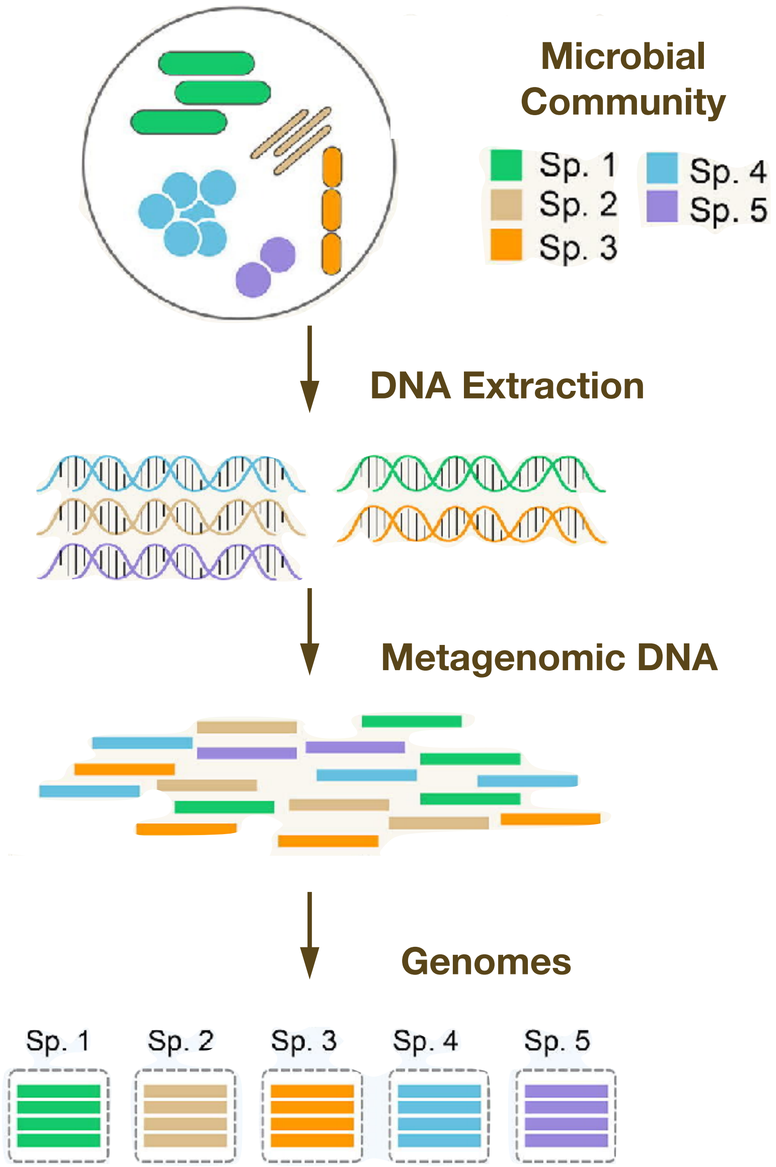
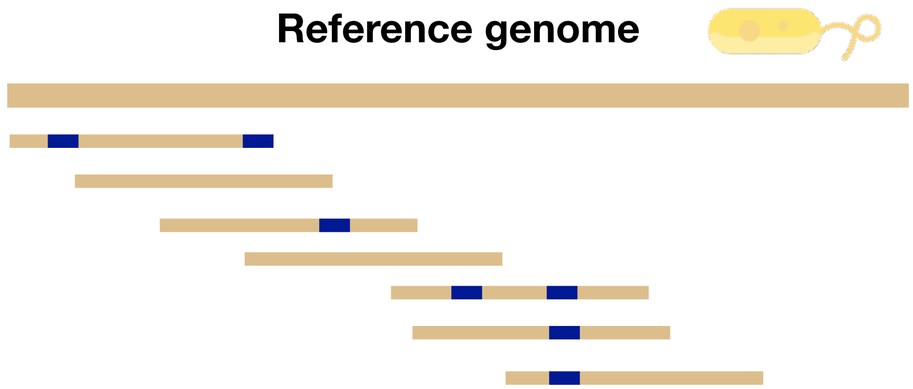
Sourmash + branchwater for large-scale sequence search and profiling

Tessa Pierce Ward, PhD
University of California, Davis

What organisms are in my metagenome?





align reads (or map contigs) allowing mismatches

But, how do you find the *right* reference genomes?

Can we speed up analysis?

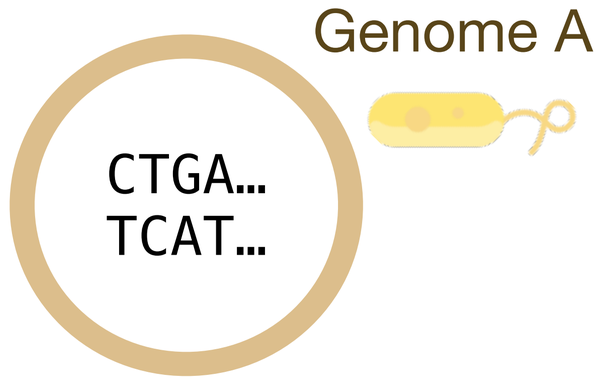
General approach: k-mers for sequence comparisons

ACTACGCCTTCATGACTC

ACTA
CTAC
TACG
ACGC

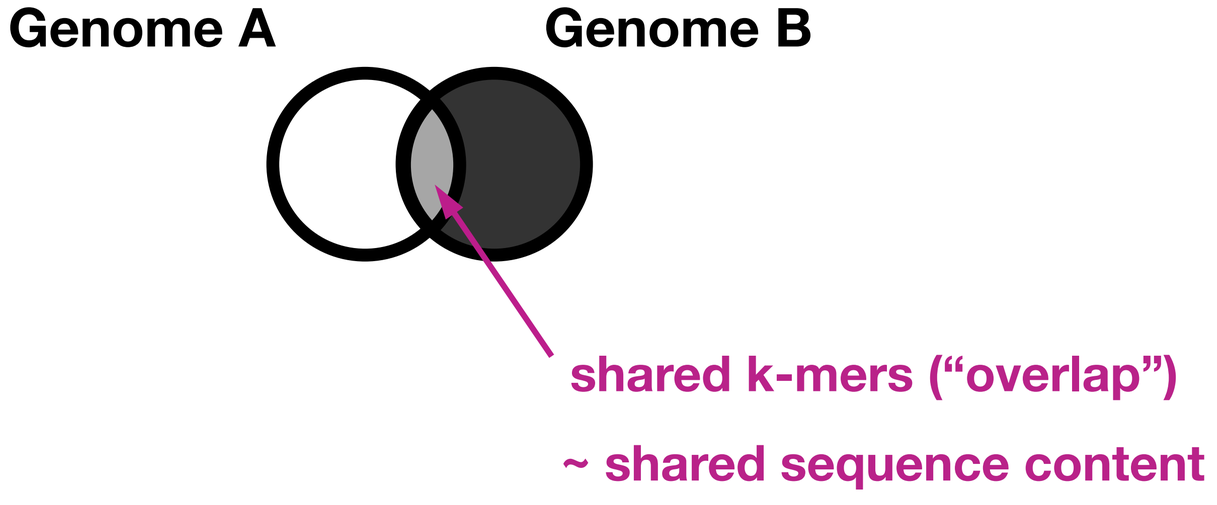
k-mers of length 4
(4-mers)

Represent datasets as
collections of their k-mers



Compare datasets
with set operations

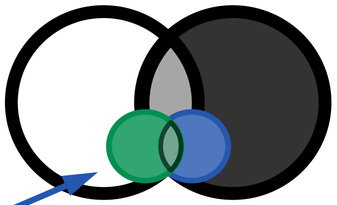
(exact matching)



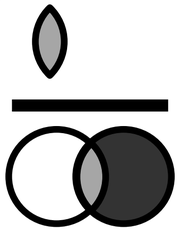
k-mer comparisons capture genomic similarity

Longer k-mers can be *specific*

Genome A **Genome B**

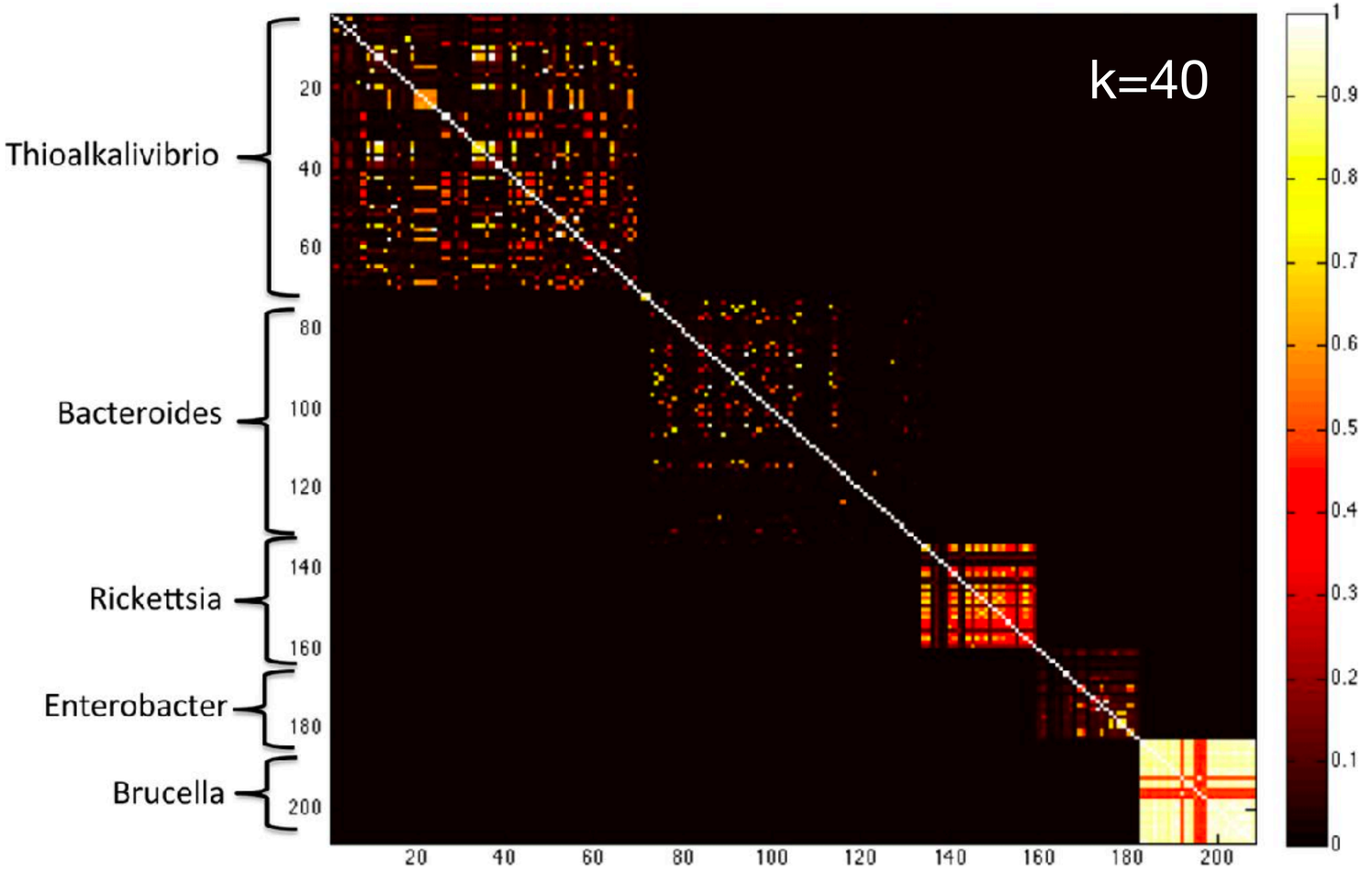


representative subsampling

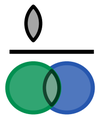


% shared sequence between A and B

Jaccard Index



Koslicki and Falush 2016
10.1128/mSystems.00020-16



This works even when you select a representative subsample of k-mers ("sketch")

Why use a k-mer approach?

- Assembly-independent: works with or without good assemblies
    - Particularly helpful for metagenomes
- Lightweight: developments in sketching and storage have significantly accelerated computation
- Highly sensitive and specific

sourmash: a k-mer analysis multitool

sourmash

sourmash.readthedocs.io

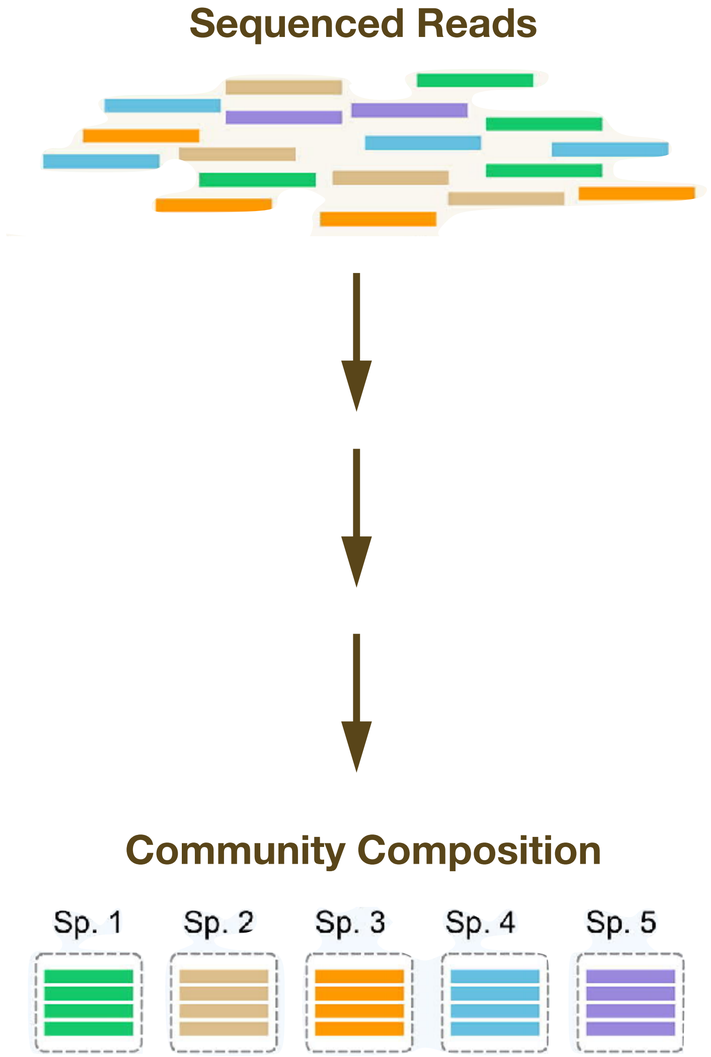
github.com/sourmash-bio/sourmash

gitter.im/sourmash-bio/community

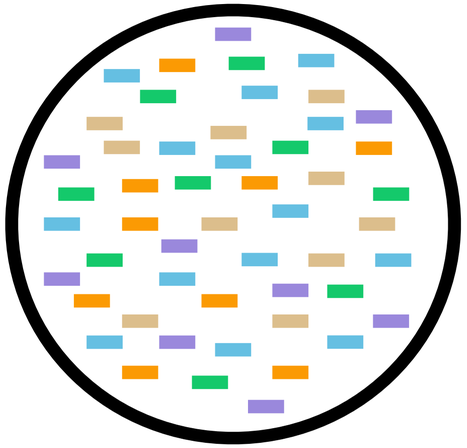
- Implements **FracMinHash**, which randomly and systematically selects a **fractional subset** of k-mers ***(default 0.1%)***
- Sketches can be used for **fast and lightweight** sequence comparisons
- Search, compare, analyze **genomic** and **metagenomic** datasets
- Open source, openly developed :)

Irber et al. (2024) 10.21105/joss.06830
Pierce et al. (2019) 10.12688/f1000research.19675.1

What organisms are in my metagenome?



Sketch metagenome into k-mer collection



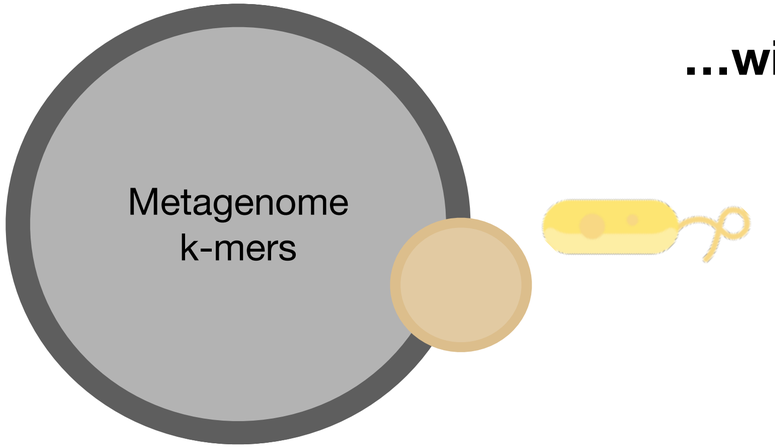
…with k-mers

What organisms are in my metagenome?

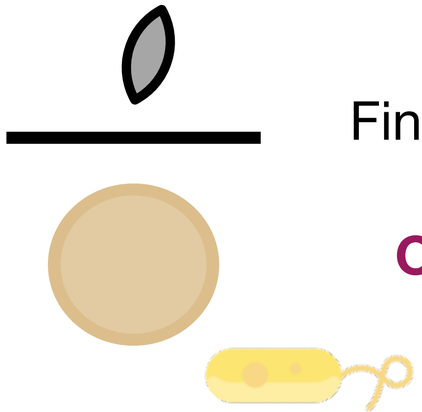
Sequenced Reads

Sketch metagenome into k-mer collection

…with k-mers



Community Composition



Find k-mers shared with reference genome

Containment Index

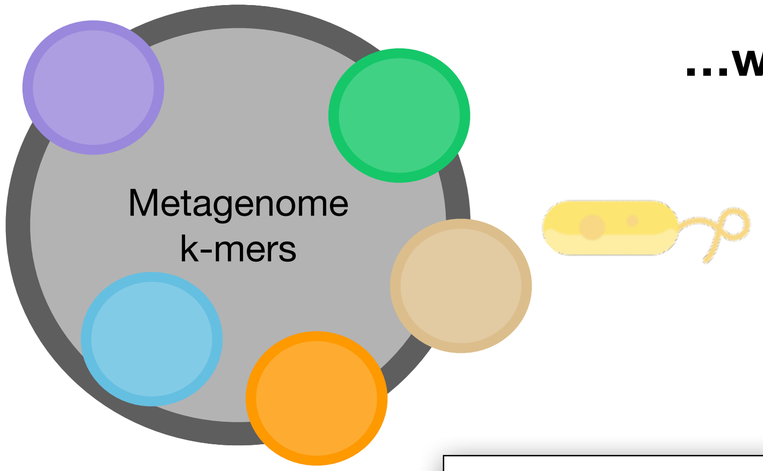
FracMinHash

Ondov et al., 2019 10.1186/s13059-019-1841-x
Koslicki and Zabeti 2019 10.1016/j.amc.2019.02.018
Irber et al., 2022 10.1101/2022.01.11.475838

What *genomes* are in my metagenome?

Sequenced Reads

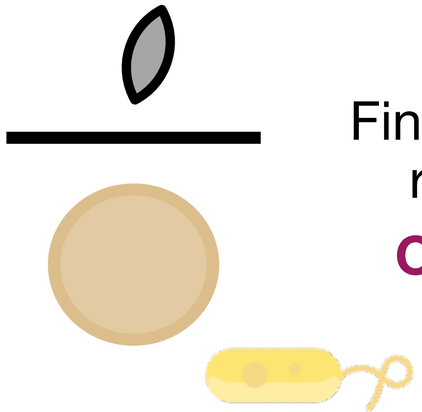
Sketch metagenome into k-mer collection



…with k-mers

Screen large databases for genome matches

Community Composition



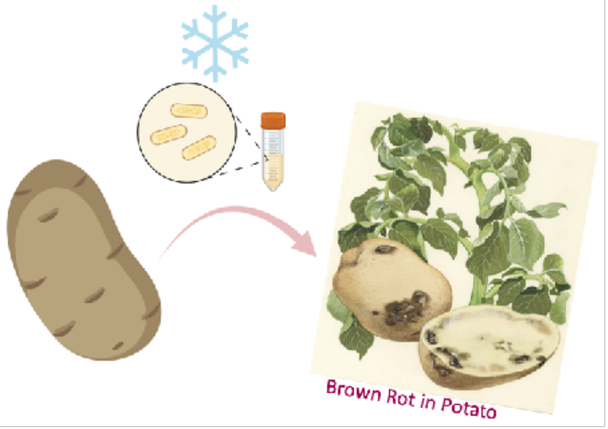
Find k-mers shared with reference genomes

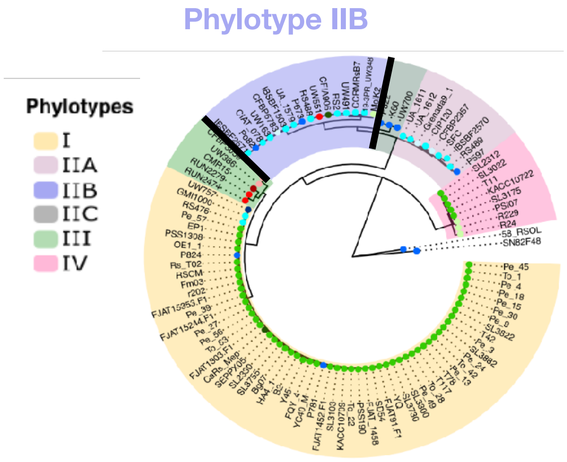
Containment Index

FracMinHash

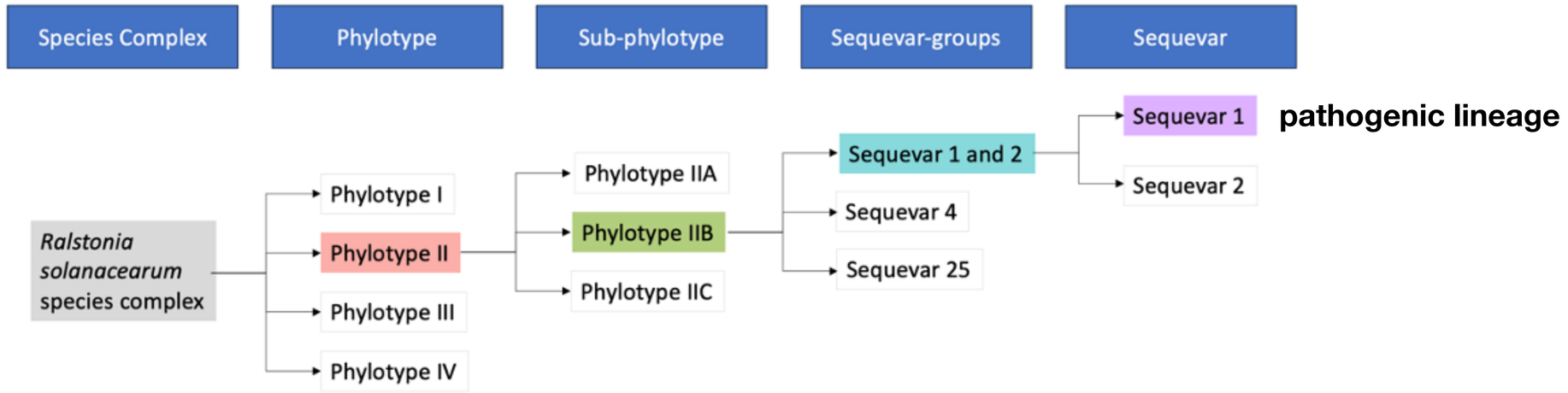
Ondov et al., 2019 10.1186/s13059-019-1841-x
Koslicki and Zabeti 2019 10.1016/j.amc.2019.02.018
Irber et al., 2022 10.1101/2022.01.11.475838

Taxonomic profiling with sourmash: Strain-level demo with *Ralstonia solanacearum*





Species Complex

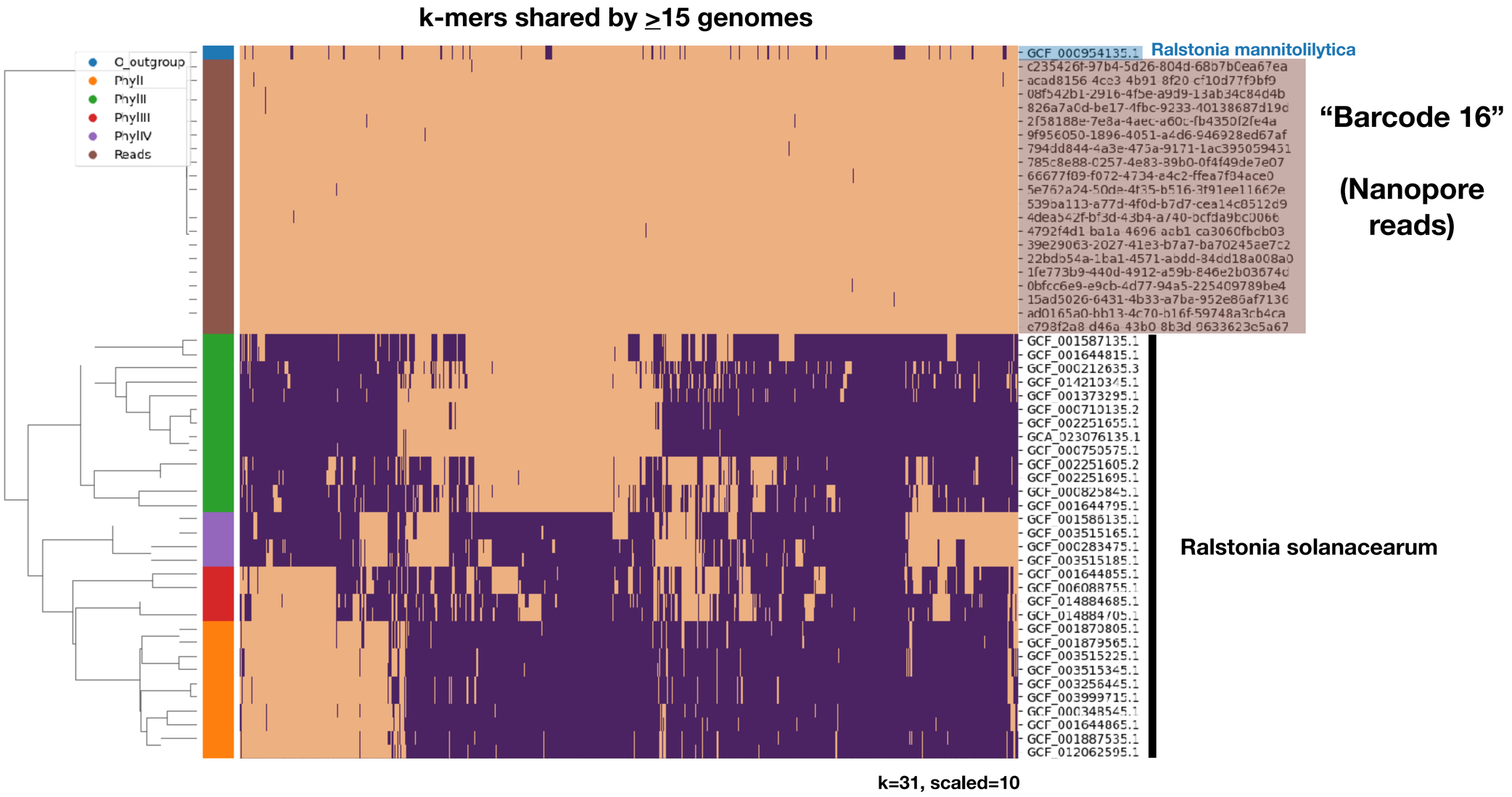


Parul Sharma, PhD

Ralstonia solanacearum genomes share many k-mers

32 genome Ralstonia database

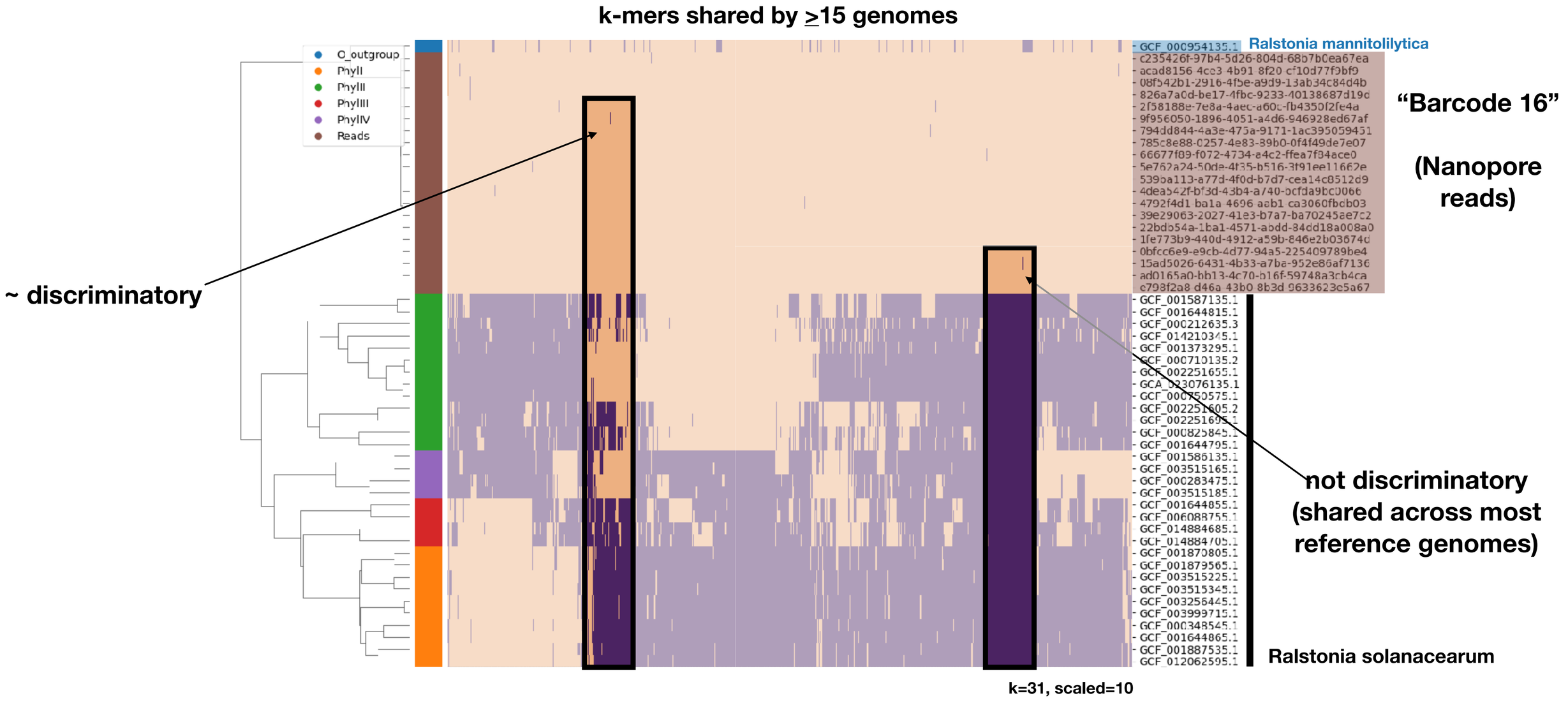
Each read represents a small portion of a genome



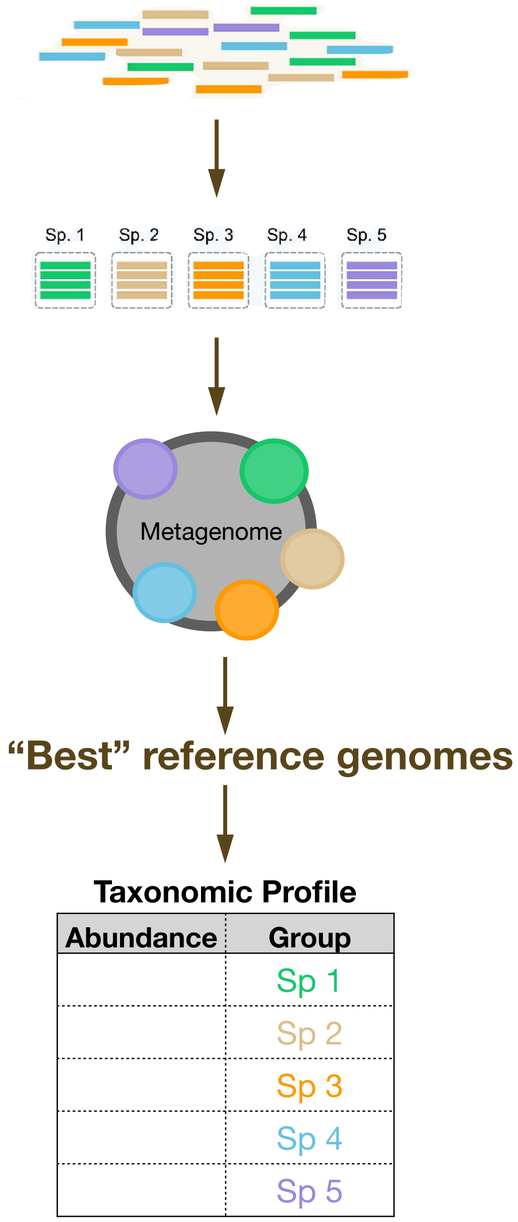
“Barcode 16”

(Nanopore reads)

Only some reads originate from discriminatory regions



Approach: Genome-Resolved Taxonomic profiling

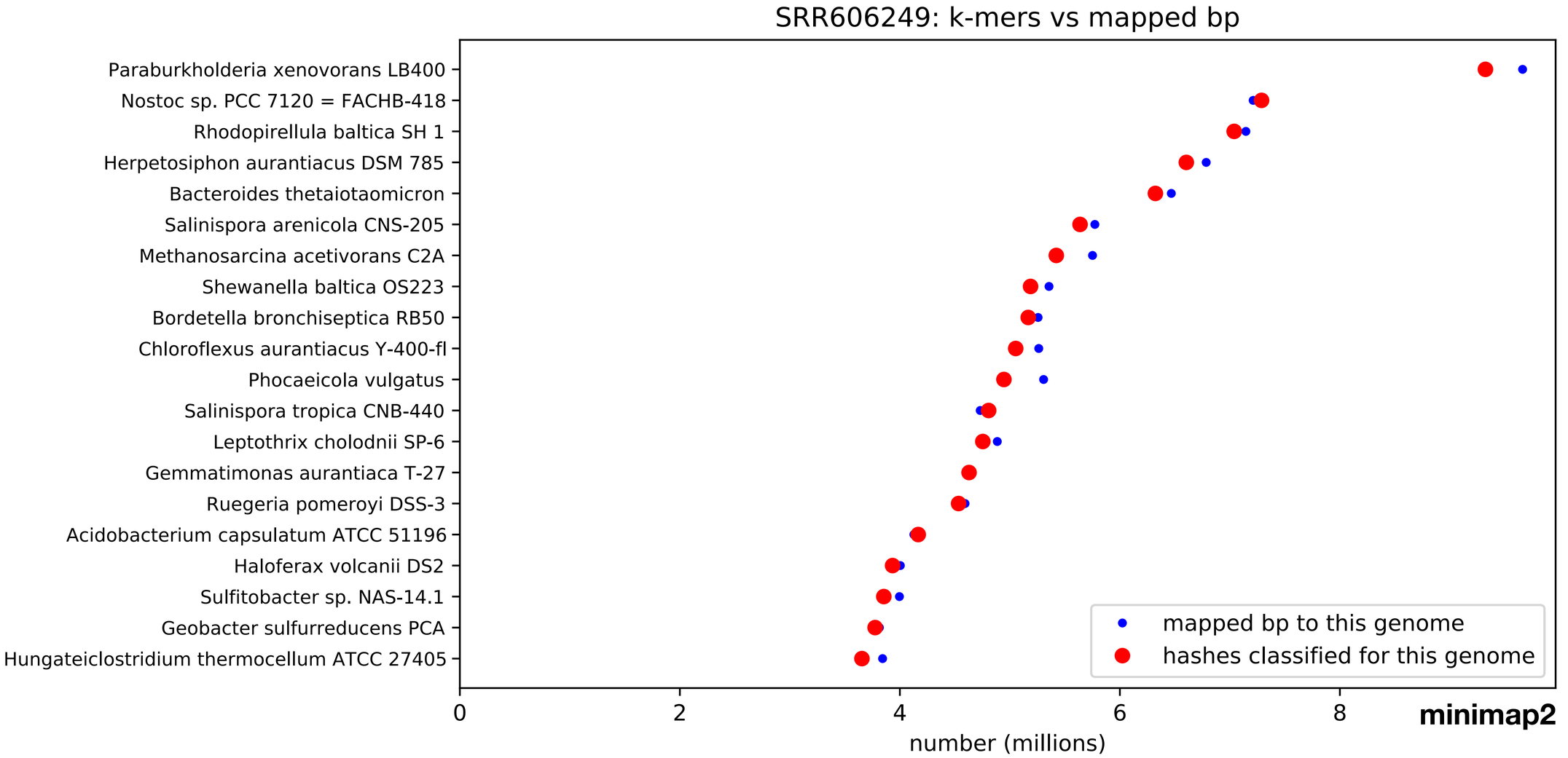


- **Rather than considering each read independently, sourmash can use information from the full file**
  - Larger combinations of k-mers can more robustly distinguish between closely-related genomes
    - particularly helpful for short reads
    - Enables faster search of large databases (e.g. GenBank, AllTheBacteria)

- **`gather` uses a greedy implementation of minimum set cover to find "best" reference genomes**

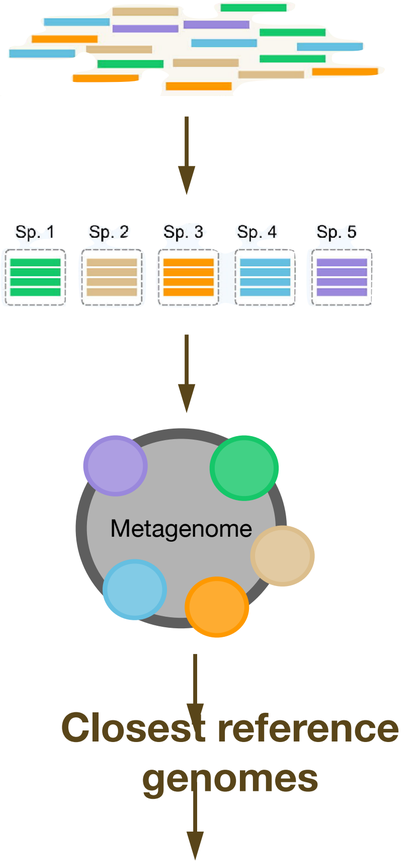
Irber et al (2022) 10.1101/2022.01.11.475838

Mapping metagenome reads to sourmash gather genome results closely matches gather k-mer estimates



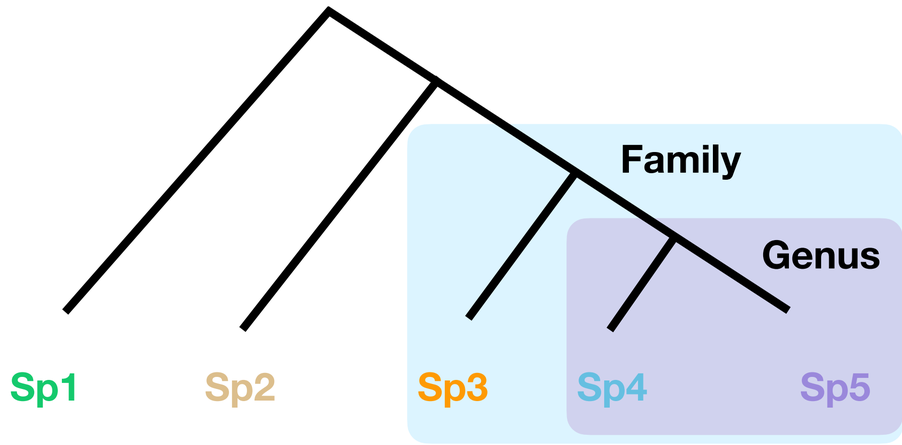
Irber et al (2022) 10.1101/2022.01.11.475838

Approach: Integrate Taxonomy after resolving genomes



- **sourmash taxonomy**:
  - use gather's non-overlapping genome matches to add taxonomic information
  - Aggregate with LCA if needed

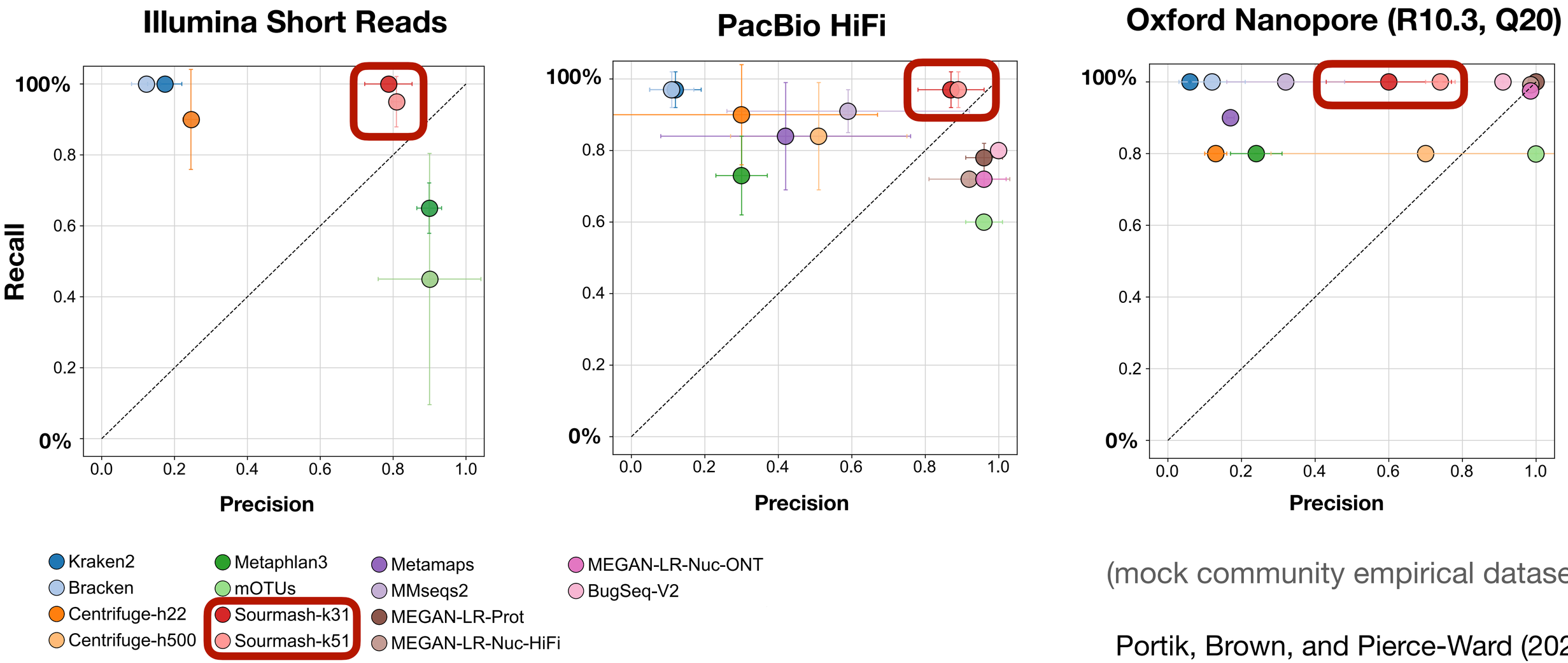
Lowest Common Ancestor (LCA) Approach



works with rank-based taxonomies (GTDB, NCBI, ICTV) and LINs/LINgroups

Portik, Brown, and Pierce-Ward (2022)
10.1186/s12859-022-05103-0

Sourmash taxonomic profiling performs very well (species-level)



(mock community empirical datasets)

Portik, Brown, and Pierce-Ward (2022)
10.1186/s12859-022-05103-0

Strain-level demo: infected field sample “barcode 16”

(nanopore)

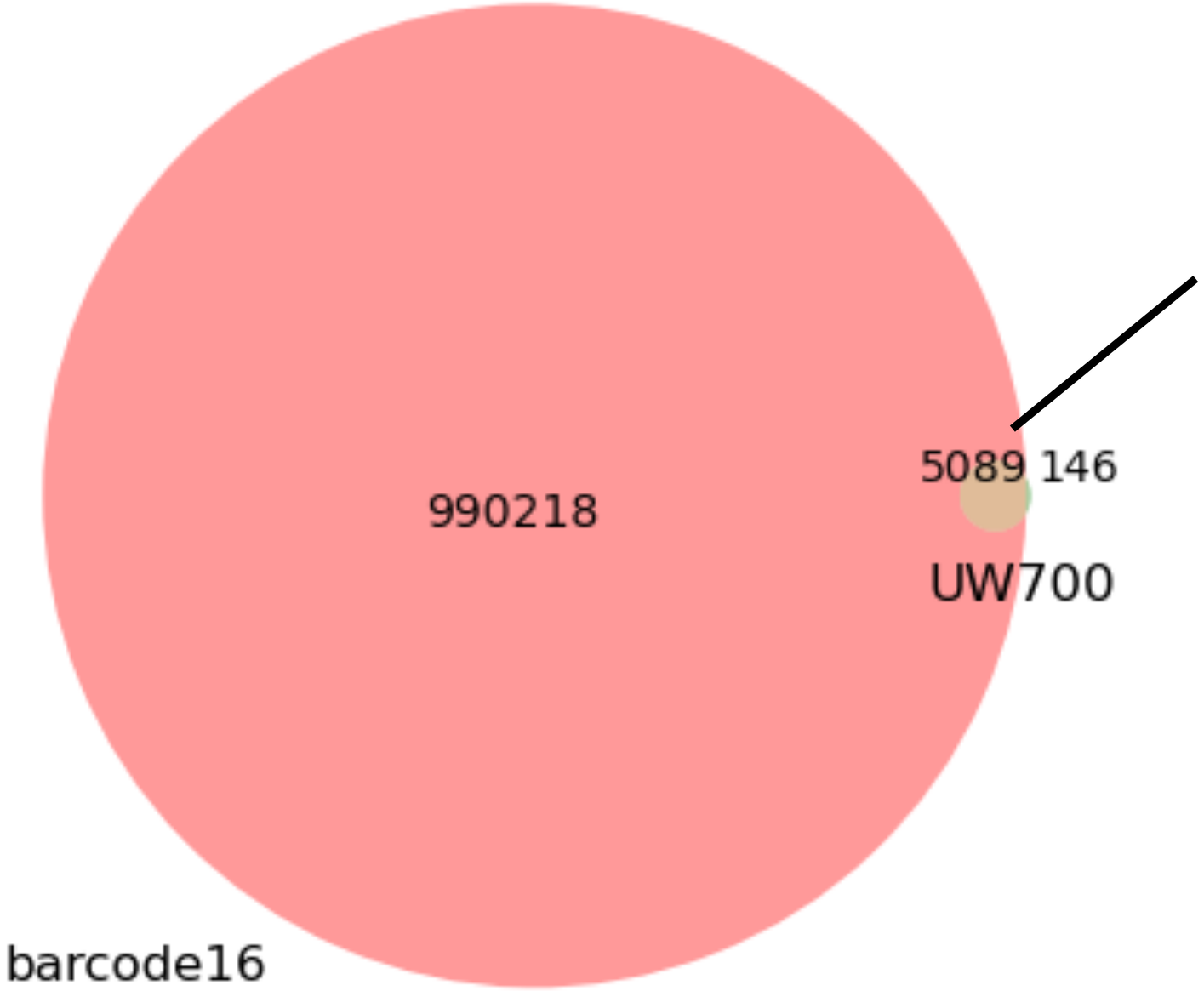
```
sourmash gather inputs/bc16.scaled1000.zip databases/ralstonia.zip \
                        -k 31 --output barcode16.k31.gather.csv
```

```
Starting prefetch sweep across databases.
Prefetch found 32 signatures with overlap >= 50.0 kbp.
Doing gather to generate minimum metagenome cover.

overlap     p_query p_match avg_abund
---------   ------- ------- ---------
5.1 Mbp       17.8%   97.2%      57.5    GCF_002251605.2 Ralstonia solanacearum UW700
2.7 Mbp        0.3%    4.0%      21.2    GCF_000825845.1 Ralstonia solanacearum Grenada9_1
2.2 Mbp        0.0%    2.1%       5.0    GCF_001644815.1 Ralstonia solanacearum CFBP6783
0.8 Mbp        0.0%    1.6%       2.0    GCF_012062595.1 Ralstonia solanacearum Pe_13
4.9 Mbp        0.2%    1.2%      44.8    GCF_002251695.1 Ralstonia solanacearum K60
2.1 Mbp        0.0%    1.0%       1.4    GCF_000212635.3 Ralstonia solanacearum MolK2
found less than 50.0 kbp in common. => exiting

found 6 matches total;
the recovered matches hit 18.3% of the abundance-weighted query.
the recovered matches hit 0.6% of the query k-mers (unweighted).
```

p_query: % of metagenome query
p_match: % of reference genome

18% of barcode16 matches Ralstonia UW700

Strain-level demo: infected field sample “barcode 16”



unique (non-abundance weighted) k-mers shared with Ralstonia solanacearum UW700

Strain-level demo: infected field sample “barcode 16”

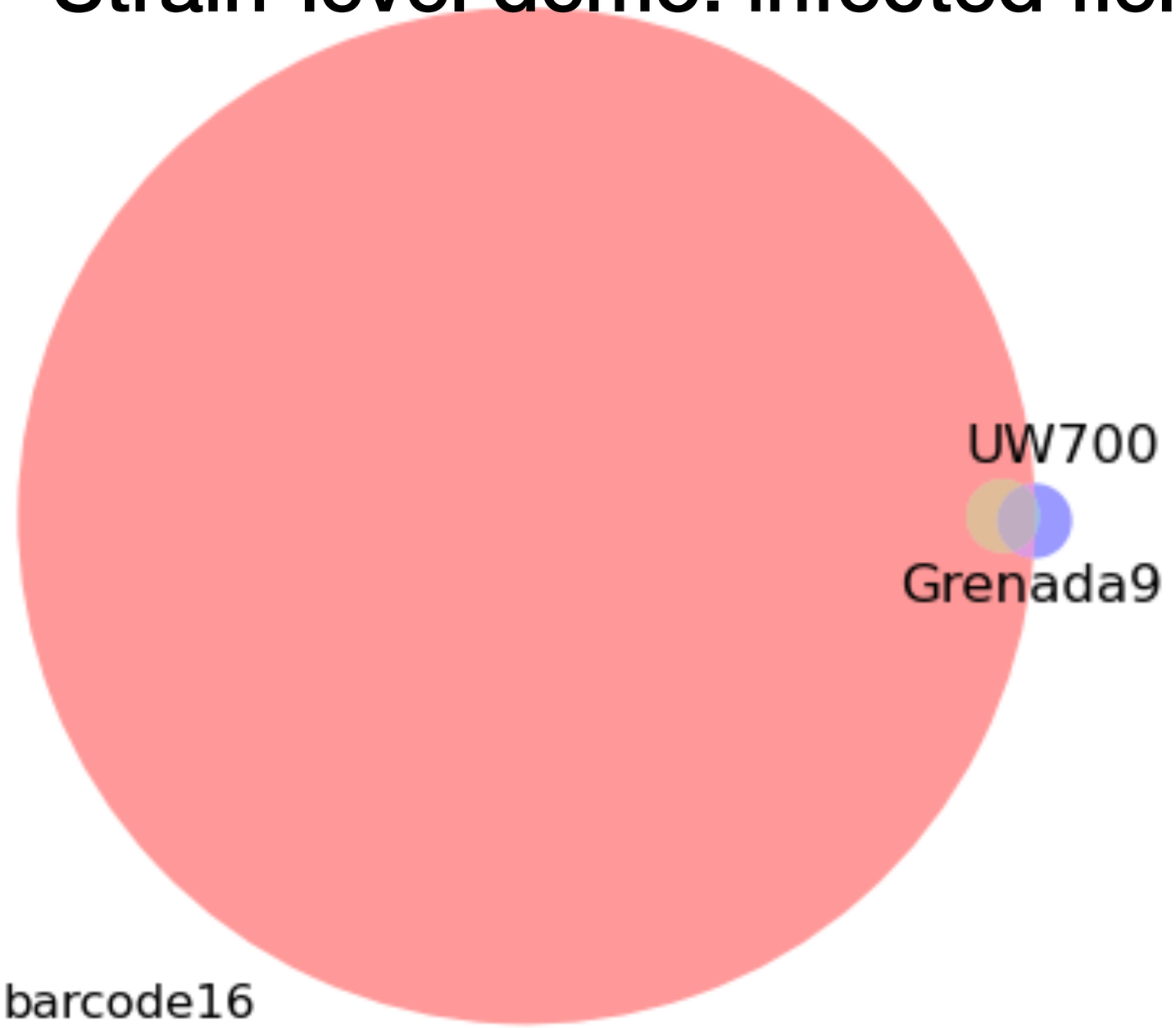
```
sourmash gather inputs/bc16.scaled1000.zip databases/ralstonia.zip \
                        -k 31 --output barcode16.k31.gather.csv
```

```
Starting prefetch sweep across databases.
Prefetch found 32 signatures with overlap >= 50.0 kbp.
Doing gather to generate minimum metagenome cover.

overlap     p_query p_match avg_abund
---------   ------- ------- ---------
5.1 Mbp       17.8%   97.2%      57.5    GCF_002251605.2 Ralstonia solanacearum UW700
2.7 Mbp        0.3%    4.0%      21.2    GCF_000825845.1 Ralstonia solanacearum Grenada9_1
2.2 Mbp        0.0%    2.1%       5.0    GCF_001644815.1 Ralstonia solanacearum CFBP6783
0.8 Mbp        0.0%    1.6%       2.0    GCF_012062595.1 Ralstonia solanacearum Pe_13
4.9 Mbp        0.2%    1.2%      44.8    GCF_002251695.1 Ralstonia solanacearum K60
2.1 Mbp        0.0%    1.0%       1.4    GCF_000212635.3 Ralstonia solanacearum MolK2
found less than 50.0 kbp in common. => exiting

found 6 matches total;
the recovered matches hit 18.3% of the abundance-weighted query.
the recovered matches hit 0.6% of the query k-mers (unweighted).
```

p_query: % of metagenome query
p_match: % of reference genome

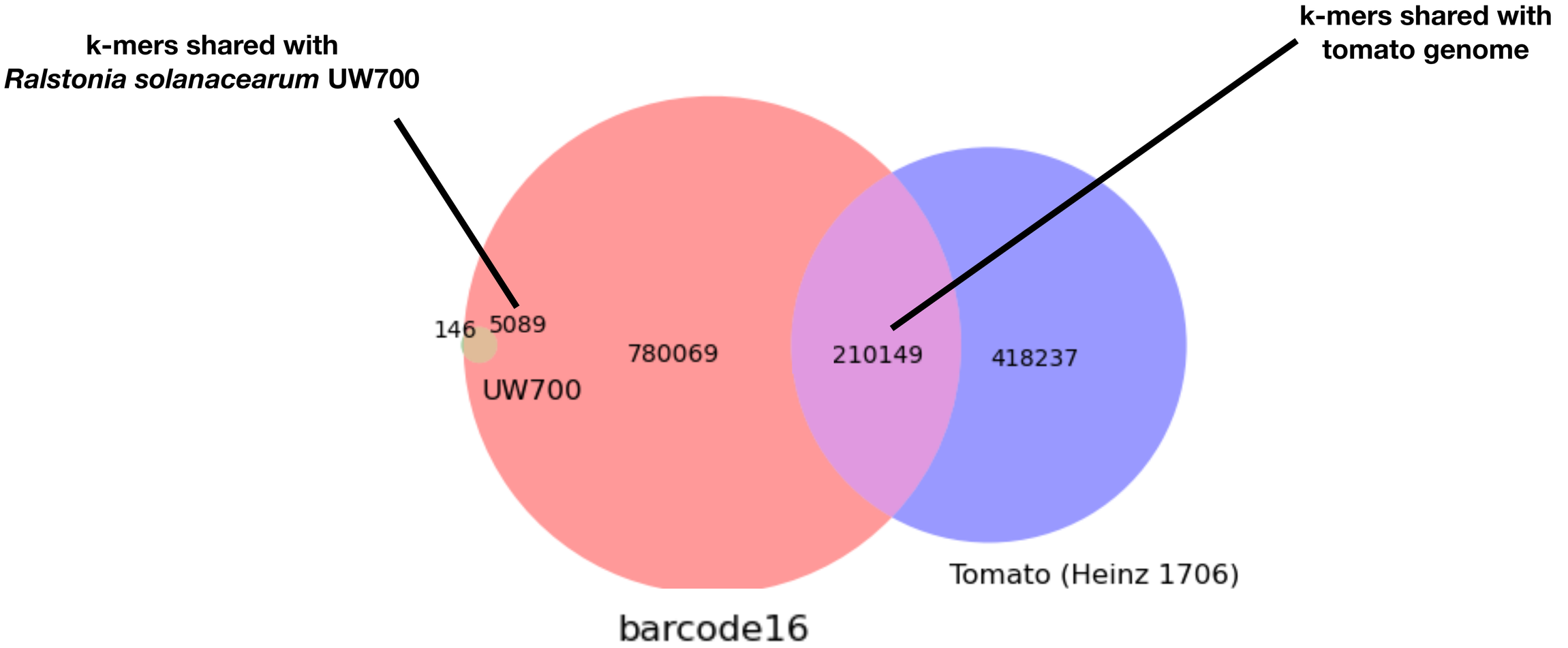
After matching UW700, only 0.3% of barcode16 is left that matches Grenada9_1

Strain-level demo: infected field sample “barcode 16”



Most of the k-mers shared with Grenada9_1 are already assigned to UW700, which is a better match

What are the remaining k-mers?



Integrating taxonomy

```
sourmash tax metagenome -g barcode16.k31.gather.csv \
                        -t databases/ralstonia32.lin-taxonomy.csv \
                        --lins --lingroup databases/ralstonia.lingroups.csv
```

```
Trying to read LIN taxonomy assignments.
loaded 1 gather results from 'barcode16.k31.gather.wt.csv'.
loaded results for 1 queries from 1 gather CSVs
of 7 gather results, lineage assignments for 1 results were missed.
The following are missing from the taxonomy information: GCF_000188115
Read 20 lingroup rows and found 20 distinct lingroup prefixes.
name    lin     percent_containment     num_bp_contained
A_Total_reads;B_PhylII  14;1;0;0;0;3;0  18.33   300671000
A_Total_reads;B_PhylII;C_IIC    14;1;0;0;0;3;0;2        18.01   295389000
A_Total_reads;B_PhylII;C_IIA    14;1;0;0;0;3;0;1        0.28    4629000
A_Total_reads;B_PhylII;C_IIB    14;1;0;0;0;3;0;0        0.04    653000
A_Total_reads;B_PhylII;C_IIB;D_seq4     14;1;0;0;0;3;0;0;1;0;0;0;0;0    0.04    579000
A_Total_reads;B_PhylI   14;1;0;0;0;0;0;0;0;0    0.01    182000
A_Total_reads;B_PhylI;C_seq15   14;1;0;0;0;0;0;0;0;0;2  0.01    182000
```

18% of sample assigned to Phylotype IIC

Strain-level demo: simulated metagenomes

Simulated Samples: Ralstonia + tomato host:

- `Sample0` - no Ralstonia
- `Sample-II` - Ralstonia solanacearum, PhyIIIB
- `SampleIV` - Ralstonia solanacearum, PhyI-IV

Sample II gather command:

```
sourmash gather inputs/Sample-II.sc1000.zip \
                databases/ralstonia.zip -k 31 \
                --output Sample-II.k31.gather.csv
```

Sample II gather output:

```
overlap     p_query p_match avg_abund
---------   ------- ------- ---------
1.3 Mbp        3.9%   26.6%       1.2    GCF_001373295.1 Ralstonia solanacearum RS2
found less than 50.0 kbp in common. => exiting
```

Sample II tax output:

```
name    lin     percent_containment     num_bp_contained
A_Total_reads;B_PhylII  14;1;0;0;0;3;0  3.94    1464000
A_Total_reads;B_PhylII;C_IIB    14;1;0;0;0;3;0;0        3.94    1464000
A_Total_reads;B_PhylII;C_IIB;D_seq1&seq2        14;1;0;0;0;3;0;0;0;0;1;0;0;0;0  3.94    1464000
A_Total_reads;B_PhylII;C_IIB;D_seq1&seq2;E_seq1 14;1;0;0;0;3;0;0;0;0;1;0;0;0;0;0;0      3.94    1464000
```

Phyl IIB - sequevar 1 (pathogenic lineage)

Demo: https://mybinder.org/v2/gh/bluegenes/2024-icppb/HEAD?labpath=sourmash-icppb-demo.ipynb

Considerations for sourmash profiling

- Anchoring groups of k-mers to reference genomes can provide accurate taxonomic profiling
    - As sourmash gather provides ***genome*-level** results, resolution can be ~as fine-grained as your taxonomic structure,
        - BUT must choose resolution (scaled value) and k-mer size appropriately. *size/speed trade-off*
- Too slow / many samples? Try **sourmash_plugin_branchwater** for multithreaded, multiplexed implementations (fastgather, fastmultigather!)
- Questions? Ask us!

https://github.com/sourmash-bio/sourmash_plugin_branchwater

gitter.im/sourmash-bio/community
github.com/sourmash-bio/sourmash/issues

Teaser for Tuesday, 2:40pm: Where in the world is my favorite organism?

https://branchwater.jgi.doe.gov

Branchwater Metagenome Query

Real-time search for a genome within metagenomes in the SRA.

Curious what publicly available metagenomes contain an organism you're interested in?

Start by selecting your genome below (fasta format only) or checking out our examples.

Select nucleotides FastA files:

Choose Files... Browse

Suzanne Fleishman

Adam Rivers

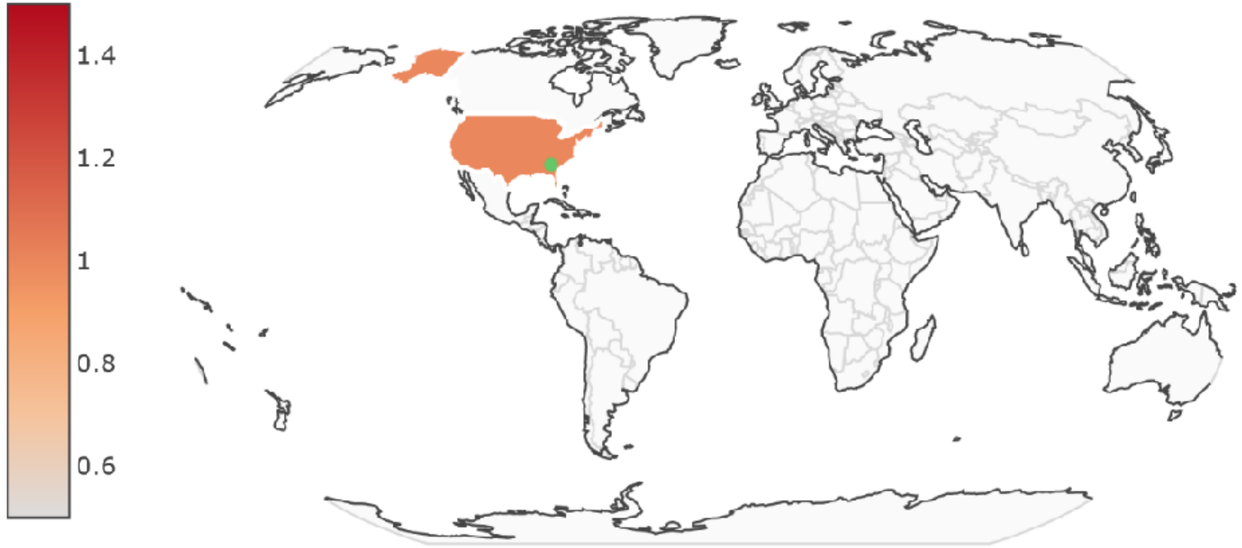
UC DAVIS UNIVERSITY OF CALIFORNIA

JGI JOINT GENOME INSTITUTE

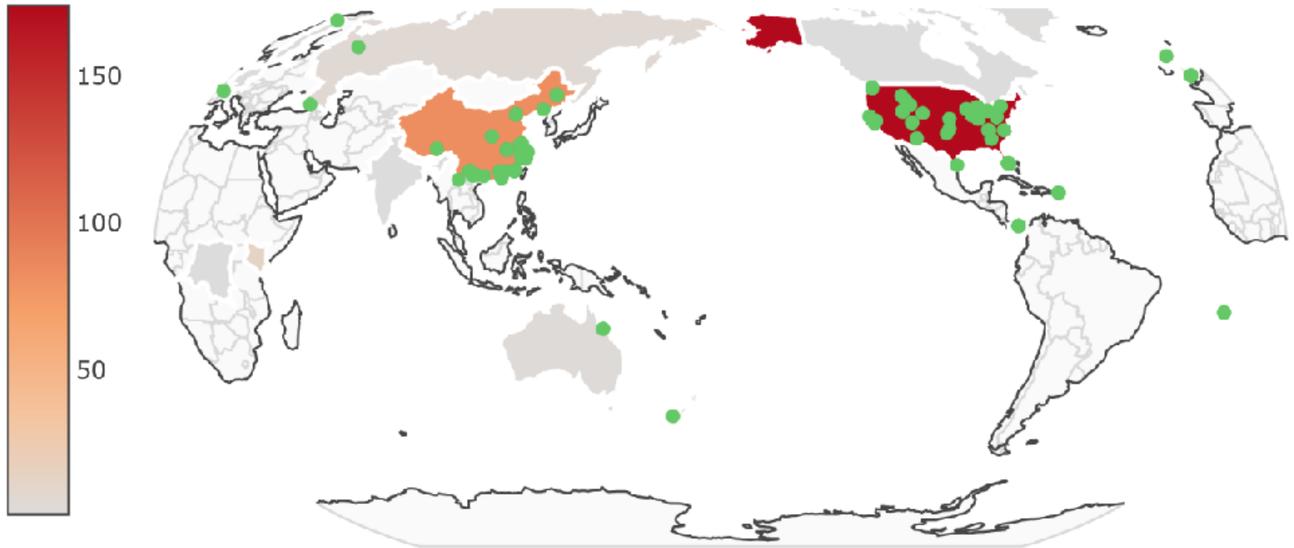
USDA

Branchwater: Search MAG from barcode 16

https://branchwater.jgi.doe.gov



>=99% ANI



>=95% ANI

Sahar Abdelrazek, PhD

Questions?