Scalable methods for genome-wide & phylogeny-aware sequence analysis

Ali Osman Berk Şapcı

UC San Diego
Bioinformatics & Systems Biology

UC San Diego
Electrical and Computer Engineering
JACOBS SCHOOL OF ENGINEERING

BIO 310 at Sabancı University - 11 May 2026

About me

BSc & MSc (2022) @ Sabancı University both advised by Öznur Taştan :)

About me

NLP & active learning (co-advised by Reyyan Yeniterzi)

BSc & MSc (2022) @ Sabancı University both advised by Öznur Taştan :)

computational biology/ethology & ML (co-advised by Sündüz Keleş)

About me

NLP & active learning (co-advised by Reyyan Yeniterzi)

BSc & MSc (2022) @ Sabancı University both advised by Öznur Taştan :)

computational biology/ethology & ML (co-advised by Sündüz Keleş)

Now a PhD student @ UC San Diego, working with Siavash Mirarab

About me

NLP & active learning (co-advised by Reyyan Yeniterzi)

BSc & MSc (2022) @ Sabancı University both advised by Öznur Taştan :)

computational biology/ethology & ML (co-advised by Sündüz Keleş)

Now a PhD student @ UC San Diego, working with Siavash Mirarab

Mostly algorithmic:

- Traditional phylogenetics & phylogenomics
- Sequence analysis (+ comparative genomics)
- Metagenomics (+ applied phylogenetics)

About me

NLP & active learning (co-advised by Reyyan Yeniterzi)

BSc & MSc (2022) @ Sabancı University both advised by Öznur Taştan :)

computational biology/ethology & ML (co-advised by Sündüz Keleş)

Now a PhD student @ UC San Diego, working with Siavash Mirarab

Mostly algorithmic:

- Traditional phylogenetics & phylogenomics
- Sequence analysis (+ comparative genomics)
- Metagenomics (+ applied phylogenetics)

Identifying the taxa present in biological samples is central to many applications.

Identifying the taxa present in biological samples is central to many applications.

The Microbiome and Its Myth-Making Machine

If you have heard something very specific about the microbiome, odds are it's wrong

Article https://doi.org/10.1038/s41467-022-34416-0

Ancient oral microbiomes support gradual Neolithic dietary shifts towards agriculture

Received: 3 March 2022
Accepted: 25 October 2022
Published on
Check fo

Andrea Quagliariello[1], Alessandra Modi[2], Gabriel Innocenti[1], Valentina Zaro[2], Cecilia Conati Barbaro[3], Annamaria Ronchitelli[4], Francesco Boschin[4], Claudio Cavazzuti[5], Elena Dellù[6], Francesca Radina[6],

Paleoanthropological insights

Identifying the taxa present in biological samples is central to many applications.

The Microbiome and Its Myth-Making Machine

If you have heard something very specific about the microbiome, odds are it's wrong

Gut microbiome research

Article https://doi.org/10.1038/s41467-022-34416-0

Ancient oral microbiomes support gradual Neolithic dietary shifts towards agriculture

Received: 3 March 2022
Accepted: 25 October 2022
Published on
Check fo

Andrea Quagliariello[1], Alessandra Modi[2], Gabriel Innocenti[1], Valentina Zaro[2], Cecilia Conati Barbaro[3], Annamaria Ronchitelli[4], Francesco Boschin[4], Claudio Cavazzuti[5], Elena Dellù[6], Francesca Radina[6],

Paleoanthropological insights

Identifying the taxa present in biological samples is central to many applications.

Article | Open access | Published: 08 July 2023

Atlantic water influx and sea-ice cover drive taxonomic and functional shifts in Arctic marine bacterial communities

Taylor Priest, Wilken-Jon von Appen, Ellen Oldenburg, Ovidiu Popa, Sinhué Torres-Valdés, Christina Bienhold, Katja Metfies, William Boulton, Thomas Mock, Bernhard M. Fuchs, Rudolf Amann, Antje Boetius & Matthias Wietz

Monitoring effects of climate change

The Microbiome and Its Myth-Making Machine

If you have heard something very specific about the microbiome, odds are it's wrong

Gut microbiome research

Article

https://doi.org/10.1038/s41467-022-34416-0

Ancient oral microbiomes support gradual Neolithic dietary shifts towards agriculture

Received: 3 March 2022

Accepted: 25 October 2022

Published on

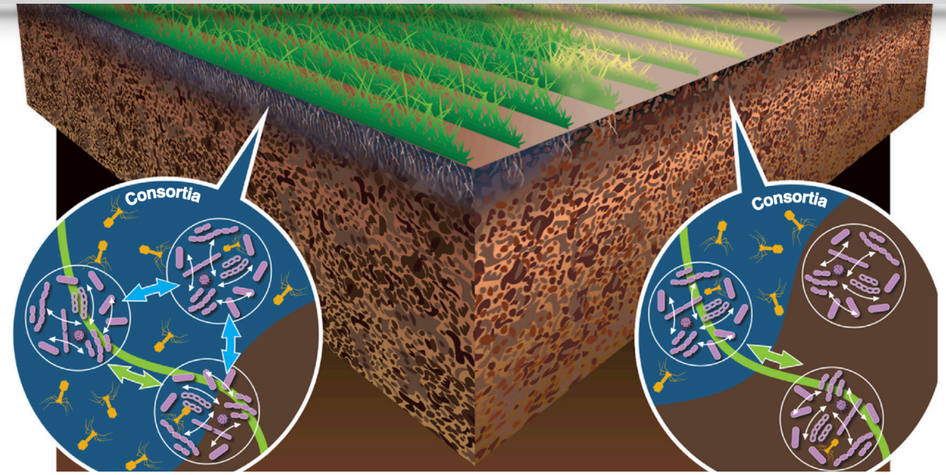
Check fo

Andrea Quagliariello[1], Alessandra Modi[2], Gabriel Innocenti[1], Valentina Zaro[2], Cecilia Conati Barbaro[3], Annamaria Ronchitelli[4], Francesco Boschin[4], Claudio Cavazzuti[5], Elena Dellù[6], Francesca Radina[6],

Paleoanthropological insights

Identifying the taxa present in biological samples is central to many applications.

Soil microbiome: sustainable agriculture



The soil microbiome—from metagenomics to metaphenomics

Janet K Jansson [1], Kirsten S Hofmockel [1 2]

Article | Open access | Published: 08 July 2023

Atlantic water influx and sea-ice cover drive taxonomic and functional shifts in Arctic marine bacterial communities

Taylor Priest, Wilken-Jon von Appen, Ellen Oldenburg, Ovidiu Popa, Sinhué Torres-Valdés, Christina Bienhold, Katja Metfies, William Boulton, Thomas Mock, Bernhard M. Fuchs, Rudolf Amann, Antje Boetius & Matthias Wietz

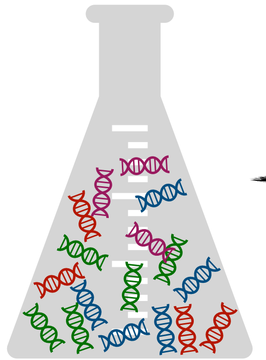
Monitoring effects of climate change

The Microbiome and Its Myth-Making Machine

If you have heard something very specific about the microbiome, odds are it's wrong

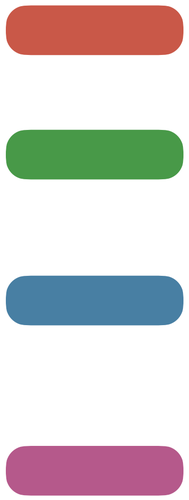
Gut microbiome research

Analyzing metagenomic samples



sample

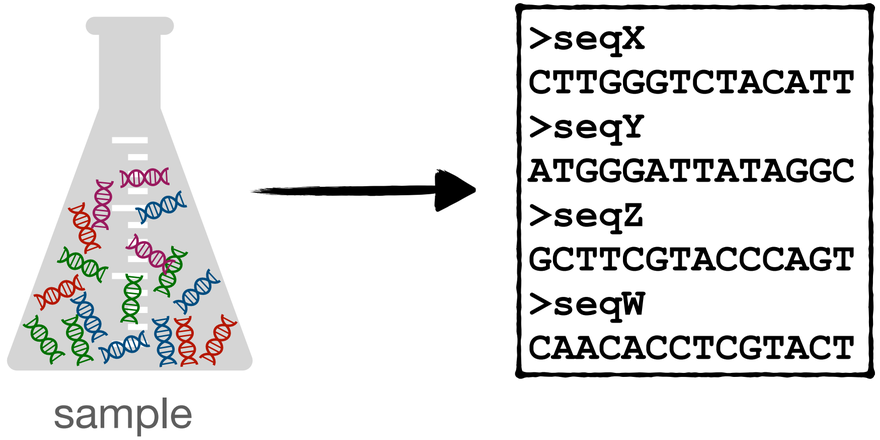
```
>seqX
CTTGGGTCTACATT
>seqY
ATGGGATTATAGGC
>seqZ
GCTTCGTACCCAGT
>seqW
CAACACCTCGTACT
```

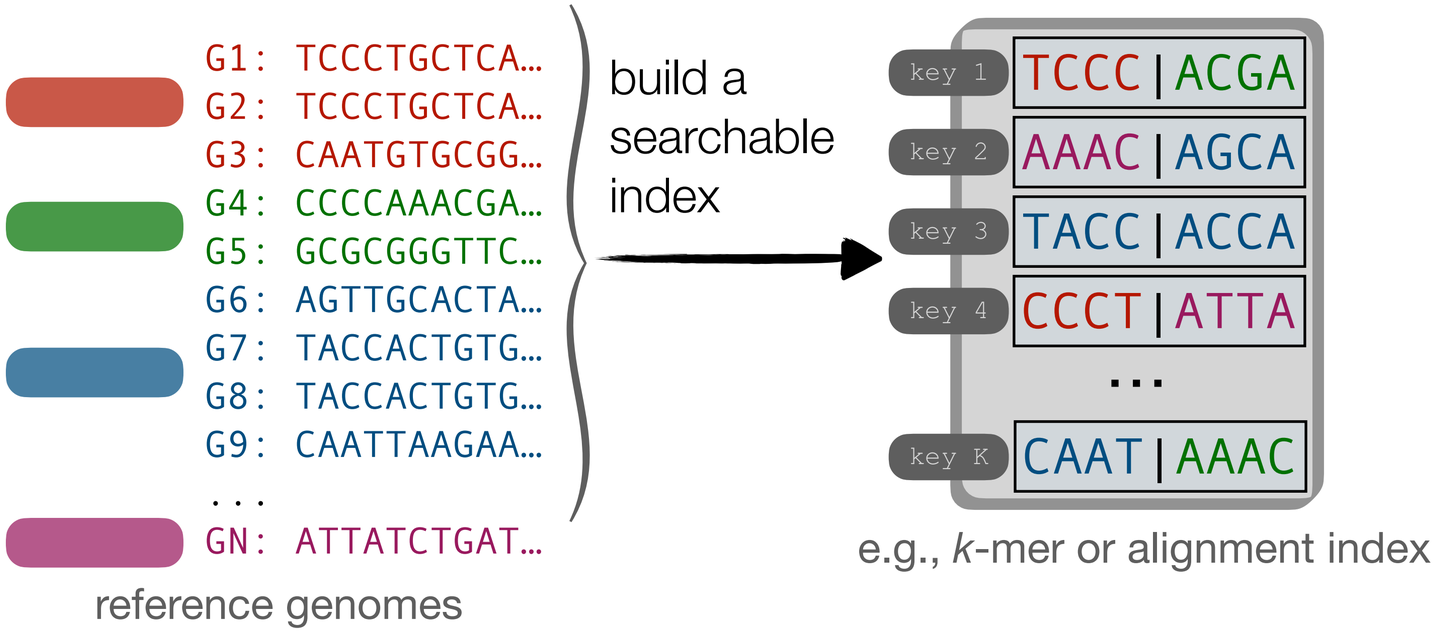


```
G1: TCCCTGCTCA…
G2: TCCCTGCTCA…
G3: CAATGTGCGG…
G4: CCCCAAACGA…
G5: GCGCGGGTTC…
G6: AGTTGCACTA…
G7: TACCACTGTG…
G8: TACCACTGTG…
G9: CAATTAAGAA…
...
GN: ATTATCTGAT…
```

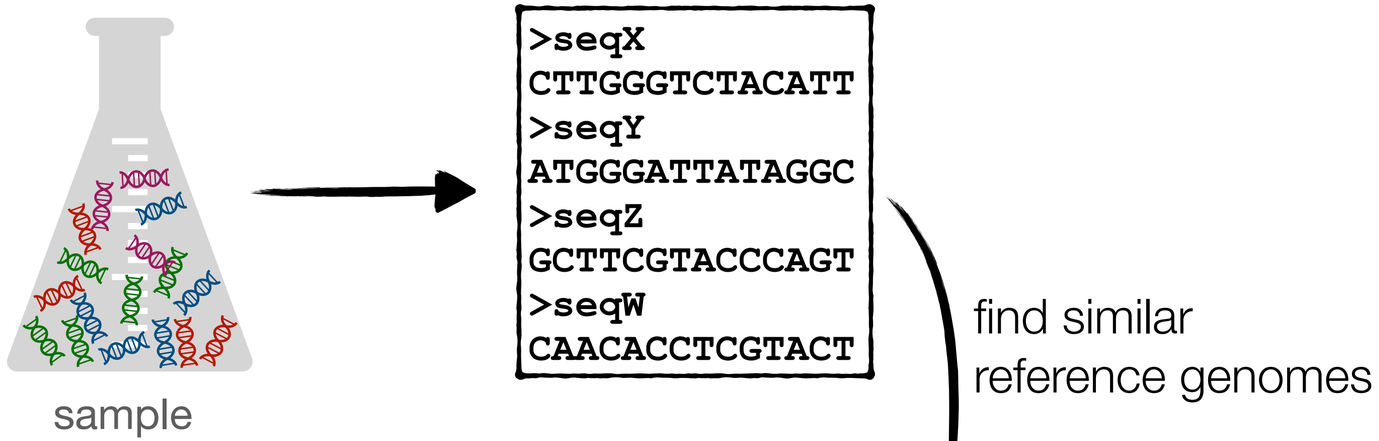
reference genomes

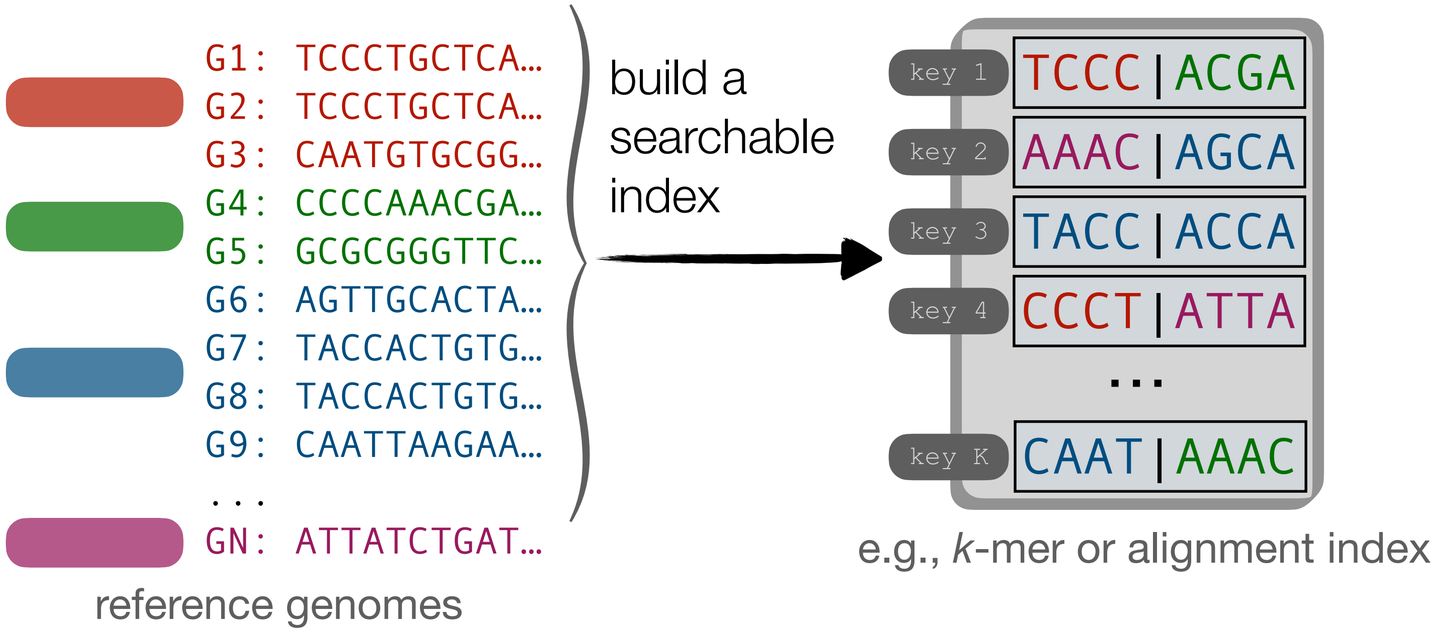
Analyzing metagenomic samples



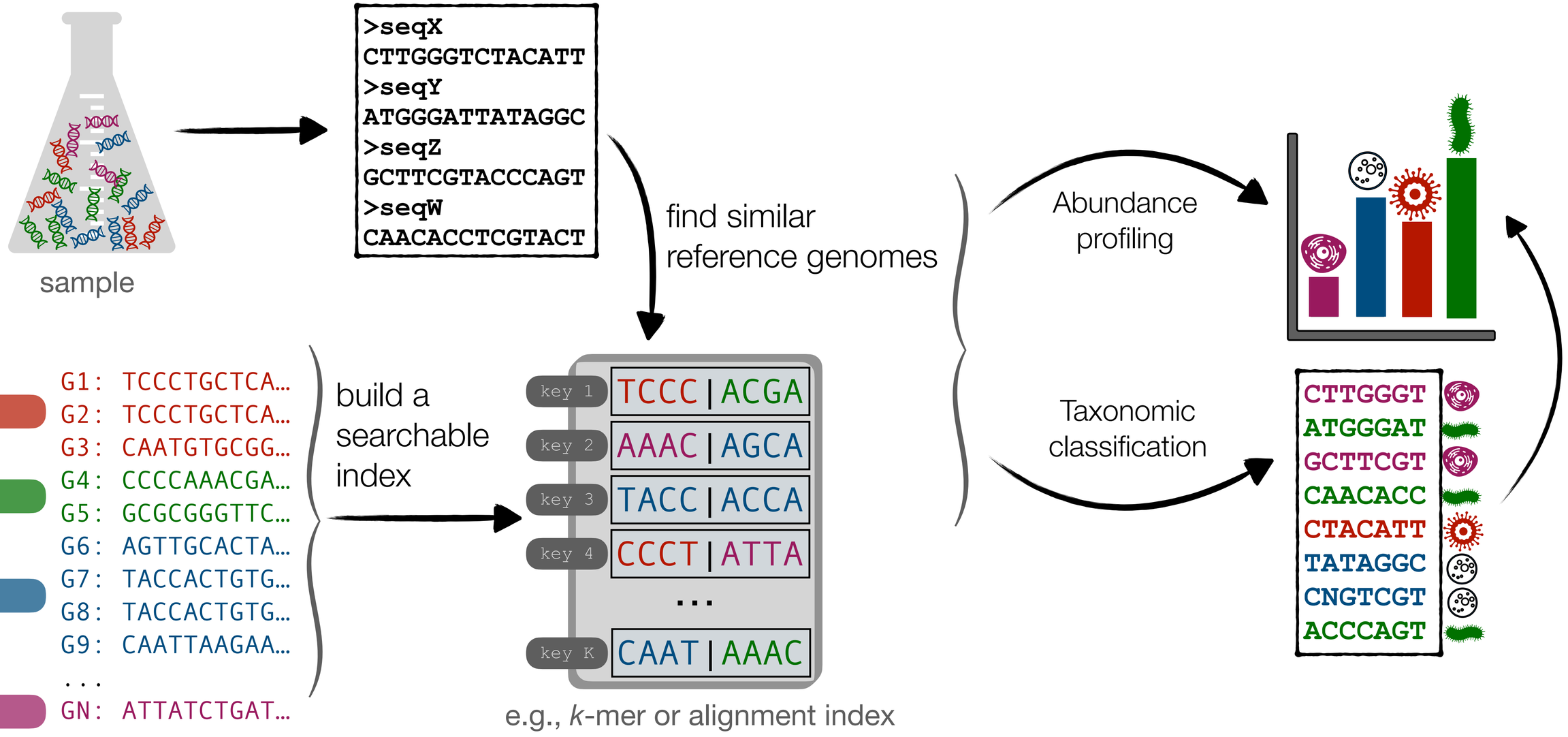


Analyzing metagenomic samples





Analyzing metagenomic samples

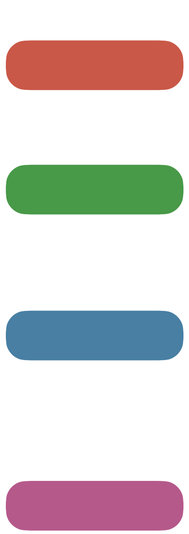


Analyzing metagenomic samples

sample

```
>seqX
CTTGGGTCTACATT
>seqY
ATGGGATTATAGGC
>seqZ
GCTTCGTACCCAGT
>seqW
CAACACCTCGTACT
```

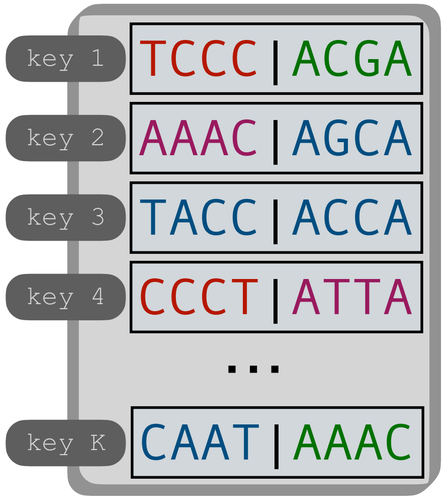
find similar reference genomes



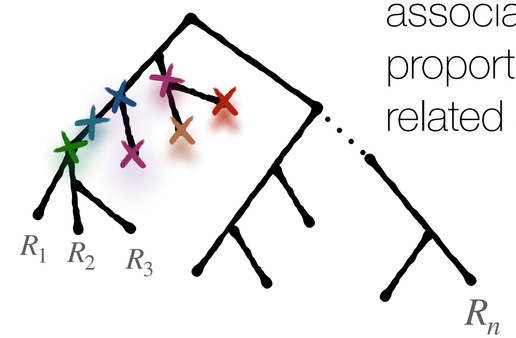
```
G1:  TCCCTGCTCA…
G2:  TCCCTGCTCA…
G3:  CAATGTGCGG…
G4:  CCCCAAACGA…
G5:  GCGCGGGTTC…
G6:  AGTTGCACTA…
G7:  TACCACTGTG…
G8:  TACCACTGTG…
G9:  CAATTAAGAA…
...
GN:  ATTATCTGAT…
```

reference genomes

build a searchable index



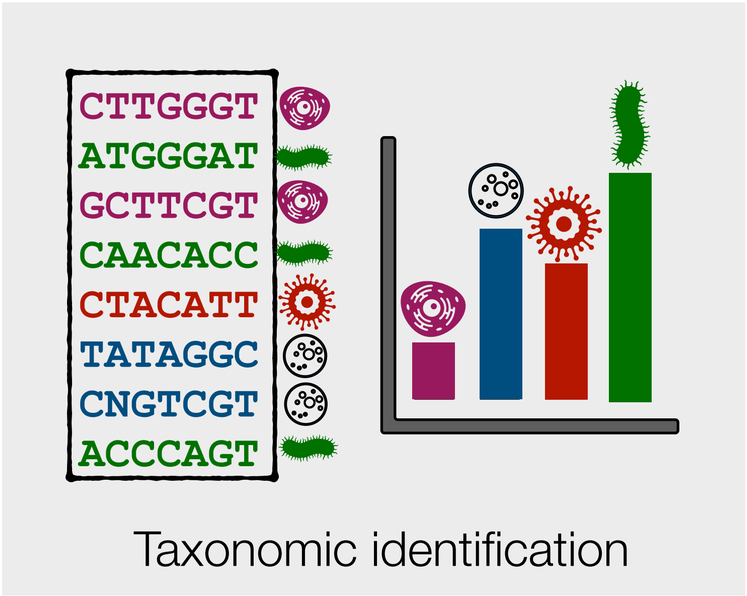
e.g., *k*-mer or alignment index



associate some proportion with related edges

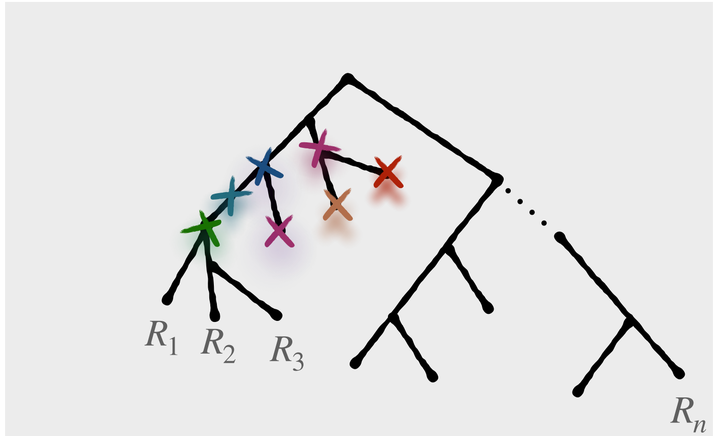
+ Phylogenetic Placements

traditional hierarchy (morphology, genetic)



Taxonomic identification

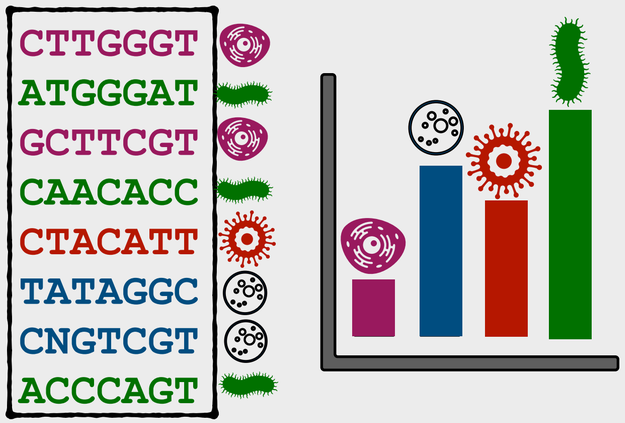
evolutinary relations & branch lengths



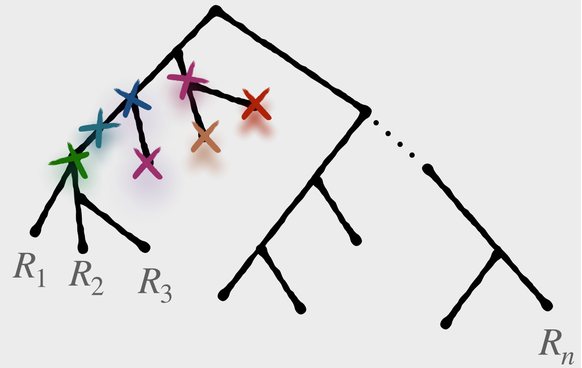
Phylogenetic Placement (PP)

traditional hierarchy (morphology, genetic) **evolutinary relations & branch lengths**

higher resolution
better evolutionary signal



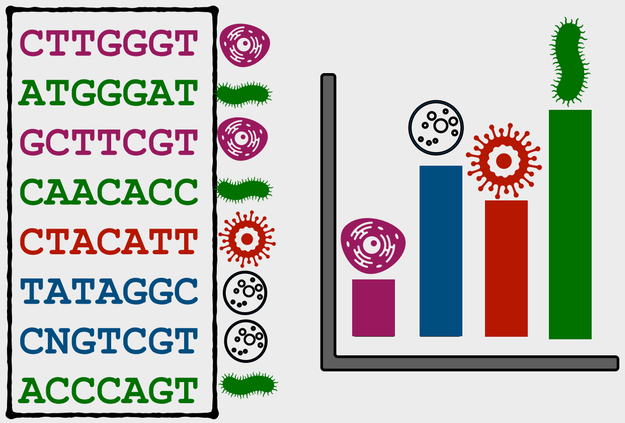
Taxonomic identification



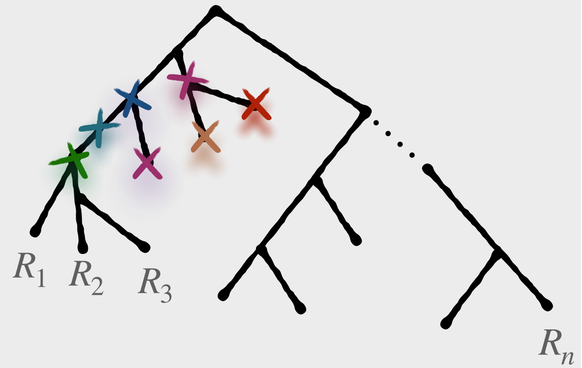
+ Phylogenetic Placement (PP)

traditional hierarchy (morphology, genetic) **evolutinary relations & branch lengths**

higher resolution
better evolutionary signal



Taxonomic identification



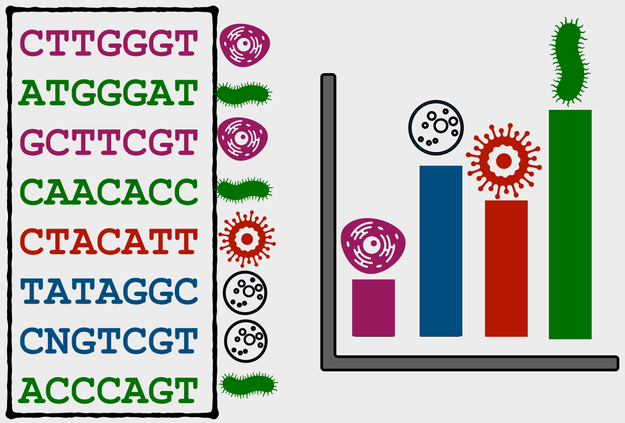
+ Phylogenetic Placement (PP)

simpler: better scalability
utilizing larger ref. sets

traditional hierarchy (morphology, genetic) **evolutinary relations & branch lengths**

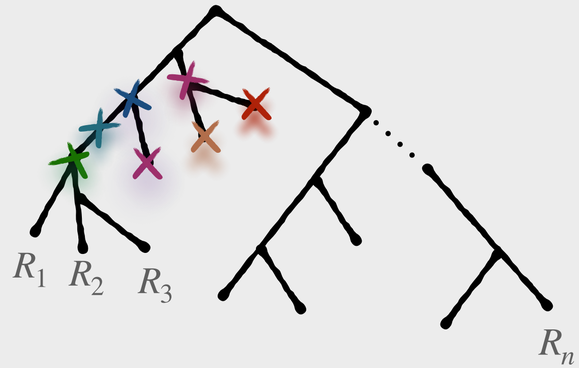
higher resolution
better evolutionary signal

methods are limited



Taxonomic identification

k-mer based methods



\+ Phylogenetic Placement (PP)

simpler: better scalability
utilizing larger ref. sets

We will use *k*-mers

for sequence distance estimation &
phylogenetic placement

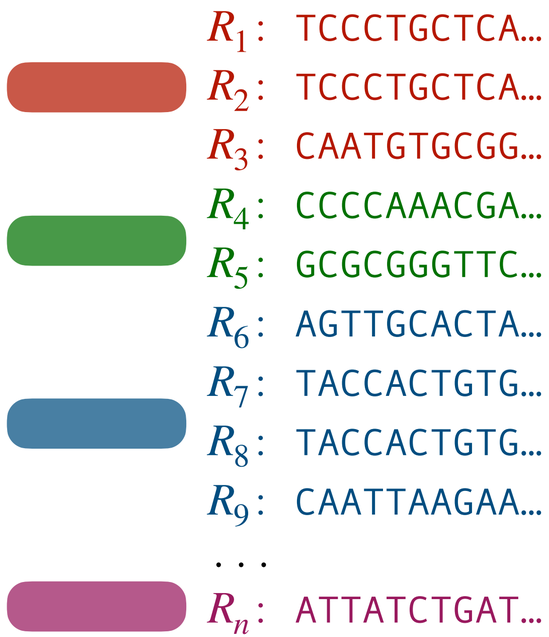
Two methods: CONSULT and krepp

for taxonomic classification,
not too techincal: warm-up

Core problem: searching similar sequences

> Query sequence Q
ATGGGATTATAGGCATAGGCATTAGTGGC

Alignment is not scalable when $N \approx 100{,}000$



reference genomes $\mathscr{R}$

Core problem: searching similar sequences

> Query sequence Q
ATGGGATTATAGGCATAGGCATTAGTGGC

$H($**AGGCATAGGC**$) \rightarrow$ `0xc95f5e06b6f9f52f`

Idea: extract all subsequences of a fixed length k (*k*-mers) from $\mathscr{R}$ and Q

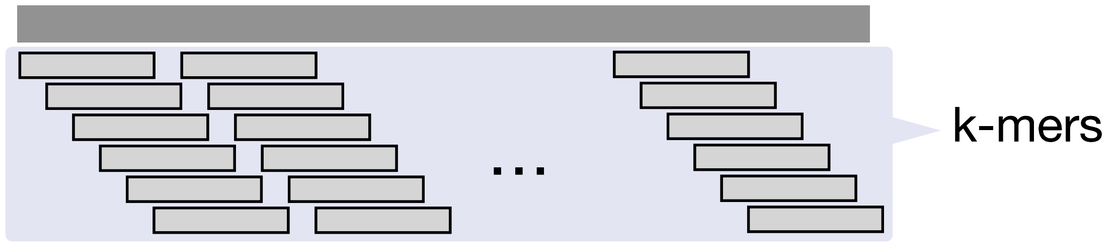
Compute **hash** values for each *k*-mer

Look for presence/absence of *k*-mers using hash tables in *constant time*

Alignment is not scalable when $N \approx 100{,}000$

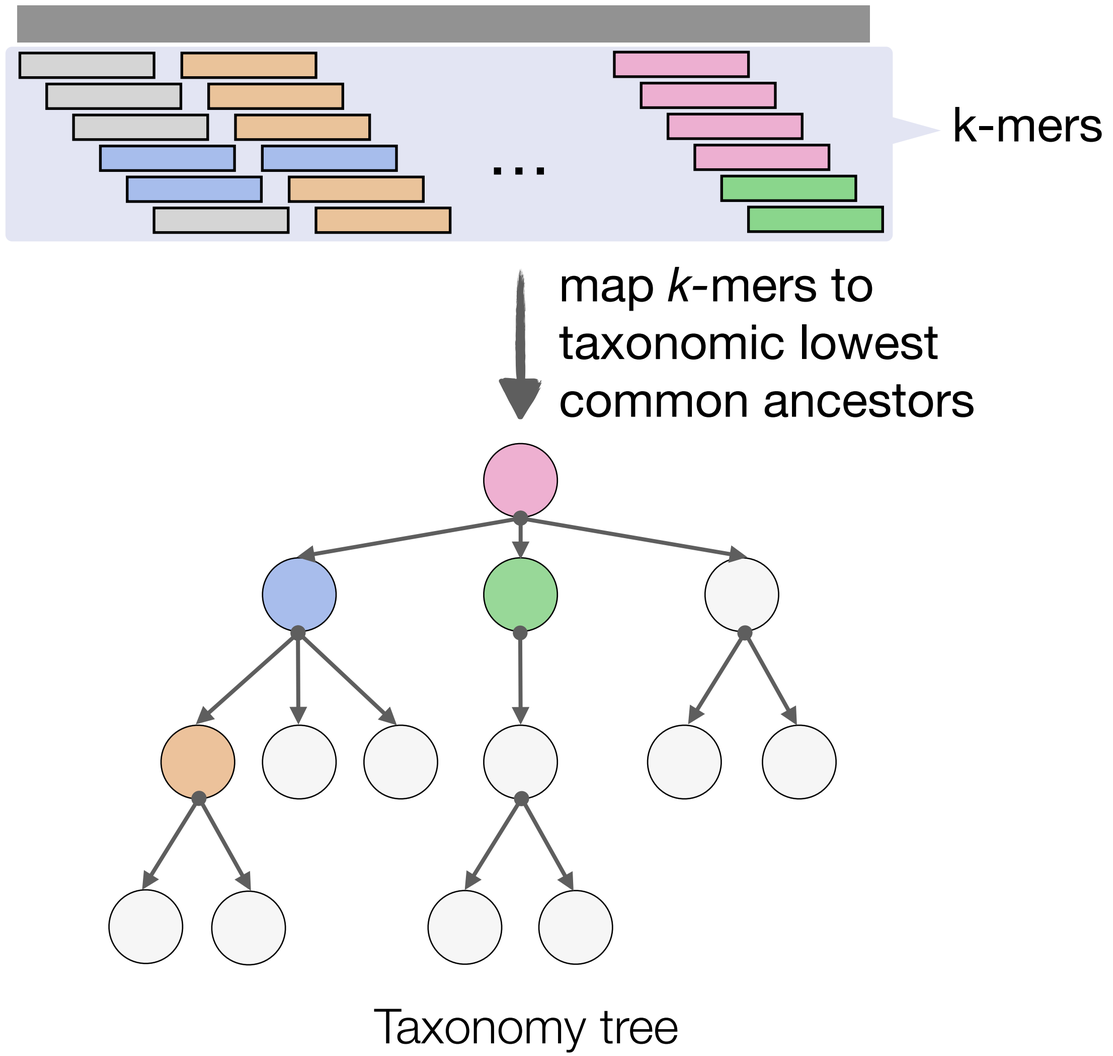
reference genomes $\mathscr{R}$

Taxonomic classification using k-mer presence/absence



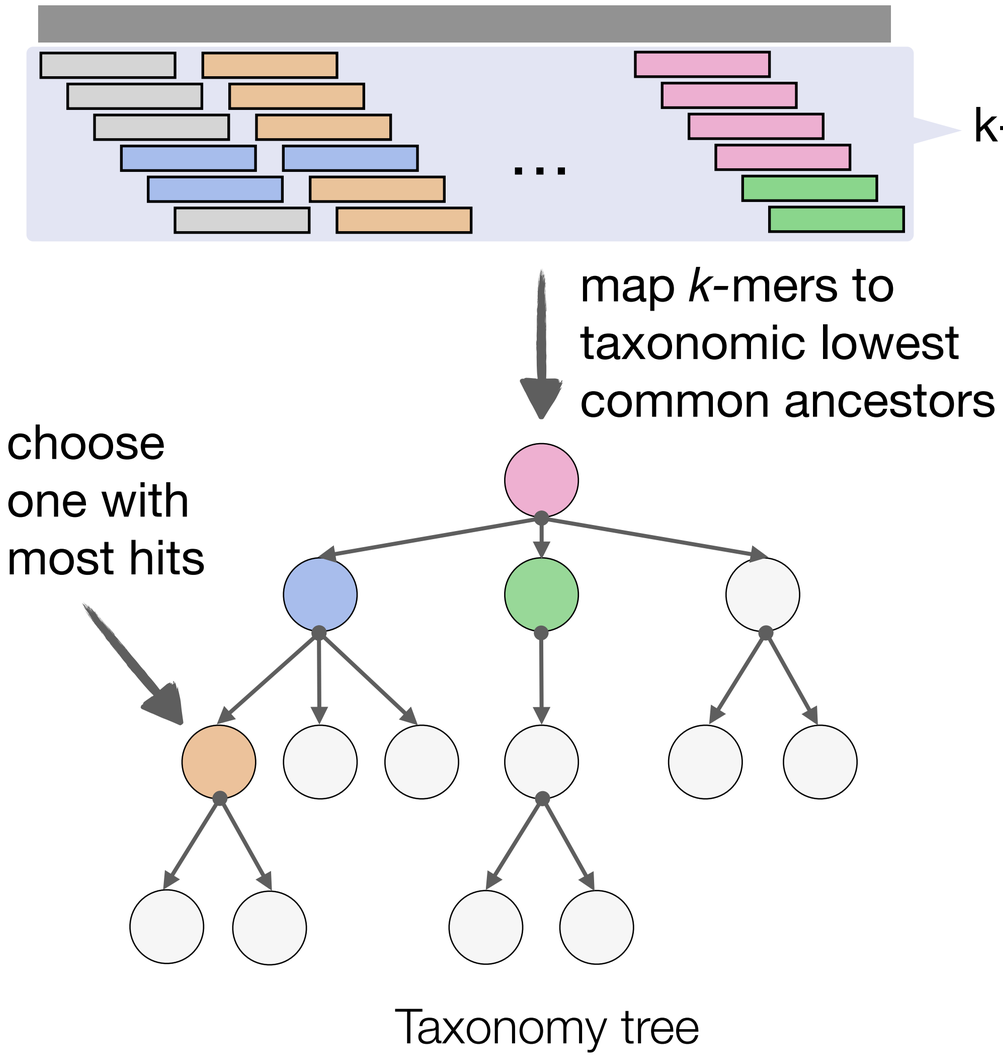
- Use hashing to test **presence/absence** of query *k*-mers in w.r.t. an index

Taxonomic classification using k-mer presence/absence



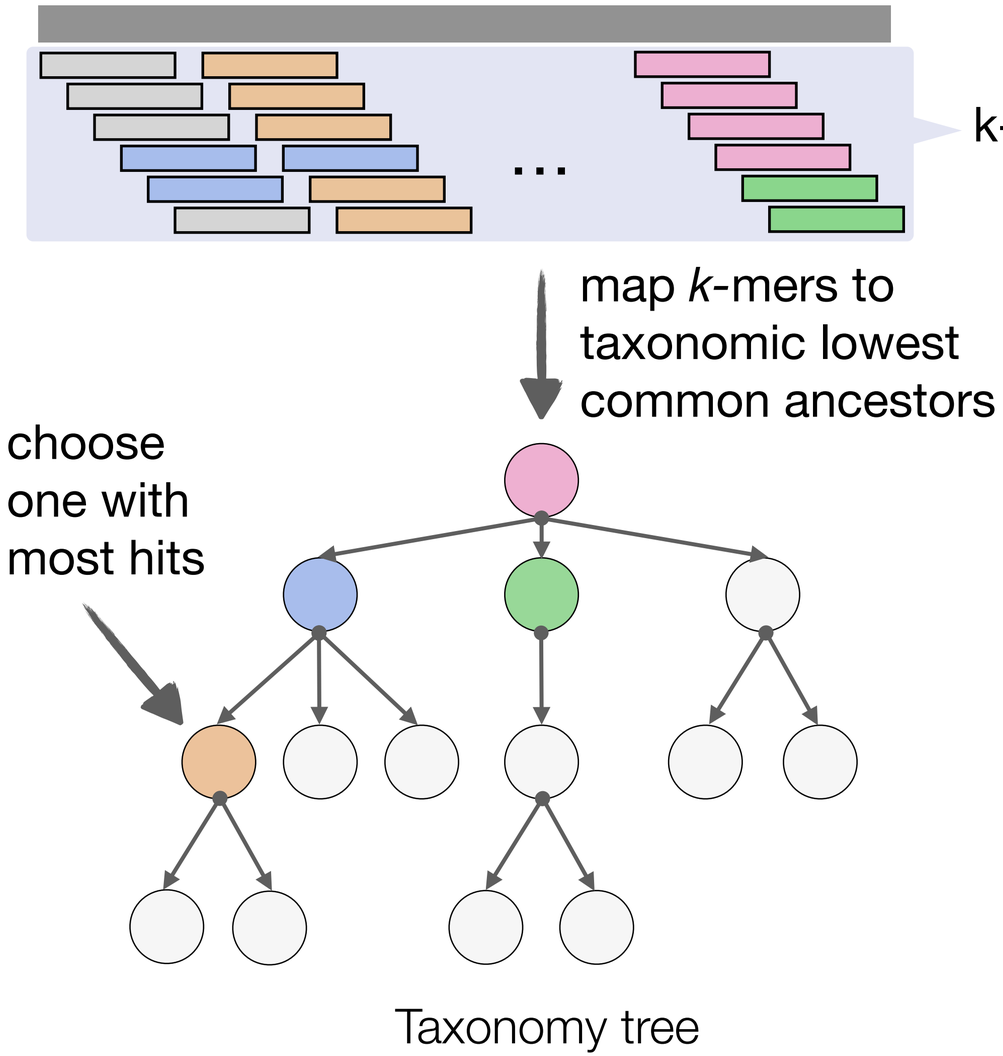
- Use hashing to test **presence/absence** of query *k*-mers in w.r.t. an index
- Map reference *k*-mers to taxonomic groups (e.g., **lowest common ancestor**)

Taxonomic classification using k-mer presence/absence



- Use hashing to test **presence/absence** of query *k*-mers in w.r.t. an index
- Map reference *k*-mers to taxonomic groups (e.g., **lowest common ancestor**)
- Use **heuristics** to summarize *k*-mer matches into taxonomic labels

Taxonomic classification using k-mer presence/absence



- Use hashing to test **presence/absence** of query *k*-mers in w.r.t. an index
- Map reference *k*-mers to taxonomic groups (e.g., **lowest common ancestor**)
- Use **heuristics** to summarize *k*-mer matches into taxonomic labels
- Kraken, CLARK, Kraken2

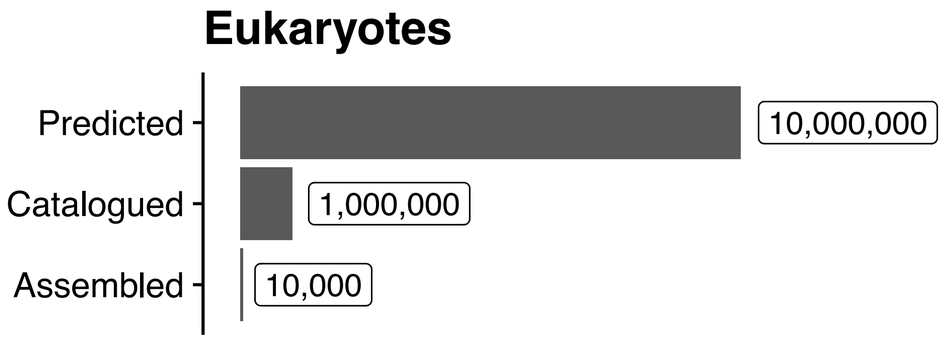
  [Wood and Salzberg 2014]

  [Ounit et al. 2015]

  [Wood et al. 2019]

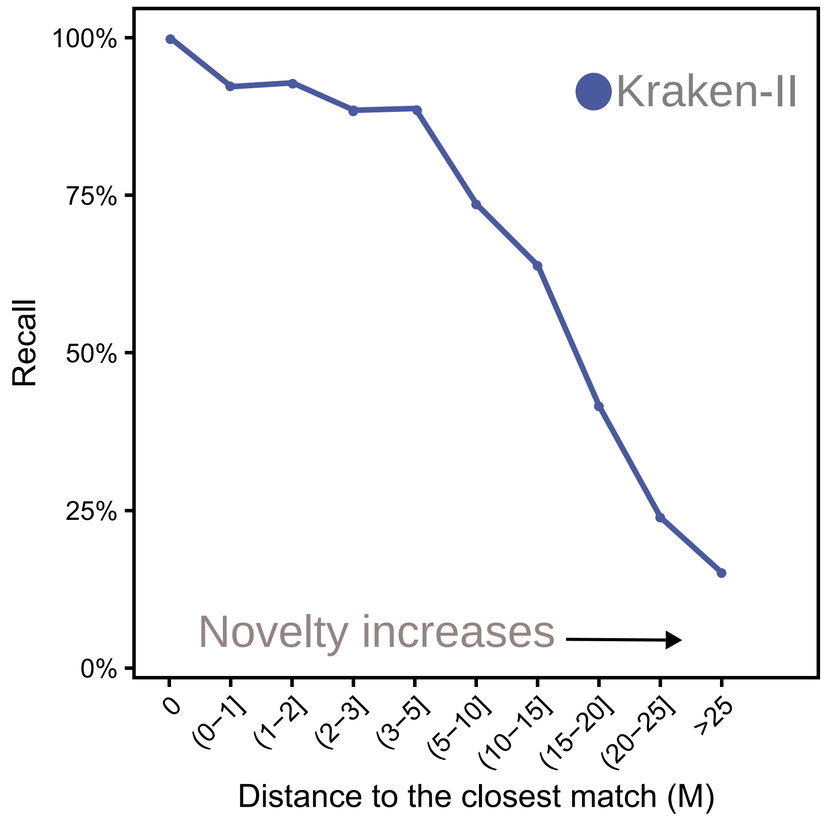
Novel queries & incomplete references challenge popular tools

- Reference databases **remain incomplete** compared to the biodiversity of earth…



- **Novel sequences:** sequences which lack a close representation in the database

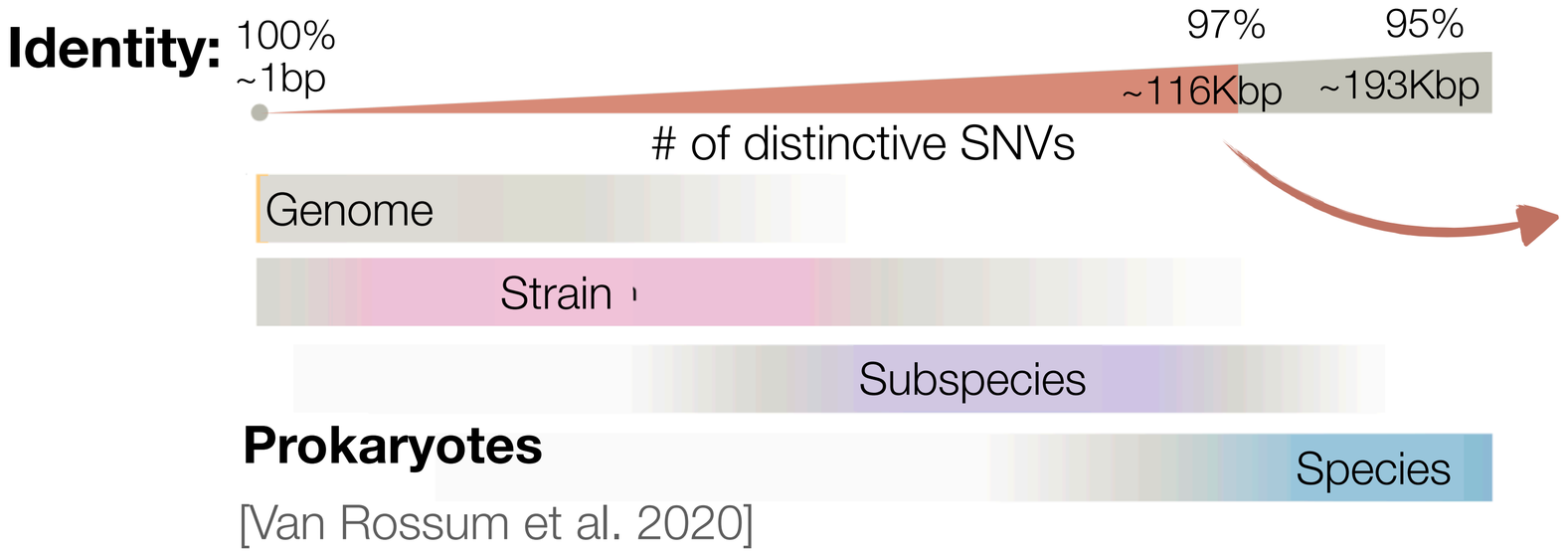
[Rachtman et al. 2019]



~% distance in substituions

Better utilization of the evolutionary signal for distant queries

- Diversity within taxon: **need for quantifying similarities**



Only 12% of **Earth's microbiome** can be identified at this level [Zhang et al. 2020]

Remaining 88%: novel sequences? Ocean, soil, novel environments…

CONSULT-II:

Can we the evolutionary aspect and keep using k-mers?

Computing Hamming distances between homologous k-mers using LSH

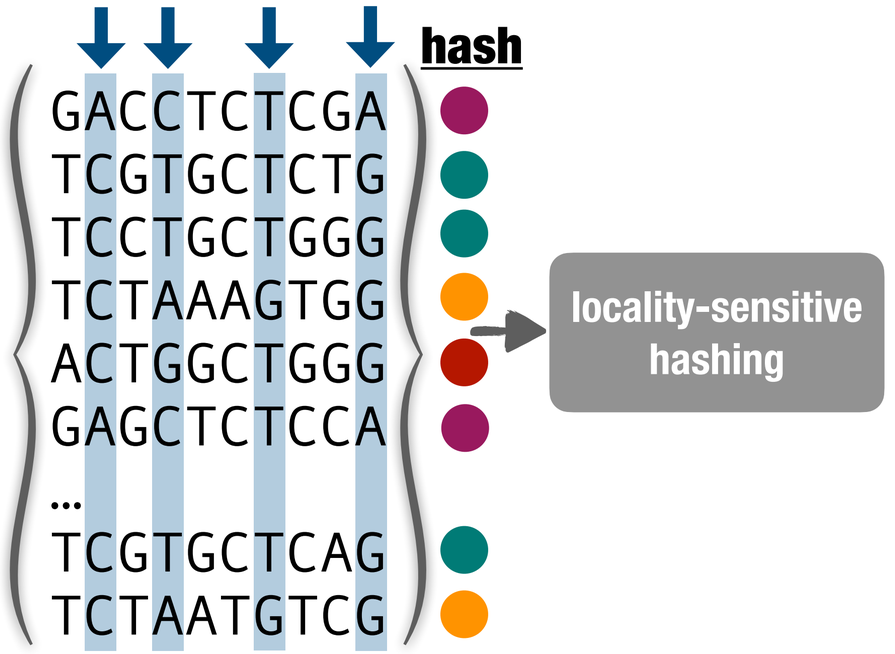
Goal: allow mismatches in *k*-mers, measure the **Hamming distance**

reference *k*-mer set

Computing Hamming distances between homologous k-mers using LSH

Goal: allow mismatches in *k*-mers, measure the **Hamming distance**

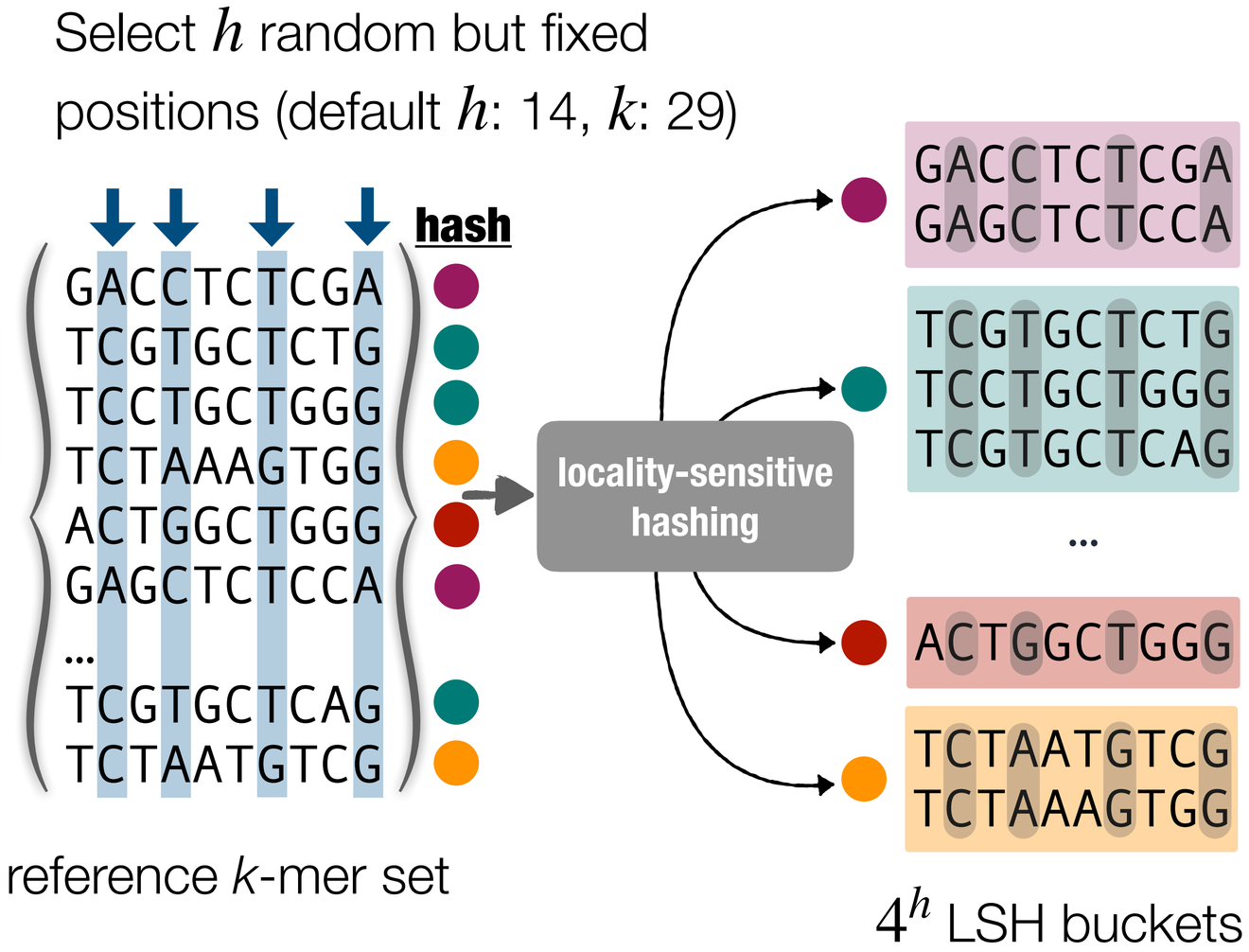
Select h random but fixed positions (default h: 14, k: 29)



reference *k*-mer set

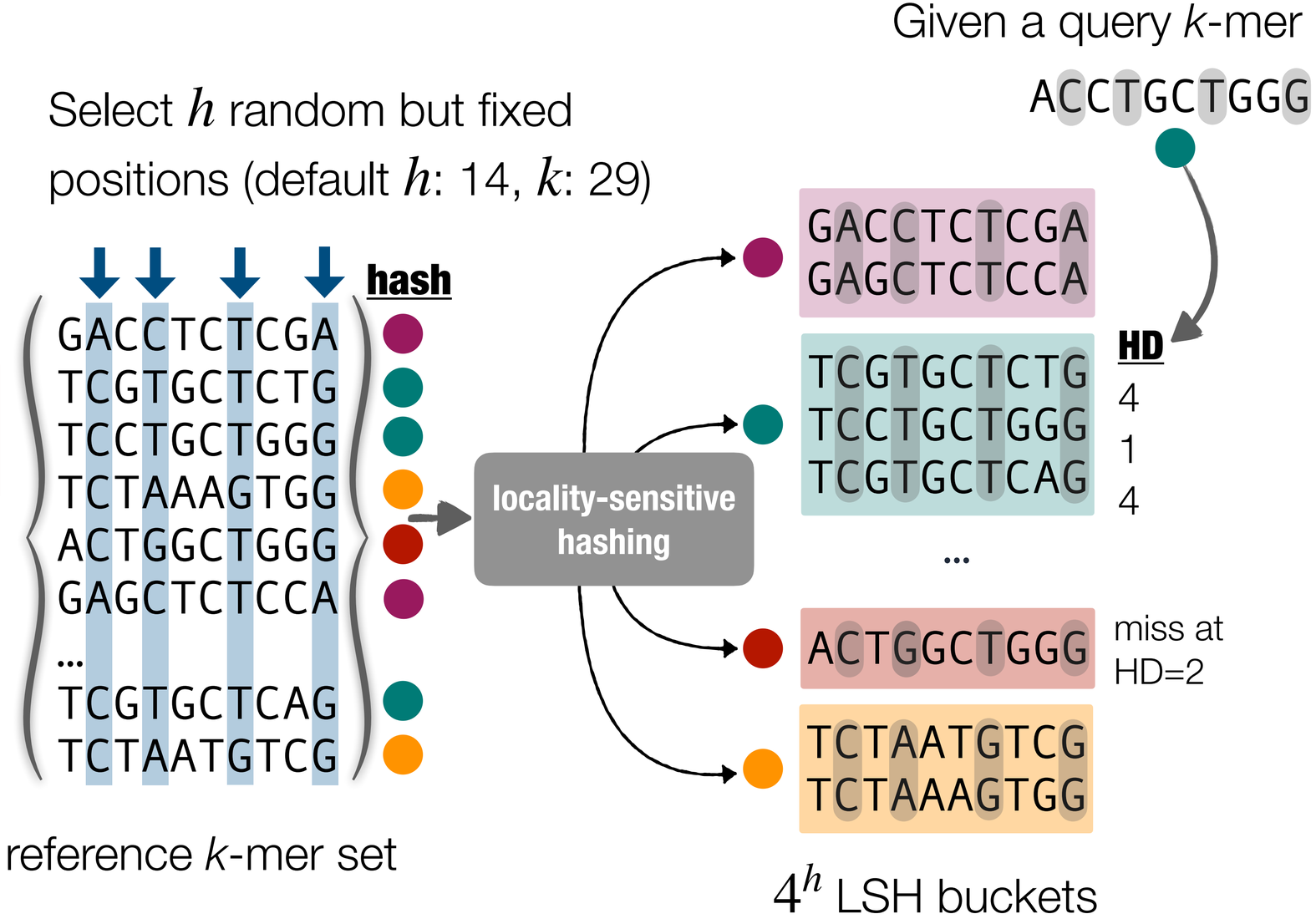
Computing Hamming distances between homologous k-mers using LSH

Goal: allow mismatches in *k*-mers, measure the **Hamming distance**



Computing Hamming distances between homologous k-mers using LSH

Goal: allow mismatches in *k*-mers, measure the **Hamming distance**

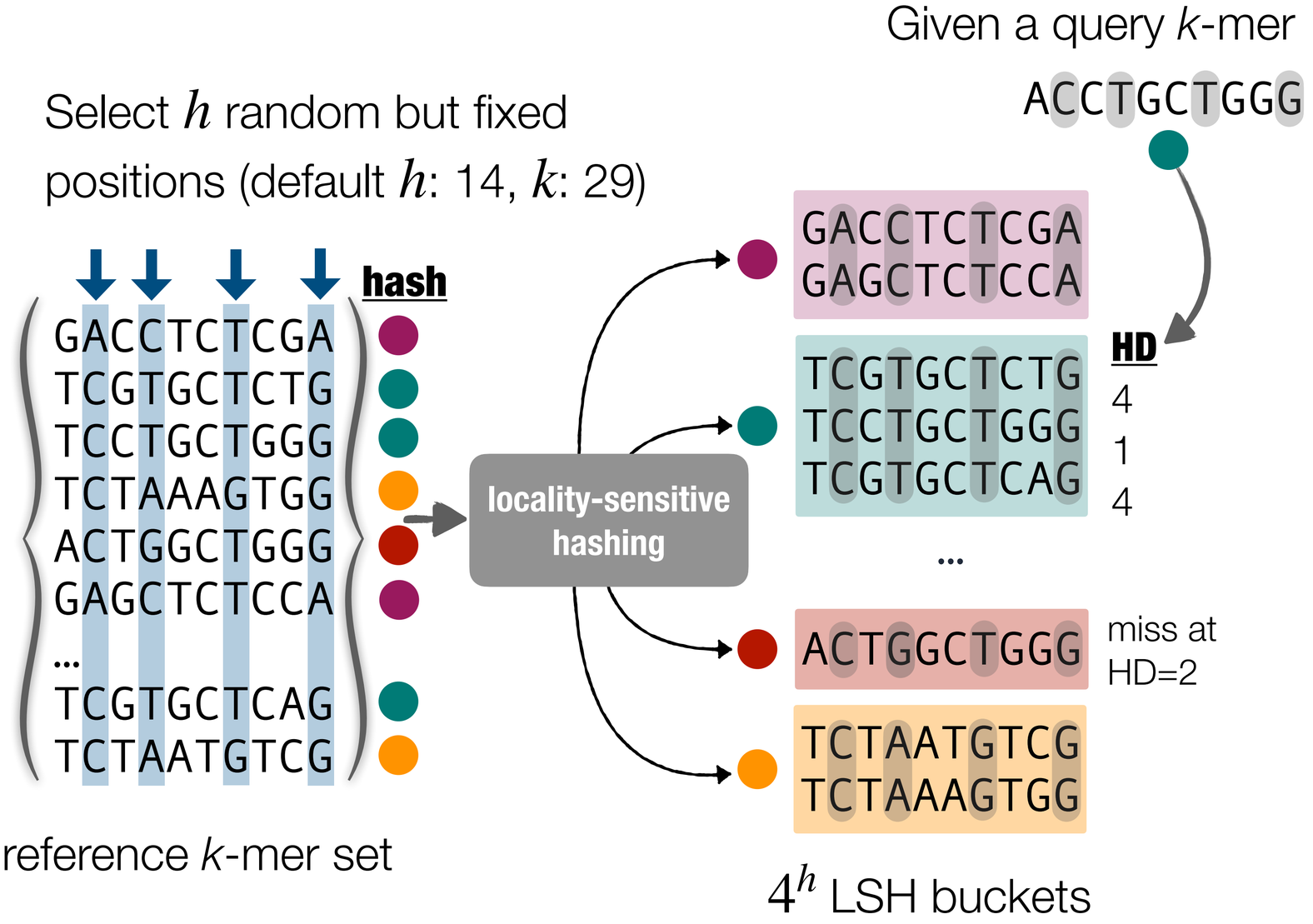


Computing Hamming distances between homologous k-mers using LSH

Goal: allow mismatches in *k*-mers, measure the **Hamming distance**

CONSULT-II: accurate taxonomic identification and profiling using locality-sensitive hashing

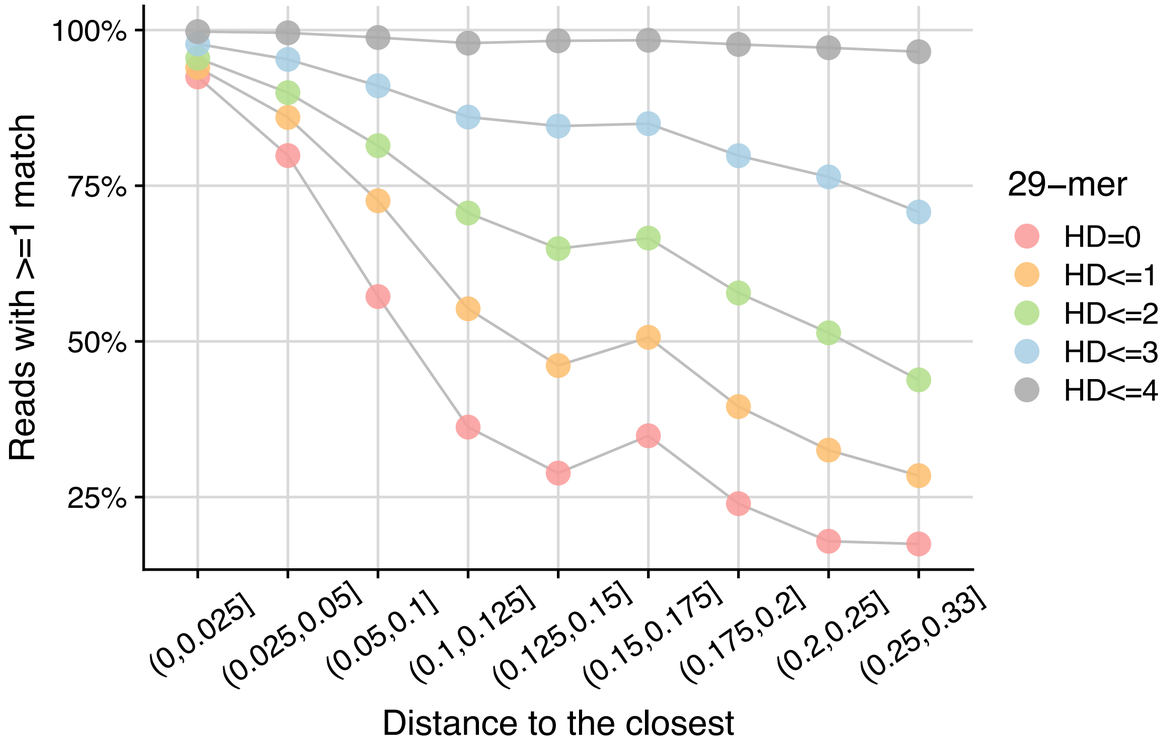
Ali Osman Berk Şapcı [1], Eleonora Rachtman [1], Siavash Mirarab [1,2,*]



CONSULT-II goes beyond exact k-mers

CONSULT-II goes beyond exact k-mers

- **Fast:** limited number of HD computations
- **More sensitive** than exact *k*-mer search



Lowest common ancestor is sensitive to outliers

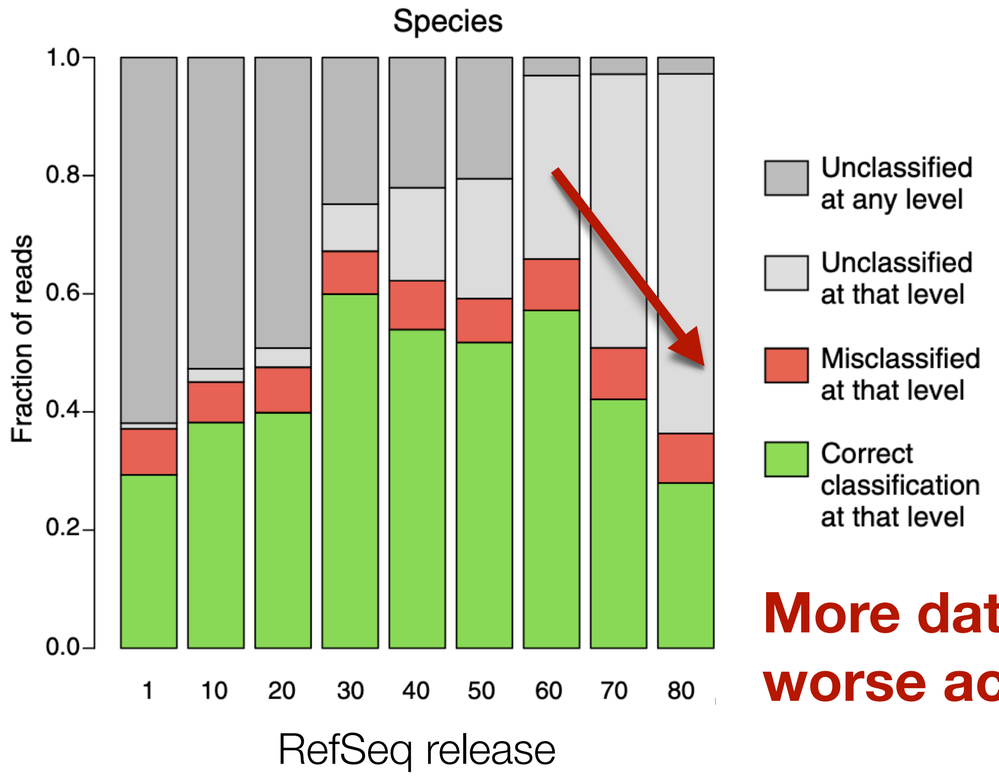
reference errors, contamination, complex evolutionary processes (e.g., HGT)

Lowest common ancestor is sensitive to outliers

reference errors, contamination, complex evolutionary processes (e.g., HGT)

RefSeq database growth influences the accuracy of *k*-mer-based lowest common ancestor species identification

Daniel J. Nasko[1], Sergey Koren[2], Adam M. Phillippy[2] and Todd J. Treangen[3*]

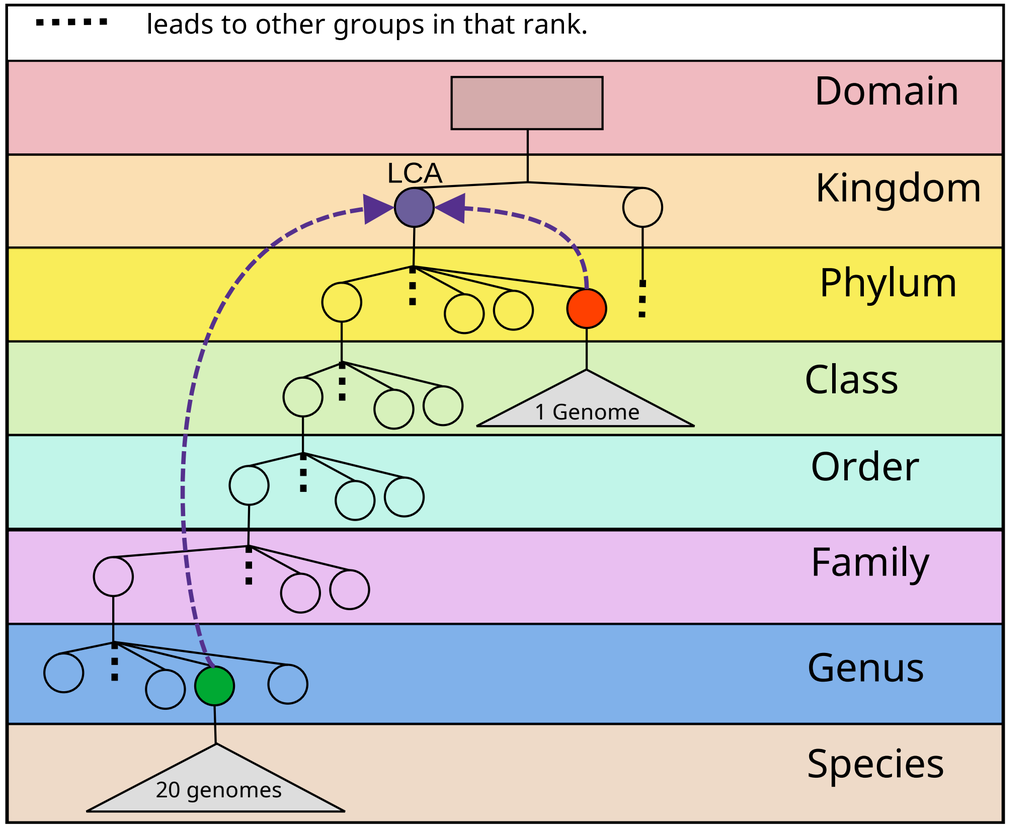


More data; worse accuracy

Lowest common ancestor is sensitive to outliers

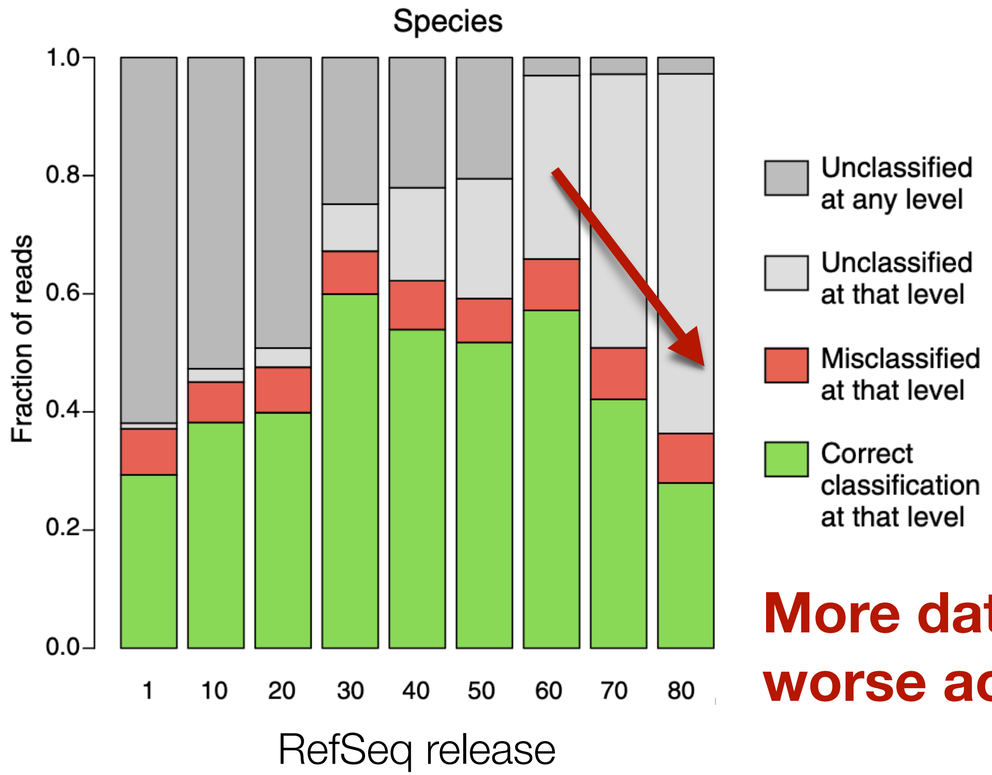
reference errors, contamination, complex evolutionary processes (e.g., HGT)

- **Correct genus:** 20 genomes
- **Erroneous phylum:** 1 genome
- Kraken would push the LCA to the **kingdom**



RefSeq database growth influences the accuracy of *k*-mer-based lowest common ancestor species identification

Daniel J. Nasko[1], Sergey Koren[2], Adam M. Phillippy[2] and Todd J. Treangen[3*]



More data; worse accuracy

CONSULT-II:

Can we make LCA less sensitive to outliers?

Soft-LCA of CONSULT-II is less sensitive to outliers

- **Idea:** ignore each genome with some probability
- **Intuition:** a *k*-mer has to appear in *sufficiently* many genomes under a taxon to have an effect

Soft-LCA of CONSULT-II is less sensitive to outliers

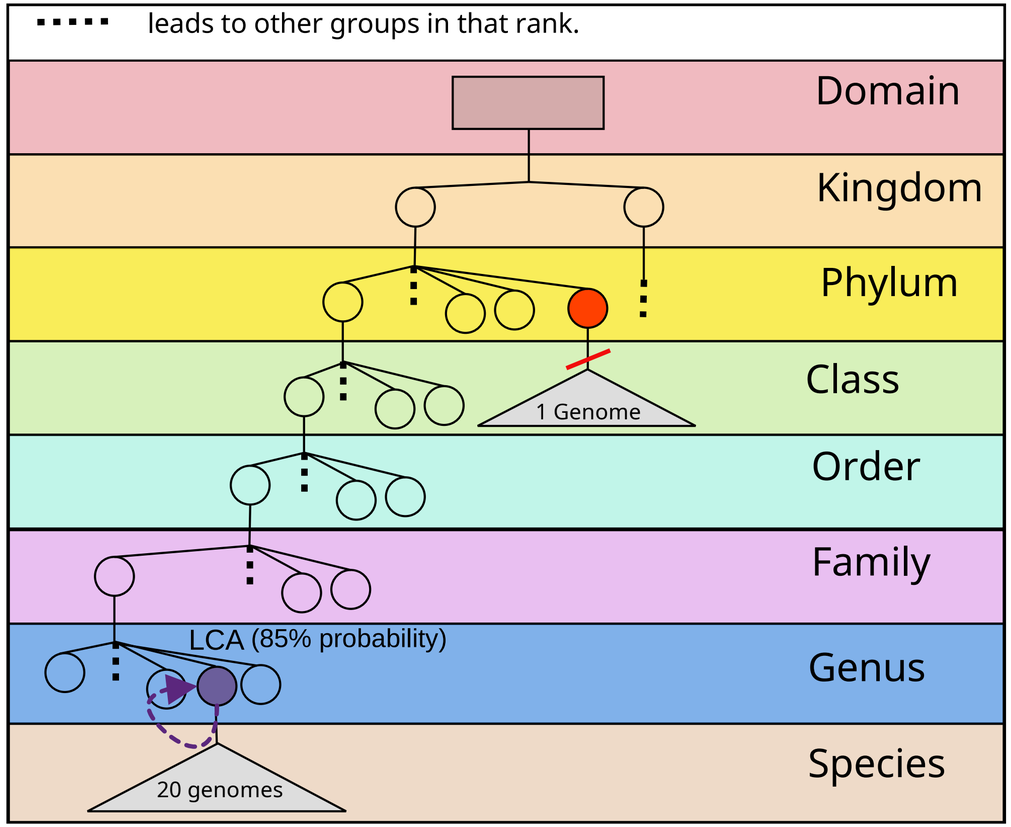
- **Idea:** ignore each genome with some probability
- **Intuition:** a *k*-mer has to appear in *sufficiently* many genomes under a taxon to have an effect
- $p(n)$: success prob. for a *k*-mer appearing in n genomes
  - frequent *k*-mers → many times
  - rare *k*-mers → a few would suffice

Soft-LCA of CONSULT-II is less sensitive to outliers

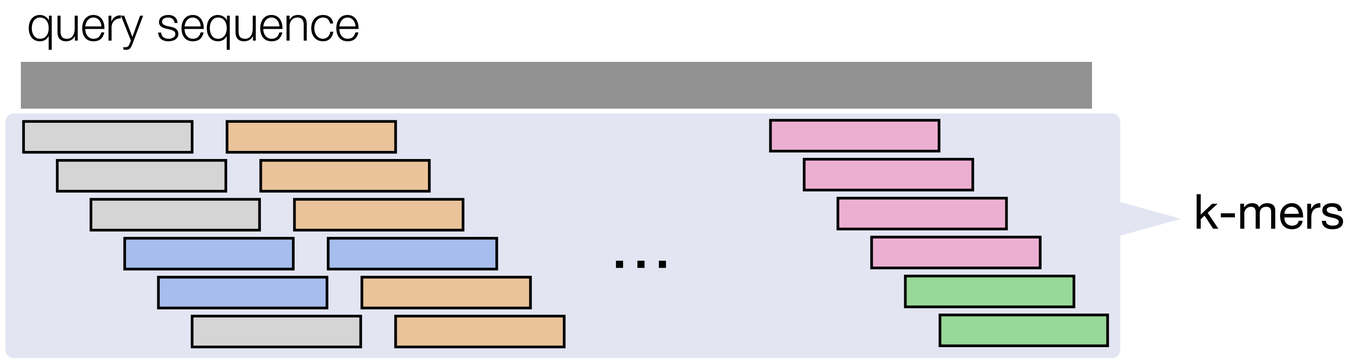
$$P(\text{at least 1 out of 20 is not ignored})P(\text{1 is ignored}) \approx 0.85$$

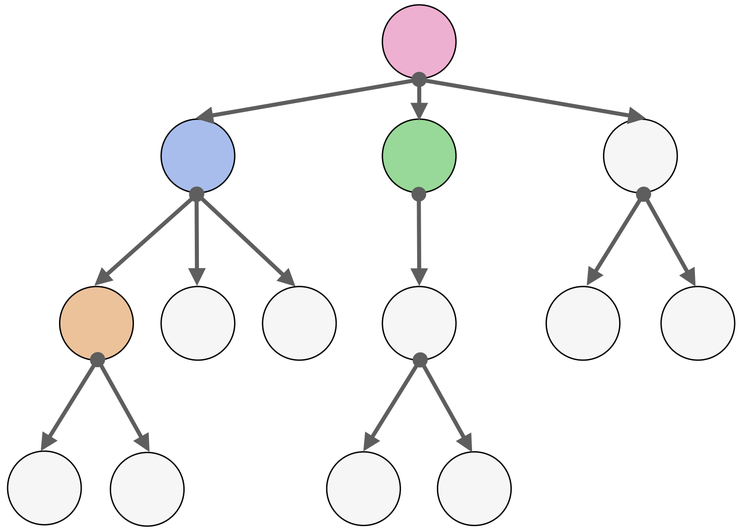
$$1-(1-p(21))^{20} \qquad 1-p(21)$$

- **Idea:** ignore each genome with some probability
- **Intuition:** a *k*-mer has to appear in *sufficiently* many genomes under a taxon to have an effect
- $p(n)$: success prob. for a *k*-mer appearing in n genomes
  - frequent *k*-mers → many times
  - rare *k*-mers → a few would suffice



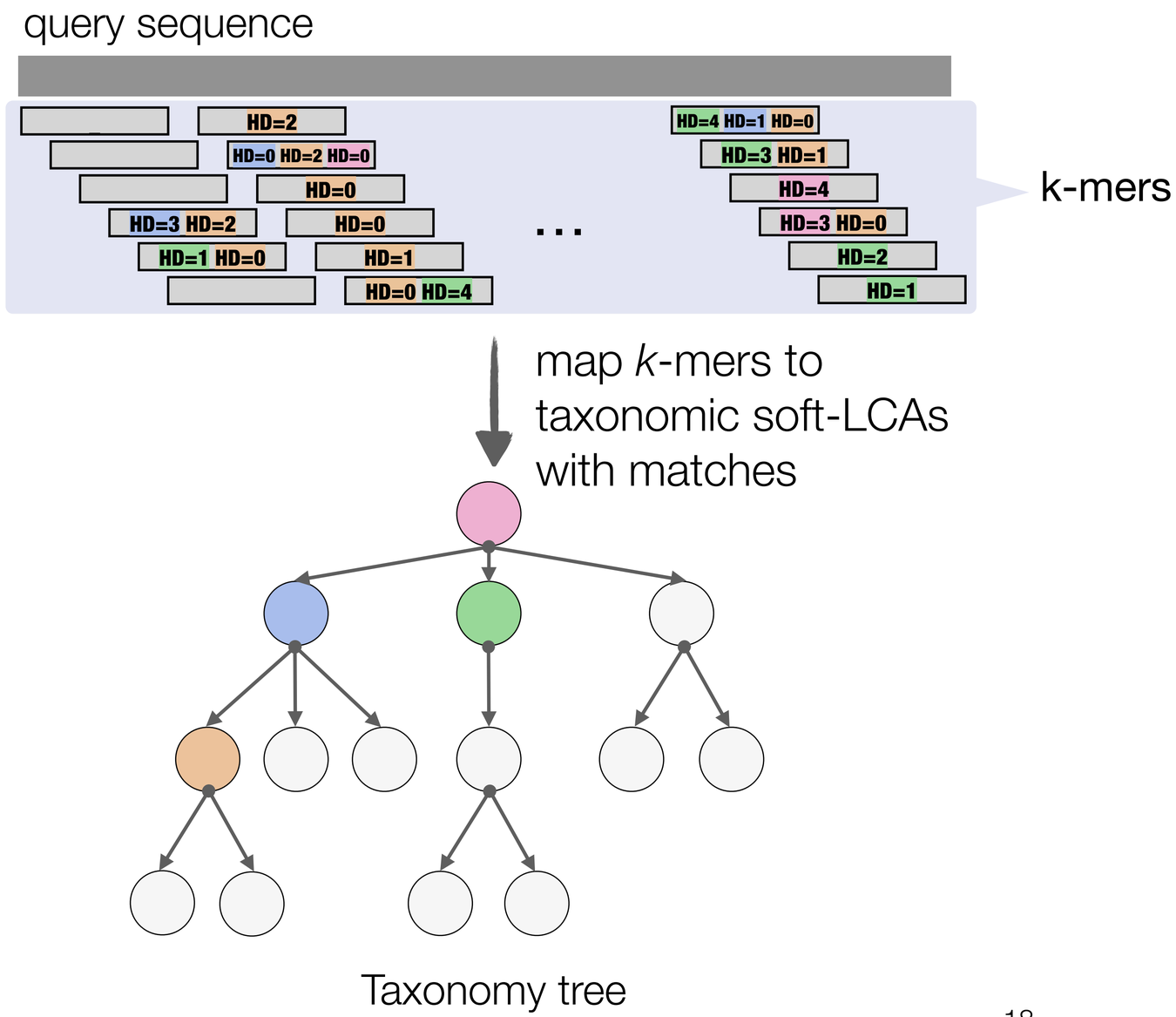
k-mers vote for matching taxa according to HDs





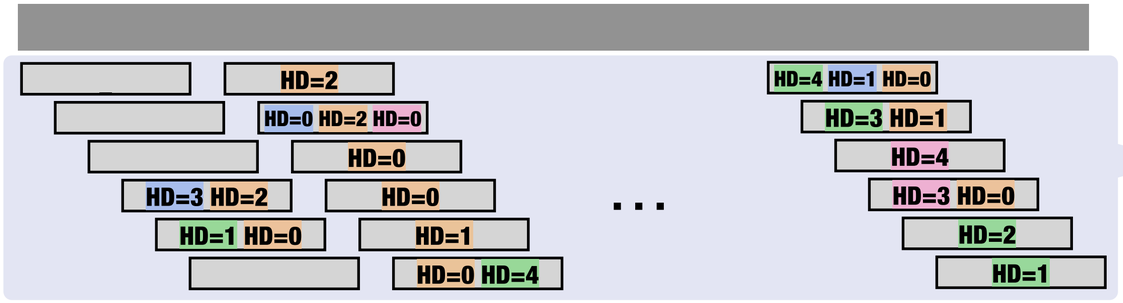
Taxonomy tree

k-mers vote for matching taxa according to HDs



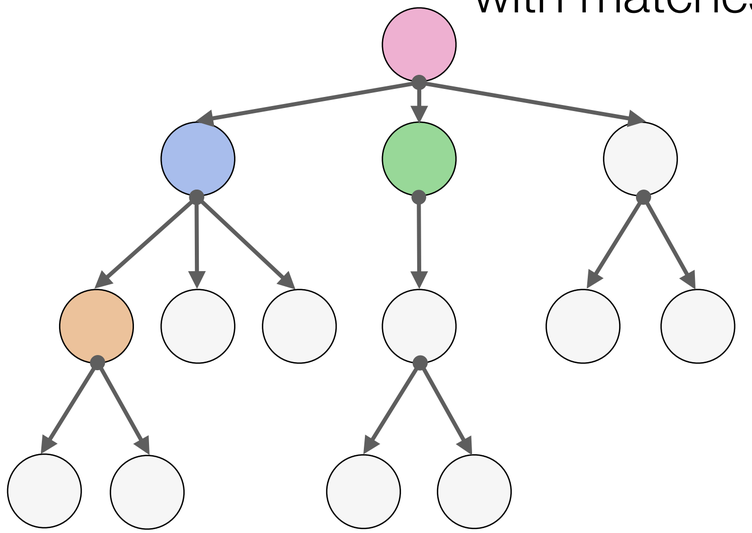
k-mers vote for matching taxa according to HDs

query sequence



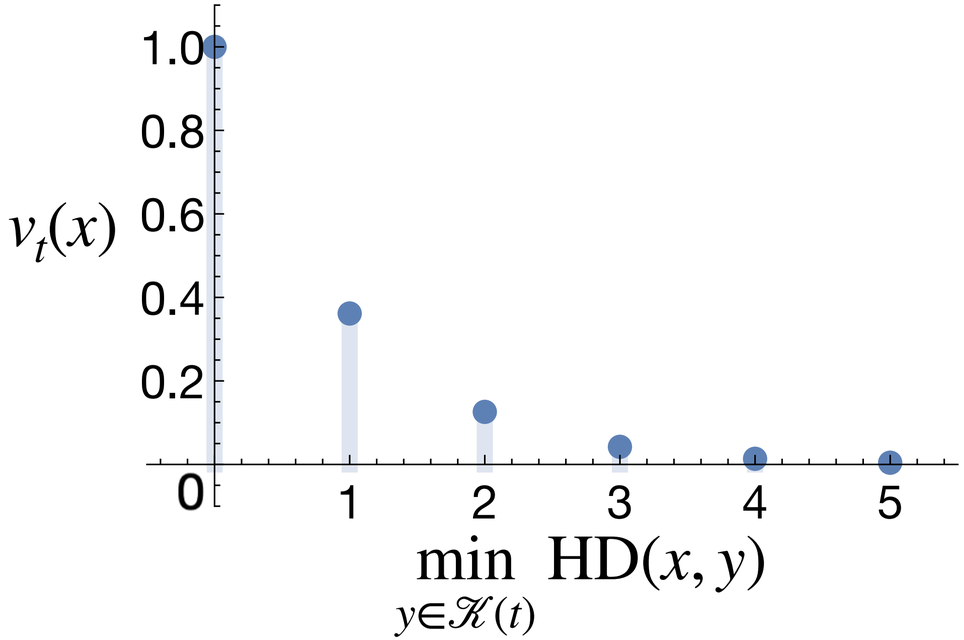
k-mers

map *k*-mers to taxonomic soft-LCAs with matches



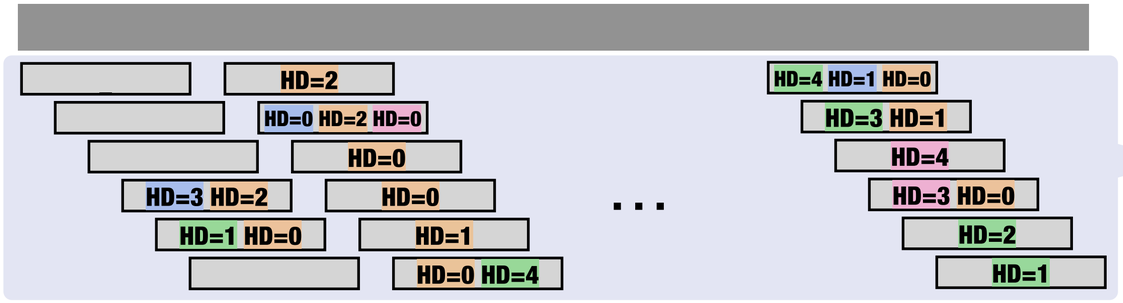
Taxonomy tree

Each match votes to taxa and their parents weighted disproportionately by HDs



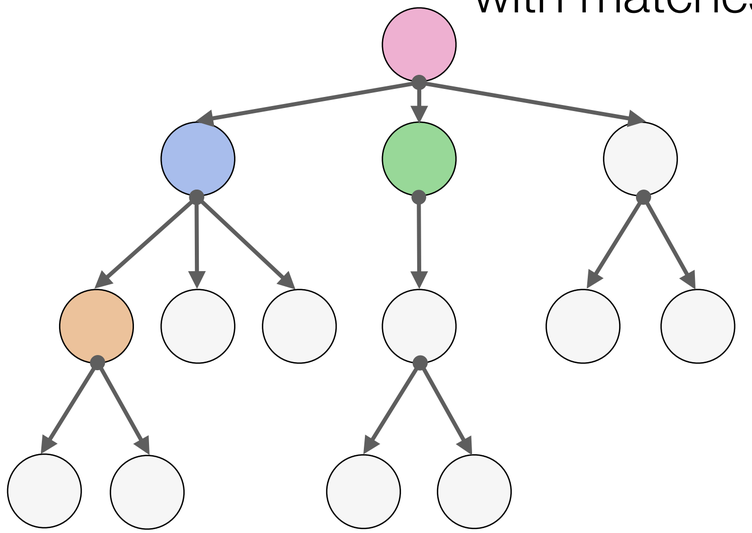
k-mers vote for matching taxa according to HDs

query sequence



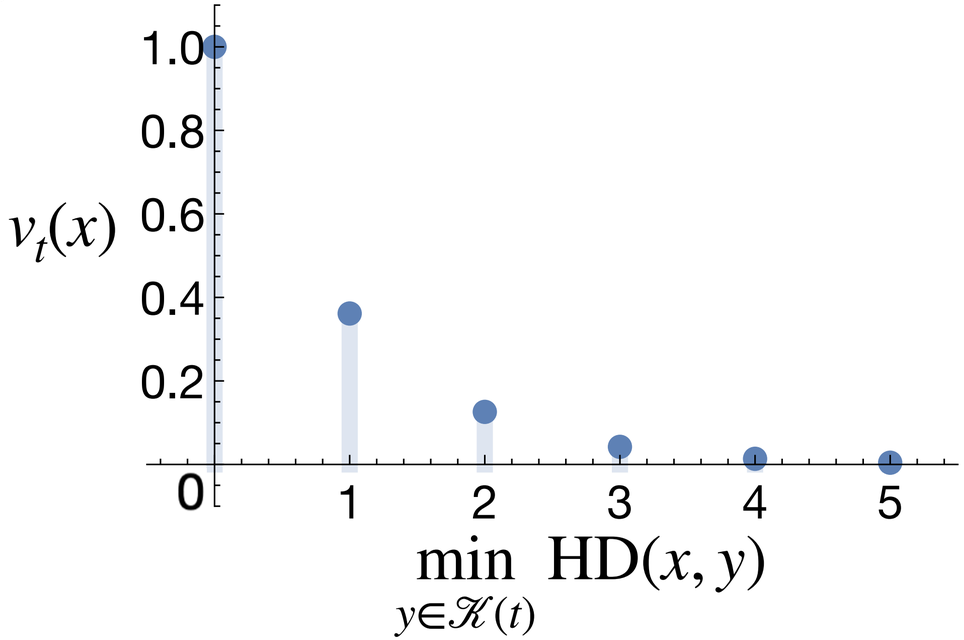
k-mers

map *k*-mers to taxonomic soft-LCAs with matches



Taxonomy tree

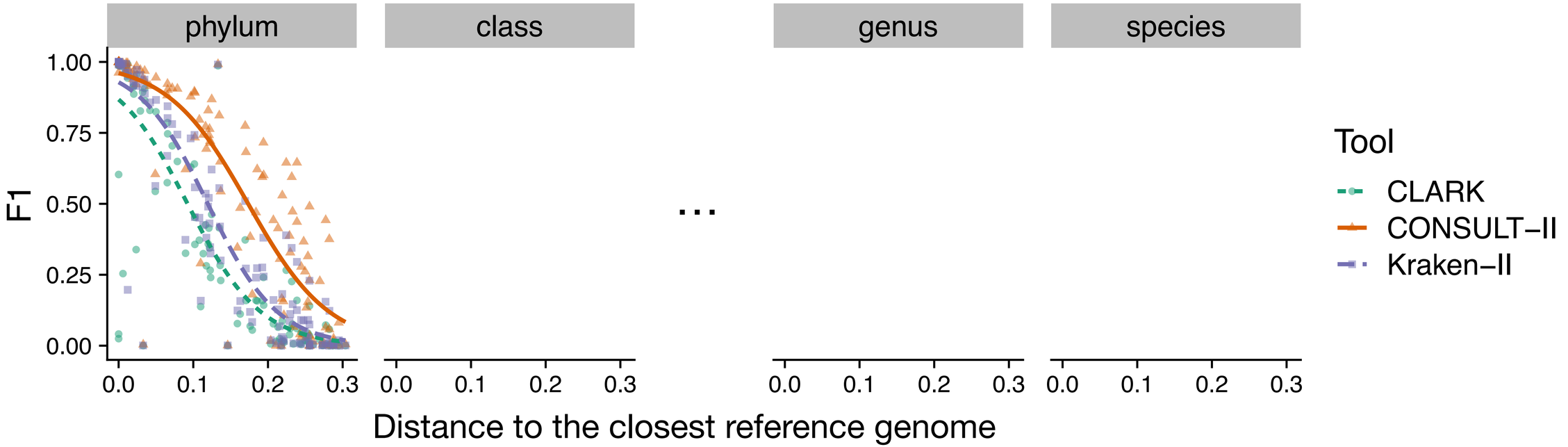
Each match votes to taxa and their parents weighted disproportionately by HDs



- **Classification:** require a majority vote (half of the vote at the root)
- **Profiling:** normalize votes at each rank

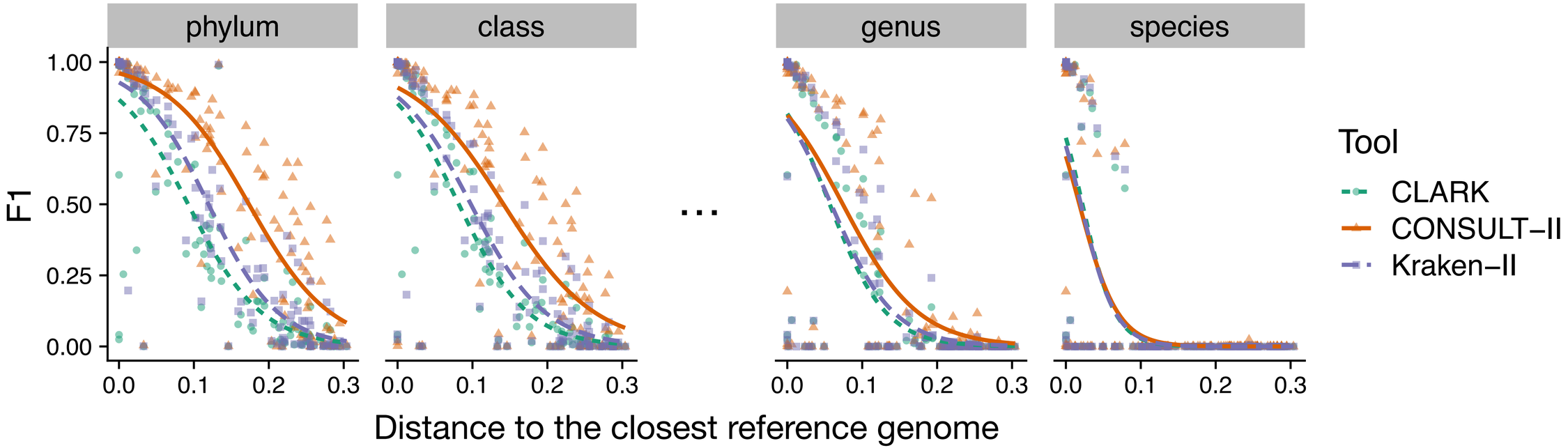
CONSULT-II can classify reads from distant genomes

- Query genomes spanning different **novelty levels** using Mash [Ondov et al., 2016]
- **Short reads** simulated from 120 bacterial & 100 archaeal genomes with errors
- WoL reference dataset: 10,595 microbial genomes



CONSULT-II can classify reads from distant genomes

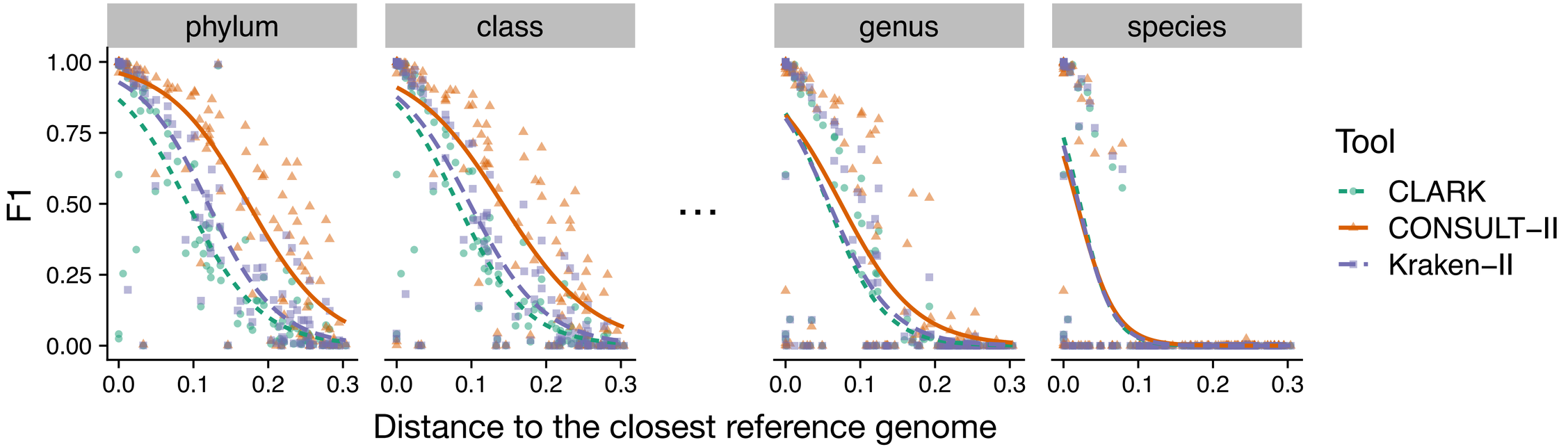
- Query genomes spanning different **novelty levels** using Mash [Ondov et al., 2016]
- **Short reads** simulated from 120 bacterial & 100 archaeal genomes with errors
- WoL reference dataset: 10,595 microbial genomes



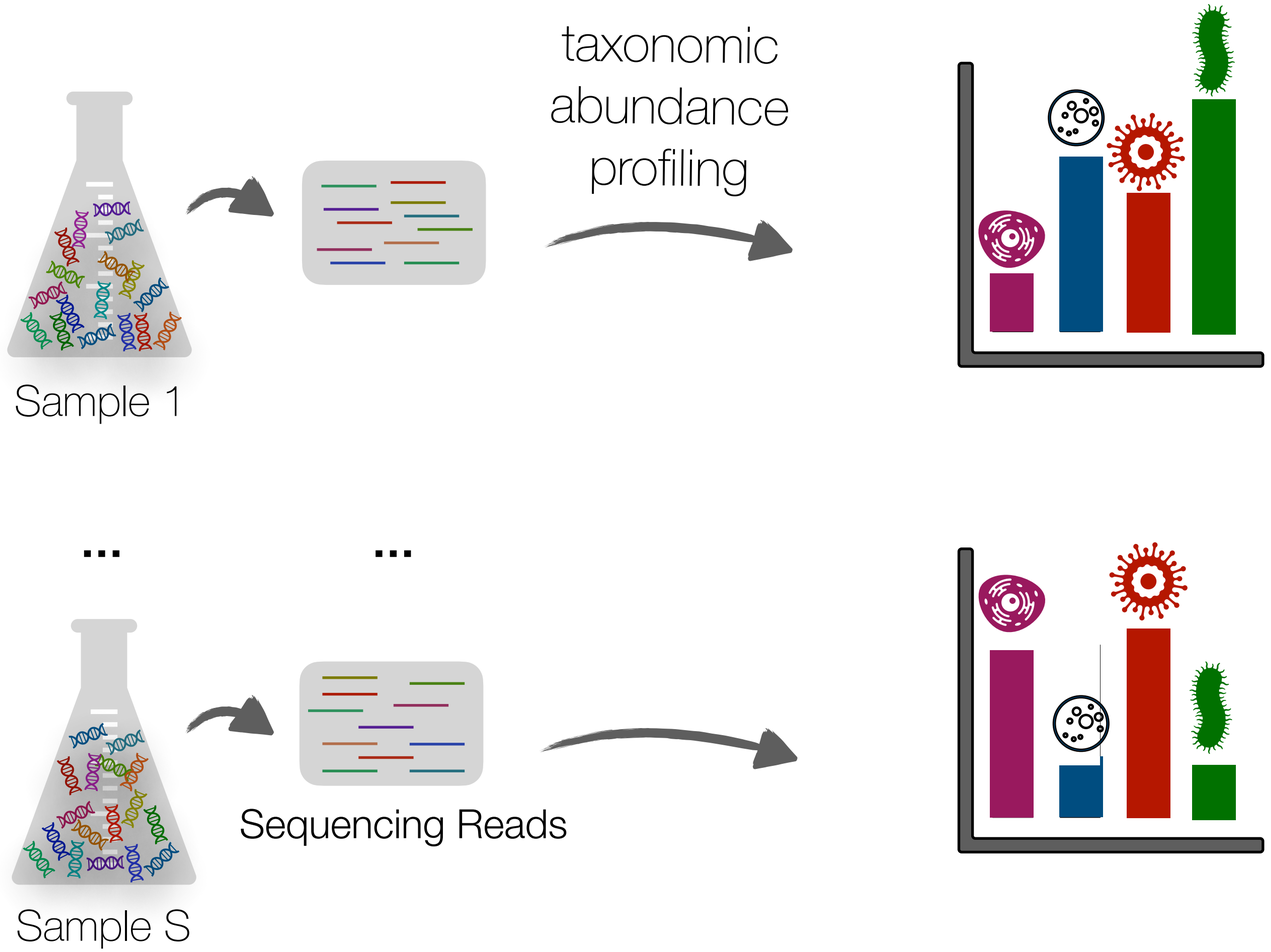
CONSULT-II can classify reads from distant genomes

- Query genomes spanning different **novelty levels** using Mash [Ondov et al., 2016]
- **Short reads** simulated from 120 bacterial & 100 archaeal genomes with errors
- WoL reference dataset: 10,595 microbial genomes

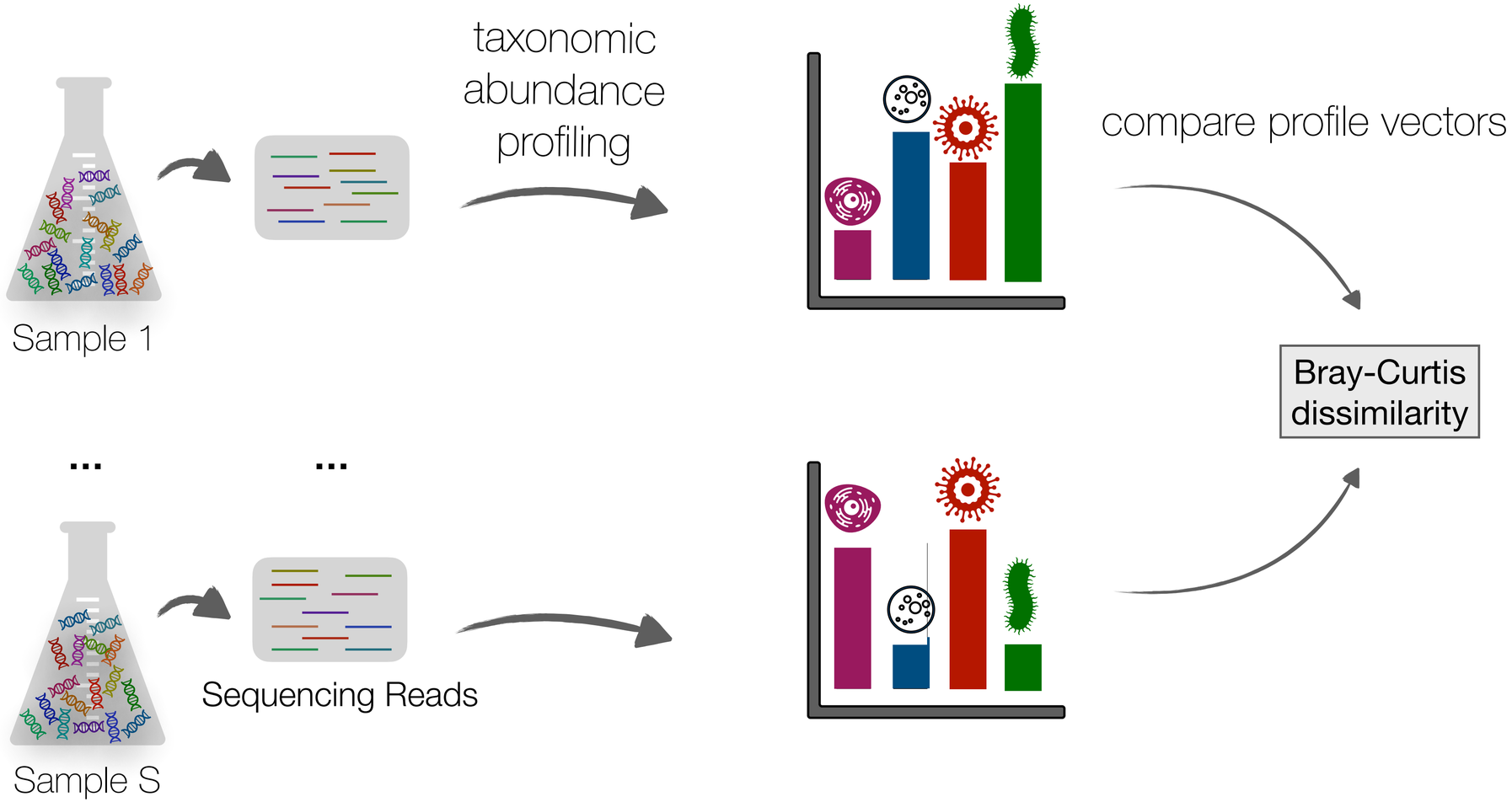
Better recall with no expense of precision!



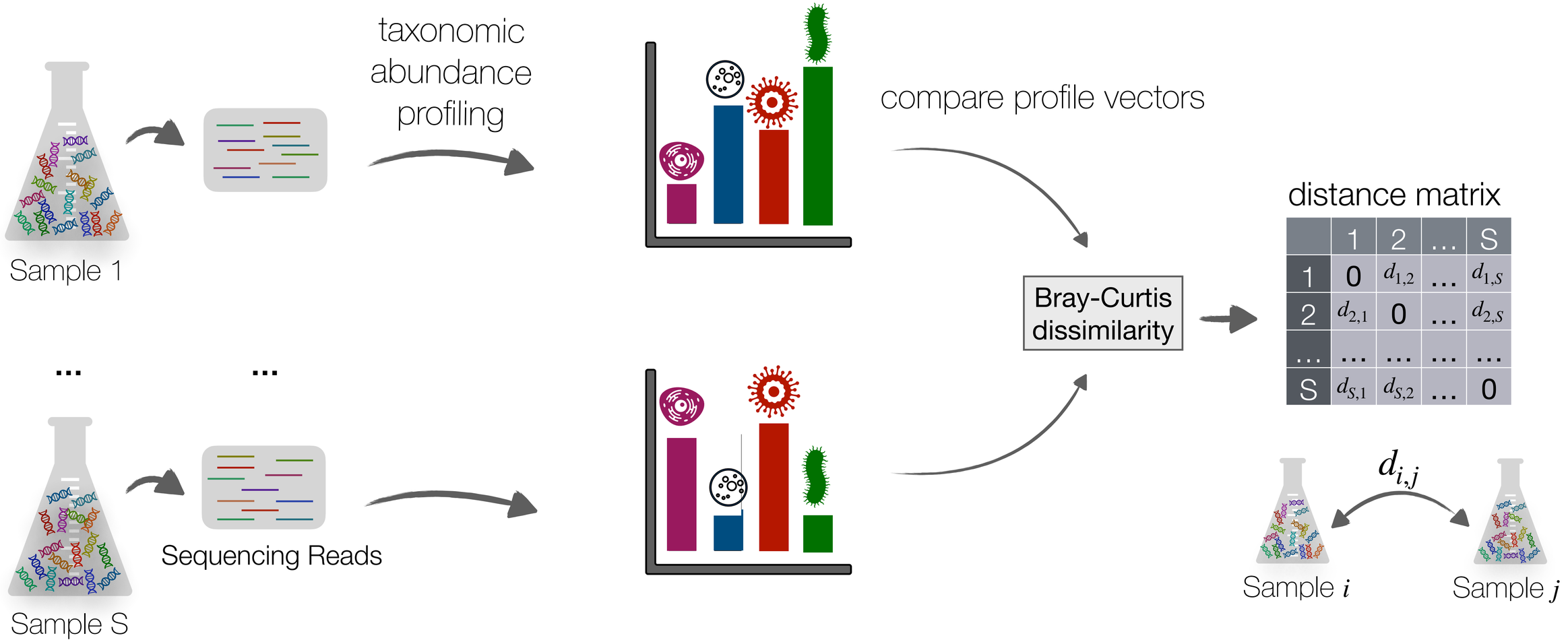
A practical aspect: comparing samples downstream



A practical aspect: comparing samples downstream

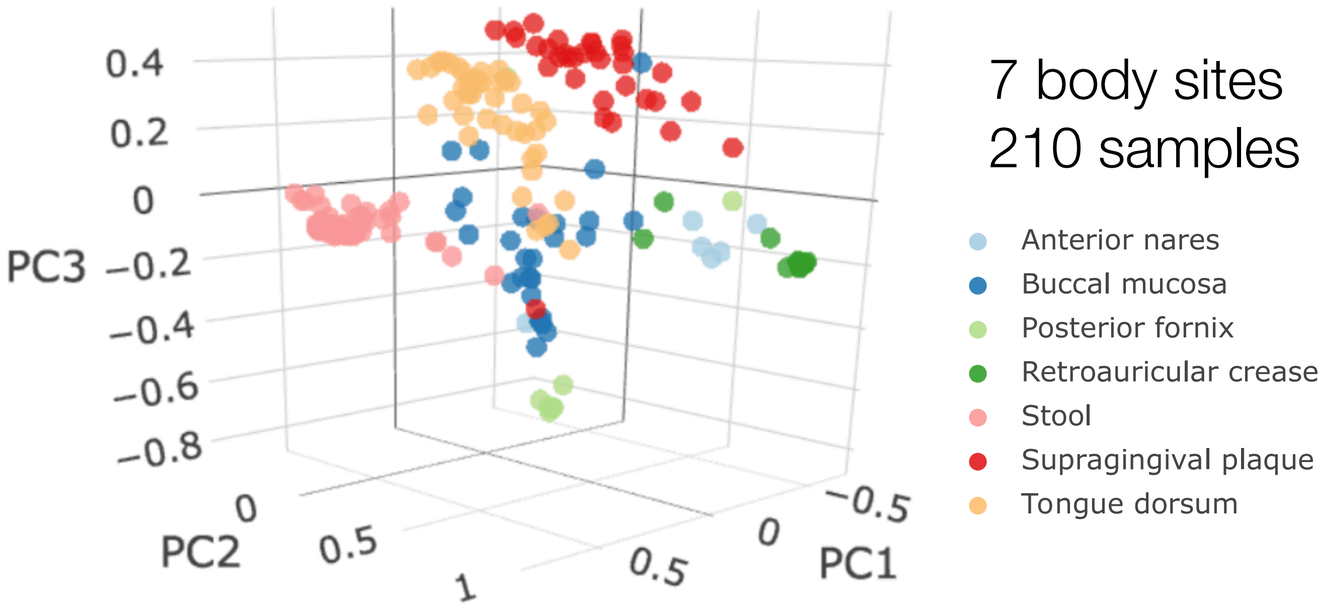


A practical aspect: comparing samples downstream



Comparing microbiomes from different body sites

Visualizing the distance matrix with PCoA

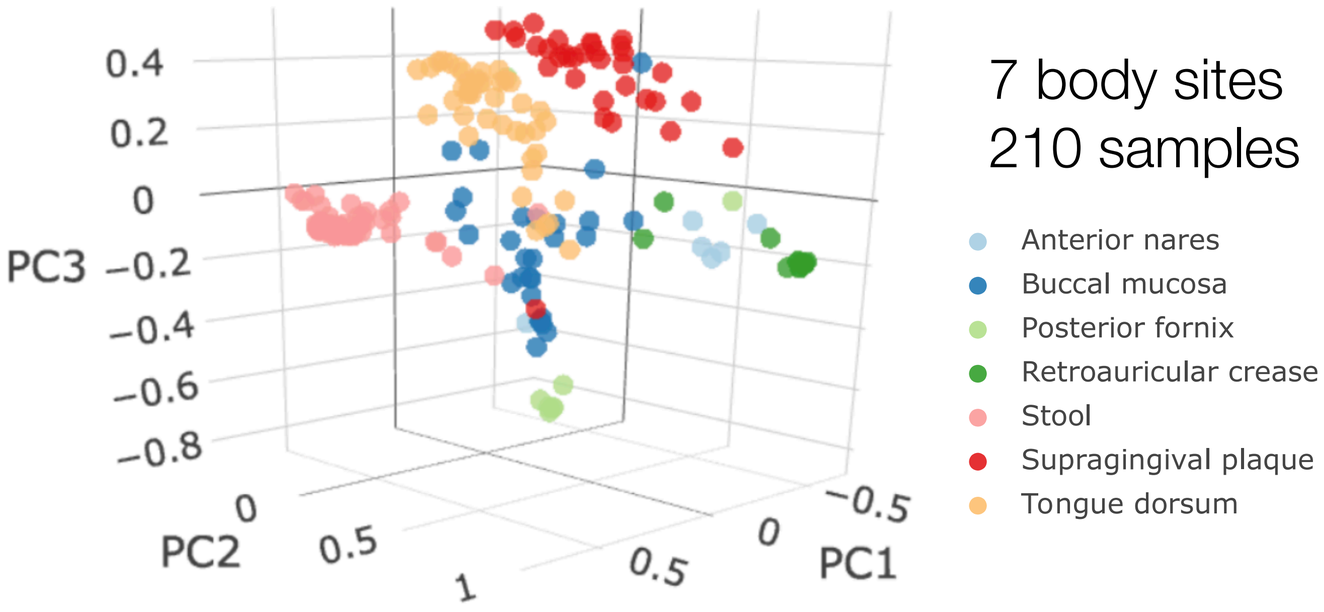


7 body sites
210 samples

- Anterior nares
- Buccal mucosa
- Posterior fornix
- Retroauricular crease
- Stool
- Supragingival plaque
- Tongue dorsum

Comparing microbiomes from different body sites

Visualizing the distance matrix with PCoA

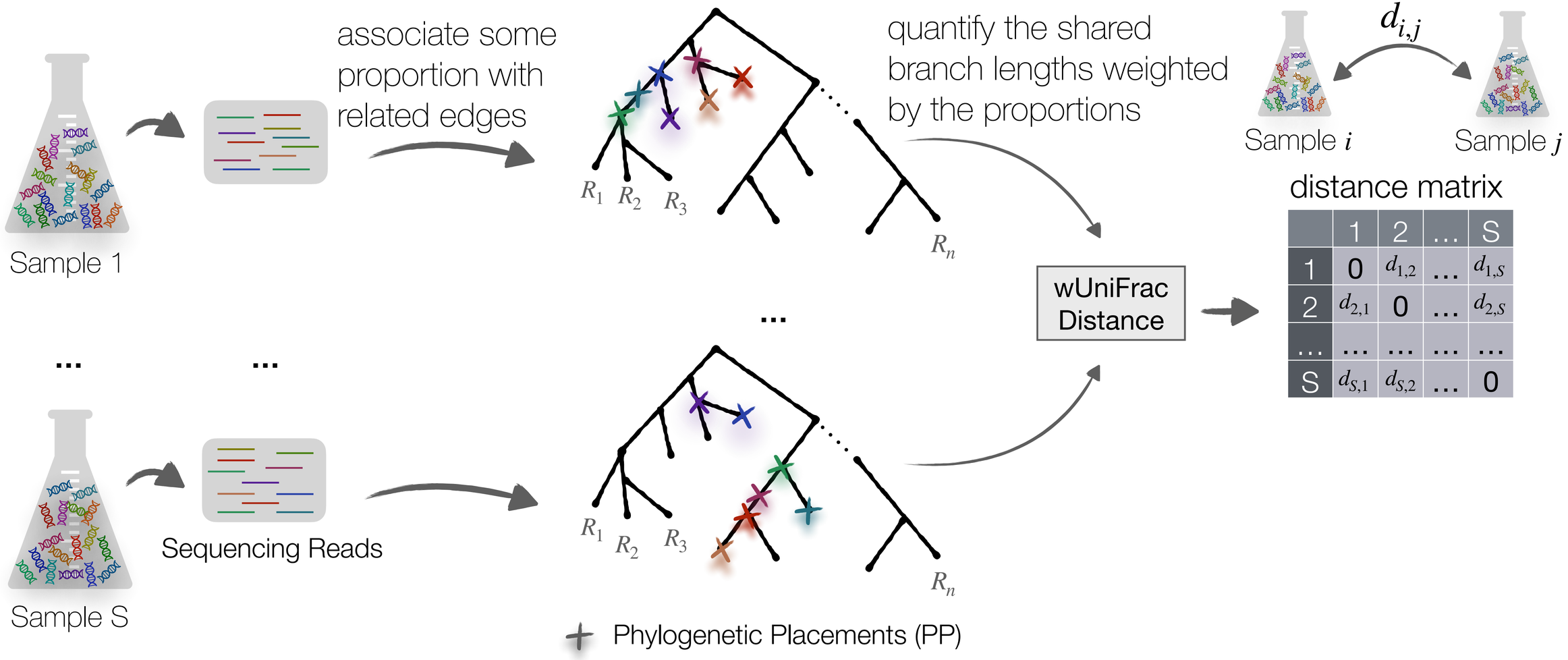


7 body sites
210 samples

- Anterior nares
- Buccal mucosa
- Posterior fornix
- Retroauricular crease
- Stool
- Supragingival plaque
- Tongue dorsum

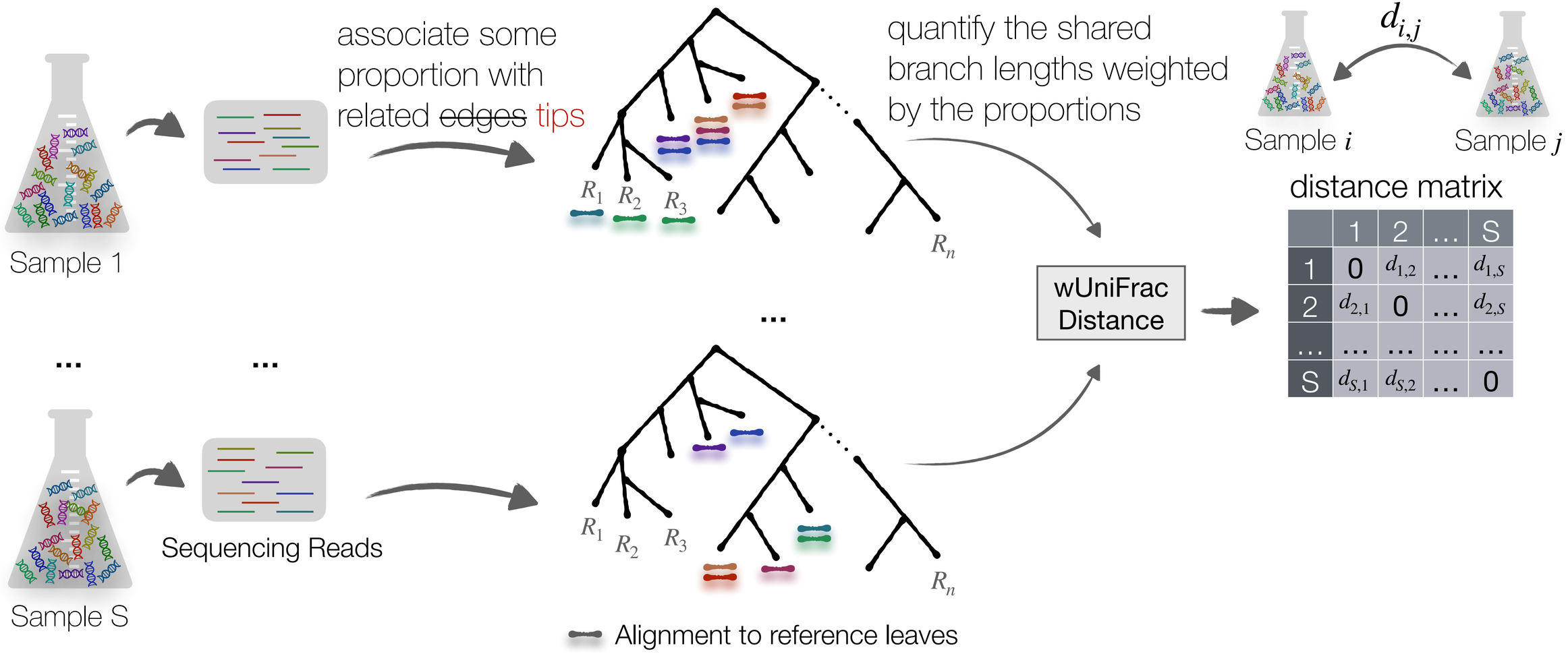
Reference: Web of Life (v2)
16,000 microbial genomes

What if we also have a reference phylogeny?



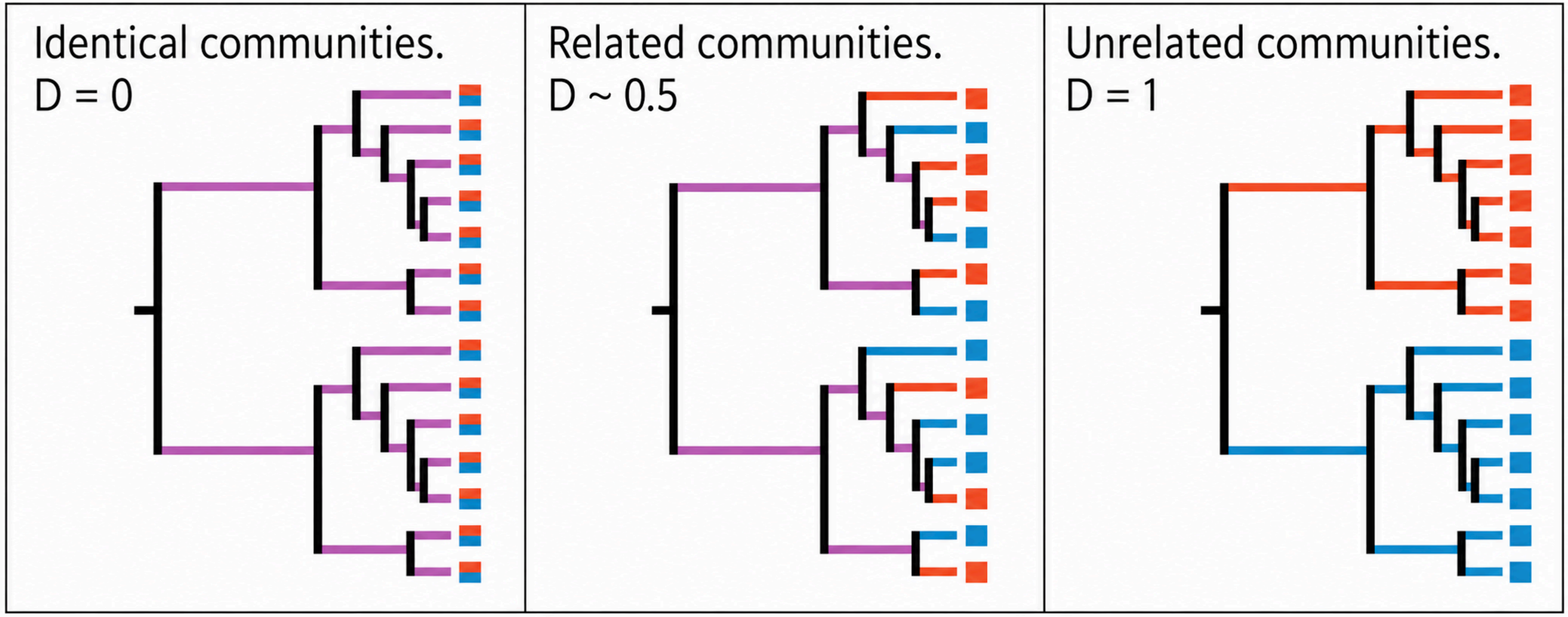
	1	2	…	S
1	0	$d_{1,2}$	…	$d_{1,S}$
2	$d_{2,1}$	0	…	$d_{2,S}$
…	…	…	…	…
S	$d_{S,1}$	$d_{S,2}$	…	0

What if we also have a reference 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

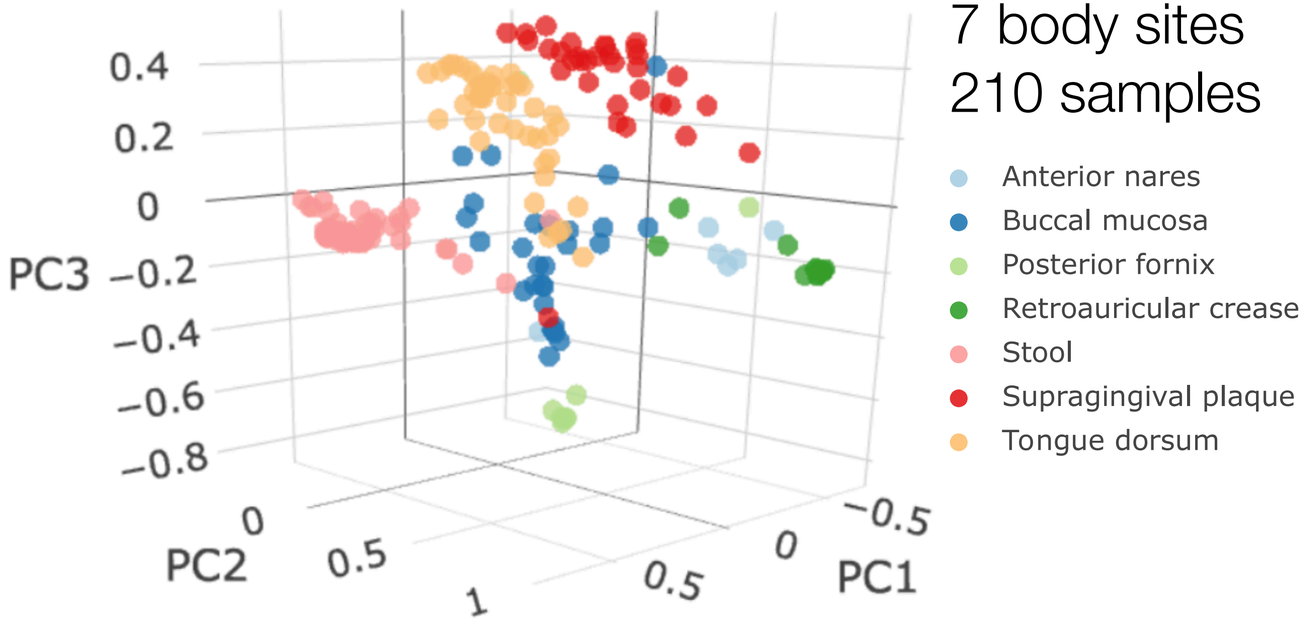
Unique Fraction (UniFrac) metric



[Lozupone et al. 2005]

Better resolution with alignment and UniFrac

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



7 body sites
210 samples

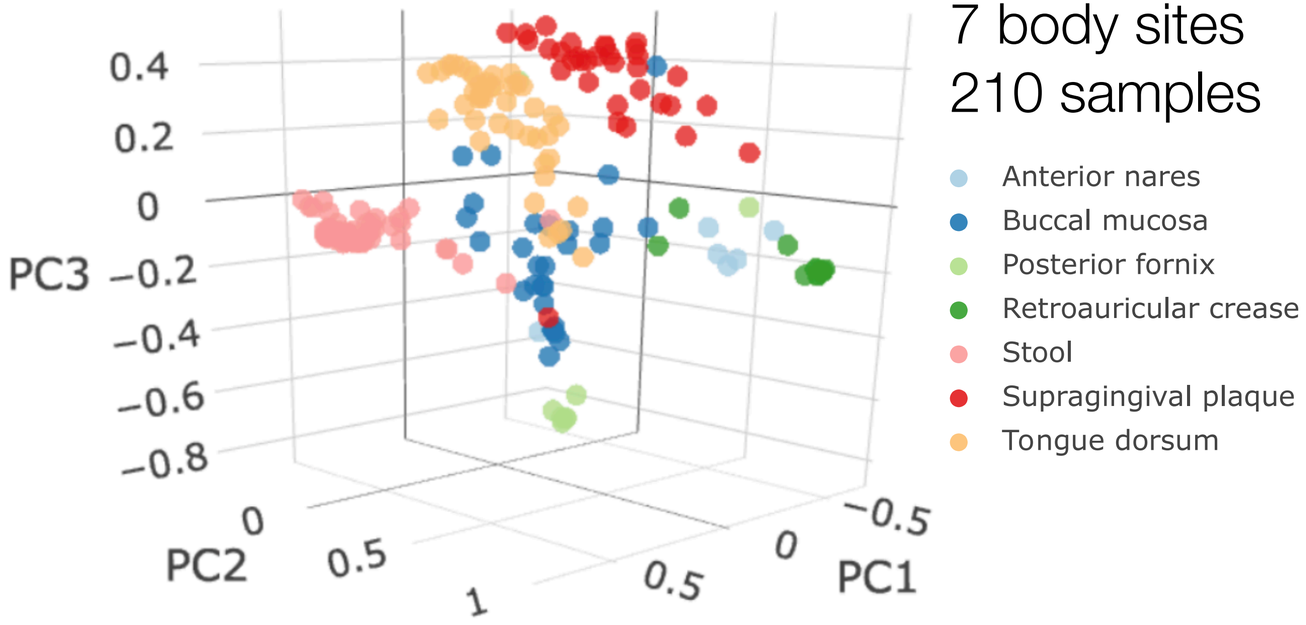
- Anterior nares
- Buccal mucosa
- Posterior fornix
- Retroauricular crease
- Stool
- Supragingival plaque
- Tongue dorsum

Evaluating distances:

pseudo *F* statistic: compare within group versus across group distances

Better resolution with alignment and UniFrac

UniFrac achieves much better separation



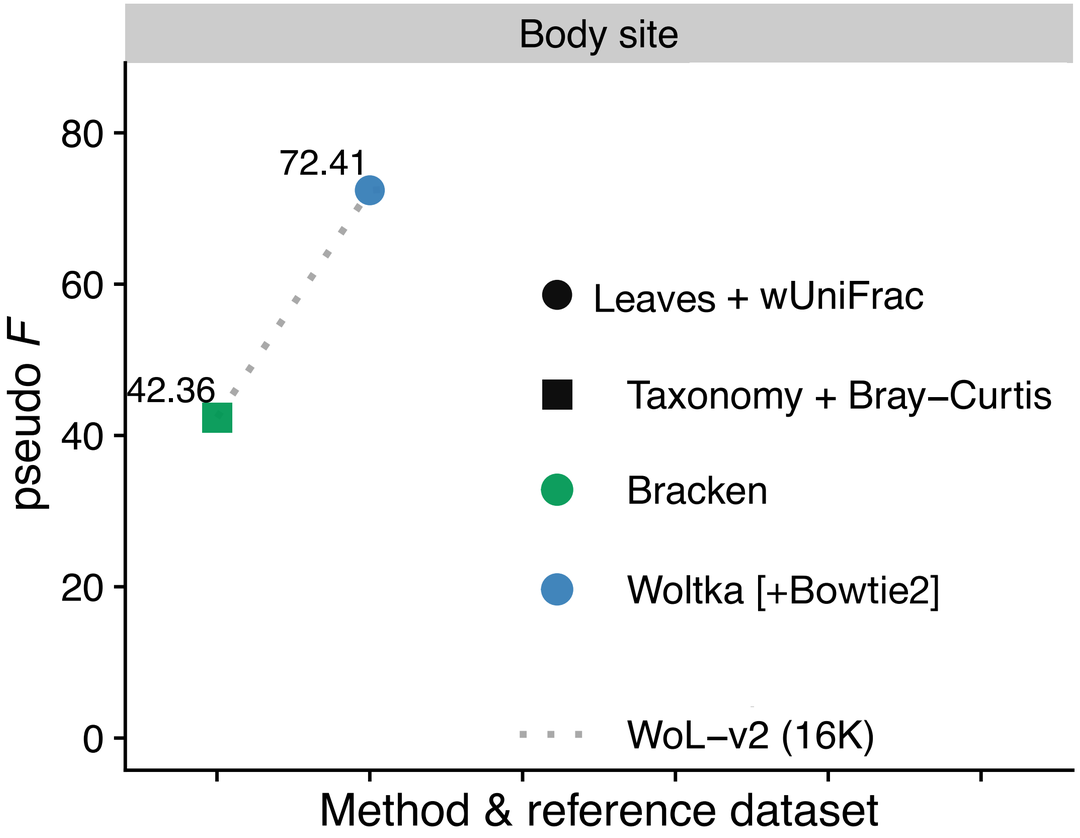
7 body sites
210 samples

- Anterior nares
- Buccal mucosa
- Posterior fornix
- Retroauricular crease
- Stool
- Supragingival plaque
- Tongue dorsum

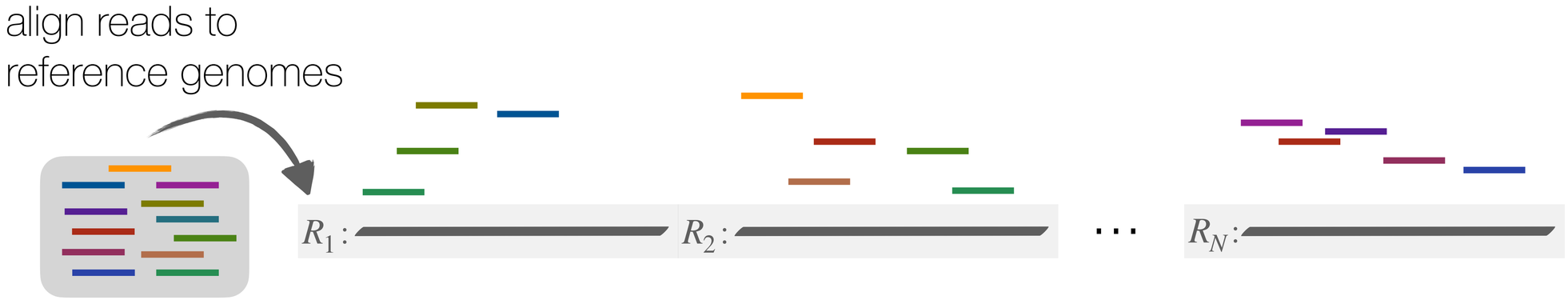
Evaluating distances:

pseudo *F* statistic: compare within group versus across group distances

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

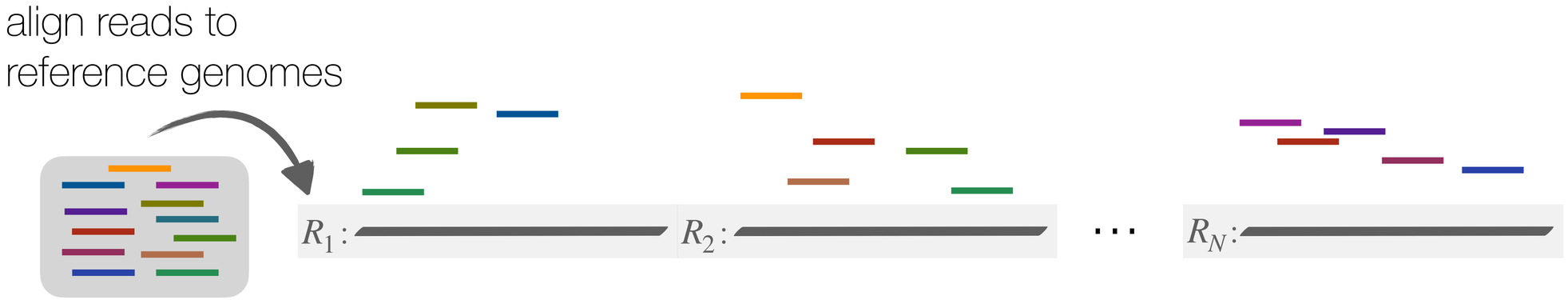


OGU: challenges of aligning reads



(+) quantifying similarity — as detailed as it gets

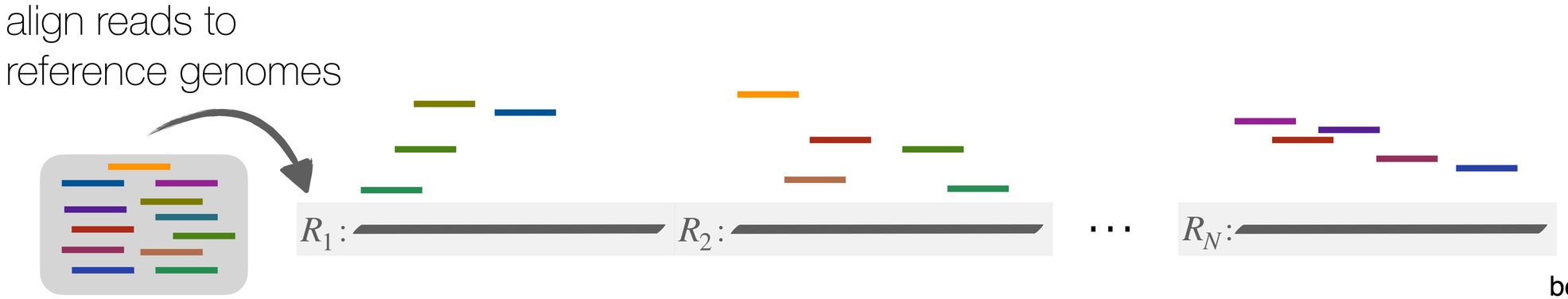
OGU: challenges of aligning reads



(+) quantifying similarity — as detailed as it gets

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

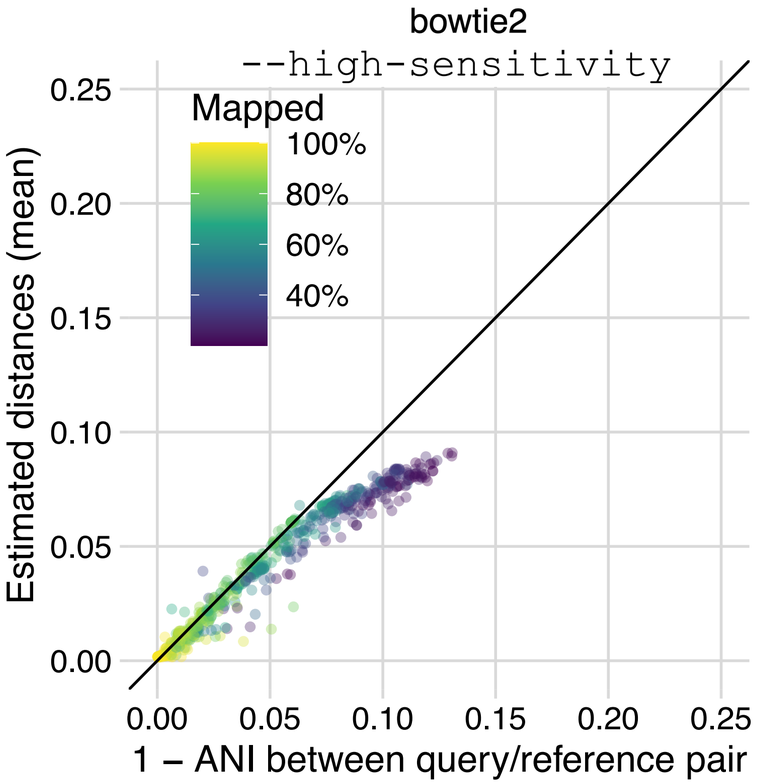
OGU: challenges of aligning reads

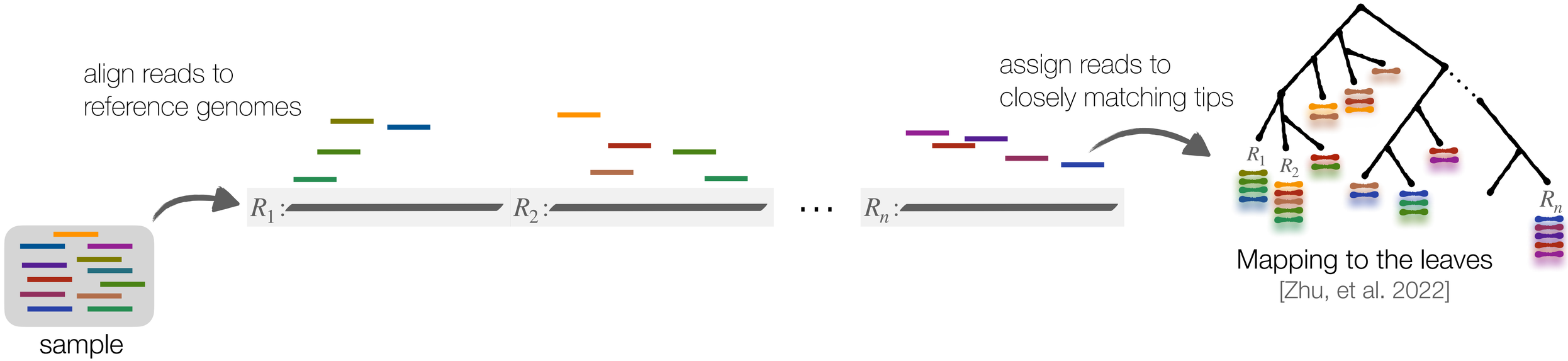


(+) quantifying similarity — as detailed as it gets

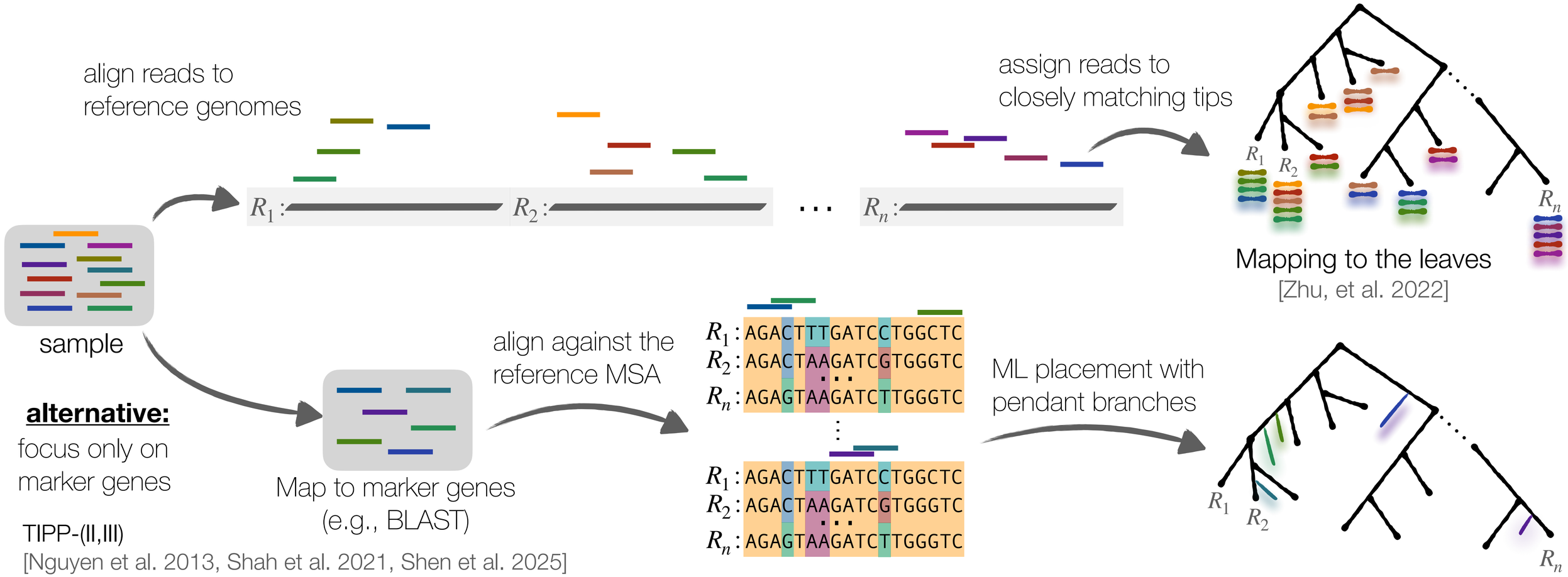
(-) not scalable for large N, even with efficient indexes

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




align reads to
reference genomes
R_1 :
R_2 :
...
R_n :
sample
assign reads to
closely matching tips
R_1
R_2
R_n
Mapping to the leaves
[Zhu, et al. 2022]

Phylogenetic placement has been (mostly) limited to marker genes *or* small phylogenies.

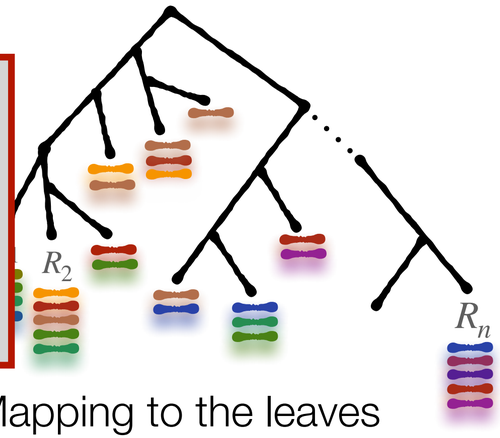


TIPP-(II,III)
[Nguyen et al. 2013, Shah et al. 2021, Shen et al. 2025]

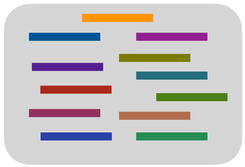
Phylogenetic placement has been (mostly) limited to marker genes *or* small phylogenies.

align reads to reference ge

Can we estimate accurate read to genome distances without alignment?



Mapping to the leaves
[Zhu, et al. 2022]



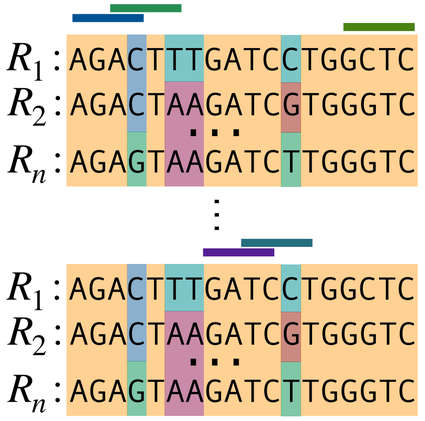
sample

alternative:
focus only on marker genes

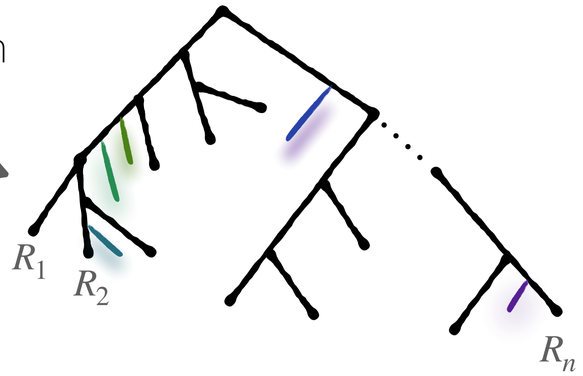
Map to marker genes (e.g., BLAST)

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

align against the reference MSA



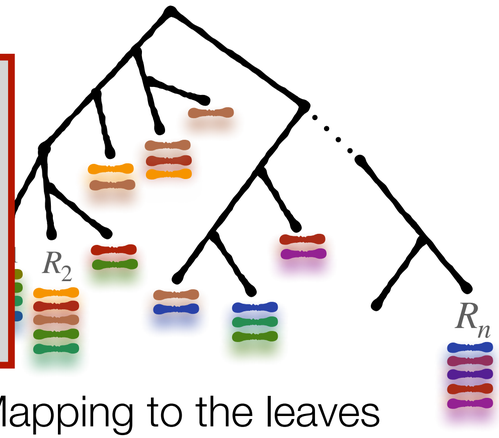
ML placement with pendant branches



Phylogenetic placement has been (mostly) limited to marker genes *or* small phylogenies.

align reads to
reference ge

Can we estimate accurate read to genome distances without alignment?



Mapping to the leaves
[Zhu, et al. 2022]

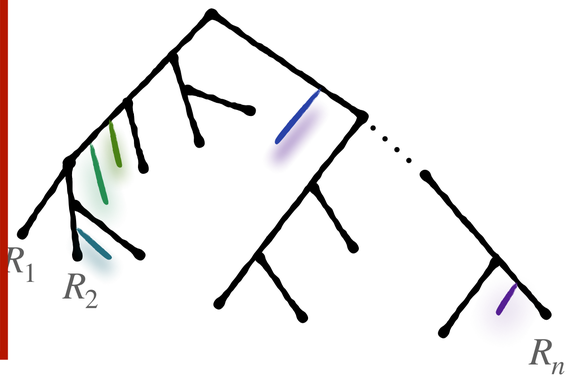
sample

alternative:
focus only on
marker genes

TIPP-(II,III)
[Nguyen et al. 2013, Shah et al. 2021, Shen et al. 2023]

Can we place genome-wide reads on an ultra-large phylogeny?

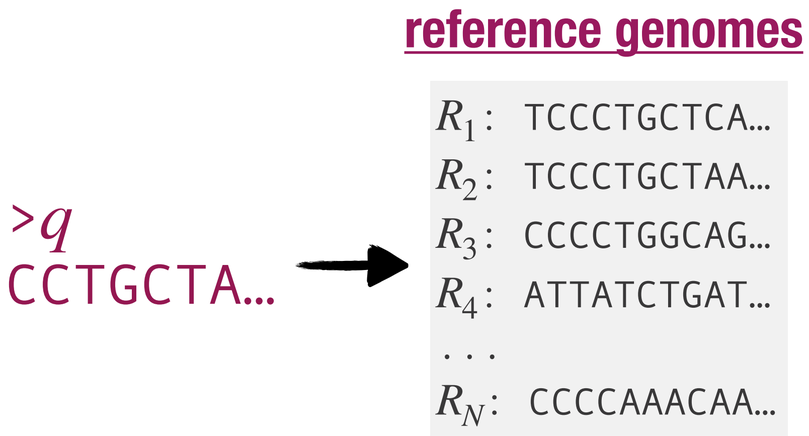
No (scalable) existing method was designed for this!



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:

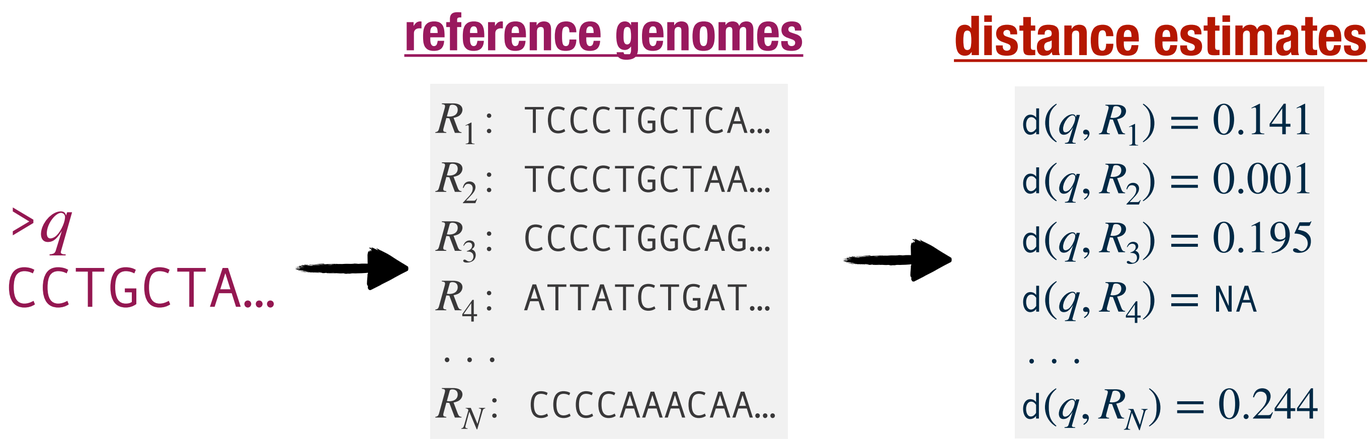
- query sequence q
- set of references $\mathscr{R} = \{R_1, \ldots, R_N\}$
- a backbone phylogeny T

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



- from each read to all sufficiently similar genomes

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

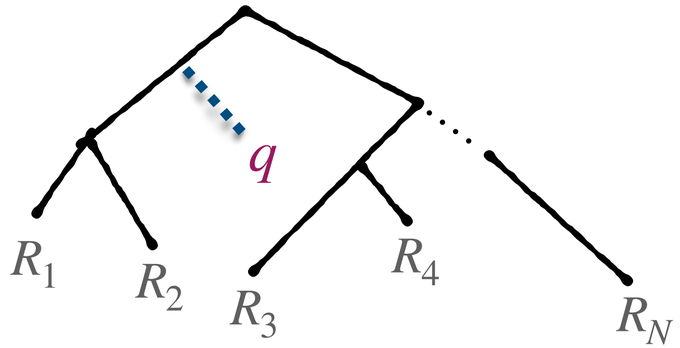
distance estimates

phylogenetic placement:

$>q$
CCTGCTA…

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

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



- from each read to all sufficiently similar genomes
- 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

krepp solves both of these problems for genome-wide reads (distances & PP)

$>q$
CCTGCTA...

reference genomes

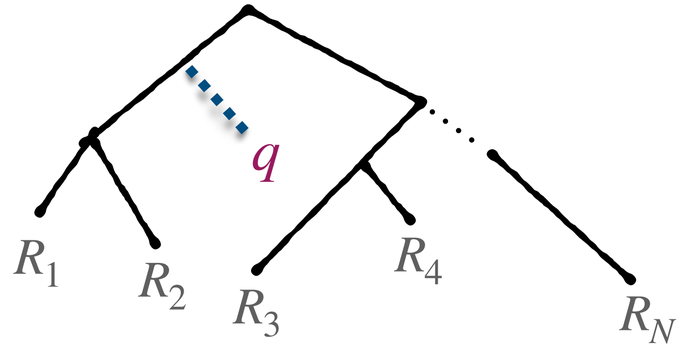
R_1: TCCCTGCTCA...
R_2: TCCCTGCTAA...
R_3: CCCCTGGCAG...
R_4: ATTATCTGAT...
...
R_N: CCCCAAACAA...

distance estimates

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

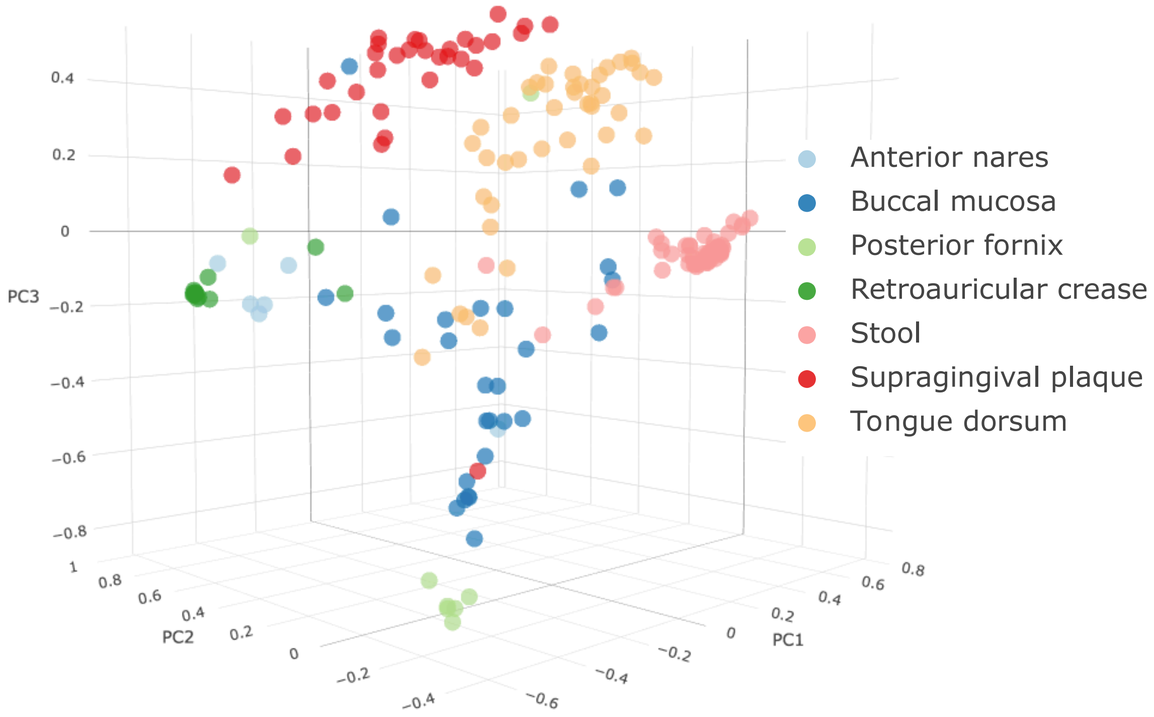
- from each read to all sufficiently similar genomes

phylogenetic placement:



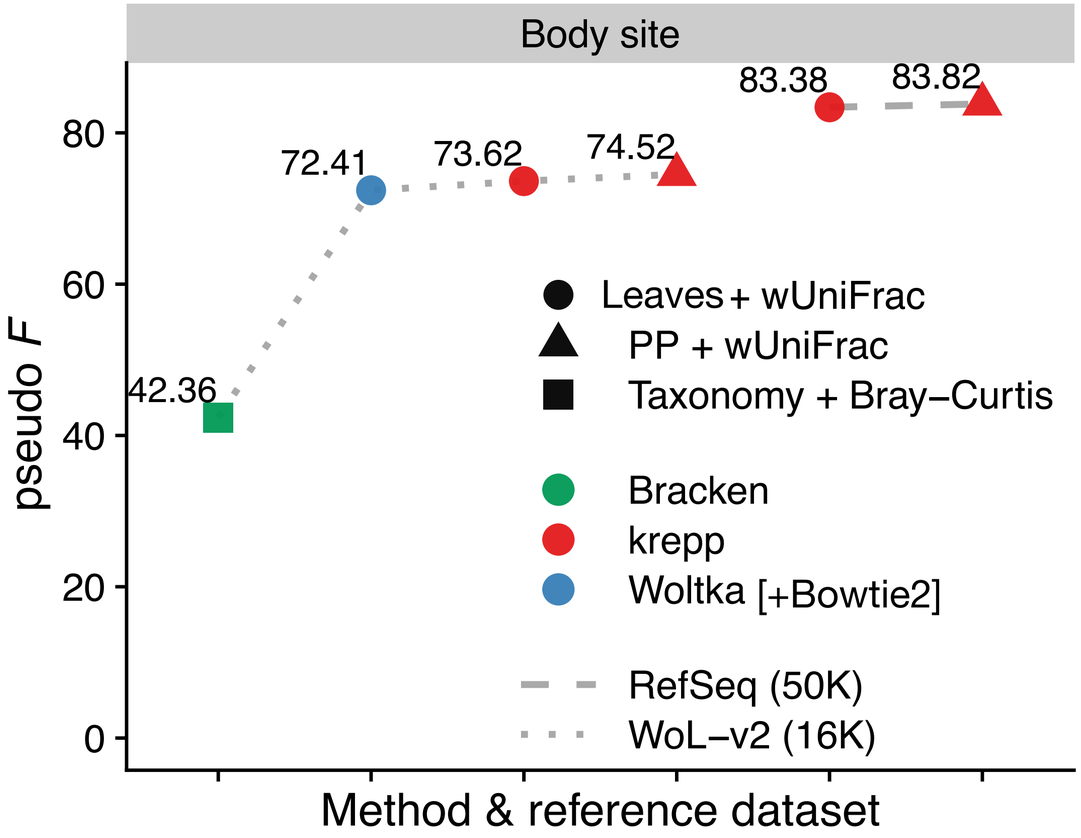
- go beyond markers
- ultra-large phylogenies

krepp can estimate distances & phylogenetically place reads



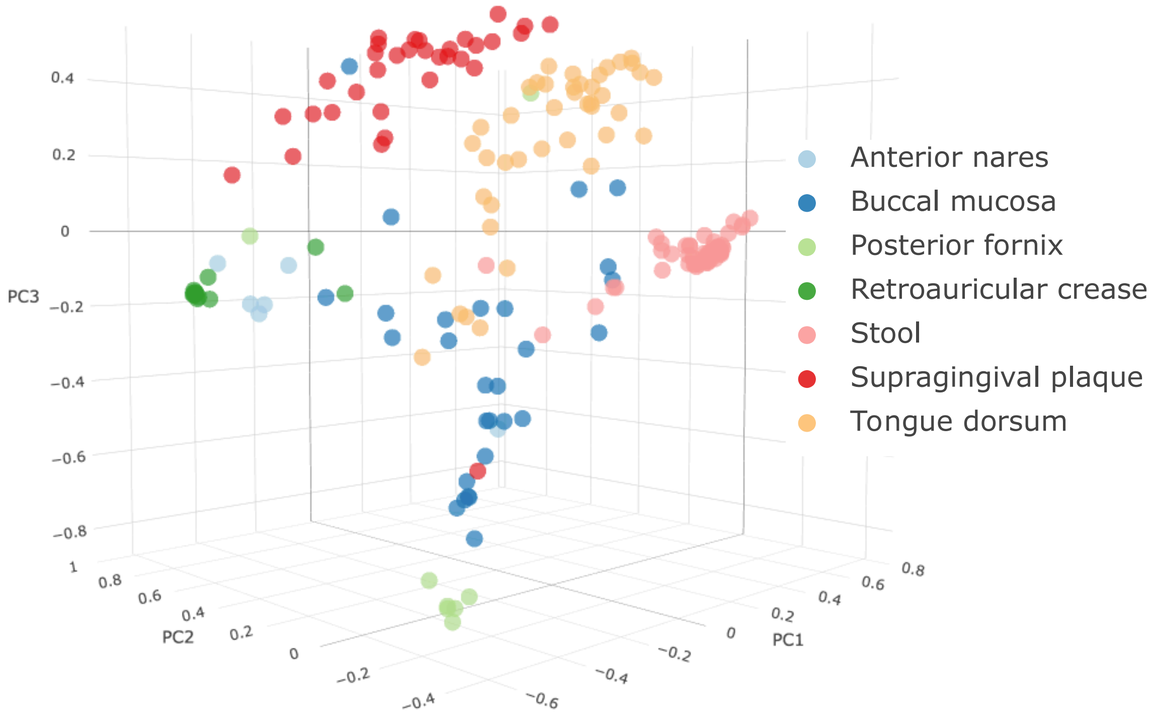
7 body sites, 210 samples

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



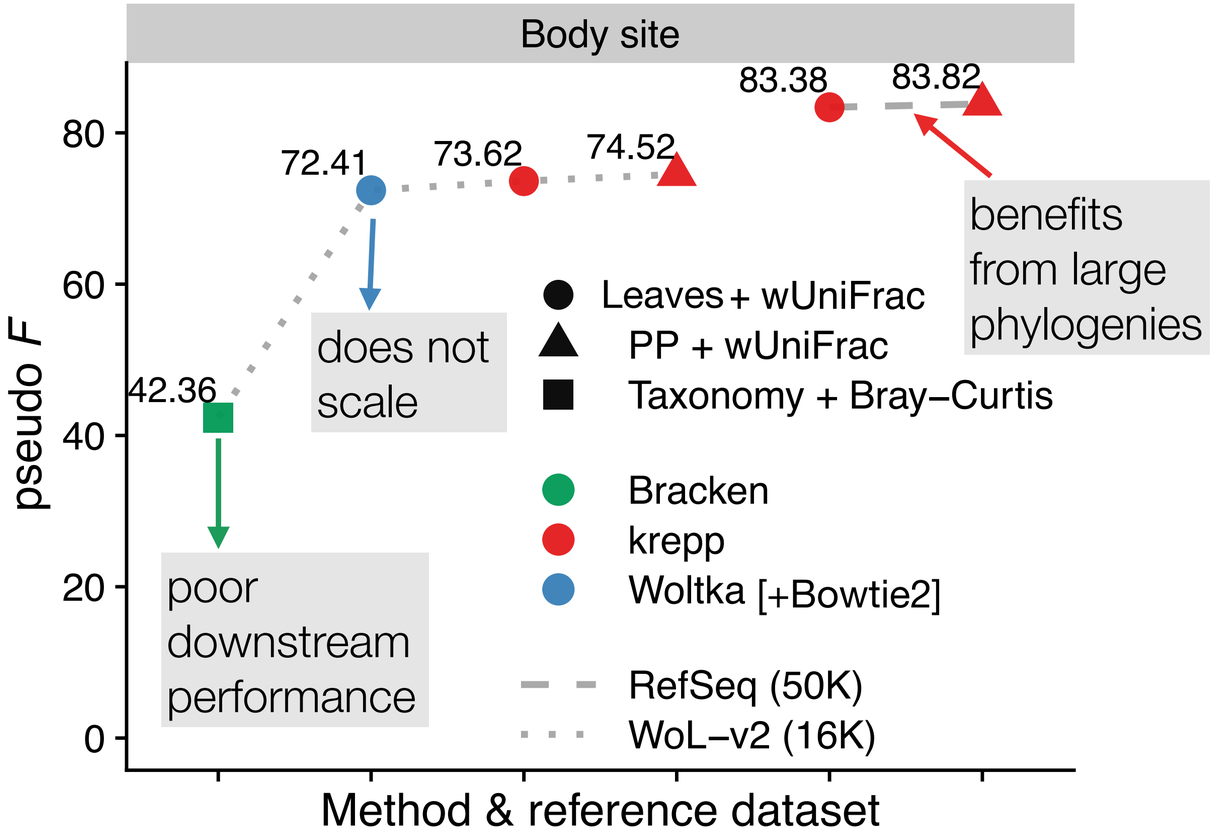
krepp can estimate distances & phylogenetically place reads

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



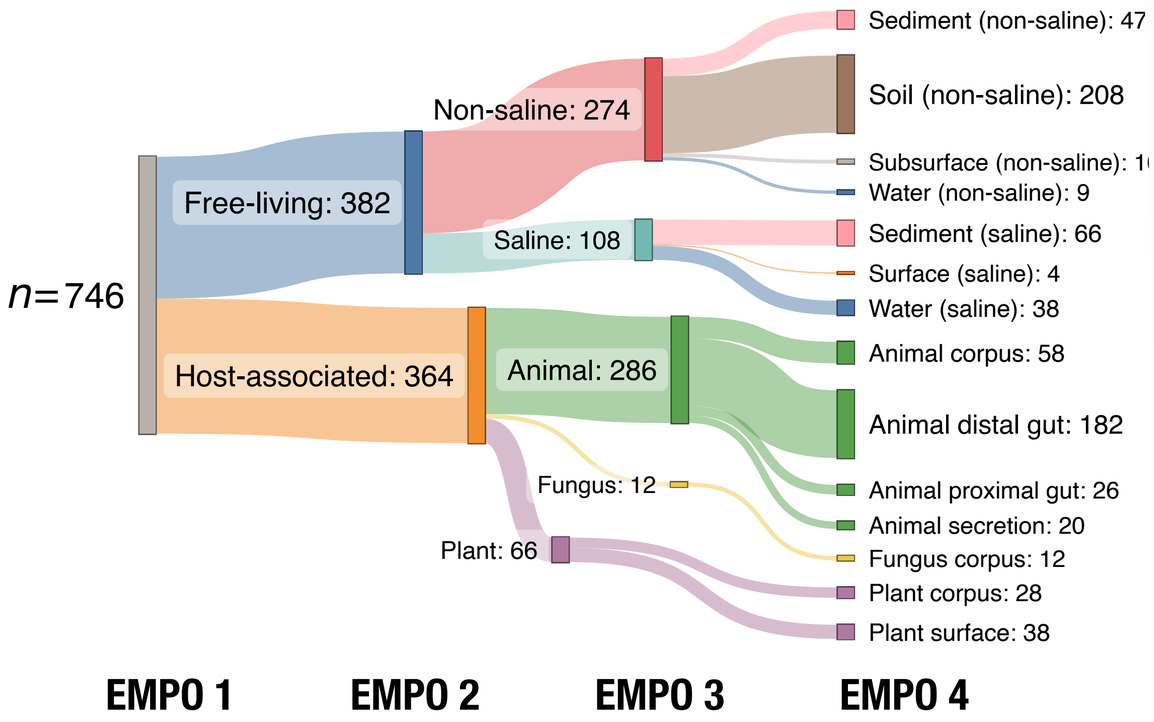
7 body sites, 210 samples

Reference: RefSeq subset
50,000 microbial genomes

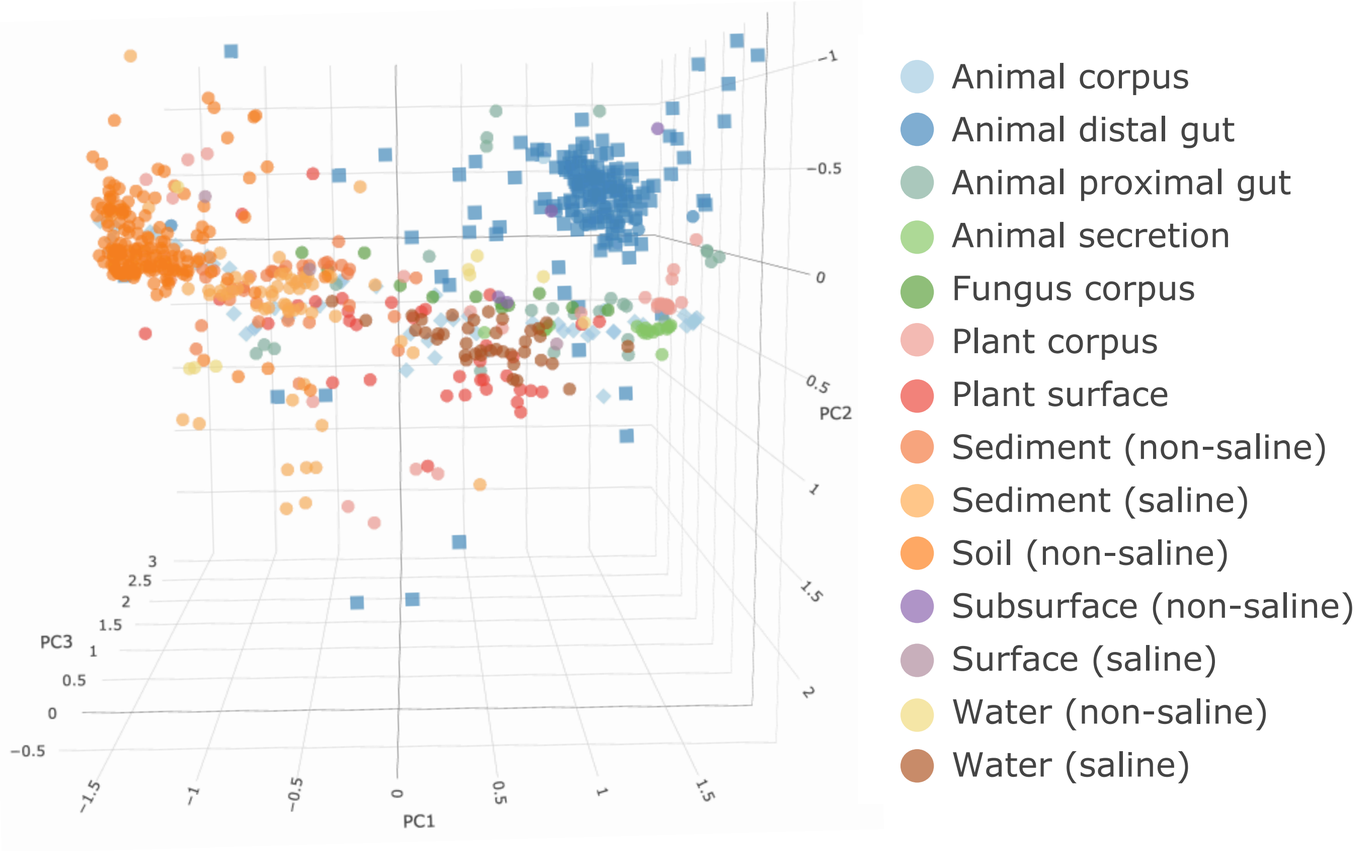


Better characterization of less-studied microbiome of earth

Hierarchical categorization
of earth microbiome samples

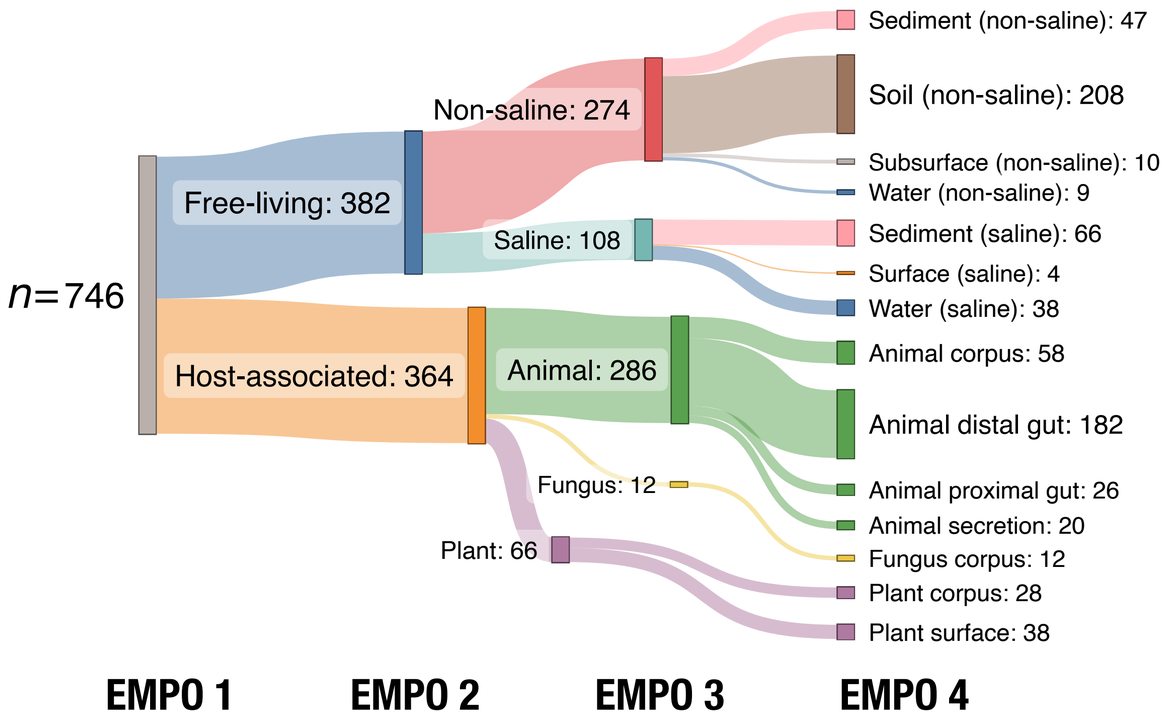


Reference: Web of Life (v1)
11,000 microbial genomes

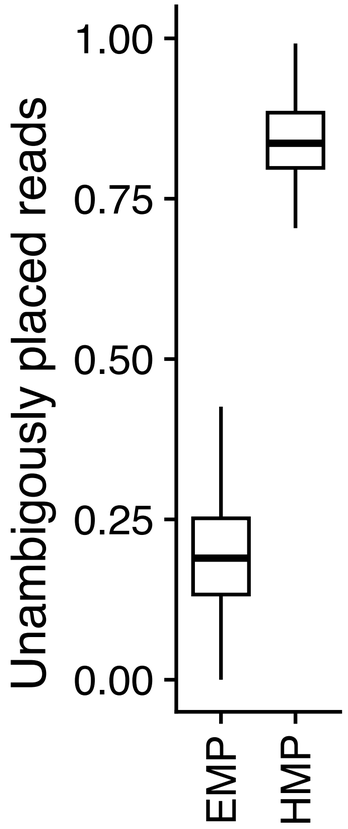


Better characterization of less-studied microbiome of earth

Hierarchical categorization
of earth microbiome samples



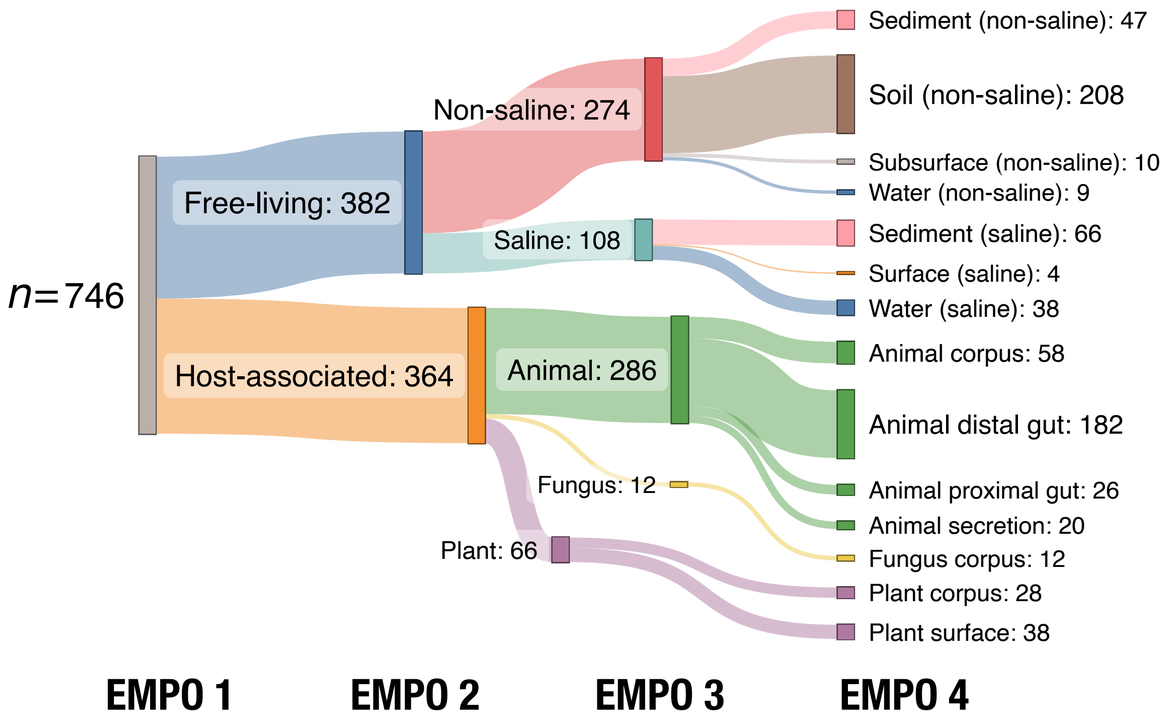
Higher novelty
& uncertainty



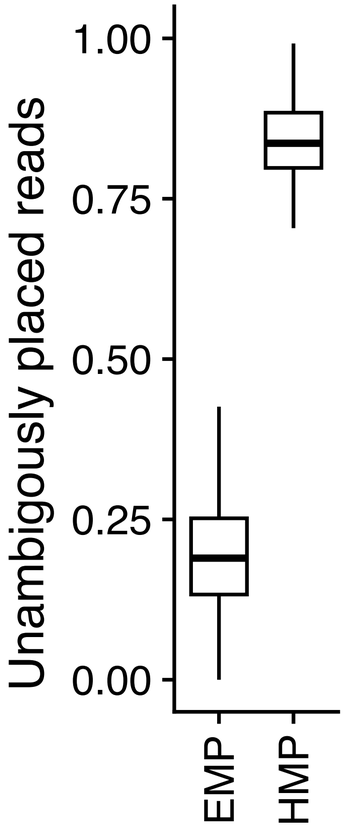
Reference: Web of Life (v1)
11,000 microbial genomes

Better characterization of less-studied microbiome of earth

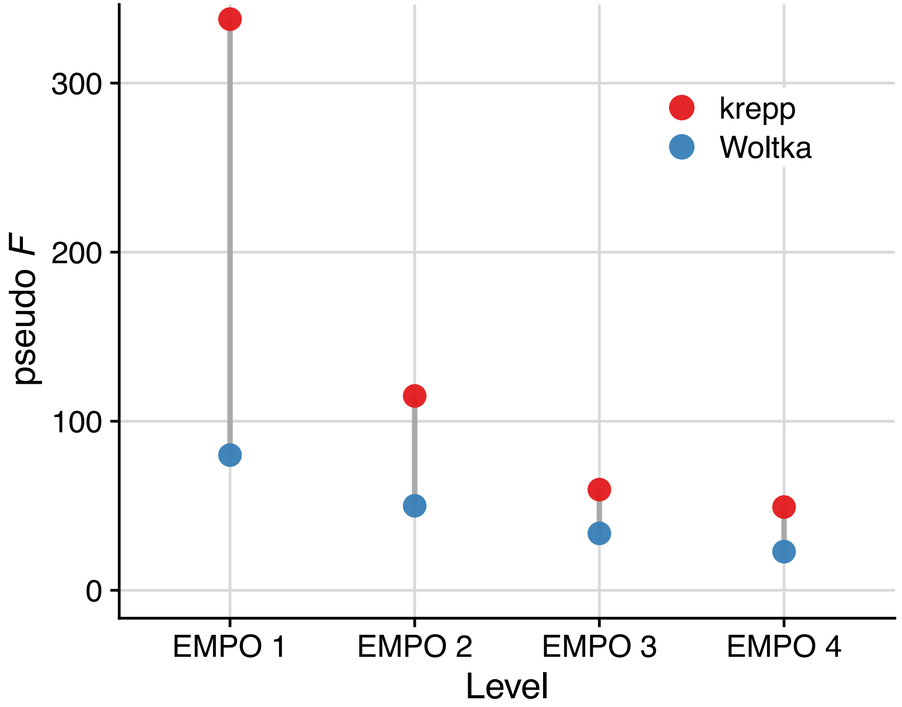
Hierarchical categorization
of earth microbiome samples



Higher novelty
& uncertainty



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

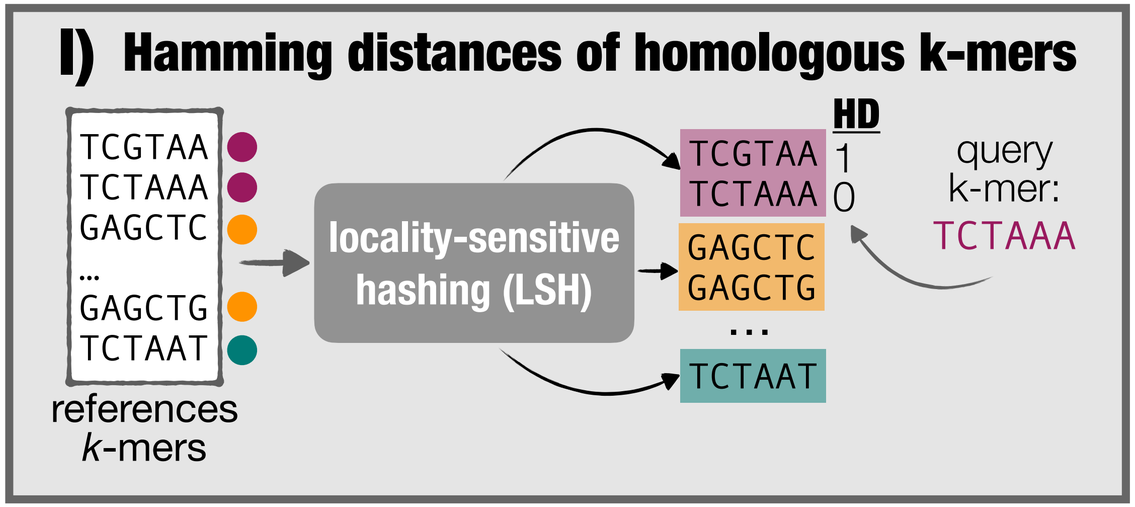


Four different computational (sub)problems

krepp: k-mer-based read phylogenetic placement

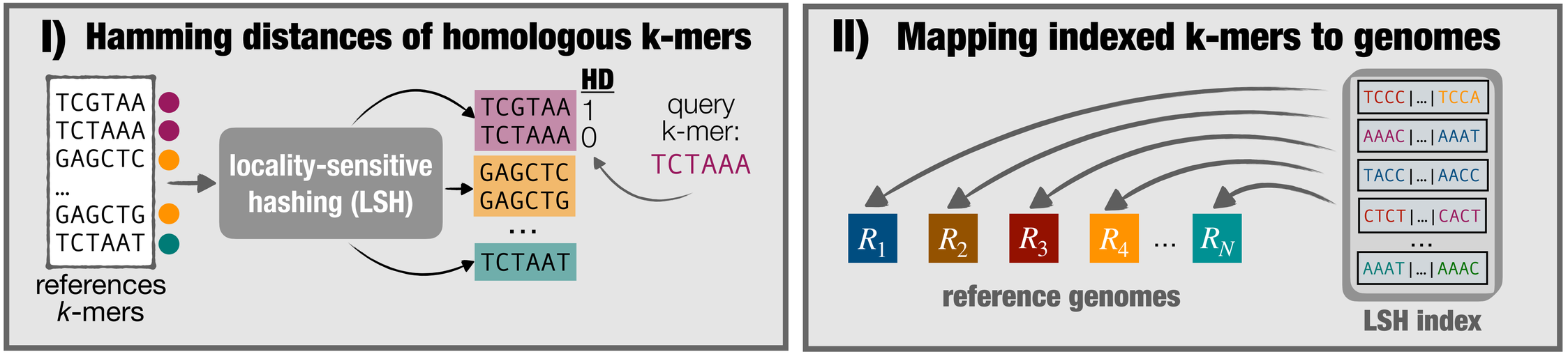
Four different computational (sub)problems

krepp: k-mer-based read phylogenetic placement



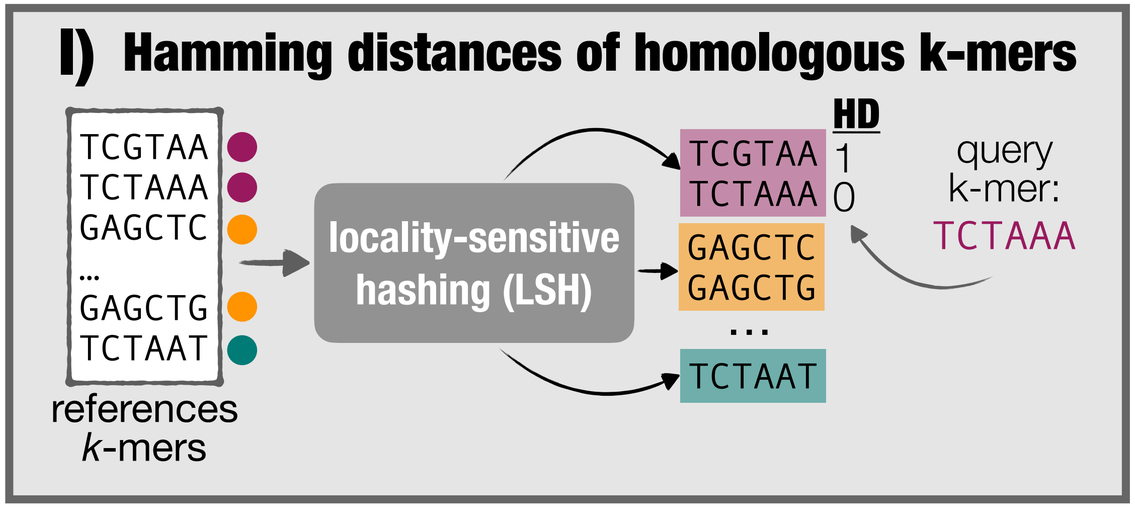
Four different computational (sub)problems

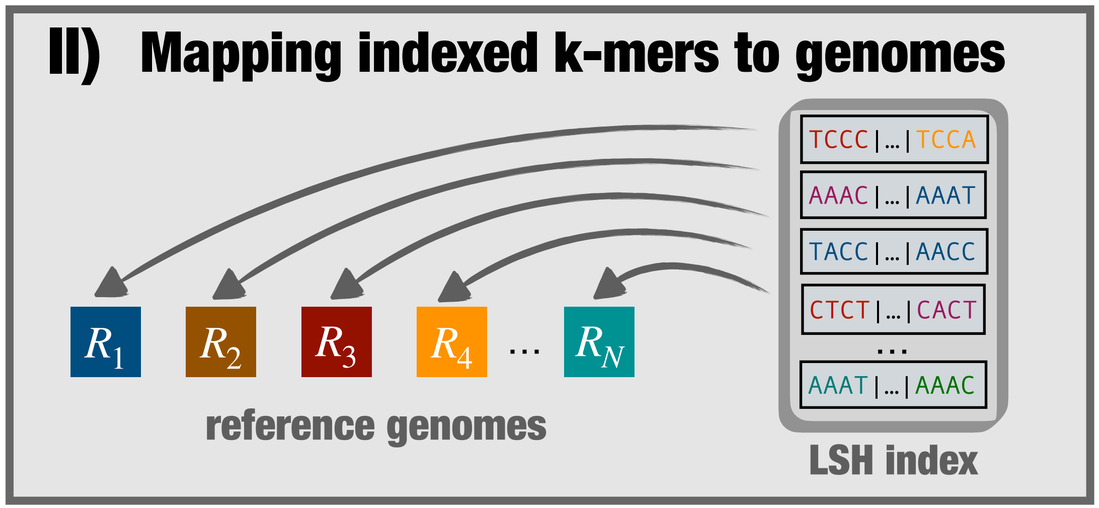
krepp: k-mer-based read phylogenetic placement

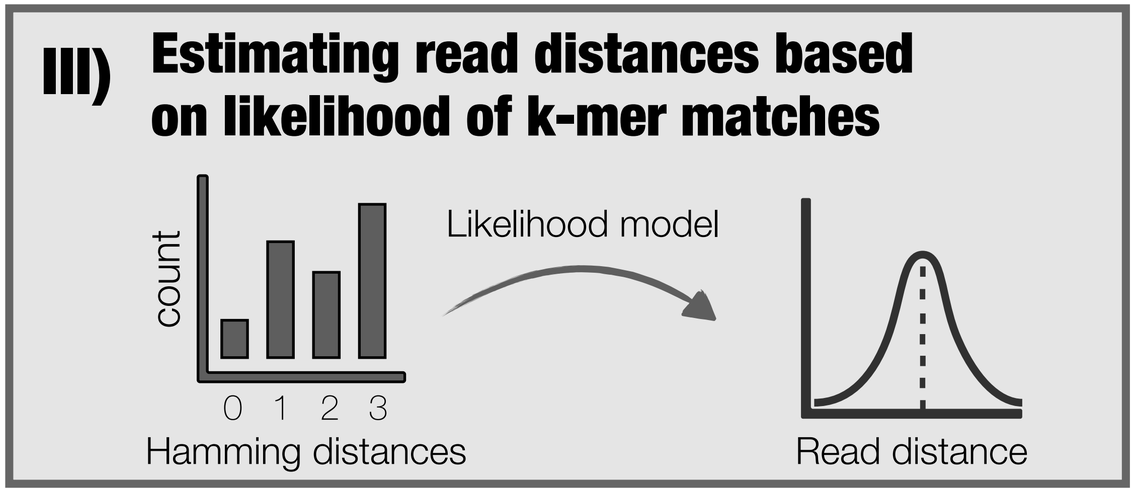


Four different computational (sub)problems

krepp: k-mer-based read phylogenetic placement

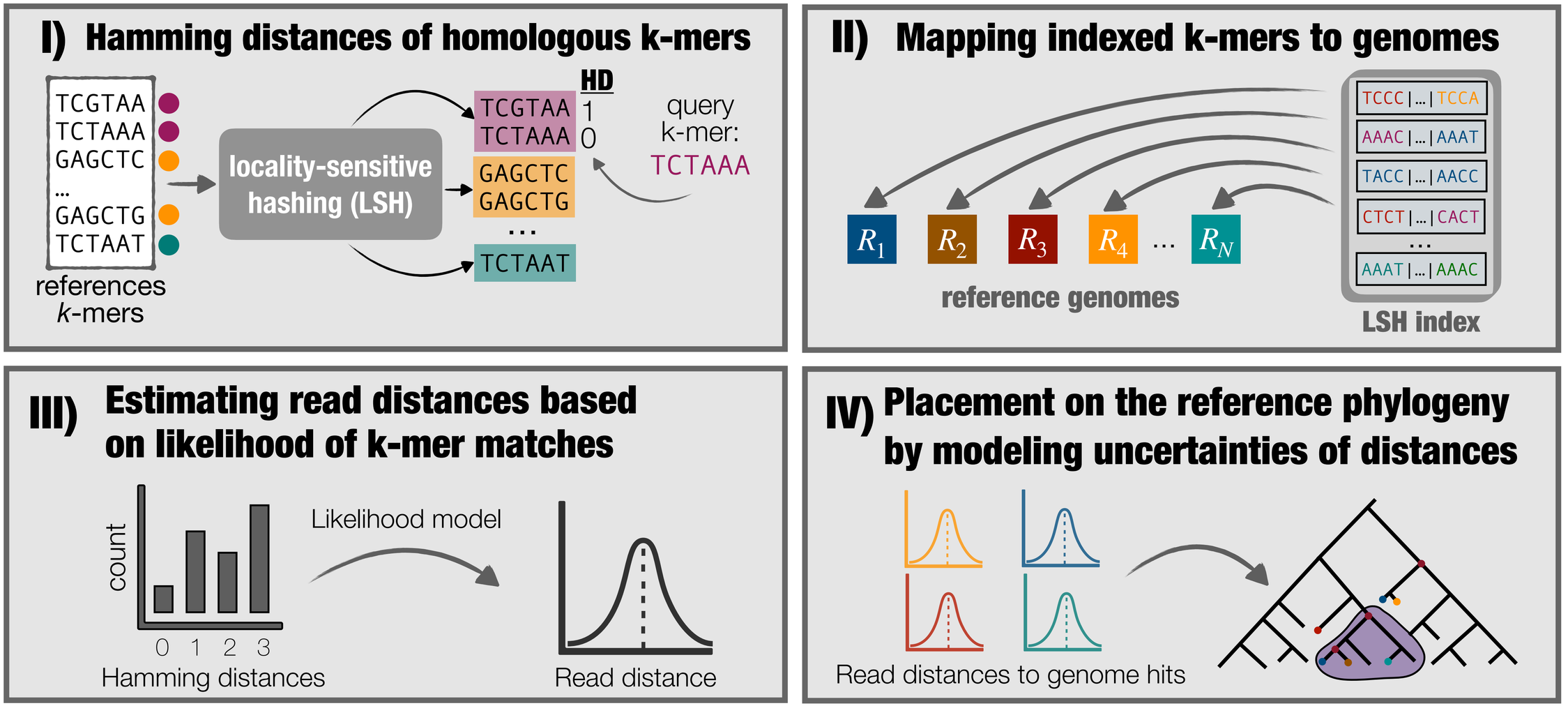






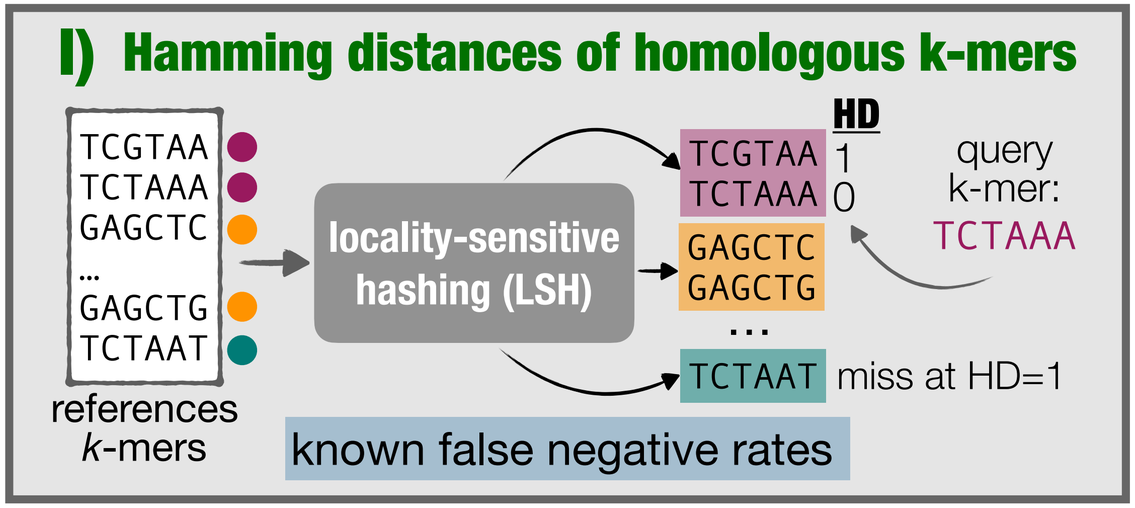
Four different computational (sub)problems

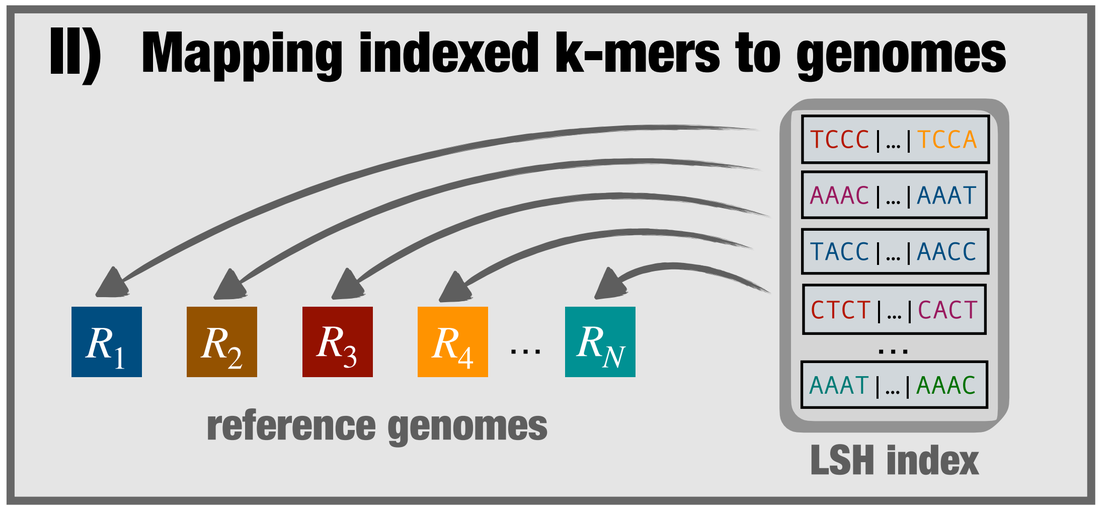
krepp: k-mer-based read phylogenetic placement

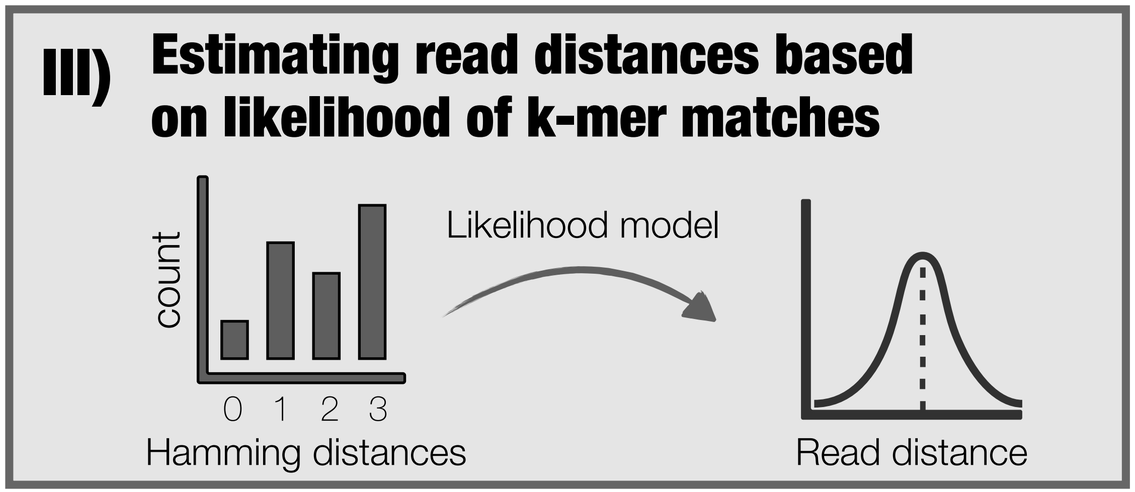


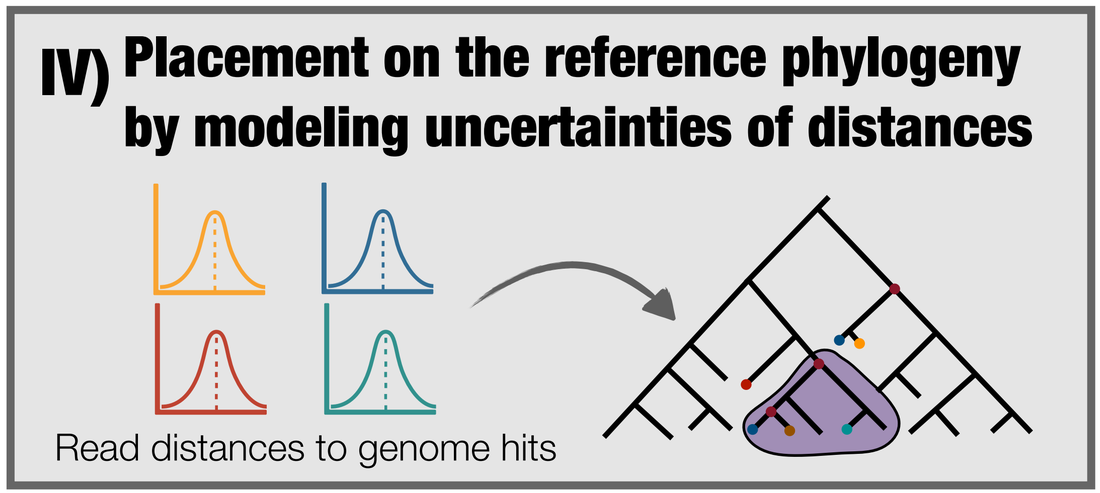
Four different computational (sub)problems

krepp: k-mer-based read phylogenetic placement

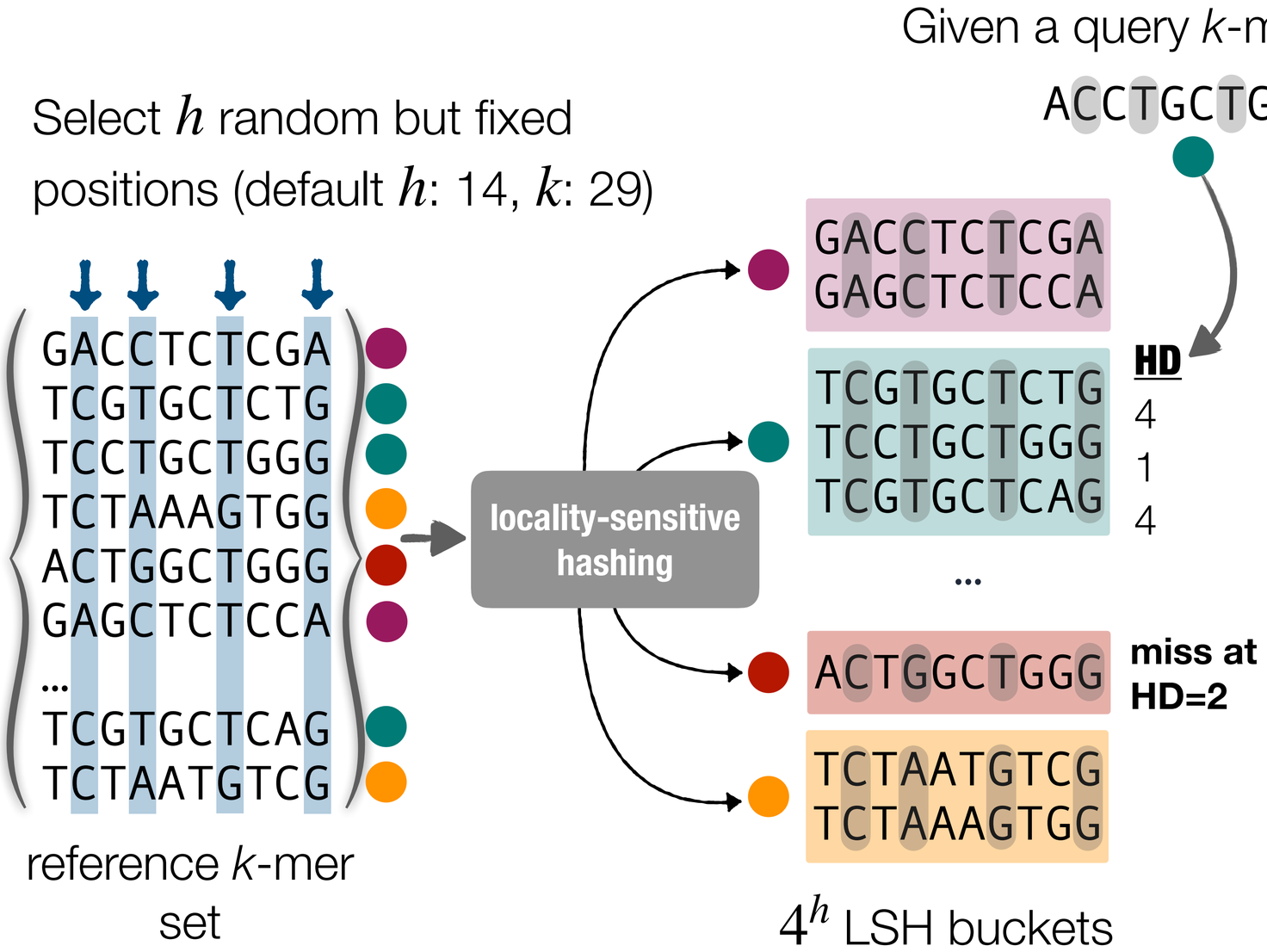








Problem I: computing false negative rates of LSH

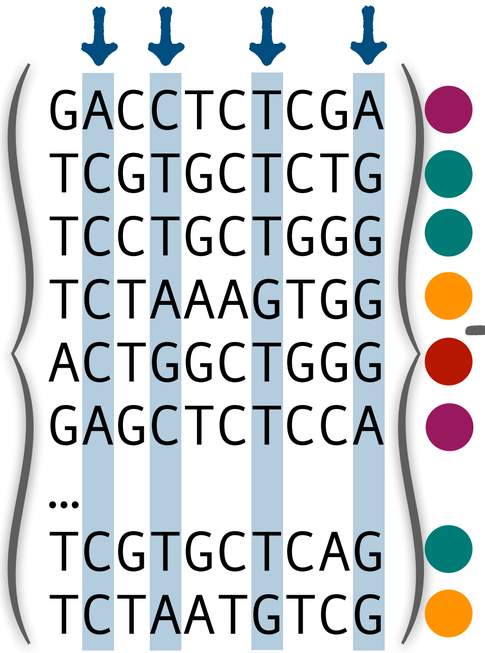


Problem I: computing false negative rates of LSH

Given a query *k*-mer

ACCTGCTGGG

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



reference *k*-mer set

locality-sensitive hashing

GACCTCTCGA
GAGCTCTCCA

TCGTGCTCTG
TCCTGCTGGG
TCGTGCTCAG

…

ACTGGCTGGG

TCTAATGTCG
TCTAAAGTGG

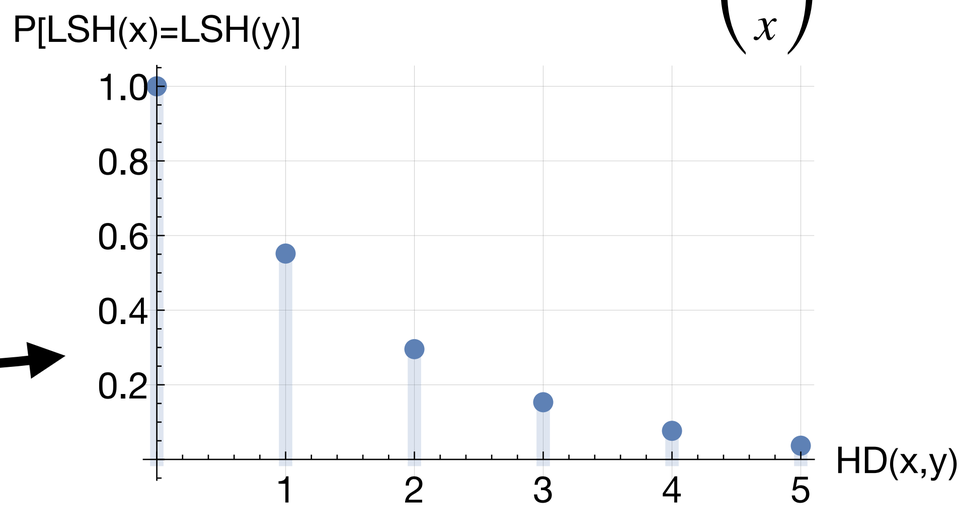
4^h LSH buckets

HD
4
1
4

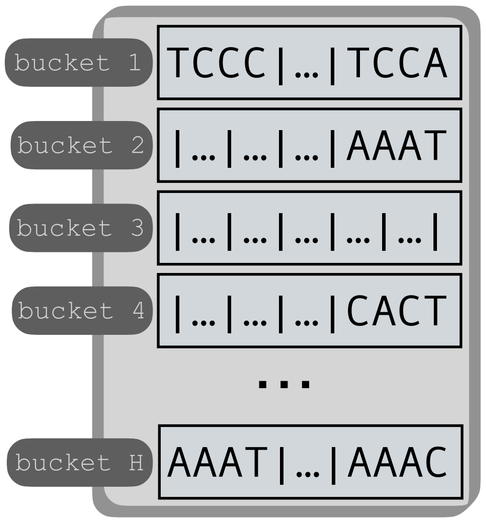
miss at HD=2

False negative rate:

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



Problem II: mapping indexed k-mers to reference genomes



LSH index

R_1

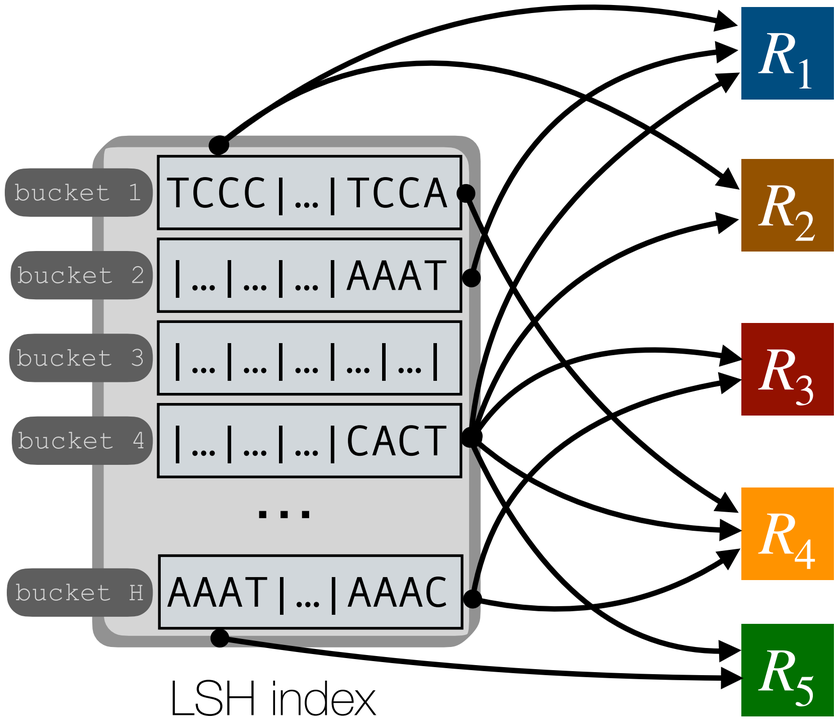
R_2

R_3

R_4

R_5

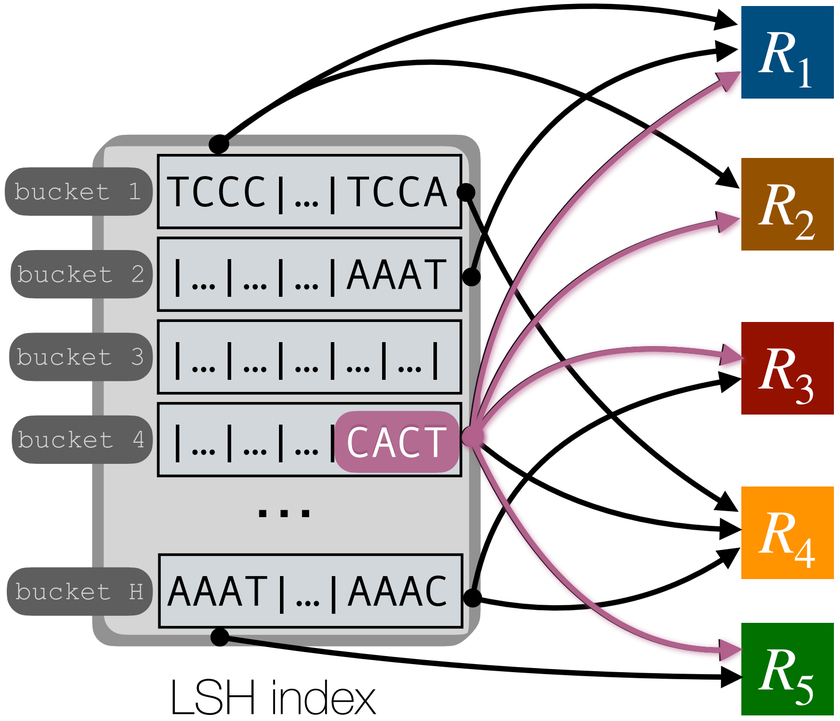
Problem II: mapping indexed k-mers to reference genomes



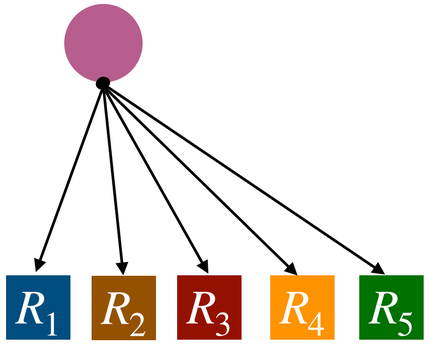
well studied **colored *k*-mer** problem

color: a subset of references
(including singletons)

Problem II: mapping indexed k-mers to reference genomes



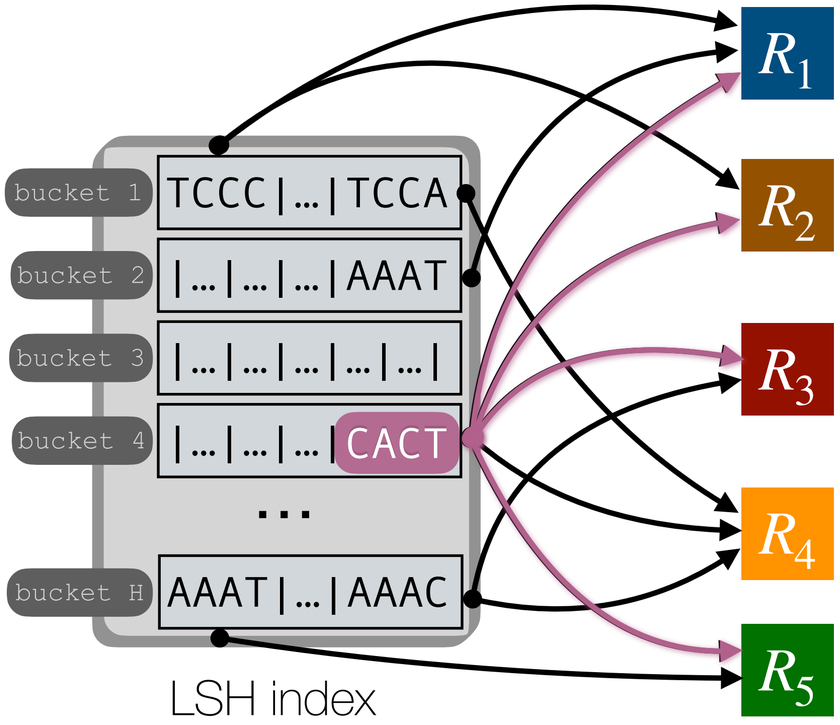
LSH index



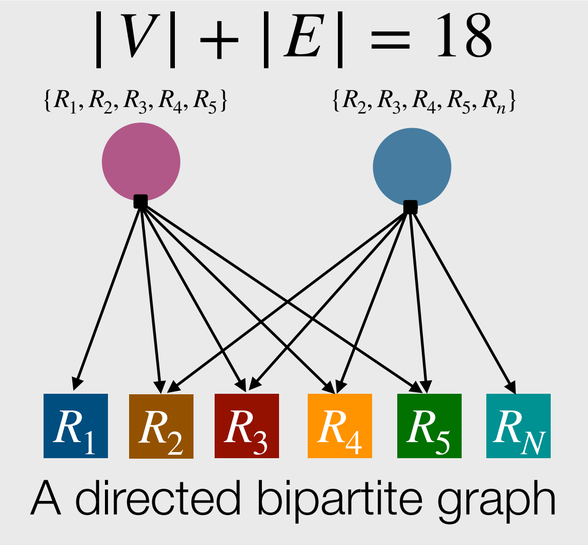
well studied **colored *k*-mer** problem

color: a subset of references
(including singletons)

Problem II: mapping indexed k-mers to reference genomes



LSH index



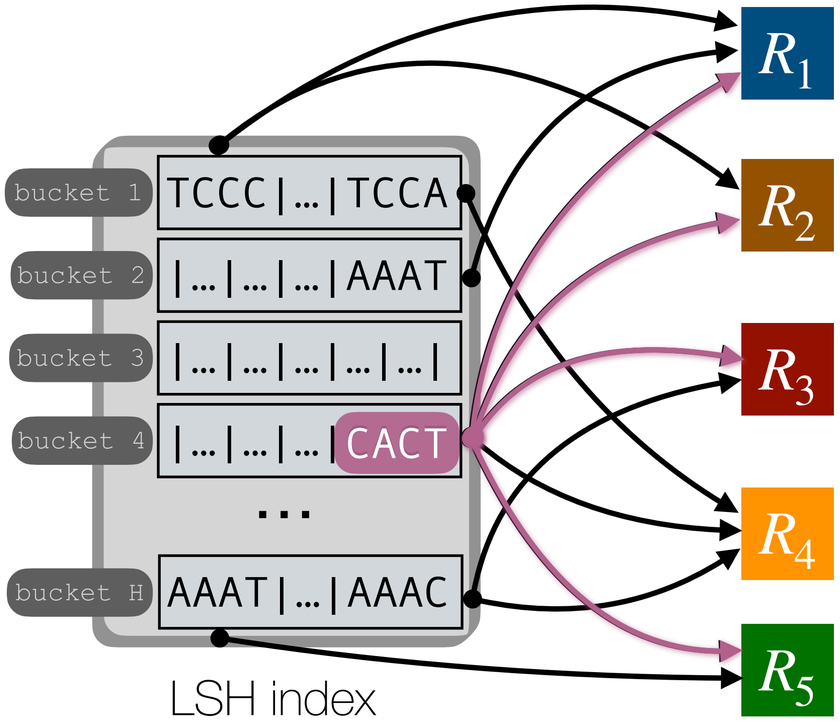
A directed bipartite graph

represent each non-singleton as a union of others

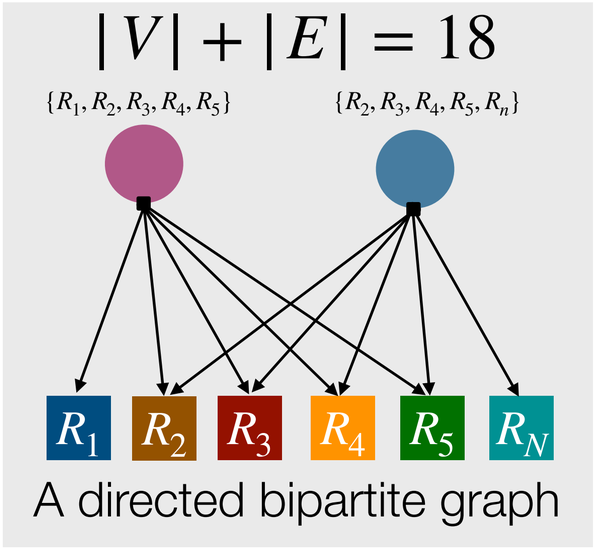
well studied **colored *k*-mer** problem

color: a subset of references (including singletons)

Problem II: mapping indexed k-mers to reference genomes



LSH index



A directed bipartite graph

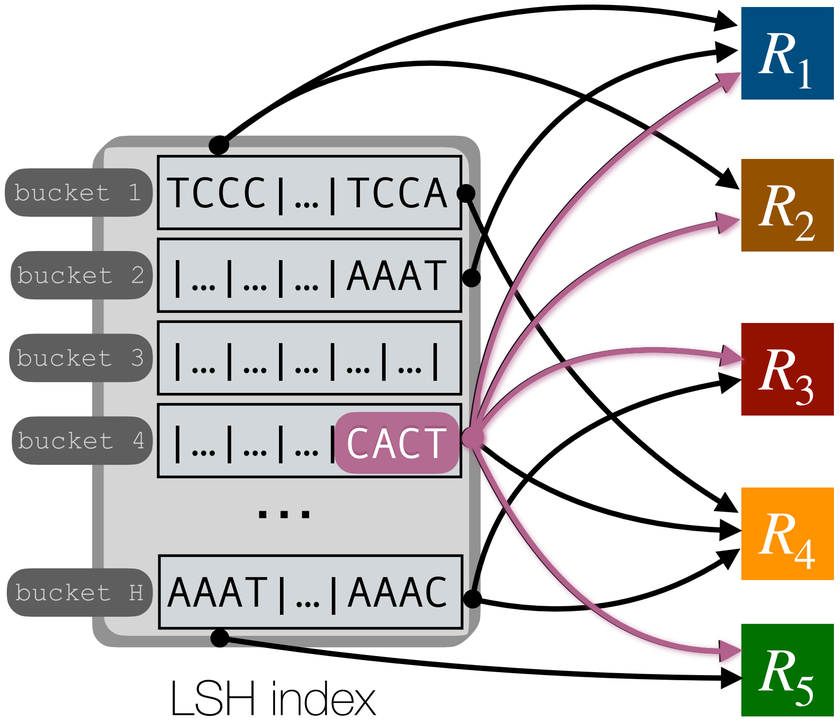
represent each non-singleton as a union of others

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

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

color: a subset of references (including singletons)

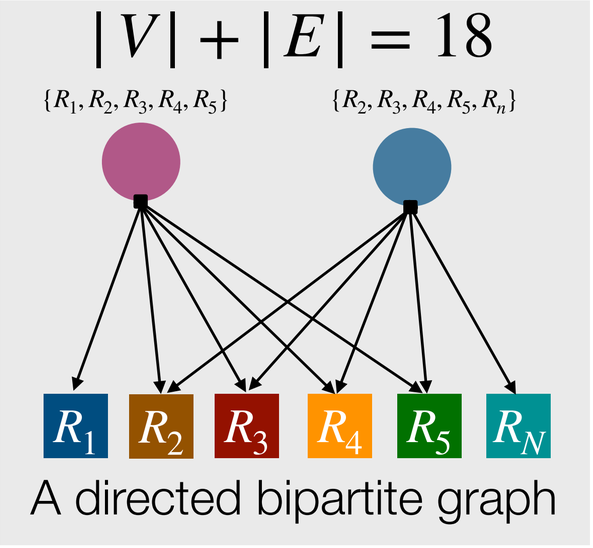
Problem II: mapping indexed k-mers to reference genomes



LSH index

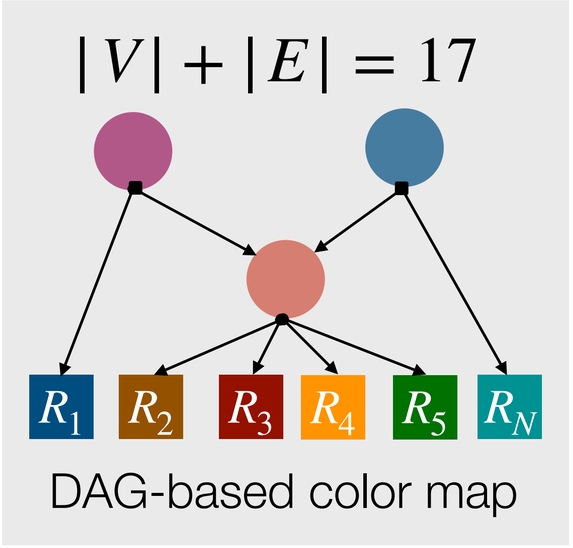
well studied **colored *k*-mer** problem

color: a subset of references (including singletons)



A directed bipartite graph

represent each non-singleton as a union of others



DAG-based color map

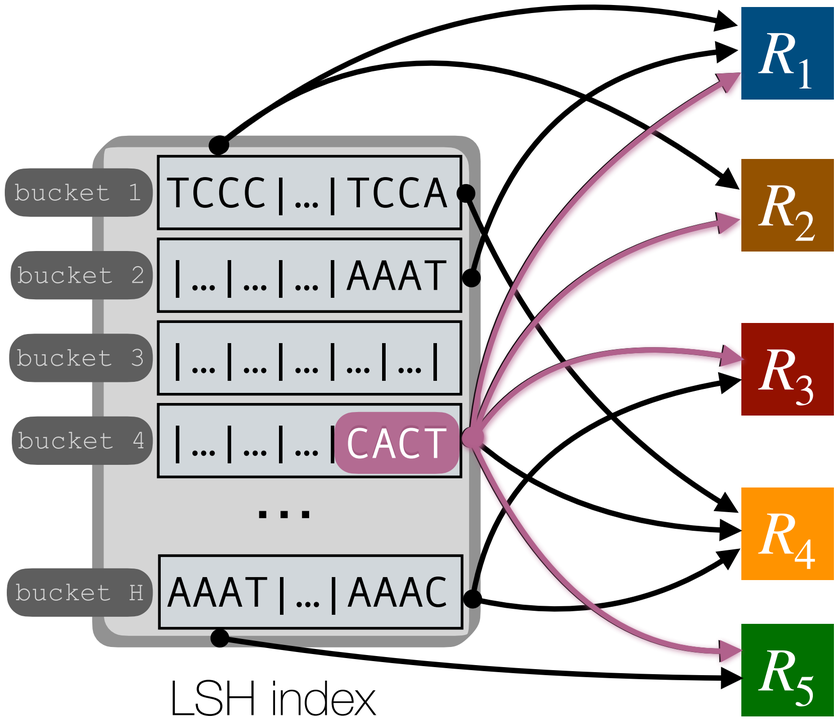
Minimize $|V| + |E|$?

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

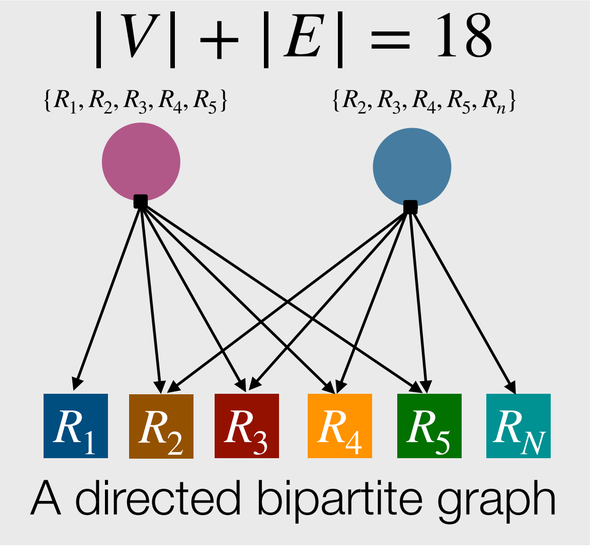
Problem II: mapping indexed k-mers to reference genomes



LSH index

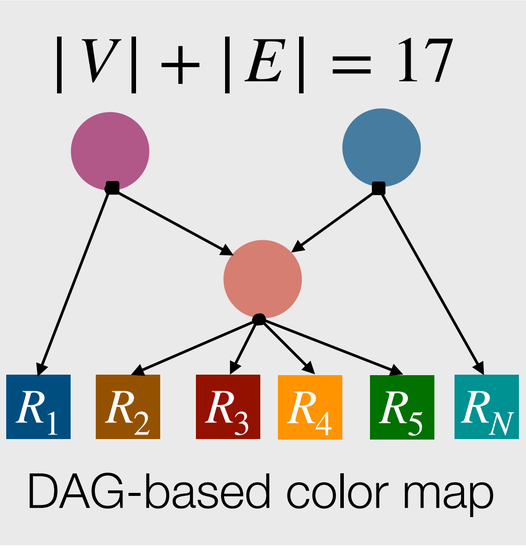
well studied **colored *k*-mer** problem

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*.

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
```

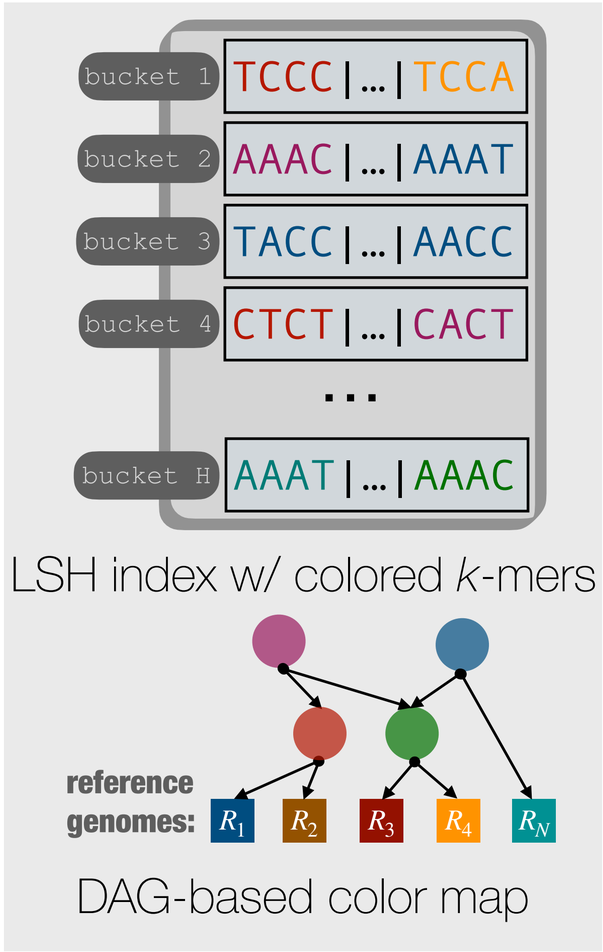
Finding homologous k-mers of reference genomes

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



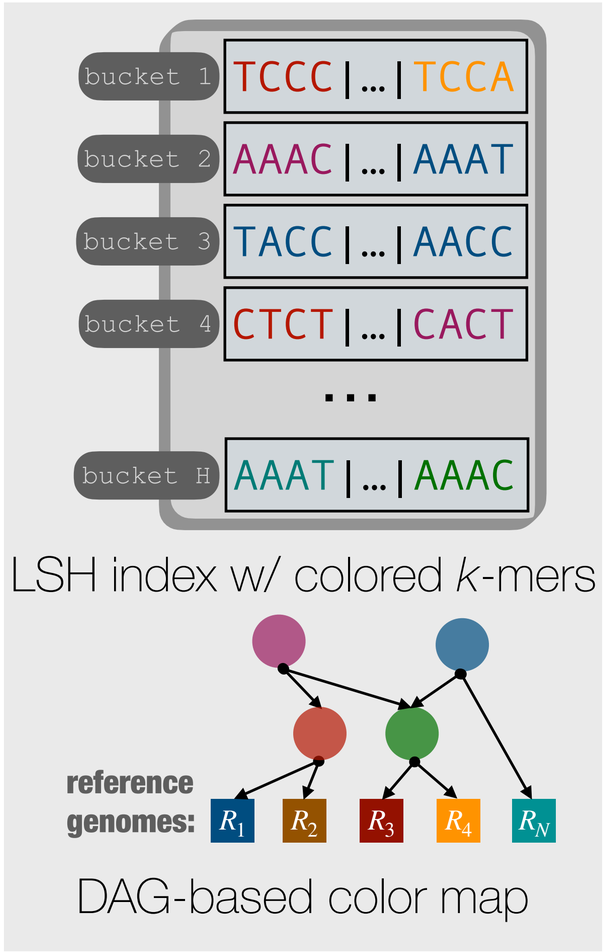
Finding homologous k-mers of reference genomes

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

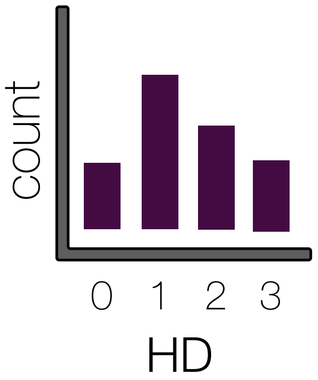
Estimating a distance from the homologous k-mers

	HD
1	0
2	1
3	-
4	-
5	0
…	…
L-k+1	3

Estimating a distance from the homologous k-mers

Summarize as a histogram:

	HD
1	0
2	1
3	-
4	-
5	0
...	...
L - k+1	3



matches: $\mathbf{v} = [v_0, v_1, \ldots, v_\delta]$

misses: $u = (j - i) - \sum_{d=0}^{\delta} v_d$

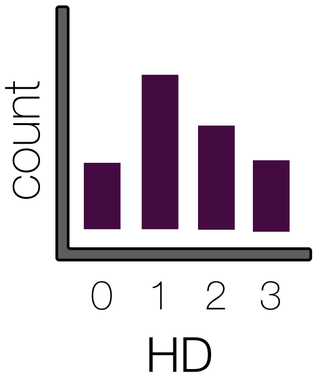
Independence assumption:

treat $Q_{i:j}$ as a bag of $j - i$ k-mers

Estimating a distance from the homologous k-mers

Summarize as a histogram:

	HD
1	0
2	1
3	-
4	-
5	0
...	...
L - k+1	3



matches: $\mathbf{v} = [v_0, v_1, \ldots, v_\delta]$

misses: $u = (j - i) - \sum_{d=0}^{\delta} v_d$

Independence assumption:

treat $Q_{i:j}$ as a bag of $j - i$ *k*-mers

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

$$\mathscr{L}(D; k, h, \delta, u, \mathbf{v}) = P_{miss}(D; k, h, \delta)^u \prod_{d=0}^{\delta} P_{match}(D; d, k, h)^{v_d}$$

Probability of having u missses in total

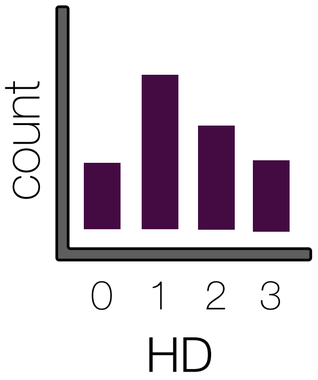
Probability of having v_d matches at HD $= d$

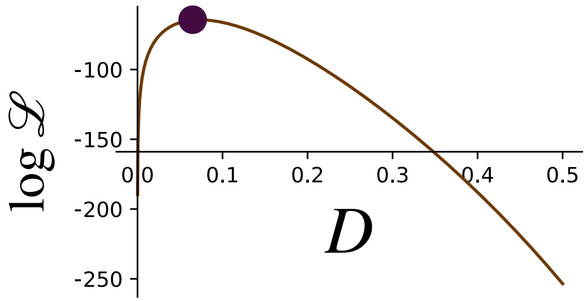
Estimating a distance from the homologous k-mers

Compute the likelihood of Q having distance D to R

Summarize as a histogram:

	HD
1	0
2	1
3	-
4	-
5	0
...	...
L - k+1	3





matches: $\mathbf{v} = [v_0, v_1, \ldots, v_\delta]$

misses: $u = (j - i) - \sum_{d=0}^{\delta} v_d$

Likelihood of distance D to reference R

: a product over all *k*-mers

$$\mathscr{L}(D; k, h, \delta, u, \mathbf{v}) = P_{miss}(D; k, h, \delta)^u \prod_{d=0}^{\delta} P_{match}(D; d, k, h)^{v_d}$$

Probability of having u missses in total

Probability of having v_d matches at HD $= d$

Independence assumption:

treat $Q_{i:j}$ as a bag of $j - i$ *k*-mers

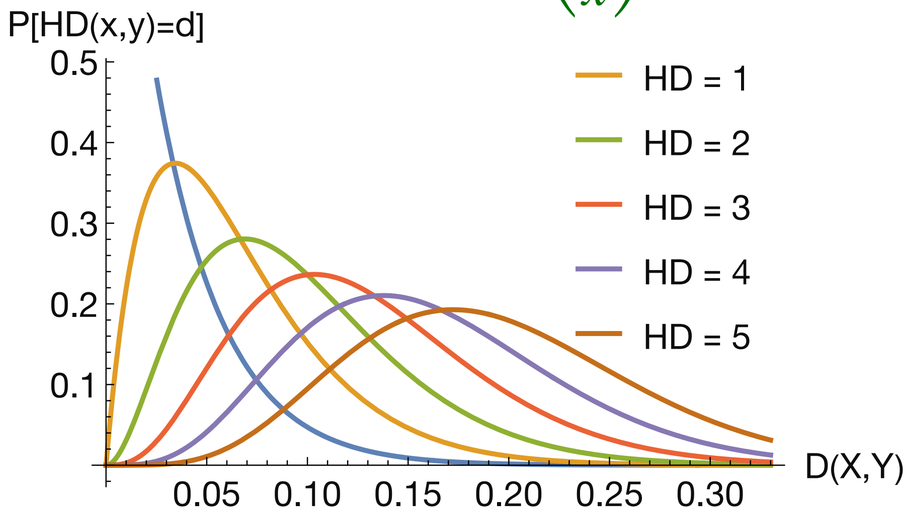
Observing k-mers matches with varying HDs

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

Observing k-mers matches with varying HDs

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

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

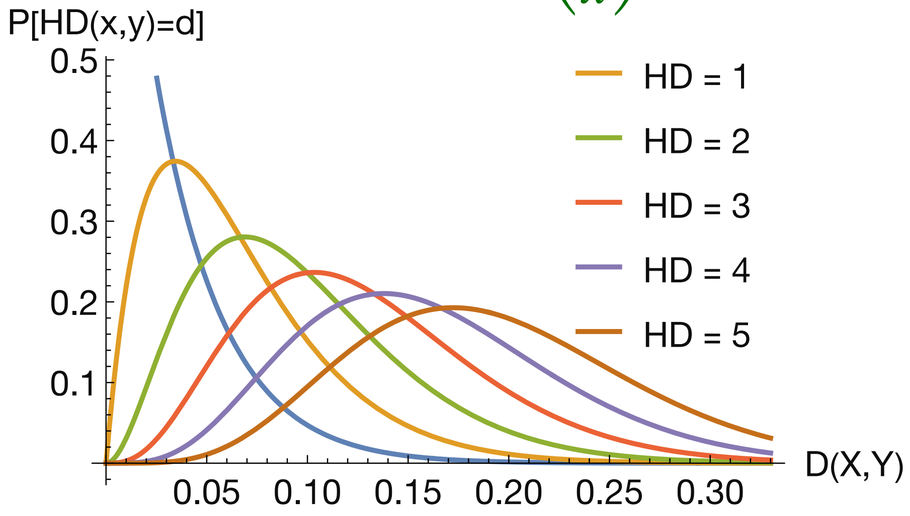


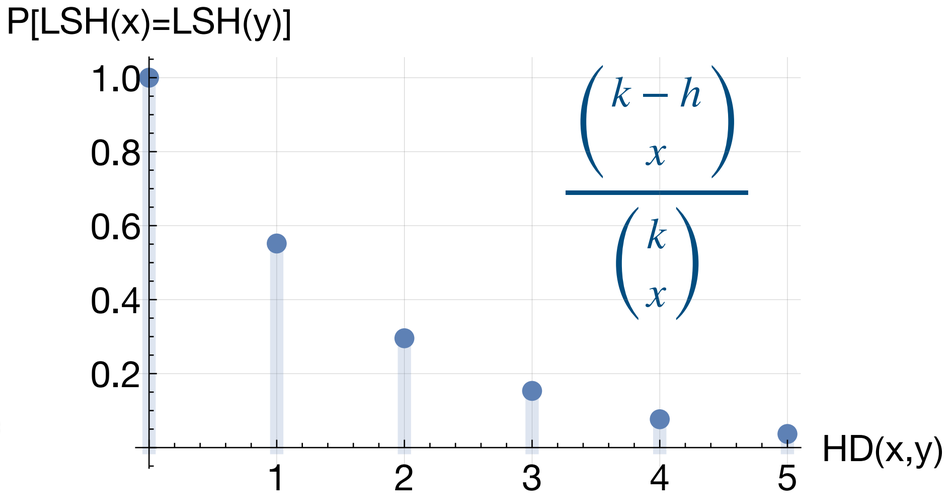
Observing k-mers matches with varying HDs

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

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

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





Observing k-mers matches with varying HDs

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

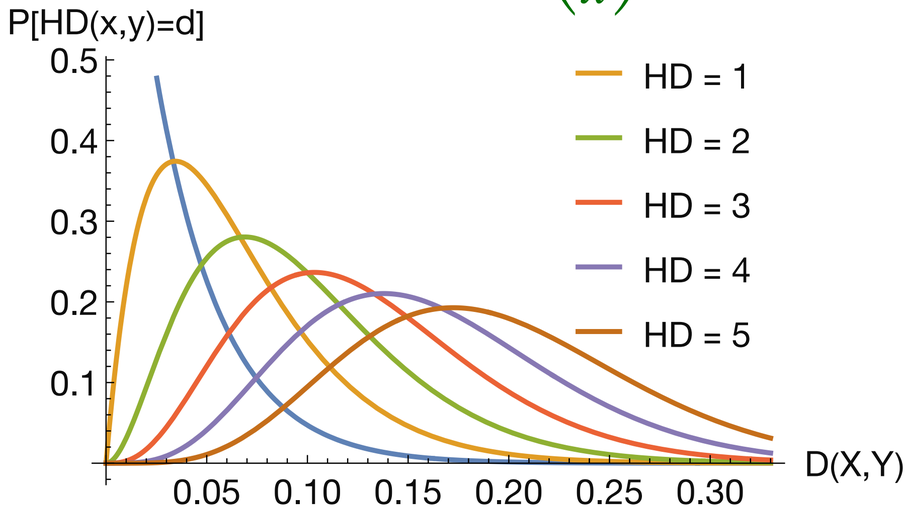
$$\rho_i = \frac{\text{\# of subsampled}}{\text{\# of distinct}}$$

precomputed for R_i

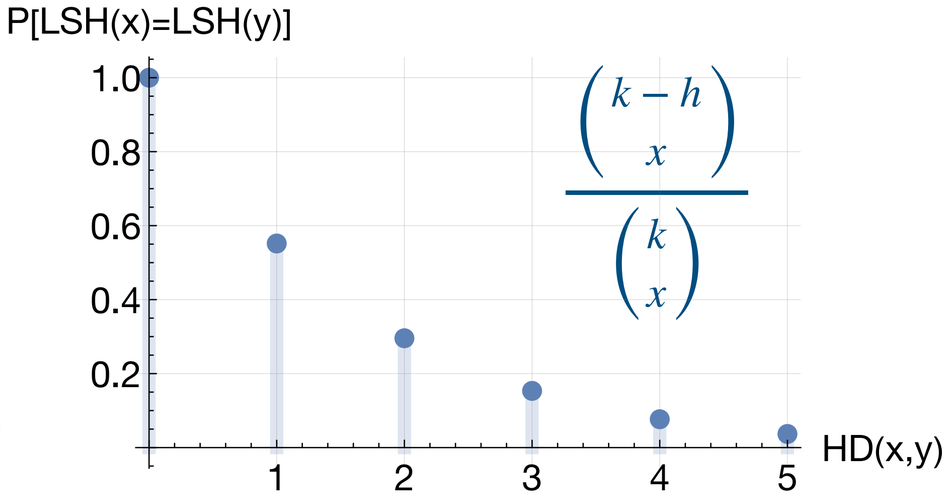
→ not all *k*-mers have to be indexed:

- 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:

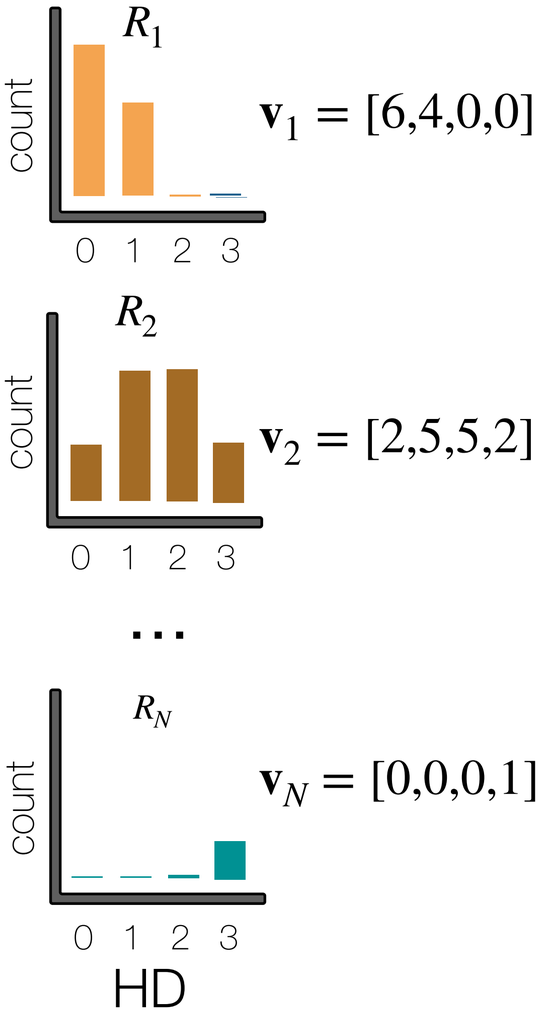
i) $\mathrm{HD}(a, b) > \delta$: $\sum_{x=\delta+1}^{k} P_{mutate}(D; x, k)$ ***or***

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

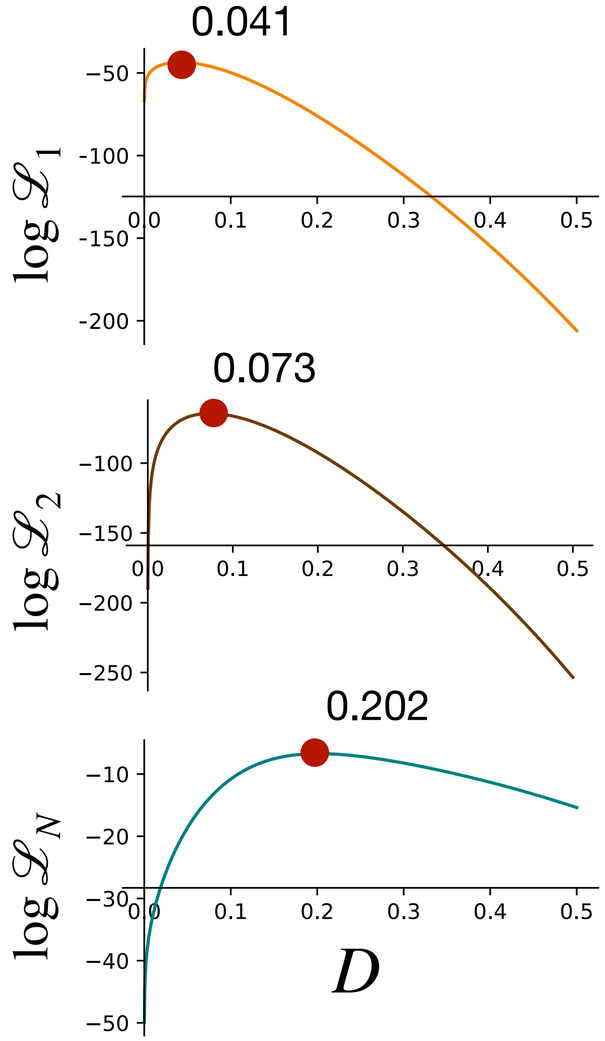


Optimize $-\log \mathscr{L}_i$ w.r.t. D

for each hitting reference R_i:

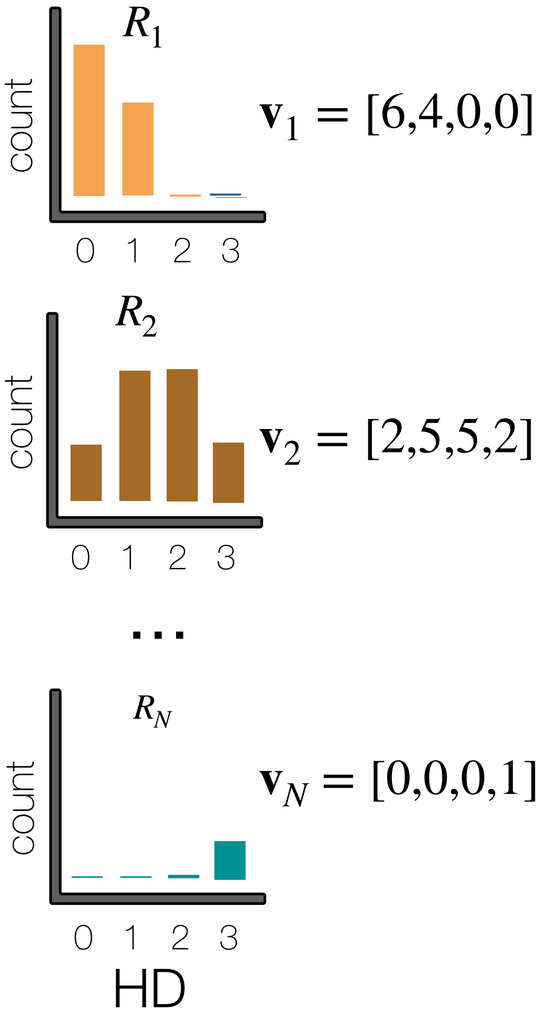
$$\arg\max_D \; P_{miss}(D;k,h,\delta)^{u_i} \prod_{x=0}^{\delta} P_{match}(D;x,k,h)^{v_{i,x}}$$

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

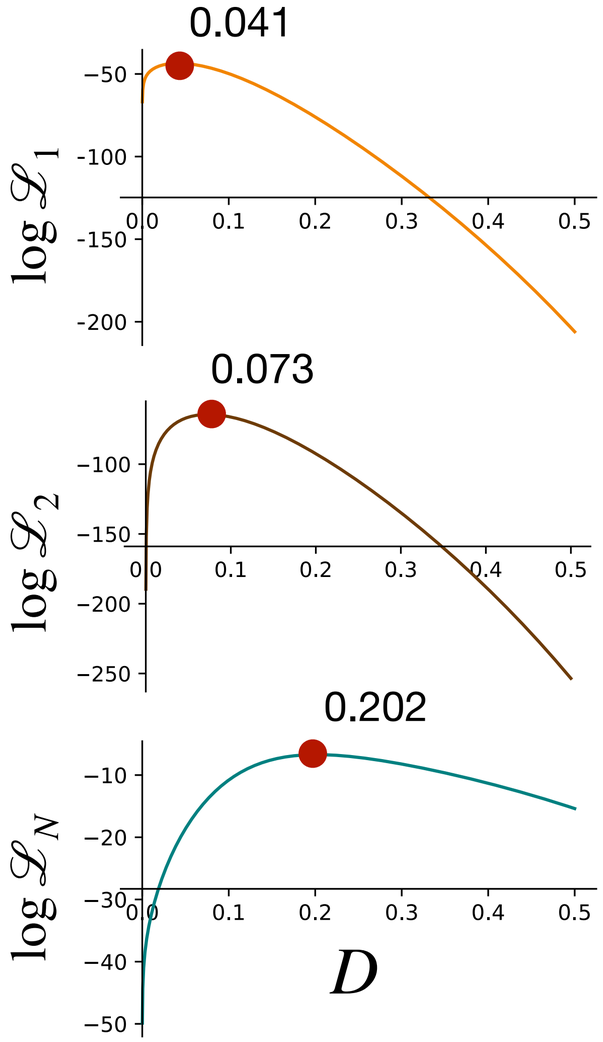


Maximum likelihood estimation of distances

Hamming distance histograms



Are maximum likelihood distances accurate?



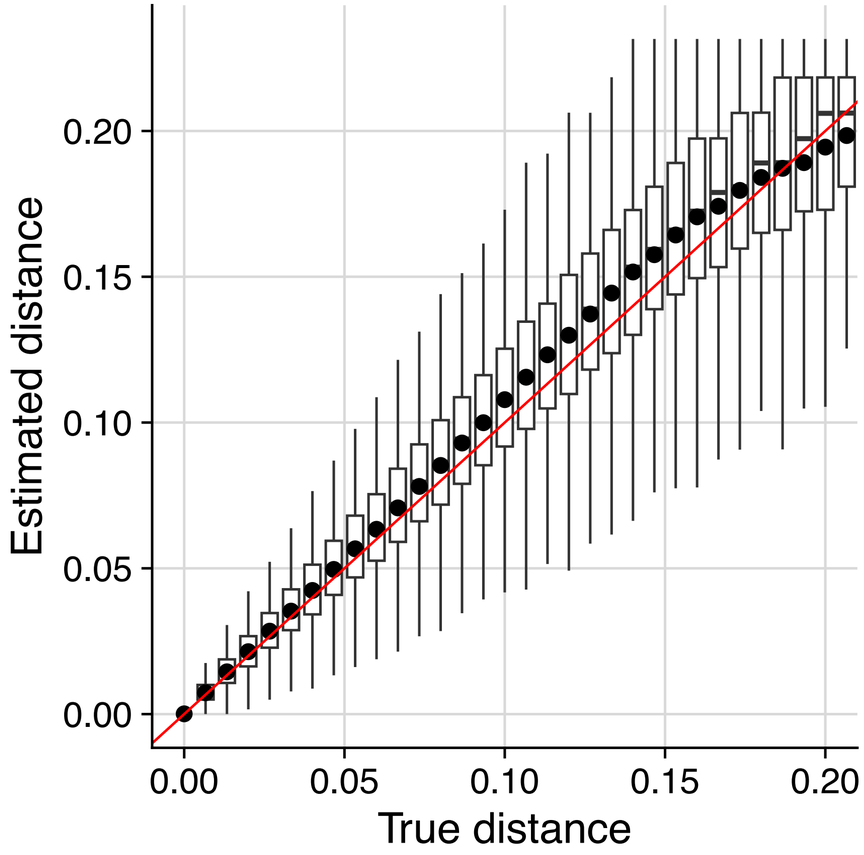
krepp estimates distances accurately at the read-level

default: 29-mer
minimizers of 35-mers

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

~150 bp short reads

(Hamming distance) / (seq. length)



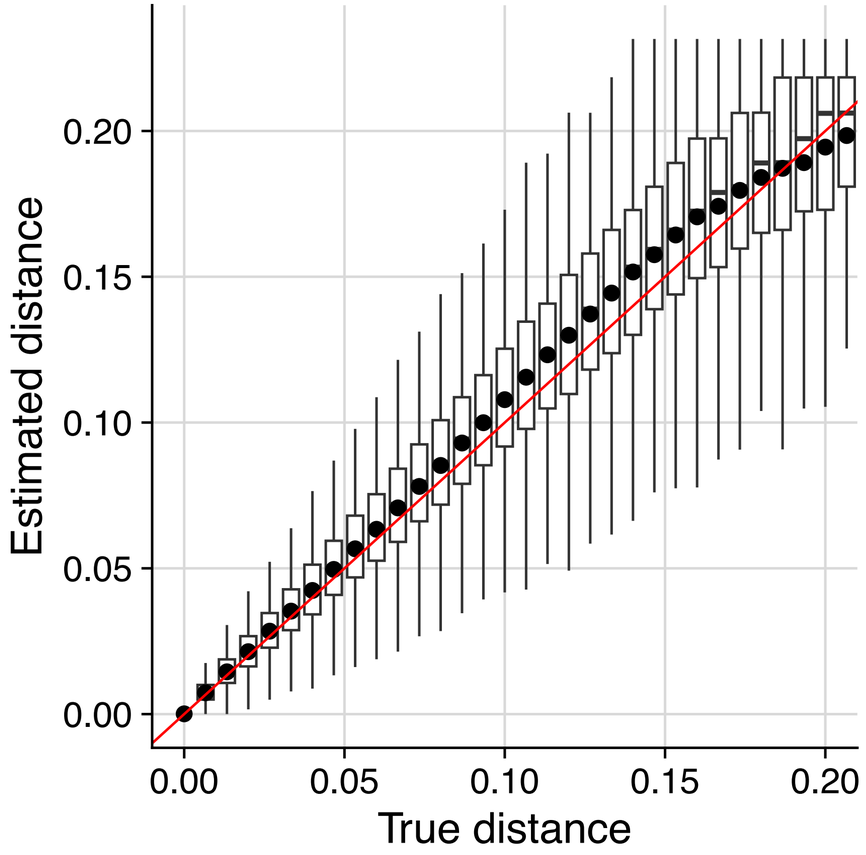
krepp estimates distances accurately at the read-level

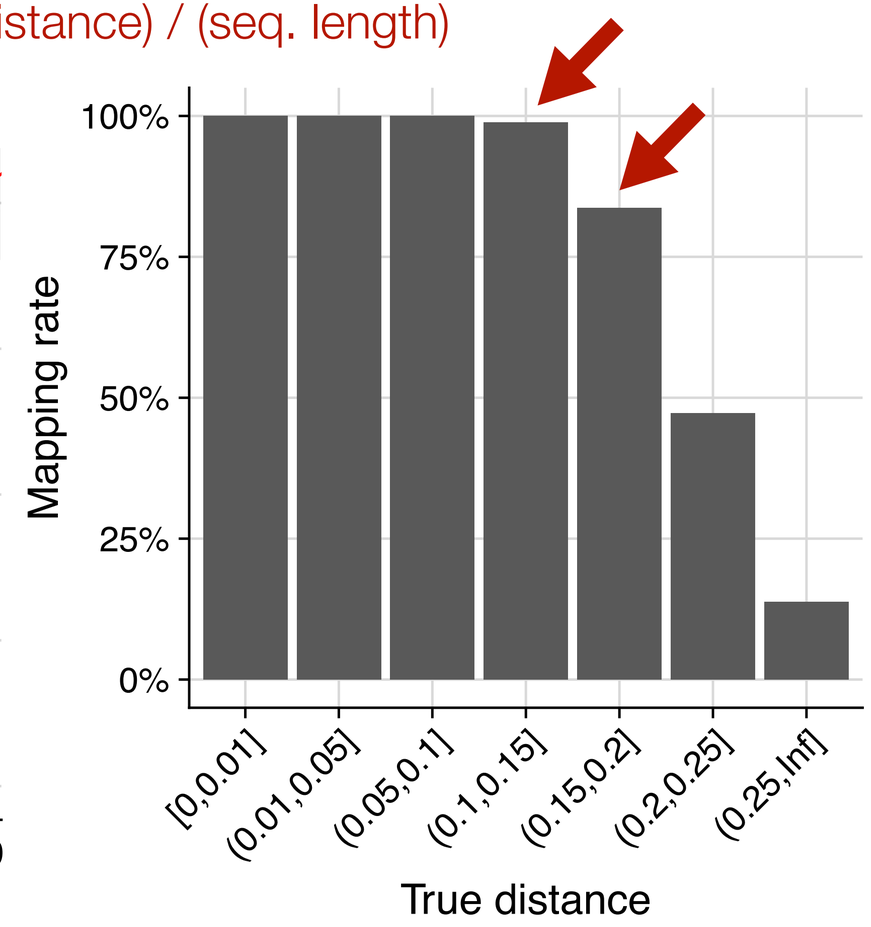
default: 29-mer
minimizers of 35-mers

- 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)

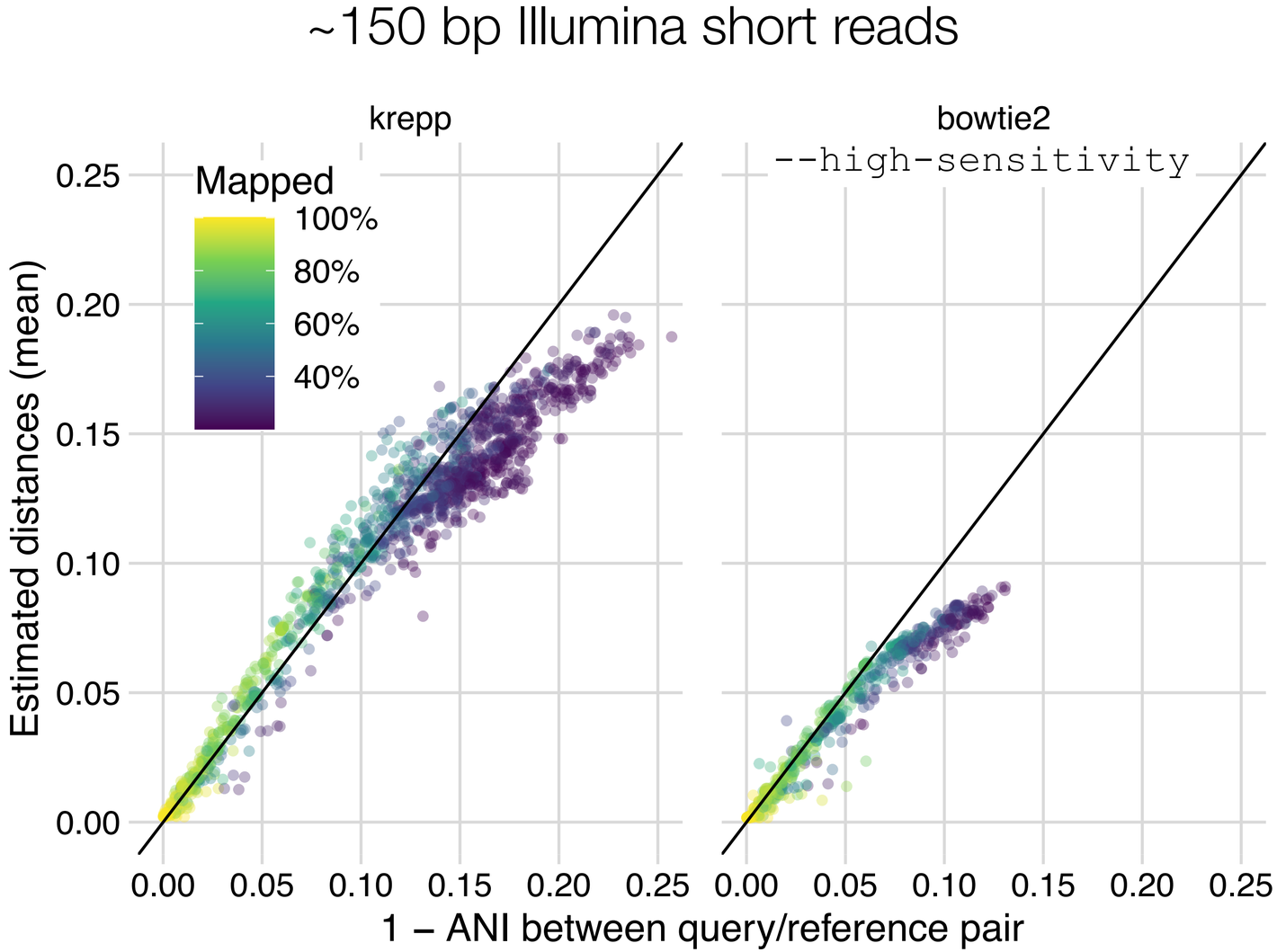




krepp matches nucleotide identity on average for real genomes

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

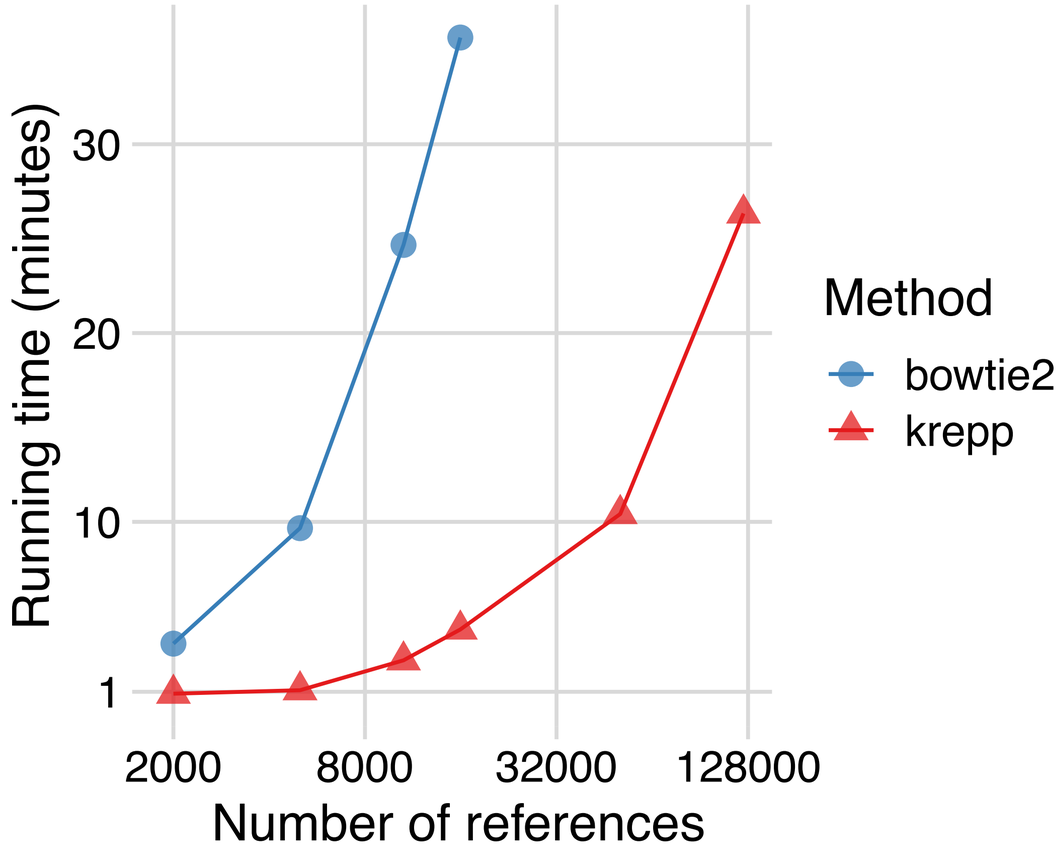
- Real query/reference genomes (pairs with >20% mapping rate)
- *krepp* extends to distant (>10%) reference genomes accurately
- Increasing the sensitivity by relaxing the alignment is costly



Scalability:
Avoiding alignment & effective parallelization

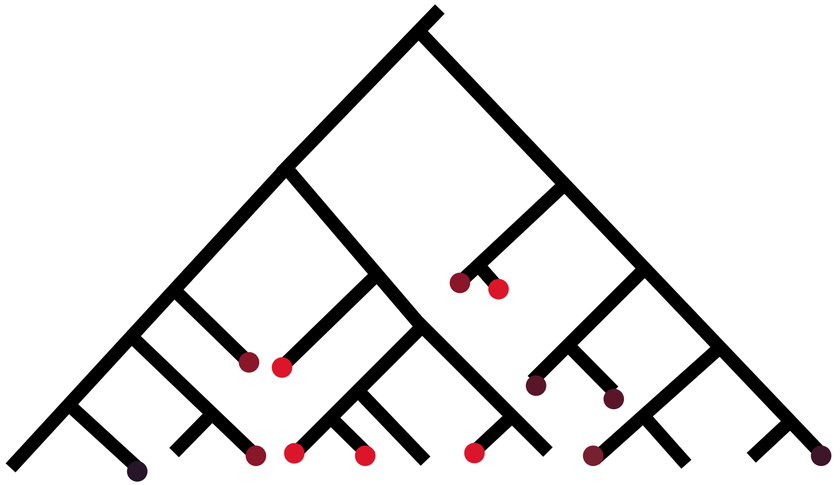
- *krepp* is already **>10x faster, scales well w/ large references**
- **>3x faster indexing** compared bowtie2

~150 bp Illumina short reads
Mapping 10M reads (16 threads):



Problem IV: distance-based placement

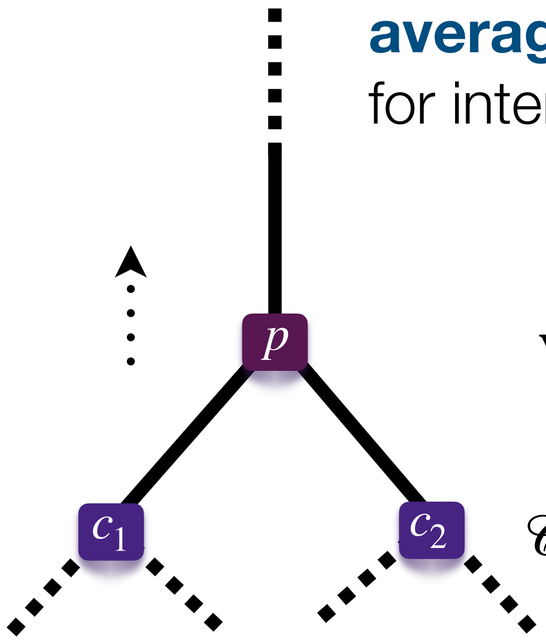
Given $d(q, R_i)$ for many R_is, find the "best" placement of q on T



Challenge:

- Short reads — **low signal**
  - Small differences in distances may not be meaningful (**statistical distinguishability**)

Defining a notion of distance for clades

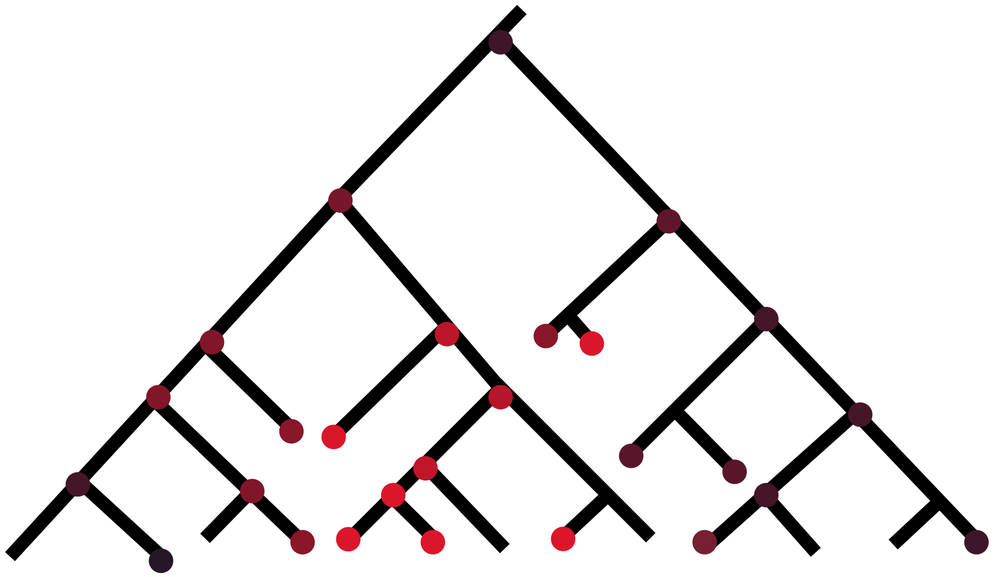


recursively compute
average HD histograms
for internal nodes

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

$\mathscr{C}(p)$: 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
w.r.t. 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
w.r.t. the closest reference:

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

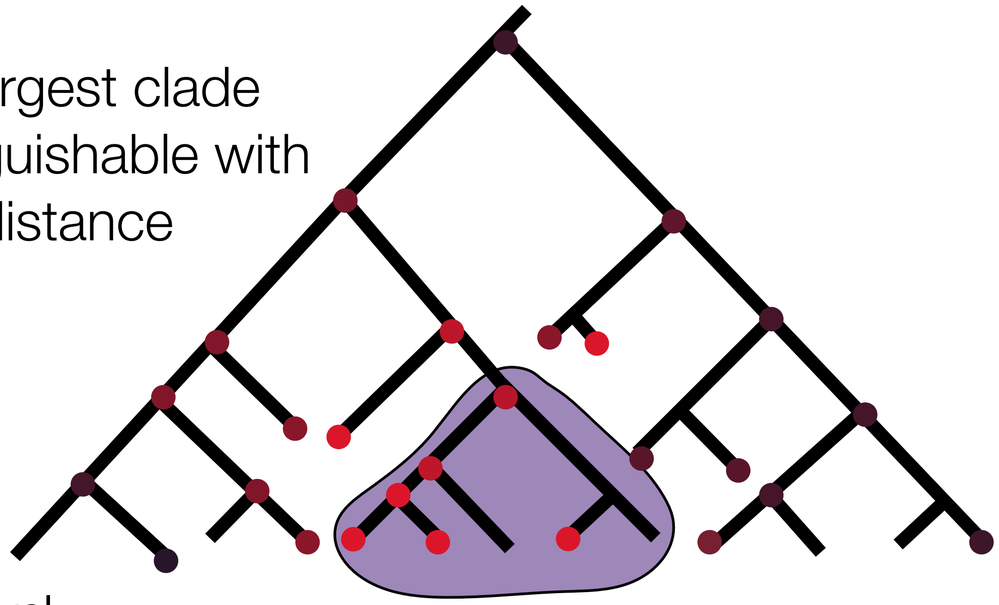
D: alternative distance

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

$i*$: closest reference

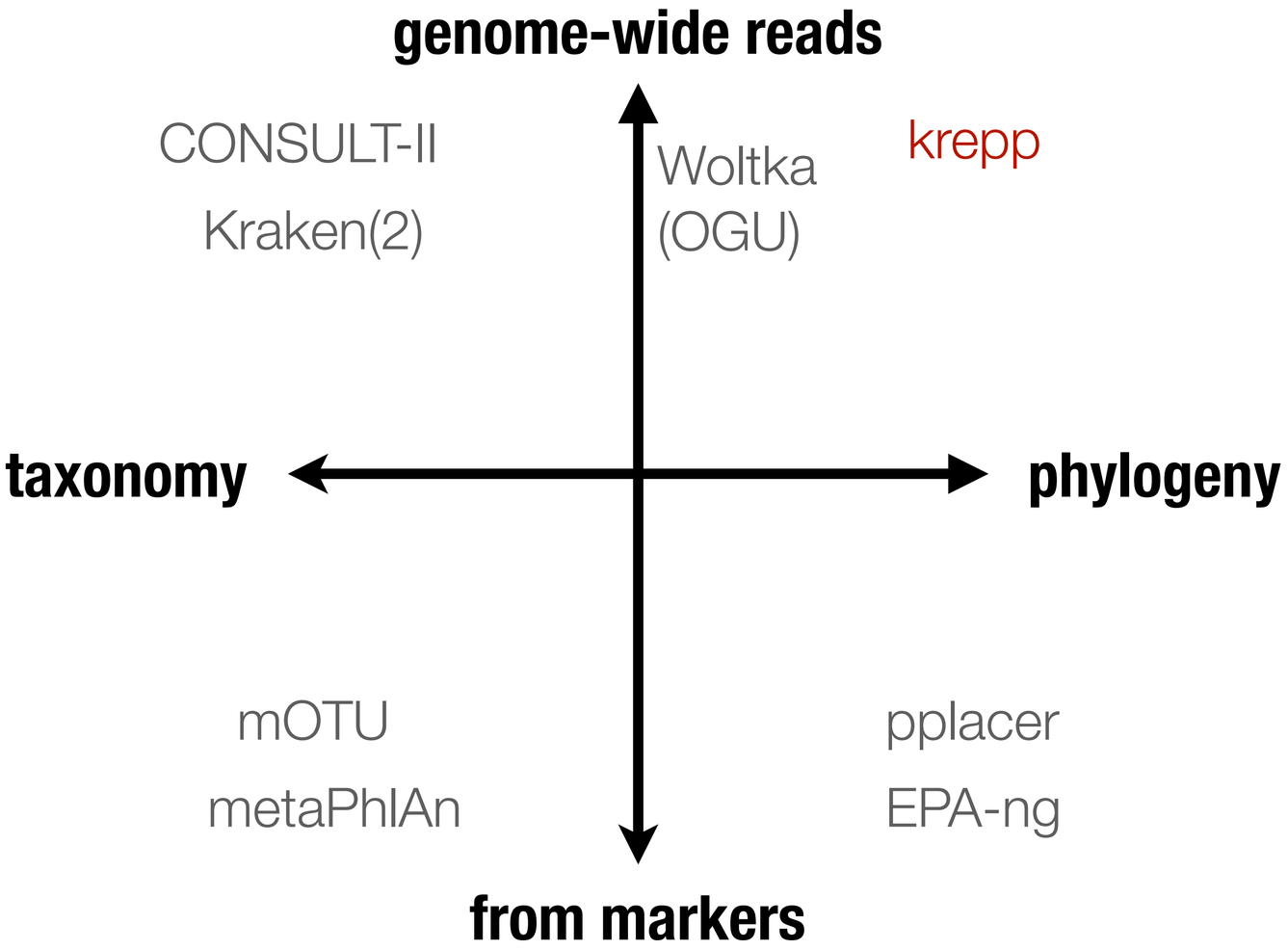
$\lambda_{LR} \sim \chi^2$

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



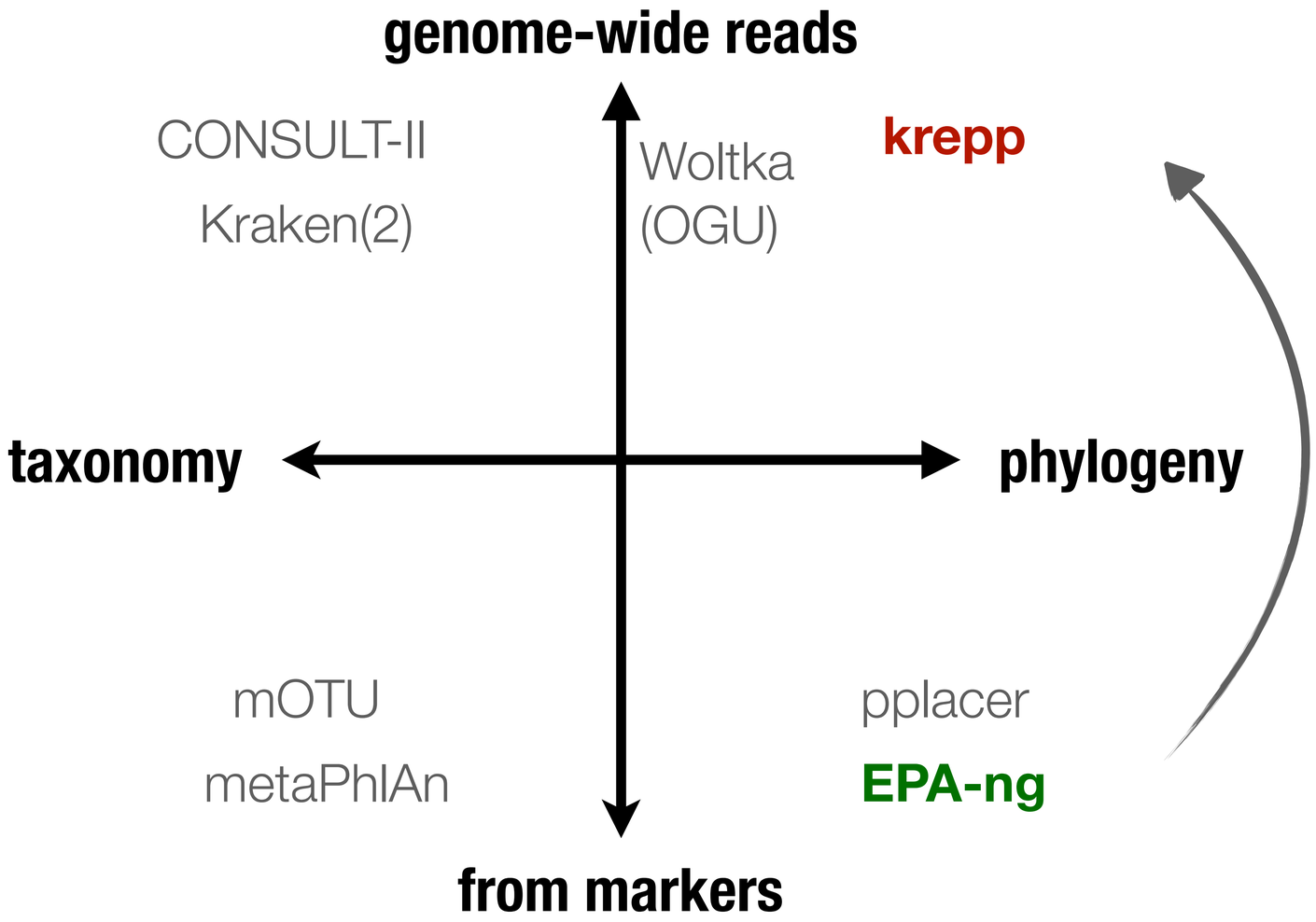
indistinguishable w.r.t. the closest reference

Comparing krepp and marker-based ML placement



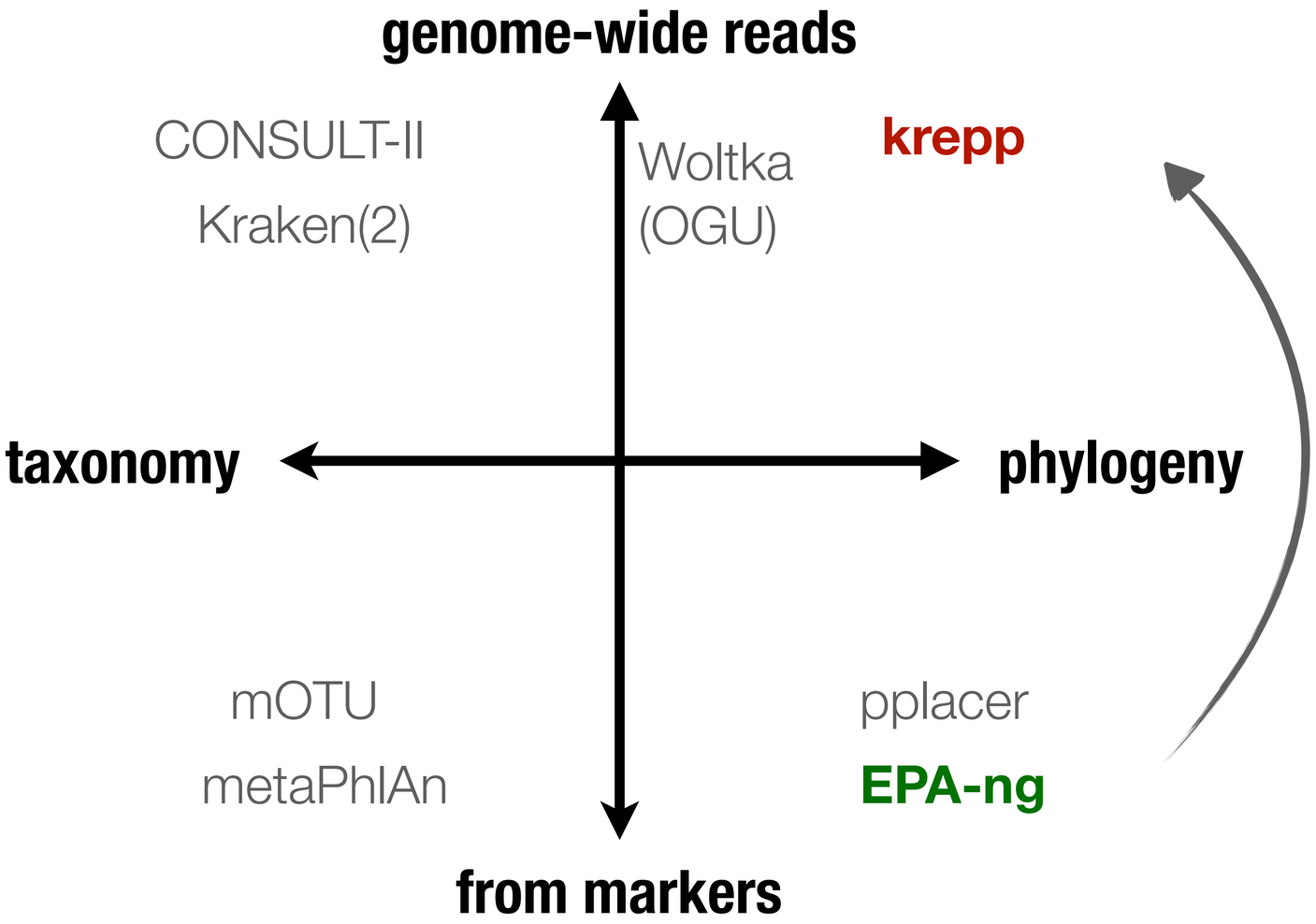
Comparing krepp and marker-based ML placement

- EPA-ng: needs a MSA; only markers



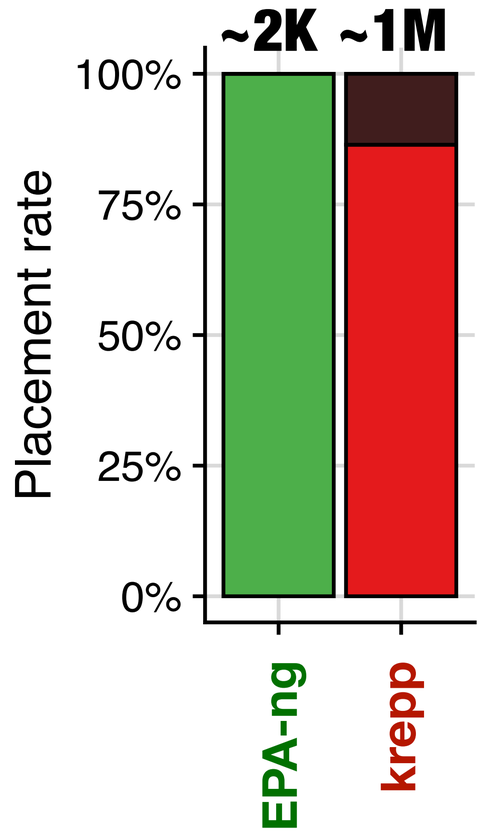
several orders of magnitude more reads analyzed

Comparing krepp and marker-based ML placement



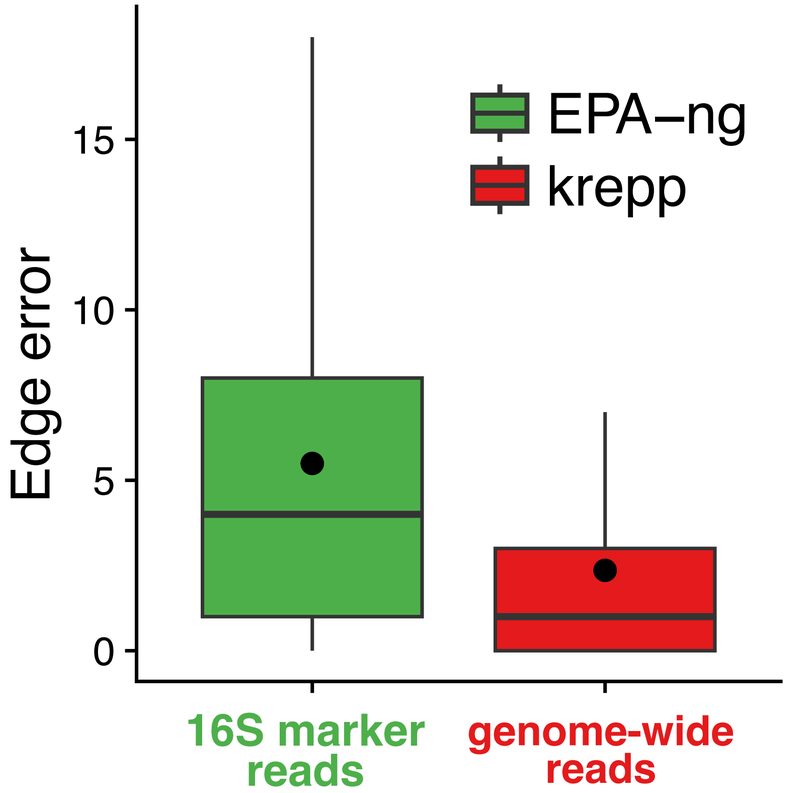
- EPA-ng: needs a MSA; only markers
- *krepp* places 86% of all reads

several orders of magnitude more reads analyzed



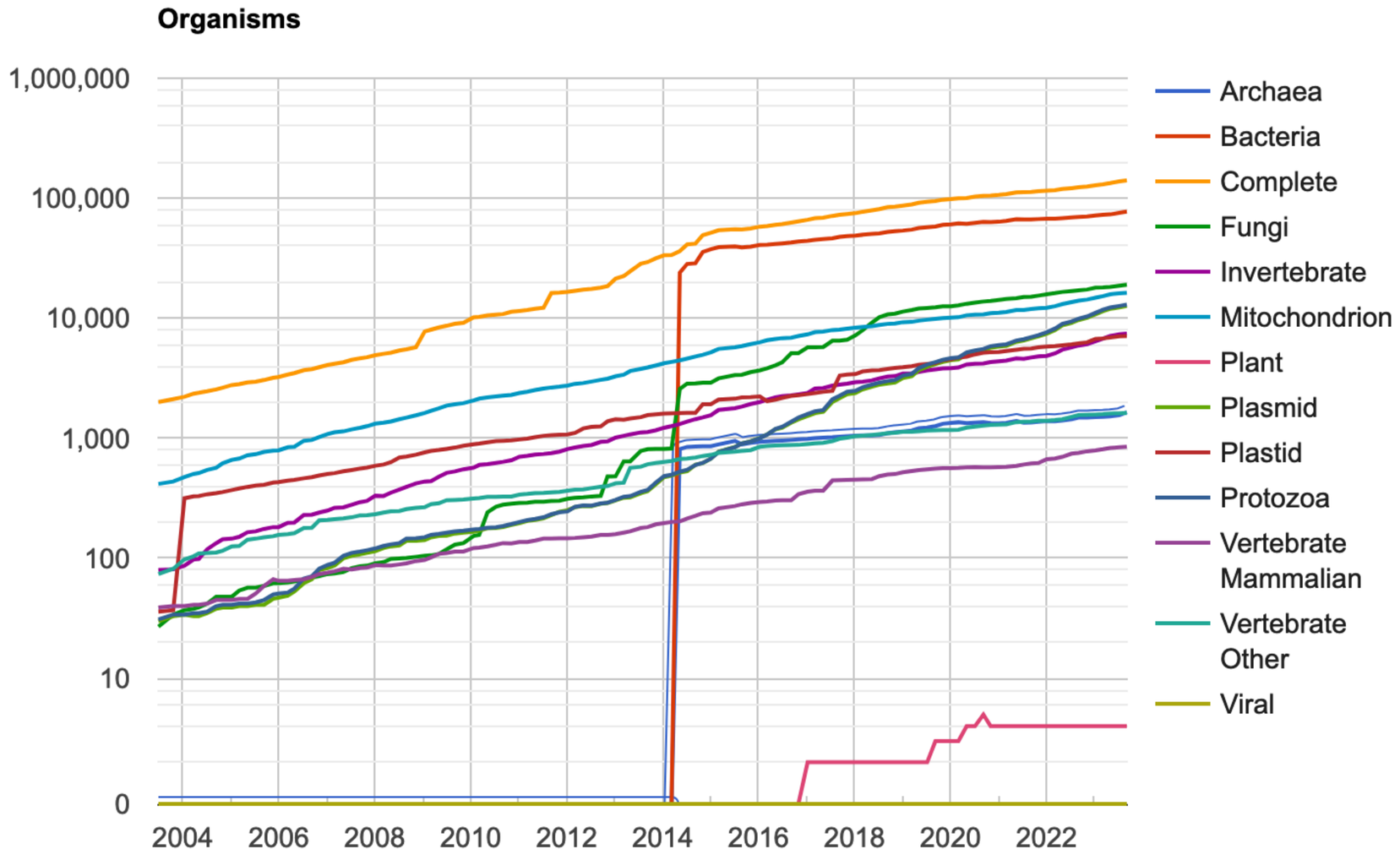
krepp places genome-wide reads more accurately than ML-based 16S placement

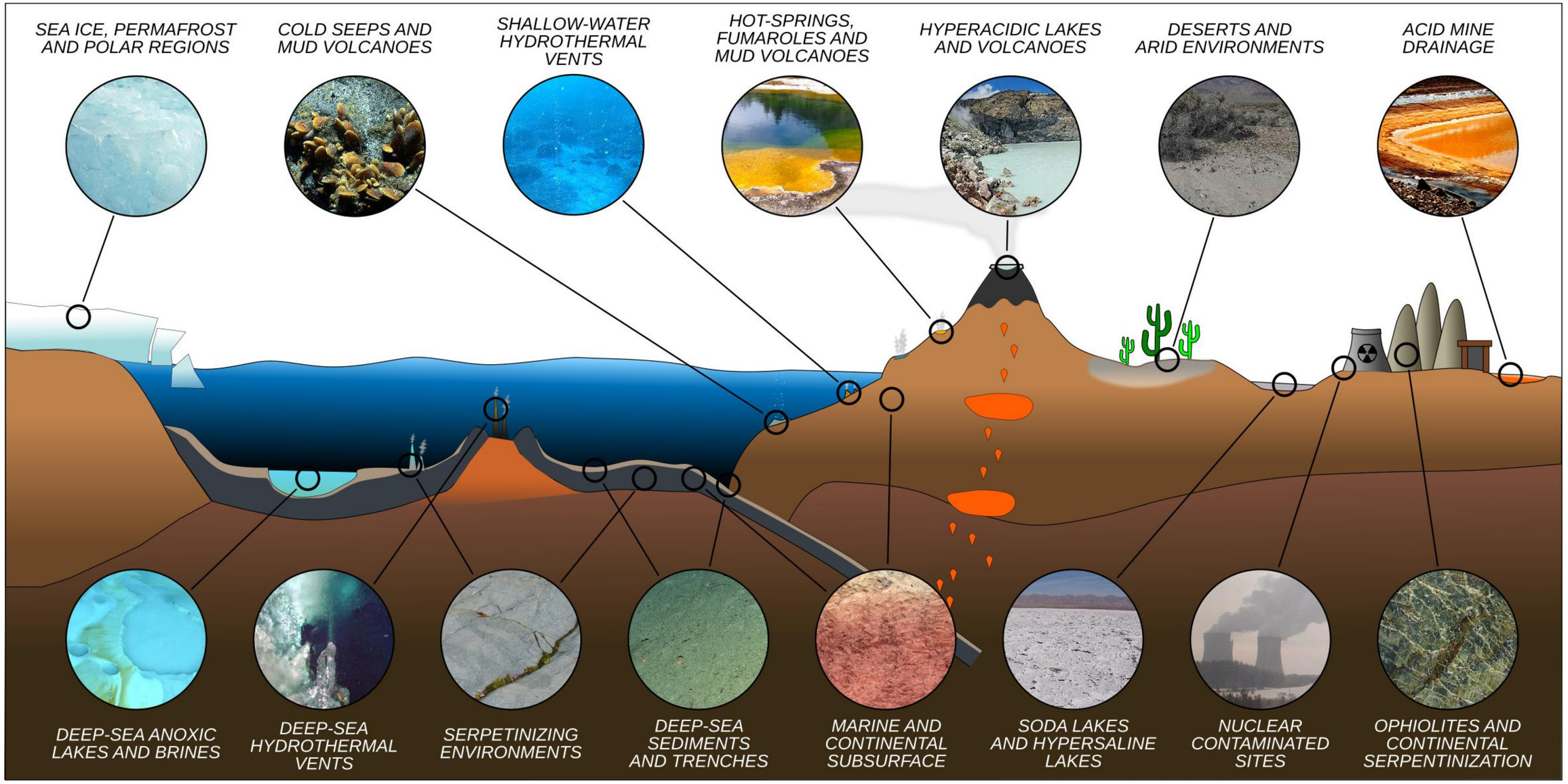
- Leave all out —100 queries from 11,000 taxa (WoLv1)
- **2.4 vs. 5.6 edge error (average)**



Summary and a few takeaways

- **Metagenomics faces many challenges:**
  - Novel sequences, low quality or absent references, scalibility
- Good practices and established protocols but **new concepts can emerge**
- **Phylogenies & better algorithms can alleviate some of these problems:**
- We have good tools — but these problems are not solved yet.
  - Ever growing databases!
  - Novel and extreme enviroments!


Organisms
1,000,000
100,000
10,000
1,000
100
10
0
2004
2006
2008
2010
2012
2014
2016
2018
2020
2022
Archaea
Bacteria
Complete
Fungi
Invertebrate
Mitochondrion
Plant
Plasmid
Plastid
Protozoa
Vertebrate Mammalian
Vertebrate Other
Viral


SEA ICE, PERMAFROST AND POLAR REGIONS
COLD SEEPS AND MUD VOLCANOES
SHALLOW-WATER HYDROTHERMAL VENTS
HOT-SPRINGS, FUMAROLES AND MUD VOLCANOES
HYPERACIDIC LAKES AND VOLCANOES
DESERTS AND ARID ENVIRONMENTS
ACID MINE DRAINAGE
DEEP-SEA ANOXIC LAKES AND BRINES
DEEP-SEA HYDROTHERMAL VENTS
SERPETINIZING ENVIRONMENTS
DEEP-SEA SEDIMENTS AND TRENCHES
MARINE AND CONTINENTAL SUBSURFACE
SODA LAKES AND HYPERSALINE LAKES
NUCLEAR CONTAMINATED SITES
OPHIOLITES AND CONTINENTAL SERPENTINIZATION

Metagenomics and sequence analysis

CONSULT-II: accurate taxonomic identification and profiling using locality-sensitive hashing

Ali Osman Berk Şapcı [1], Eleonora Rachtman [1], Siavash Mirarab [1,2,*]

Home > Genome Biology > Article

krepp: a *k*-mer-based maximum pseudo-likelihood method for estimating read distances and genome-wide phylogenetic placement

Genome Biology

Ali Osman Berk Şapcı & Siavash Mirarab

Metagenomics and sequence analysis

CONSULT-II: accurate taxonomic identification and profiling using locality-sensitive hashing

Ali Osman Berk Şapcı [1], Eleonora Rachtman [1], Siavash Mirarab [1,2,*]

Memory-bound *k*-mer selection for large and evolutionarily diverse reference libraries

Ali Osman Berk Şapcı[1] and Siavash Mirarab[1,2]

Home > Genome Biology > Article

krepp: a *k*-mer-based maximum pseudo-likelihood method for estimating read distances and genome-wide phylogenetic placement

Genome Biology

Ali Osman Berk Şapcı & Siavash Mirarab

Metagenomics and sequence analysis

CONSULT-II: accurate taxonomic identification and profiling using locality-sensitive hashing

Ali Osman Berk Şapcı [1], Eleonora Rachtman [1], Siavash Mirarab [1,2,*]

Memory-bound *k*-mer selection for large and evolutionarily diverse reference libraries

Ali Osman Berk Şapcı[1] and Siavash Mirarab[1,2]

Home > Genome Biology > Article

krepp: a *k*-mer-based maximum pseudo-likelihood method for estimating read distances and genome-wide phylogenetic placement

Genome Biology

Ali Osman Berk Şapcı & Siavash Mirarab

Population genetics and ecology

SPrUCE: Utilizing Ultraconserved Elements of DNA for Population-Level Genetic Diversity Estimation

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

Metagenomics and sequence analysis

CONSULT-II: accurate taxonomic identification and profiling using locality-sensitive hashing

Ali Osman Berk Şapcı [1], Eleonora Rachtman [1], Siavash Mirarab [1,2,*]

Memory-bound *k*-mer selection for large and evolutionarily diverse reference libraries

Ali Osman Berk Şapcı[1] and Siavash Mirarab[1,2]

Home > Genome Biology > Article

krepp: a *k*-mer-based maximum pseudo-likelihood method for estimating read distances and genome-wide phylogenetic placement

Genome Biology

Ali Osman Berk Şapcı & Siavash Mirarab

Population genetics and ecology

SPrUCE: Utilizing Ultraconserved Elements of DNA for Population-Level Genetic Diversity Estimation

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

Phylogenetics

Phlag: Scalable detection of genomics regions with unexplained phylogenetic heterogeneity

Ali Osman Berk Şapcı,[1] Shayesteh Arasti,[2] Edward L. Braun[3] and Siavash Mirarab[1,4*]

Deconvolving Phylogenetic Distance Mixtures

Shayesteh Arasti,[1] Ali Osman Berk Şapcı,[2] Eleonora Rachtman,[4] Mohammed El-Kebir,[3] and Siavash Mirarab[4*]

Siavash Mirarab

Our building @ UCSD

Daira, Eduardo, Nora, Shayesteh, Yueyu…

Supplementary Slides

CONSULT-II

Adding taxonomic information for each k-mer

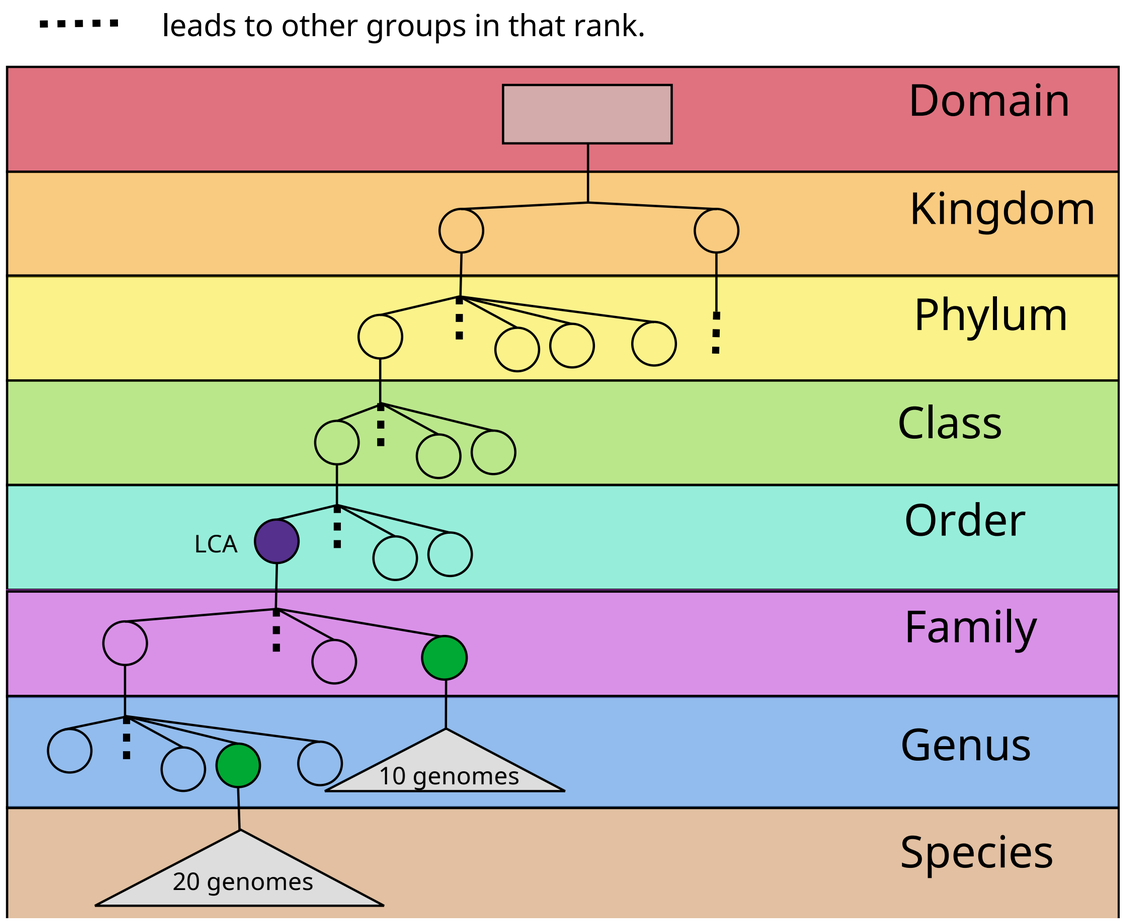
- Keeping a list of genomes per *k*-mer: infeasible!
- CONSULT already uses 120GB for 8B *k*-mers

- Keep only the lowest common taxonomic ancestor (LCA)!

an idea used by other tools too

- 2 bytes for LCA taxon ID: 16GB in total for 8B *k*-mers

memory-wise, manageable

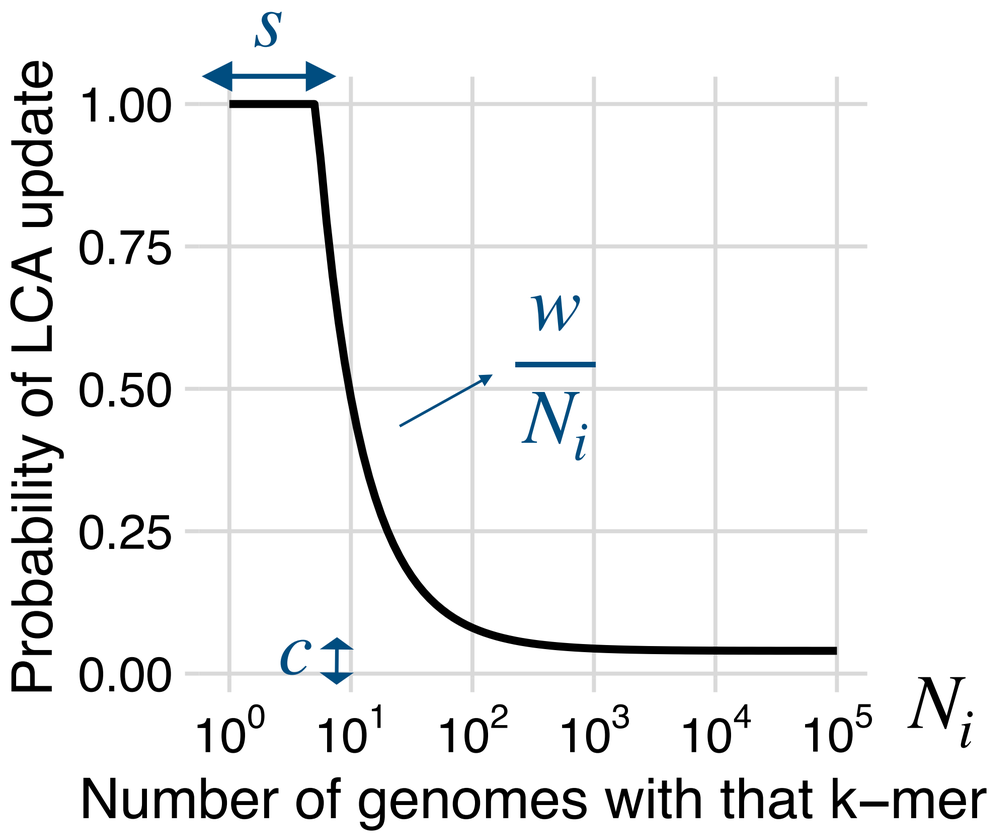


Soft-LCA approach of CONSULT-II

- **Idea:** Ignore each genome with some probability while computing the LCA taxon.
- **Intuition:** A k-mer should appear sufficiently many times in a group to affect the LCA taxon.
  - Frequent k-mers → many times.
  - Rare k-mers → a few would be enough.
- For each genome having the k-mer i, update the LCA taxon with probability $p_u(N_i)$:

$$p_u(N_i) = w \log_2(2^{\frac{N_i - 1}{s}} + 2)^{-1}$$

Two parameters: $s = 5$ and $w = 2$

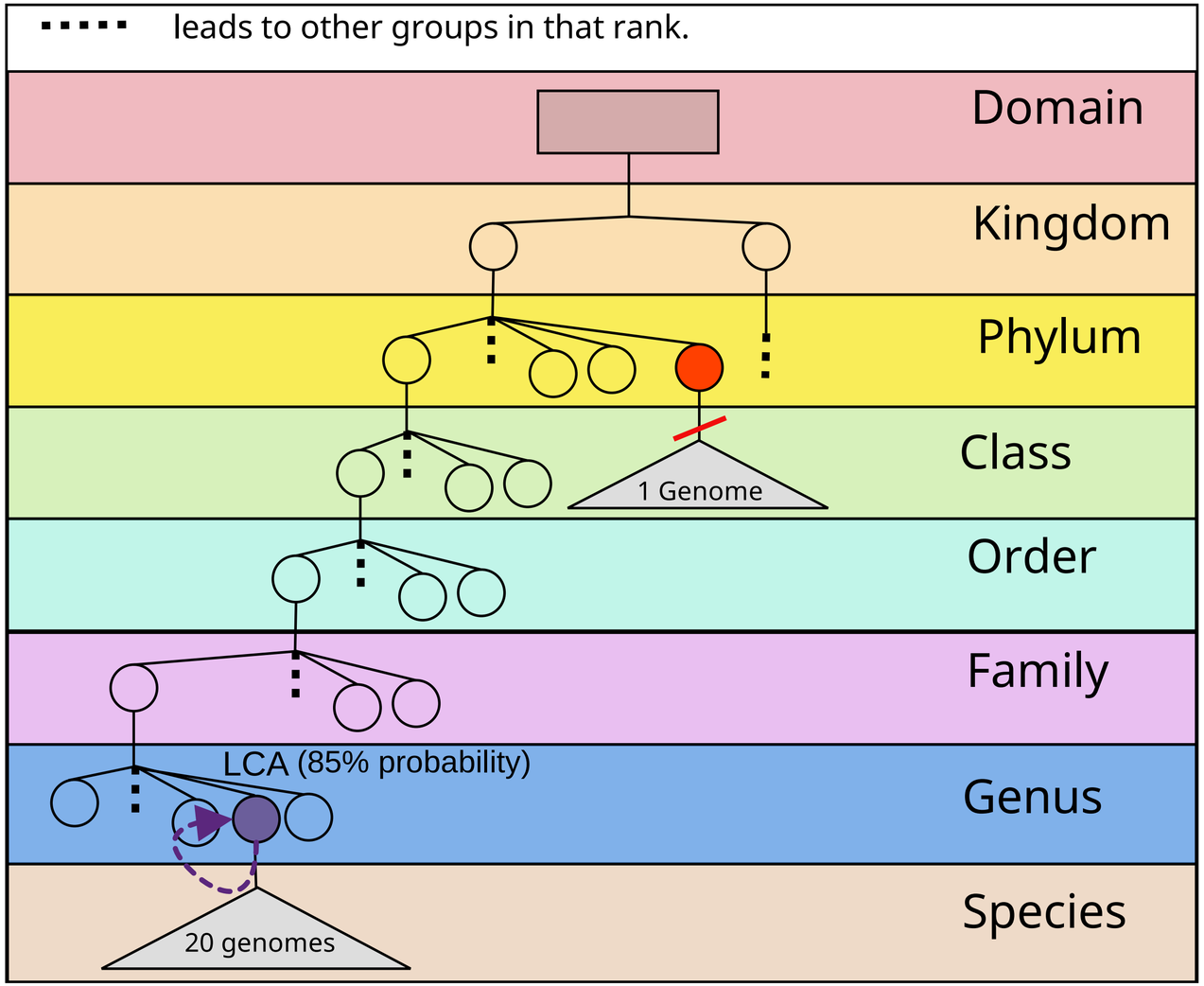


Back to the example - soft LCA of CONSULT-II

- **Example:**
  - 20 genomes in the **green genus**
  - 1 genome in the erroneous **red phylum**
- 85% probability → LCA taxon is the **correct genus**

$$(1-(1-p_u(21))^{20})(1-p_u(21)) \approx 0.85$$

(at least 1 out of 20 is not ignored)(1 is ignored)

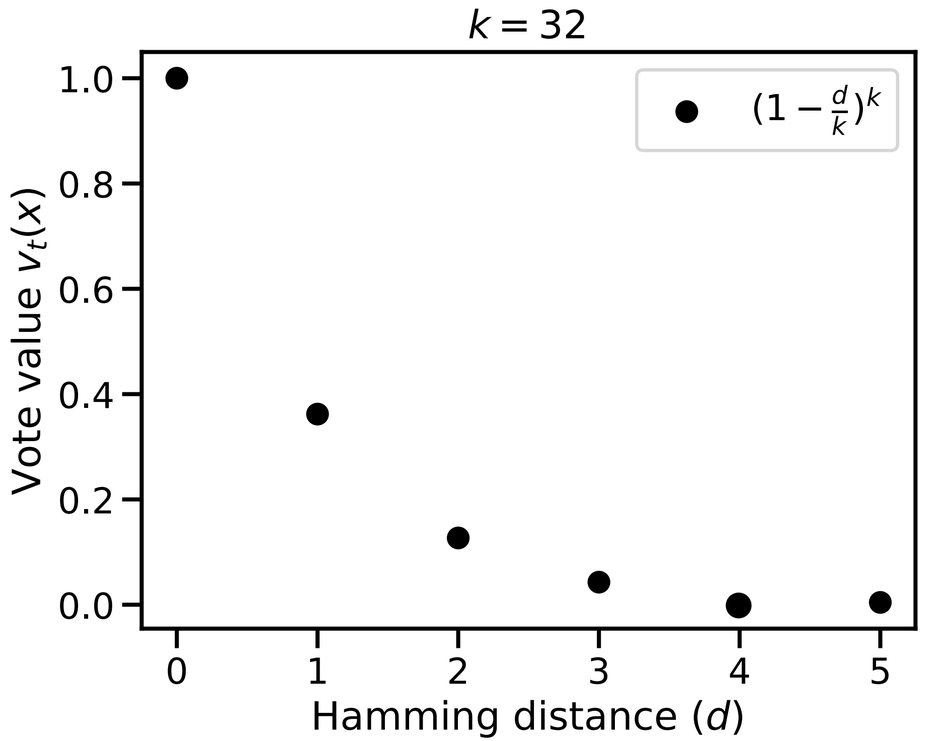


A vote-based taxonomic identification approach

- **Idea:** each *k*-mer match will vote to the corresponding taxon, weighted by its distance

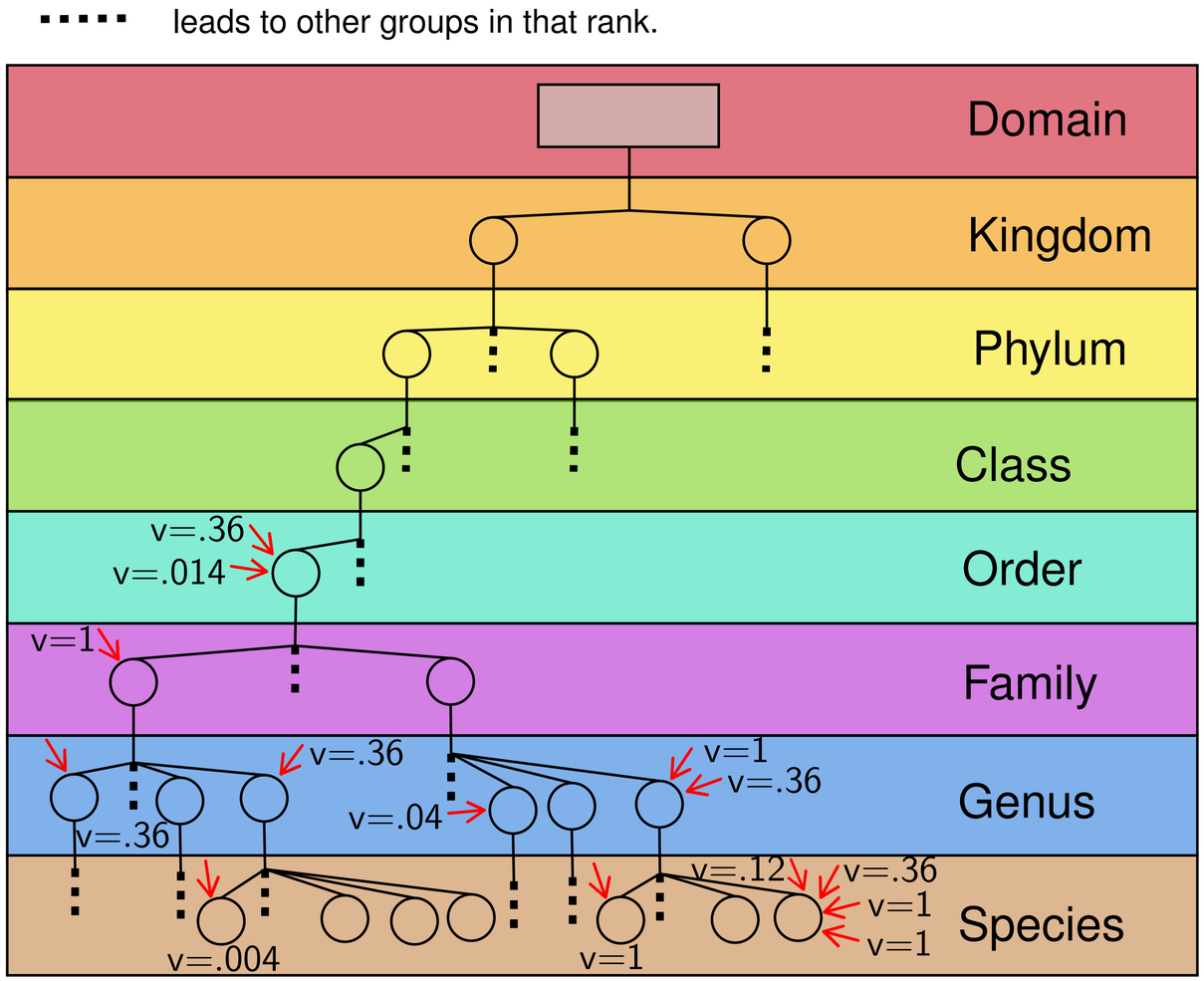
Heuristic: vote values decrease exponentially w.r.t. Hamming distance

$$v_x(t) = \left(1 - \frac{\min_{y \in \mathcal{K}(t)} hd(x, y)}{k}\right)^k$$



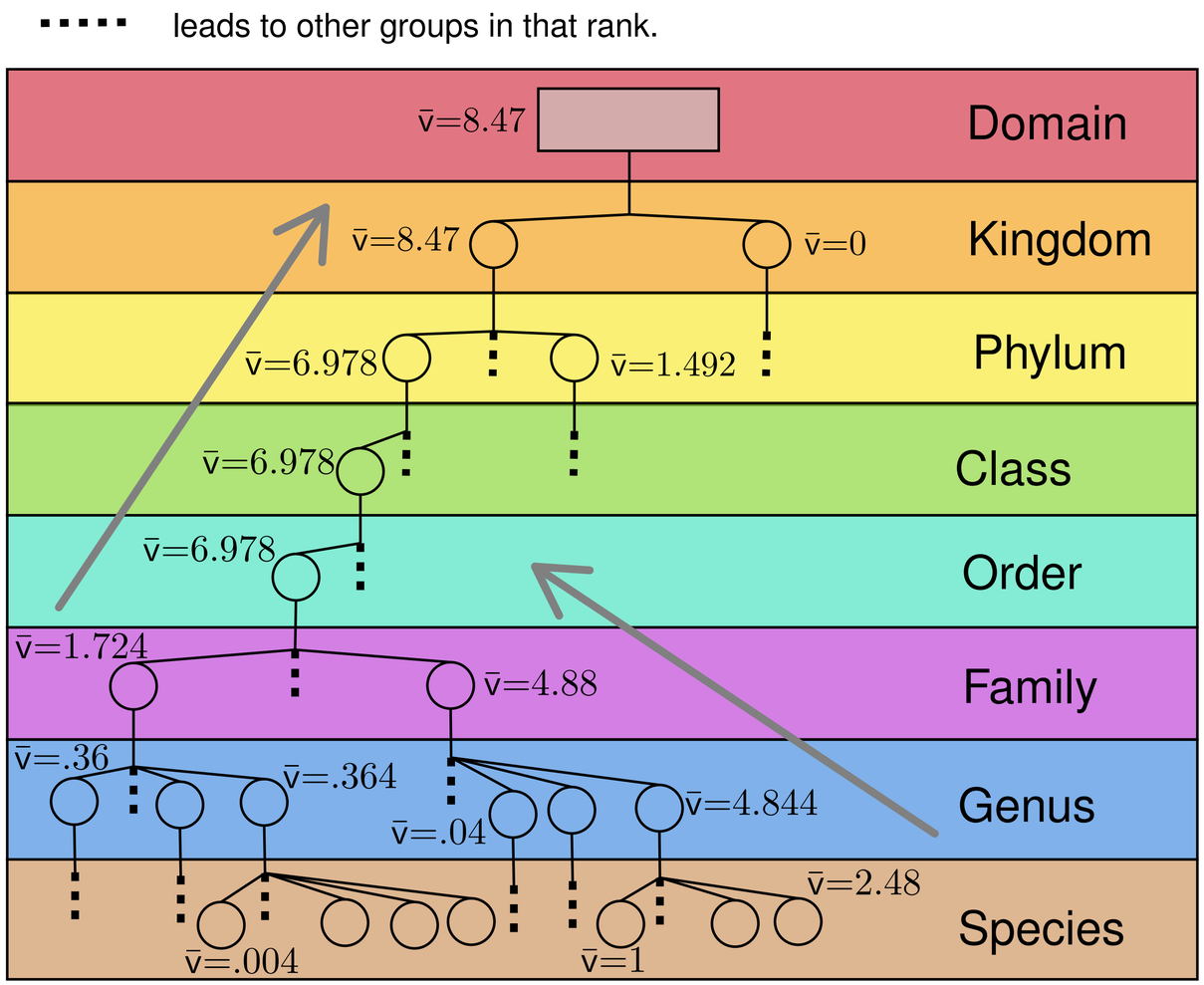
A vote-based taxonomic identification approach

- Vote of *k*-mer x for the taxon t:
  - d=0→$v_x(t)$=1.0
  - d=1→$v_x(t)$=0.36
  - d=2→$v_x(t)$=0.12
  - d=3→$v_x(t)$=0.04
  - d=4→$v_x(t)$=0.014
  - d=5→$v_x(t)$=0.004



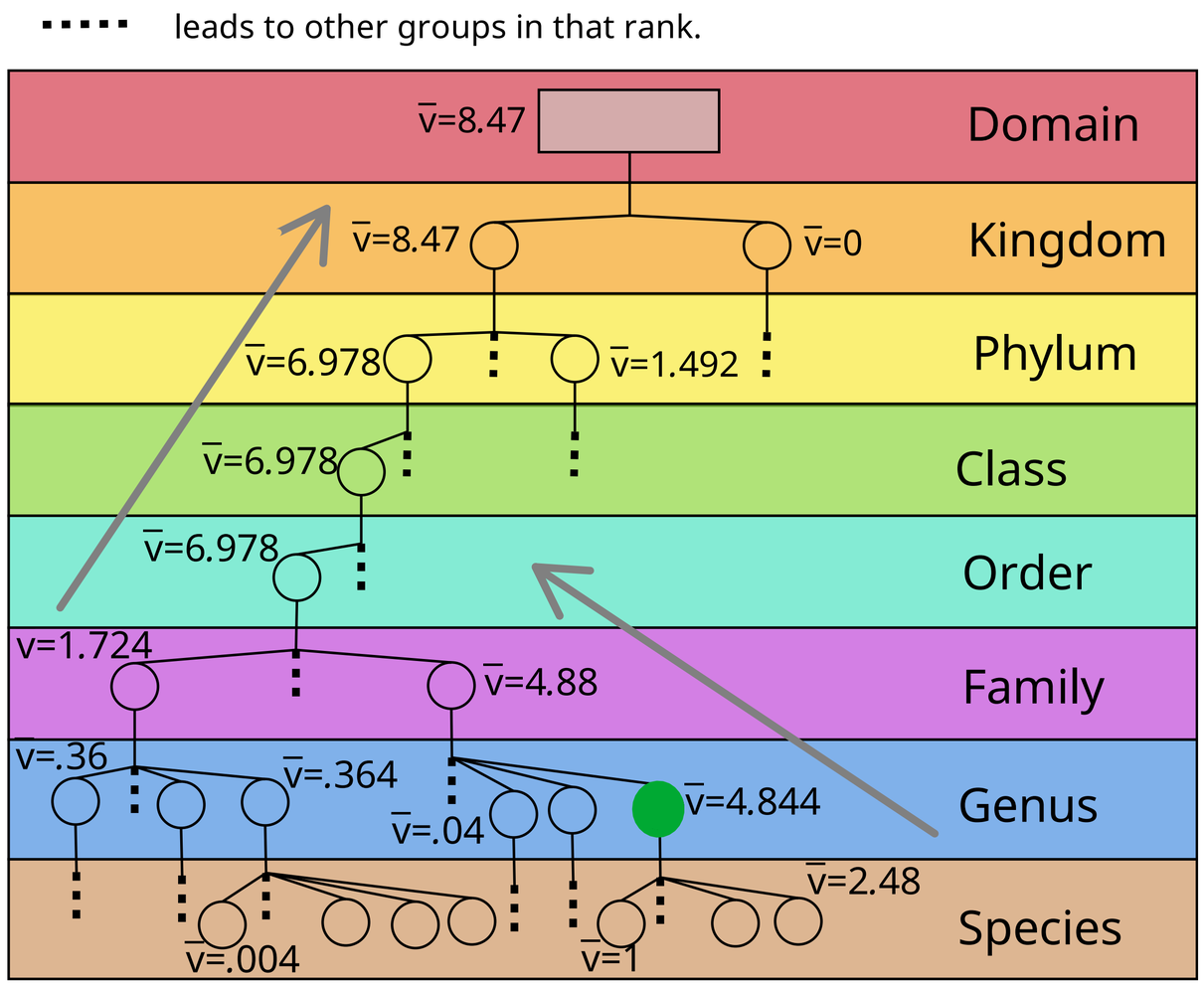
A vote-based taxonomic identification approach

- Incorporate the **hierarchical structure** between taxa:
  - use the tree: recursively sum in a bottom-up manner



A vote-based taxonomic identification approach

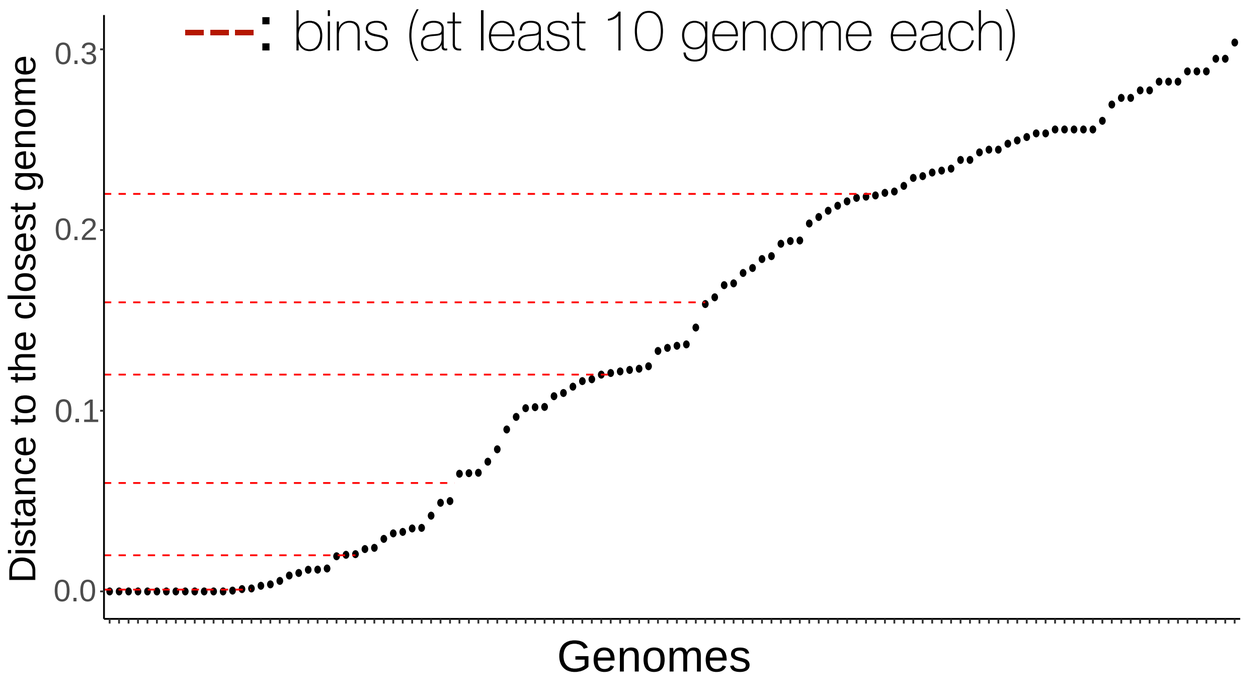
- Incorporate the **hierarchical structure** between taxa:
  - use the tree: recursively sum in a bottom-up manner
- Balance **specificity & sensitivity**:
  - require a majority vote (half of the vote at the root)
  - choose the lowest taxon exceeding the threshold



Controlled novelty benchmarking

Using WoL reference dataset (Zhu et al., 2019) → Moderate size: 9,000+ species, 10,000+ genomes

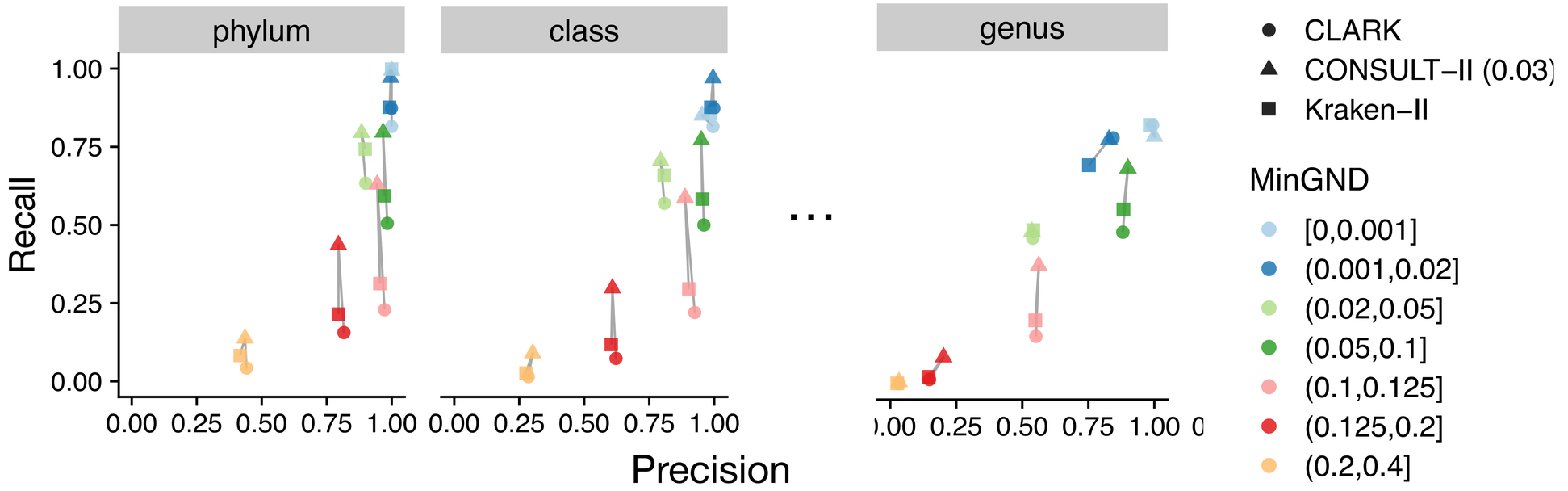
- Selected query genomes spanning a wide range of **novelty levels** (using Mash [Ondov et al., 2016])
- Short reads simulated from 120 bacterial & 100 archaeal genomes with Illumina error profiles
- Comparison with popular *k*-mer-based tools: Kraken 2 & CLARK



CONSULT-II can classify reads from distant genomes

- CONSULT-II significantly outperforms especially for novel queries
- Improvements are more palpable for upper levels (e.g., phylum, class)
- CONSULT-II has universally higher recall, precision levels are comparable

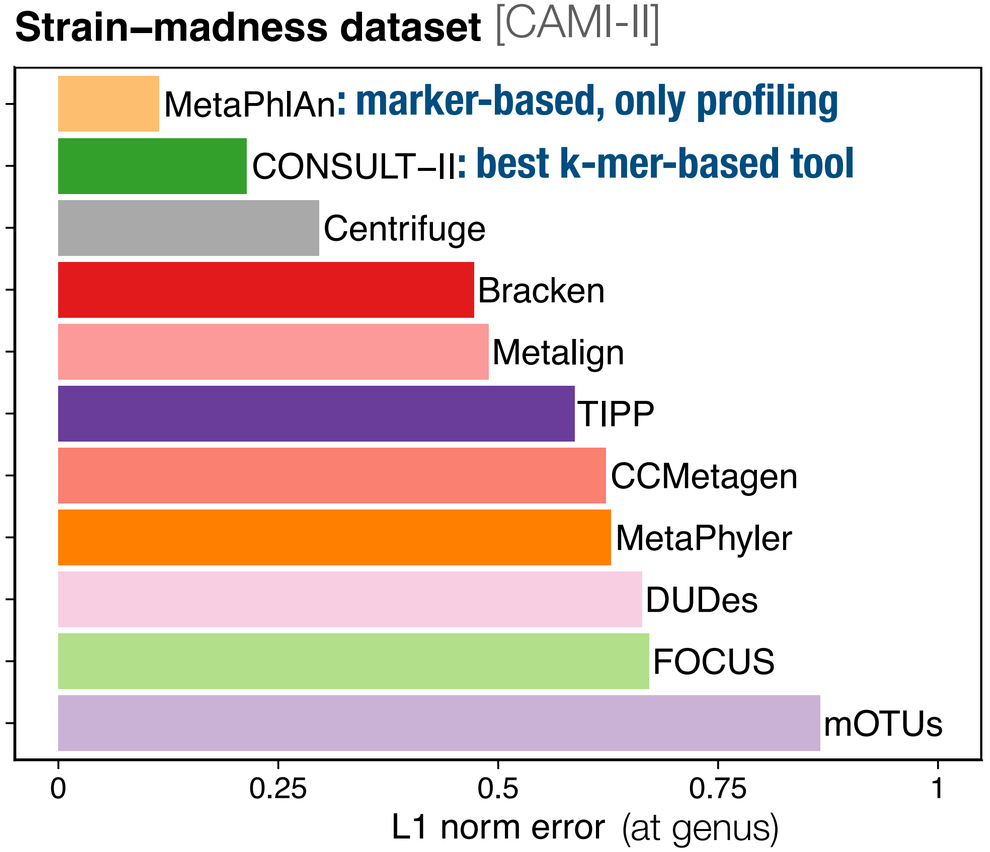
Better recall with no expense of precision!



CAMI-II: benchmarking challenge for metagenomics

- **Abundance profiling:** estimating the taxonomic composition of a given sample
- Using >130k reference genomes (RefSeq)
- CONSULT-II: best *k*-mer based tool, and second-best overall

Trade-off: marker-based methods are less flexible and cannot perform read classification



(using a RefSeq snapshot from 2019 with ~130k genomes)

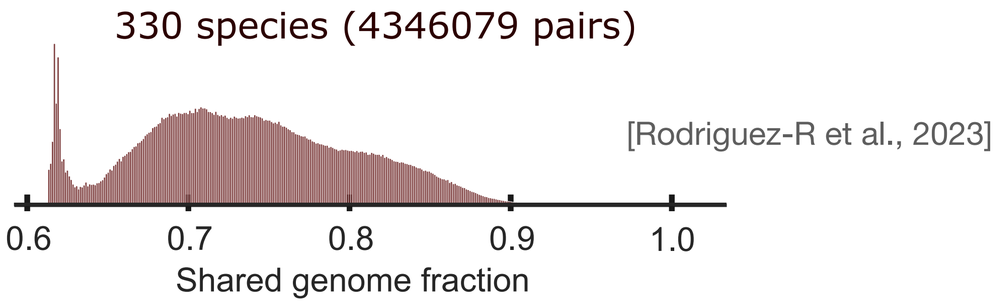
KRANK

Reducing the reference set by selecting k-mers

Baseline: random selection

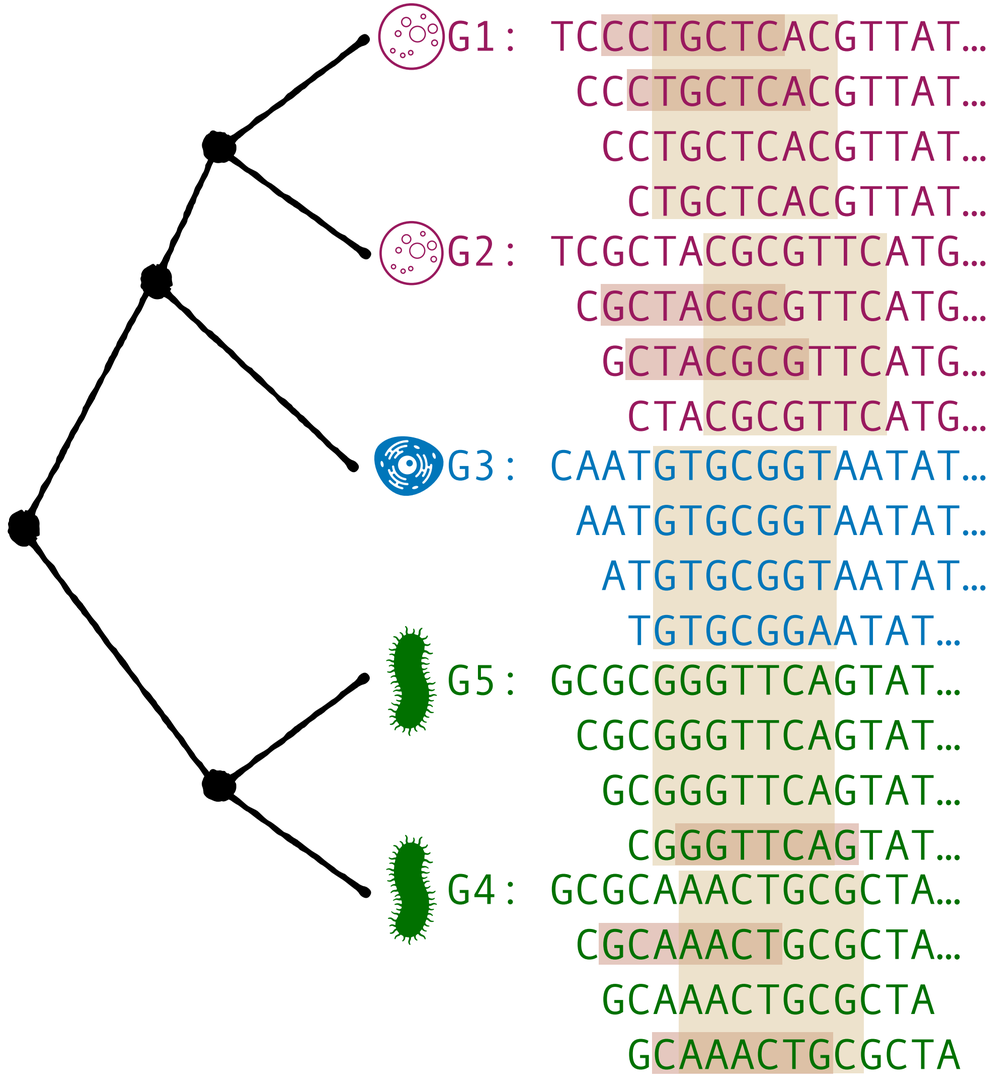
Minimizers: local sampling; select among overlapping *k*-mers with a sliding window

Even with minimizers, number of distinct *k*-mers grows fast with the number of genomes



[Rodriguez-R et al., 2023]

Idea: don't treat each genome independent; exploit the evolutionary dimension

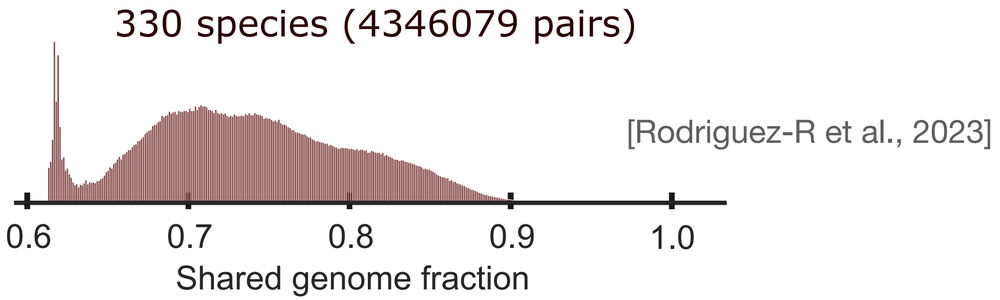


Reducing the reference set by selecting k-mers

Baseline: random selection

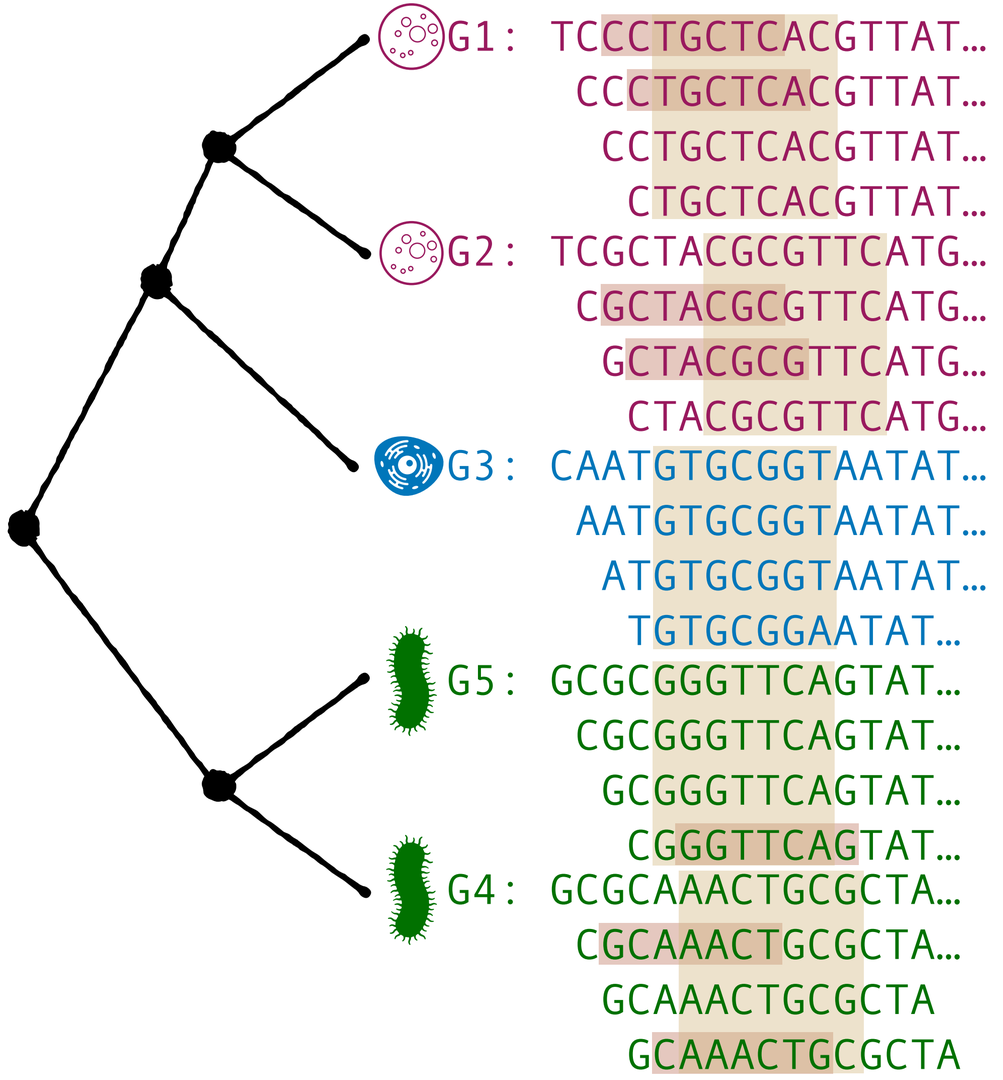
Minimizers: local sampling; select among overlapping *k*-mers with a sliding window

Even with minimizers, number of distinct *k*-mers grows fast with the number of genomes



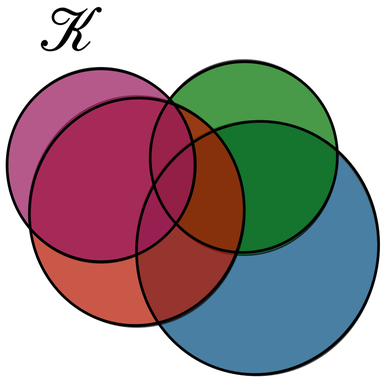
[Rodriguez-R et al., 2023]

Idea: don't treat each genome independent; exploit the evolutionary dimension

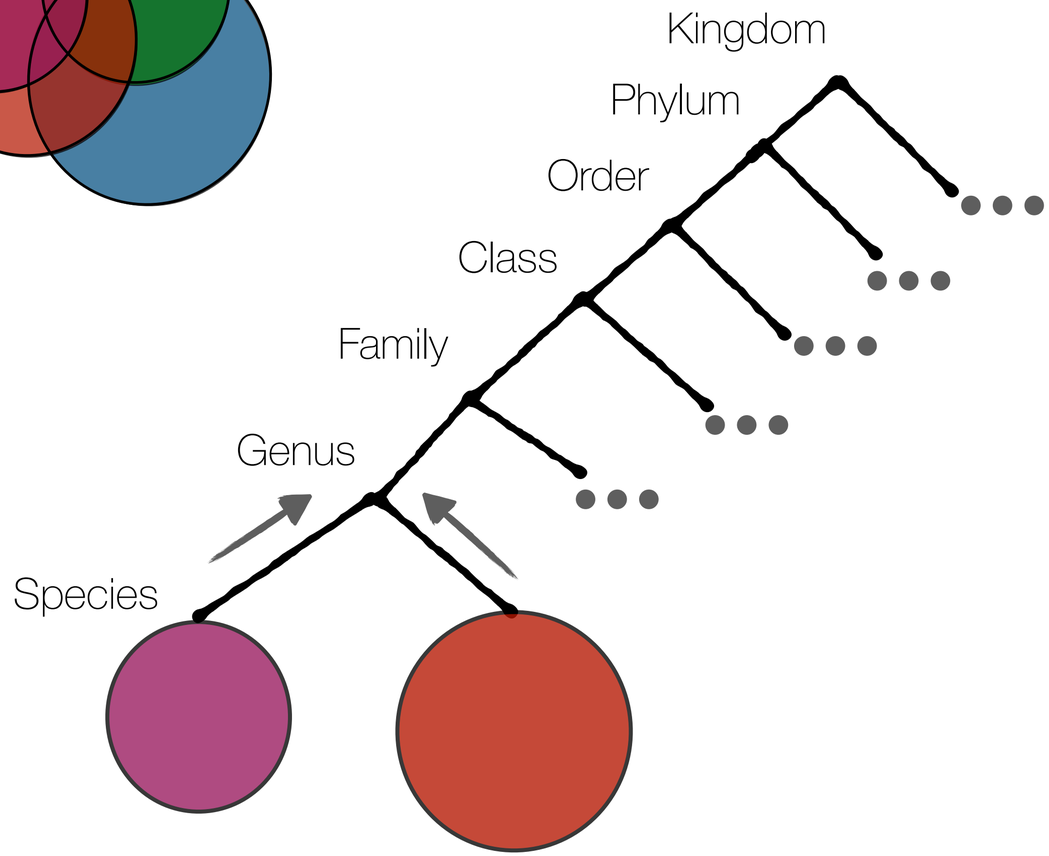


KRANK selects a representative *k*-mer subset in a memory-bound manner!

Challenge: explicitly computing *k*-mer set intersections across taxa is expensive

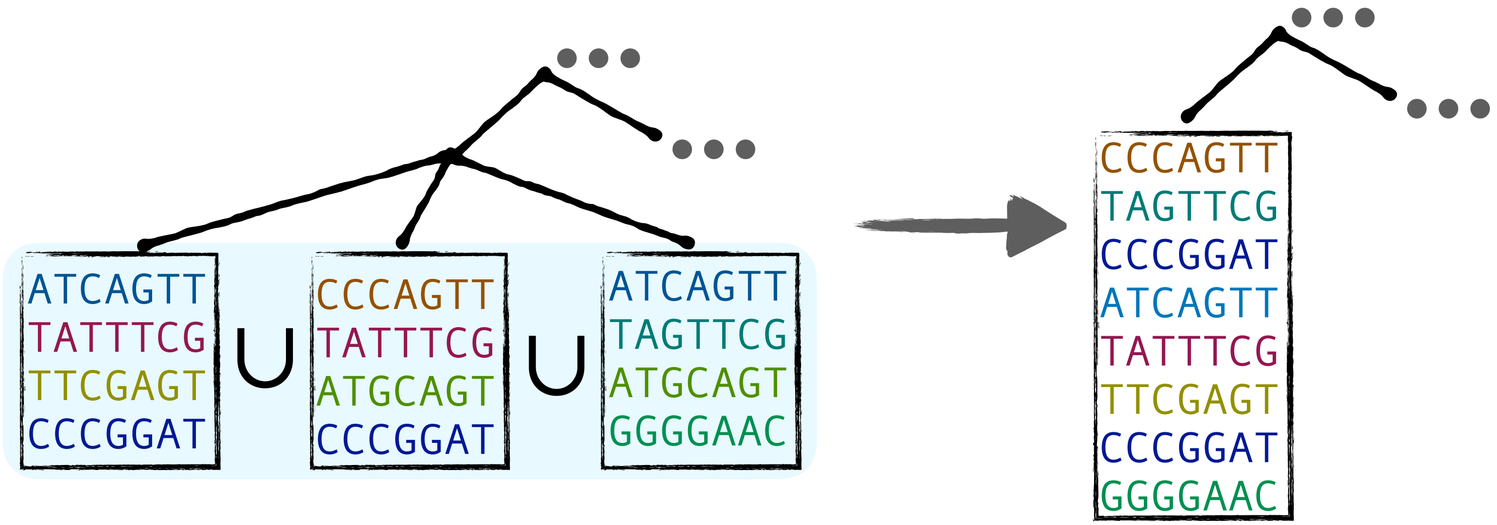


Core idea: hierarchical subsampling through a post order traversal of the taxonomic tree



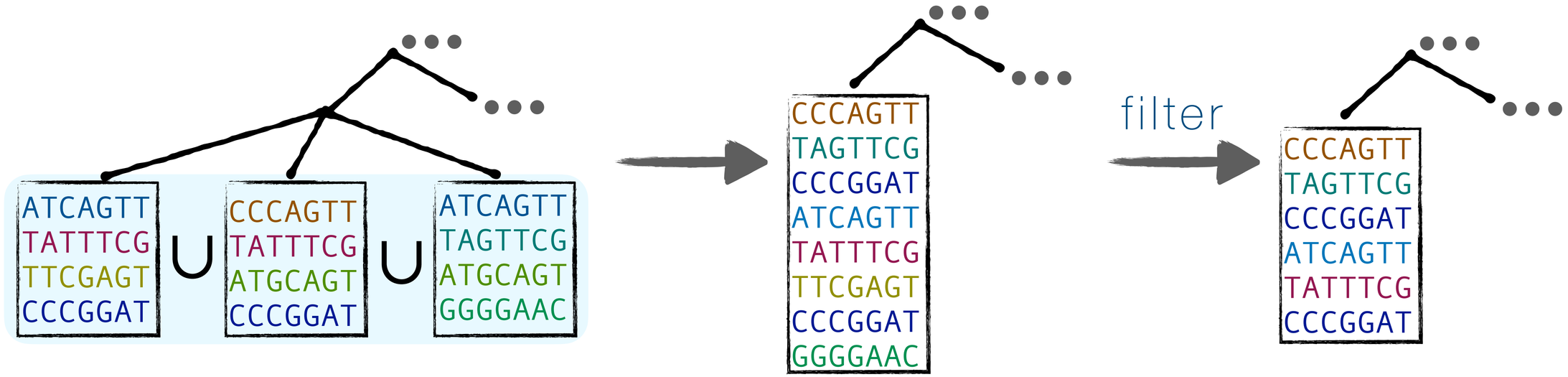
Gradual filtering of k-mers at internal nodes

- Recursively take the union of sibling taxa



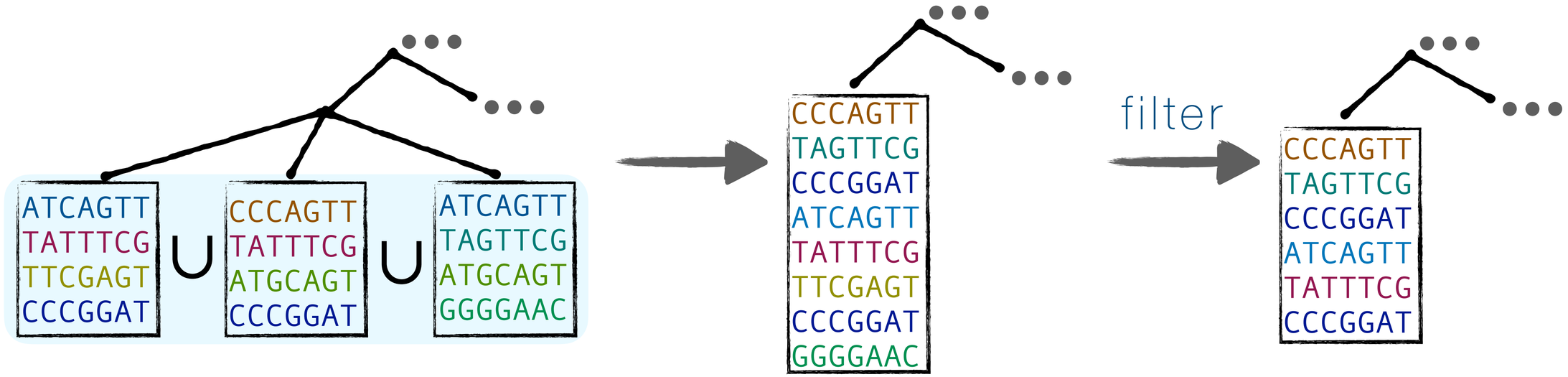
Gradual filtering of k-mers at internal nodes

- Recursively take the union of sibling taxa
- Filter *some number* of *k*-mers based on *a ranking*



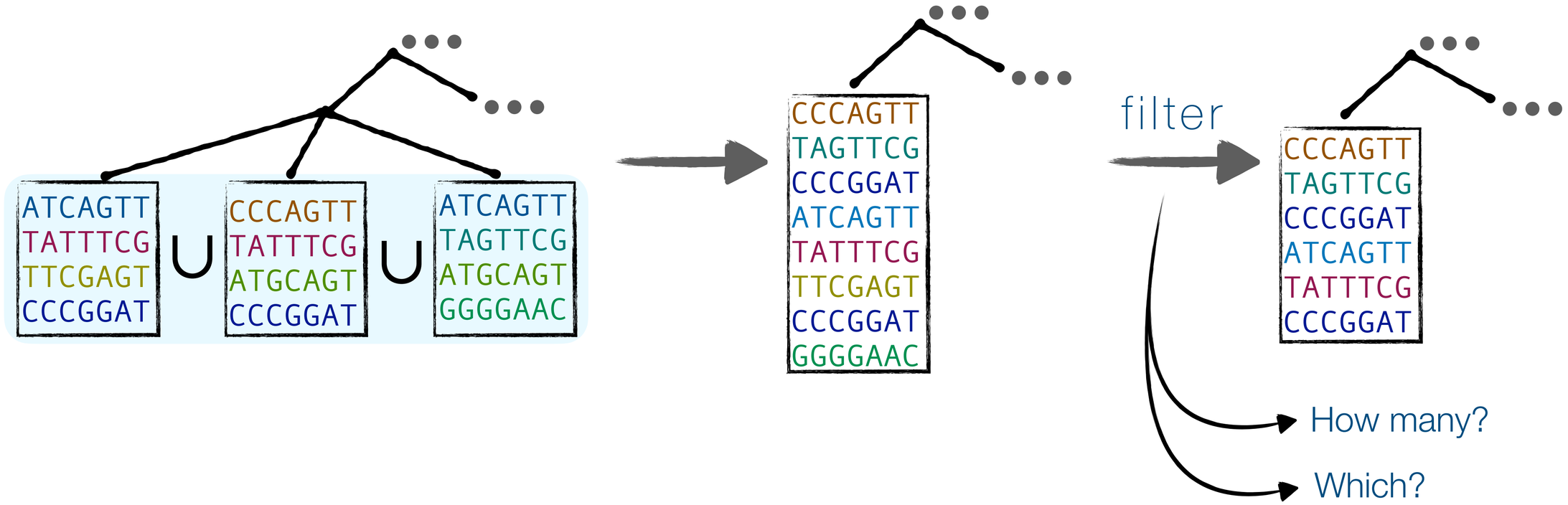
Gradual filtering of k-mers at internal nodes

- Recursively take the union of sibling taxa
- Filter *some number* of *k*-mers based on *a ranking*
- At the root, we obtain the final library with size budget M



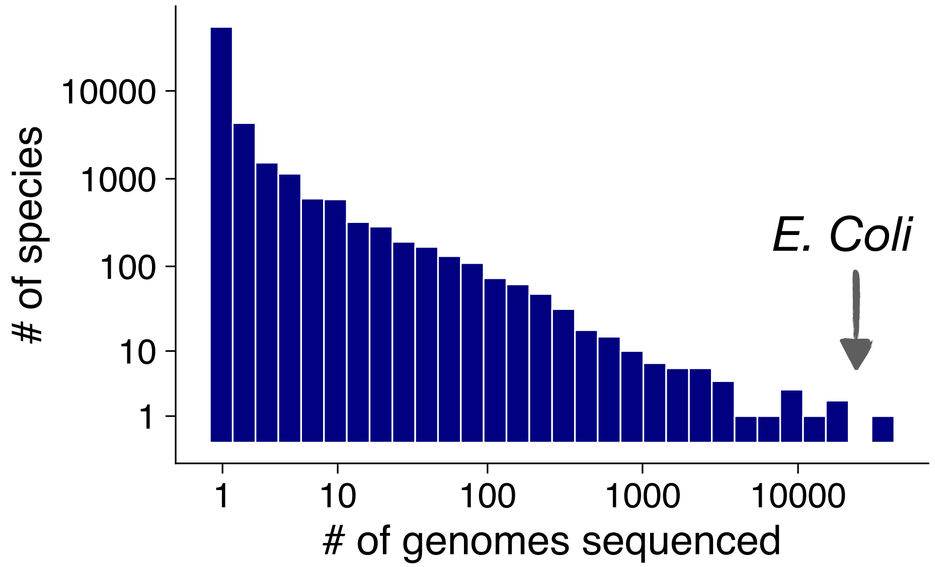
Gradual filtering of k-mers at internal nodes

- Recursively take the union of sibling taxa
- Filter *some number* of *k*-mers based on *a ranking*
- At the root, we obtain the final library with size budget M



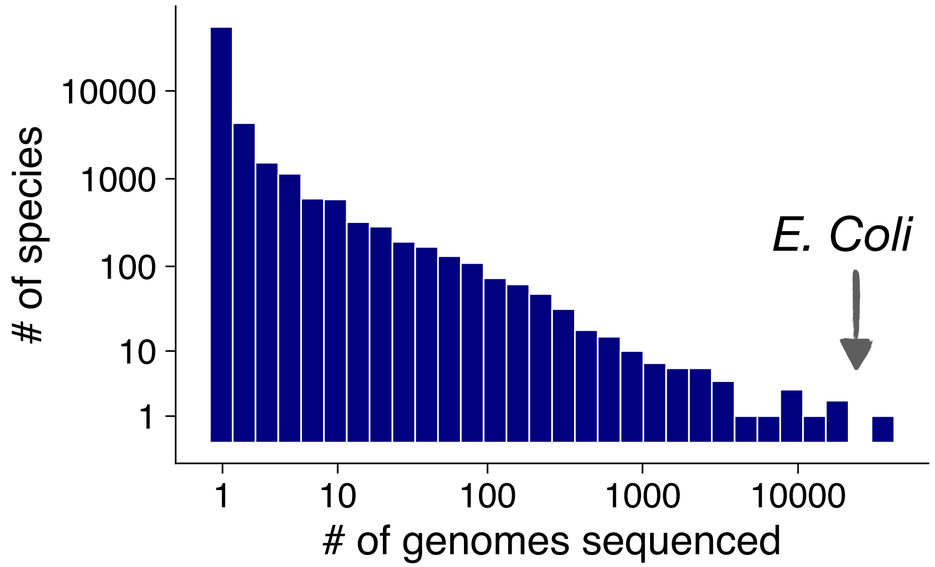
Q1: How many *k*-mers should we remove from each node/taxon?

More from highly sampled groups
Less from underrepresented groups



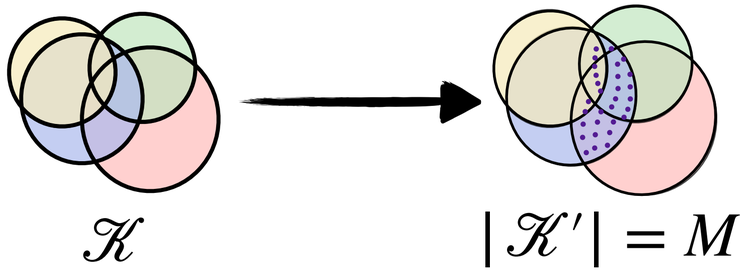
Q1: How many *k*-mers should we remove from each node/taxon?

More from highly sampled groups
Less from underrepresented groups



Q2: How do we rank *k*-mers to assess which one(s) should be kept?

Make sure to cover all taxa
Try to find informative *k*-mers



- **Baseline:** no gradual filtering — wait & select M randomly at the root

- **Baseline:** no gradual filtering — wait & select M randomly at the root

Given total budget M,
$\mathbb{E}[$# of selected k-mers for a taxon $t]$ is

$$M \frac{|\mathcal{K}_t|}{|\mathcal{K}|}$$

$|\mathcal{K}_t|$: set of k-mers under the taxon t

$|\mathcal{K}|$: set of all reference k-mers

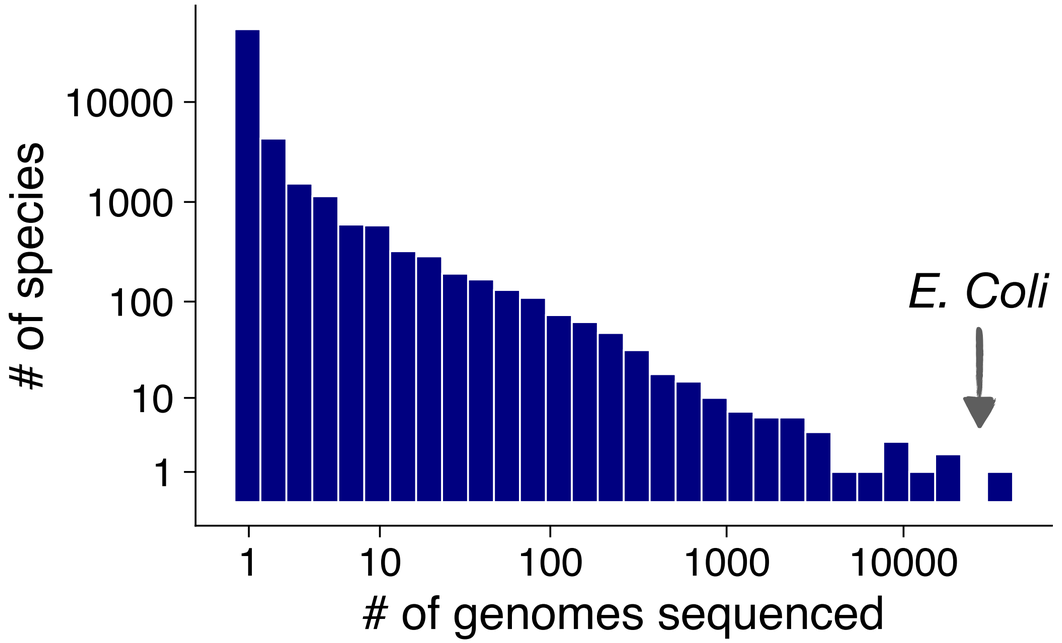
- **Baseline:** no gradual filtering — wait & select M randomly at the root

Given total budget M,
$\mathbb{E}[$# of selected k-mers for a taxon $t]$ is

$$M \frac{|\mathcal{K}_t|}{|\mathcal{K}|}$$

- $|\mathcal{K}_t|$: set of k-mers under the taxon t
- $|\mathcal{K}|$: set of all reference k-mers

- Proportional contribution →
  - taxa with low sampling get little representation
  - highly-sampled groups dominates



Gradual filtering is making some decisions earlier

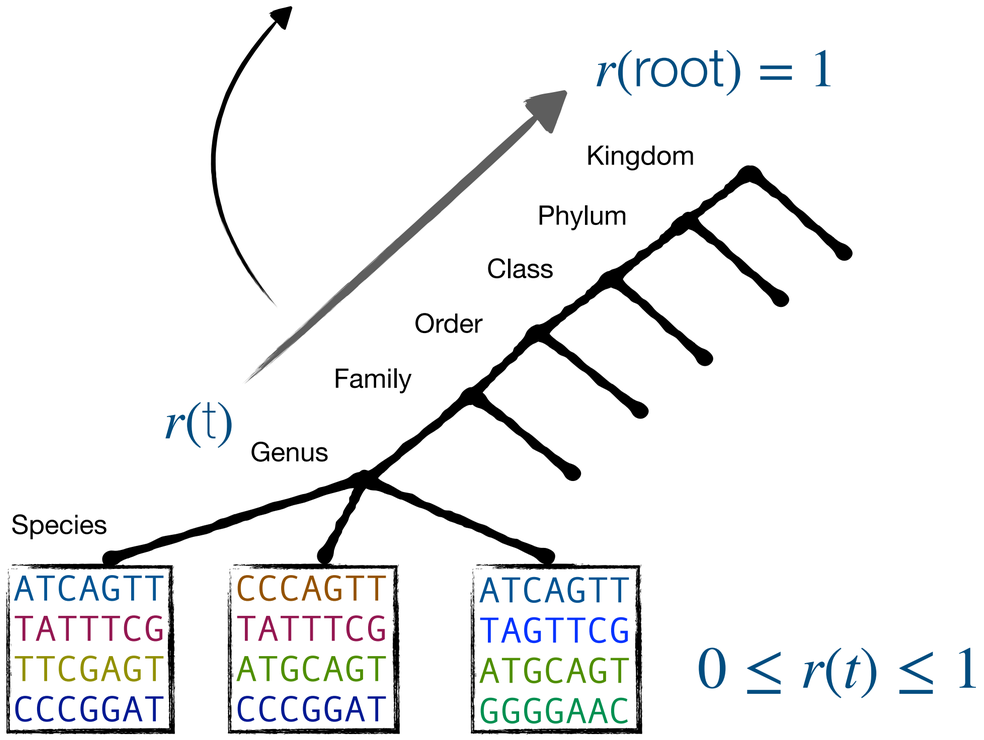
Goal: remove *k*-mers from bloated taxa earlier & delay decisions for smaller taxa

Gradual filtering is making some decisions earlier

Goal: remove *k*-mers from bloated taxa earlier & delay decisions for smaller taxa

- Adaptive size constraint, $r(t)M$,
  on internal nodes

increases as we go up in the tree!



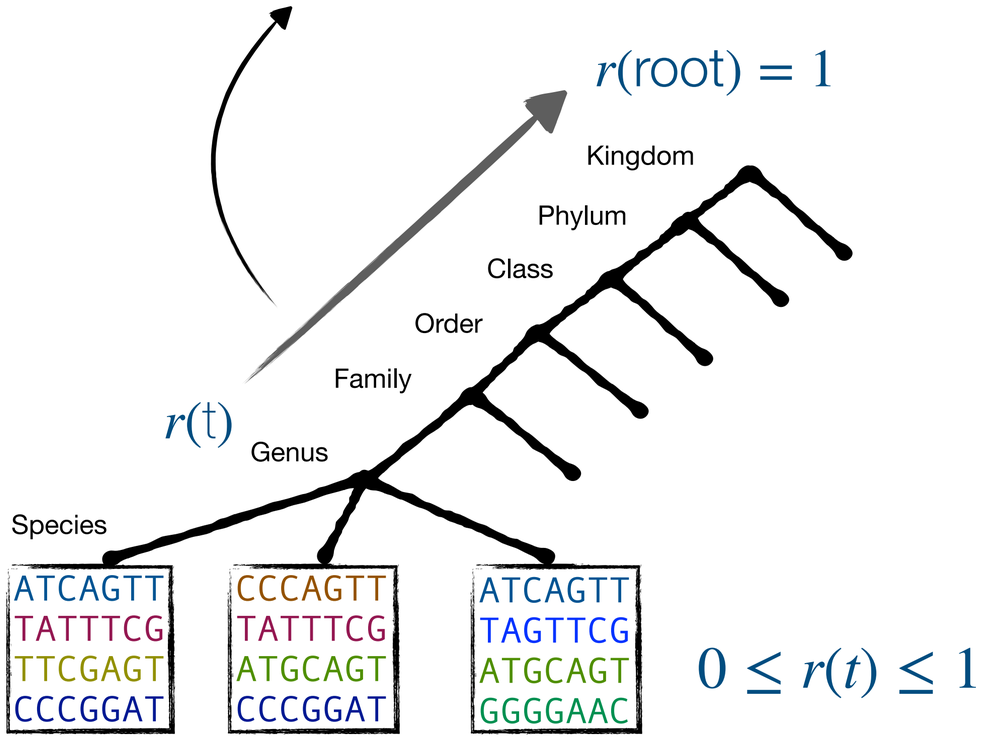
$0 \leq r(t) \leq 1$

Gradual filtering is making some decisions earlier

Goal: remove *k*-mers from bloated taxa earlier & delay decisions for smaller taxa

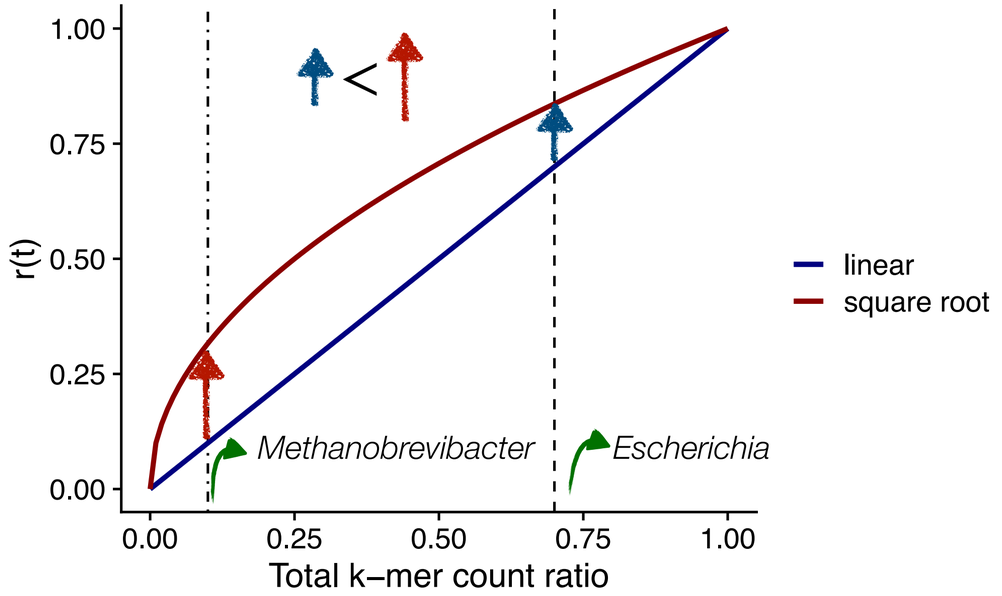
- Adaptive size constraint, $r(t)M$, on internal nodes

increases as we go up in the tree!

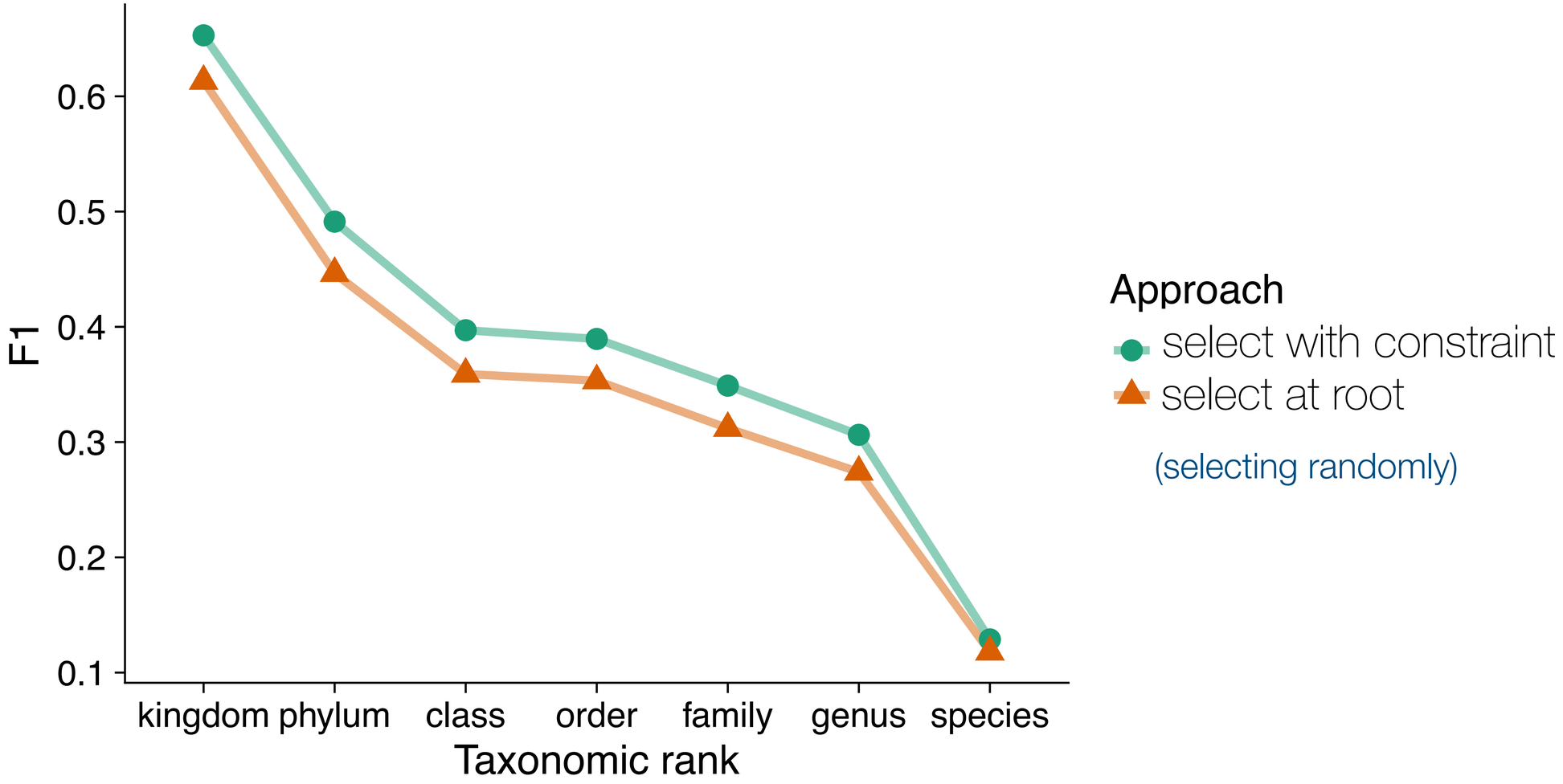


$0 \leq r(t) \leq 1$

- $r(t)$ is a heuristic:
  square root of ratio of *k*-mers under t
- Concavity of $r(t)$ favors taxa with fewer *k*-mers (less diversity or sparsely sampled)



Adaptive size constraint improves classification



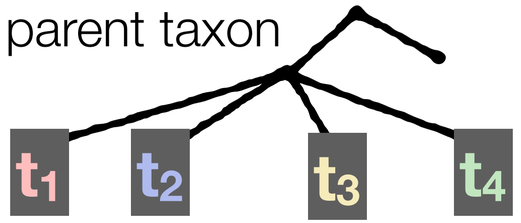
(empirical analysis using 3.2Gb, in WoL-v1 with 9k species, 10k genomes)

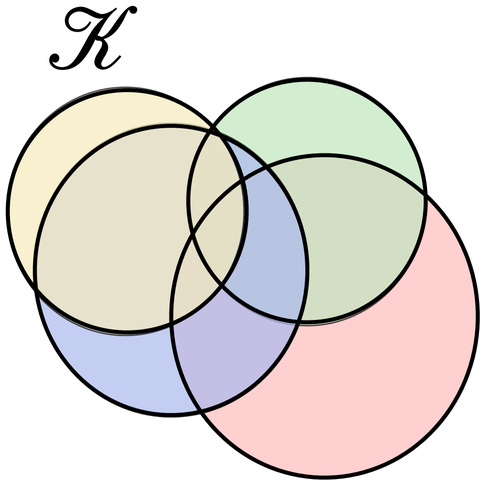
Q1: How many *k*-mers should we remove from each node/taxon?

Q2: How do we rank *k*-mers to assess which one(s) should be kept?

Which k-mers would provide better representation?

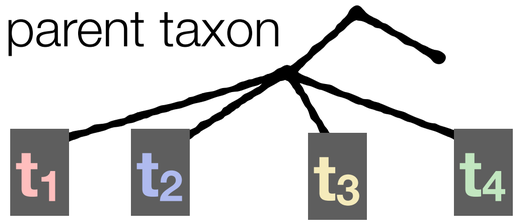
Baseline: selecting randomly until the constraint is satisfied

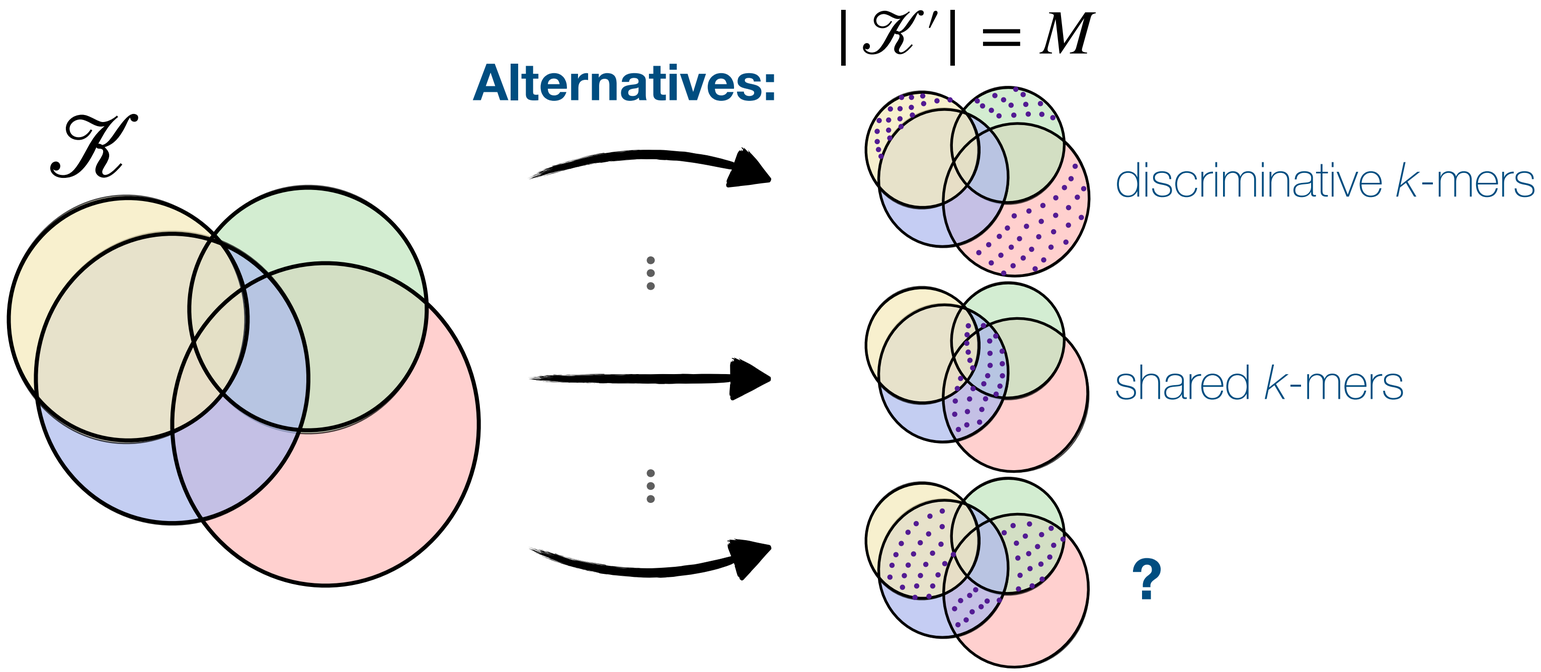




Which k-mers would provide better representation?

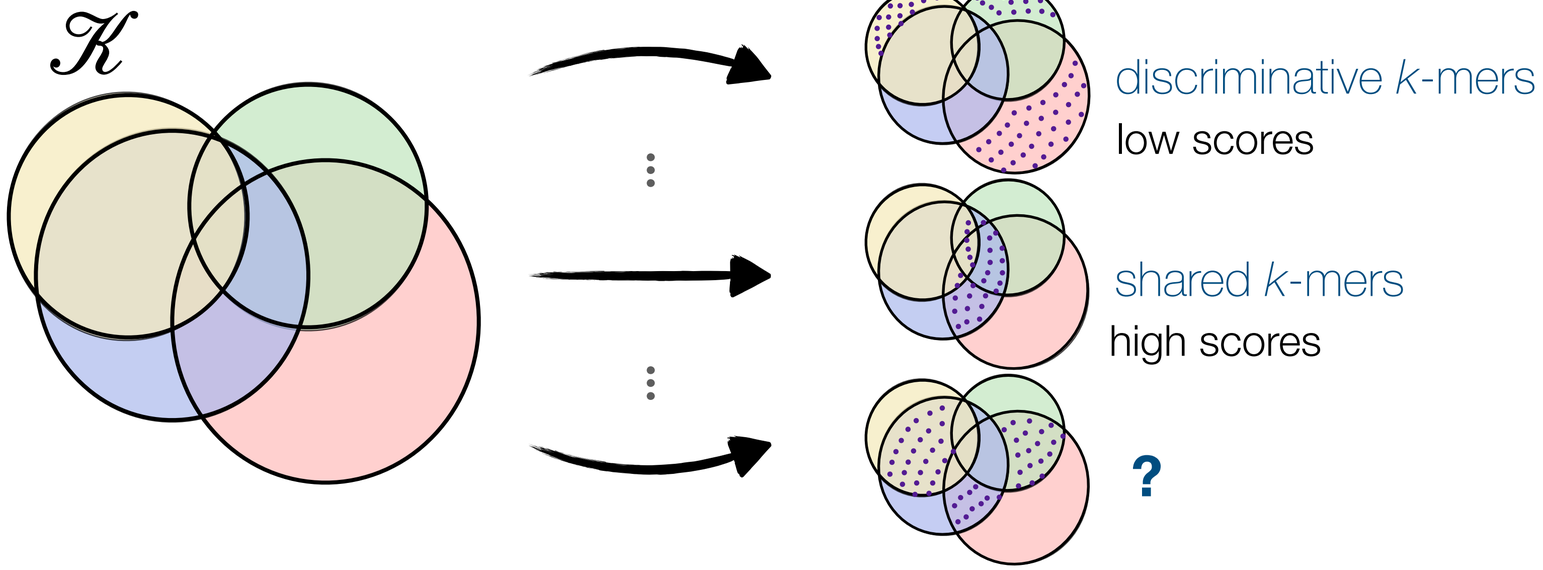
Baseline: selecting randomly until the constraint is satisfied

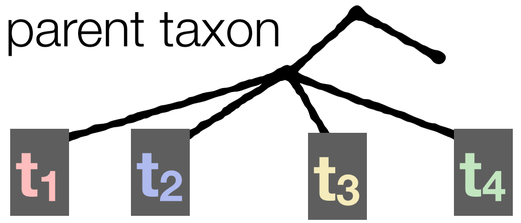




Which k-mers would provide better representation?

Baseline: selecting randomly until the constraint is satisfied





# of species under t with k-mer x

	x_1	x_2	x_3	...	$x_{\lvert\mathcal{K}\rvert}$
t_1	4	7	0	...	3
t_2	0	0	2	...	0
t_3	0	0	1	...	1
t_4	2	2	1	...	0
Score:	**6**	**9**	**4**	**...**	**4**

The case against discriminative k-mers

- **Problem:** considerably small portion of *k*-mers are shared within a group! (it gets worse for upper ranks)

Given a query genome, what is the expected portion of shared *k*-mers in a reference set with N genomes within $2d$ distance?

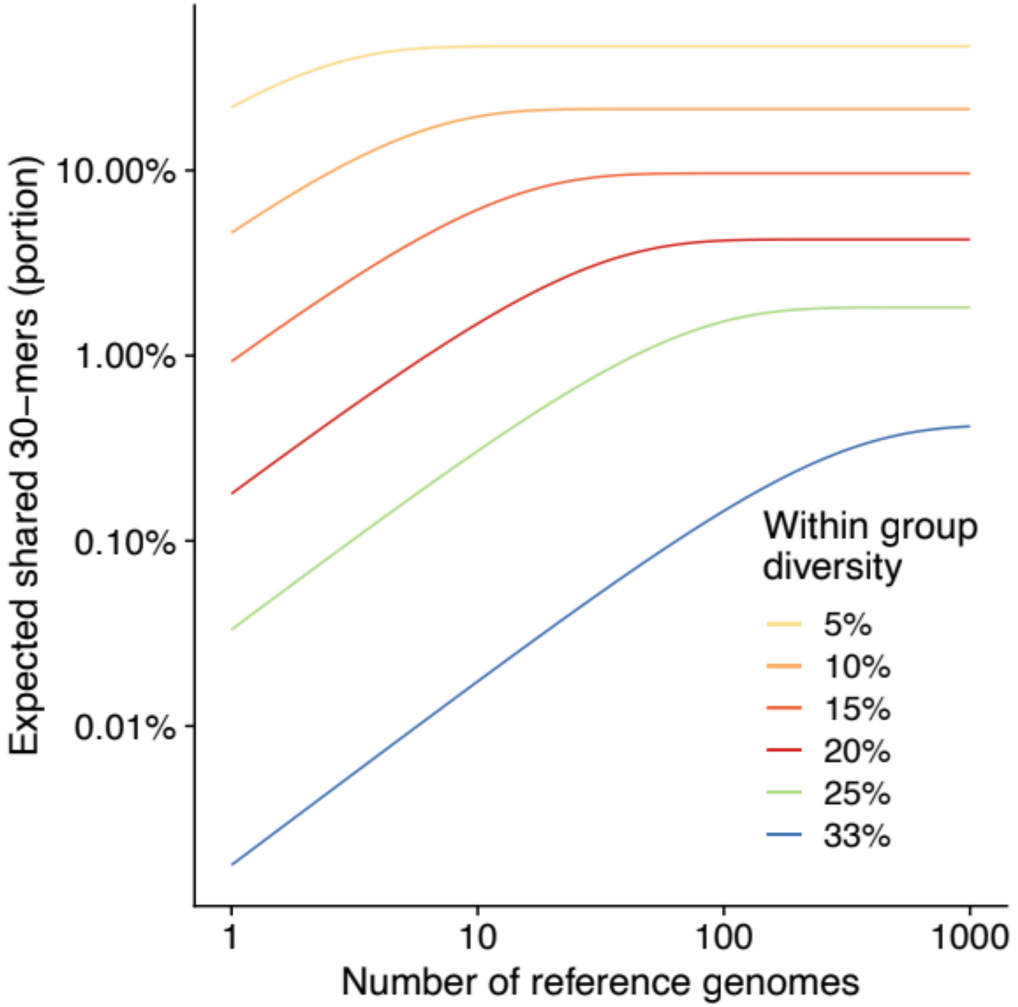
$$(1-d)^k\left(1-(1-(1-d)^k)^N\right)$$

- $(1-d)^k$: *k*-mer from the ancestor stays same
- $(1-(1-d)^k)^N$: *k*-mer from the ancestor changes in all N

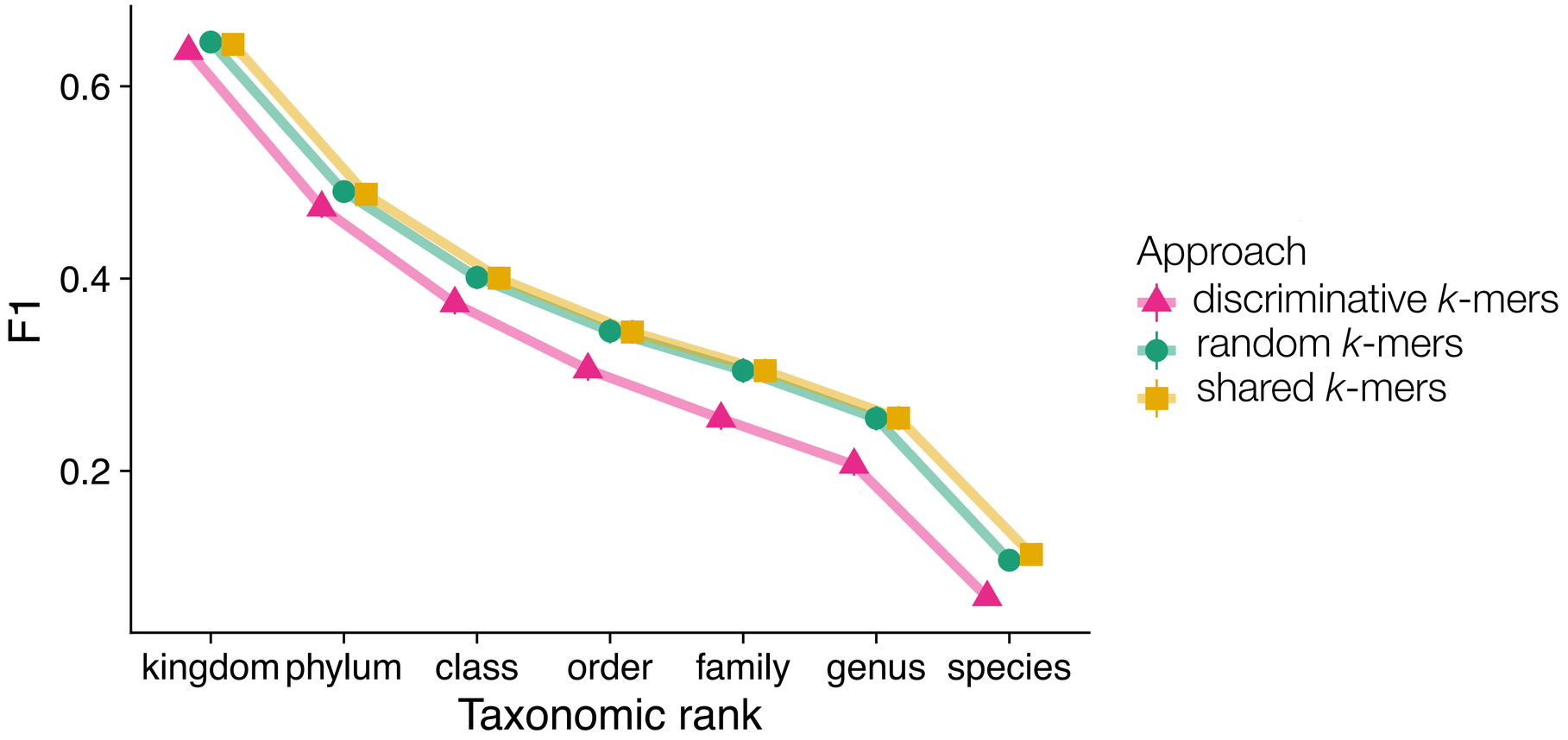
Example: within d = 20% diversity (~genus)

- $N = 5$: 0.7% of query 30-mers,
- $N \to \infty$: 4.2% of query 30-mers,

will be found in at least one reference.



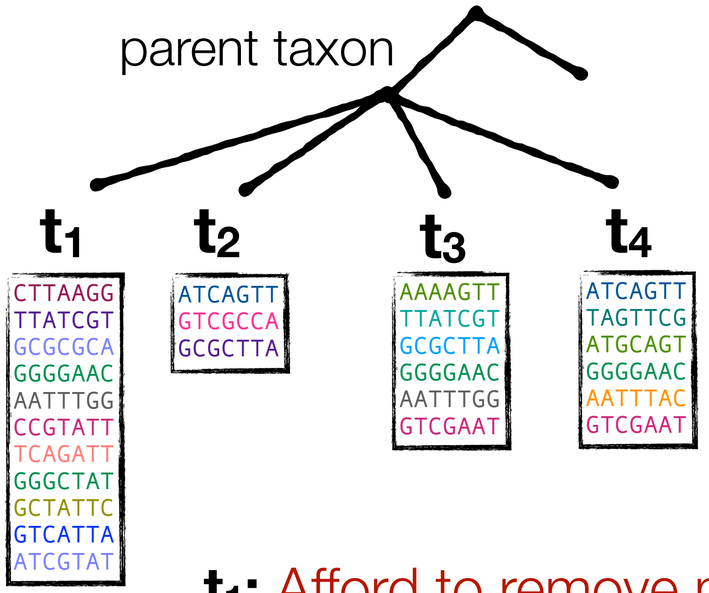
Neither discriminative nor shared k-mers improve the baseline



(empirical analysis using 3.2Gb, in WoL-v1 with 9k species, 10k genomes)

Incorporating taxon coverage in ranking

Intuition: keep shared *k*-mers but ensure no group is left uncovered



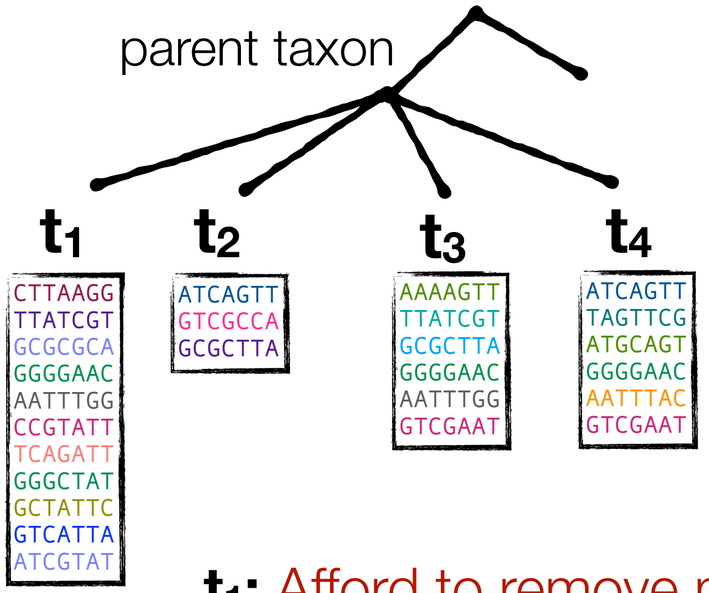
t_1: Afford to remove more!

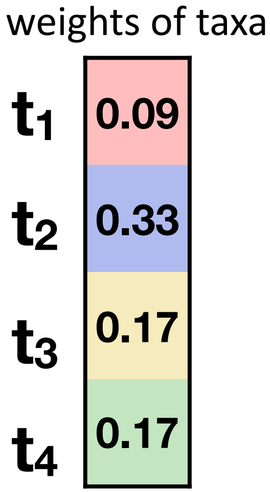
t_2: Needs to be prioritized!

Incorporating taxon coverage in ranking

Intuition: keep shared *k*-mers but ensure no group is left uncovered

Scalable heuristic: down-weight the impact of taxa that are highly covered among surviving k-mers



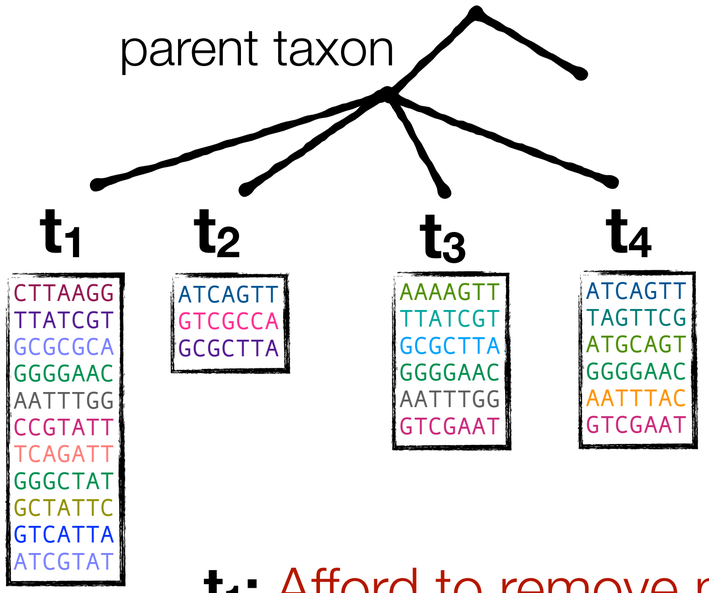


t_1: Afford to remove more!

t_2: Needs to be prioritized!

Incorporating taxon coverage in ranking

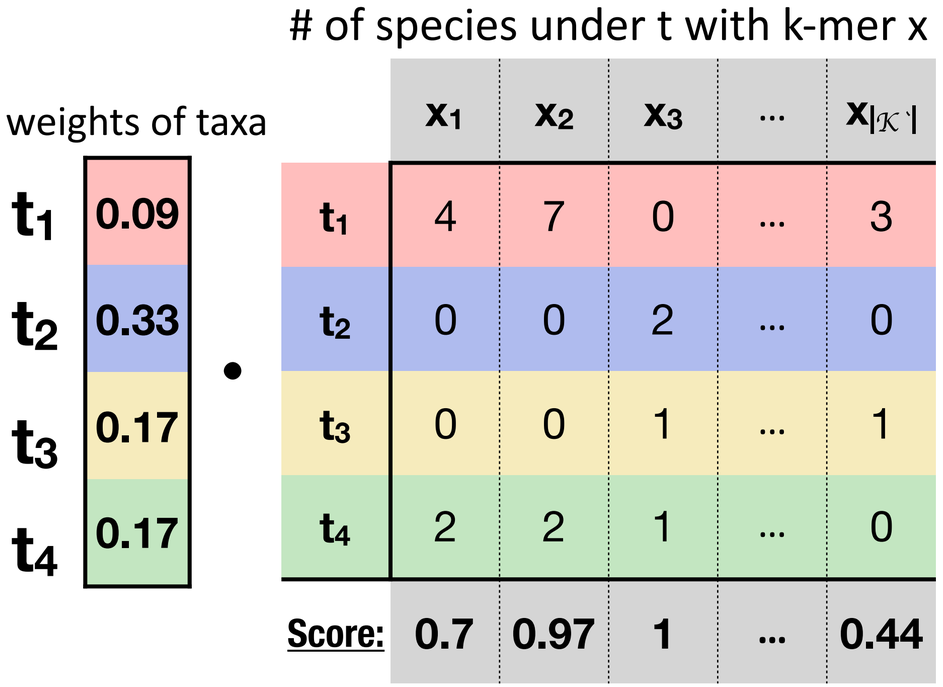
Intuition: keep shared *k*-mers but ensure no group is left uncovered



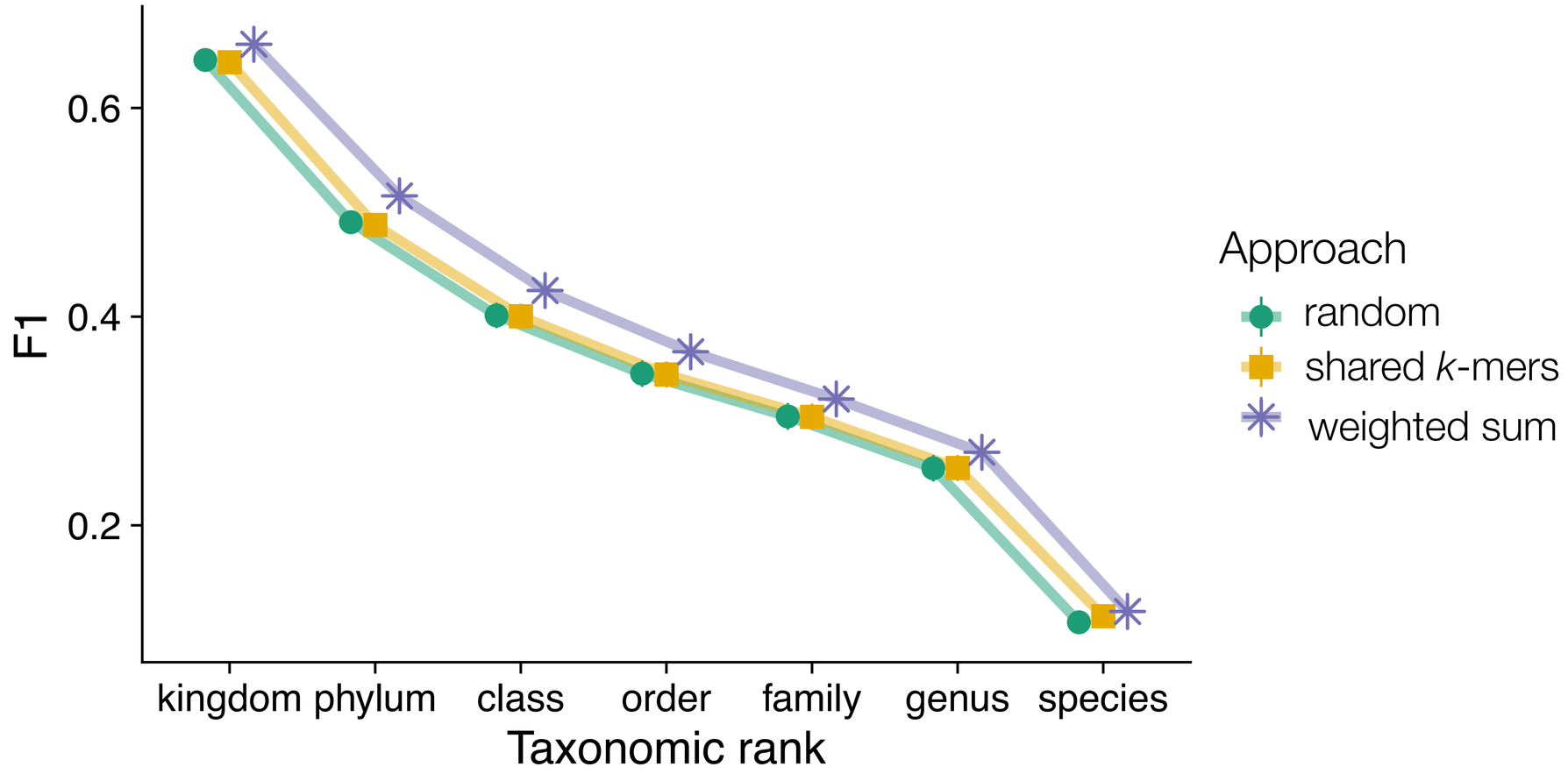
t_1: Afford to remove more!

t_2: Needs to be prioritized!

Scalable heuristic: down-weight the impact of taxa that are highly covered among surviving k-mers



Neither discriminative nor shared k-mers improve the baseline



(empirical analysis using 3.2Gb, in WoL-v1 with 9k species, 10k genomes)

- **KRANK** puts all these heuristics together:
  - weighted-sum ranking + adaptive size constraint
  - other minor tricks
  - highly optimized and scalable implementation

Bonus: compact k-mer encodings

CONSULT-II used 2 bits per letter: 64bit for 32-mers.

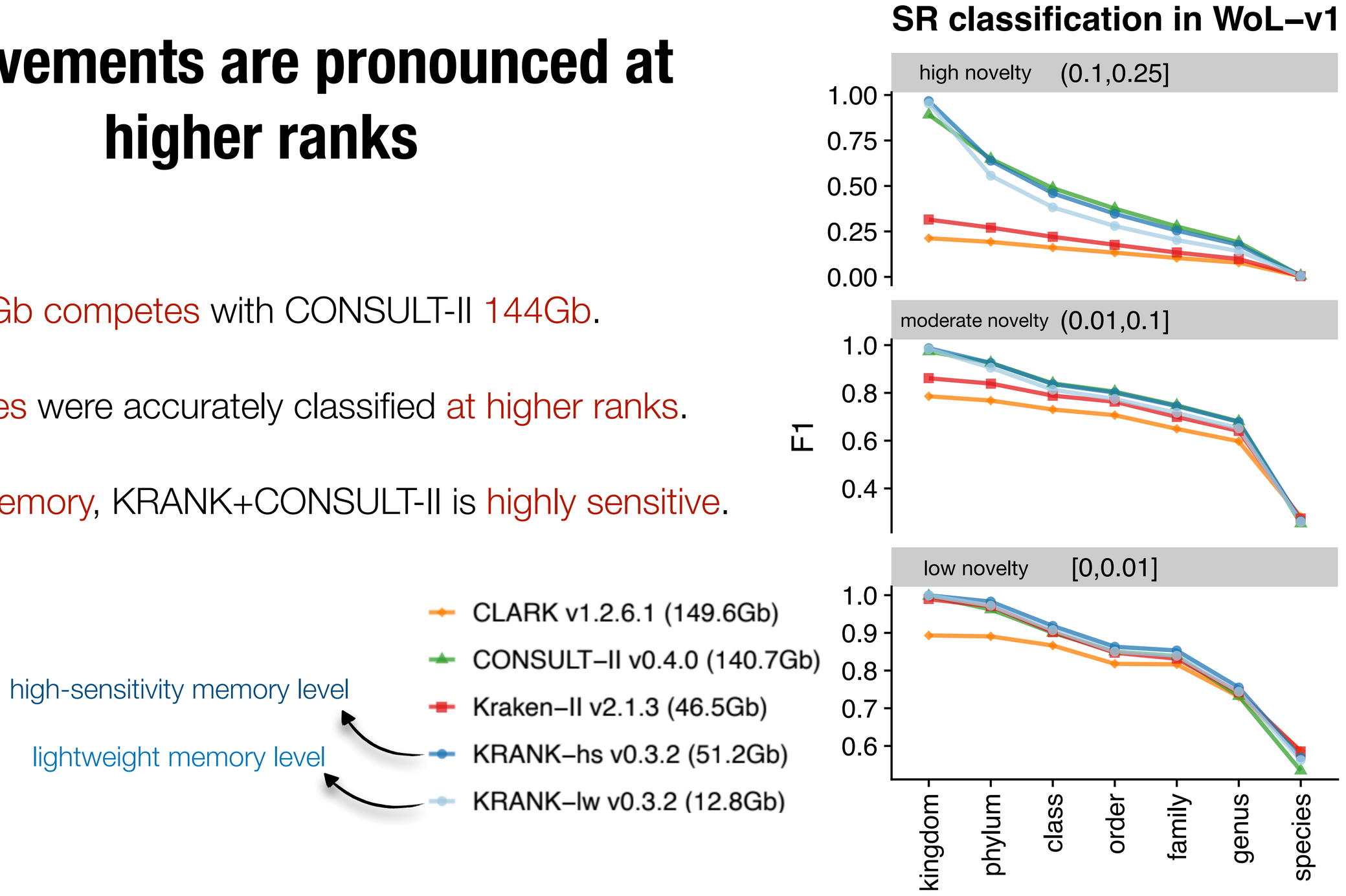
We only compute HD between *k*-mers that have the same hash value!

We do not need h positions used to compute LSH; they are already the same!

Just drop LSH positions and store the rest: $k = 32, h = 16 \rightarrow$ 32bit

Improvements are pronounced at higher ranks

- KRANK 13Gb competes with CONSULT-II 144Gb.
- Novel queries were accurately classified at higher ranks.
- With little memory, KRANK+CONSULT-II is highly sensitive.



krepp

Weighted UniFrac

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.

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

pseudo-F statistic

$$SS_{\text{total}} = \frac{1}{N}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^2 \qquad SS_{\text{within}} = \frac{1}{N}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^2 \delta_{ij}$$

$$SS_{across} = SS_{\text{total}} - SS_{\text{within}}$$

$$F = \frac{\left(\frac{SS_{\text{across}}}{p-1}\right)}{\left(\frac{SS_{\text{within}}}{N-p}\right)}$$

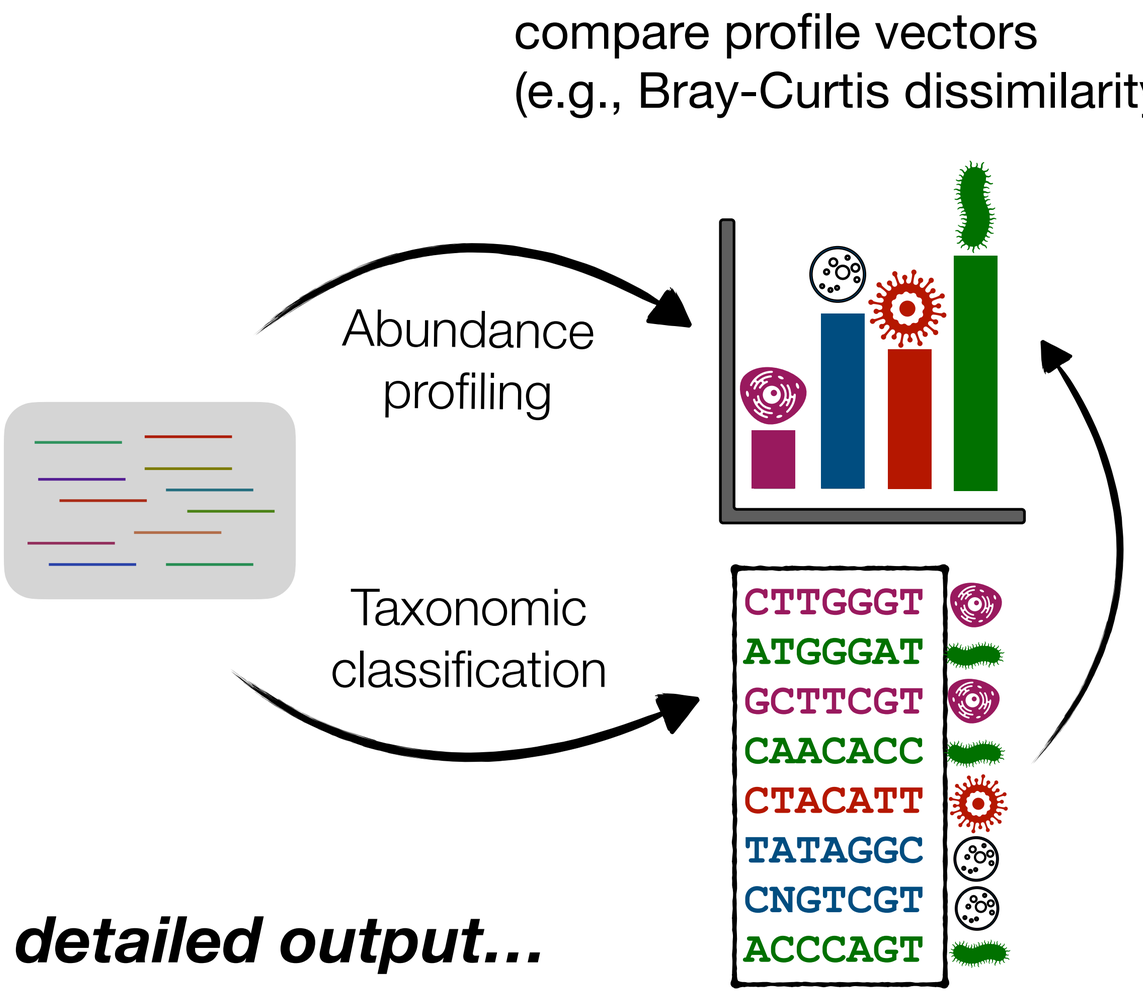
Multiple permutations to get a p-value!

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

Poor man's solution: taxonomic profiling

- **Fast and scalable methods**
  based on *k*-mer search

- **Limited:**
  - has low resolution
  - often ambiguous (e.g., HGT)
  - omits within-group diversity
  - no notion of *distance*, novelty?



We need a more detailed output…

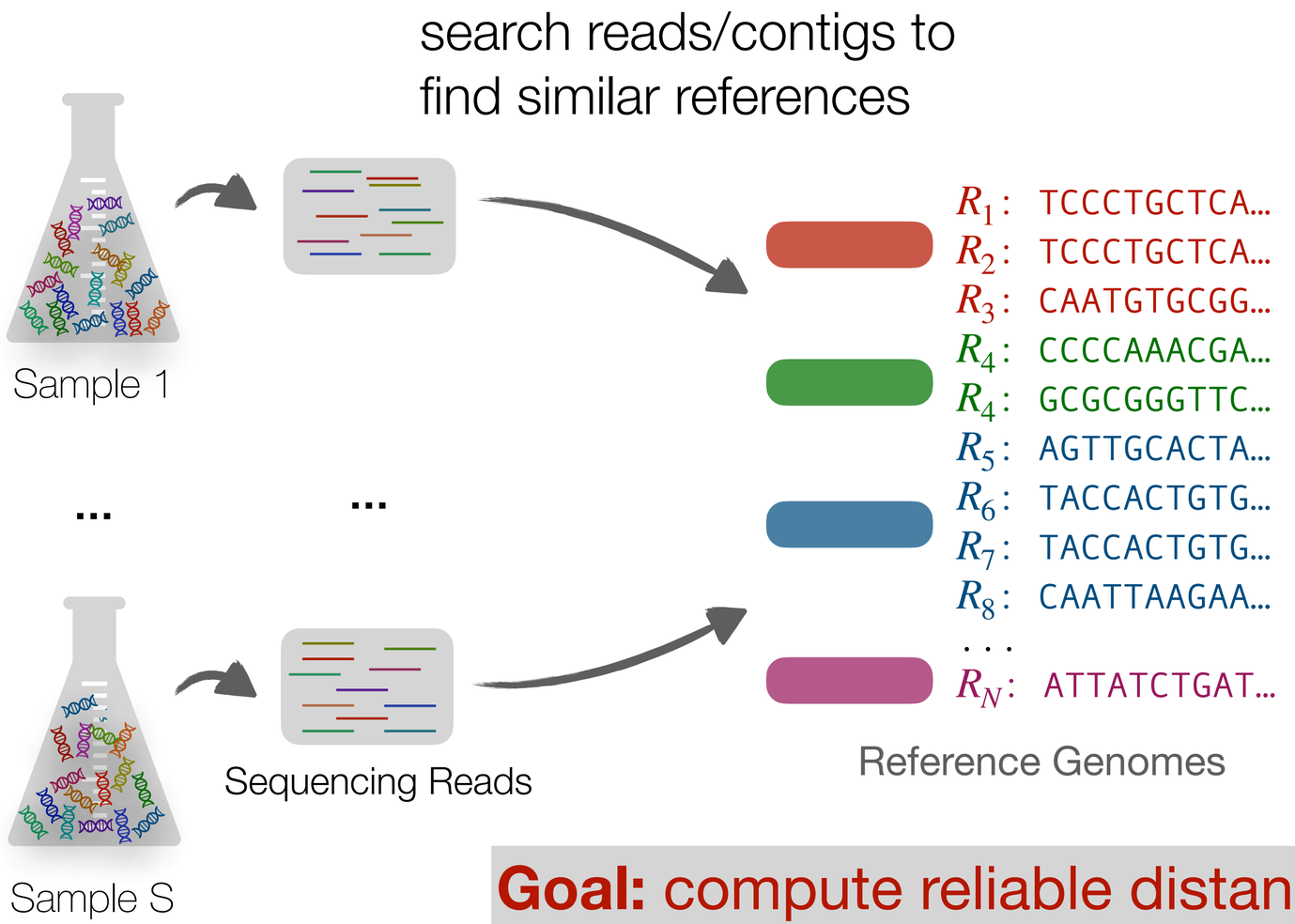
What alternatives do we have?

- **Sample-wide containment analysis (using MinHash)?**
  - higher resolution compared to taxonomic profiles **(+)**
  - no assignments for individual sequences **(-)**
  - no distances btw. queries and references **(-)**

- **Using marker genes?**
  - aligning to a MSA is possible & and distances **(+)**
  - limited and potentially biased **(-)**
  - inability to capture novel sequences and queries **(-)**

- **Aligning reads to all references?**

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…

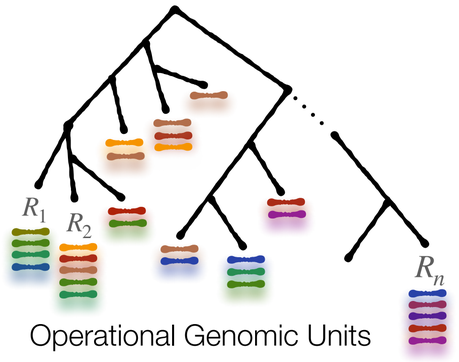
Existing Pipelines

align reads to
reference genomes

sample

R_1 : R_2 : ... R_n :

assign reads to
closely matching tips



Operational Genomic Units

alternatively

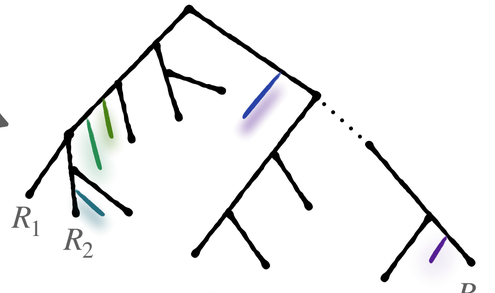
focus only on
marker genes

e.g., 16S reads

align against the
reference MSA

R_1 : AGACTTTGATCCTGGCTC
R_2 : AGACTAAGATCGTGGGTC
...
R_n : AGAGTAAGATCTTGGGTC

ML placement with
pendant branches

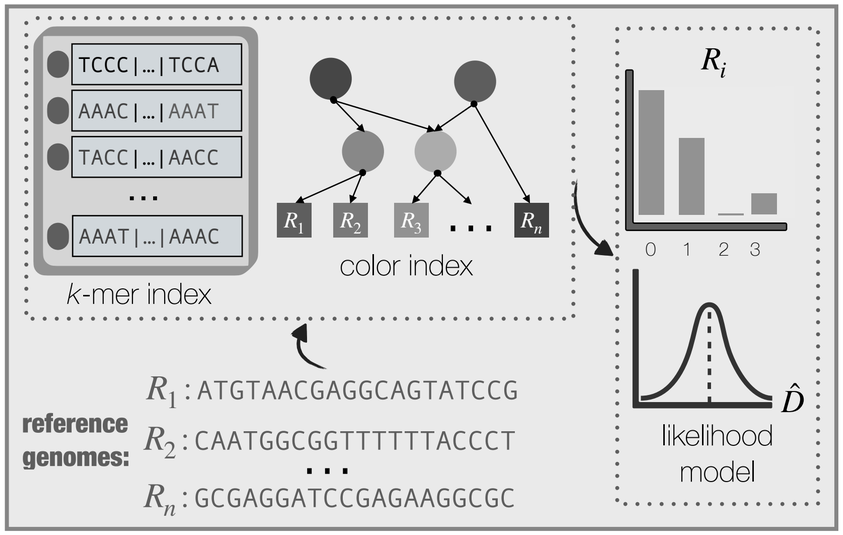


Phylogenetic Placements

krepp

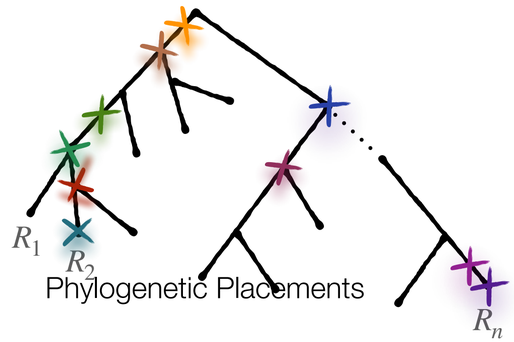
compute HD of
k-mer matches

sample



i) estimate distances of
reads to *all* references

R_1	R_2	...	R_n
0.01	0.02	...	NA
0.03	0.12	...	0.06
NA	0.05	...	0.02
0.18	0.04	...	0.21
0.02	NA	...	0.05
NA	0.19	...	0.01
NA	NA	...	0.02



Phylogenetic Placements

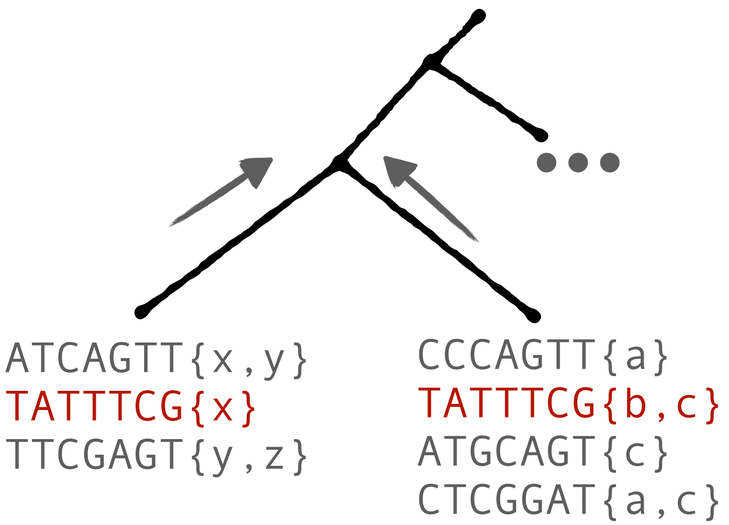
ii) place reads anywhere on
the backbone phylogeny

Constructing the color multi-tree

- Start with *k*-mer sets of all reference
- Initialize the multi-tree as unconnected singletons
- Label *k*-mers with singletons

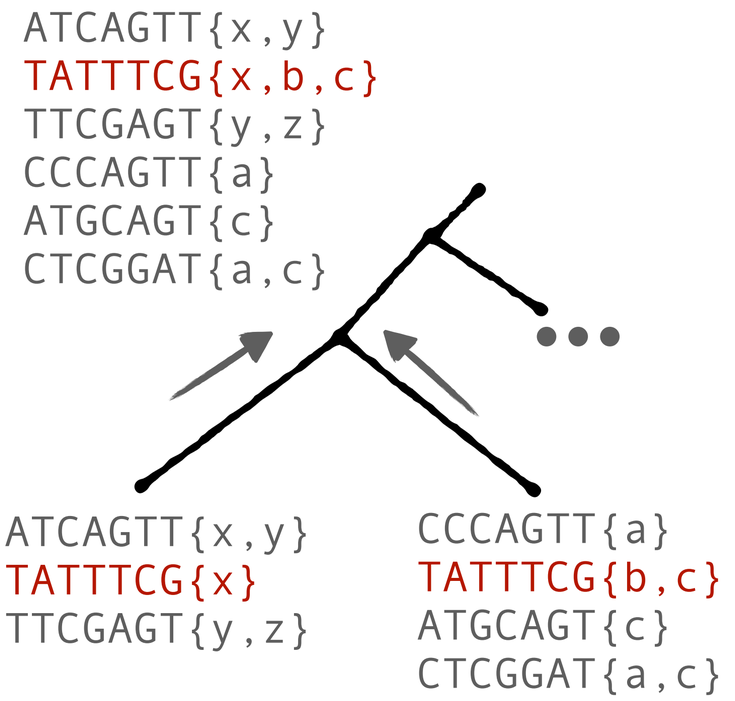
Constructing the color multi-tree

- Start with *k*-mer sets of all reference
- Initialize the multi-tree as unconnected singletons
- Label *k*-mers with singletons
- During the traversal, add a color for the union if a *k*-mer exist in both children and update its label



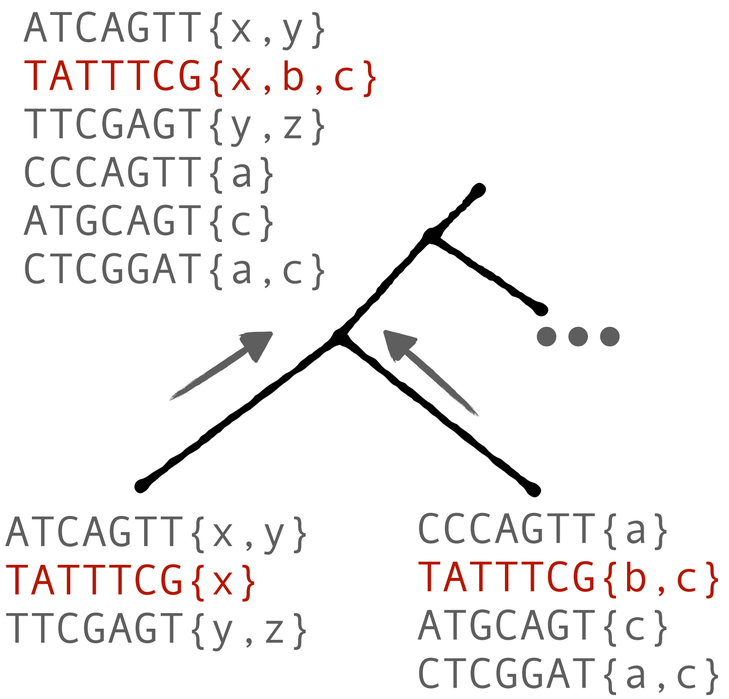
Constructing the color multi-tree

- Start with *k*-mer sets of all reference
- Initialize the multi-tree as unconnected singletons
- Label *k*-mers with singletons
- During the traversal, add a color for the union if a *k*-mer exist in both children and update its label



Constructing the color multi-tree

- Start with *k*-mer sets of all reference
- Initialize the multi-tree as unconnected singletons
- Label *k*-mers with singletons
- During the traversal, add a color for the union if a *k*-mer exist in both children and update its label



Challenges:

- How to intersections of large sets? **LSH partitions**
- How to test if a color is already added? **Abelian group hashing**
- How to represent set labels of *k*-mers? **Abelian group hashing**

krepp estimates distances accurately at the read-level

default: 29-mer
minimizers of 35-mers

~150 bp short reads

(Hamming distance) / (seq. length)

- Simulation experiments
  (true read distances)

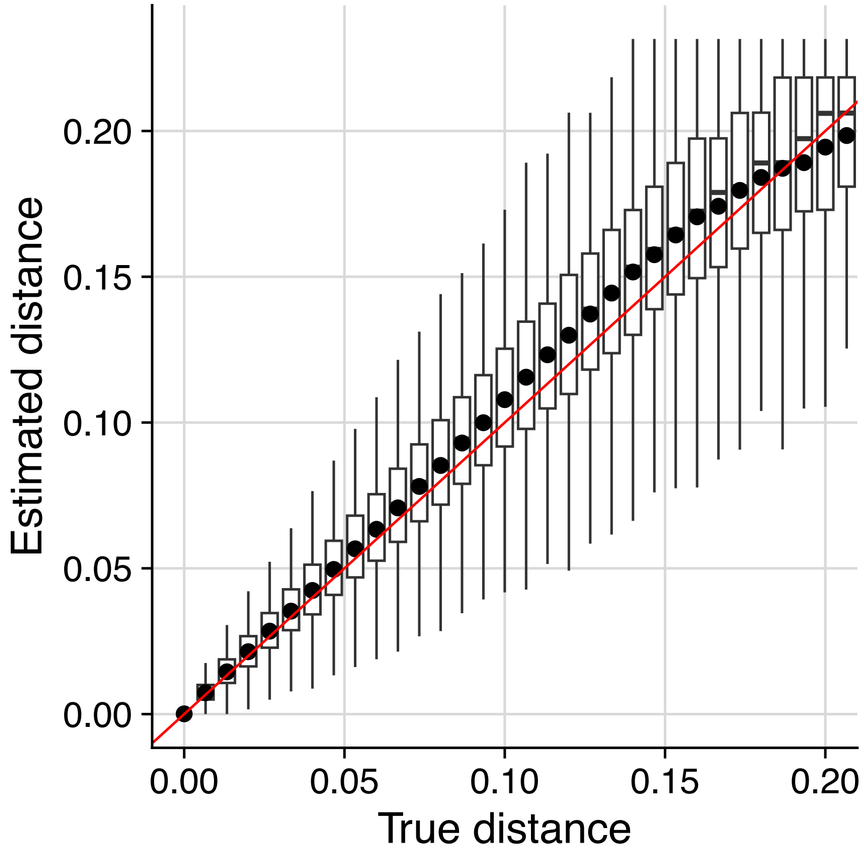
krepp estimates distances accurately at the read-level

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)



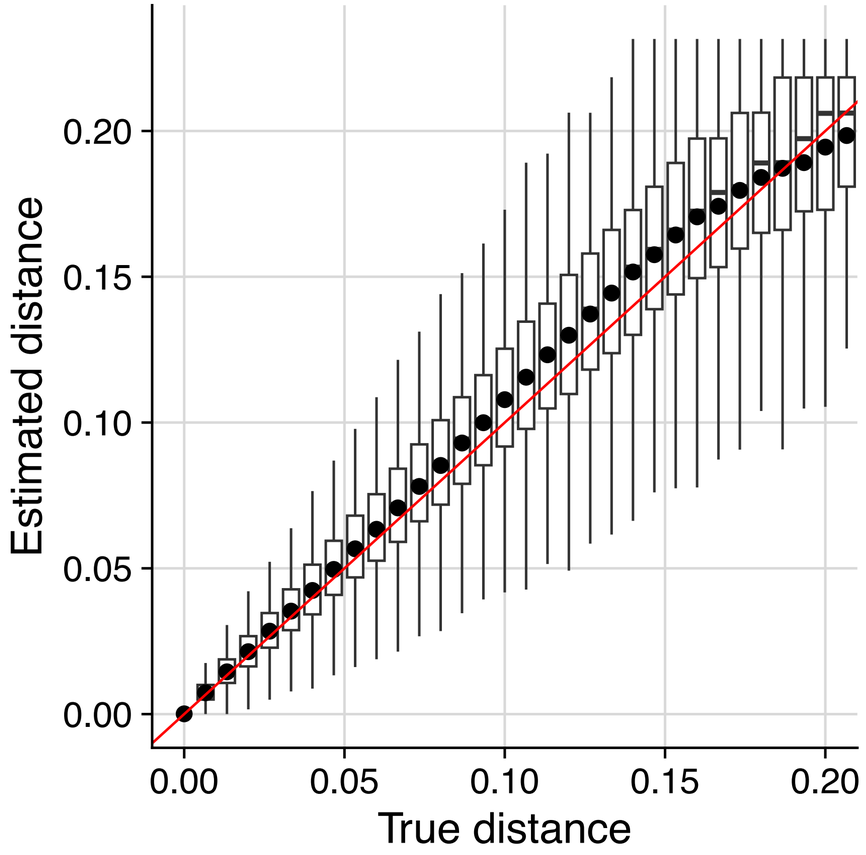
krepp estimates distances accurately at the read-level

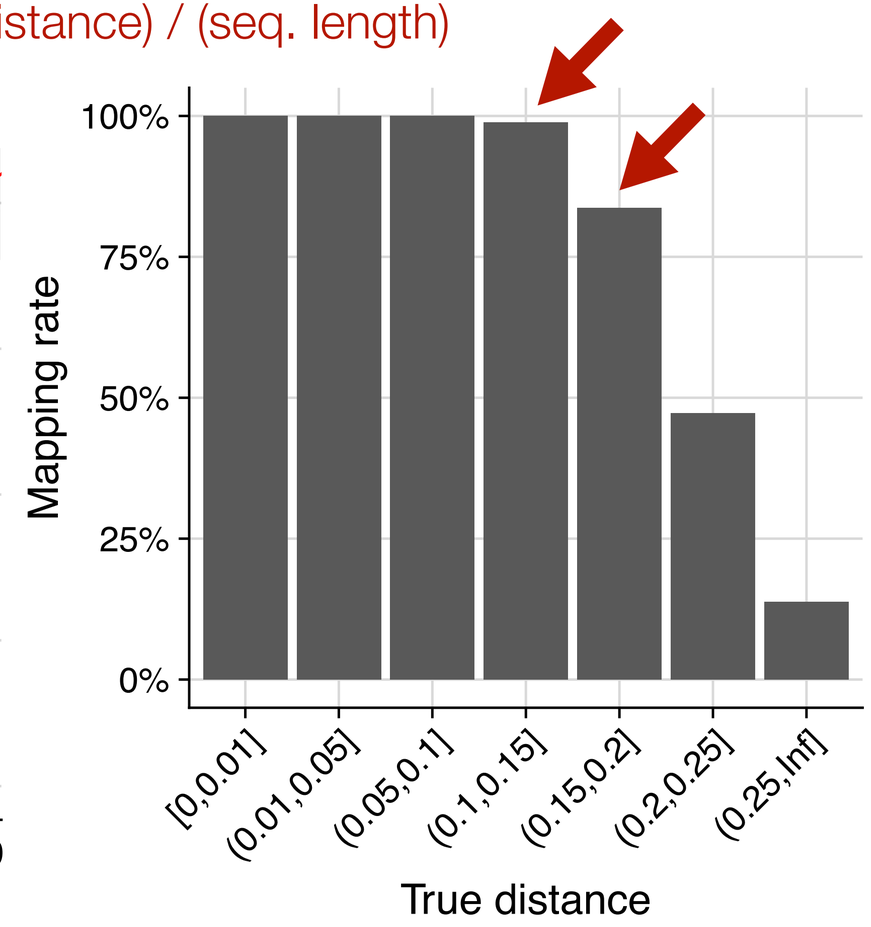
default: 29-mer
minimizers of 35-mers

- 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)





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:

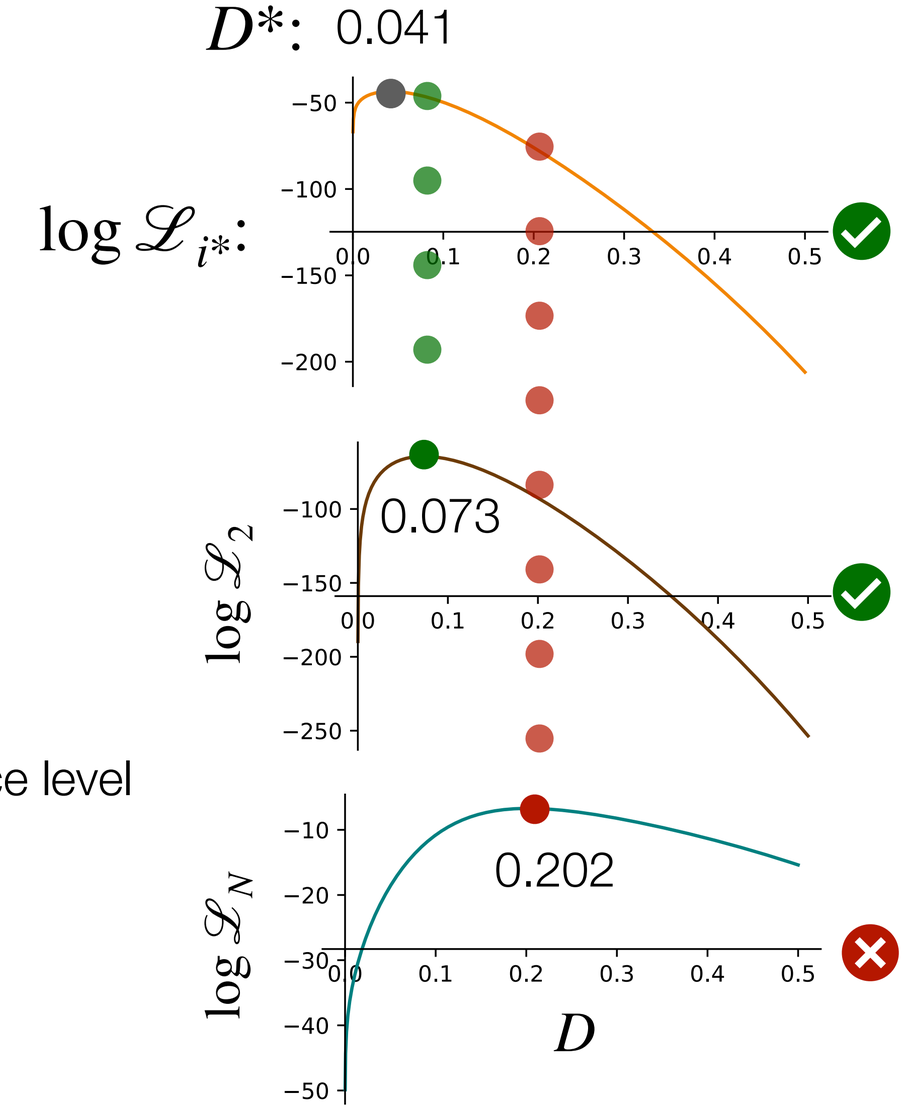
$$\lambda_{LR} = \frac{\mathscr{L}_{i*}(D; k, h, \delta, u_{i*}, \mathbf{v}_{i*})}{\mathscr{L}_{i*}(D^*; k, h, \delta, u_{i*}, \mathbf{v}_{i*})}$$

D: alternative distance

$i*$: closest reference

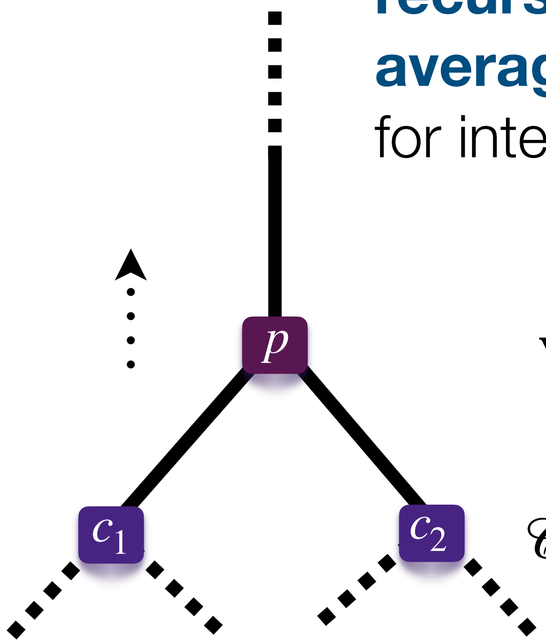
$$\lambda_{LR} \sim \chi^2$$

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



Defining clade distances & branch length agnostic placement

A notion of distance for internal nodes:



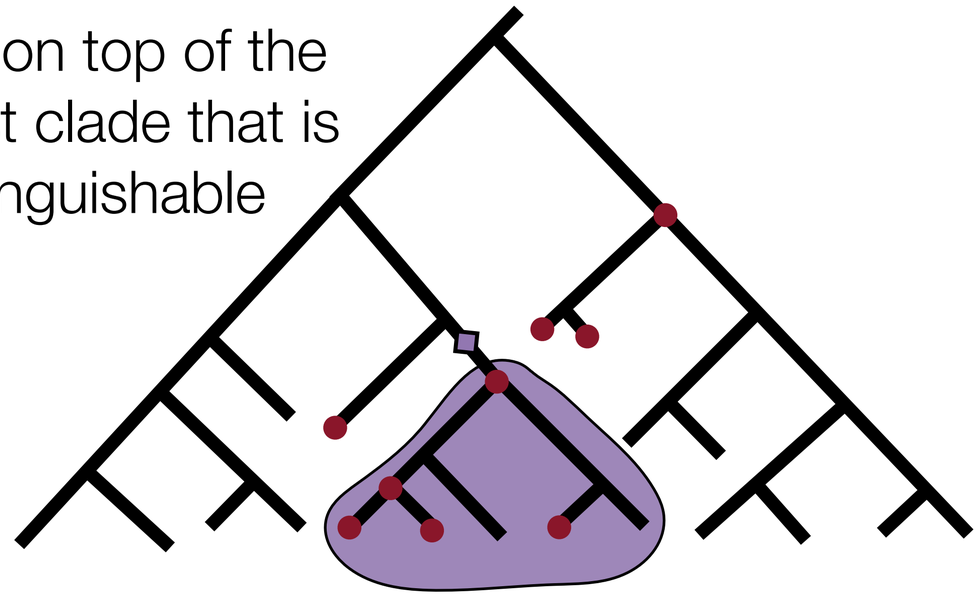
recursively compute average HD histograms for internal nodes

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

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

use the same likelihood model and log-likelihood ratio test

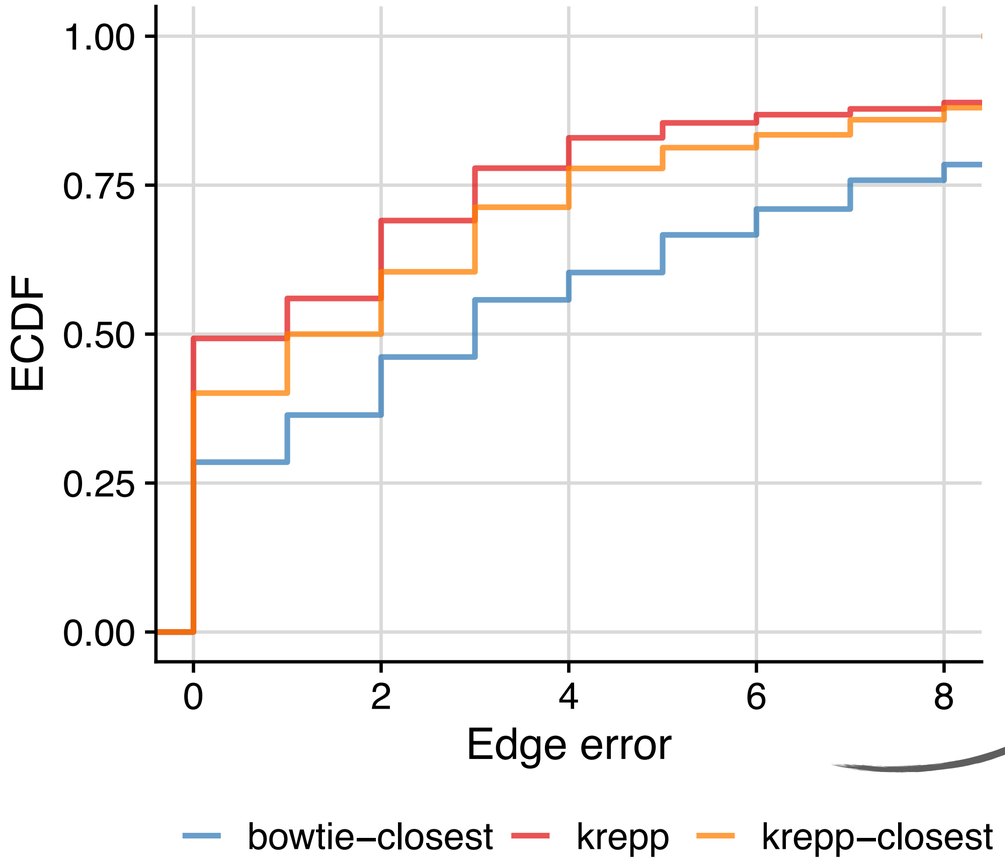
place on top of the largest clade that is indistinguishable



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

krepp's heuristic improves closest-tip placement

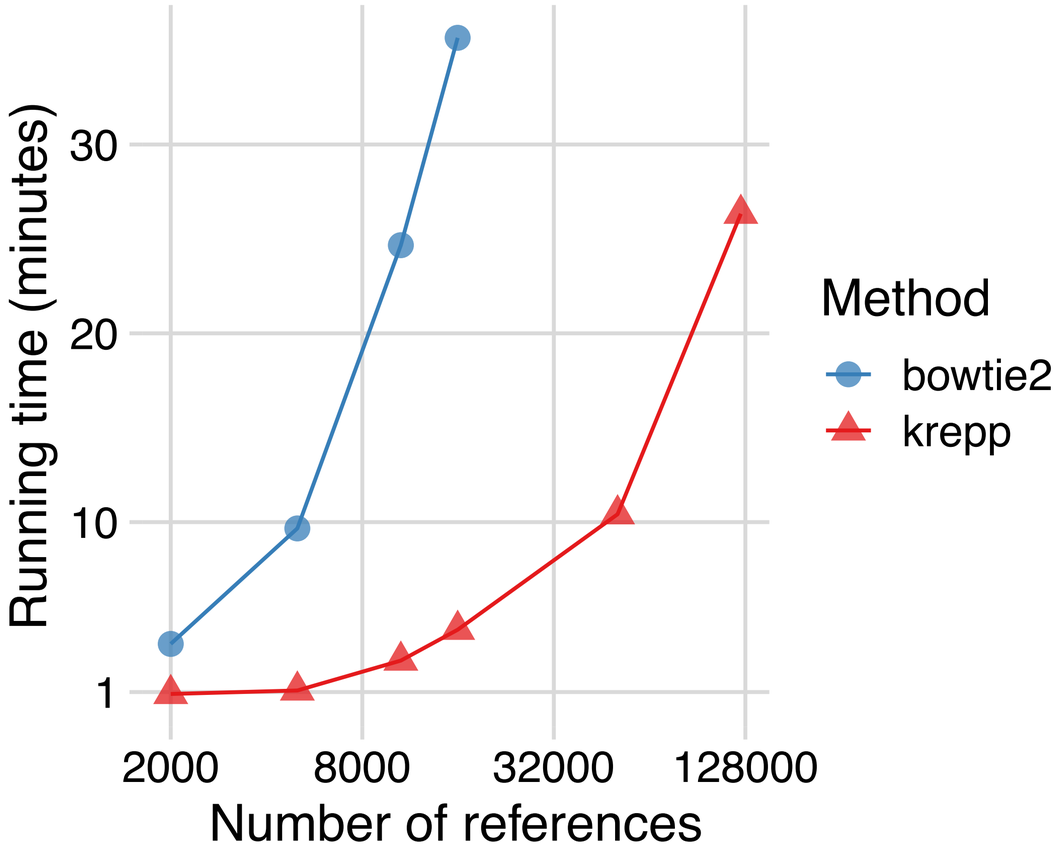
- Leave one out: 100/16,000 (WoLv2)
- Outperforms baselines:
  on the closest, on the LCA, etc.
- >80% of all reads within four edges



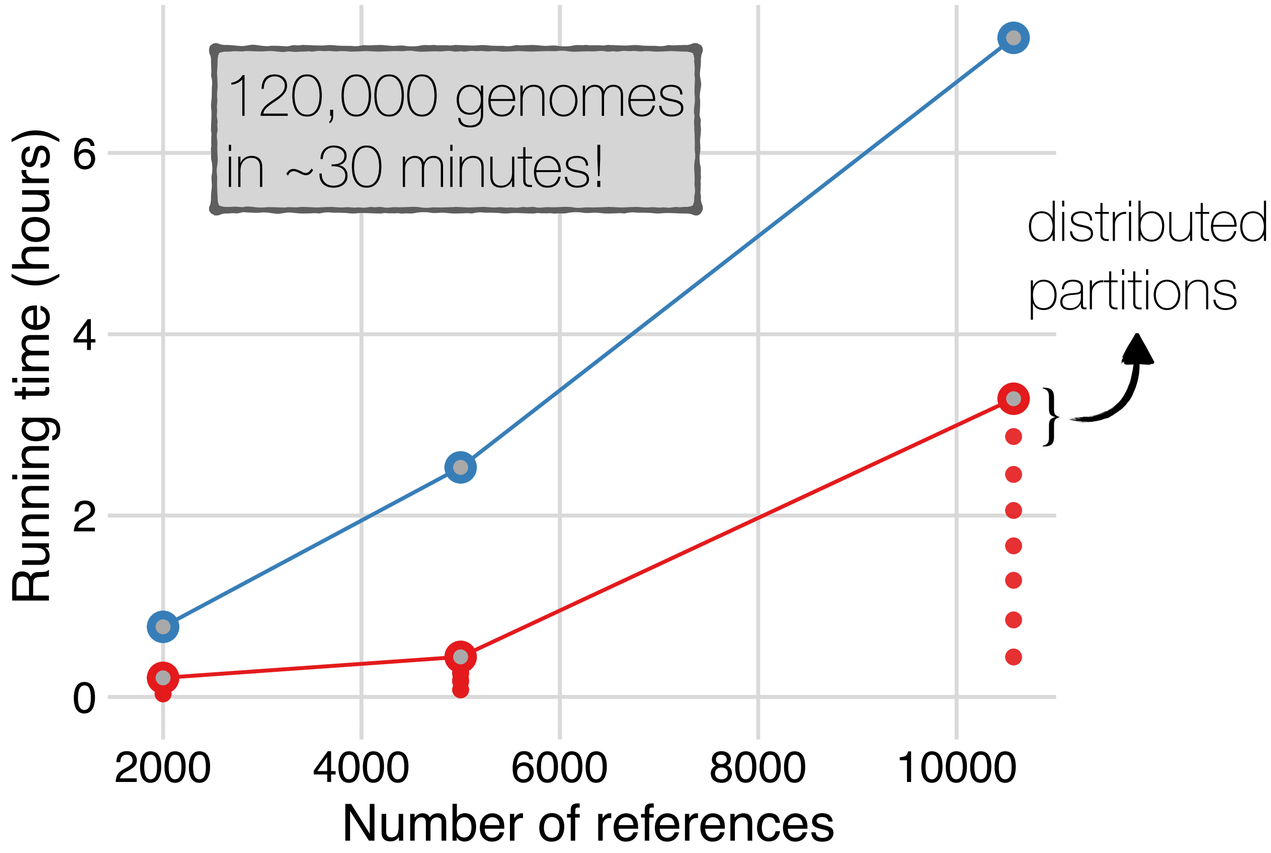
how many edges away is the placement from the correct edge?

Scalability:
Avoiding the more difficult problem & effective parallelization

Mapping 10M reads (16 threads):



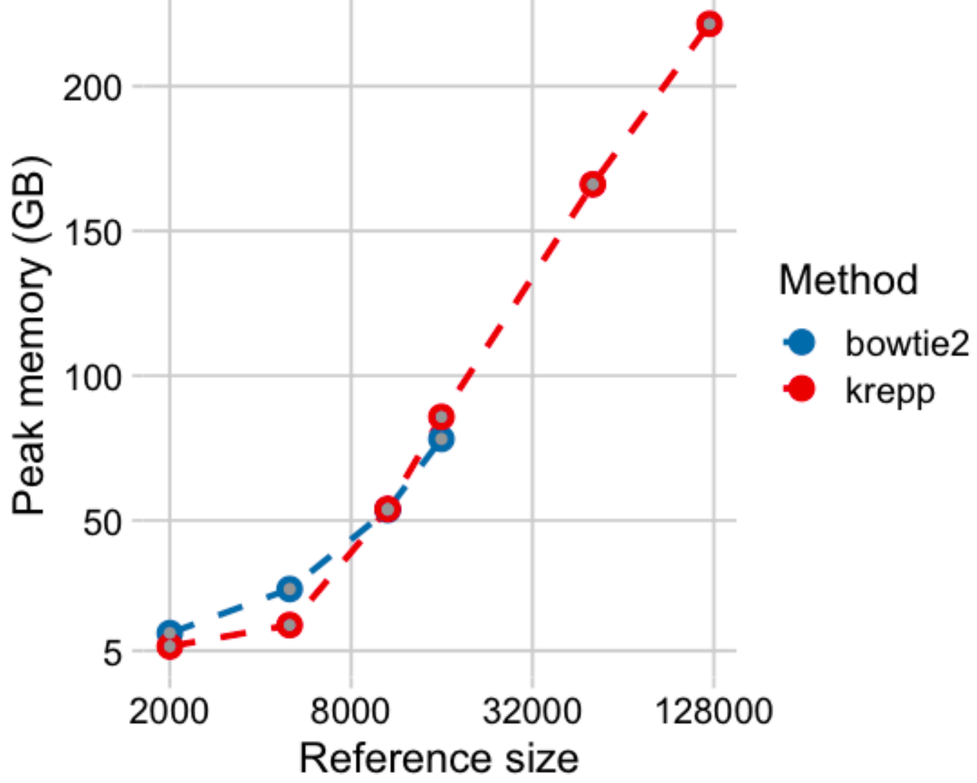
Indexing microbial genomes (32 threads):



Scalability:
krepp can be distributed and has flexible memory requirements

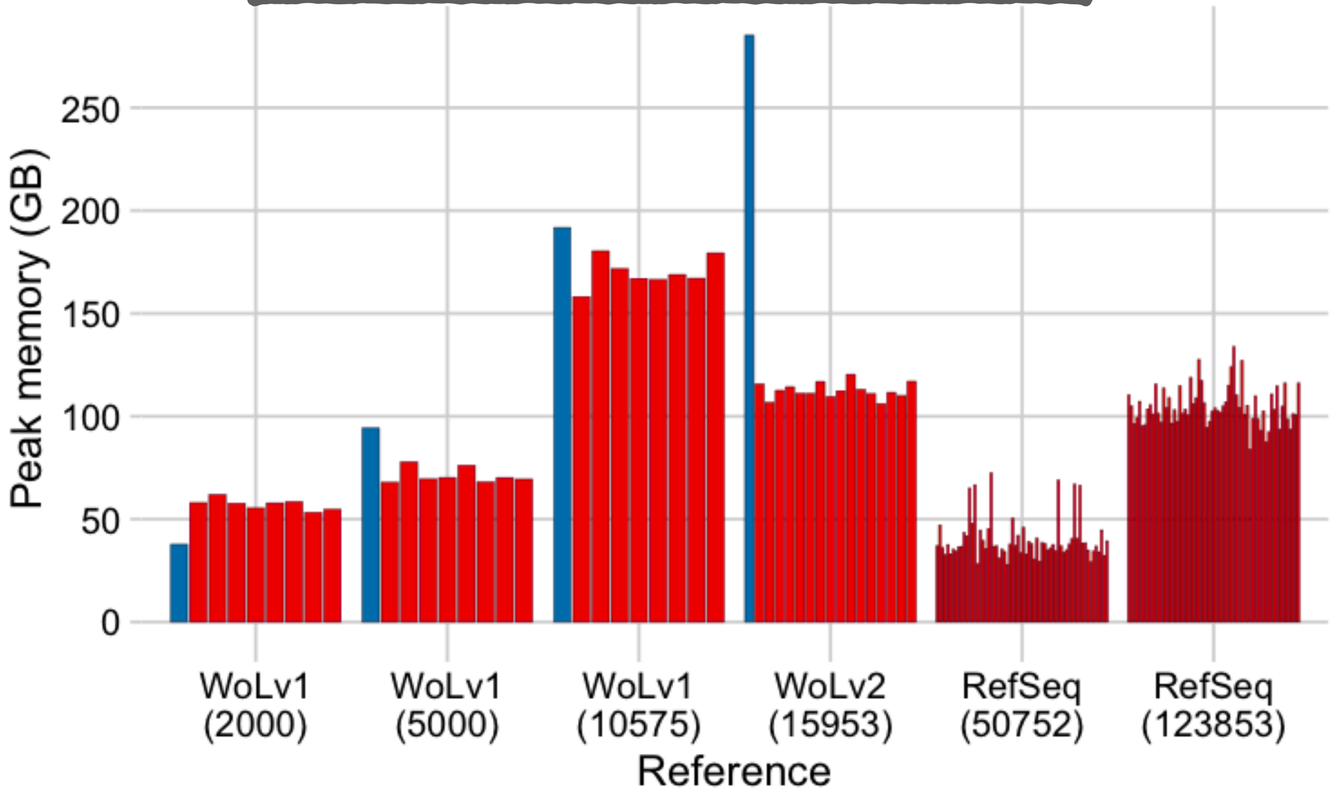
Mapping 10M reads (16 threads):

reducing memory use
w/ further subsampling...

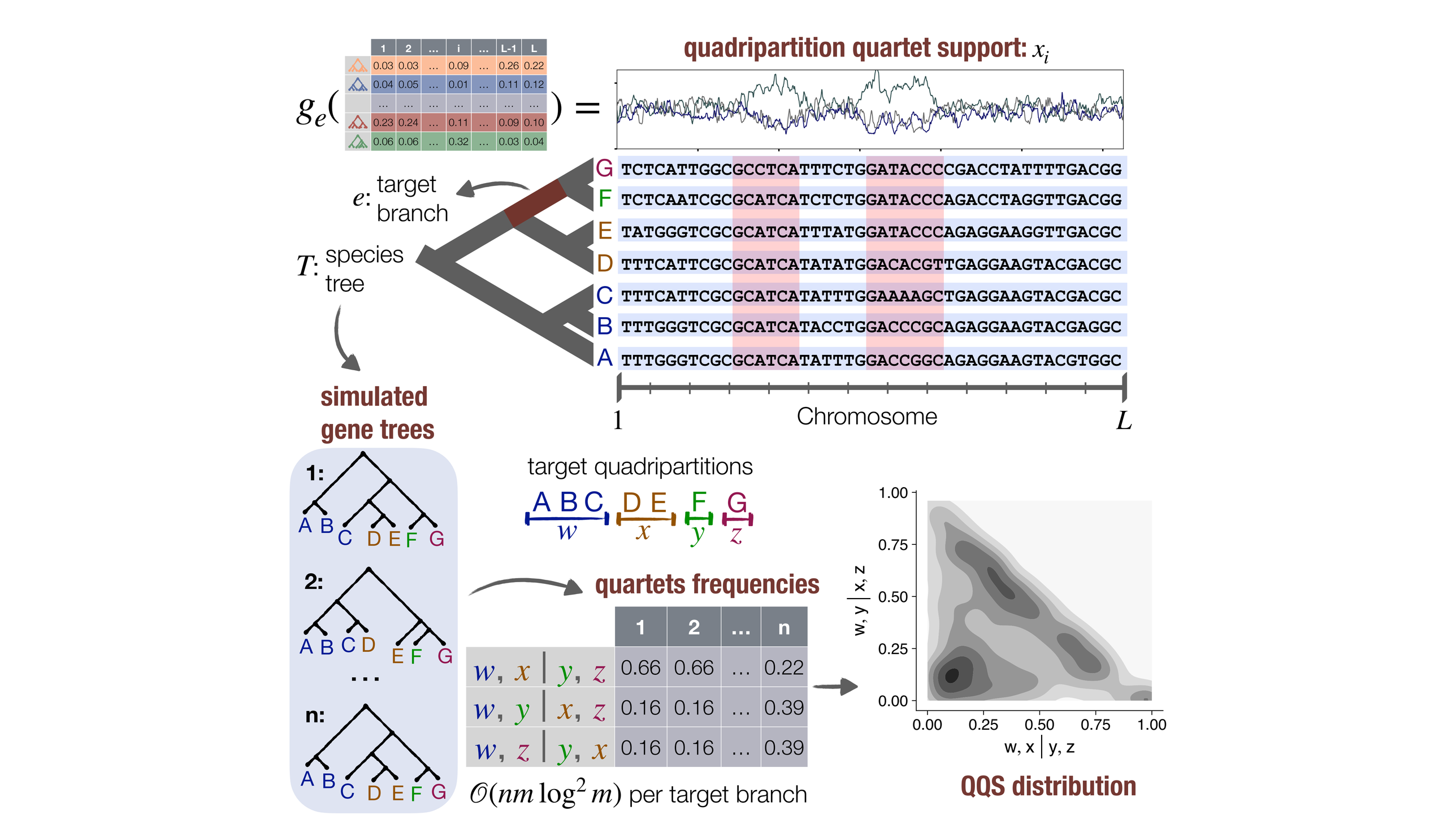


Indexing microbial genomes (32 threads):

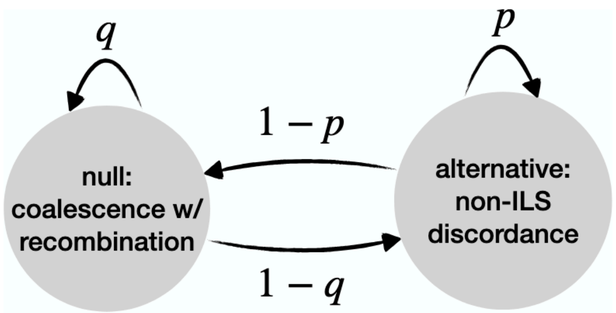
adjusting partitioning based on the
input size & available memory...



Segmentation


1 2 ... i ... L-1 L
0.03 0.03 ... 0.09 ... 0.26 0.22
0.04 0.05 ... 0.01 ... 0.11 0.12
... ...
0.23 0.24 ... 0.11 ... 0.09 0.10
0.06 0.06 ... 0.32 ... 0.03 0.04
$g_e($ $) =$
quadripartition quartet support: x_i
e: target branch
T: species tree
G TCTCATTGGCGCCTCATTTCTGGATACCCCGACCTATTTTGACGG
F TCTCAATCGCGCATCATCTCTGGATACCCAGACCTAGGTTGACGG
E TATGGGTCGCGCATCATTTATGGATACCCAGAGGAAGGTTGACGC
D TTTCATTCGCGCATCATATATGGACACGTTGAGGAAGTACGACGC
C TTTCATTCGCGCATCATATTTGGAAAAGCTGAGGAAGTACGACGC
B TTTGGGTCGCGCATCATACCTGGACCCGCAGAGGAAGTACGAGGC
A TTTGGGTCGCGCATCATATTTGGACCGGCAGAGGAAGTACGTGGC
1
Chromosome
L
simulated gene trees
1:
A B C D E F G
2:
A B C D E F G
...
n:
A B C D E F G
target quadripartitions
A B C | D E | F | G
w x y z
quartets frequencies
1 2 ... n
$w, x \mid y, z$ 0.66 0.66 ... 0.22
$w, y \mid x, z$ 0.16 0.16 ... 0.39
$w, z \mid y, x$ 0.16 0.16 ... 0.39
$\mathcal{O}(nm \log^2 m)$ per target branch
1.00
0.75
0.50
0.25
0.00
w, y | x, z
0.00 0.25 0.50 0.75 1.00
w, x | y, z
QQS distribution

An HMM w/ standard initial and transition distributions with Bayesian priors on parameters:



$$(q, 1-q) \sim \mathrm{Dir}(\beta + \kappa + \rho, \beta)$$
$$(p, 1-p) \sim \mathrm{Dir}(\beta + \rho, \kappa + \beta)$$

- **Initial distribution:** Categorical dist. with an uninformative Dirichlet conjugate prior
- **Transition matrix:** $\pi = \begin{pmatrix} q & 1-q \\ 1-p & p \end{pmatrix}$ reflects our assumptions with flexible & robust priors

Transition matrix hyperparameters

- **Challenges:**
  - noisy observations, high sampling rate
  - slightly off $p(x_i \mid T)$ (e.g., varying rates)
- **stickiness** (κ) favors continuous segments, reduces transitions btw. different states
- **sparsity** (ρ) boosts transitions to the null state and makes states imbalanced