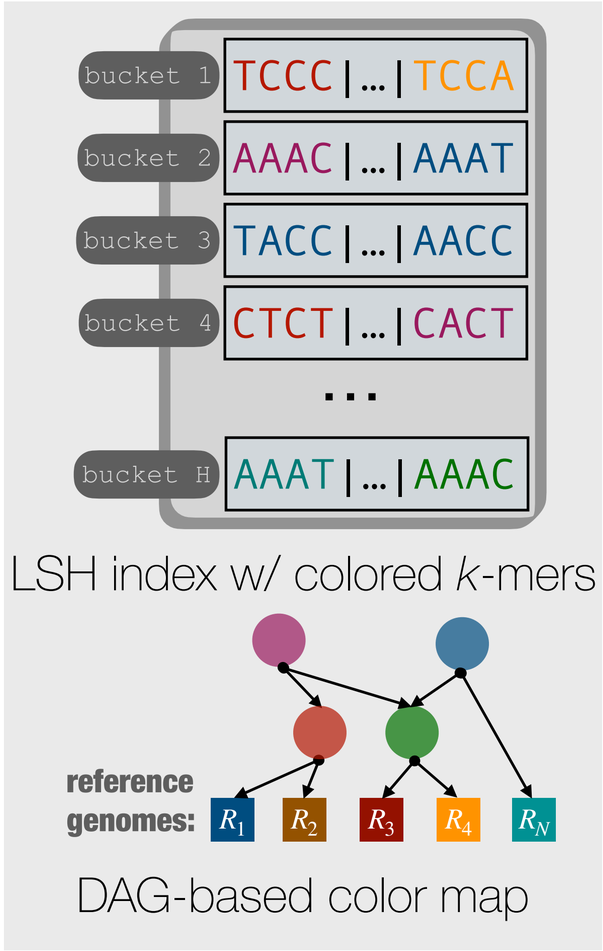
	R_1	R_2	...	R_N
1	0	1	...	-
2	1	4	...	4
3	-	-	...	-
4	-	2	...	-
5	0	-	...	3
...	...	...	...	...
L-k+1	3	0	...	4

Problem 3: Estimate distances from k-mer matches

>q
ATACCTAGGAGTACGGGAC

1: ATAC
2: TACC
3: ACCT
4: CCTA
5: CTAG
...
L-k+1: GGAC

search *k*-mer matches up to a HD threshold δ



keep the closest match as orthologous

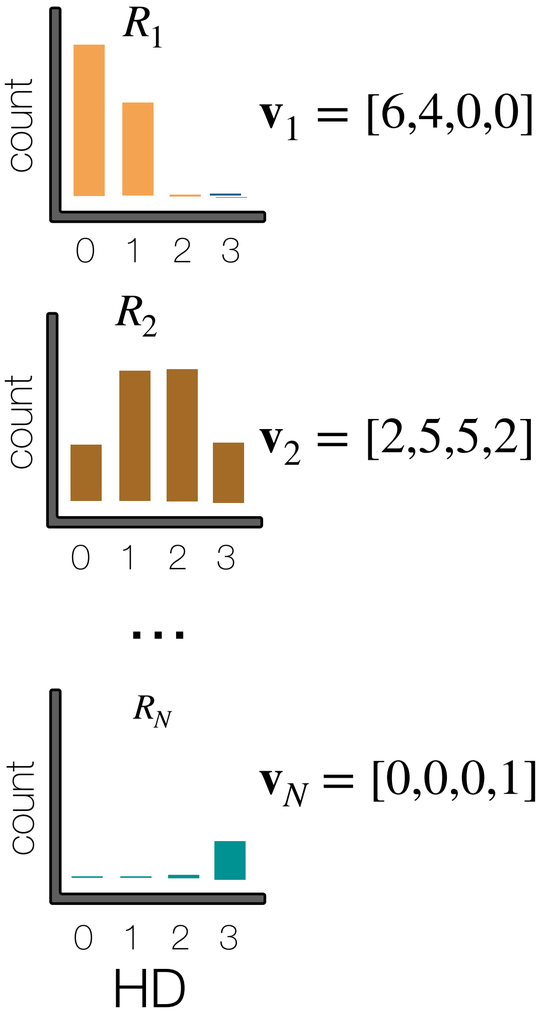
sparse table

	R_1	R_2	...	R_N
1	0	1	...	-
2	1	4	...	4
3	-	-	...	-
4	-	2	...	-
5	0	-	...	3
...	...	...	...	...
L-k+1	3	0	...	4

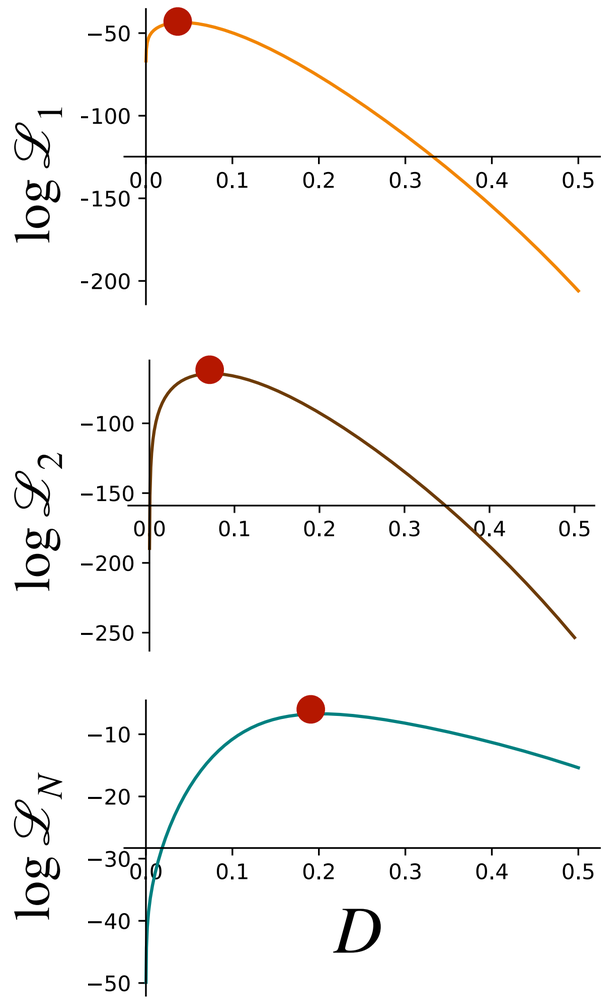
Independence assumption: treat q as a bag of independent *k*-mers (ignore overlap)

Likelihood of k-mer matches & Hamming distances

Hamming distance histograms

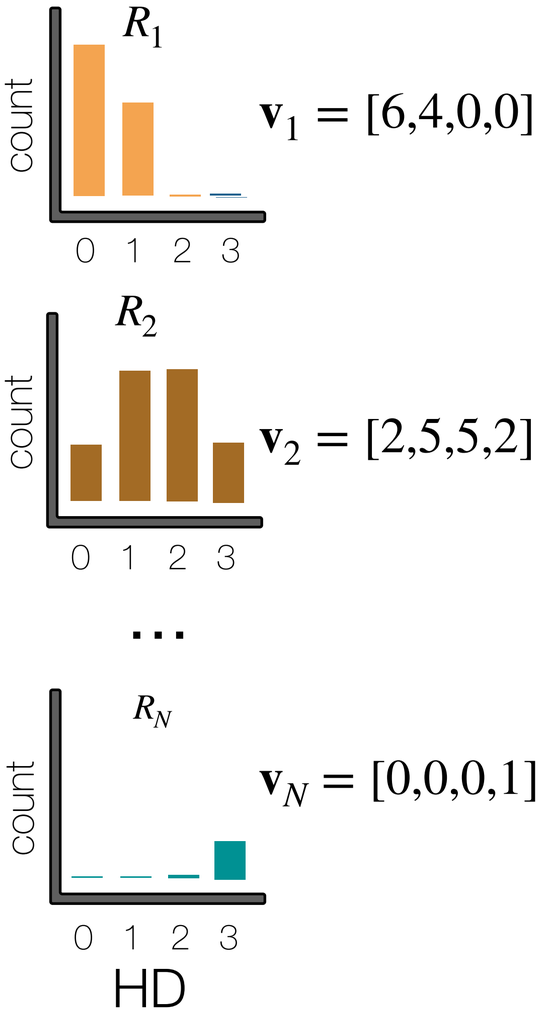


Goal: compute the likelihood of q having distance D to R_i



Likelihood of k-mer matches & Hamming distances

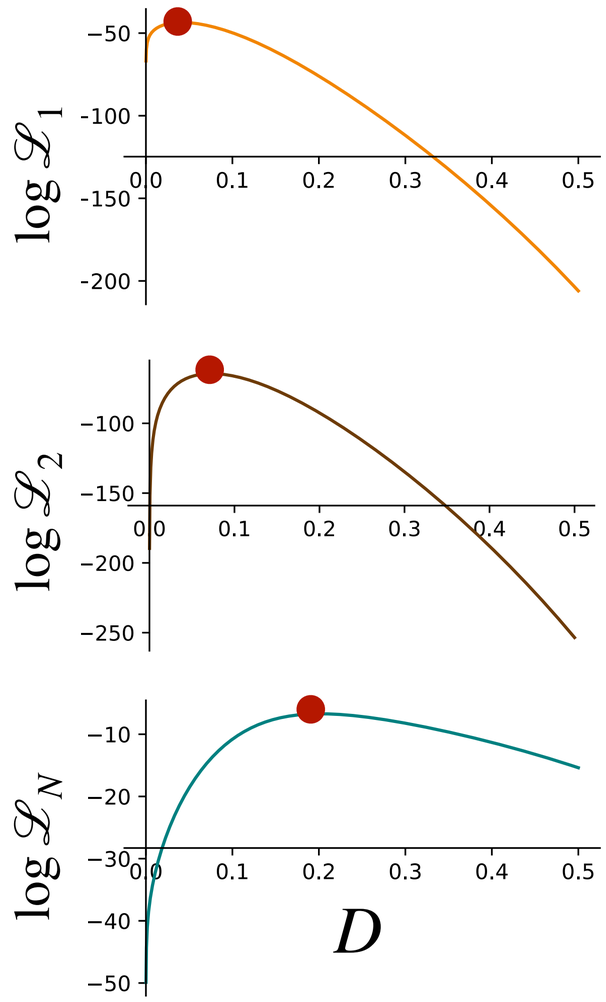
Hamming distance histograms



Goal: compute the likelihood of q having distance D to R_i

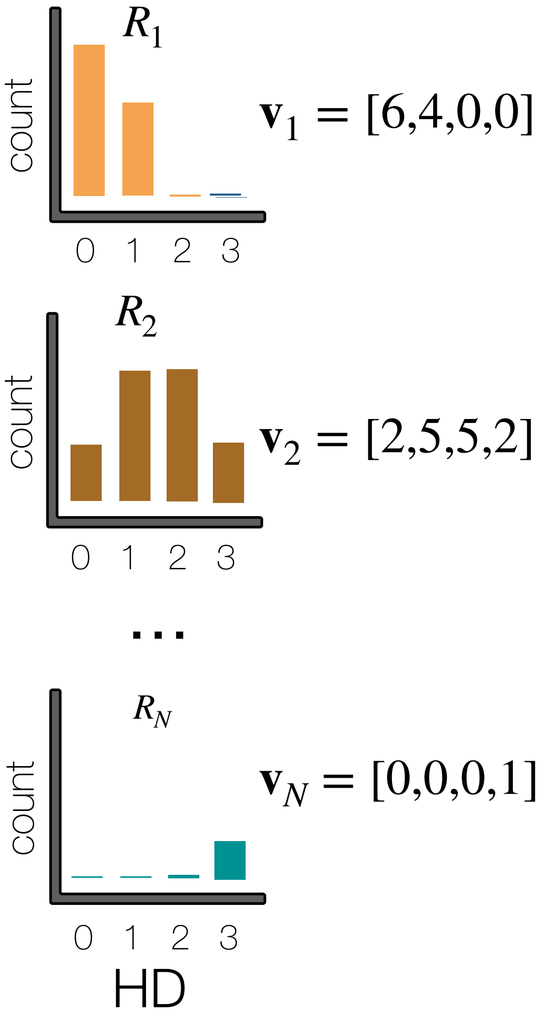
Likelihood of distance D to reference R_i : a product over all *k*-mers

$$\mathscr{L}_i(D; k, h, \delta, u_i, \mathbf{v}_i) =$$



Likelihood of k-mer matches & Hamming distances

Hamming distance histograms



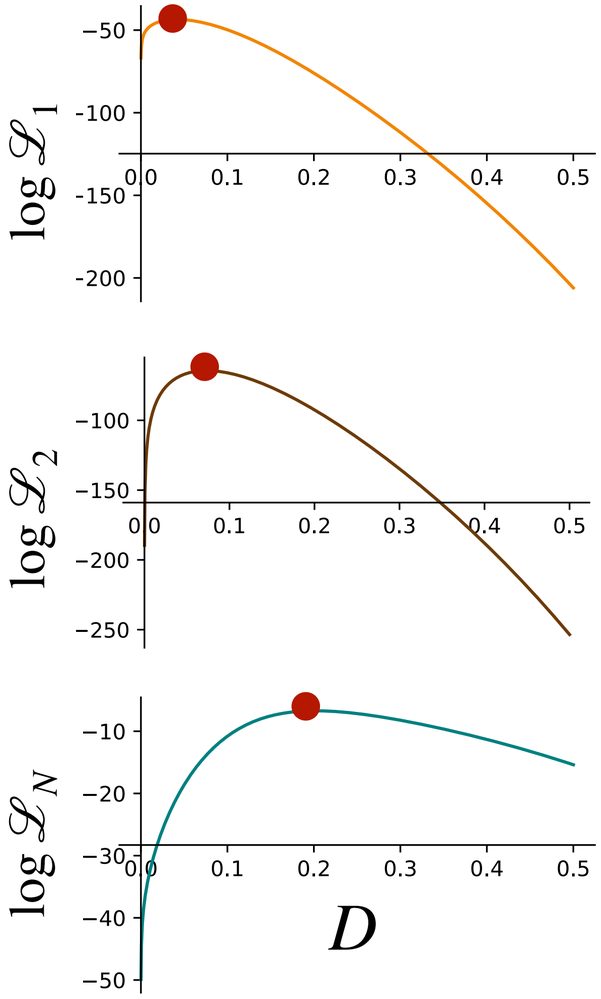
Goal: compute the likelihood of q having distance D to R_i

- $\mathbf{v}_i$: match count for each HD up to δ

Likelihood of distance D to reference R_i : a product over all *k*-mers

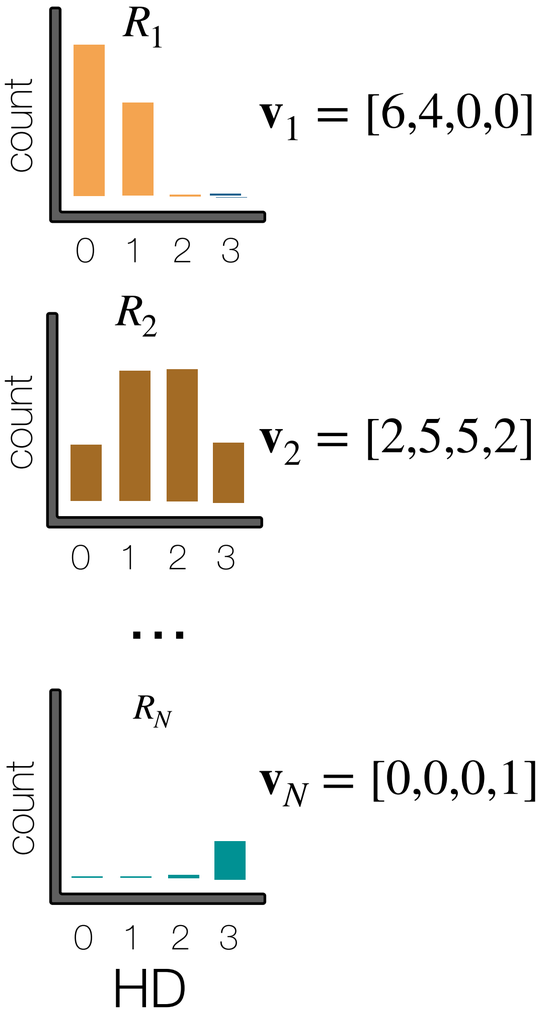
$$\mathscr{L}_i(D; k, h, \delta, u_i, \mathbf{v}_i) = \prod_{x=0}^{\delta} P_{match}(D; x, k, h)^{v_{i,x}}$$

Probability of having $v_{i,x}$ matches at HD $= x$



Likelihood of k-mer matches & Hamming distances

Hamming distance histograms



Goal: compute the likelihood of q having distance D to R_i

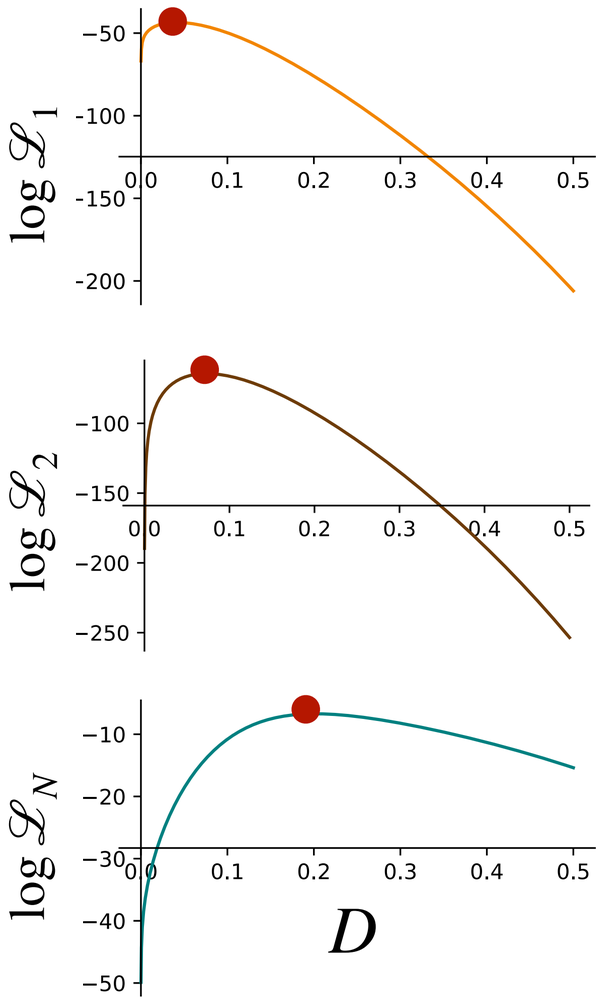
- $\mathbf{v}_i$: match count for each HD up to δ
- u_i: number of mismatches $(L-k+1) - \sum_{x=0}^{\delta} v_{i,x}$

Likelihood of distance D to reference R_i : a product over all *k*-mers

$$\mathscr{L}_i(D;k,h,\delta,u_i,\mathbf{v}_i) = P_{miss}(D;k,h,\delta)^{u_i} \prod_{x=0}^{\delta} P_{match}(D;x,k,h)^{v_{i,x}}$$

Probability of having u_i mismatches in total

Probability of having $v_{i,x}$ matches at HD $= x$



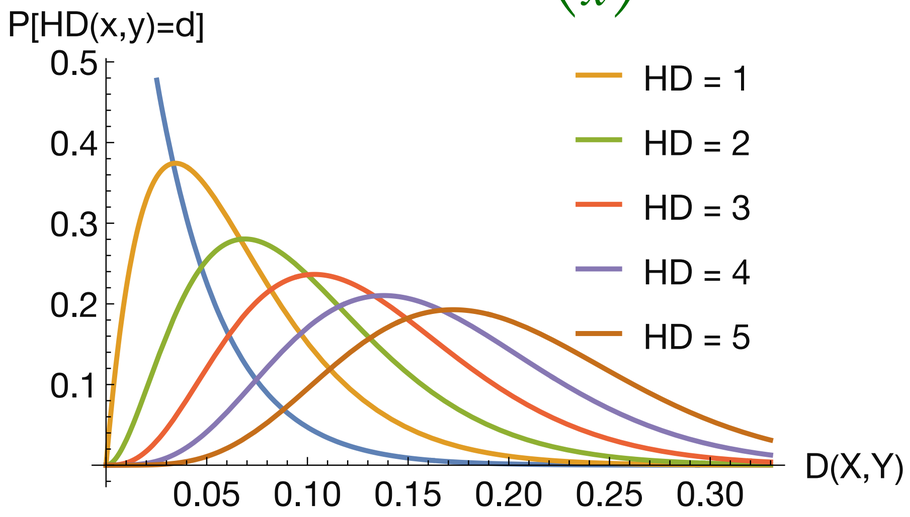
Observing k-mers matches with varying HDs

$$P_{match}(D; x, k, h) =$$

Observing k-mers matches with varying HDs

$$P_{match}(D; x, k, h) = \qquad P_{mutate}(D; x, k)$$

$$D^d(1 - D)^{(k-x)} \binom{k}{x}$$

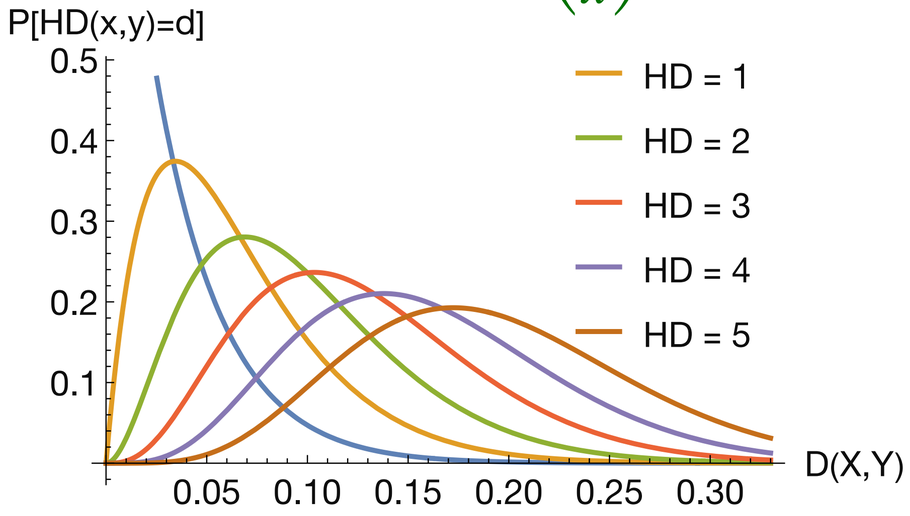


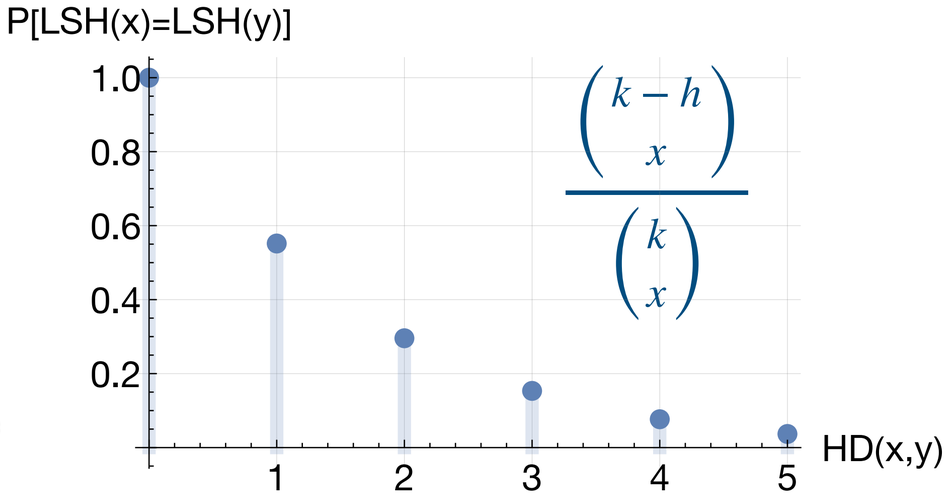
Observing k-mers matches with varying HDs

$$P_{match}(D; x, k, h) = P_{mutate}(D; x, k)\ P_{collide}(x, k, h)$$

$$D^d(1-D)^{(k-x)}\binom{k}{x}$$

1-FNR of LSH for HD $= x$





Observing k-mers matches with varying HDs

$$P_{match}(D; x, k, h) = \rho_i \; P_{mutate}(D; x, k) \; P_{collide}(x, k, h)$$

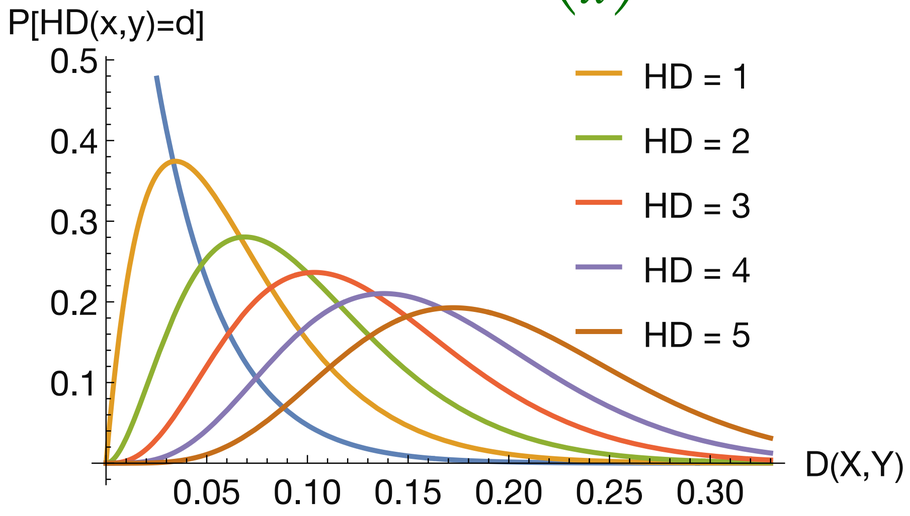
$$\rho_i = \frac{\text{\# of indexed}}{\text{\# of distinct}}$$

precomputed for R_i

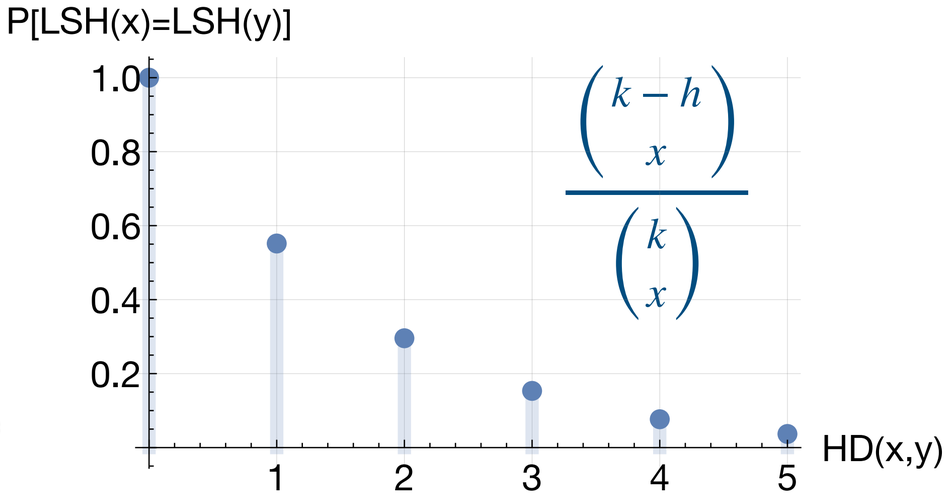
not all *k*-mers are indexed:

- minimizers
- FracMinHash
- …

$$D^d (1 - D)^{(k-x)} \binom{k}{x}$$



1-FNR of LSH for HD $= x$



Multiple events could lead to a mismatch

A mismatch occurs for a query *k*-mer a and reference (with ortholog k-mer b), iff

Multiple events could lead to a mismatch

A mismatch occurs for a query *k*-mer a and reference (with ortholog k-mer b), iff

- Reference is not indexed $(1 - \rho)$ ***or***

Multiple events could lead to a mismatch

A mismatch occurs for a query *k*-mer a and reference (with ortholog k-mer b), iff

- Reference is not indexed $(1 - \rho)$ ***or***
- Reference is indexed (ρ), but either:

i) $\mathrm{HD}(a, b) > \delta$: $\sum_{x=\delta+1}^{k} P_{mutate}(D; x, k)$ ***or***

ii) $\mathrm{HD}(a, b) \leq \delta$ ***and*** $\mathrm{LSH}(a) \neq \mathrm{LSH}(b)$: $\sum_{x=0}^{\delta} P_{mutate}(D; x, k)(1 - P_{collide}(x, k, h))$

Multiple events could lead to a mismatch

A mismatch occurs for a query *k*-mer a and reference (with ortholog k-mer b), iff

- Reference is not indexed $(1-\rho)$ ***or***
- Reference is indexed (ρ), but either:

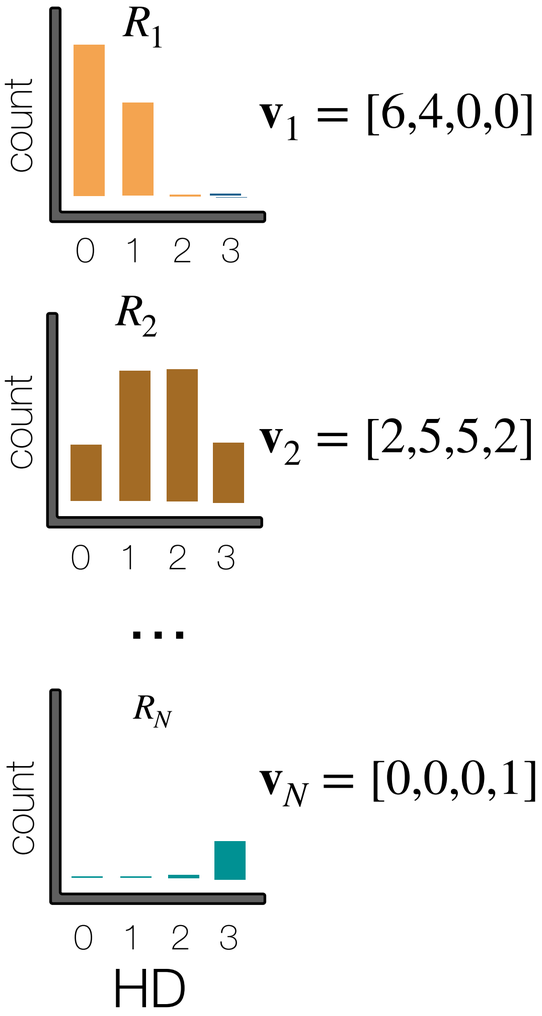
i) $\mathrm{HD}(a,b) > \delta$: $\sum_{x=\delta+1}^{k} P_{mutate}(D;x,k)$ ***or***

ii) $\mathrm{HD}(a,b) \leq \delta$ ***and*** $\mathrm{LSH}(a) \neq \mathrm{LSH}(b)$: $\sum_{x=0}^{\delta} P_{mutate}(D;x,k)(1-P_{collide}(x,k,h))$

$$P_{miss}(D;x,k,h,\delta) = (1-\rho) + \rho\left(\sum_{x=\delta+1}^{k} P_{mutate}(D;x,k) + \sum_{x=0}^{\delta} P_{mutate}(D;x,k)(1-P_{collide}(x,k,h))\right)$$

Maximum likelihood estimation of distances

Hamming distance histograms

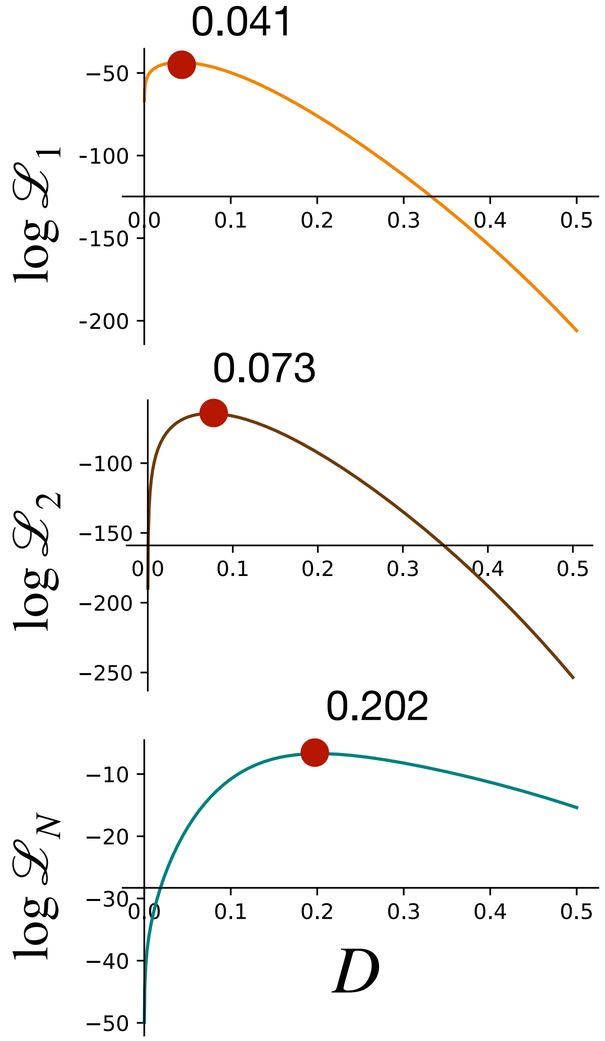


Optimize $\log \mathscr{L}_i$ w.r.t. D

for each (relevant) reference R_i

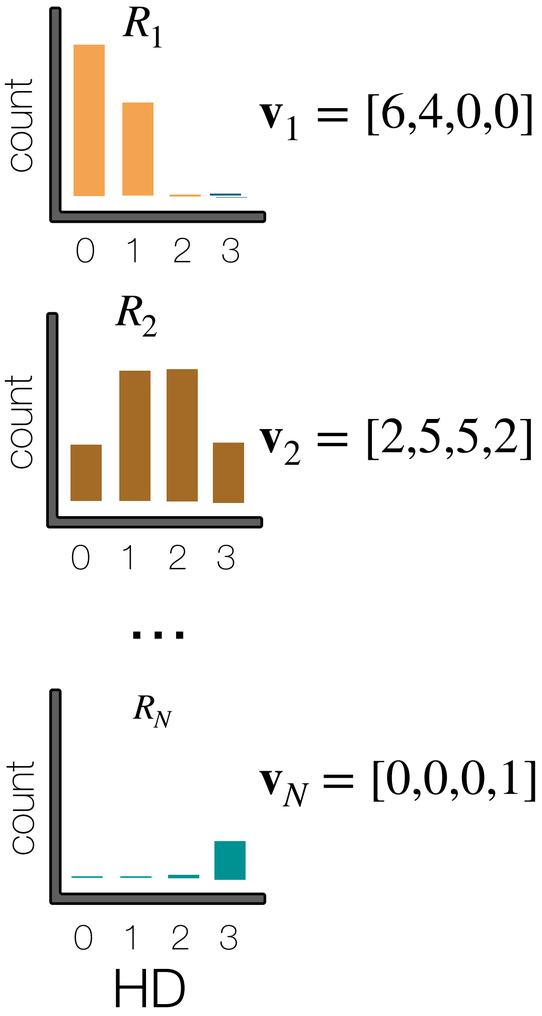
$$\arg\max_D \; P_{miss}(D;k,h,\delta)^{u_i} \prod_{x=0}^{\delta} P_{match}(D;x,k,h)^{v_{i,x}}$$

single variable & **convex** with a sensible choice of parameters

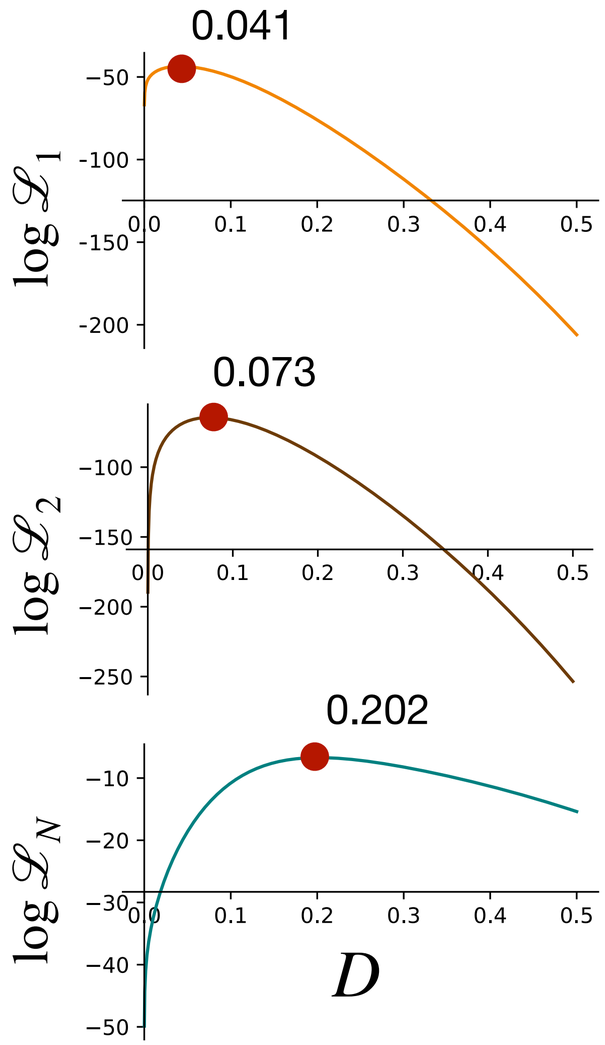


Maximum likelihood estimation of distances

Hamming distance histograms



Are maximum likelihood distances accurate?



krepp estimates distances accurately at the read-level

default: 29-mer
minimizers of 35-mers

- Simulation experiments
  (true read distances)

~150 bp short reads

(Hamming distance) / (seq. length)

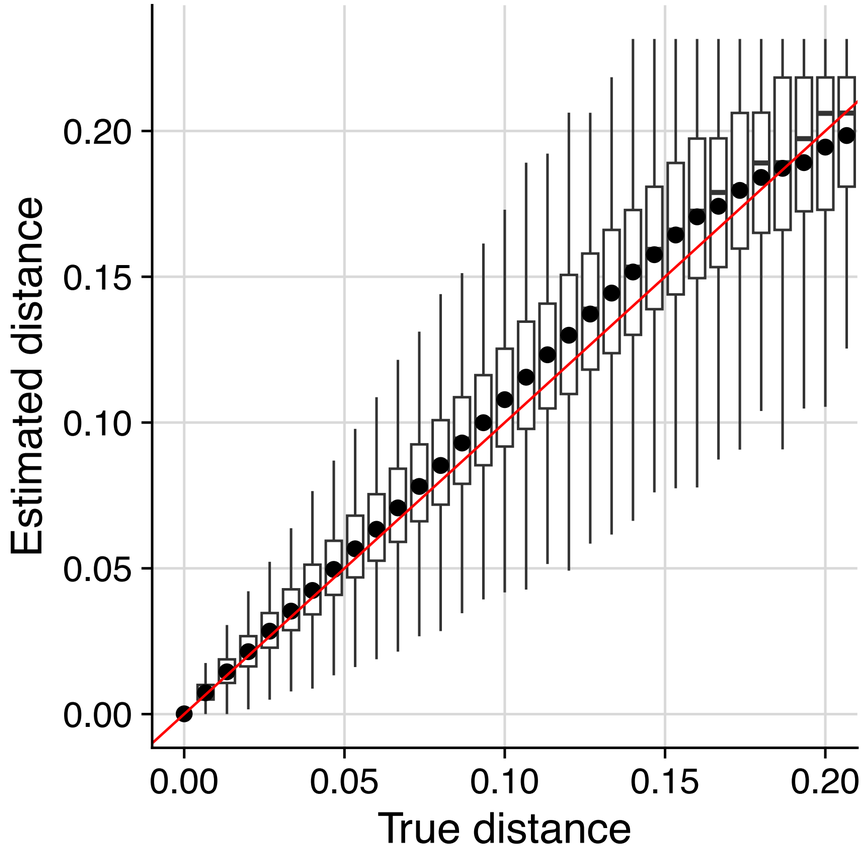
krepp estimates distances accurately at the read-level

default: 29-mer
minimizers of 35-mers

- Simulation experiments
(true read distances)
- Highly accurate
(despite some noise)
- Slight overestimation
bias for high distances

~150 bp short reads

(Hamming distance) / (seq. length)



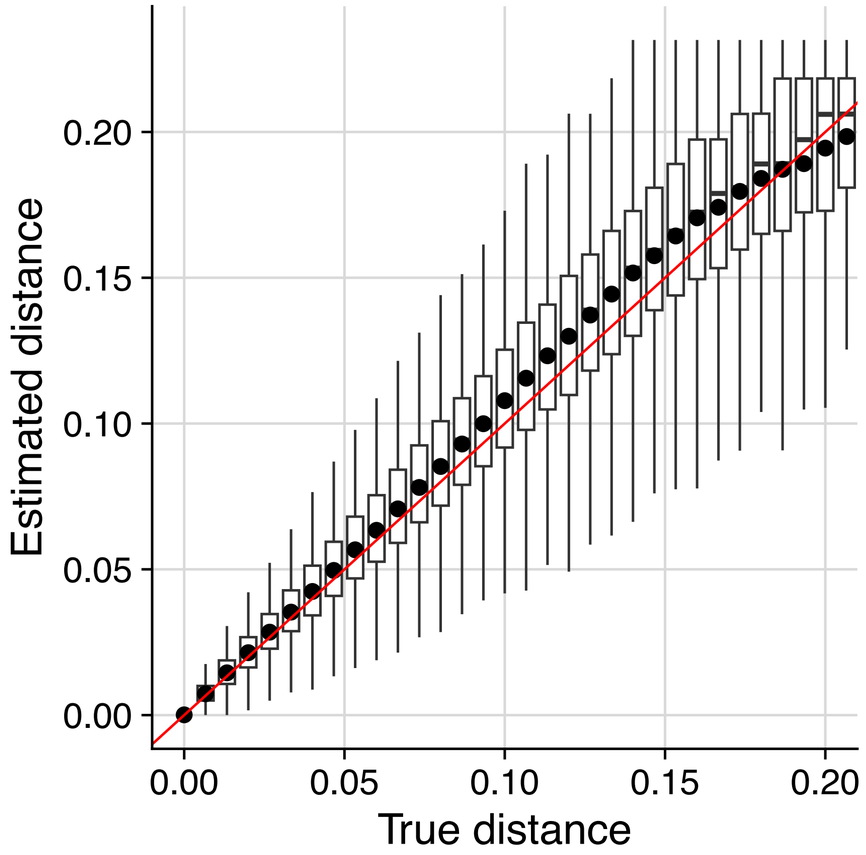
krepp estimates distances accurately at the read-level

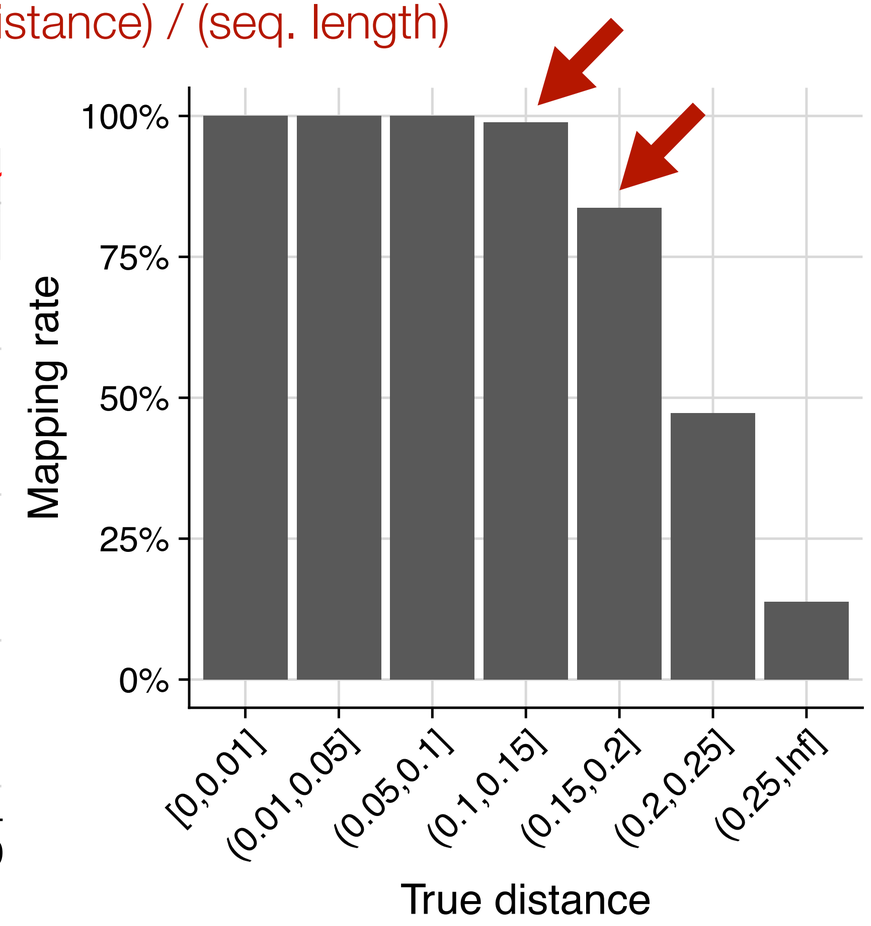
default: 29-mer
minimizers of 35-mers

- Simulation experiments (true read distances)
- Highly accurate (despite some noise)
- Slight overestimation bias for high distances
- High mapping rate even for novel reads >15%

~150 bp short reads

(Hamming distance) / (seq. length)

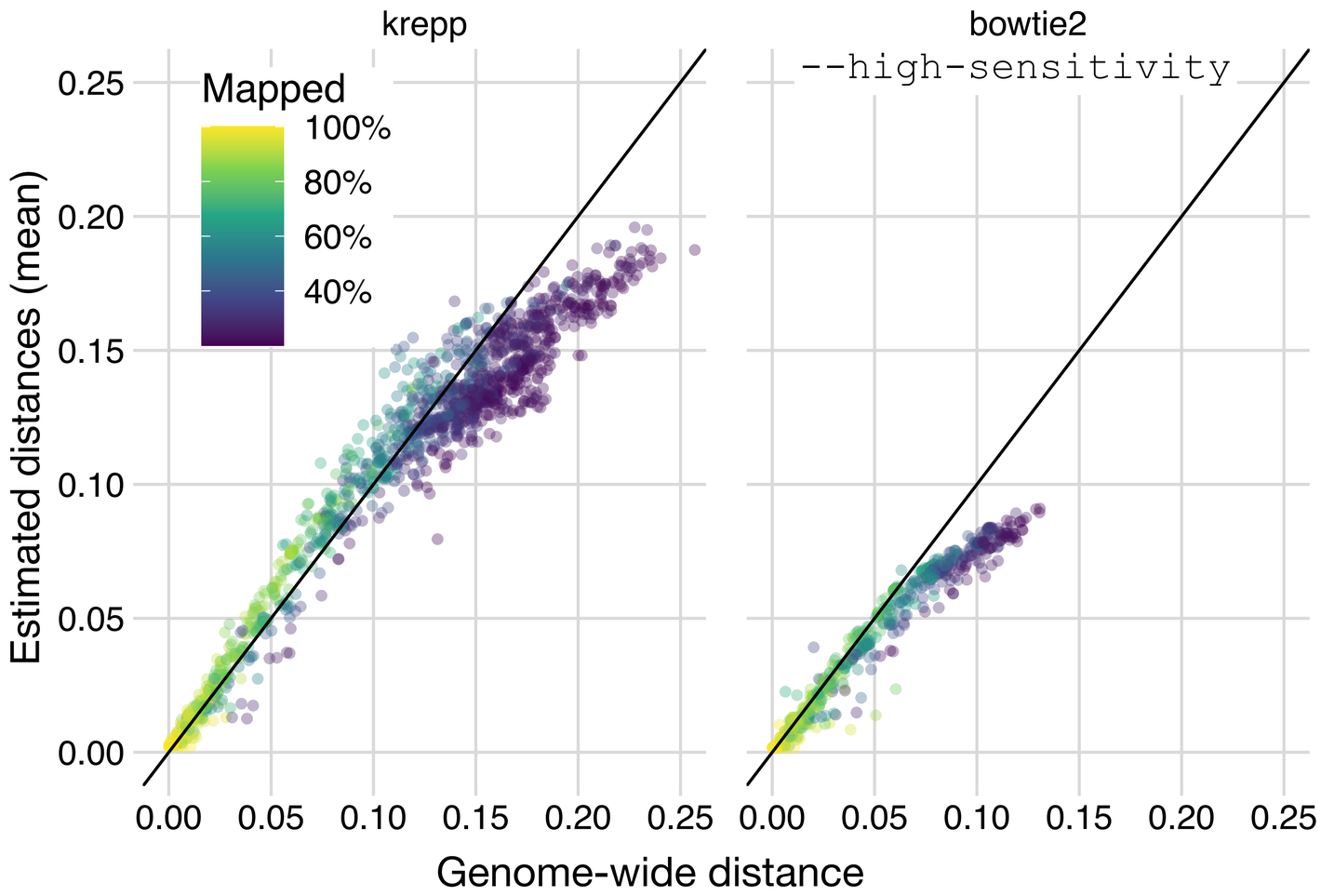




krepp matches nucleotide identity on average for real genomes

Index: Web of Life (v2)
16,000 microbial genomes

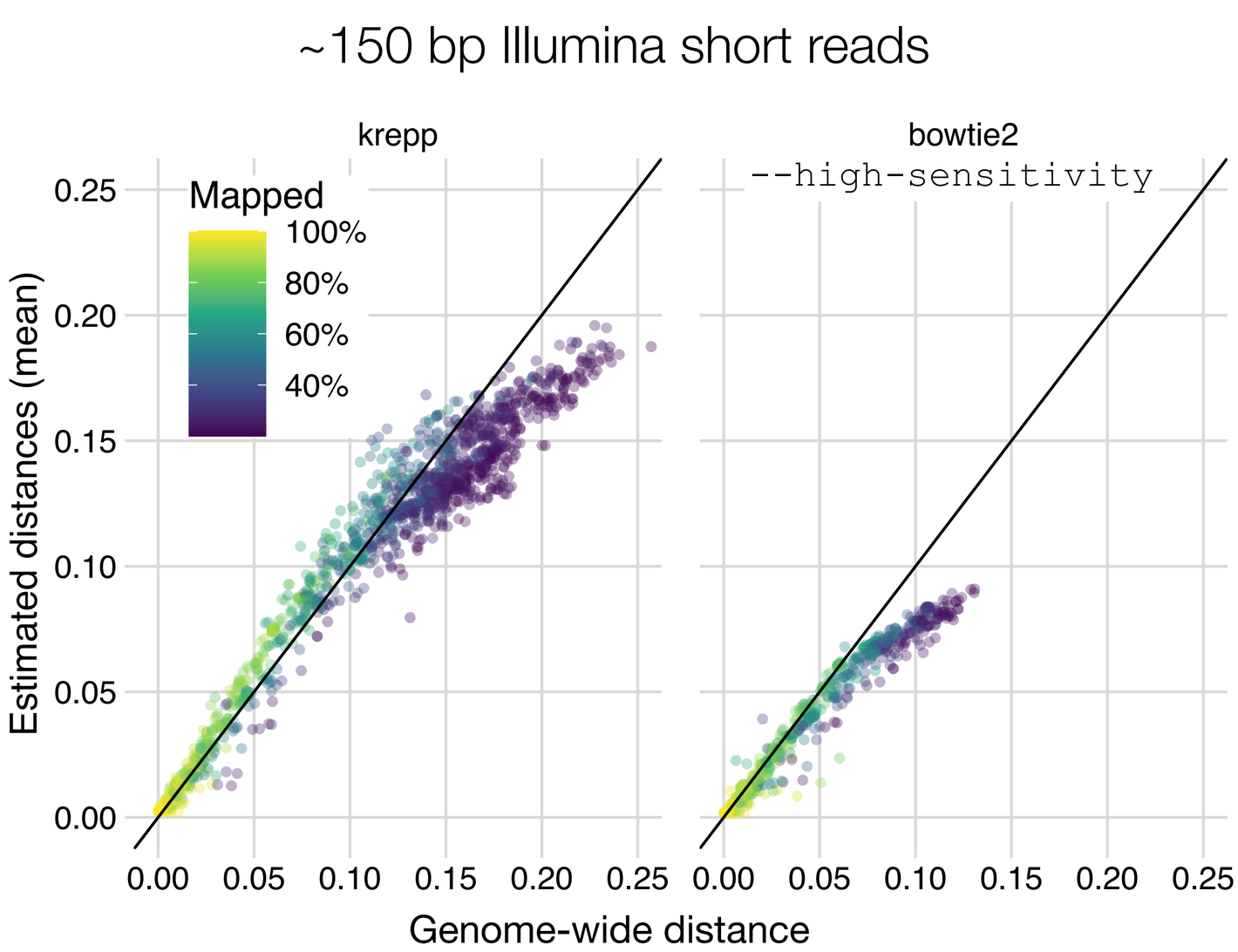
- Real query/reference genomes
  (pairs with >20% mapping rate)



krepp matches nucleotide identity on average for real genomes

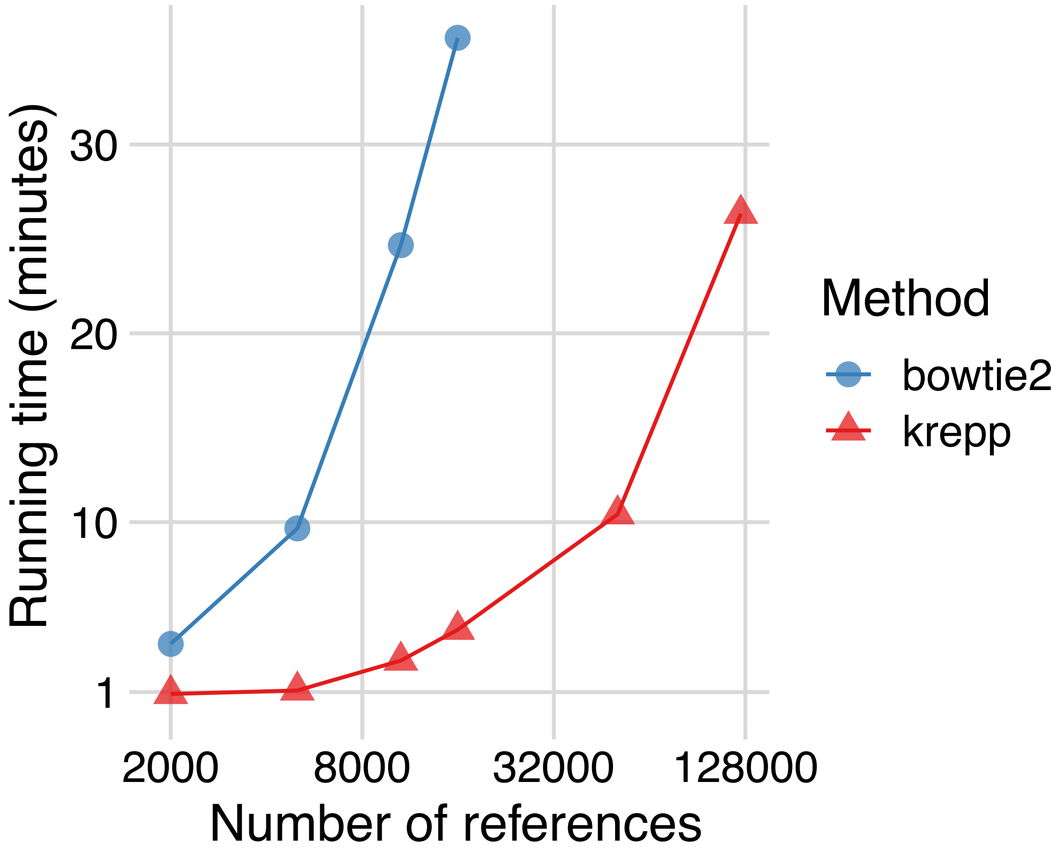
Index: Web of Life (v2)
16,000 microbial genomes

- Real query/reference genomes (pairs with >20% mapping rate)
- *krepp* extends to distant (>10%) reference genomes accurately



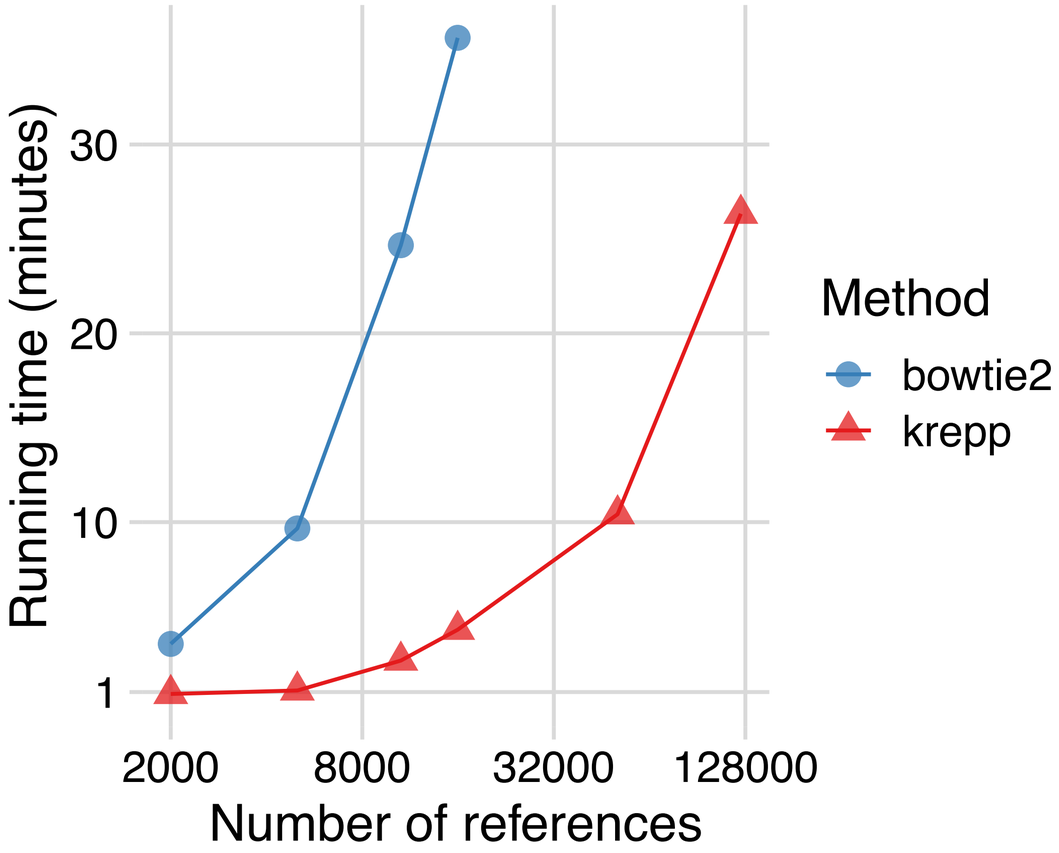
Scalability:
Avoiding alignment & effective parallelization

Mapping 10M reads (16 threads):

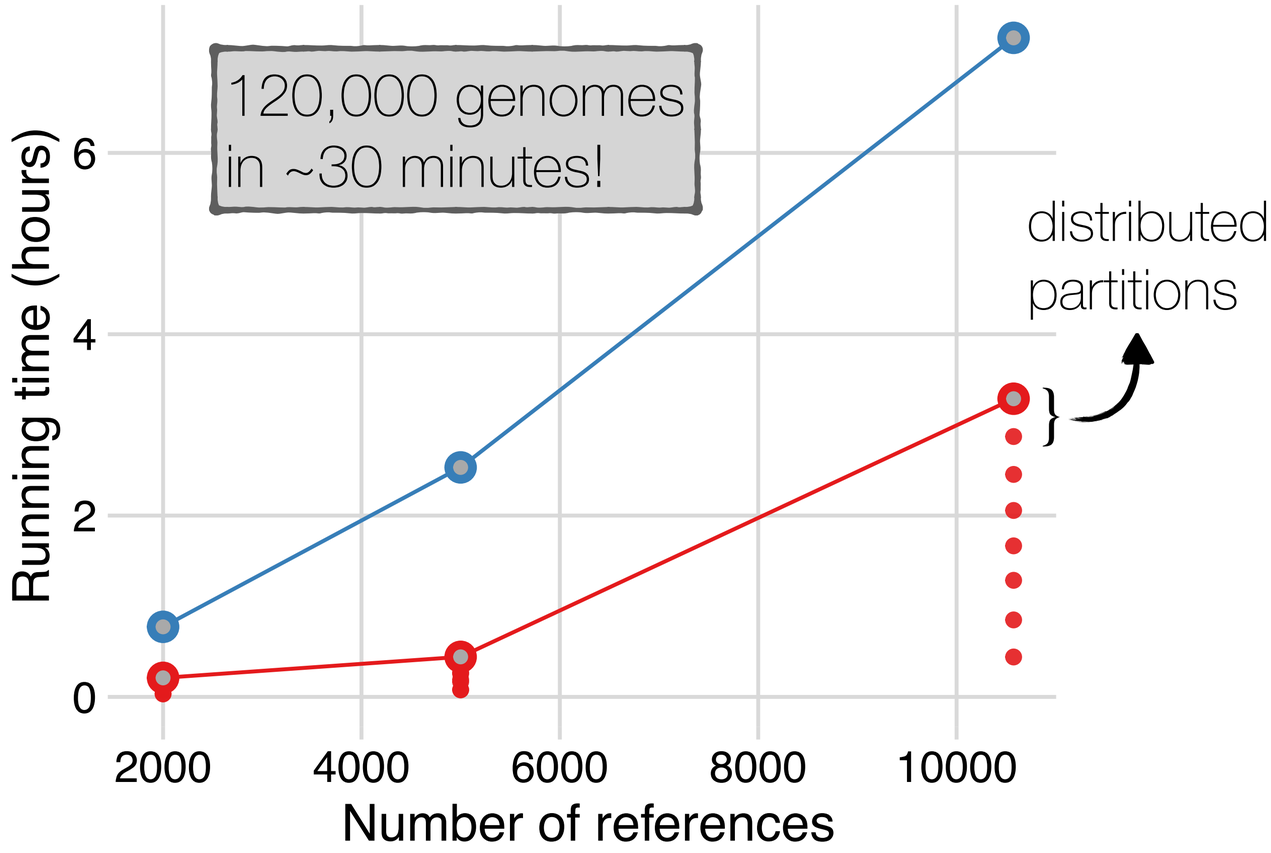


Scalability:
Avoiding alignment & effective parallelization

Mapping 10M reads (16 threads):



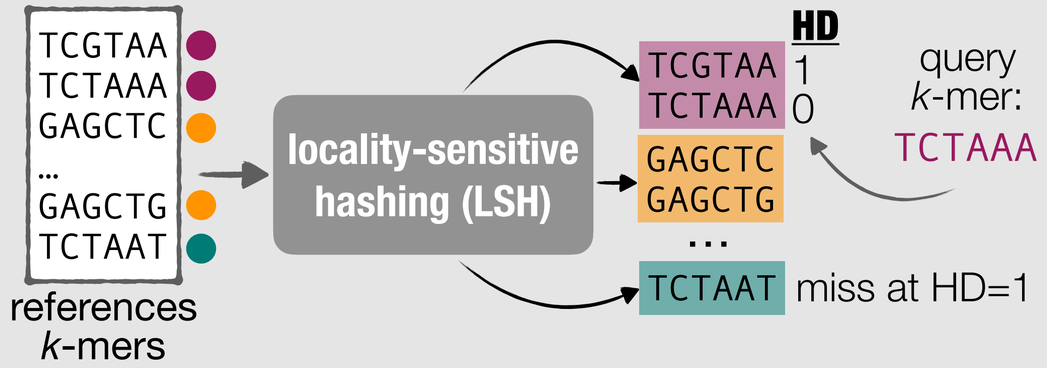
Indexing microbial genomes (32 threads):



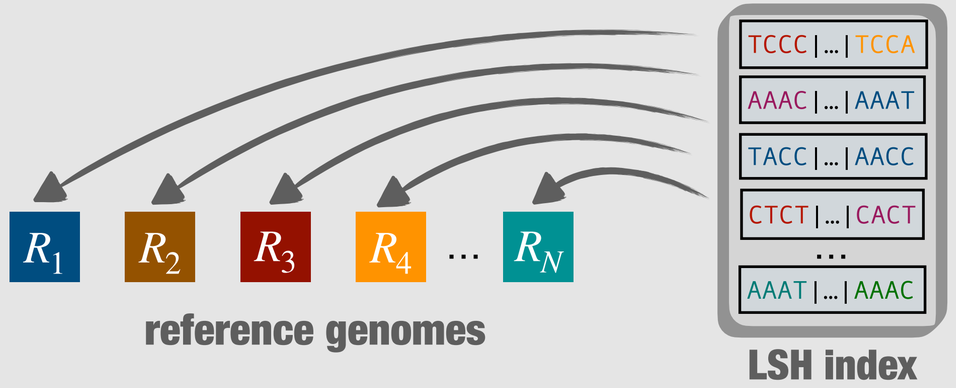
Outline of the method & subproblems we need to tackle

krepp: k-mer-based read phylogenetic placement

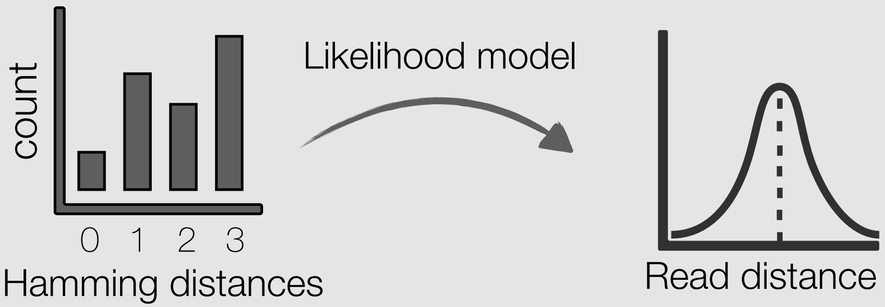
I) Hamming distances of homologous k-mers



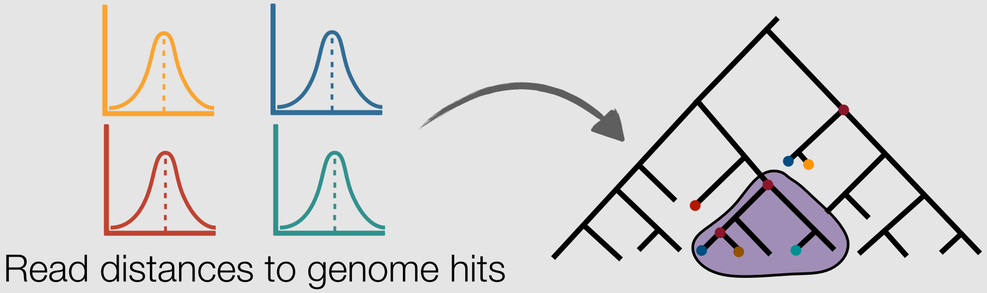
II) Mapping indexed k-mers to genomes



III) Estimating read distances based on likelihood of k-mer matches



IV) Placement on the reference phylogeny by modeling uncertainties of distances

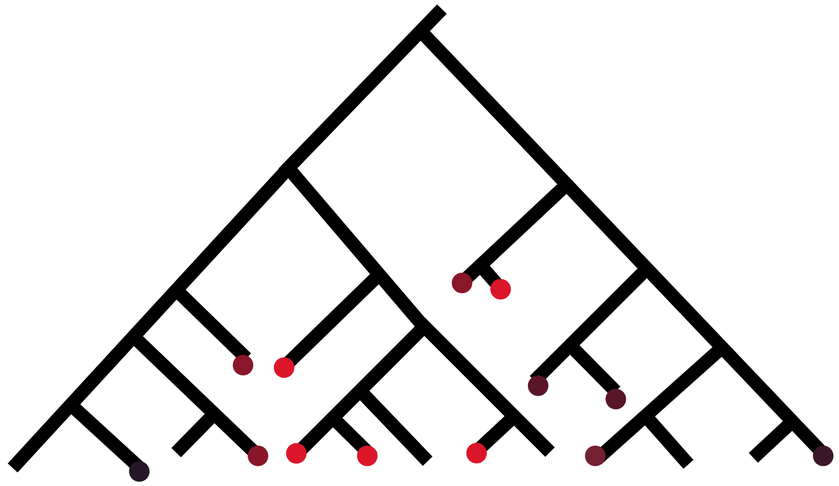


Problem 4: distance-based placement

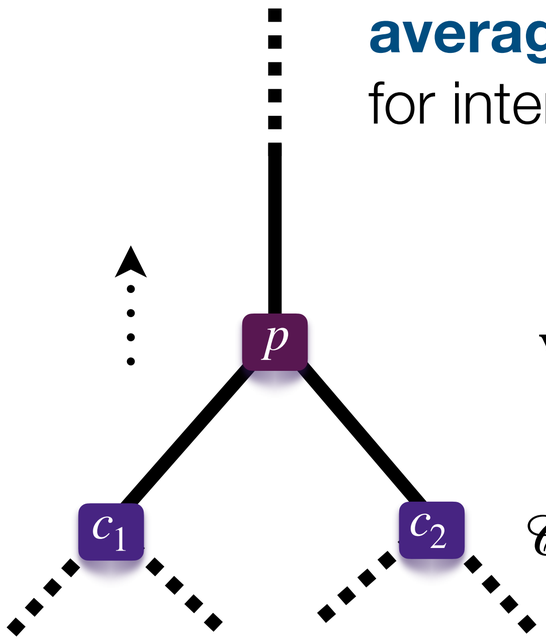
Given $d(q, R_i)$ for many R_is, find the "best" placement of q on T

Challenges:

- Short reads — **low signal**
- Only have distances to leaves, **not internal nodes**
- Distances may be in a **different unit** than the branch lengths of the reference tree T
- Small differences in distances may not be meaningful (**statistical distinguishability**)



Defining a notion of a distance to a clade

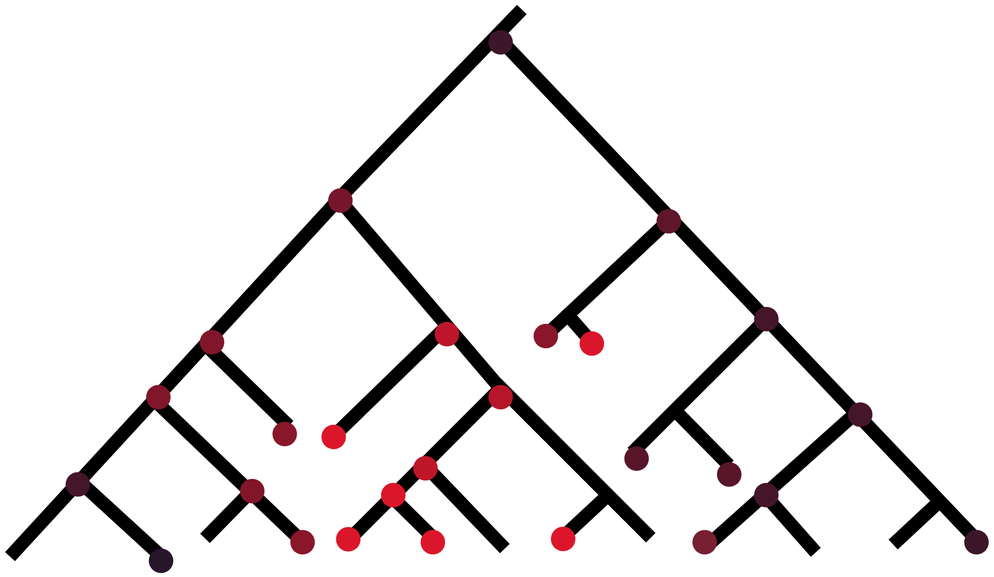


recursively compute
average HD histograms
for internal nodes

$$\mathbf{v}_p = \frac{\sum_{c \in \mathscr{C}(p)} \mathbf{v}_c}{|\mathscr{C}(p)|}$$

$\mathscr{C}(p)$: set of children of p, $\{c_1, c_2\}$

use the same likelihood model



Statistical distinguishability tests → placement

- Small differences may not be statistically meaningful
  - **test distinguishability**

Statistical distinguishability tests → placement

- Small differences may not be statistically meaningful
  - **test distinguishability**

likelihood-ratio test
with the closest reference:

D: alternative distance

$$\lambda_{LR} = \frac{\mathscr{L}_{i*}(D; k, h, \delta, u_{i*}, \mathbf{v}_{i*})}{\mathscr{L}_{i*}(D^*; k, h, \delta, u_{i*}, \mathbf{v}_{i*})}$$

$i*$: closest reference

$$\lambda_{LR} \sim \chi^2$$

- select a significance level (default: α=90%)

Statistical distinguishability tests → placement

- Small differences may not be statistically meaningful
  - **test distinguishability**

likelihood-ratio test
with the closest reference:

place on the largest clade that is indistinguishable with the minimum distance

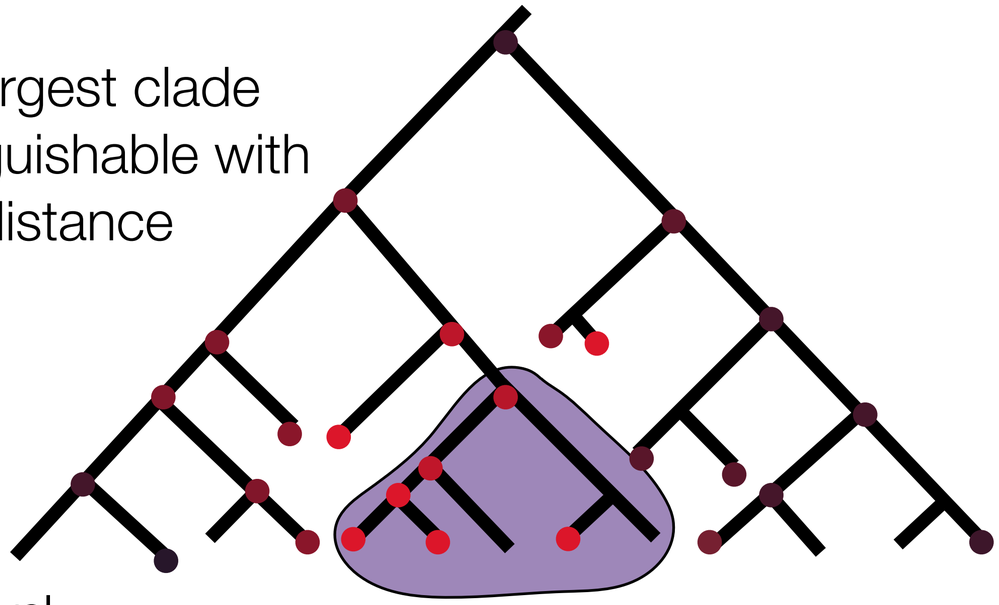
D: alternative distance

$$\lambda_{LR} = \frac{\mathscr{L}_{i*}(D; k, h, \delta, u_{i*}, \mathbf{v}_{i*})}{\mathscr{L}_{i*}(D^*; k, h, \delta, u_{i*}, \mathbf{v}_{i*})}$$

$i*$: closest reference

$\lambda_{LR} \sim \chi^2$

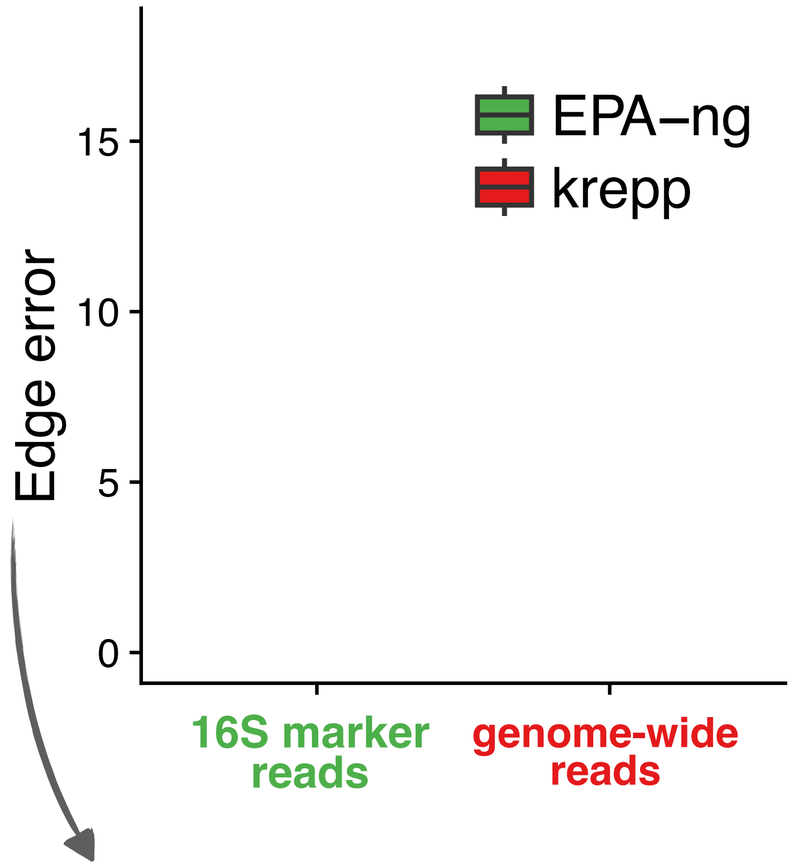
- select a significance level (default: α=90%)



indistinguishable w.r.t. the closest reference

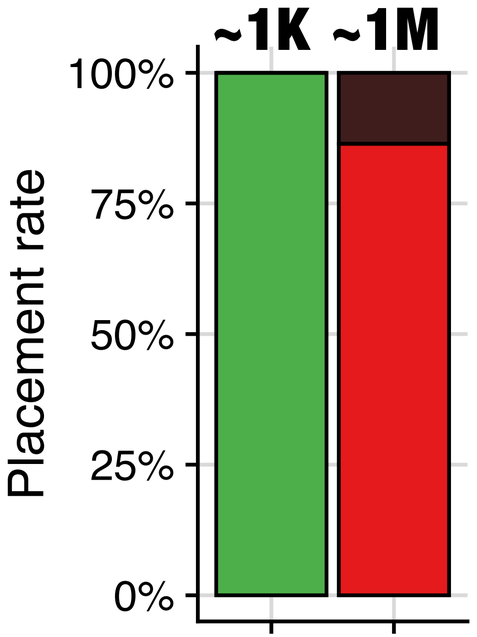
krepp places genome-wide reads more accurately than ML-based of 16S placement

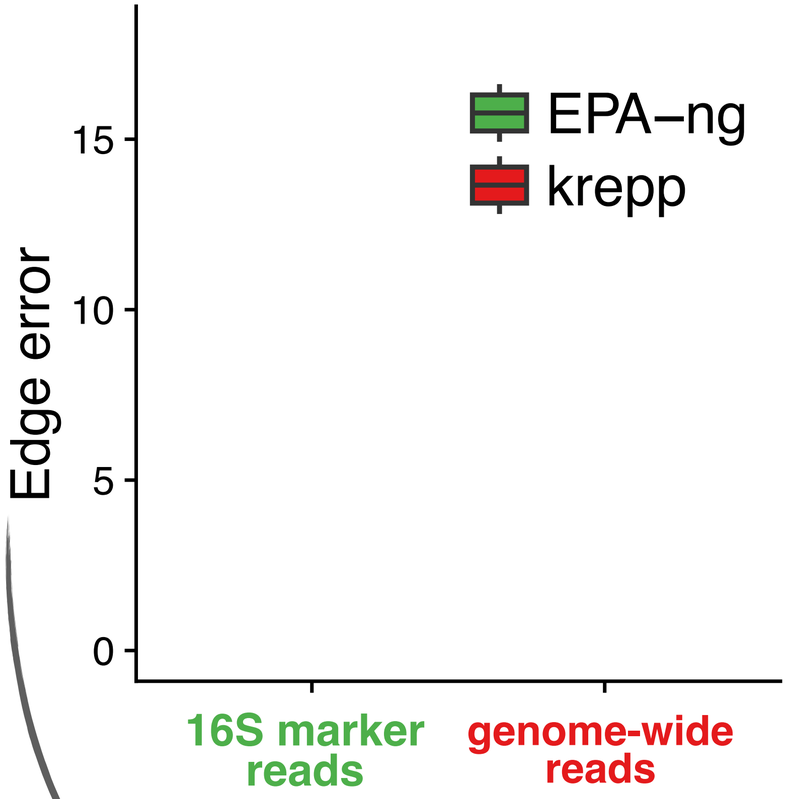
- Leave all out —100 queries from 10,500 taxa (WoLv1)
- EPA-ng: needs a MSA; only markers



krepp places genome-wide reads more accurately than ML-based of 16S placement

- Leave all out —100 queries from 10,500 taxa (WoLv1)
- EPA-ng: needs a MSA; only markers
- ***krepp*** **places 86% of all reads**

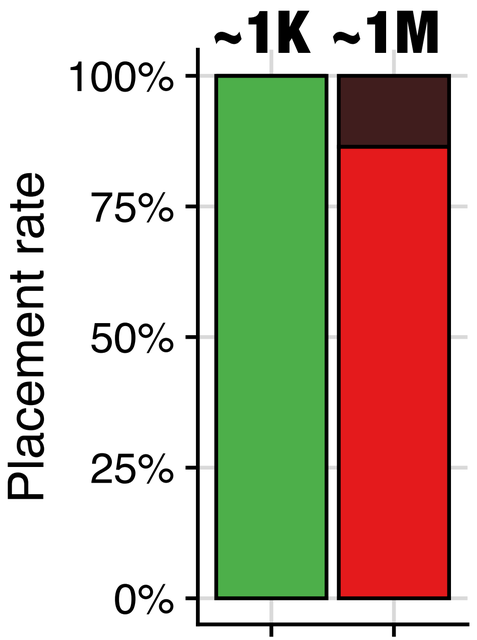


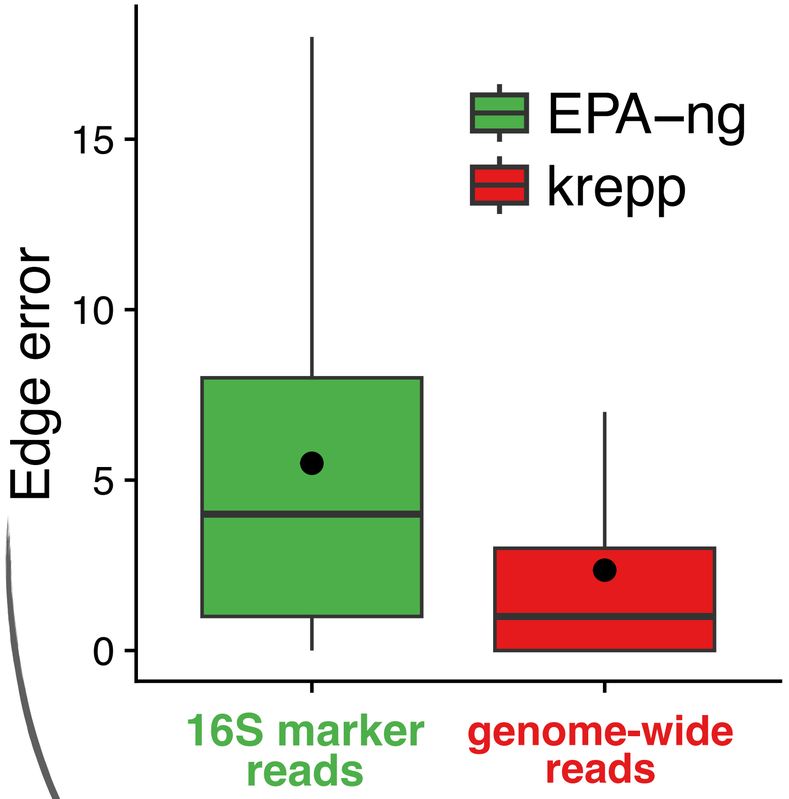


how many edges away is the placement from the correct edge?

krepp places genome-wide reads more accurately than ML-based of 16S placement

- Leave all out —100 queries from 10,500 taxa (WoLv1)
- EPA-ng: needs a MSA; only markers
- ***krepp* places 86% of all reads**
- **2.4 vs. 5.6 edge error (average)**

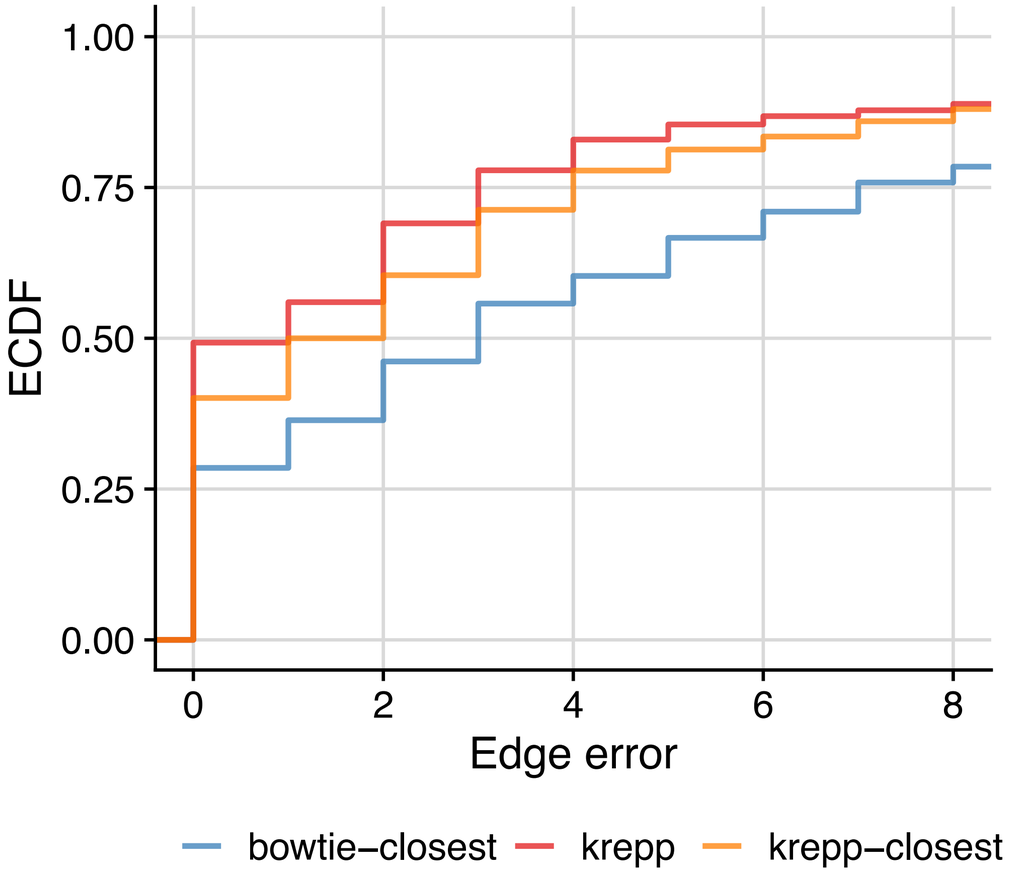




how many edges away is the placement from the correct edge?

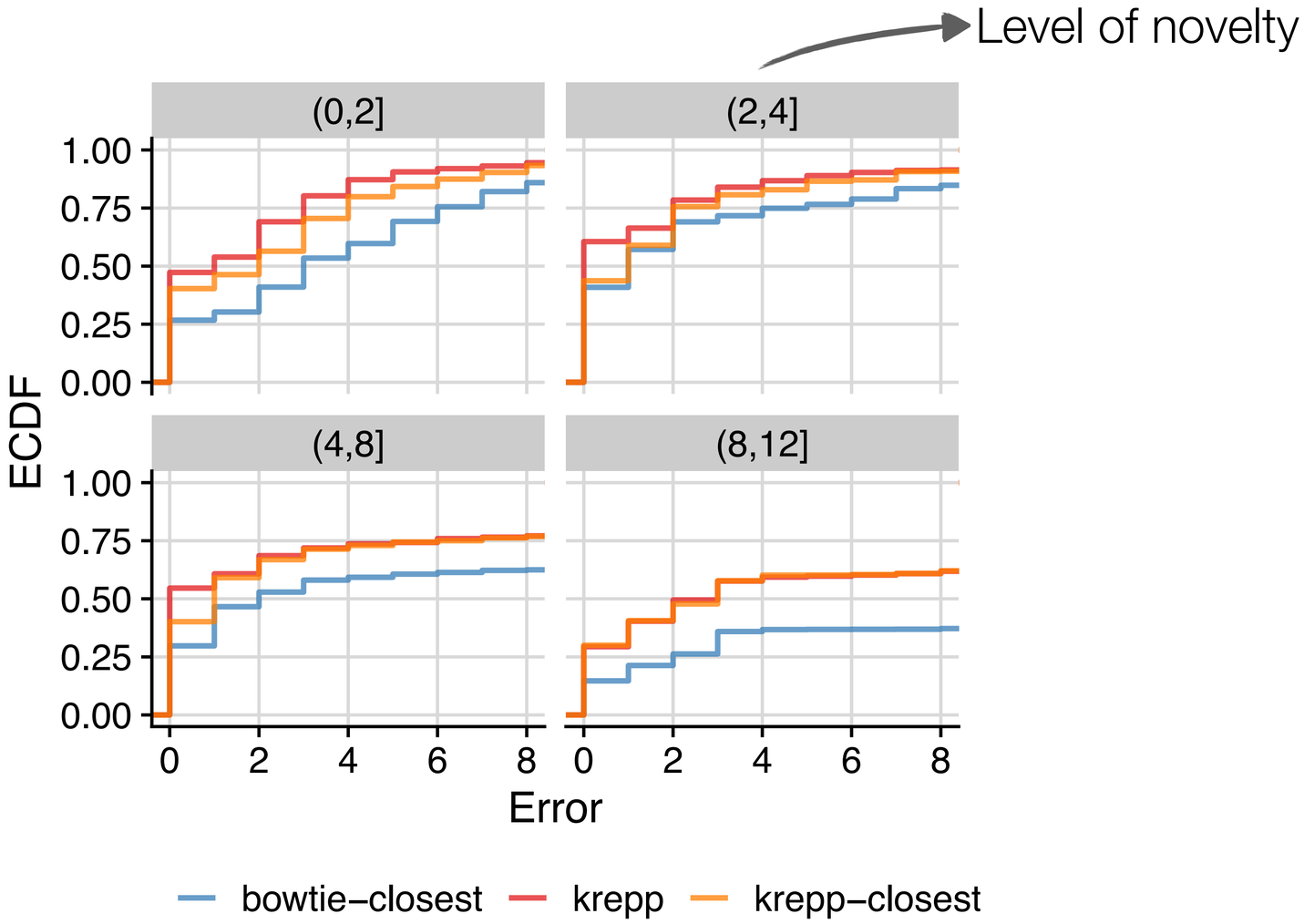
krepp's heuristic improves closest-tip placement

- Leave one out: 100/16,000 (WoLv2)
- Outperforms baselines:
  on the closest, on the LCA, etc.
- >80% of all reads within four edges



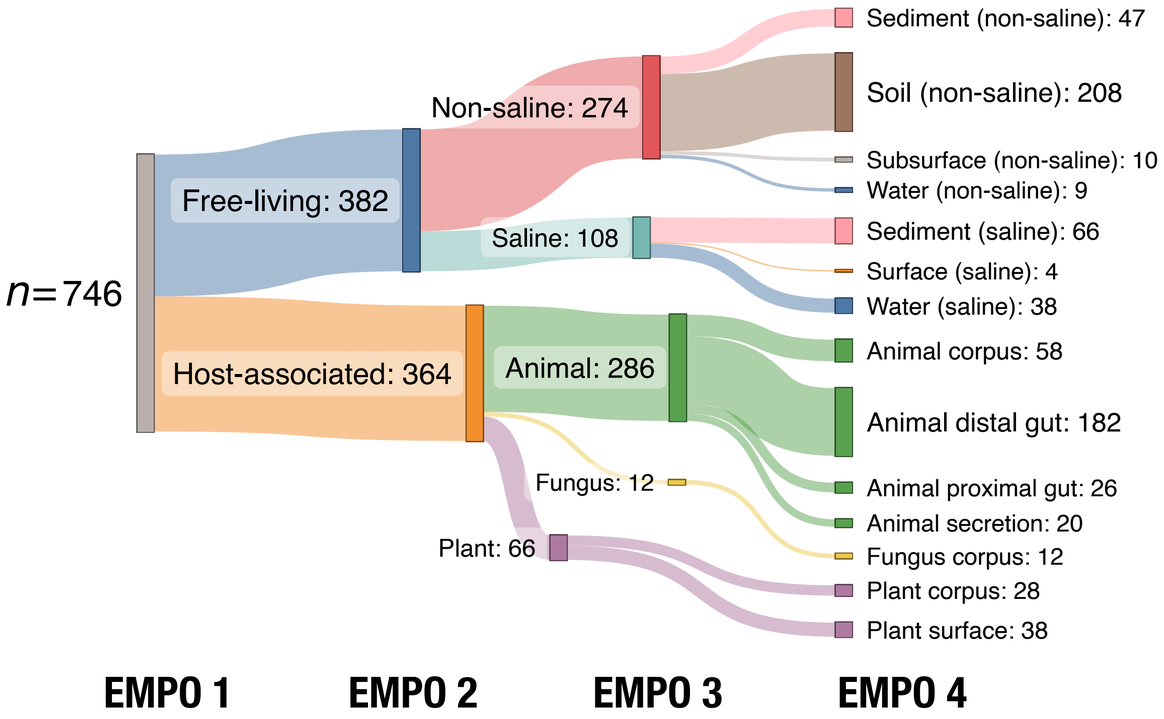
krepp's heuristic improves closest-tip placement

- Leave one out: 100/16,000 (WoLv2)
- Outperforms baselines:
  on the closest, on the LCA, etc.
- >80% of all reads within four edges

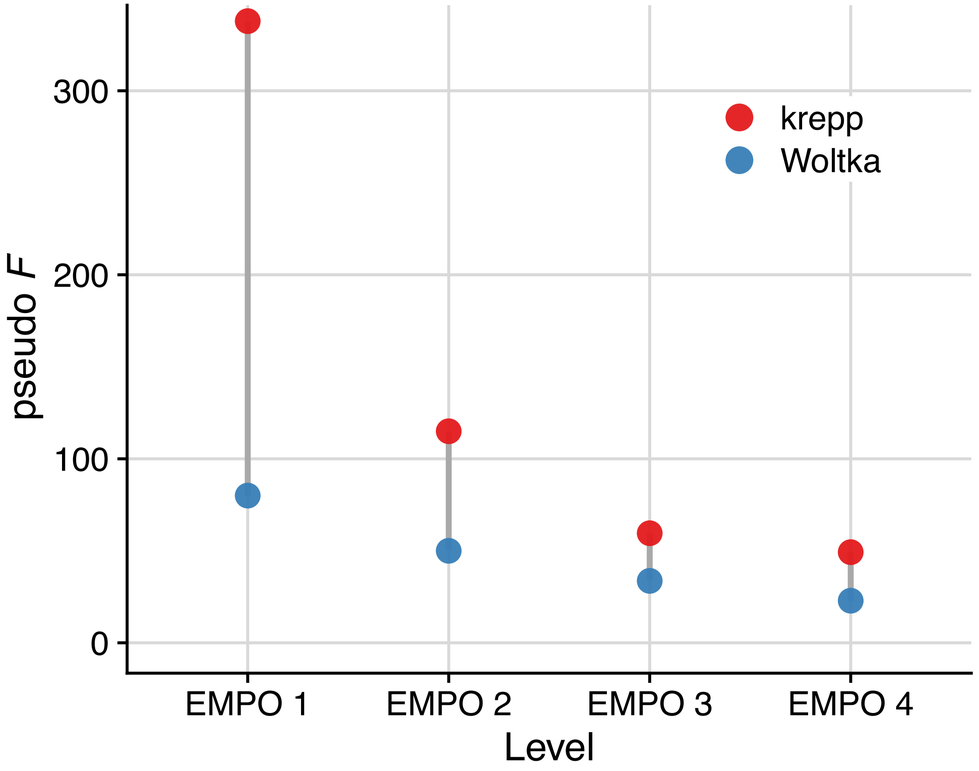


Better characterization of less-studied microbiome of earth

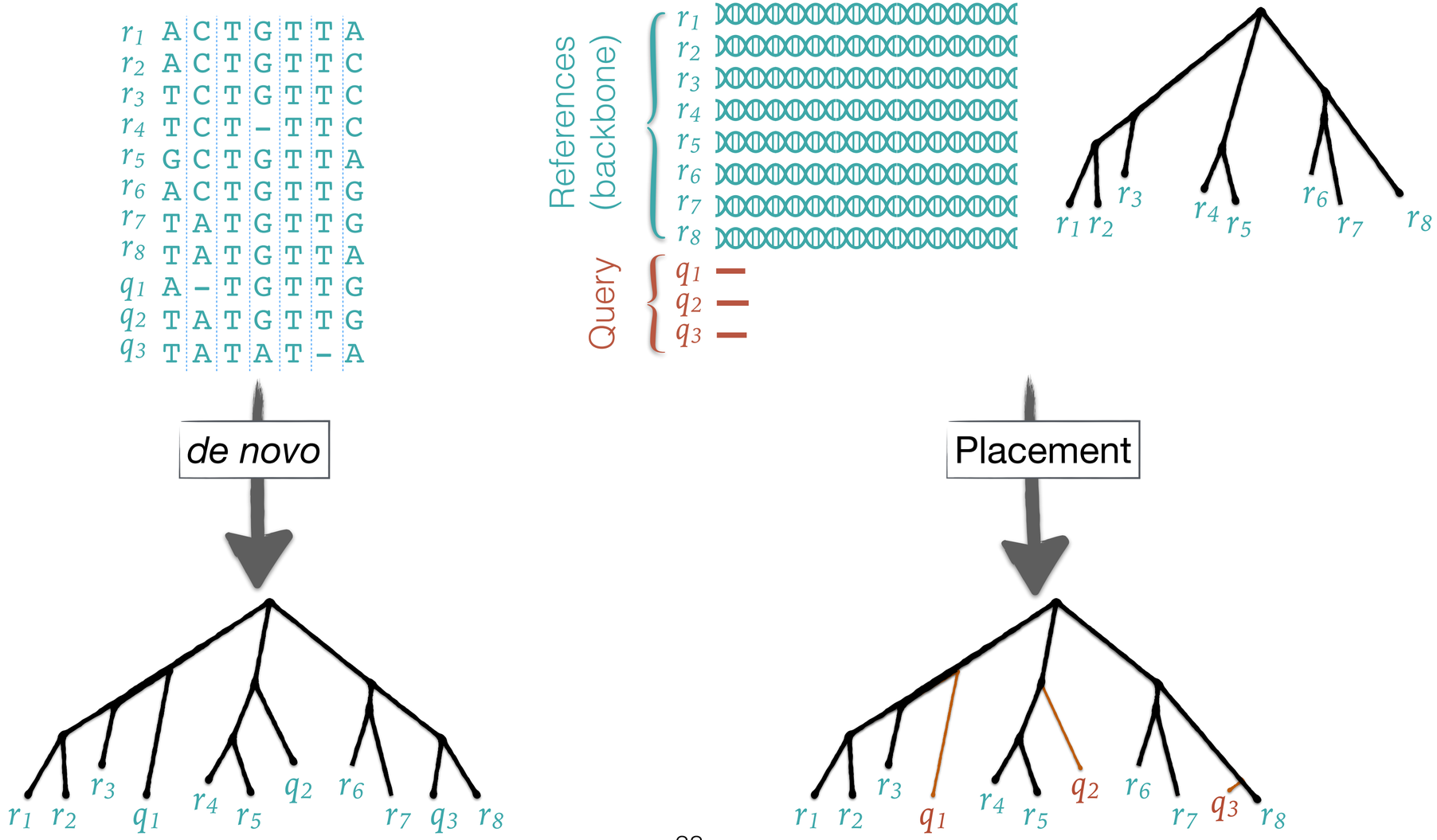
Hierarchical categorization
of earth microbiome samples



Reference: Web of Life (v1)
11,000 microbial genomes



Krepp: Phylogenetic placement (PP) of all reads



Software

software: github.com/bo1929/krepp

preprint: github.com/bo1929/krepp

Ali Osman Berk Şapcı

Many applications in ecology do not fit the marker-based alignment-based model

Place every read from anywhere on the genome (not just a handful of marker genes) on a reference tree of reference genomes

Many applications in ecology do not fit the marker-based alignment-based model

Place every read from anywhere on the genome (not just a handful of marker genes) on a reference tree of reference genomes

Place a genome skim (i.e., a bag of randomly sampled low-coverage reads) on a reference tree of reference genomes (or skims)

Place a long read or assembled contig on a reference tree of reference genomes (or skims)

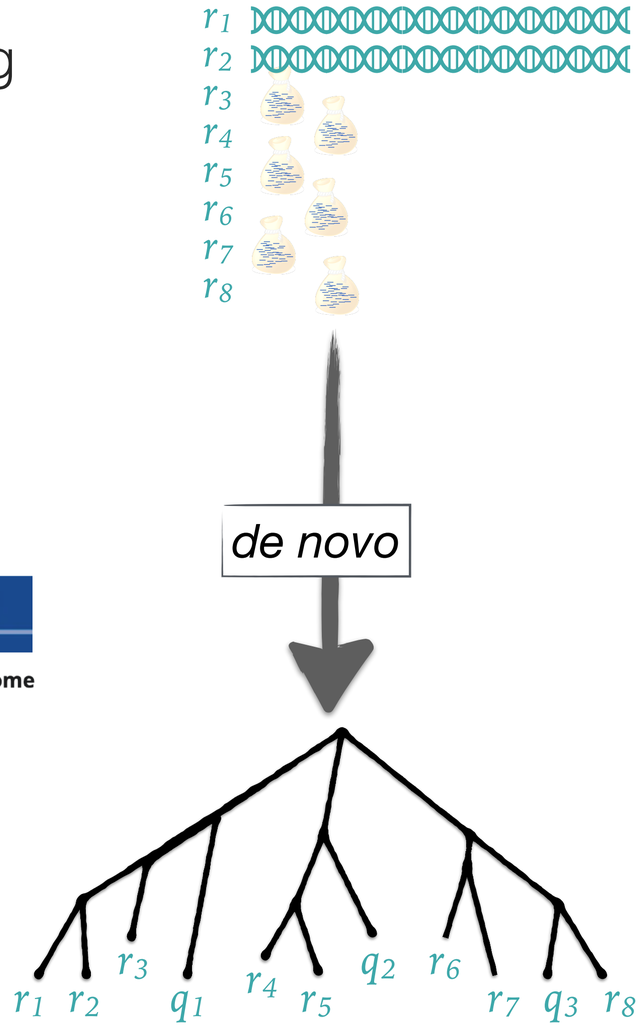
Phylogenetic placement (PP) of genome skims

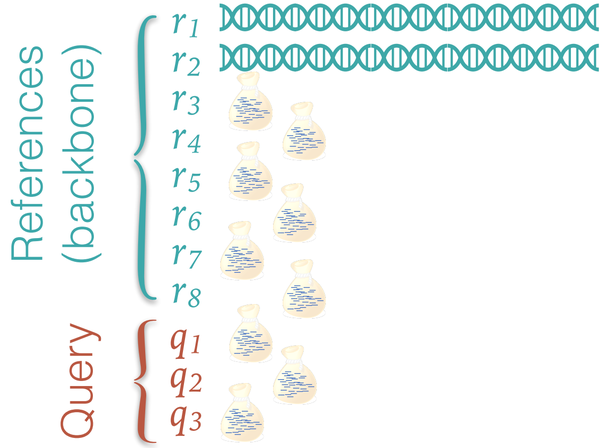
A genome skim: a bag of randomly sampled low-coverage reads

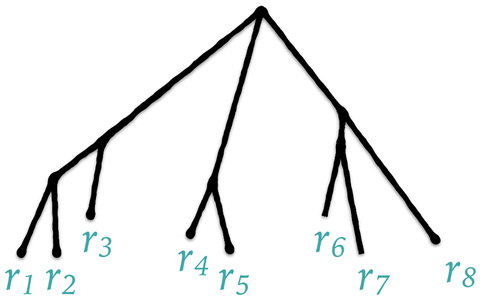
MOLECULAR ECOLOGY

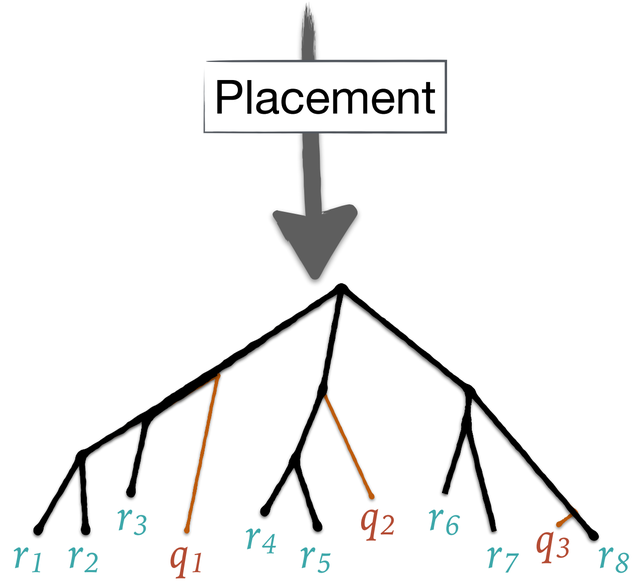
Beyond DNA barcoding: The unrealized potential of genome skim data in sample identification

Kristine Bohmann, Siavash Mirarab, Vineet Bafna, M. Thomas P. Gilbert



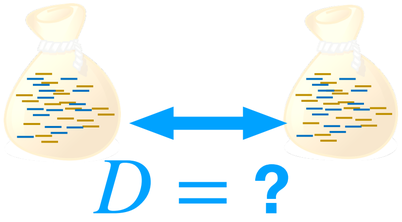






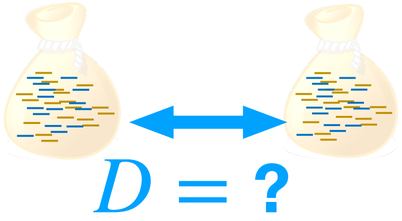
Distance calculation from k-mers

- Problem 1: What is the distance between two bags of k-mers?



Distance calculation from k-mers

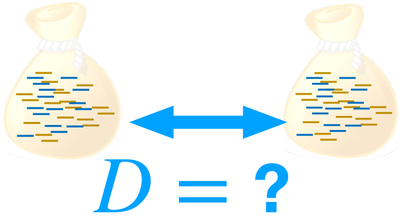
- Problem 1: What is the distance between two bags of k-mers?
  - Skmer: from Jaccard to distance:
    (Sarmashghi et el., 2019) (Balaban, et al., 2022)
  - DipSkmer: also modeling repeats:
    (Charvel et el.,under preparation)
  - ReSkmer: also modeling heterozygosity:
    (Charvel et el., under review, 2025)



Distance calculation from k-mers

- Problem 1: What is the distance between two bags of k-mers?
  - Skmer: from Jaccard to distance:
    (Sarmashghi et el., 2019) (Balaban, et al., 2022)

$$D = 1 - \left(\frac{2(\zeta_1 L_1 + \zeta_2 L_2)J}{\eta_1 \eta_2 (L_1 + L_2)(1 + J)} \right)^{1/k}$$



  - DipSkmer: also modeling repeats:
    (Charvel et el.,under preparation)

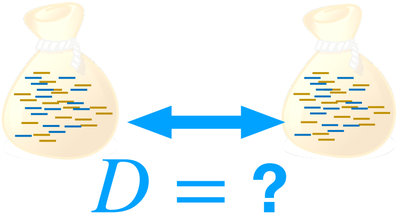
$$\sum_{i=1}^{m} r_i^{(1)} \left(1 - \left(1 - \eta^{(1)}\right)^i\right) \left(1 - \left(1 - \eta^{(2)}(1-d)^k\right)^i\right) + 3k r_i^{(1)} \left(1 - e^{-ib\lambda^{(1)}\epsilon^{(1)}(1-\epsilon^{(1)})^{k-1}}\right) \left(1 - \left((1-d)^k e^{-b\lambda^{(2)}\epsilon^{(2)}(1-\epsilon^{(2)})^{k-1}} - bd(1-d)^{k-1}\eta^{(2)} + \left(1 - (1-d)^k\right)\right)^i\right)$$

  - ReSkmer: also modeling heterozygosity:
    (Charvel et el., under review, 2025)

$$Q = 1 + 11\left(J\eta_t(2,2) - \eta_t(0,2)\eta_t(2,0)\right) / \left(J\left(4\eta_t(0,1) + 4\eta_t(1,0) + 4\eta_t(1,1) + 4\eta_t(1,2) + 4\eta_t(2,1) + \eta_t(0,2) + \eta_t(2,0) - 11\eta_t(2,2)\right) - 4\eta_t(0,1)\left(\eta_t(1,0) + \eta_t(2,0)\right) + \eta_t(0,2)\left(11\eta_t(2,0) - 4\eta_t(1,0)\right)\right)$$

Distance calculation from k-mers

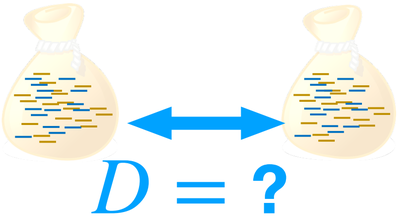
- Problem 1: What is the distance between two bags of k-mers?
  - Skmer: from Jaccard to distance:
    (Sarmashghi et el., 2019) (Balaban, et al., 2022)
  - DipSkmer: also modeling repeats:
    (Charvel et el.,under preparation)
  - ReSkmer: also modeling heterozygosity:
    (Charvel et el., under review, 2025)



Eduardo Charvel

Distance calculation from k-mers

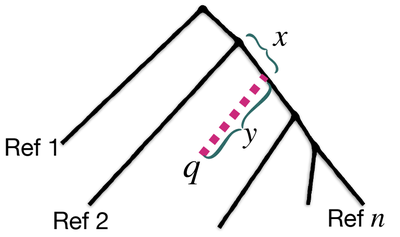
- Problem 1: What is the distance between two bags of k-mers?
  - Skmer: from Jaccard to distance:
    (Sarmashghi et el., 2019) (Balaban, et al., 2022)
  - DipSkmer: also modeling repeats:
    (Charvel et el.,under preparation)
  - ReSkmer: also modeling heterozygosity:
    (Charvel et el., under review, 2025)
- Problem 2: Given distances, place a query on a tree.



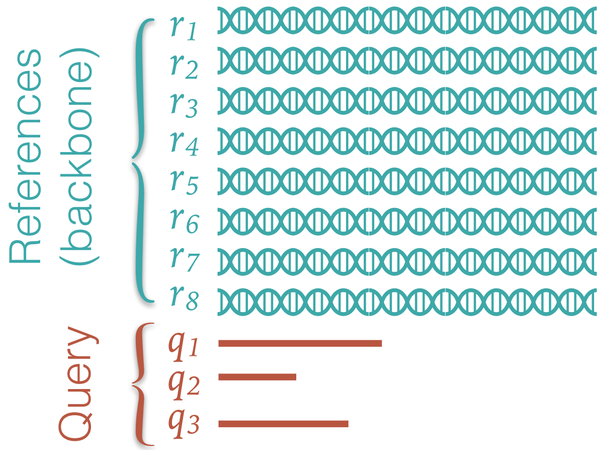
Eduardo Charvel

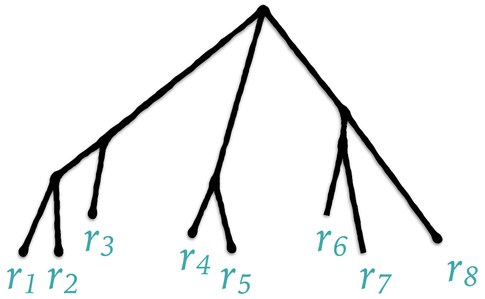
APPLES(-II)
[Balaban & Mirarab, Sys Bio, 2020]
[Balaban, et al, 2022]

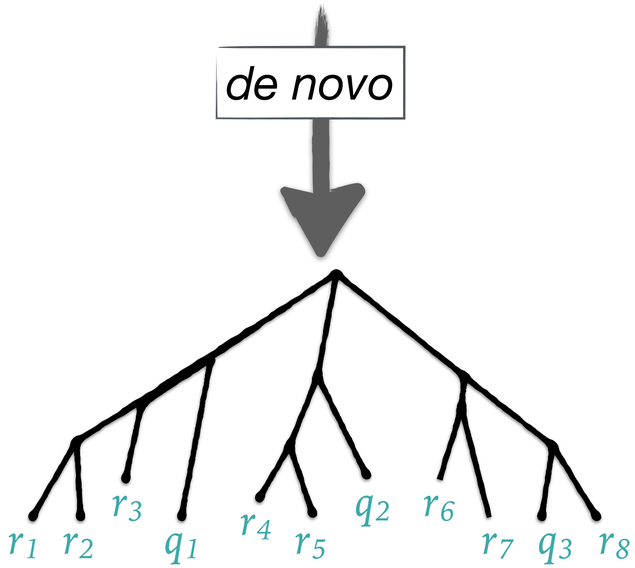
$$\text{argmin}_T \ \text{argmin}_{x,y} \sum_{i=1}^{n} w_{ij} \left(\delta_{qi} - d_T(q,i) \right)^2$$

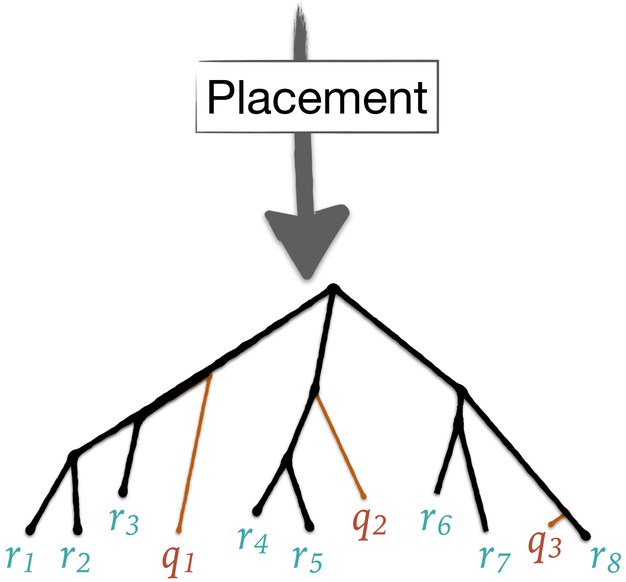


Phylogenetic placement (PP) of contigs/long read





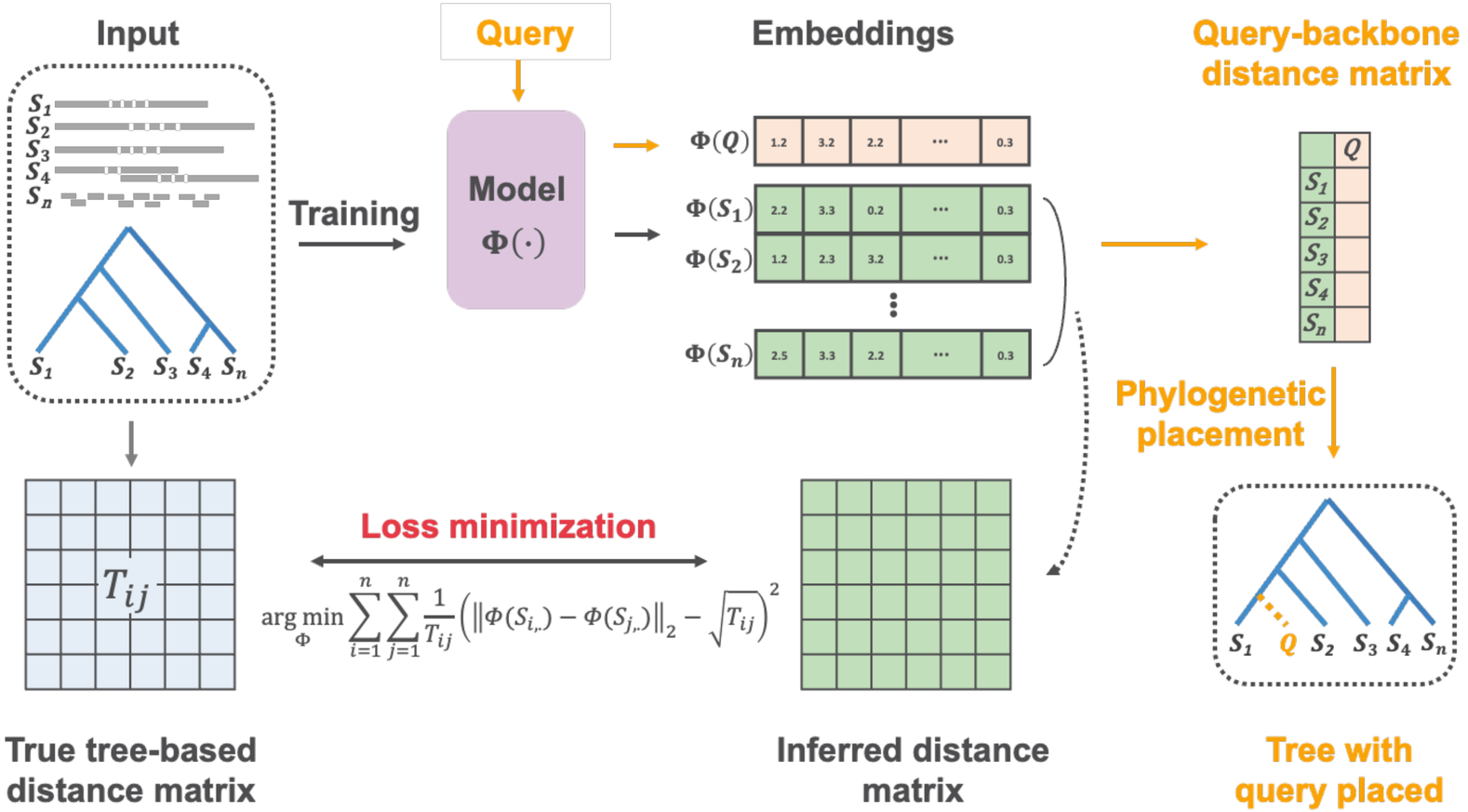




Phylogenetic Placement with k-mer frequency

Nora Rachtman

- Represent each genome, contig, or long read as 8-mer frequencies
- Train a function $\Phi : (0,1)^{4^k} \rightarrow \mathbb{R}^d$ that emulates reference tree distances
- Use the model to compute distances and place queries



Distance calculation and phylogenetic placement using k-mers

- Distance-based phylogenetics
  - Can be very fast
  - But not as accurate as ML methods 🚩

    [Nakhleh, et al., 2002] [Bruno, et al., 2010]
- Alignment-free (kmer-based) phylogenetic
  - Can be very fast
  - But not as accurate as alignment-based methods 🚩

Distance calculation and phylogenetic placement using k-mers

Yes, but …

- Versatility: Short, clean, easy-to-use pipelines, without a need for assembly or alignment
  - Enables analyses that are impossible otherwise
- Scalability
- Phylogenetic placement is noisy by necessity.

Many applications in ecology do not fit the marker-based alignment-based model

Place every read from anywhere on the genome (not just a handful of marker genes) on a reference tree of reference genomes

Place a genome skim (i.e., a bag of randomly sampled low-coverage reads) on a reference tree of reference genomes (or skims)

Place a long read or assembled contig on a reference tree of reference genomes (or skims)

Thank you!

Extra Slides

Known limitations

- Placements are too widely distributed
- Distances ignore k-mer dependencies, errors
- The placement algorithm does not enjoy any guarantees
  - For example, note the lack of any CTMC correction

Summary

An alignment-free framework for distance estimation

- based on homologous *k*-mers matches
- many potential applications including metagenomics

Summary

An alignment-free framework for distance estimation

- based on homologous *k*-mers matches
- many potential applications including metagenomics

krepp

- estimates read to genome distances >10x faster than alignment
- extends to more distant references and can map novel reads

Summary

An alignment-free framework for distance estimation

- based on homologous *k*-mers matches
- many potential applications including metagenomics

krepp

- estimates read to genome distances >10x faster than alignment
- extends to more distant references and can map novel reads
- easily scales reference sets with >100,000 microbial genomes

Summary

An alignment-free framework for distance estimation

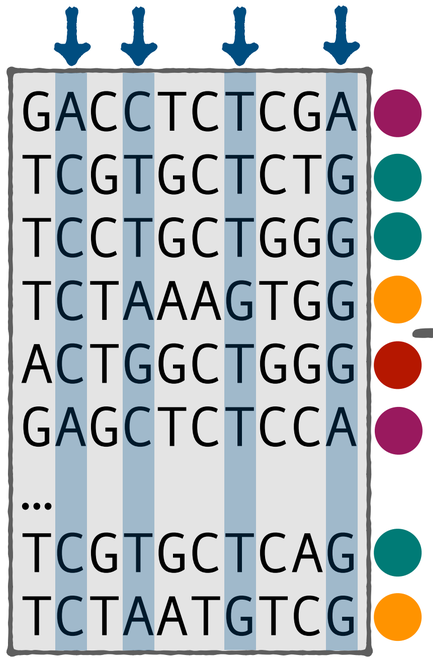
- based on homologous *k*-mers matches
- many potential applications including metagenomics

krepp

- estimates read to genome distances >10x faster than alignment
- extends to more distant references and can map novel reads
- easily scales reference sets with >100,000 microbial genomes
- places genome-wide reads on ultra-large phylogenies *(only method)*

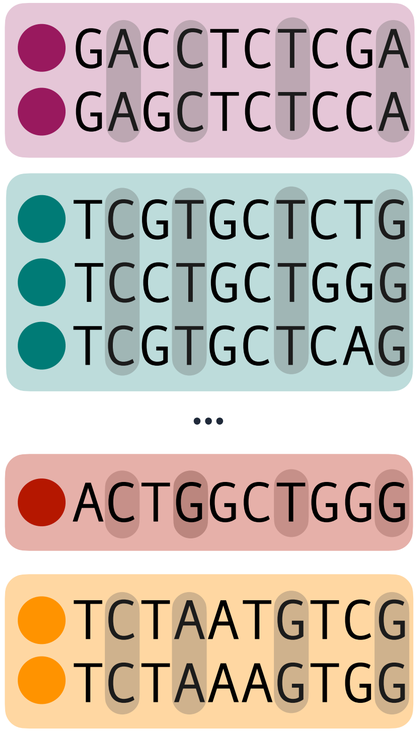
Computing Hamming distances between homologous k-mers

Select h random but fixed positions (default h: 14, k: 29)



reference *k*-mers

locality-sensitive hashing



4^h LSH buckets

HD

4

1

4

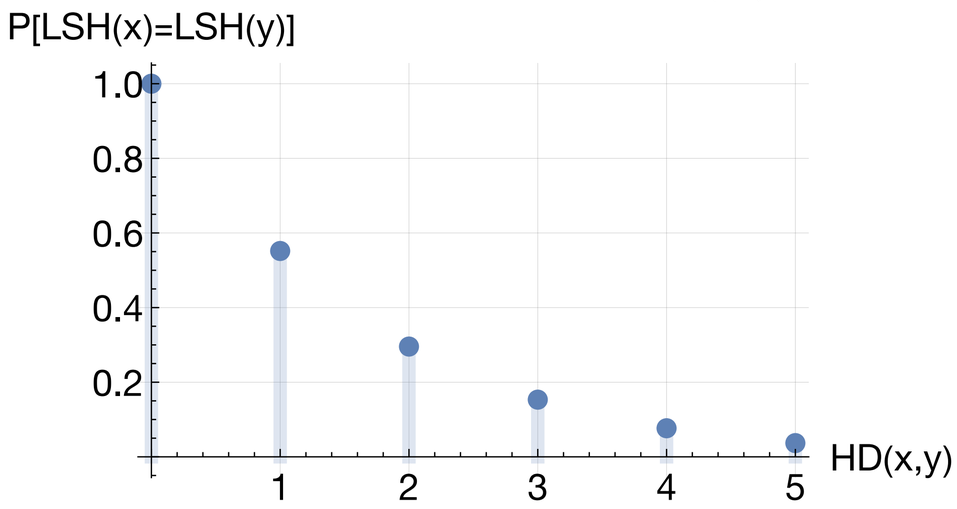
miss at HD=2

Given a query *k*-mer

ACCTGCTGGG

collides for HD=x with probability:

$$\frac{\binom{k-h}{x}}{\binom{k}{x}}$$



Dealing with uncertainty: statistically distinguishability

- short reads — **low signal**
- high distances — fewer matching *k*-mers
- small differences may not be statistically meaningful
  - **test distinguishability**

likelihood-ratio test
with the closest reference:

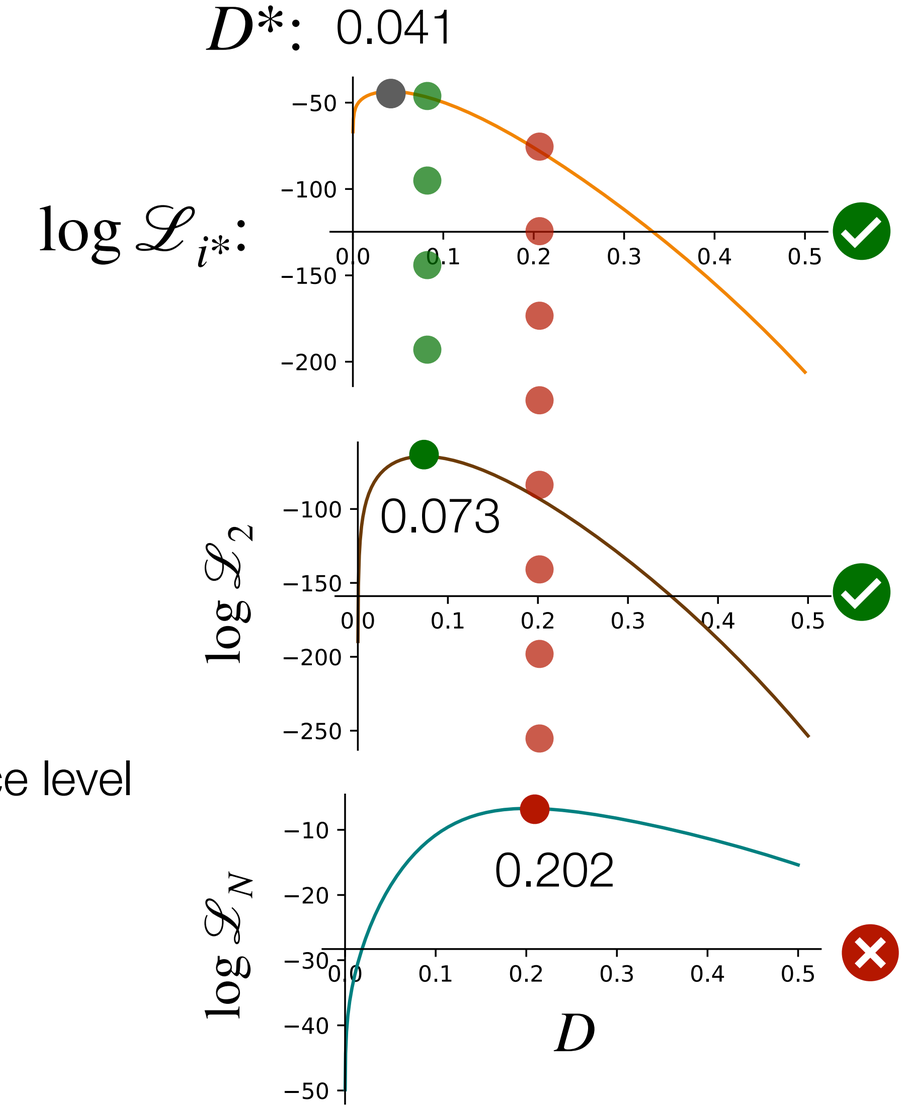
$$\lambda_{LR} = \frac{\mathscr{L}_{i*}(D; k, h, \delta, u_{i*}, \mathbf{v}_{i*})}{\mathscr{L}_{i*}(D^*; k, h, \delta, u_{i*}, \mathbf{v}_{i*})}$$

D: alternative distance

$i*$: closest reference

$$\lambda_{LR} \sim \chi^2$$

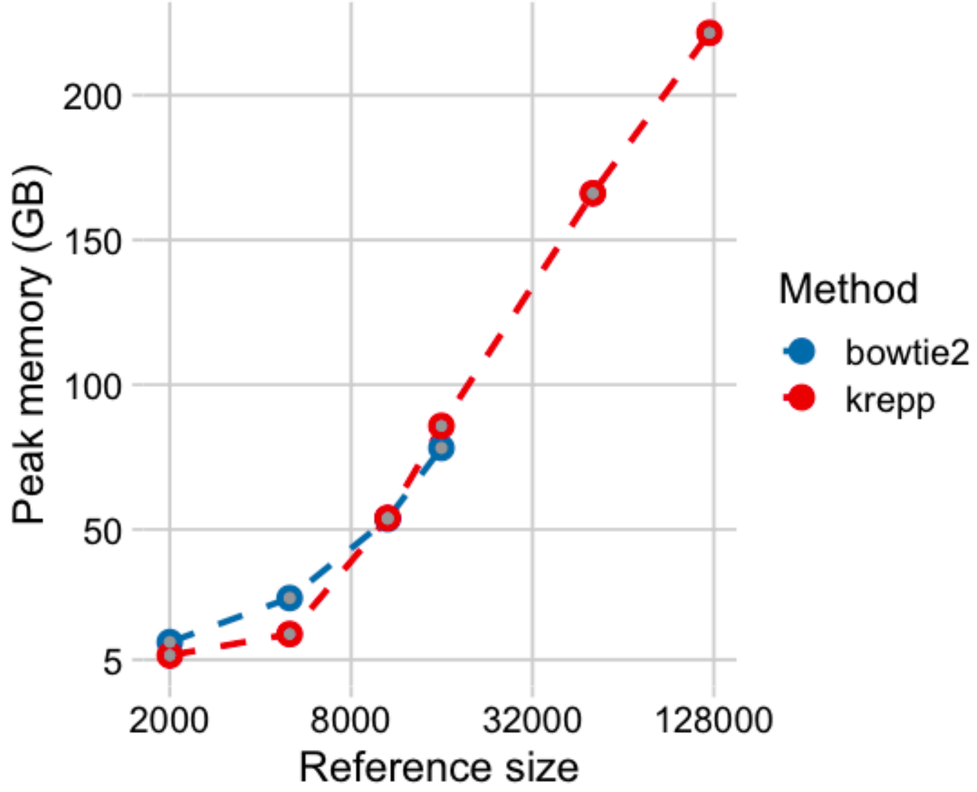
- select a significance level (default: α=90%)



Scalability:
krepp can be distributed and has flexible memory requirements

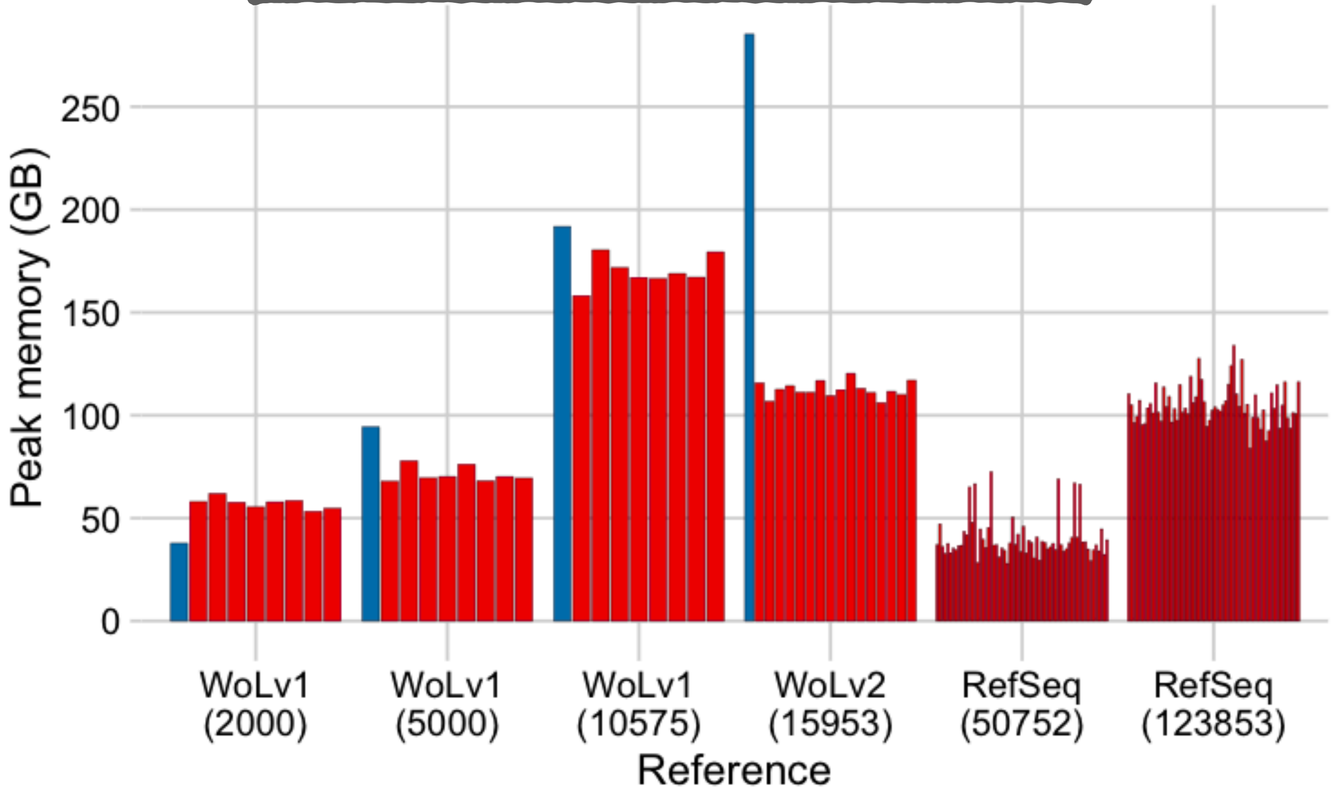
Mapping 10M reads (16 threads):

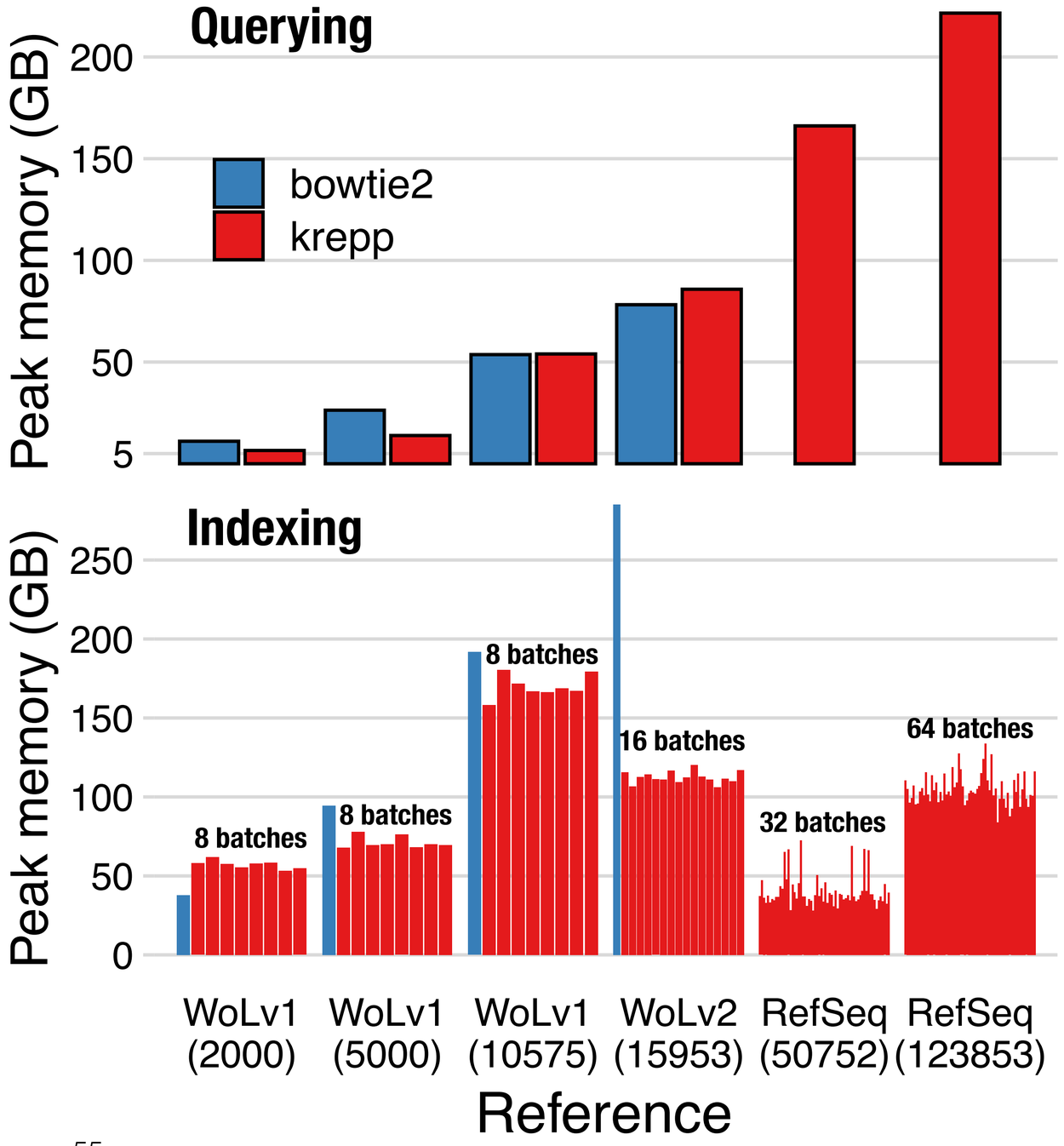
reducing memory use
w/ further subsampling...



Indexing microbial genomes (32 threads):

adjusting partitioning based on the
input size & available memory...




Querying
Peak memory (GB)
200
150
100
50
5
bowtie2
krepp
Indexing
Peak memory (GB)
250
200
150
100
50
0
8 batches
8 batches
8 batches
16 batches
32 batches
64 batches
WoLv1 (2000)
WoLv1 (5000)
WoLv1 (10575)
WoLv2 (15953)
RefSeq (50752)
RefSeq (123853)
Reference