Estimating read-genome distances using homologous *k*-mers

(and genome-wide phylogenetic placement)

Ali Osman Berk Şapcı & Siavash Mirarab
UC San Diego

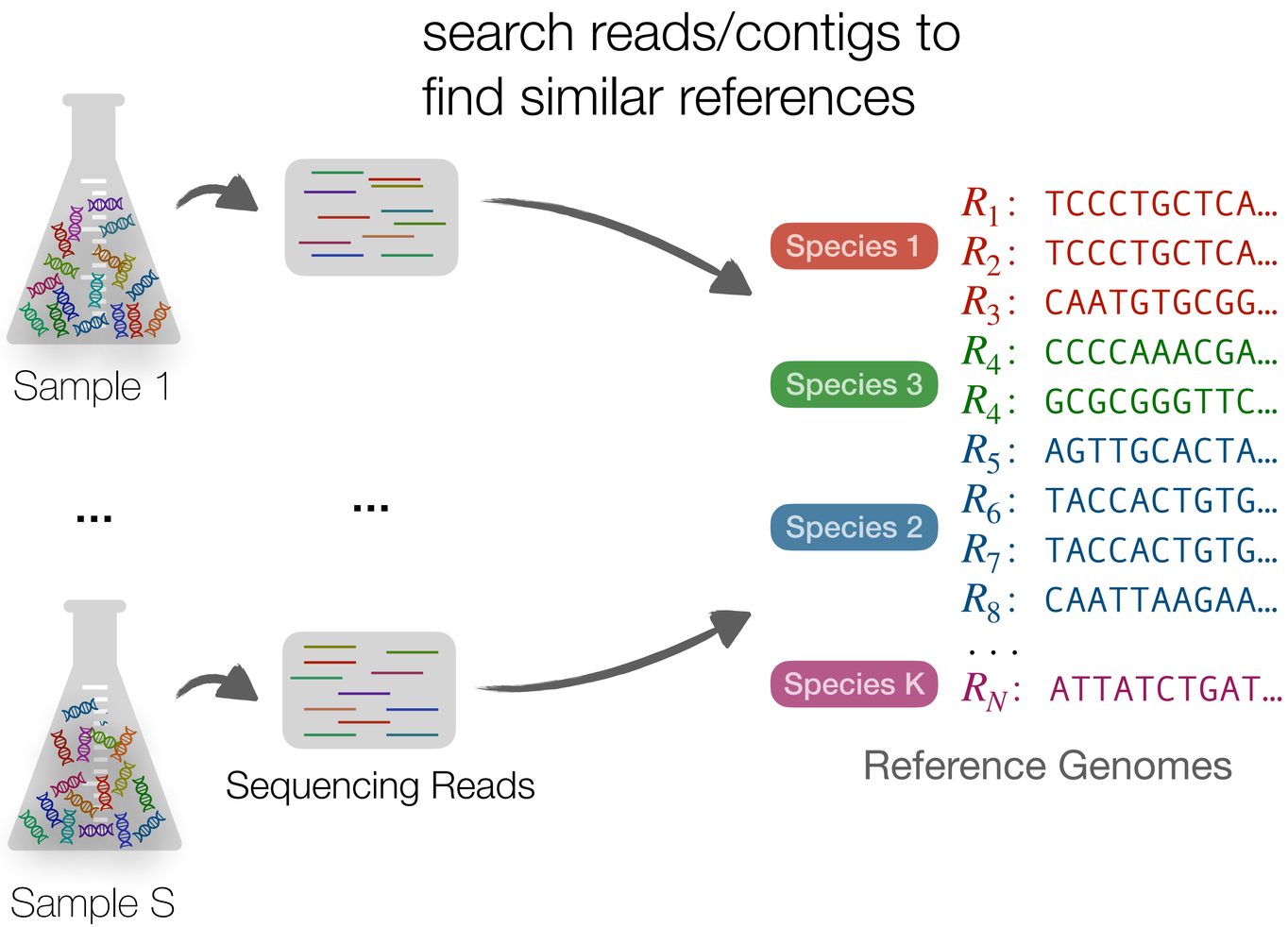
Bioinformatics & Systems Biology

UC San Diego
Electrical and Computer Engineering
JACOBS SCHOOL OF ENGINEERING

Identifying metagenomic sequences and comparing samples

Goal:
analyze the present taxa &
compare different samples

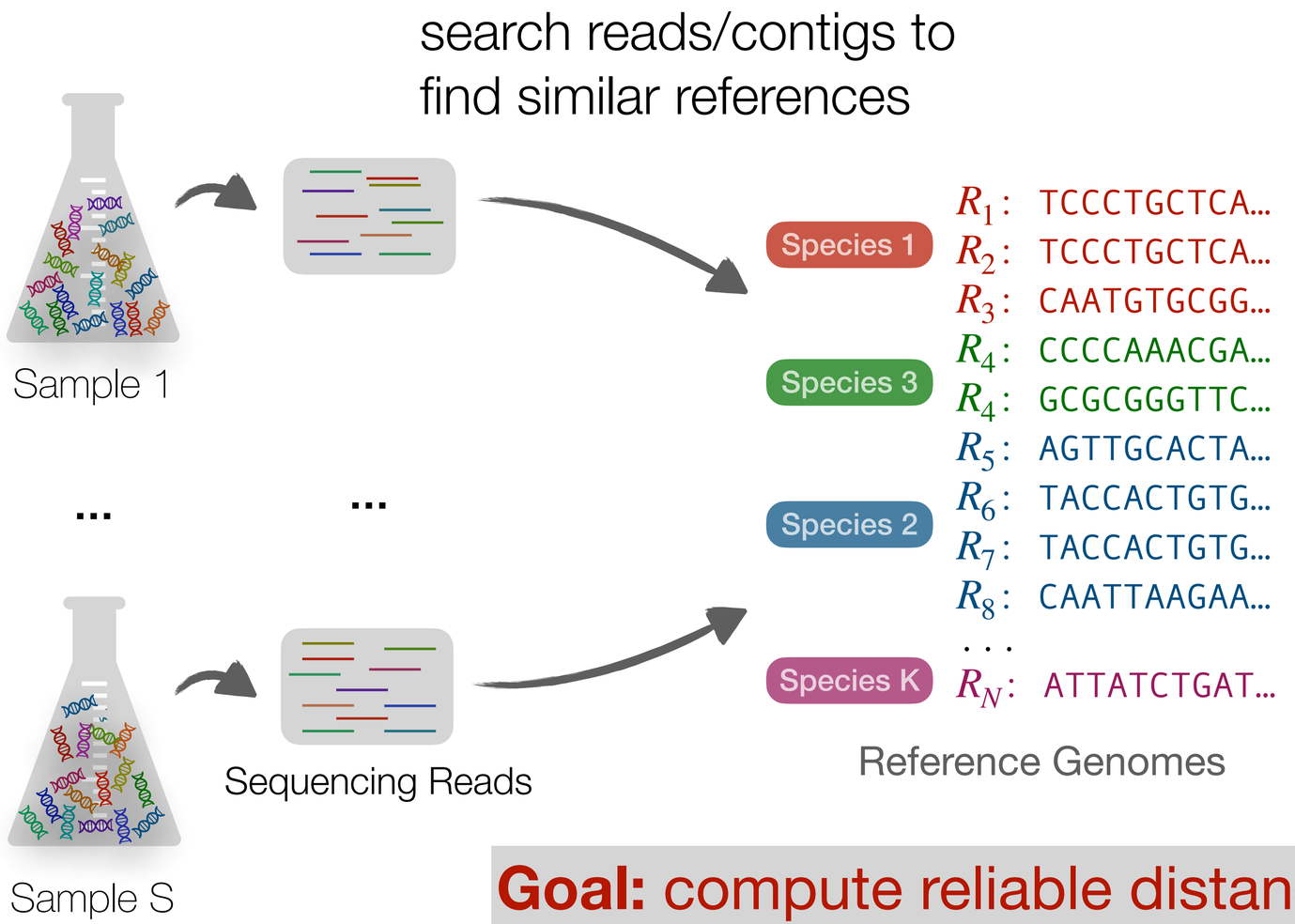
Identifying metagenomic sequences and comparing samples



Existing methods:

- Good old taxonomic profiling/binning
  → Kraken2, CONSULT-II, …
- Containment analysis using MinHash/FracMinHash
  → sourmash, mash-screen, …
- Focusing on marker genes
  → EPA-ng, mOTU, MetaPhyler, …

Identifying metagenomic sequences and comparing samples



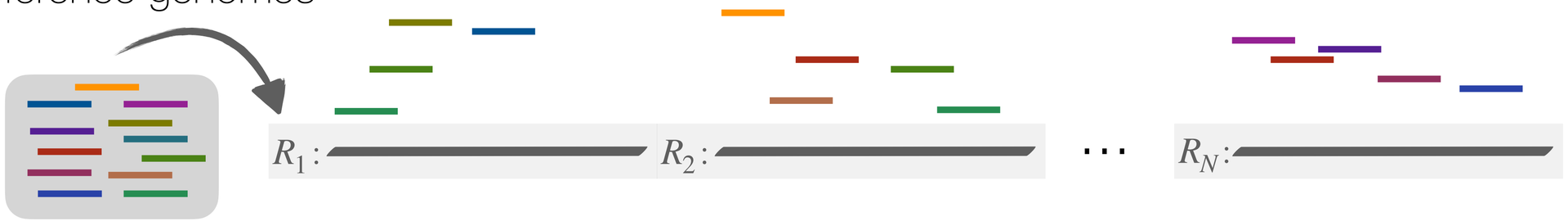
Existing methods:

- Good old taxonomic profiling/binning
  → Kraken2, CONSULT-II, …
- Containment analysis using MinHash/FracMinHash
  → sourmash, mash-screen, …
- Focusing on marker genes
  → EPA-ng, mOTU, MetaPhyler, …

Goal: compute reliable distances between every read and all related references…

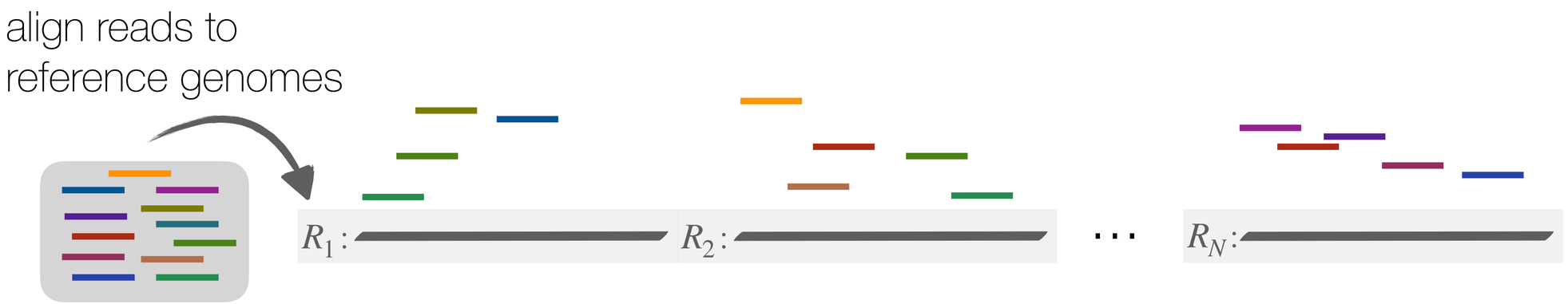
Aligning sequences of unknown origins

align reads to
reference genomes



(+) quantifying similarity — as detailed as it gets

Aligning sequences of unknown origins

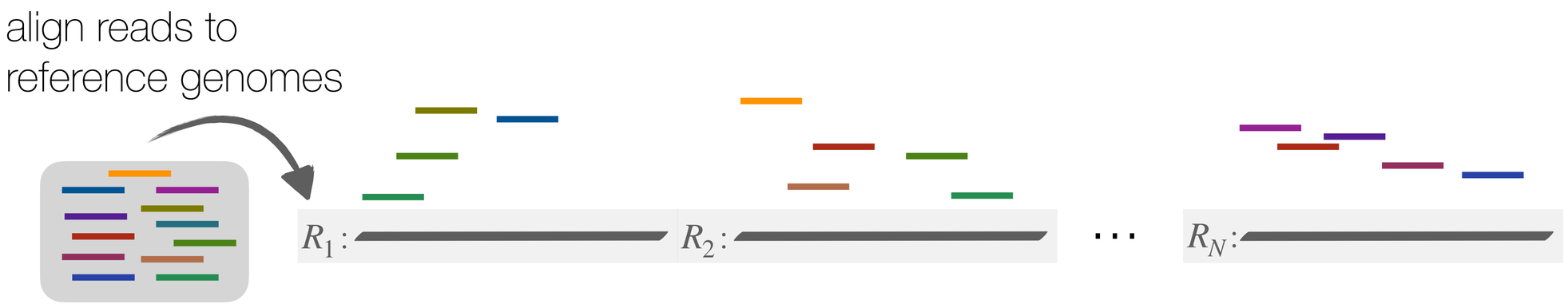


(+) quantifying similarity — as detailed as it gets

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

(-) not suitable for higher distances & novel sequences (>15%)

Aligning sequences of unknown origins



(+) quantifying similarity — as detailed as it gets

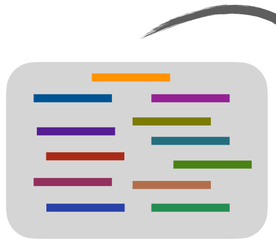
(-) not scalable for large N, even with efficient indexes

(-) not suitable for higher distances & novel sequences (>15%)

Because of these constraints, phylogenetic placement is limited to marker genes *or* small phylogenies.

Aligning sequences of unknown origins

align reads to
reference genomes



Can we estimate accurate read to genome distances without alignment?

R_1 : R_2 : ⋯ R_N :

(+) quantifying similarity — as detailed as it gets

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

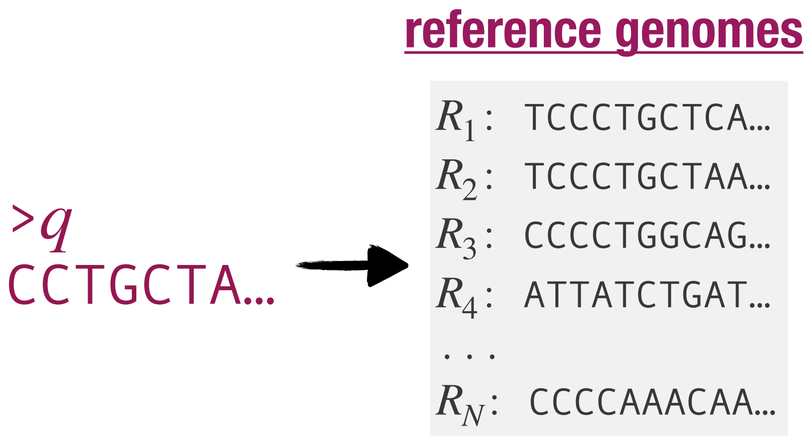
(-) not suitable for higher distances & novel sequences (>15%)

Because of these constraints, phylogenetic placement is limited to marker genes *or* small phylogenies.

Problem statement and our goals

Given:

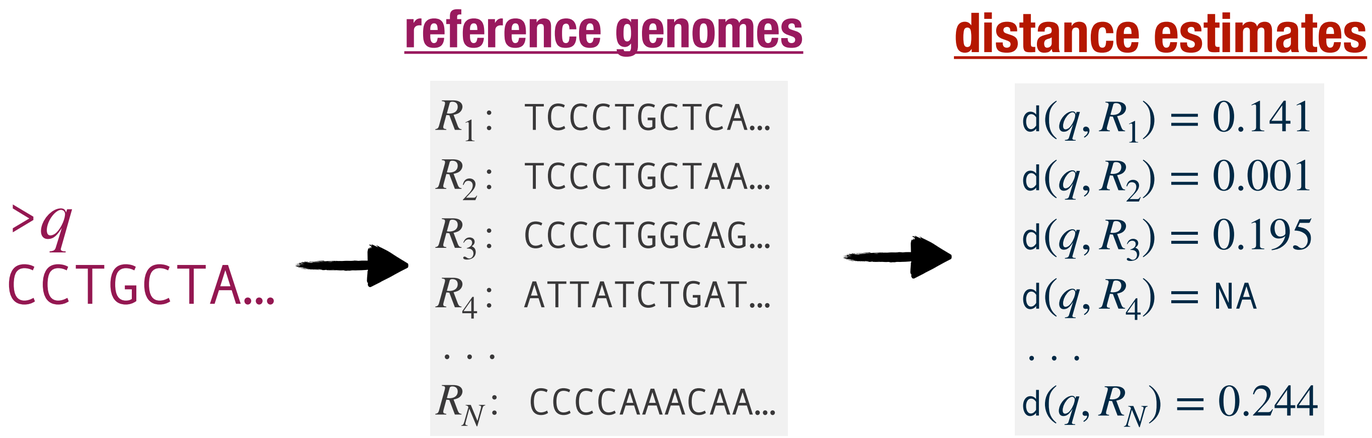
- query sequence q
- set of references $\mathscr{R} = \{R_1, \ldots, R_N\}$
- a backbone phylogeny T



Problem statement and our goals

Given:

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



- accurate & scalable with modern references

Problem statement and our goals

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

reference genomes

$>q$
CCTGCTA…

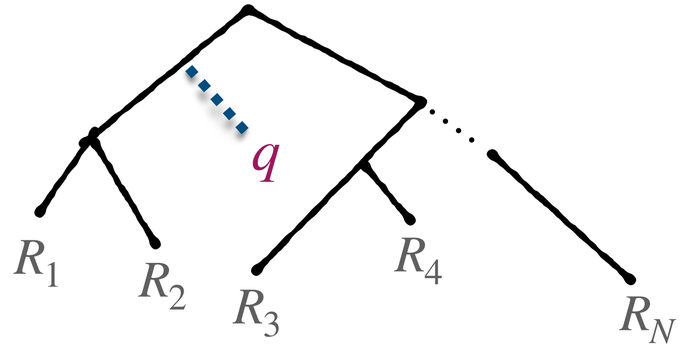
R_1: TCCCTGCTCA…
R_2: TCCCTGCTAA…
R_3: CCCCTGGCAG…
R_4: ATTATCTGAT…
. . .
R_N: CCCCAAACAA…

distance estimates

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

- accurate & scalable with modern references

phylogenetic placement:



- go beyond markers
- ultra-large phylogenies

Problem statement and our goals

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

reference genomes

$>q$
CCTGCTA…

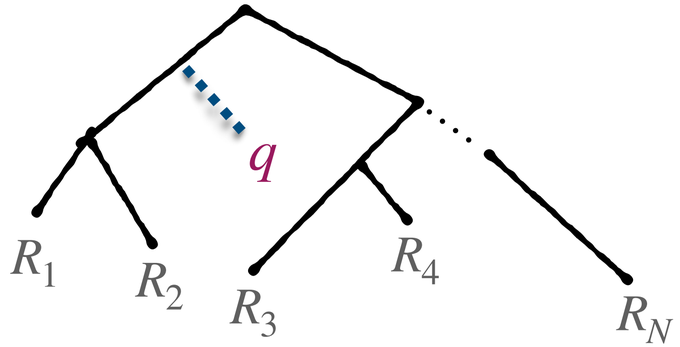
R_1: TCCCTGCTCA…
R_2: TCCCTGCTAA…
R_3: CCCCTGGCAG…
R_4: ATTATCTGAT…
. . .
R_N: CCCCAAACAA…

distance estimates

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

- accurate & scalable with modern references

phylogenetic placement:



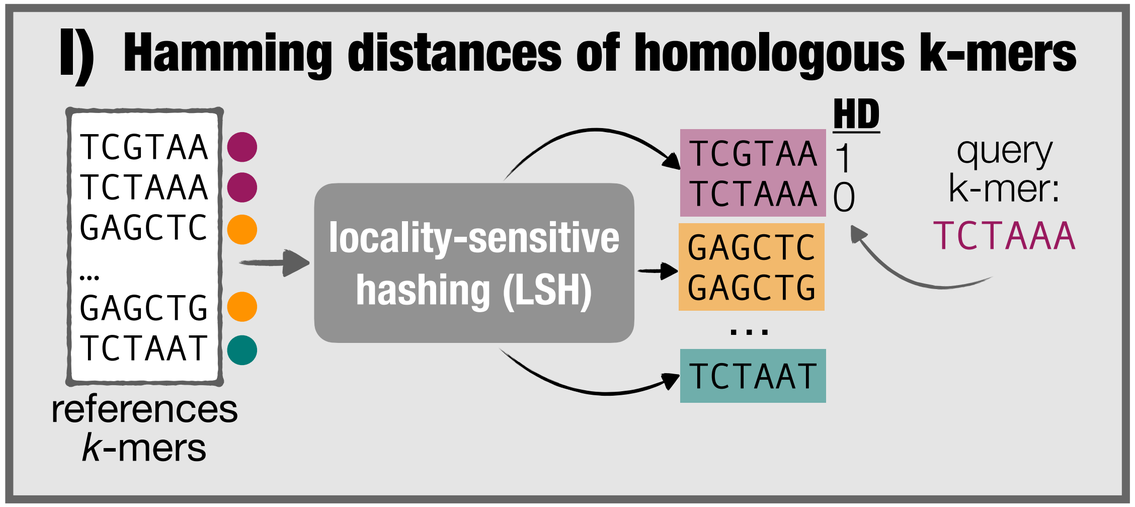
- go beyond markers
- ultra-large phylogenies

Outline of the method & subproblems we need to tackle

krepp: k-mer-based read phylogenetic placement

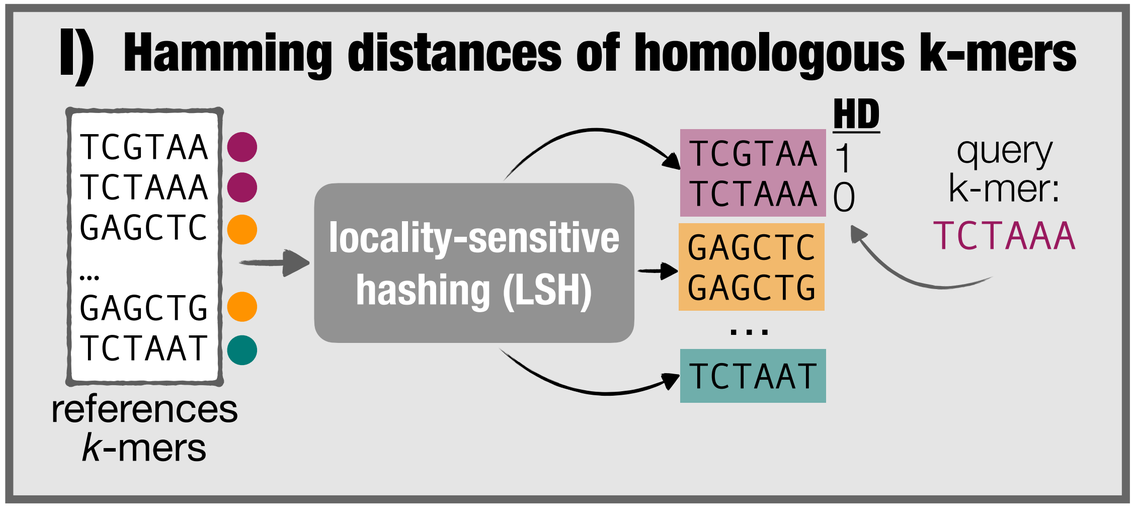
Outline of the method & subproblems we need to tackle

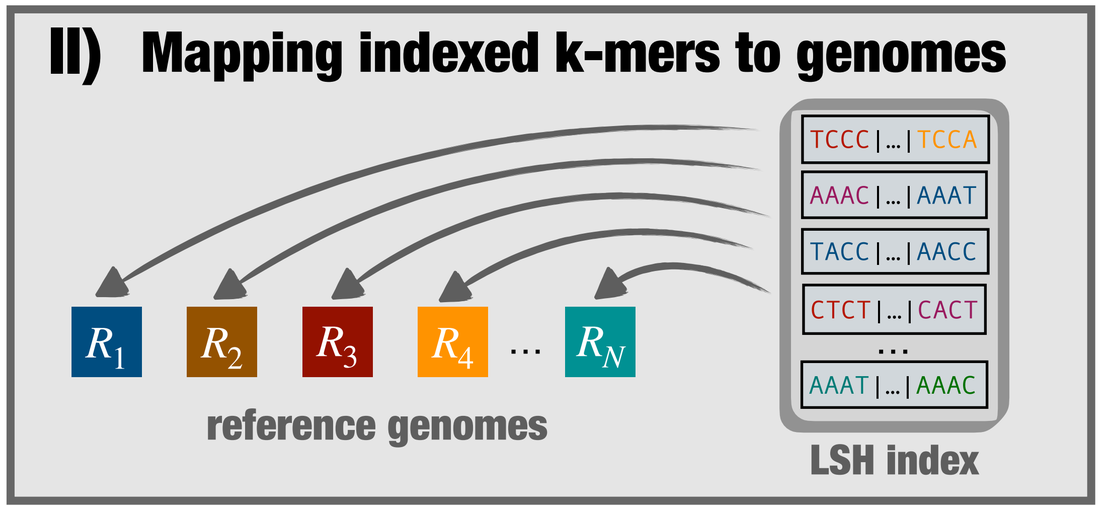
krepp: k-mer-based read phylogenetic placement



Outline of the method & subproblems we need to tackle

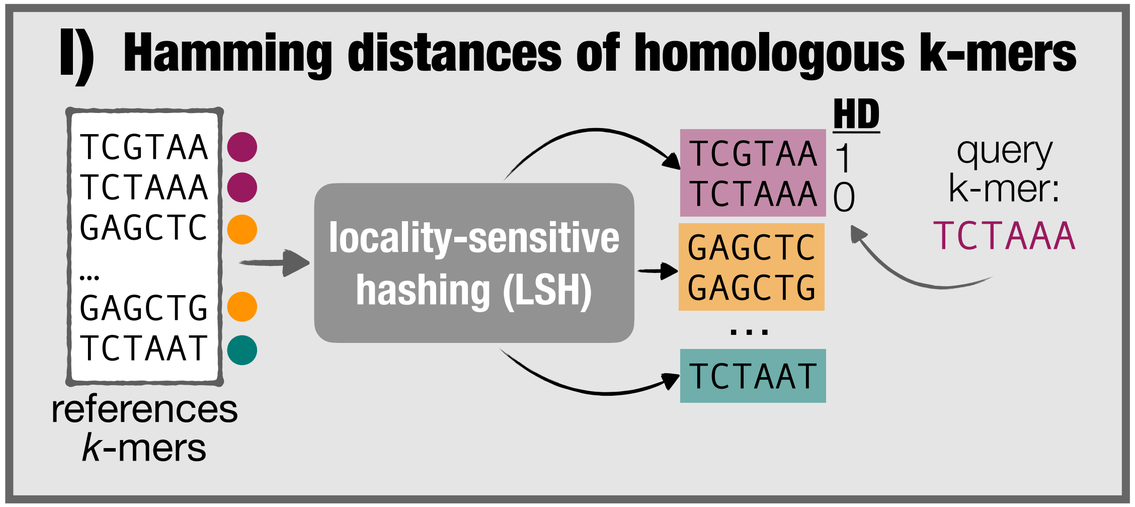
krepp: k-mer-based read phylogenetic placement

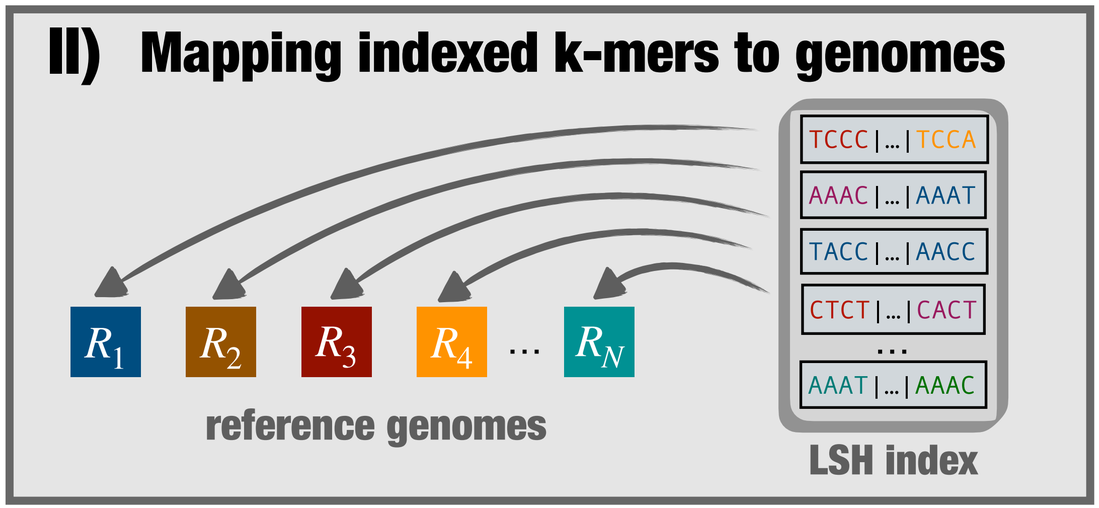


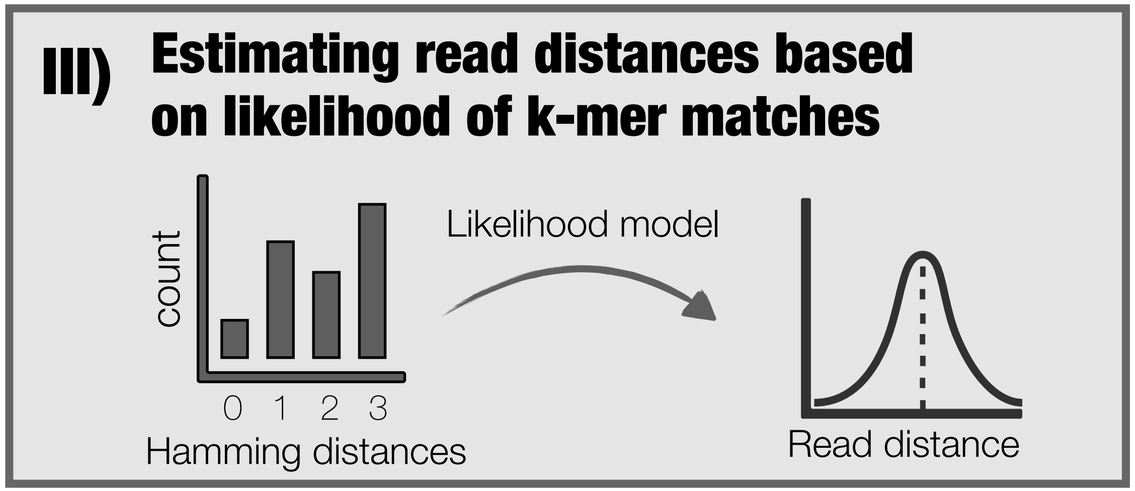


Outline of the method & subproblems we need to tackle

krepp: k-mer-based read phylogenetic placement



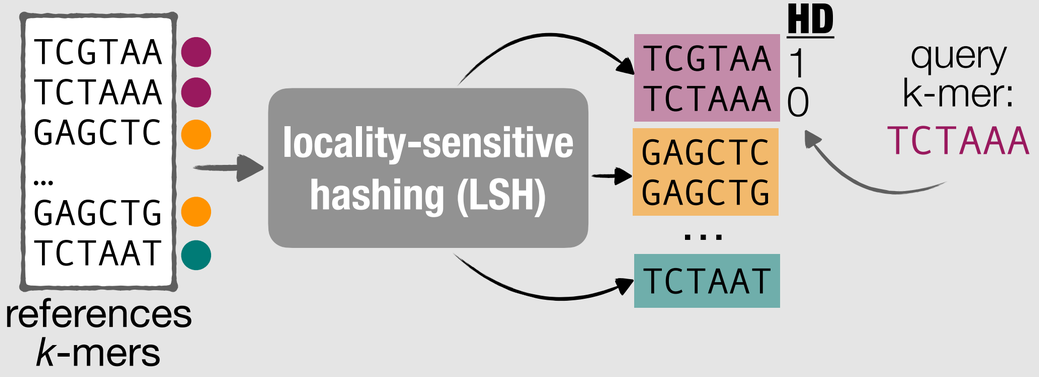




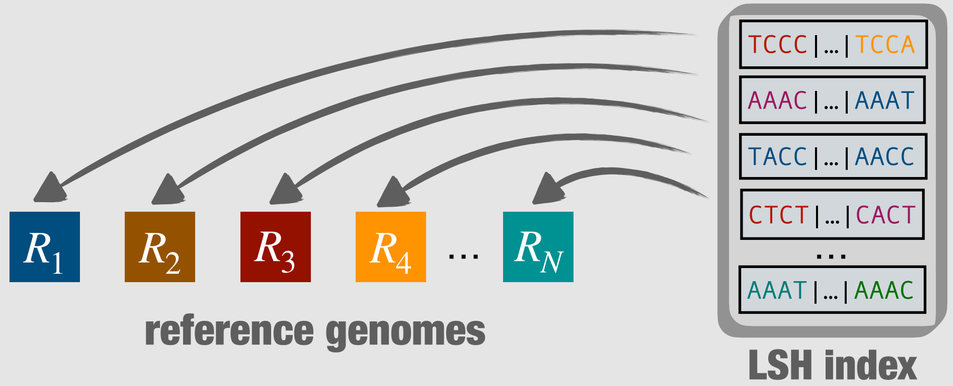
Outline of the method & subproblems we need to tackle

krepp: k-mer-based read phylogenetic placement

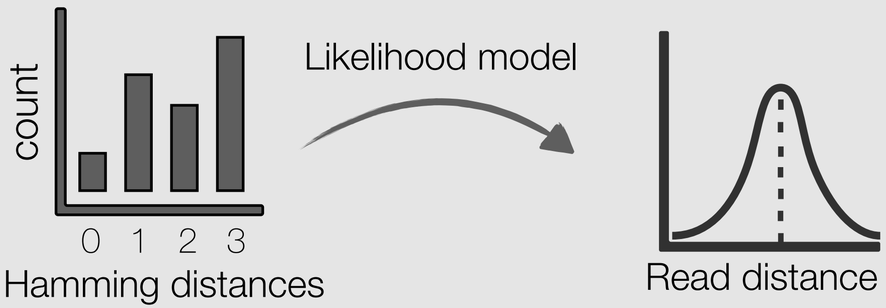
I) Hamming distances of homologous k-mers



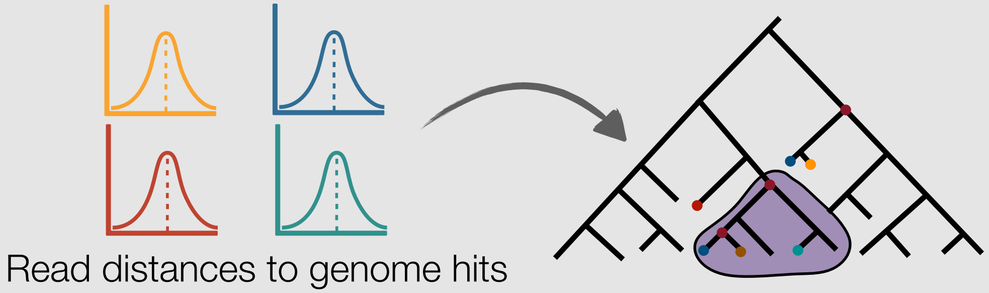
II) Mapping indexed k-mers to genomes



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



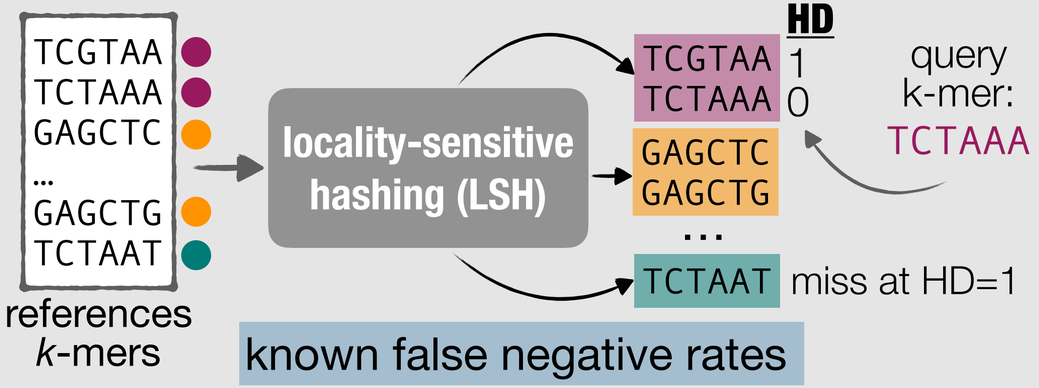
IV) Placement on the reference phylogeny by modeling uncertainties of distances



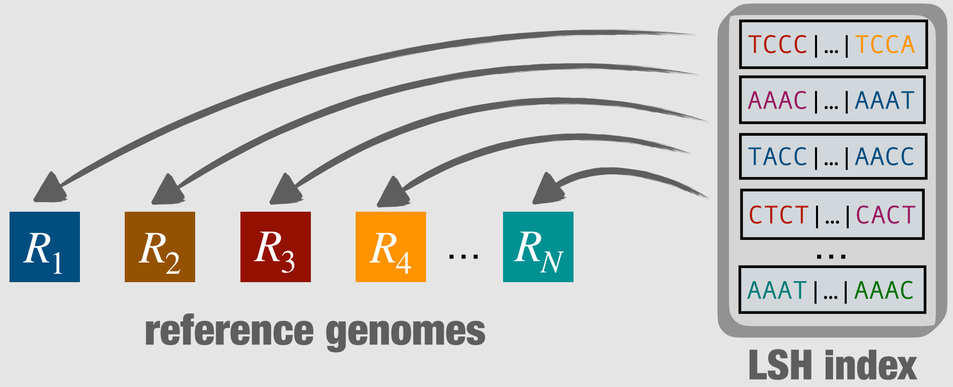
Outline of the method & subproblems we need to tackle

krepp: k-mer-based read phylogenetic placement

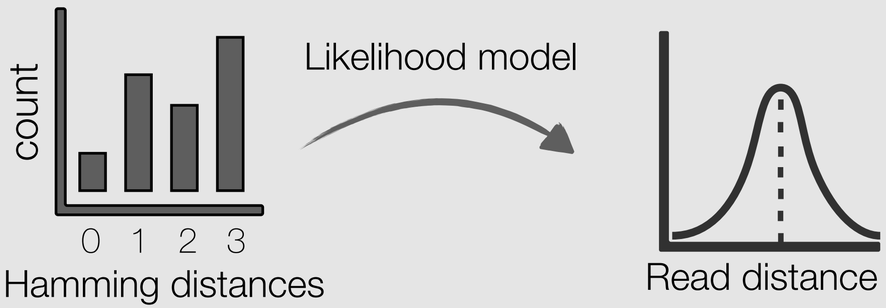
I) Hamming distances of homologous k-mers



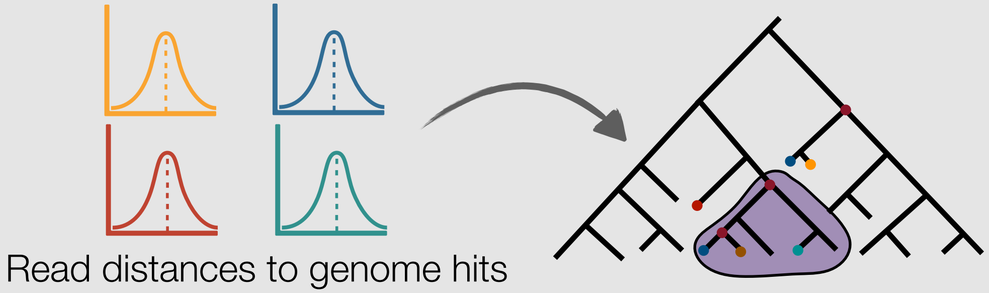
II) Mapping indexed k-mers to genomes



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



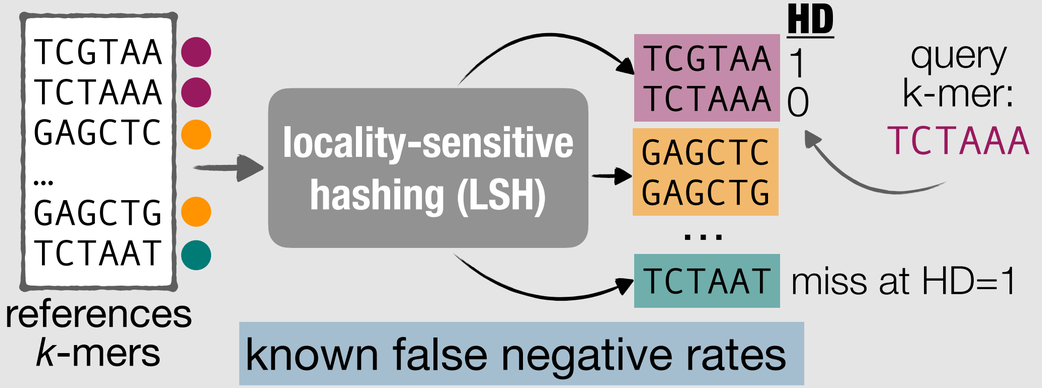
IV) Placement on the reference phylogeny by modeling uncertainties of distances



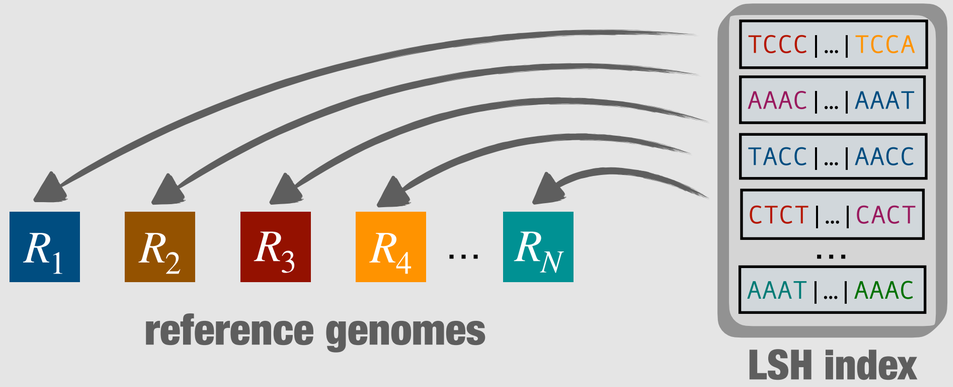
Outline of the method & subproblems we need to tackle

krepp: k-mer-based read phylogenetic placement

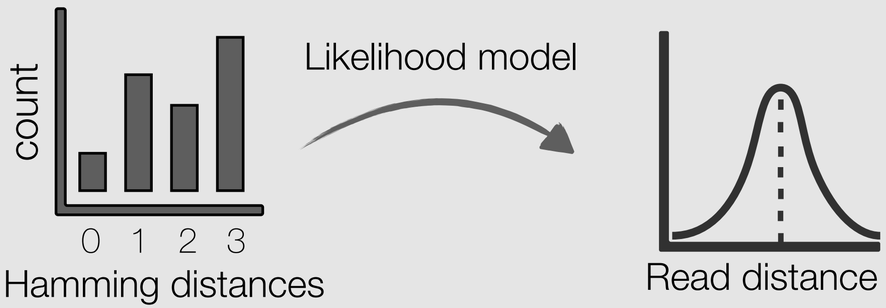
I) Hamming distances of homologous k-mers



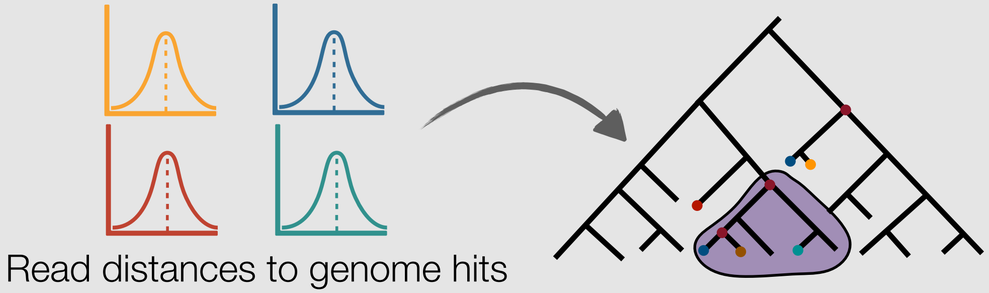
II) Mapping indexed k-mers to genomes



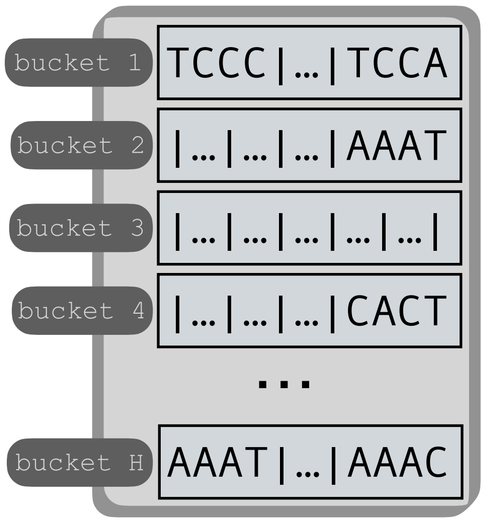
III) Estimating read distances based on likelihood of k-mer matches



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



Mapping indexed k-mers to reference genomes



LSH index

R_1

R_2

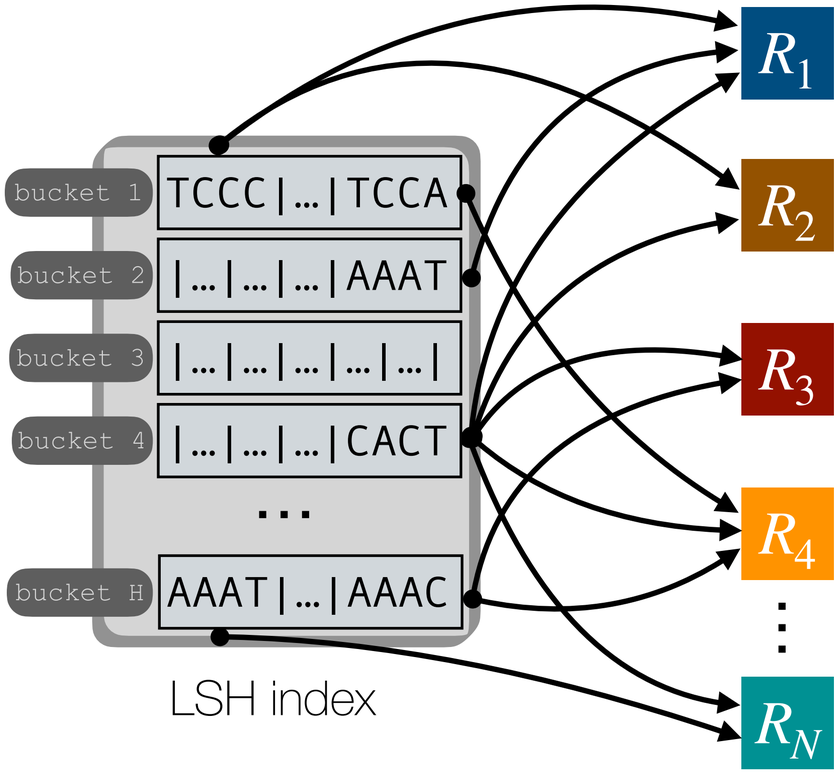
R_3

R_4

$\vdots$

R_N

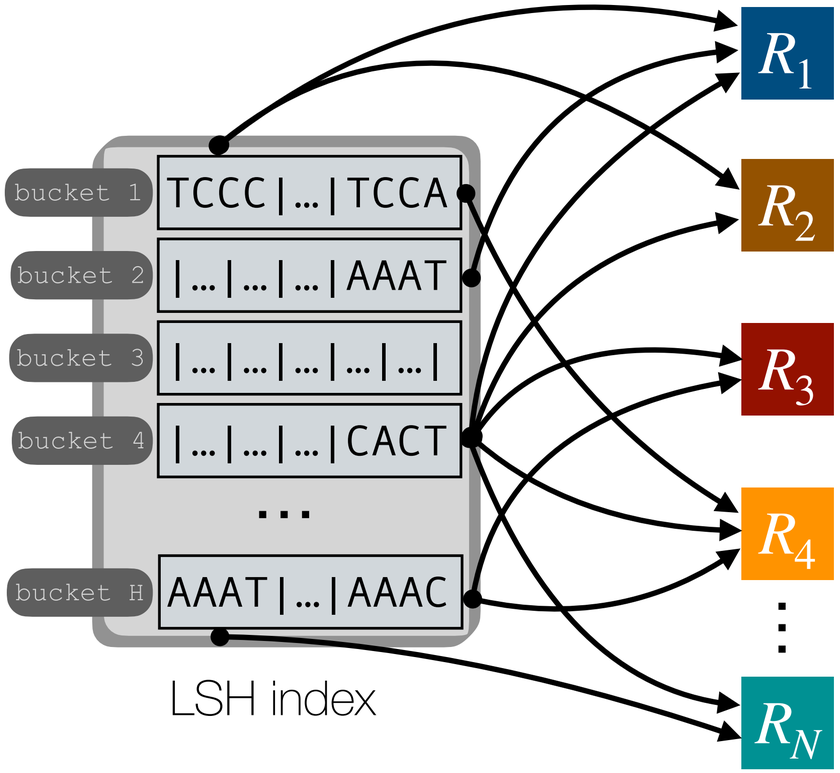
Mapping indexed k-mers to reference genomes



well studied **colored *k*-mer** problem

color: a subset of references
(including singletons)

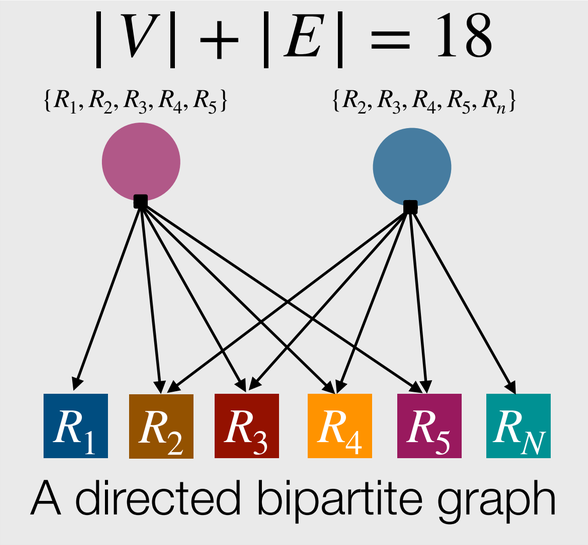
Mapping indexed k-mers to reference genomes



LSH index

well studied **colored *k*-mer** problem

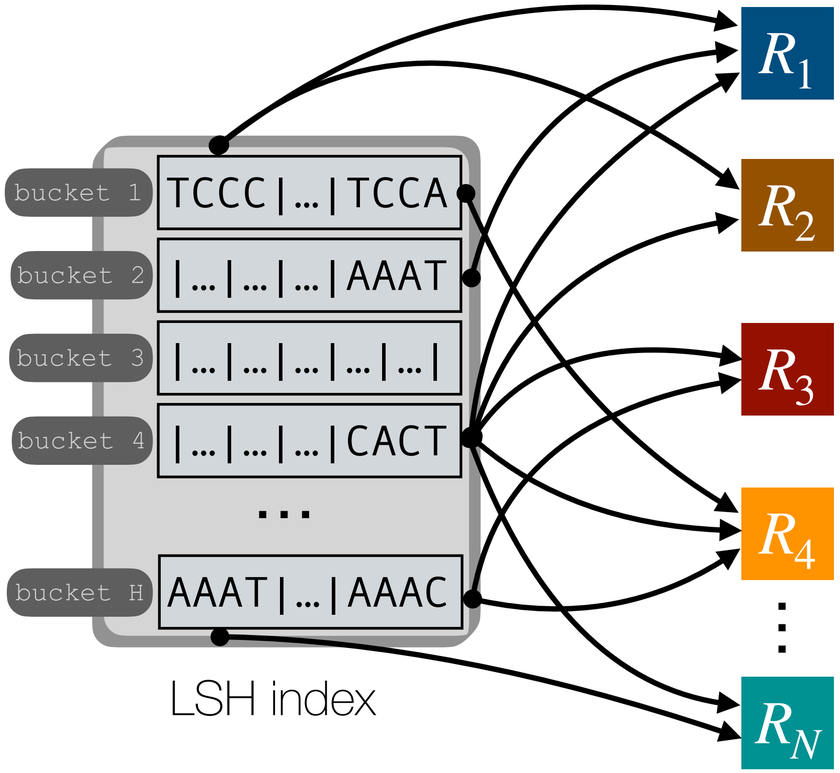
color: a subset of references
(including singletons)



A directed bipartite graph

represent each
non-singleton as
a union of others

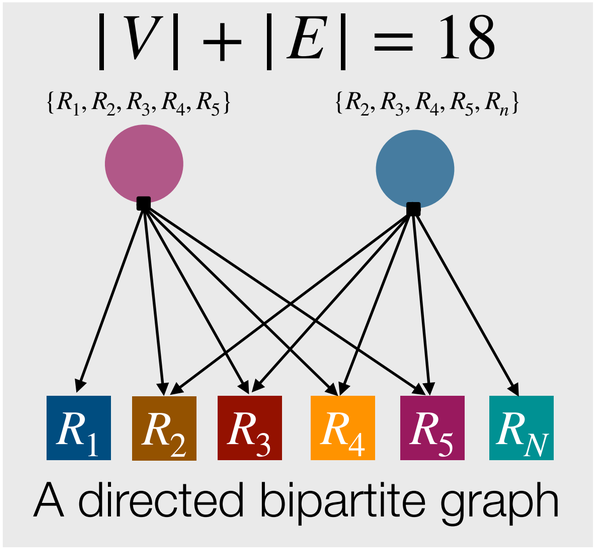
Mapping indexed k-mers to reference genomes



LSH index

well studied **colored *k*-mer** problem

color: a subset of references (including singletons)

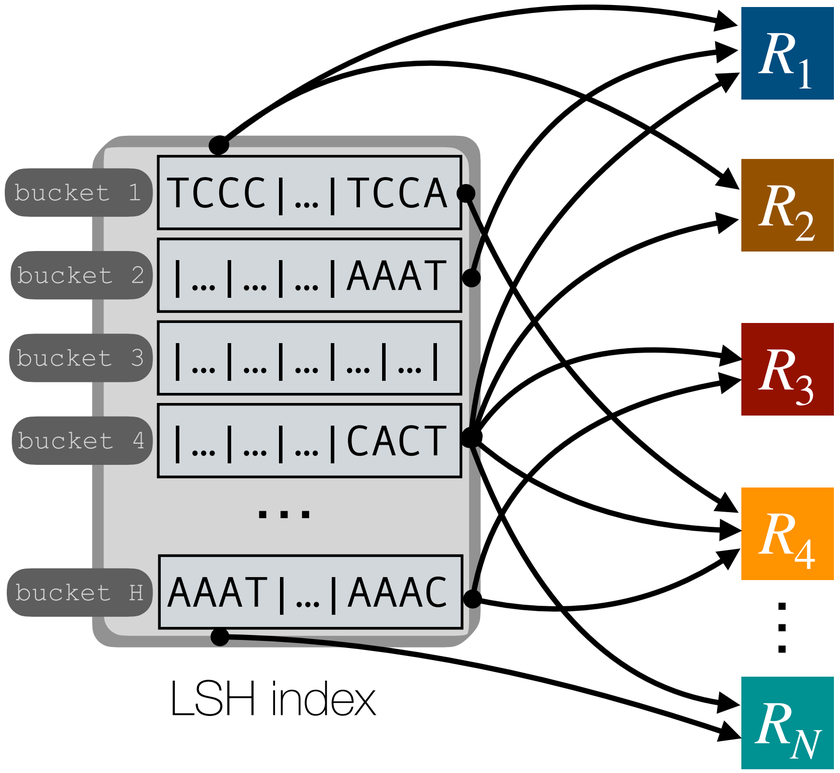


A directed bipartite graph

represent each non-singleton as a union of others

Minimize $|V| + |E|$?

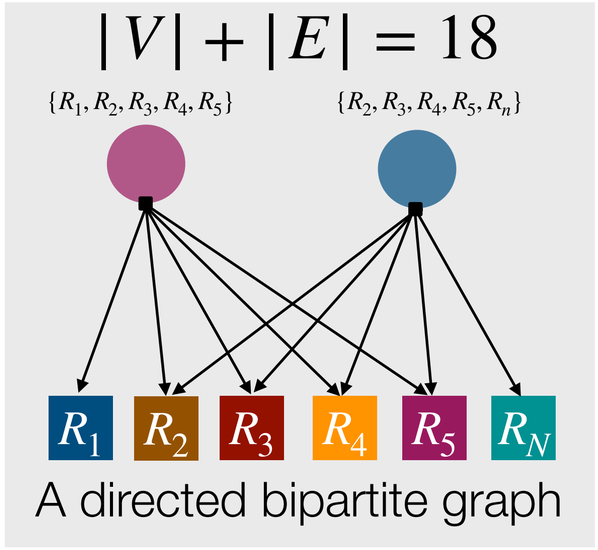
Mapping indexed k-mers to reference genomes



LSH index

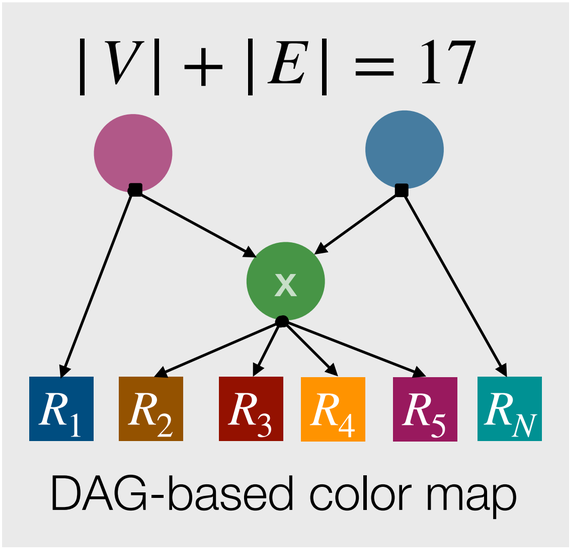
well studied **colored *k*-mer** problem

color: a subset of references (including singletons)



A directed bipartite graph

represent each non-singleton as a union of others



DAG-based color map

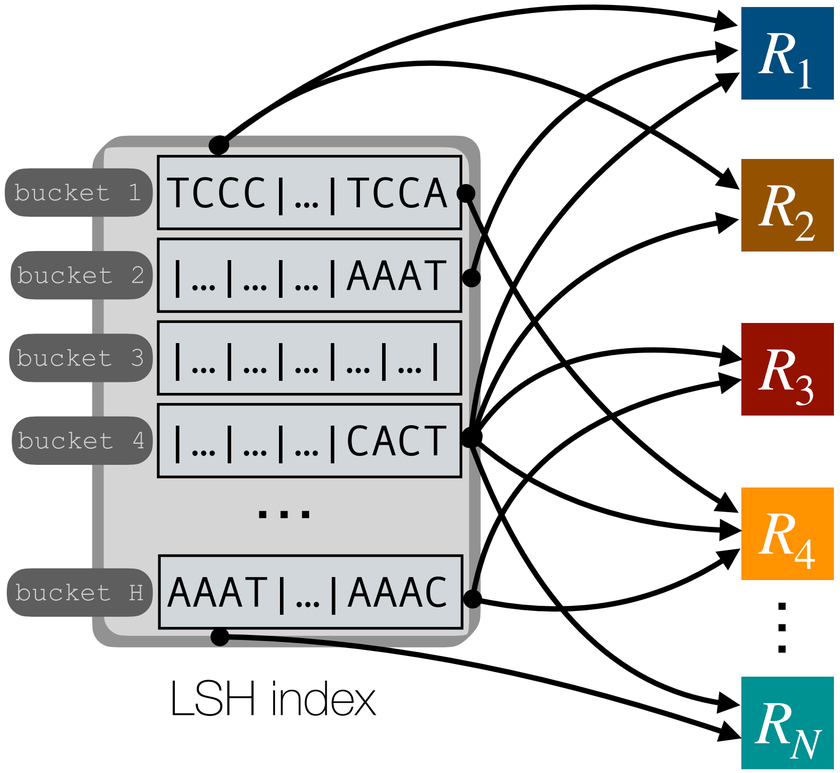
Minimize $|V| + |E|$?

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

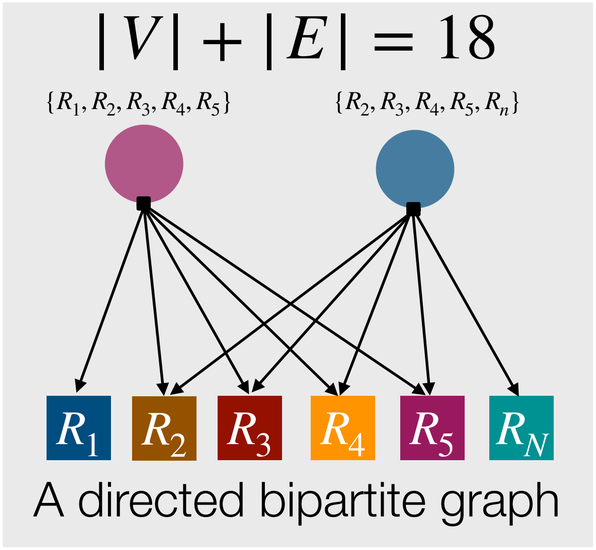
Mapping indexed k-mers to reference genomes



LSH index

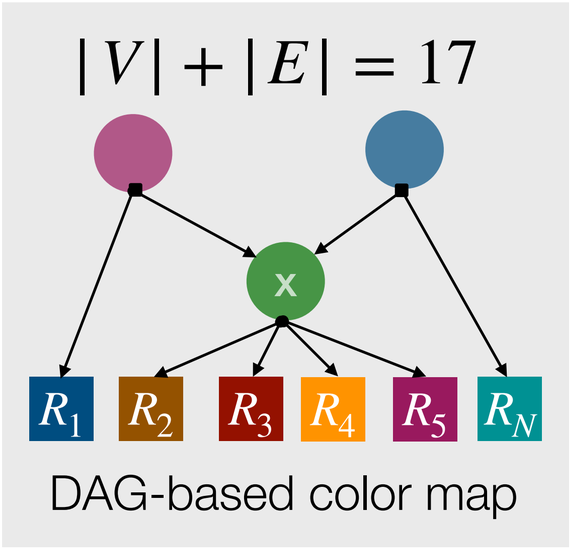
well studied **colored *k*-mer** problem

color: a subset of references (including singletons)



A directed bipartite graph

represent each non-singleton as a union of others



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

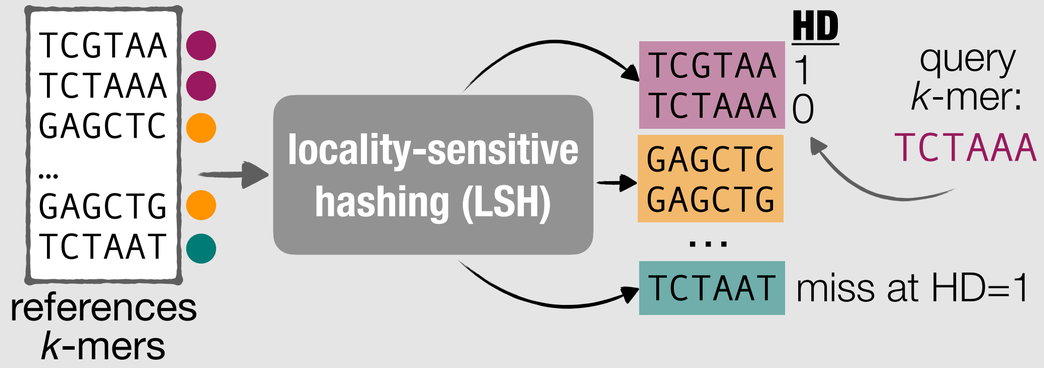
iii. follow edges to reconstruct colors

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

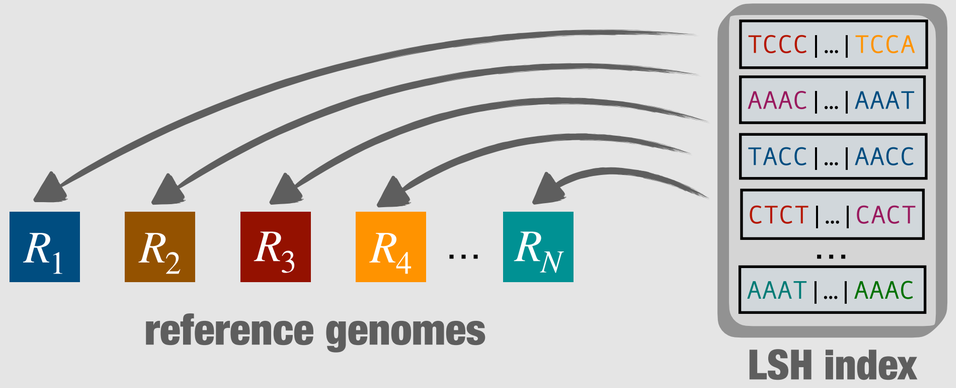
Outline of the method & subproblems we need to tackle

krepp: k-mer-based read phylogenetic placement

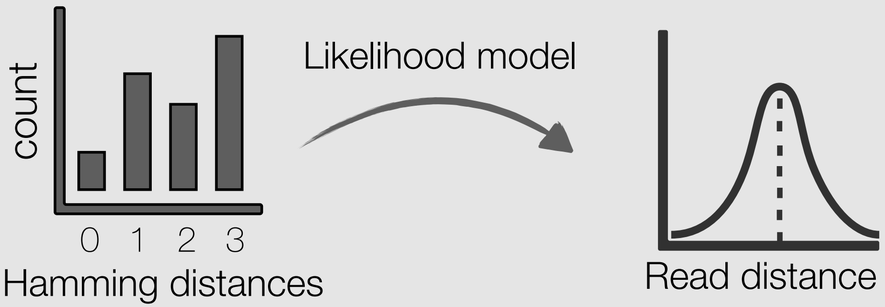
I) Hamming distances of homologous k-mers



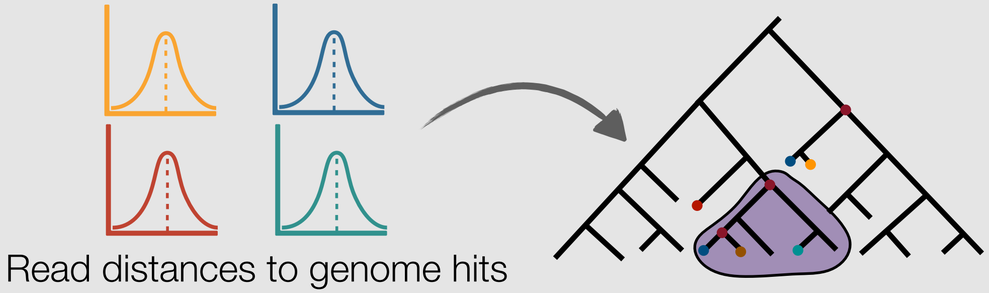
II) Mapping indexed k-mers to genomes



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



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



Finding homologous k-mers of reference genomes

Given a query sequence;

>q
ATACCTAGGAGTACGGGAC

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

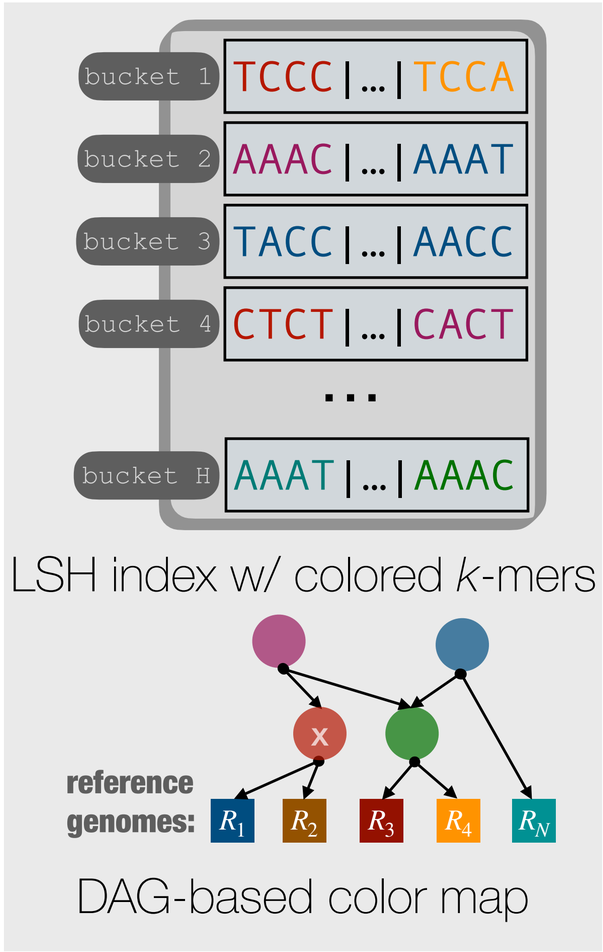
Finding homologous k-mers of reference genomes

Given a query sequence;

```
>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



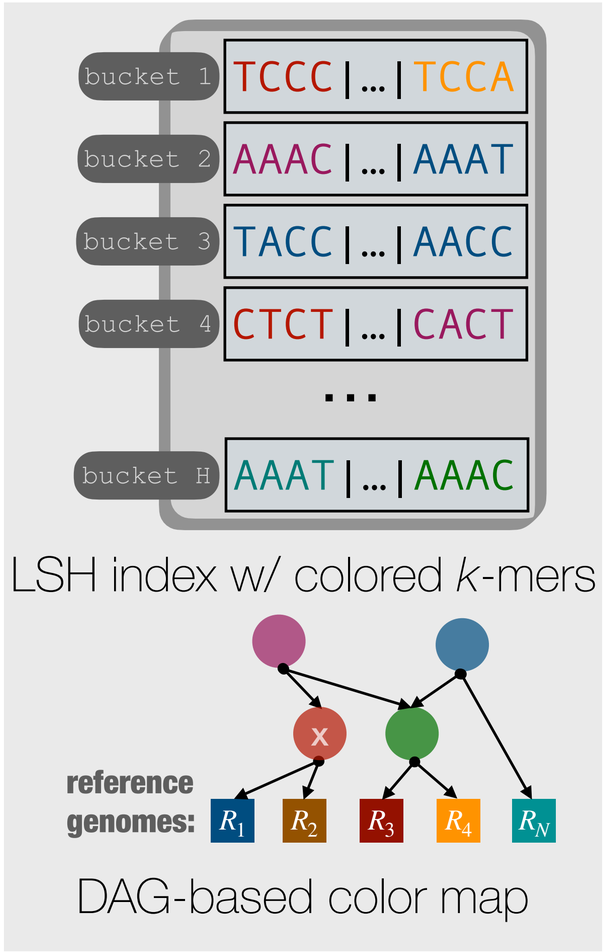
Finding homologous k-mers of reference genomes

Given a query sequence;

```
>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 homologous

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

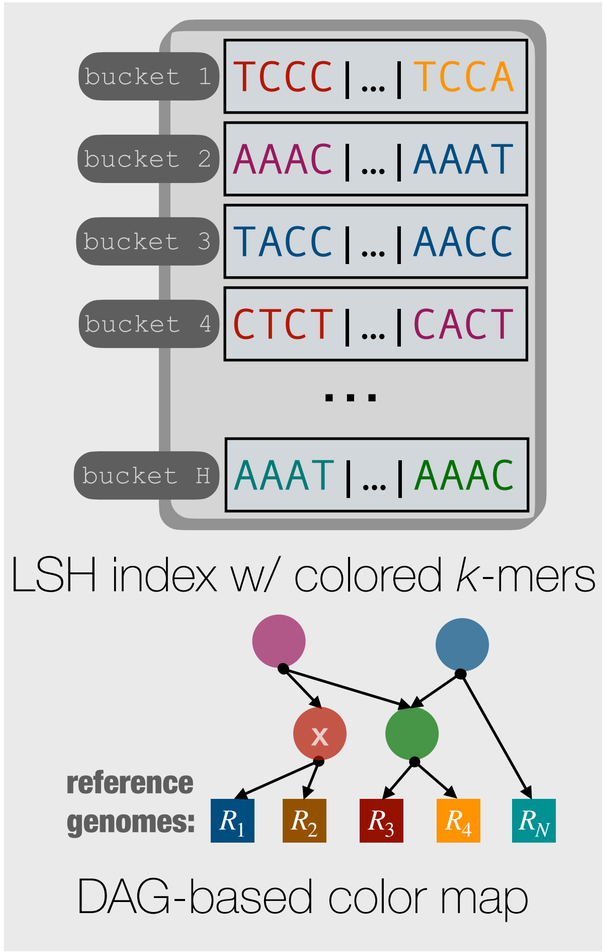
Finding homologous k-mers of reference genomes

Given a query sequence;

>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 homologous

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 $L - k + 1$ *k*-mers

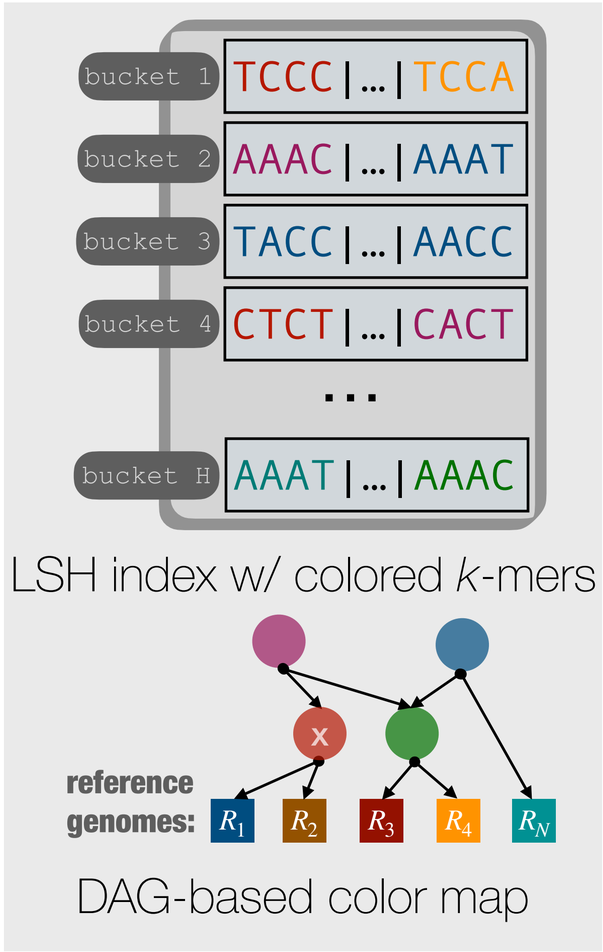
Finding homologous k-mers of reference genomes

Given a query sequence;

>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 homologous

sparse table

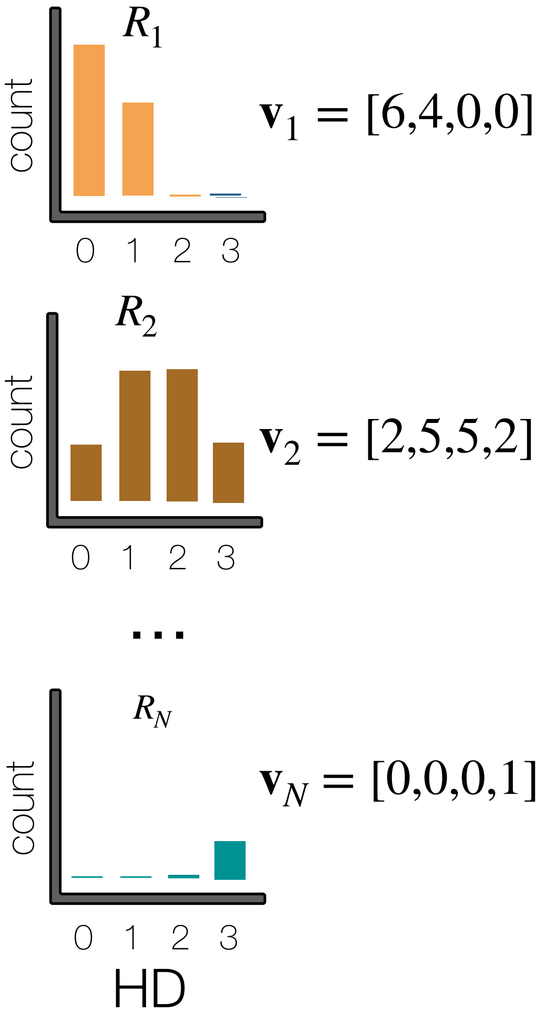
	R_1	R_2	…	R_N
1	0	1	…	-
2	1	4	…	4
3	-	-	…	-
4	-	2	…	-
5	0	-	…	3
…	…	…	…	…
L-k+1	3	0	…	4

Independence assumption: treat q as a bag of $L-k+1$ *k*-mers

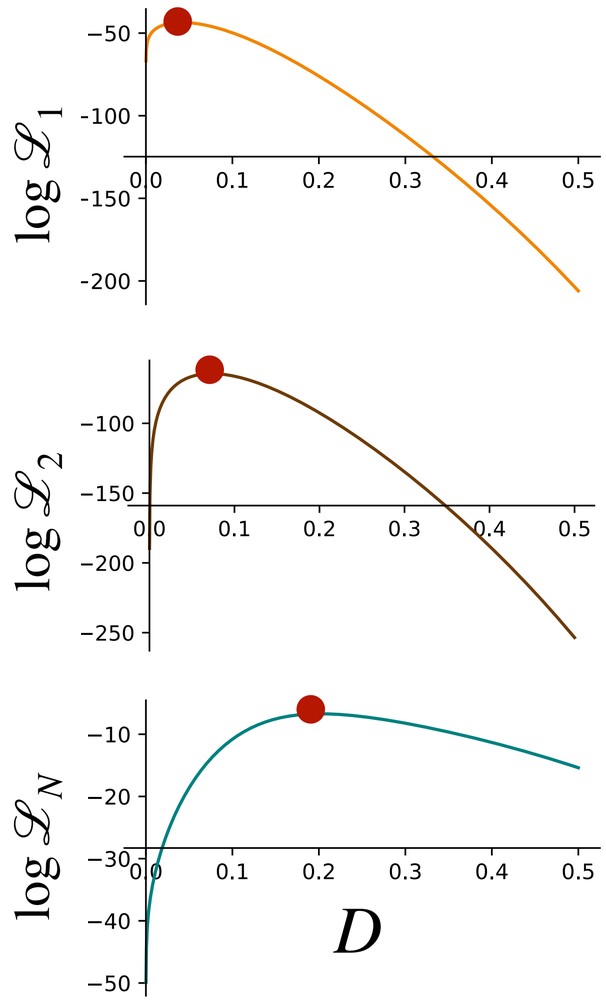
summarize each column as a histogram

Likelihood of k-mer matches & Hamming distances

Hamming distance histograms

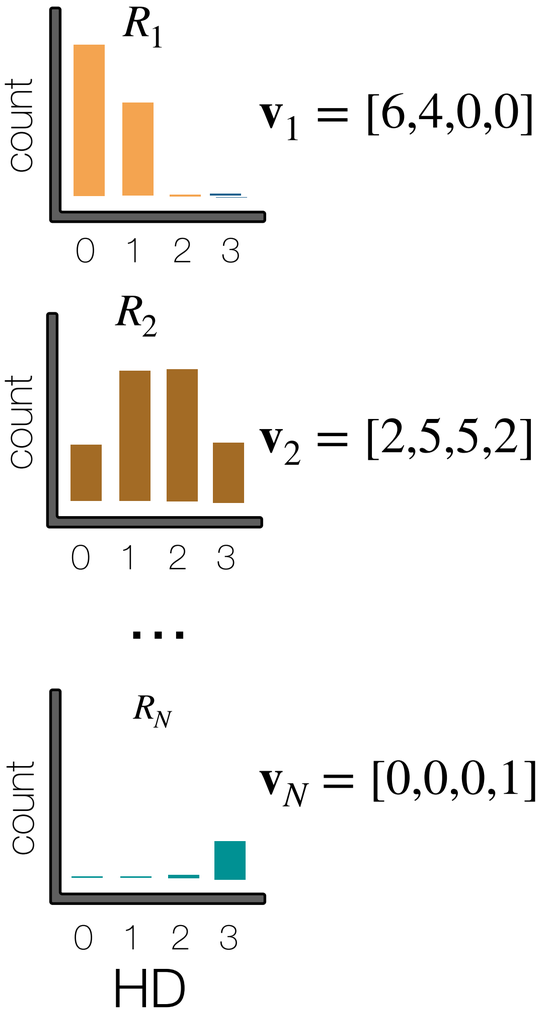


Goal: compute the likelihood of q having distance D to R_i



Likelihood of k-mer matches & Hamming distances

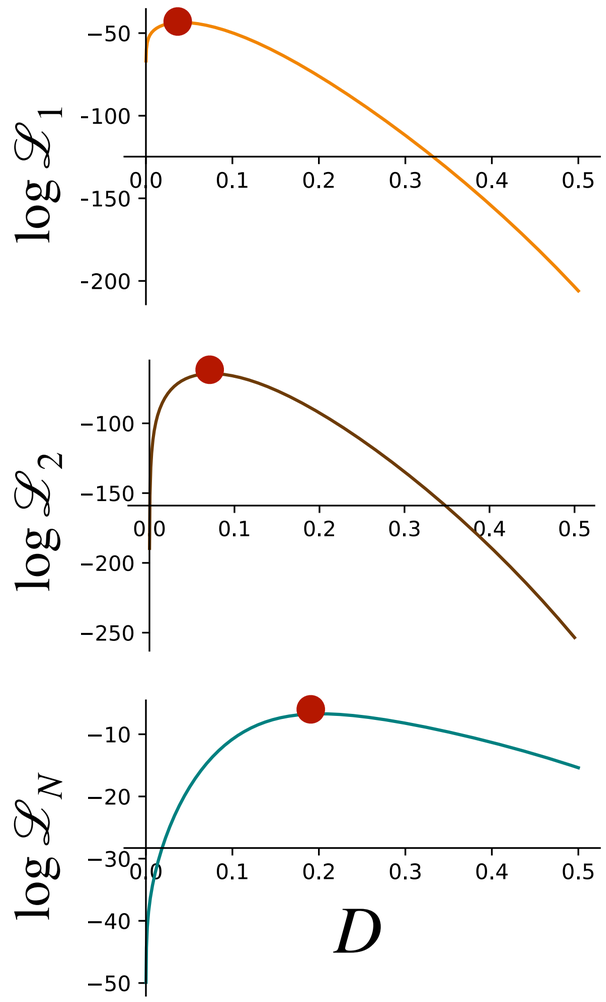
Hamming distance histograms



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

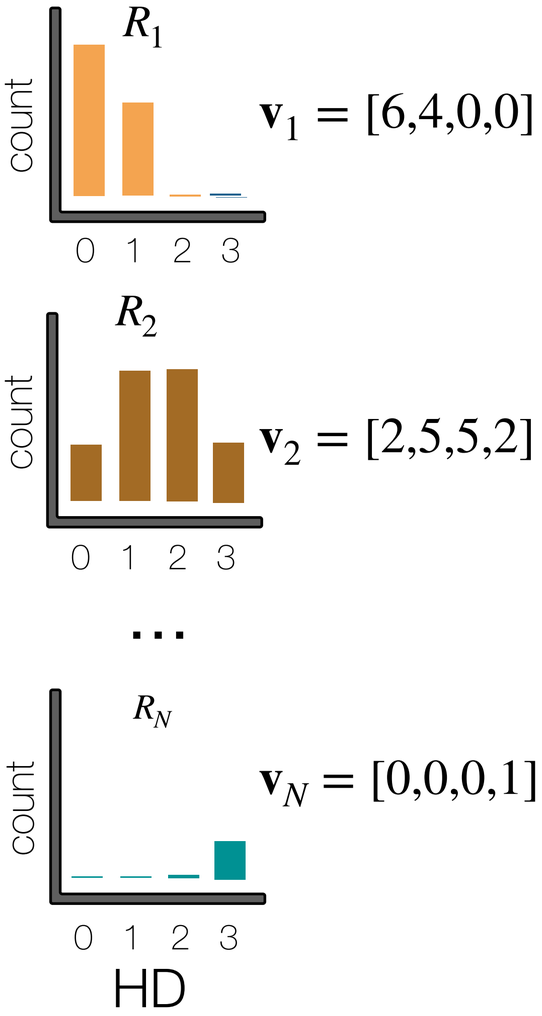
Likelihood of distance D to reference R_i : a product over all *k*-mers

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



Likelihood of k-mer matches & Hamming distances

Hamming distance histograms



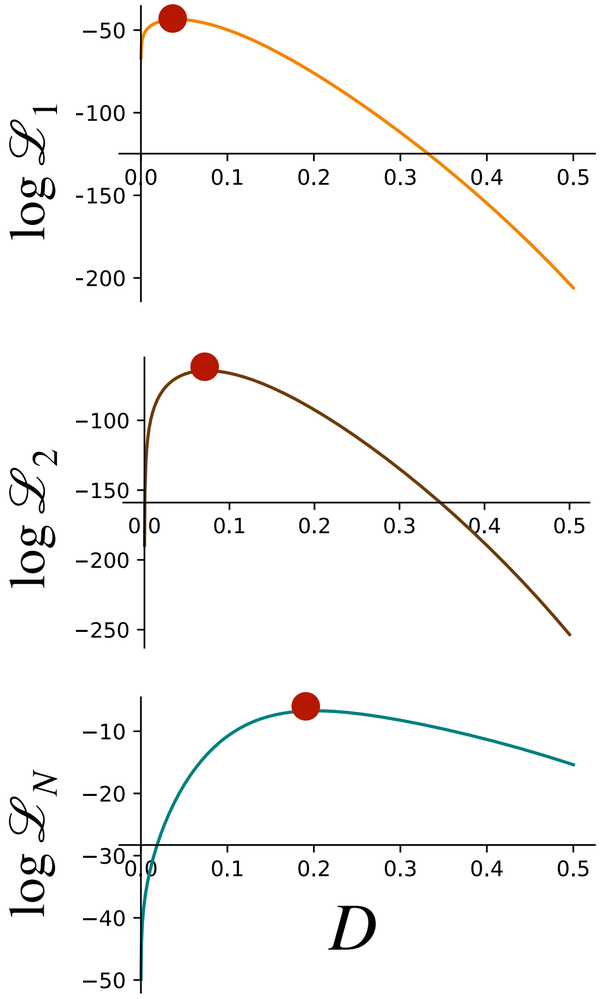
Goal: compute the likelihood of q having distance D to R_i

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

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

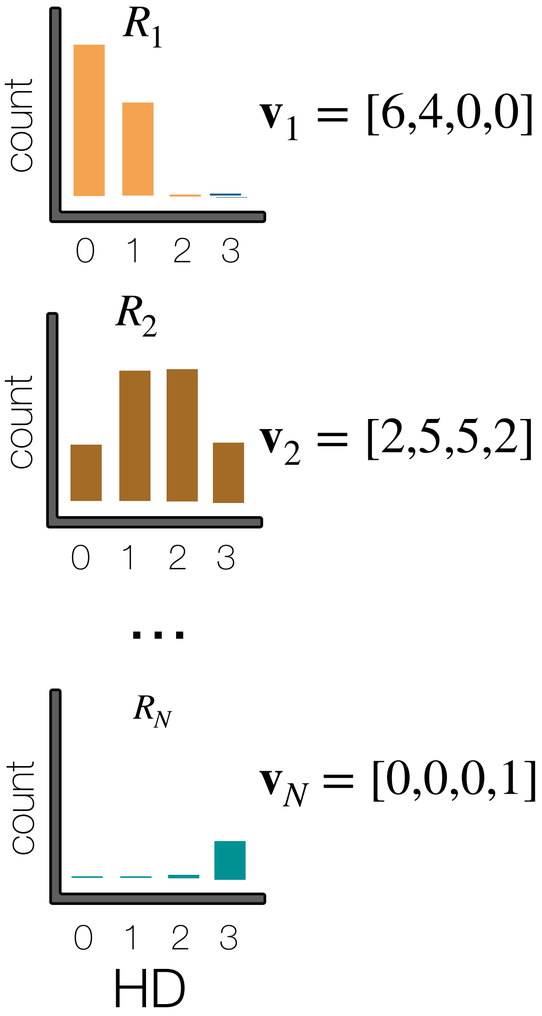
$$\mathscr{L}_i(D; k, h, \delta, u_i, \mathbf{v}_i) = \prod_{x=0}^{\delta} P_{match}(D; x, k, h)^{v_{i,x}}$$

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

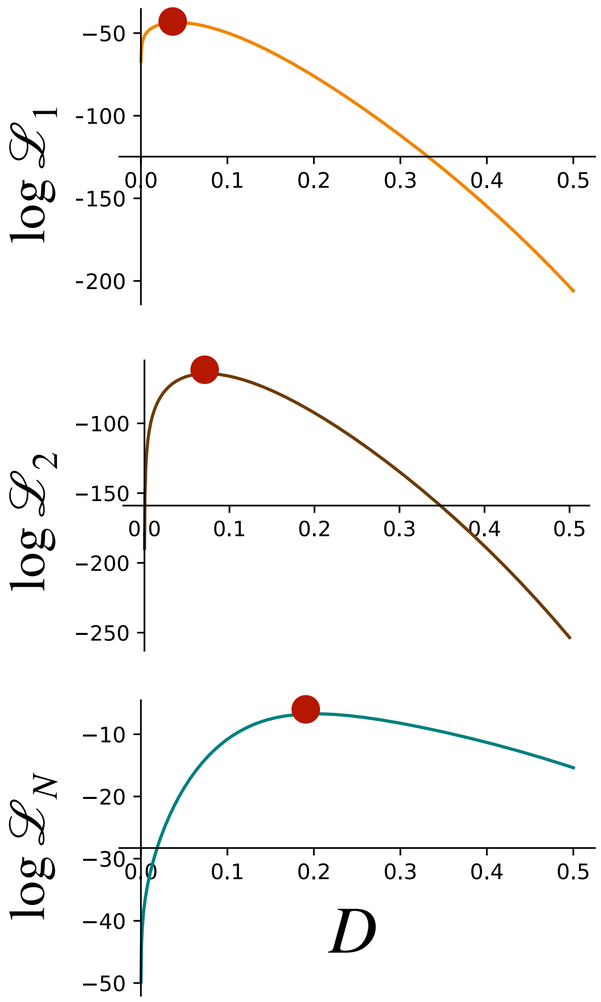
- $\mathbf{v}_i$: match count for each HD up to δ
- u_i: number of mismatches $(L-k+1) - \sum_{x=0}^{\delta} v_{i,x}$

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 $= x$



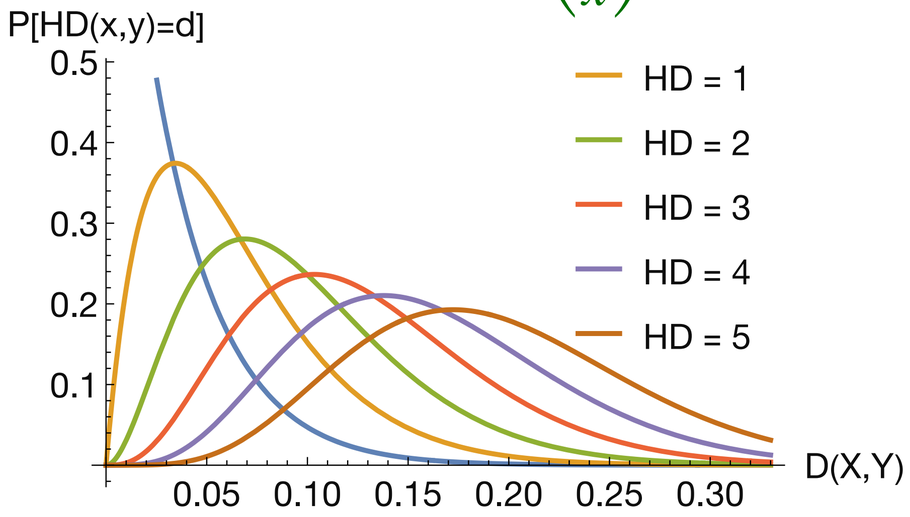
Observing k-mers matches with varying HDs

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

Observing k-mers matches with varying HDs

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

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

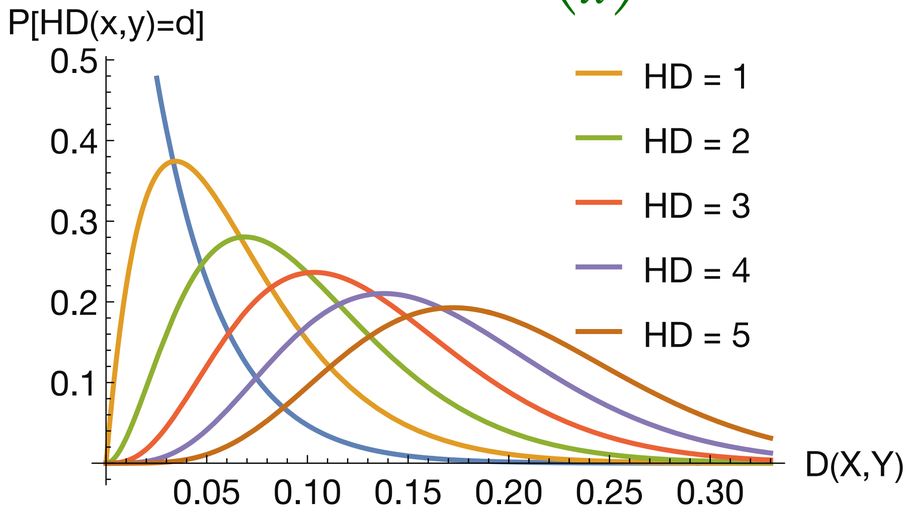


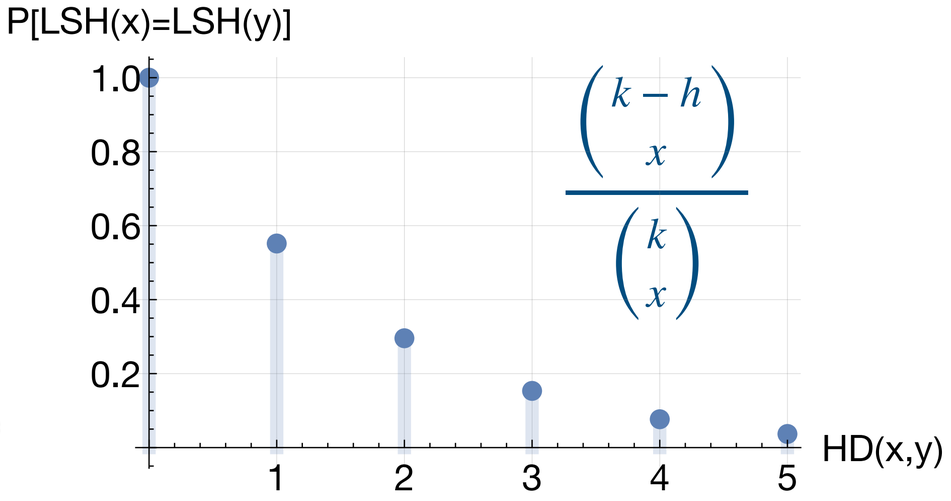
Observing k-mers matches with varying HDs

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

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

1-FNR of locality-sensitive hashing for HD $= x$





Observing k-mers matches with varying HDs

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

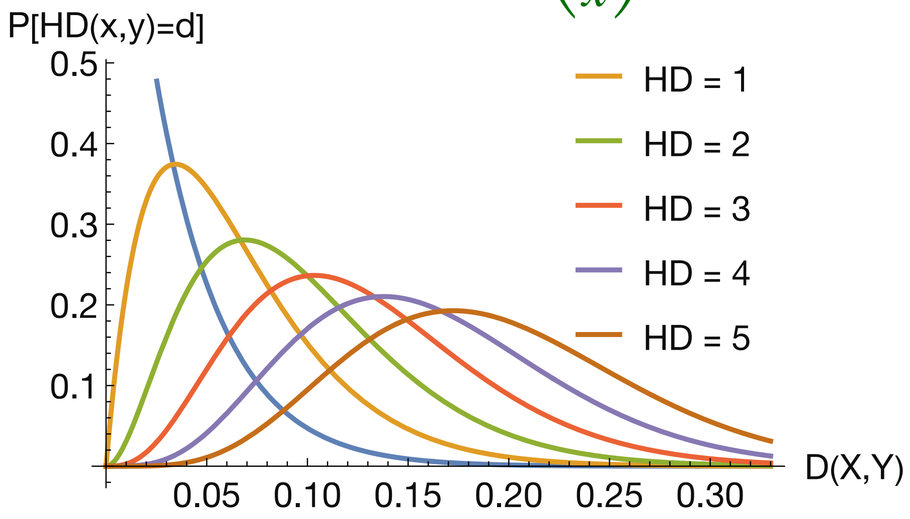
$$\rho_i = \frac{\text{\# of subsampled}}{\text{\# of distinct}}$$

precomputed for R_i

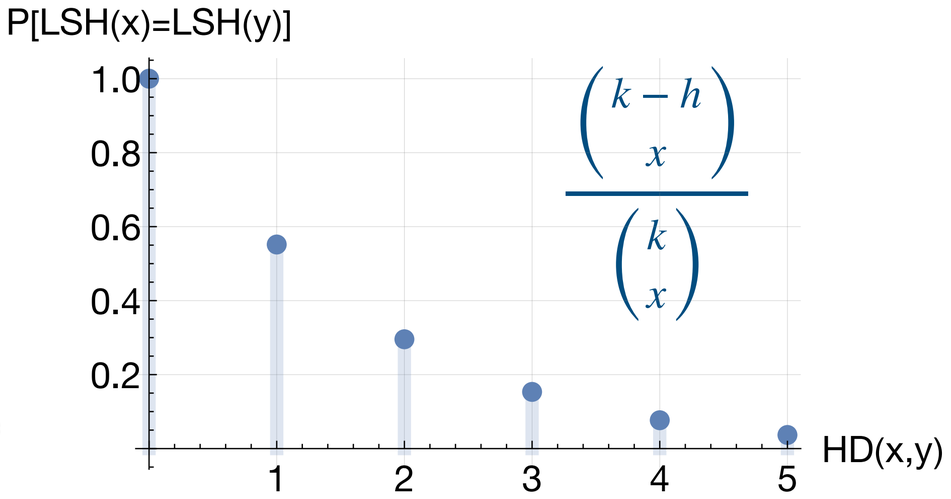
→ not all *k*-mers have to be indexed:

- minimizers
- FracMinHash
- …

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



1-FNR of locality-sensitive hashing for HD $= x$



Multiple events could lead to a mismatch

A mismatch occurs for two *k*-mers (query a and reference b), if

Multiple events could lead to a mismatch

A mismatch occurs for two *k*-mers (query a and reference b), if

- b is not indexed: $1 - \rho$ ***or***

Multiple events could lead to a mismatch

A mismatch occurs for two *k*-mers (query a and reference b), if

- b is not indexed: $1 - \rho$ ***or***
- b 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 two *k*-mers (query a and reference b), if

- b is not indexed: $1 - \rho$ ***or***
- b is indexed: ρ, but either:

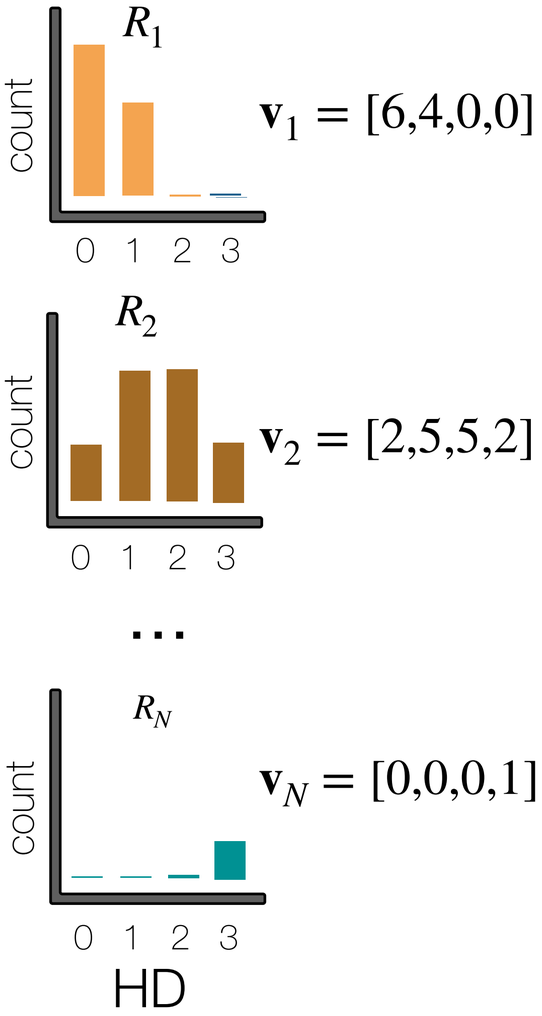
i) $\mathrm{HD}(a, b) > \delta$: $\sum_{x=\delta+1}^{k} P_{mutate}(D; x, k)$ ***or***

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

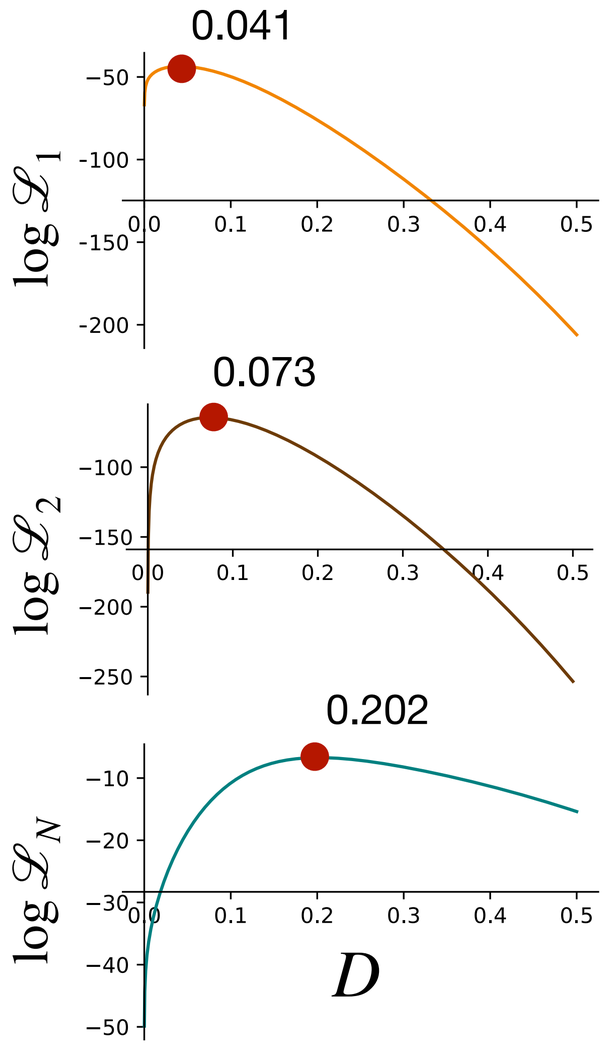


Optimize $-\log \mathscr{L}_i$ w.r.t. D

for each hitting reference R_i:

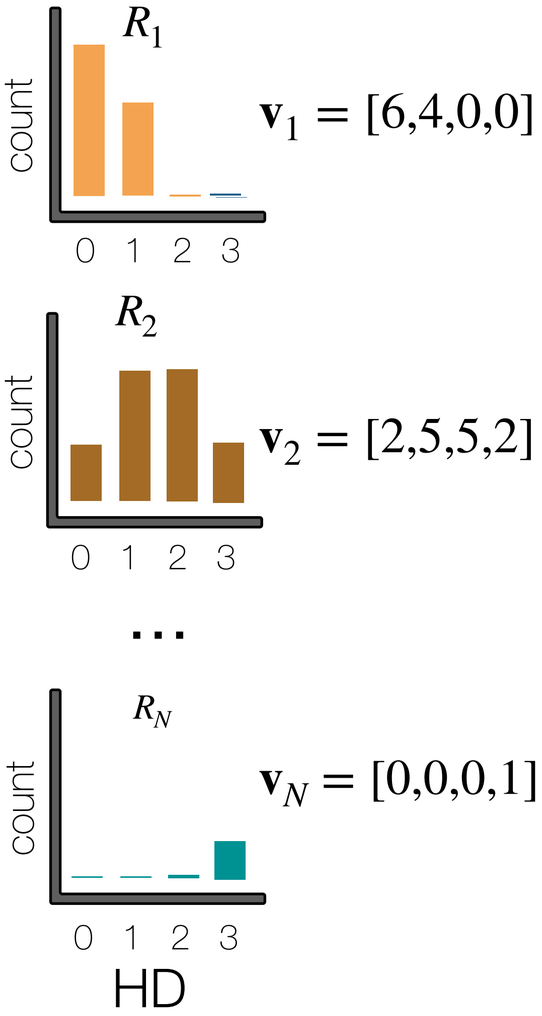
$$\arg\max_D \; P_{miss}(D;k,h,\delta)^{u_i} \prod_{x=0}^{\delta} P_{match}(D;x,k,h)^{v_{i,x}}$$

single variable & convex with a sensible choice of parameters

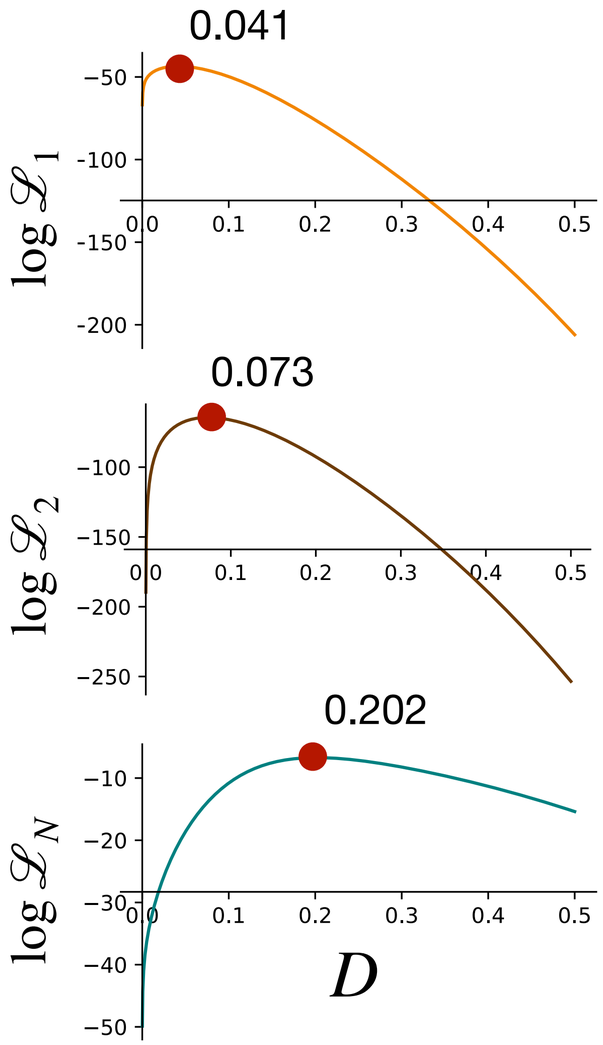


Maximum likelihood estimation of distances

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)

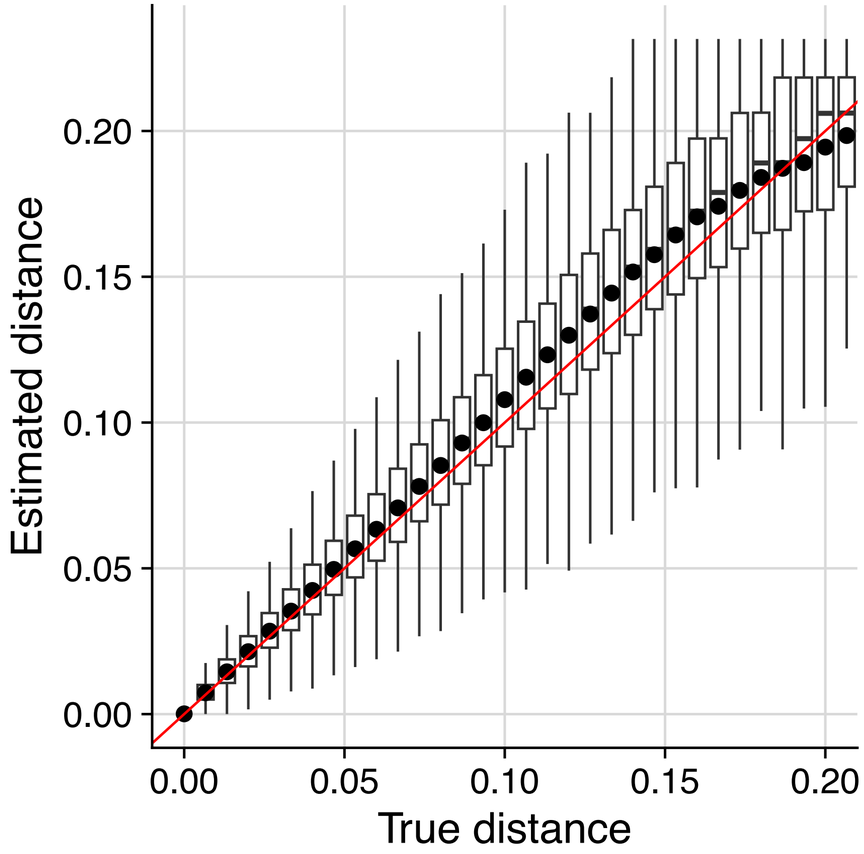
krepp estimates distances accurately at the read-level

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)



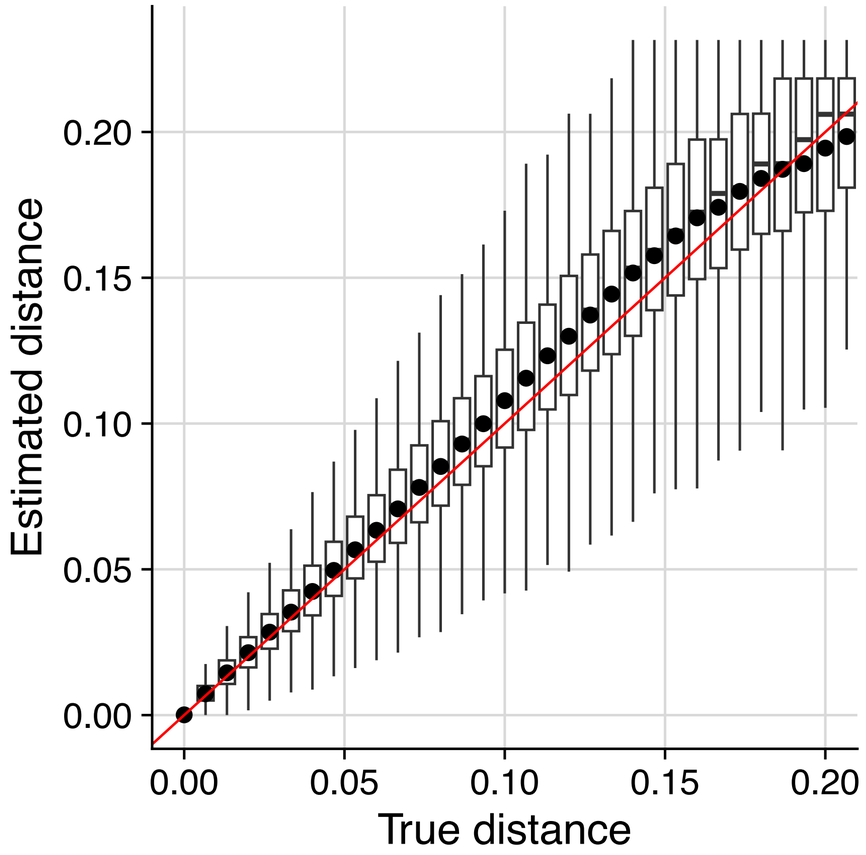
krepp estimates distances accurately at the read-level

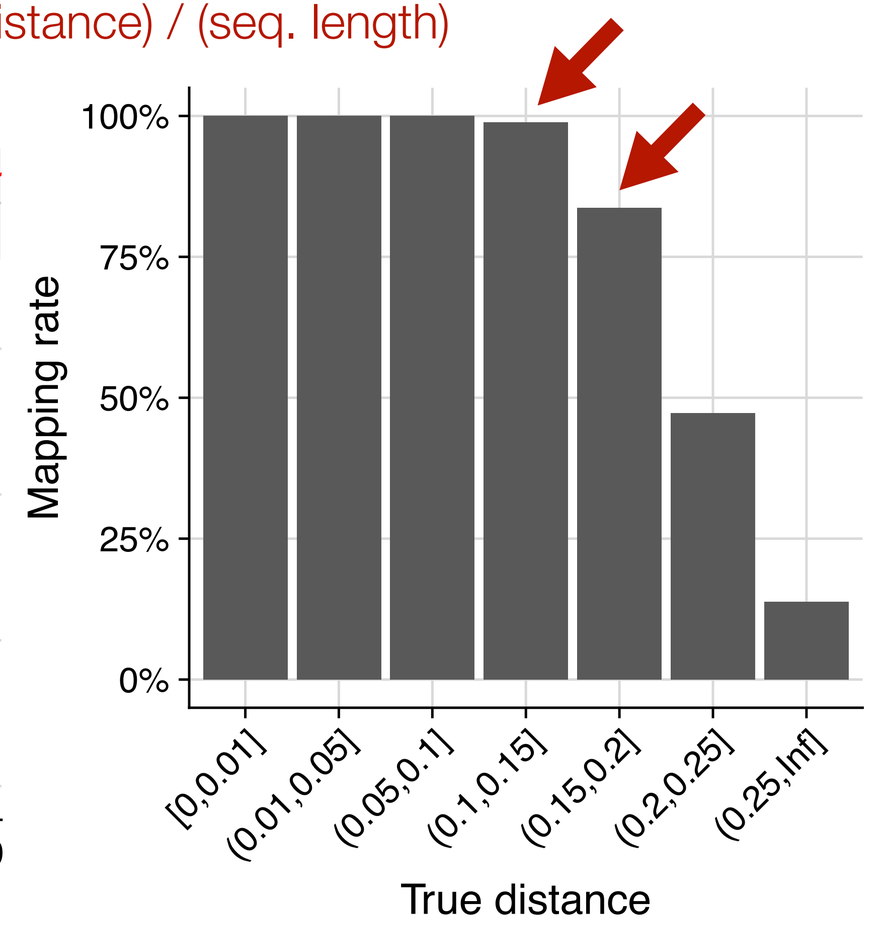
default: 29-mer
minimizers of 35-mers

- 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)



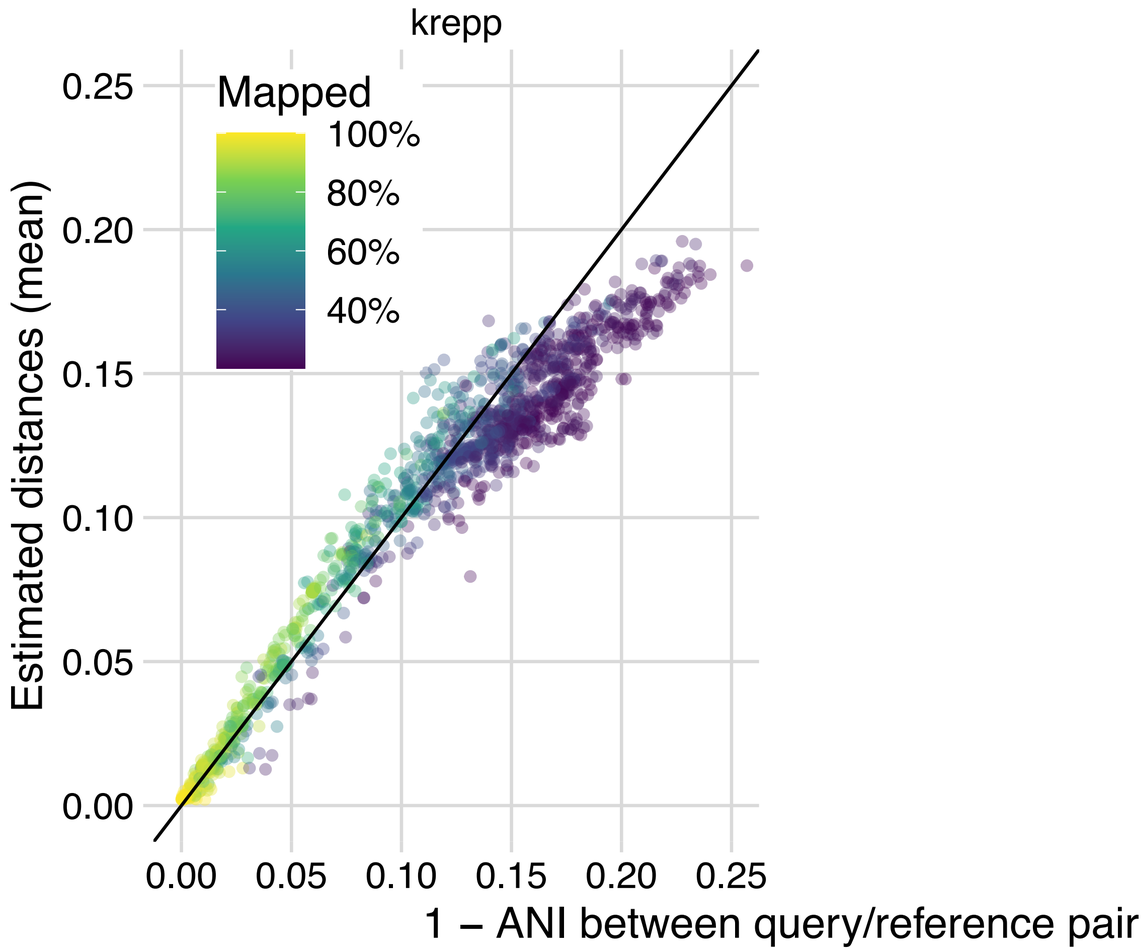


krepp matches nucleotide identity on average for real genomes

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

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

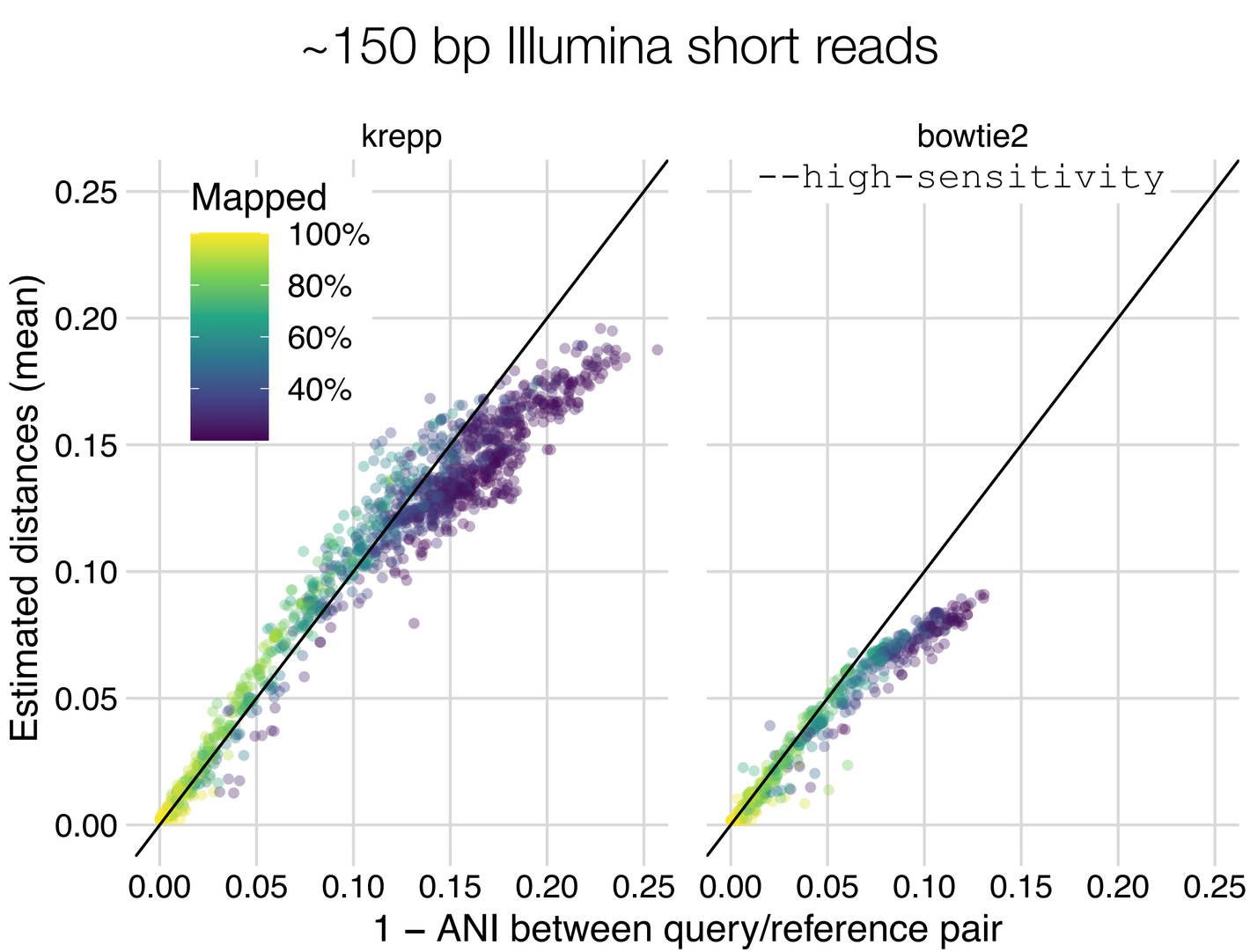
~150 bp Illumina short reads



krepp matches nucleotide identity on average for real genomes

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

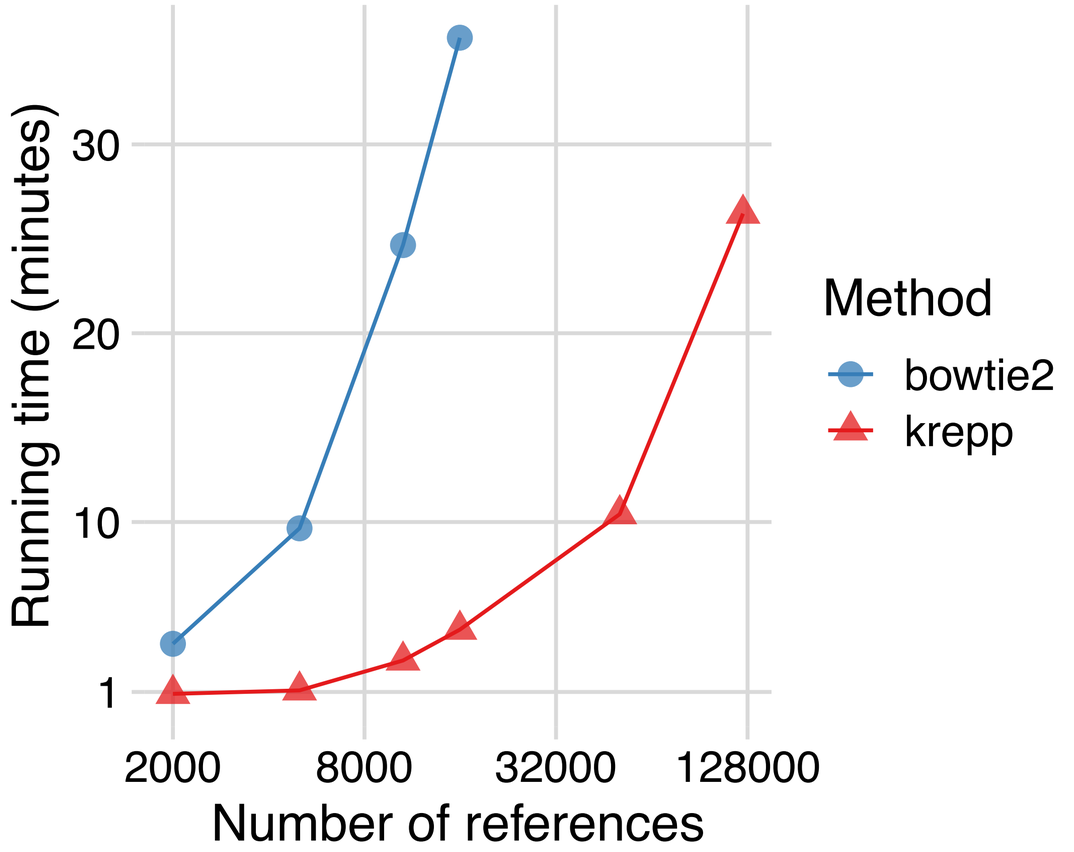
- Real query/reference genomes (pairs with >20% mapping rate)
- *krepp* extends to distant (>10%) reference genomes accurately
- Increasing the sensitivity by relaxing the alignment is costly



krepp matches nucleotide identity on average for real genomes

- Real query/reference genomes (pairs with >20% mapping rate)
- *krepp* extends to distant (>10%) reference genomes accurately
- Increasing the sensitivity by relaxing the alignment is costly
- *krepp* is already **>10x faster, scales well w/ large references**

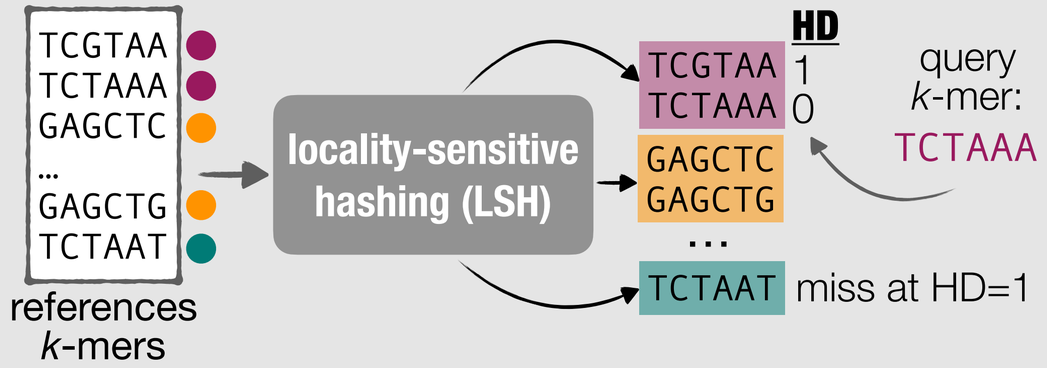
Mapping 10M reads (16 threads):



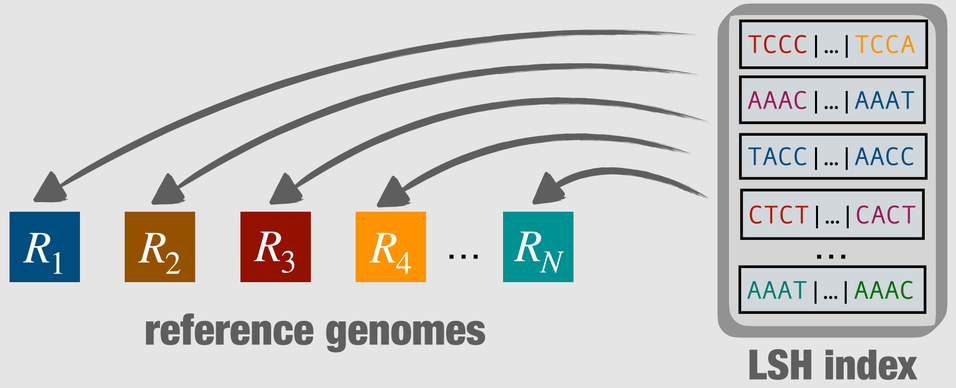
Outline of the method & subproblems we need to tackle

krepp: k-mer-based read phylogenetic placement

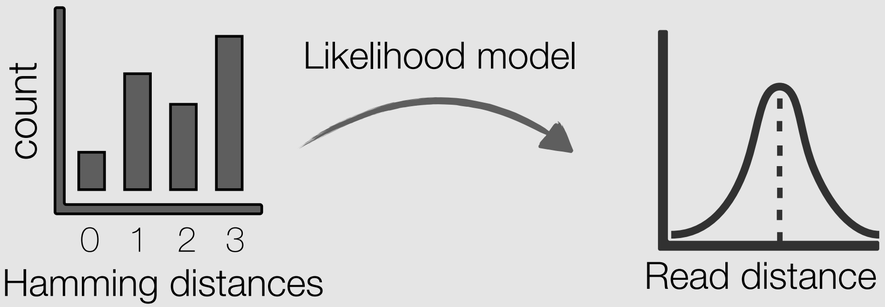
I) Hamming distances of homologous k-mers



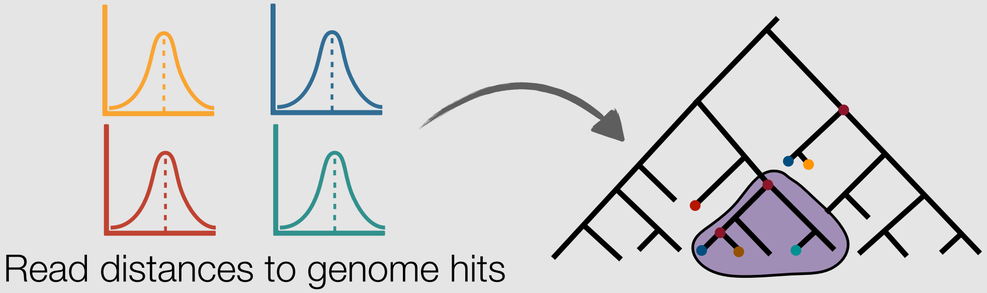
II) Mapping indexed k-mers to genomes



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



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



Statistical distinguishability and distance-based placement

- short reads — **low signal**
- high distances — many genome hits
- small differences may not be statistically meaningful
  - **test distinguishability**

Statistical distinguishability and distance-based placement

- short reads — **low signal**
- high distances — many genome hits
- 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 and distance-based placement

- short reads — **low signal**
- high distances — many genome hits
- 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

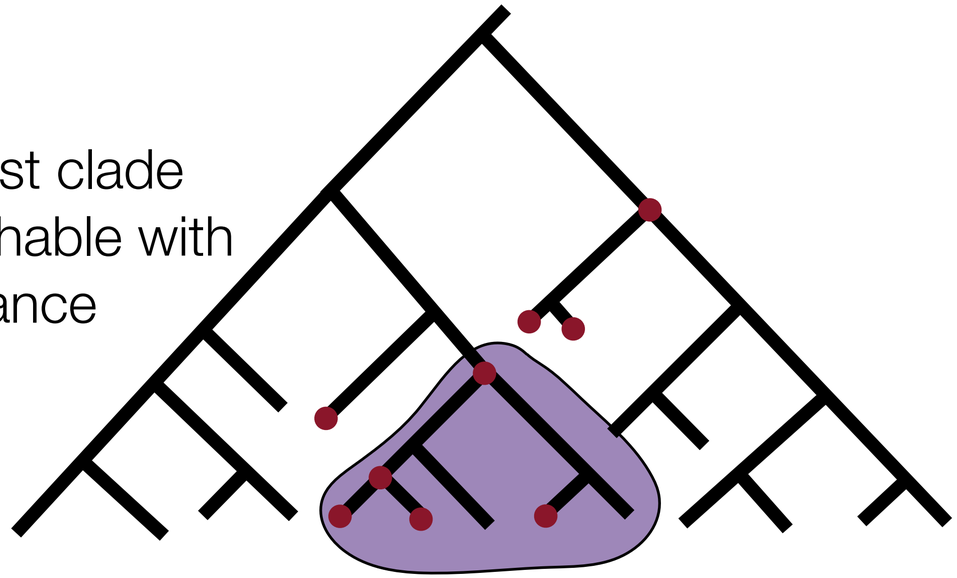
D: alternative distance

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

$i*$: closest reference

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

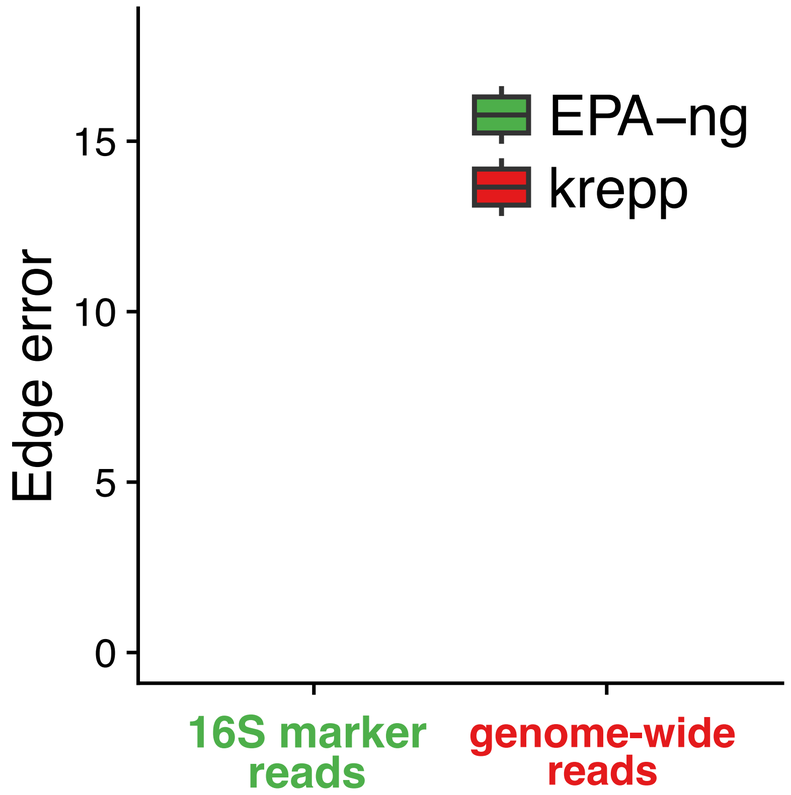
- select a significance level (default: α=90%)



● : indistinguishable w.r.t. the closest reference

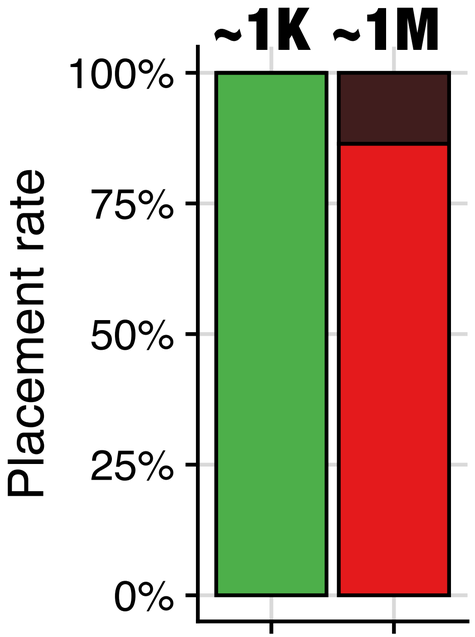
krepp places genome-wide reads more accurately than ML-based of 16S placement

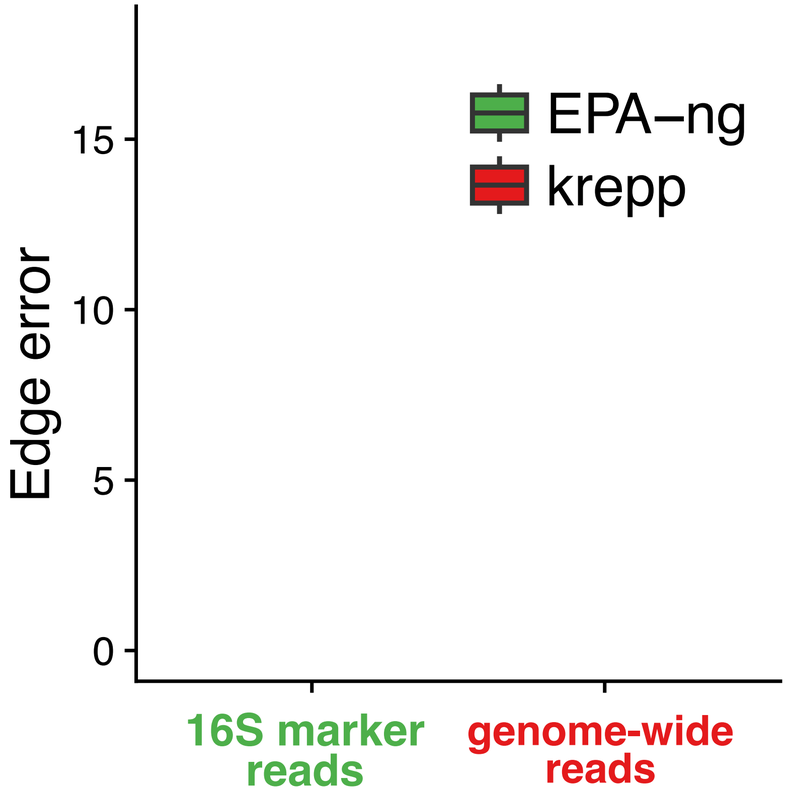
- Leave all out —100 queries from 11,000 taxa (WoLv1)
- EPA-ng: needs a MSA; only markers



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

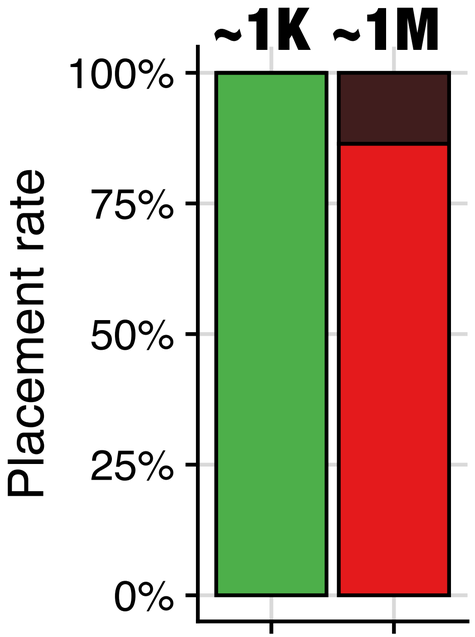
- Leave all out —100 queries from 11,000 taxa (WoLv1)
- EPA-ng: needs a MSA; only markers
- ***krepp*** **places 86% of all reads**

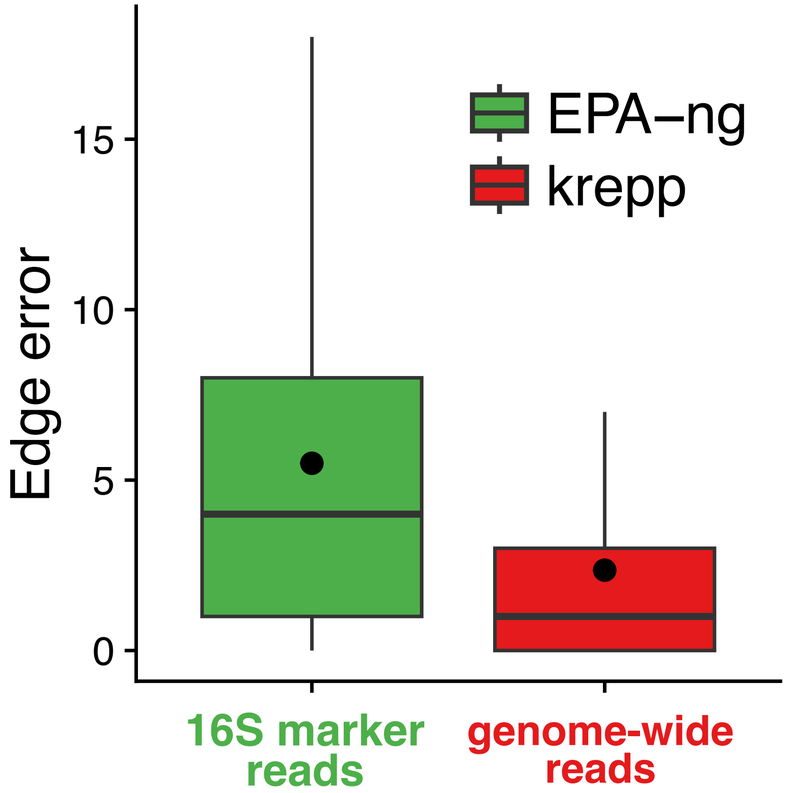




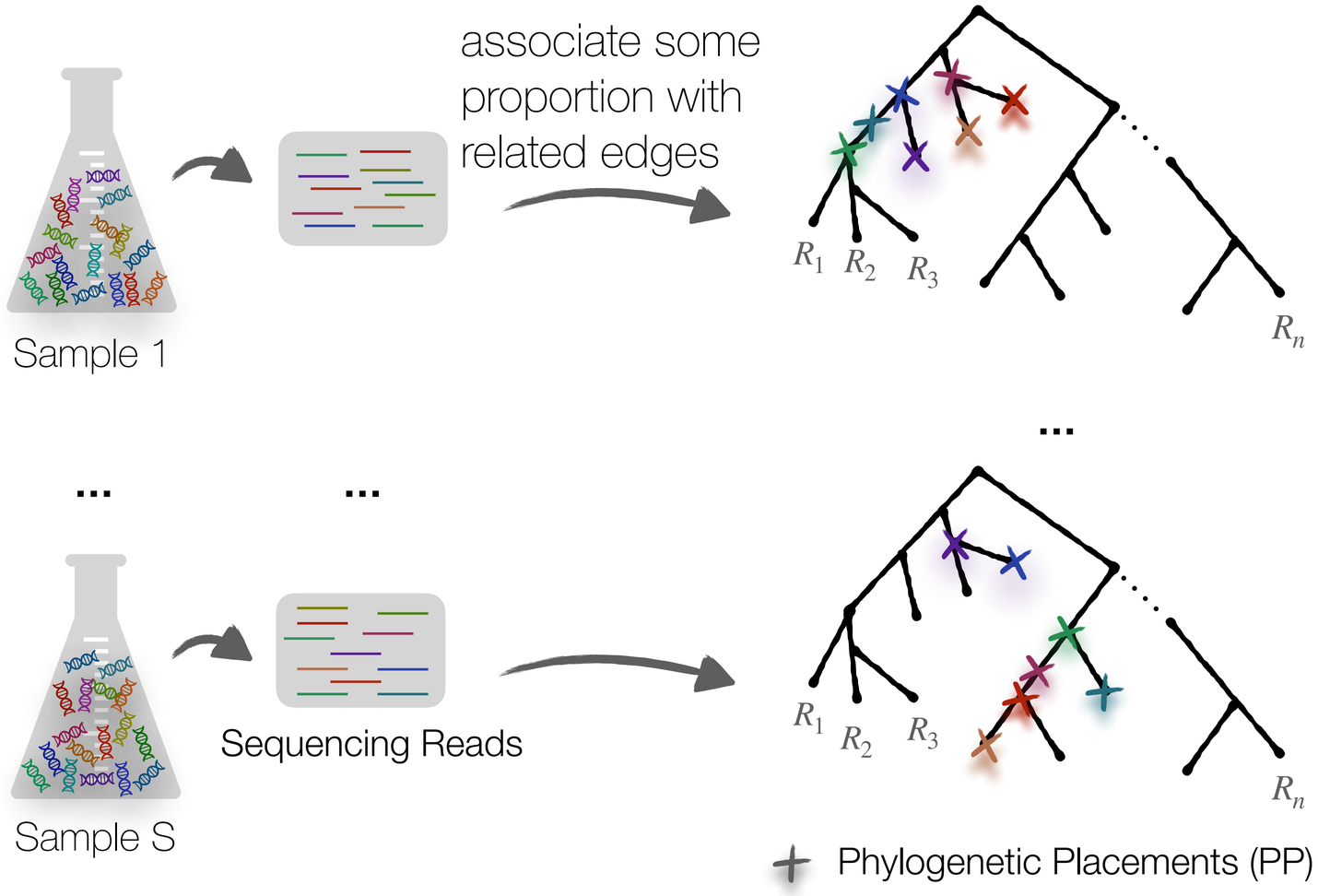
krepp places genome-wide reads more accurately than ML-based of 16S placement

- Leave all out —100 queries from 11,000 taxa (WoLv1)
- EPA-ng: needs a MSA; only markers
- ***krepp*** **places 86% of all reads**
- **2.4 vs. 5.6 edge error (average)**

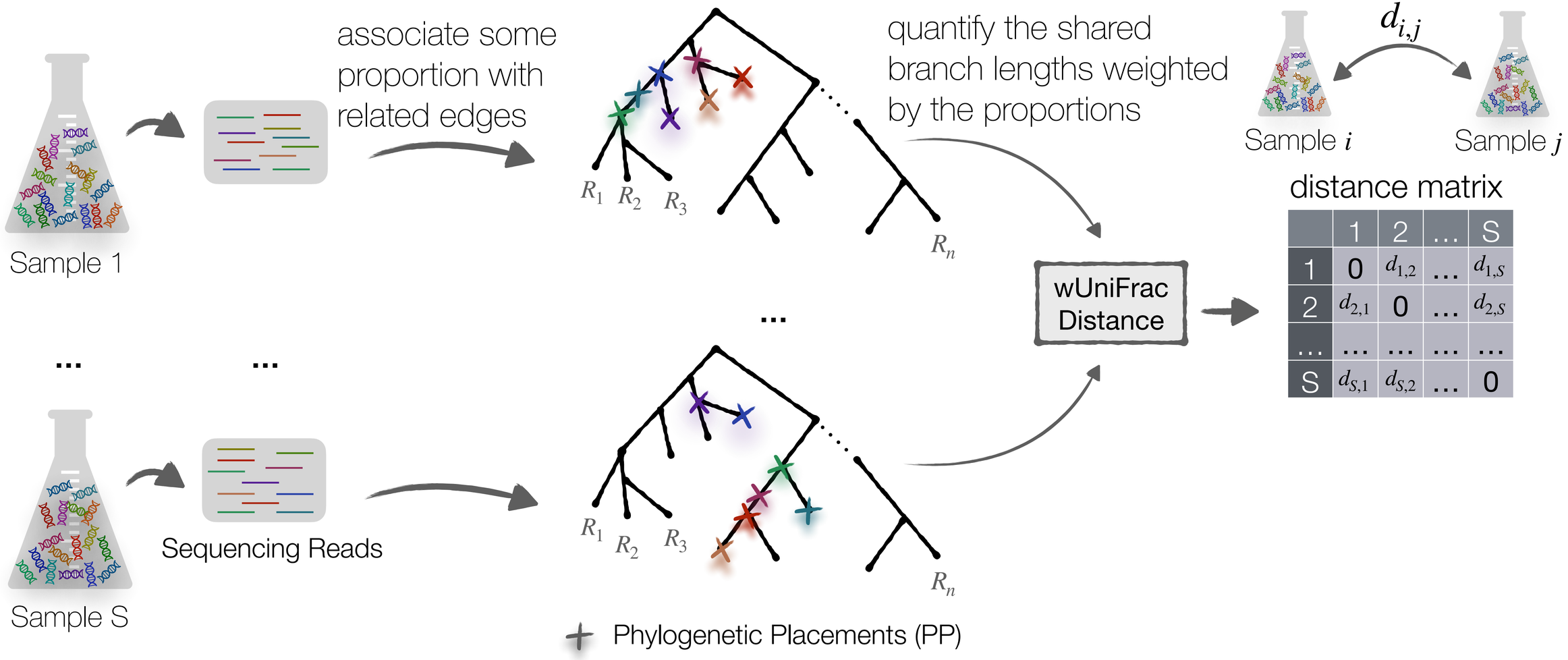




Characterization of metagenomic samples on a phylogeny

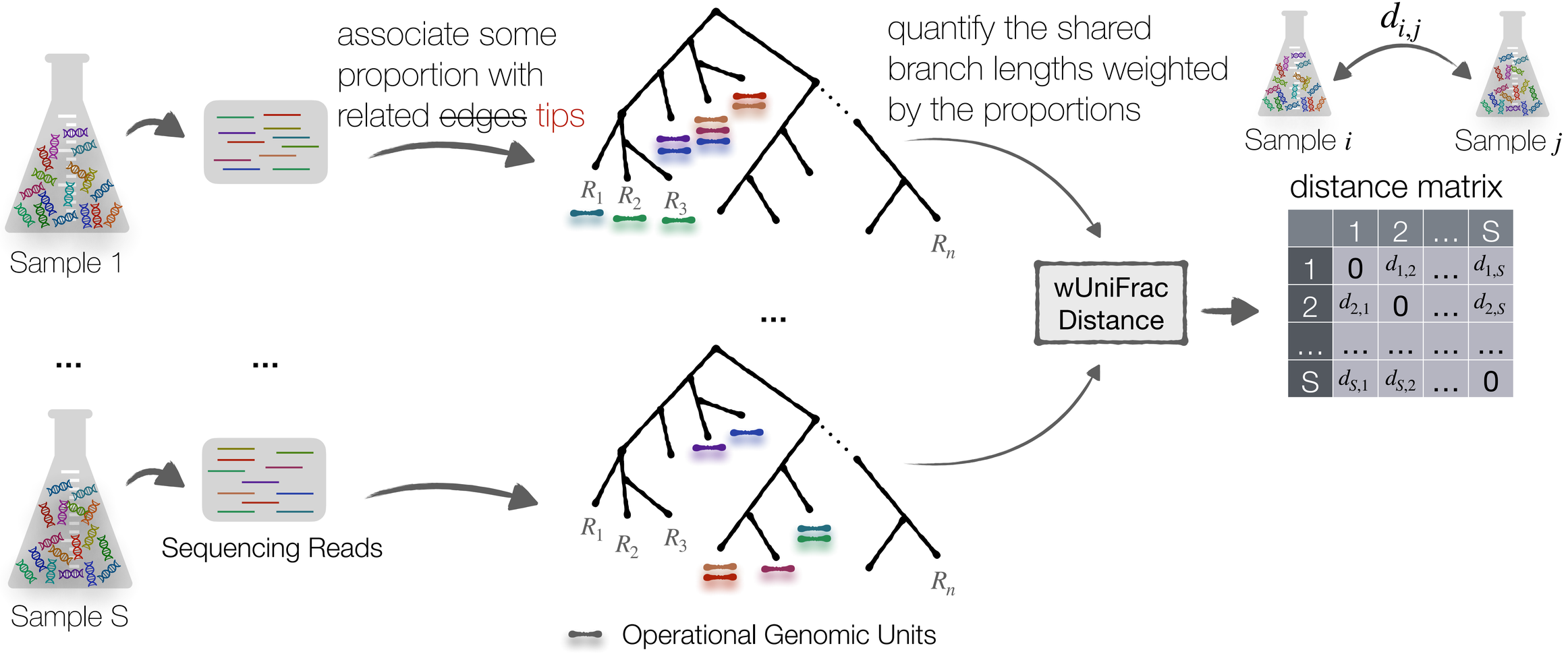


Characterization of metagenomic samples on a phylogeny



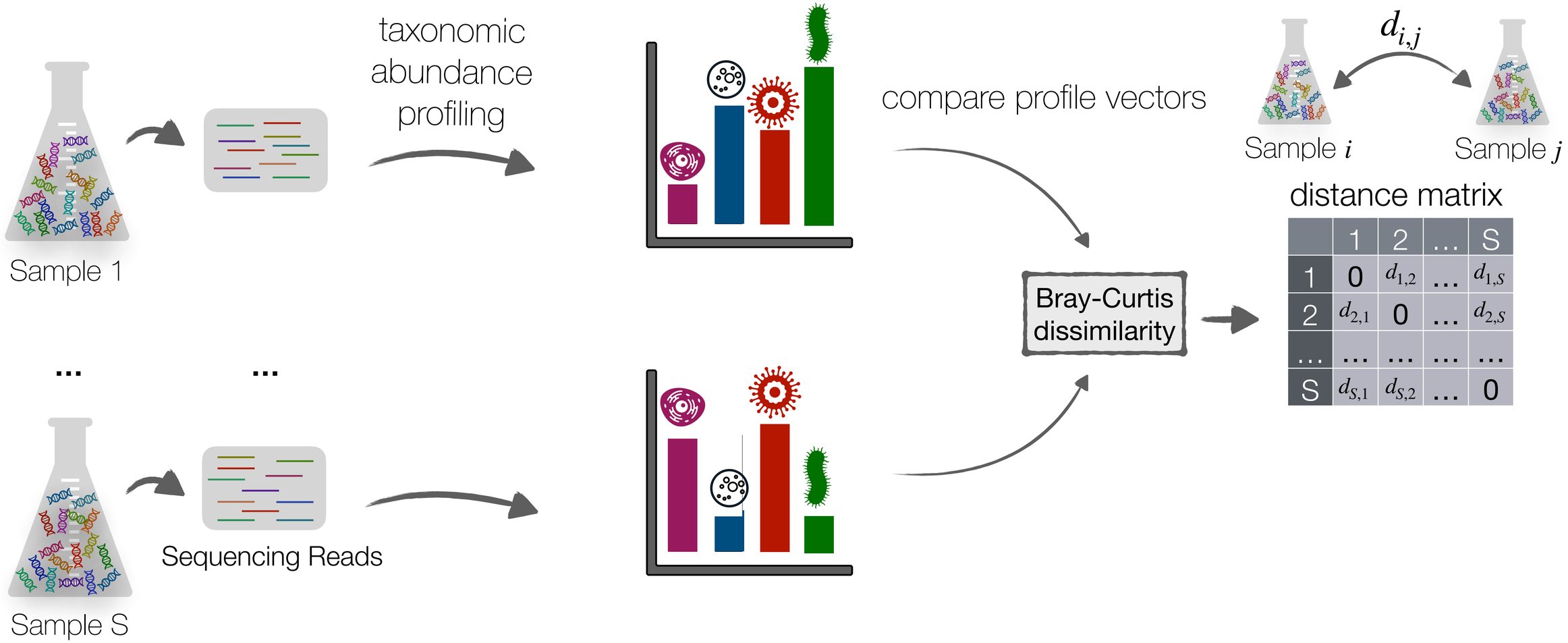
	1	2	…	S
1	0	$d_{1,2}$	…	$d_{1,S}$
2	$d_{2,1}$	0	…	$d_{2,S}$
…	…	…	…	…
S	$d_{S,1}$	$d_{S,2}$	…	0

Characterization of metagenomic samples on a phylogeny



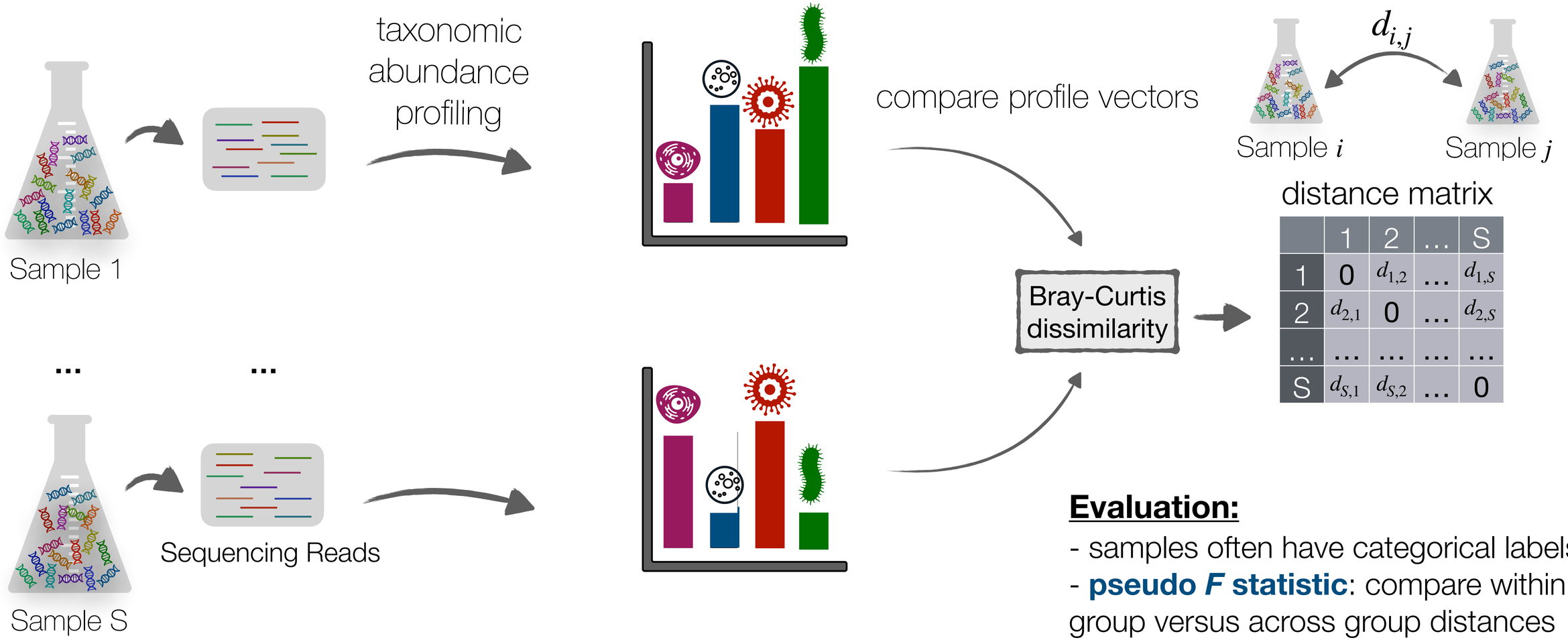
	1	2	…	S
1	0	$d_{1,2}$	…	$d_{1,S}$
2	$d_{2,1}$	0	…	$d_{2,S}$
…	…	…	…	…
S	$d_{S,1}$	$d_{S,2}$	…	0

Characterization of metagenomic samples on a phylogeny



	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

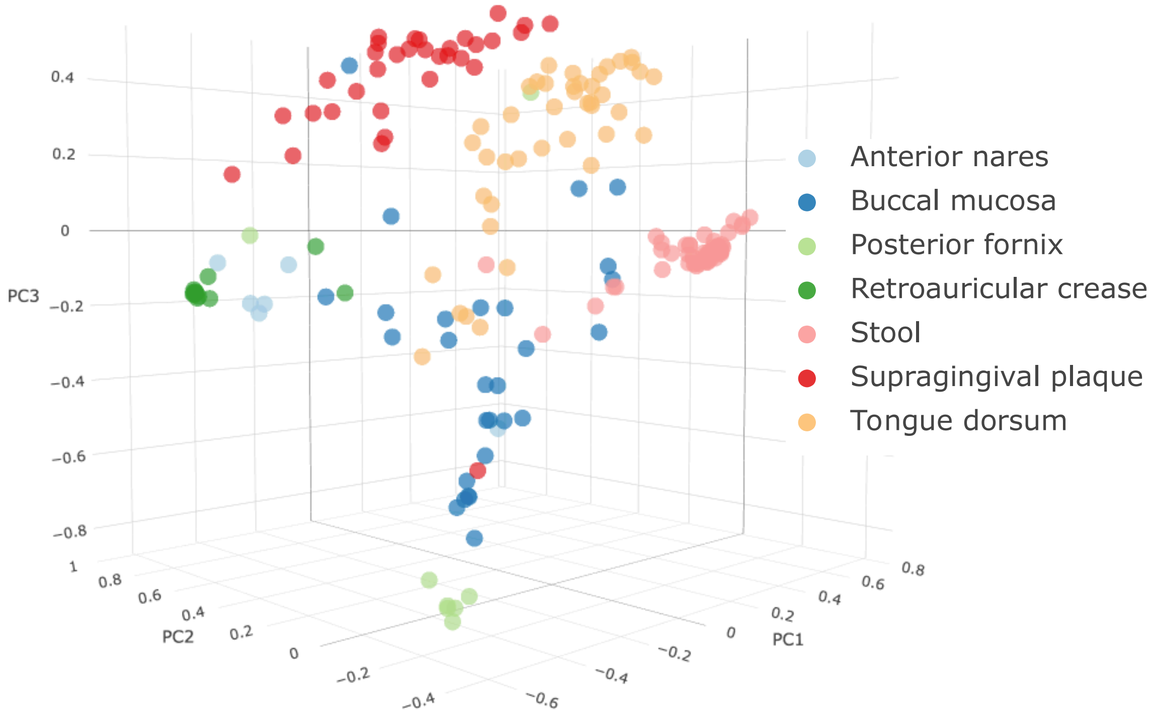
Characterization of metagenomic samples on a phylogeny



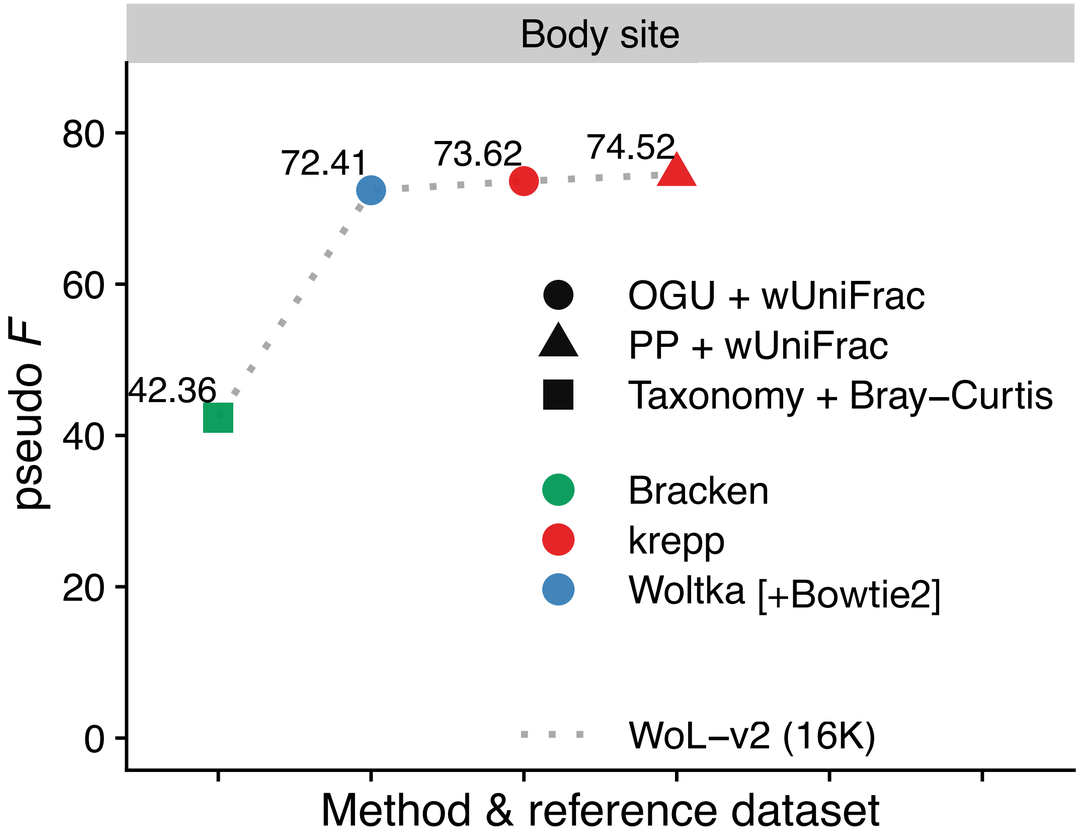
Evaluation:

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

Analyzing human microbiome with larger references

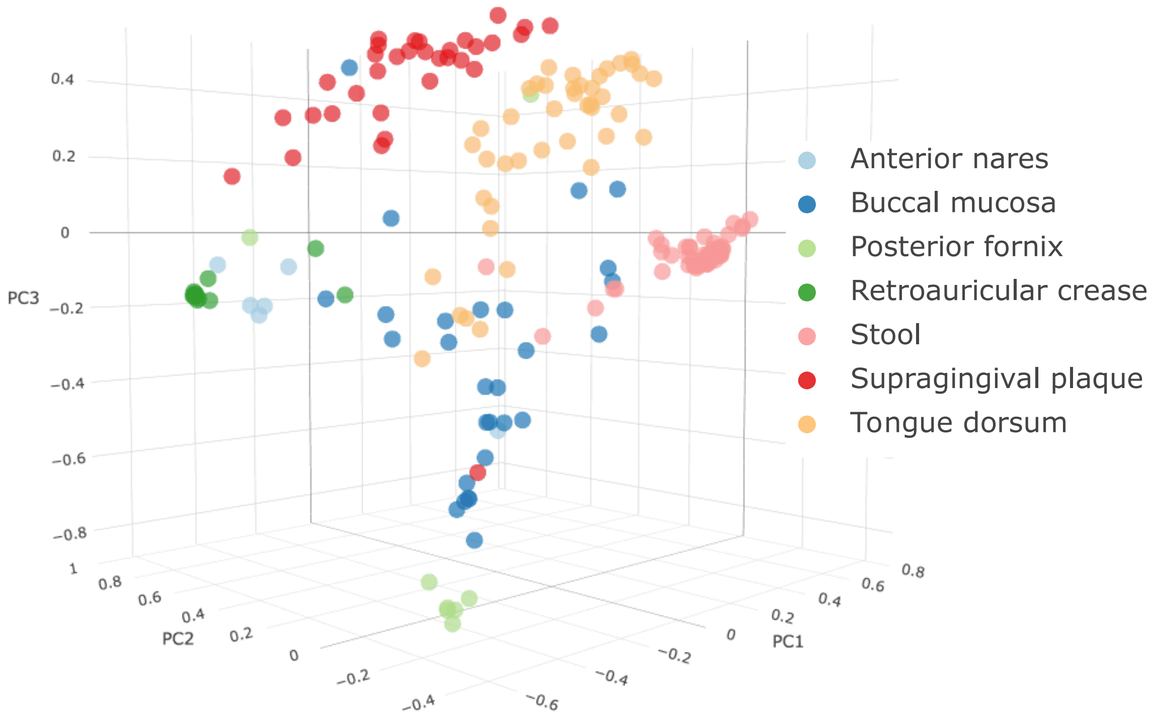


Reference: Web of Life (v2)
16,000 microbial genomes

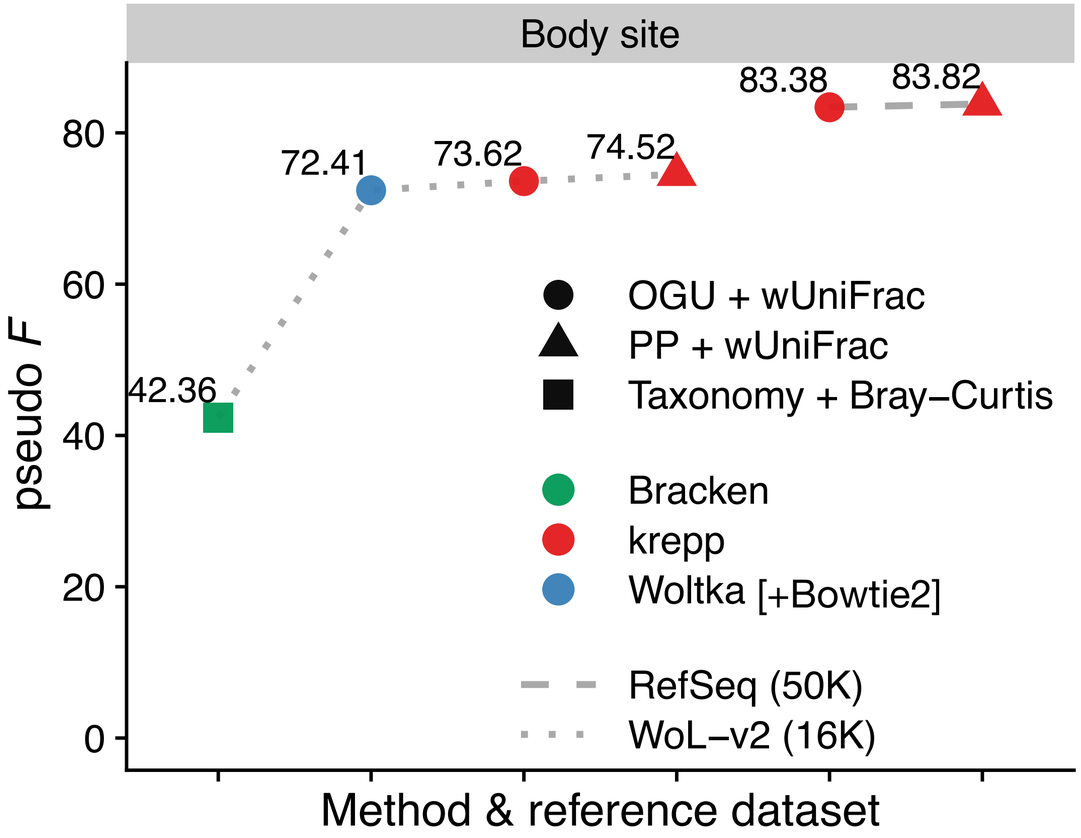


Analyzing human microbiome with larger references

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

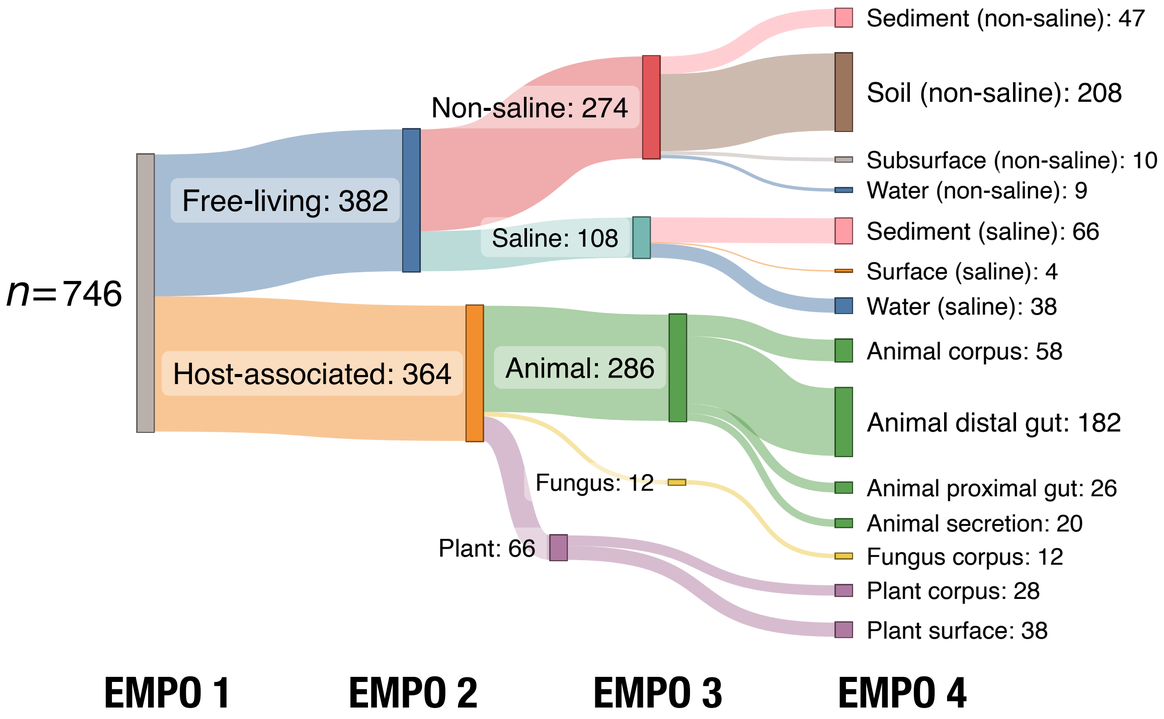


Reference: RefSeq subset
50,000 microbial genomes

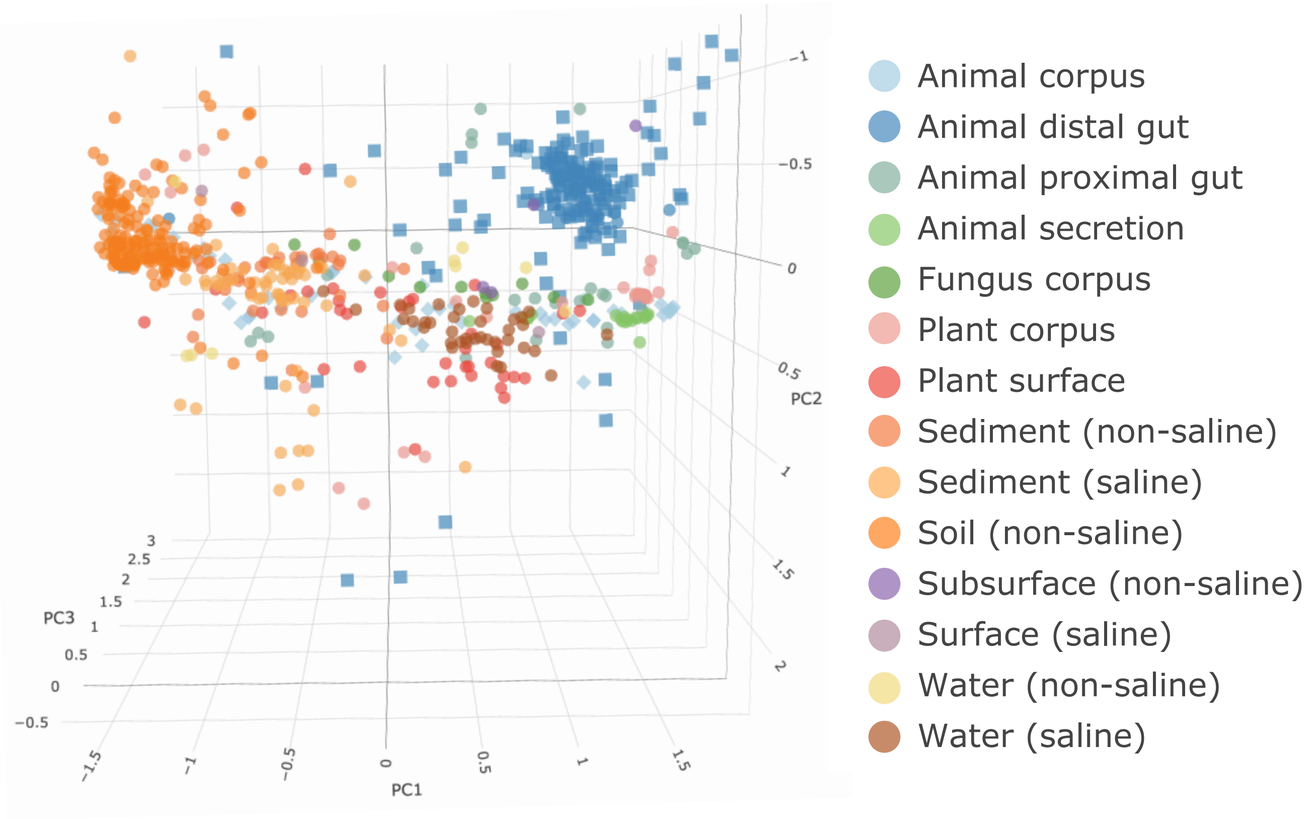


Better characterization of less-studied microbiome of earth

Hierarchical categorization
of earth microbiome samples

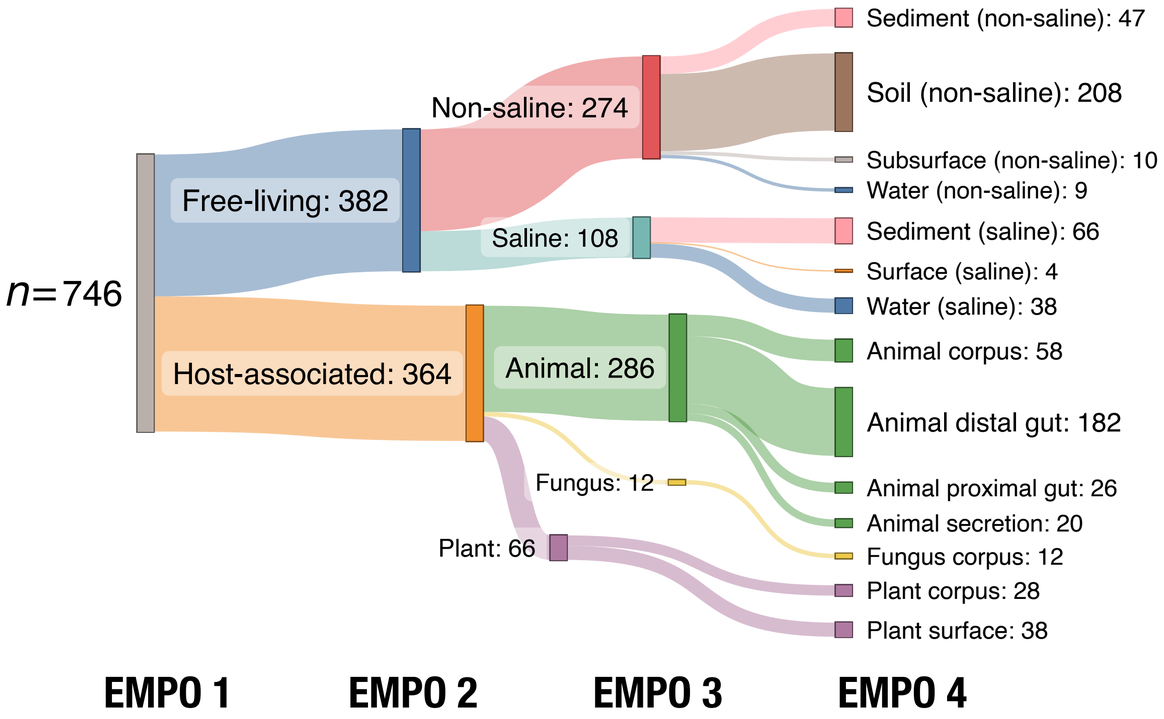


Reference: Web of Life (v1)
11,000 microbial genomes

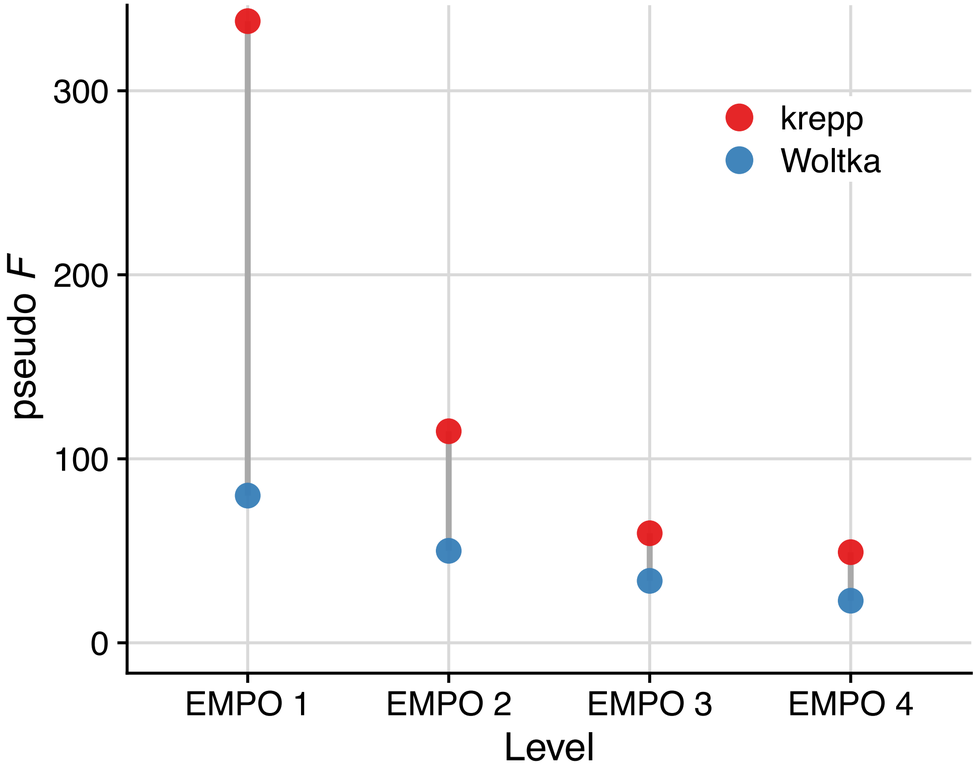


Better characterization of less-studied microbiome of earth

Hierarchical categorization
of earth microbiome samples



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



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)*

Thank you!

software: github.com/bo1929/krepp

preprint: github.com/bo1929/krepp

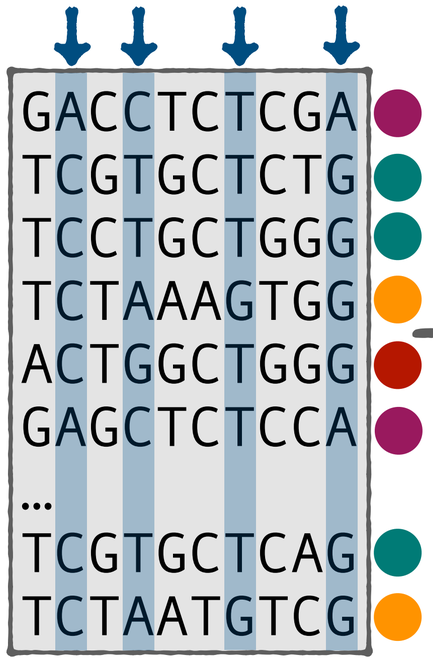
Siavash Mirarab

Funding

Extra Slides

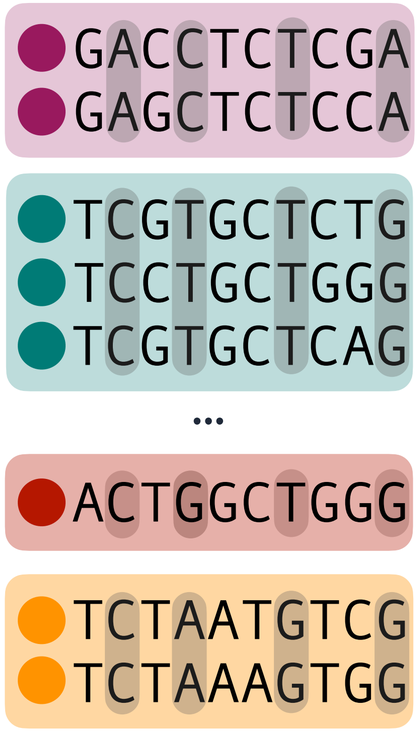
Computing Hamming distances between homologous k-mers

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

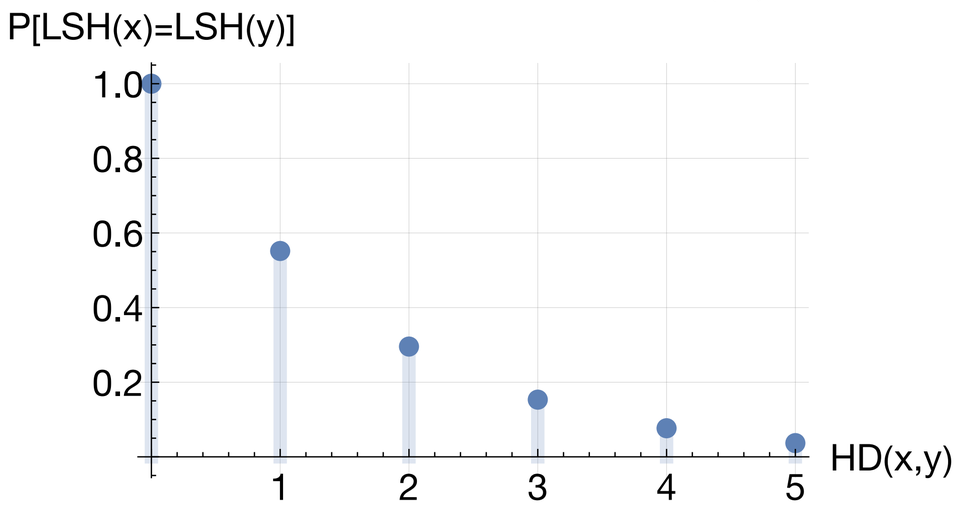
miss at HD=2

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:

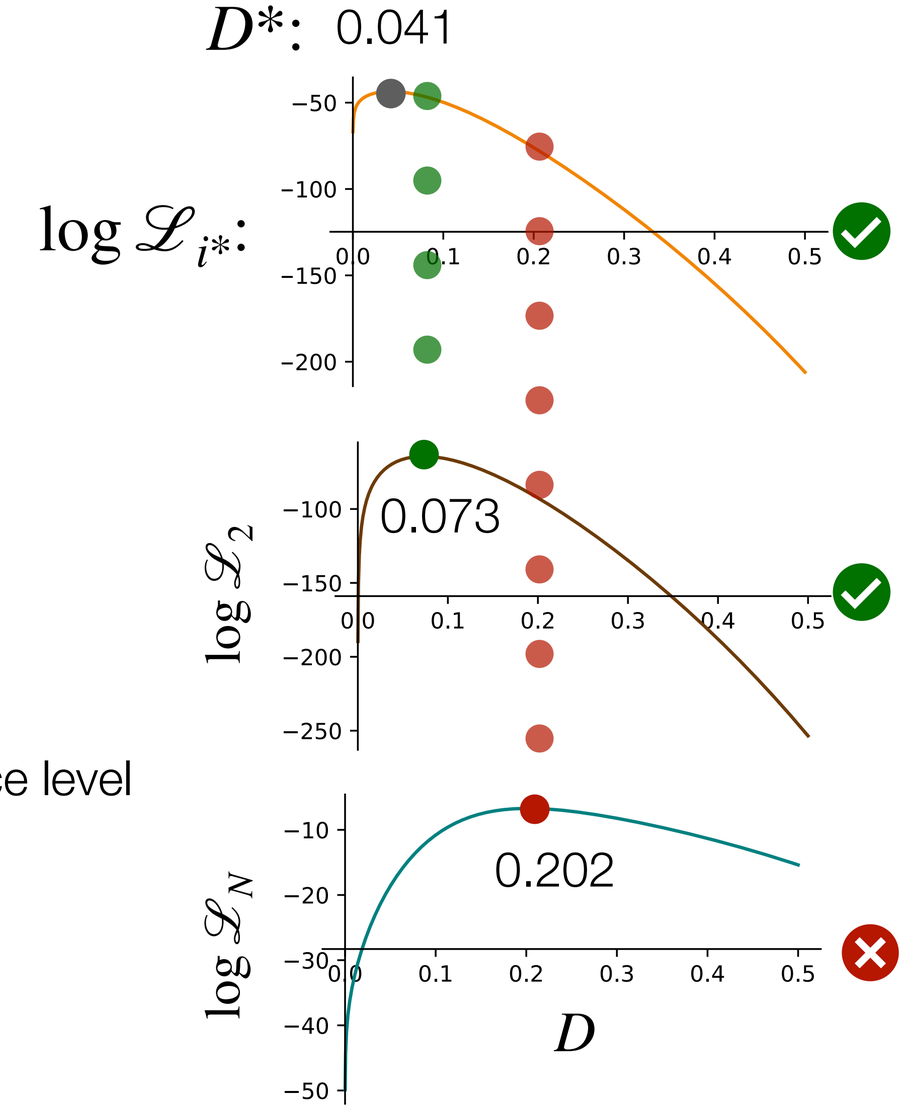
$$\lambda_{LR} = \frac{\mathscr{L}_{i*}(D; k, h, \delta, u_{i*}, \mathbf{v}_{i*})}{\mathscr{L}_{i*}(D^*; k, h, \delta, u_{i*}, \mathbf{v}_{i*})}$$

D: alternative distance

$i*$: closest reference

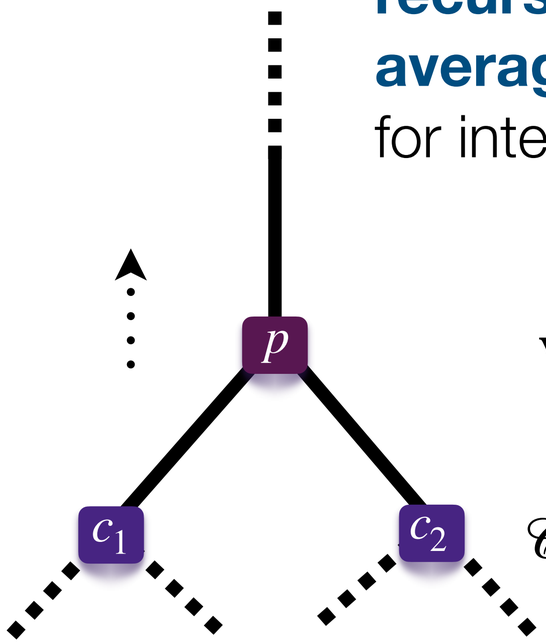
$$\lambda_{LR} \sim \chi^2$$

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



Defining clade distances & branch length agnostic placement

A notion of distance for internal nodes:



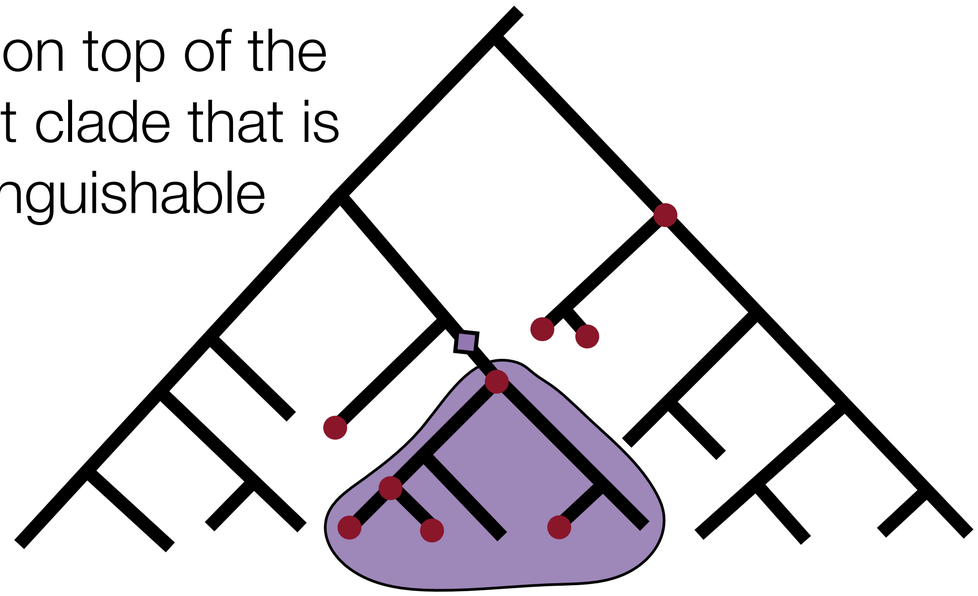
recursively compute average HD histograms for internal nodes

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

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

use the same likelihood model and log-likelihood ratio test

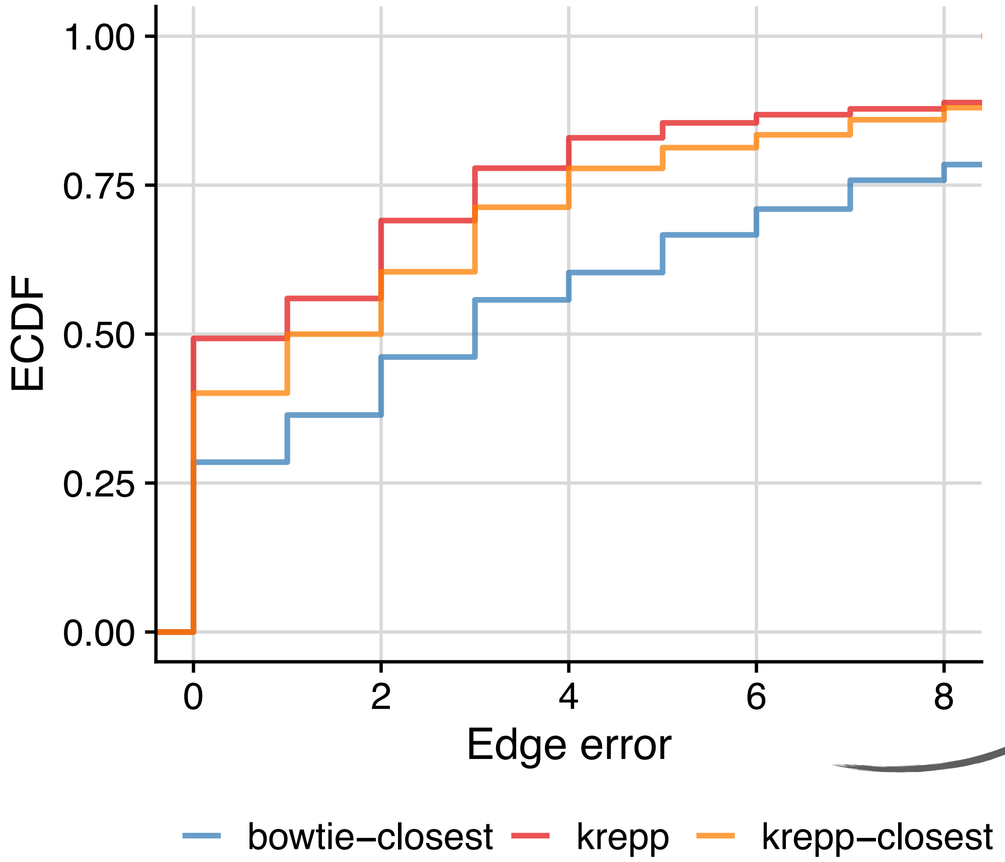
place on top of the largest clade that is indistinguishable



● : indistinguishable w.r.t. the closest reference

krepp's heuristic improves closest-tip placement

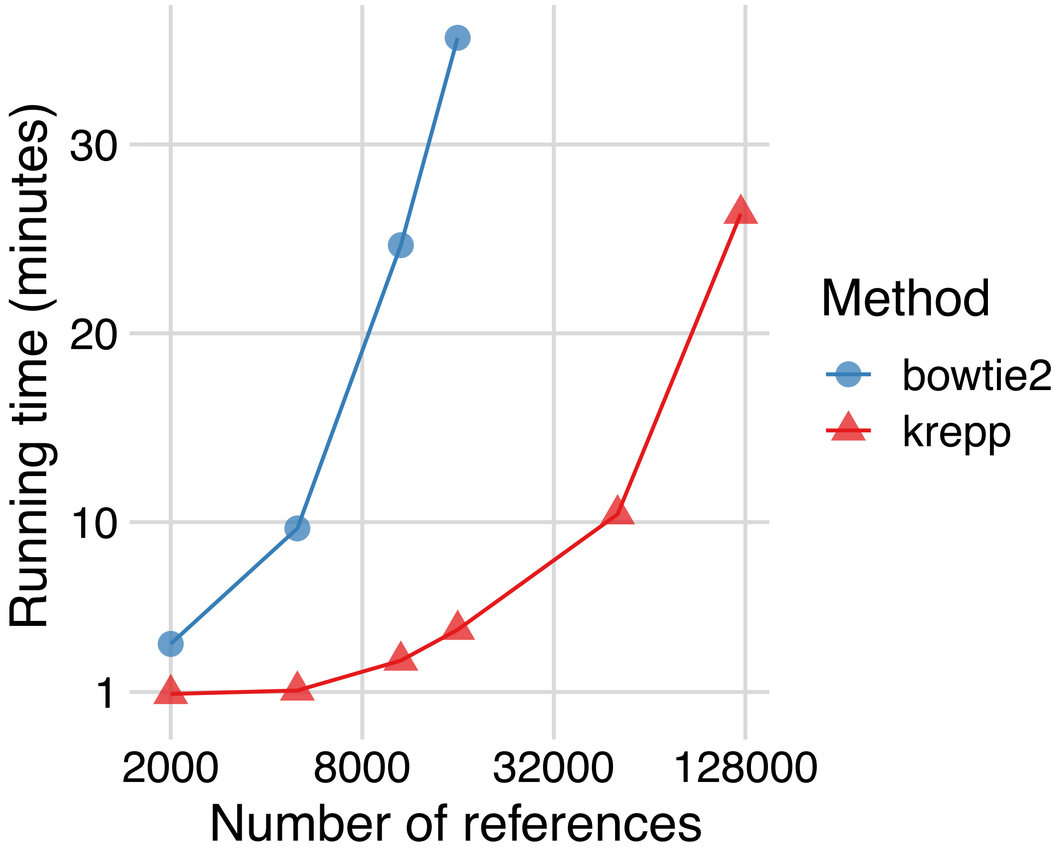
- Leave one out: 100/16,000 (WoLv2)
- Outperforms baselines:
  on the closest, on the LCA, etc.
- >80% of all reads within four edges



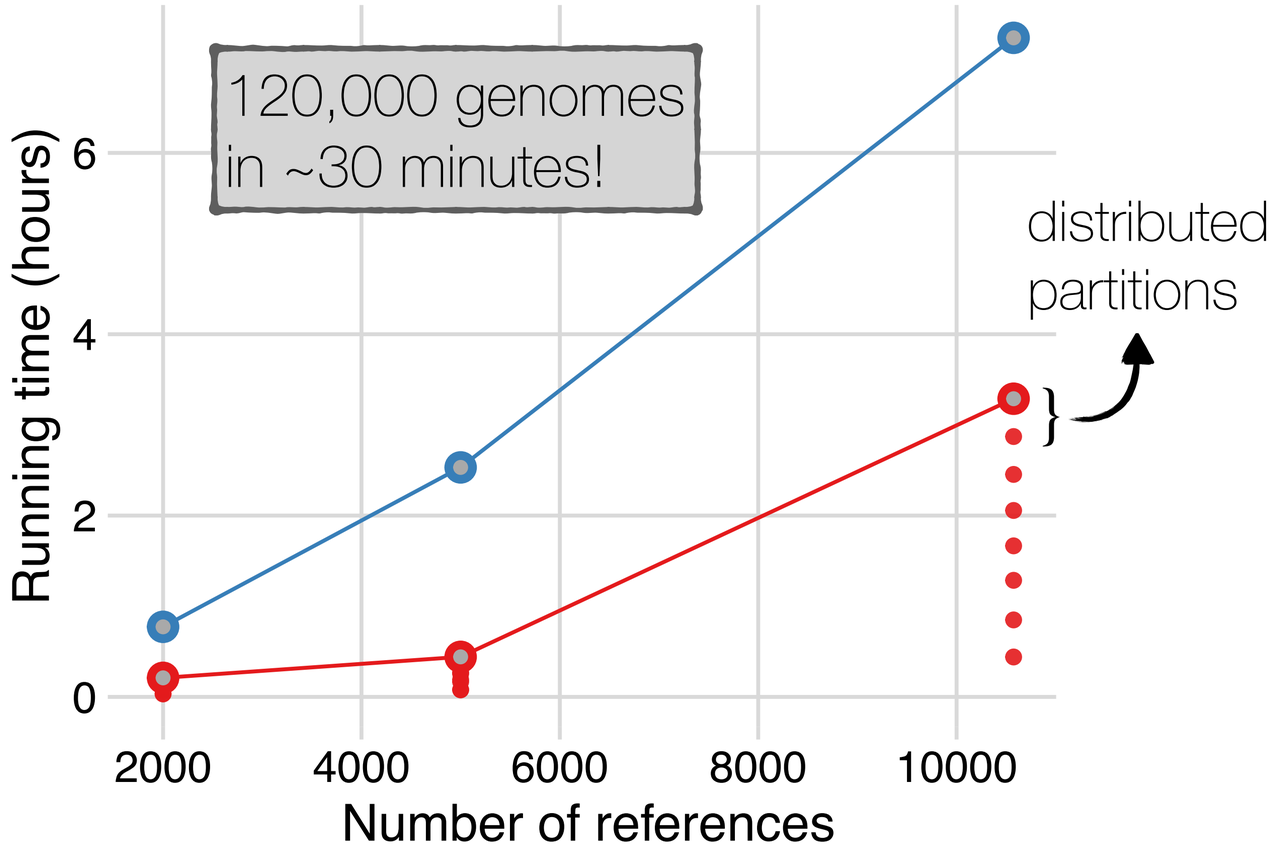
how many edges away is the placement from the correct edge?

Scalability:
Avoiding the more difficult problem & effective parallelization

Mapping 10M reads (16 threads):



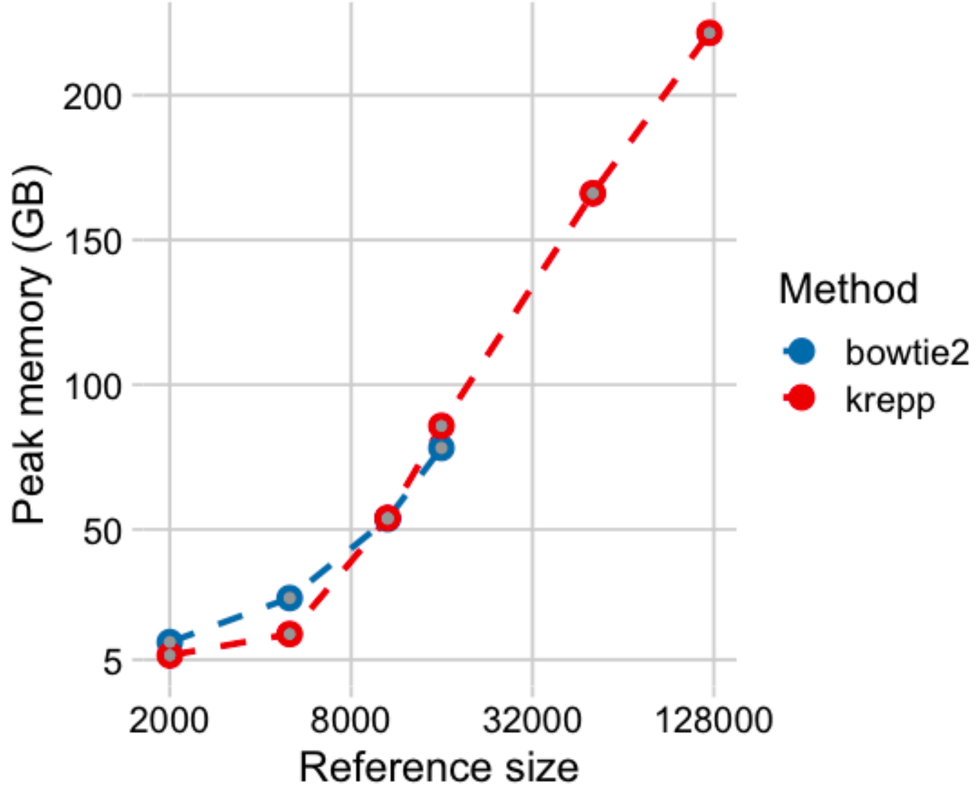
Indexing microbial genomes (32 threads):



Scalability:
krepp can be distributed and has flexible memory requirements

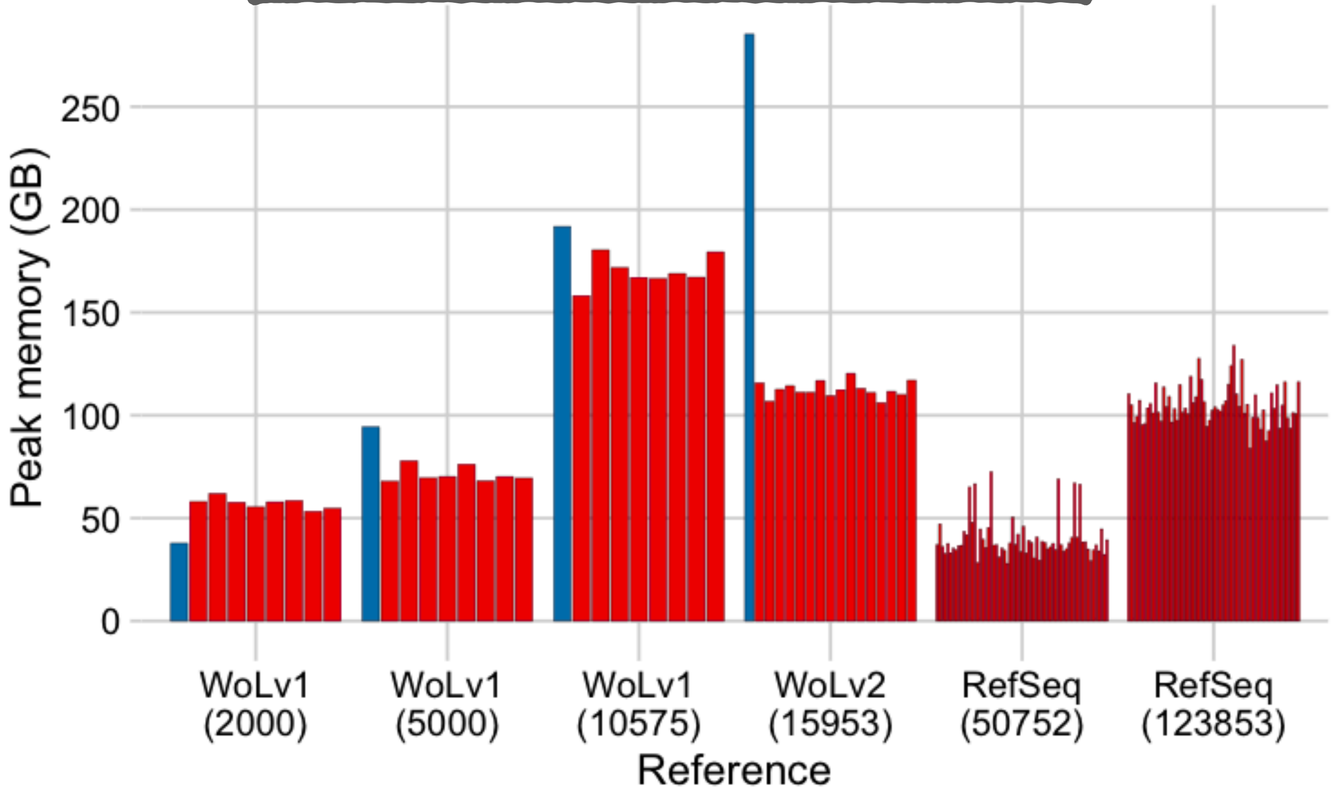
Mapping 10M reads (16 threads):

reducing memory use
w/ further subsampling...



Indexing microbial genomes (32 threads):

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



Let b_i be the length of the branch i and p_i^A and p_i^B are the taxa proportions descending from the branch i for community A and B, respectively.

$$d(A,B) = \frac{\sum_i^n b_i |p_i^A - p_i^B|}{\sum_i^n b_i (p_i^A + p_i^B)}$$

$$SS_T = \frac{1}{N}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^2 \qquad SS_W = \frac{1}{n}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^2 \delta_{ij}$$

$$SS_A = SS_T - SS_W$$

$$F = \frac{\left(\frac{SS_A}{p-1}\right)}{\left(\frac{SS_W}{N-p}\right)}$$

Multiple permutations to get a p-value!

N is the number of groups, p is the number of objects in each group.