Bioinformatics approaches for making use of microbial genome sequences for taxonomy, function, and pangenomics

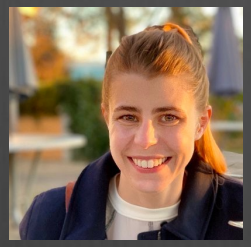
C. Titus Brown
School of Veterinary Medicine;
UC Davis

June 10, 2024

Talk slides available! google 'titus talk osf.io'

Acknowledgements

Dr. Luiz Irber



Dr. Taylor Reiter

Dr. Tessa Pierce

Dr. Phillip Brooks

This talk is *mostly* about the technology stack developed by Dr. Luiz Irber, aided and abetted by many others in the lab - most especially, Taylor, Phil, and Tessa.

Introduction

Our lab's perspective is:

- An awful lot of genomic and metagenomic data is available, and more keeps on coming!
- We need **powerful exploratory approaches** for examining this data, especially at **strain-level resolution**.
- These approaches should work with public and private data, run efficiently, and enable a wide variety of analyses.
- Importantly, they should *not* require de-replication or removal of pangenomic redundancy.

Caveats and considerations

- I'll be mostly talking about our lab's software, "sourmash".
- In part this is because I like it a lot!
- But, just as (more?) importantly, it is ***one example*** of the kind of modern analysis approaches that are increasingly available. Plan accordingly!
  - E.g. software that uses similar approaches: mash, skani

Buzzwords: "massively scalable", "exploratory", "real time"

sourmash

Quickly search, compare, and analyze genomic and metagenomic data sets

Welcome to sourmash!

sourmash is a command-line tool and Python/Rust library for **metagenome analysis** and **genome comparison** using k-mers. It supports the compositional analysis of metagenomes, rapid search of large sequence databases, and flexible taxonomic profiling with both NCBI and GTDB taxonomies (see our prepared databases for more information). sourmash works well with sequences 30kb or larger, including bacterial and viral genomes.

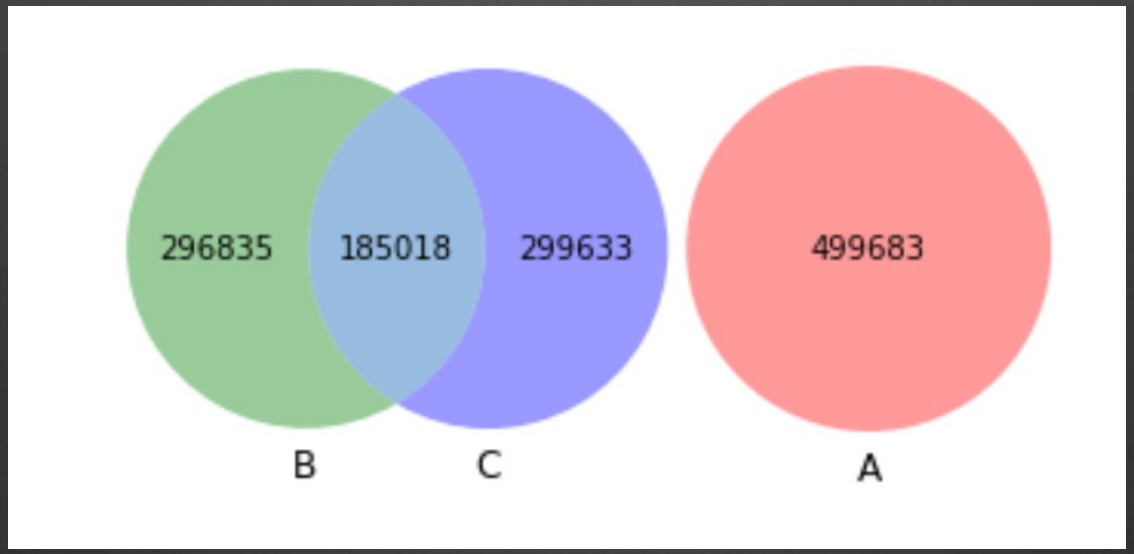
You might try sourmash if you want to -

- identify which reference genomes to use for metagenomic read mapping;
- search all Genbank microbial genomes with a sequence query;
- cluster hundreds or thousands of genomes by similarity;
- taxonomically classify genomes or metagenomes against NCBI and/or GTDB;
- search thousands of metagenomes with a query genome or sequence;

"multitool" for searching, comparing, and analyzing genomic & metagenomic data sets.

Open source software (BSD license)
Maintained, tested, and documented.
Developed "in the open" as well.

General approach: sequence comparisons with k-mer sets

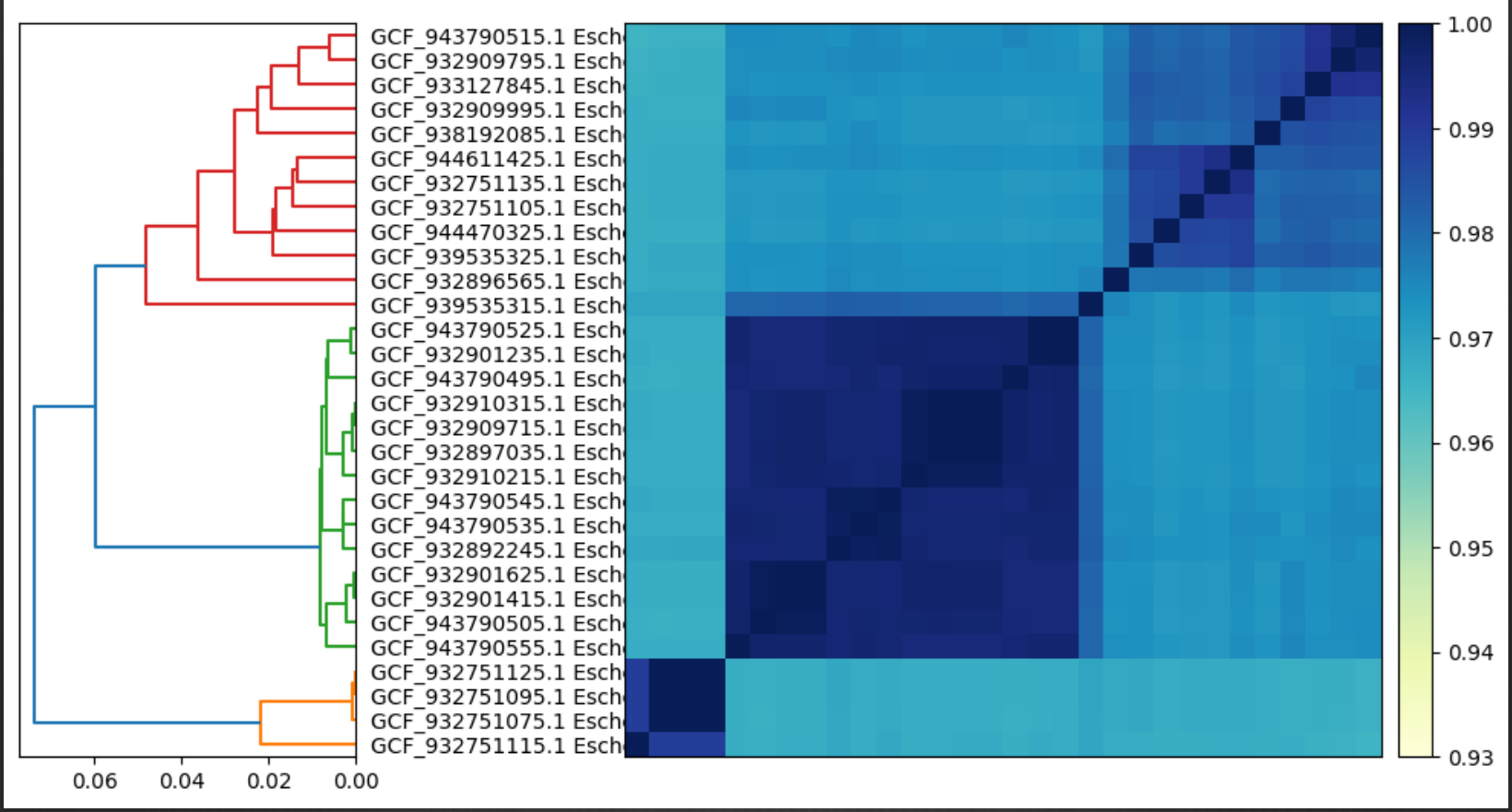


Jaccard similarity between B & C: 0.237

A k-mer based approach is *convenient*

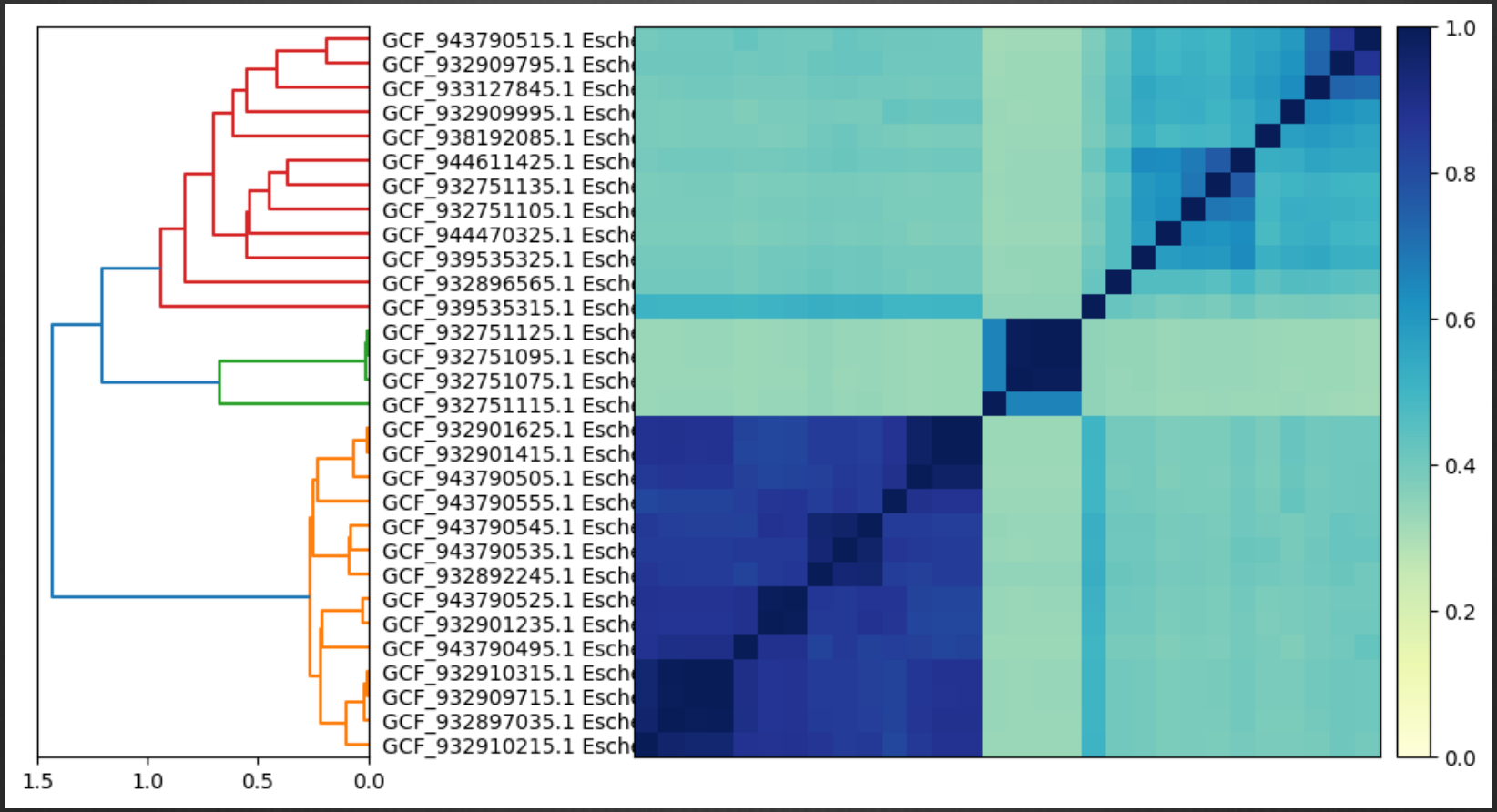
- Assembly-independent: everything works with or without good assemblies.
  - This is particularly important for metagenomes!
  - But this also enables mixture analysis for e.g. contaminants.
- Lightweight: new developments in sketching over the last decade have significantly accelerated the computation.
- Highly sensitive and specific.

Cluster *Escherichia* by ANI



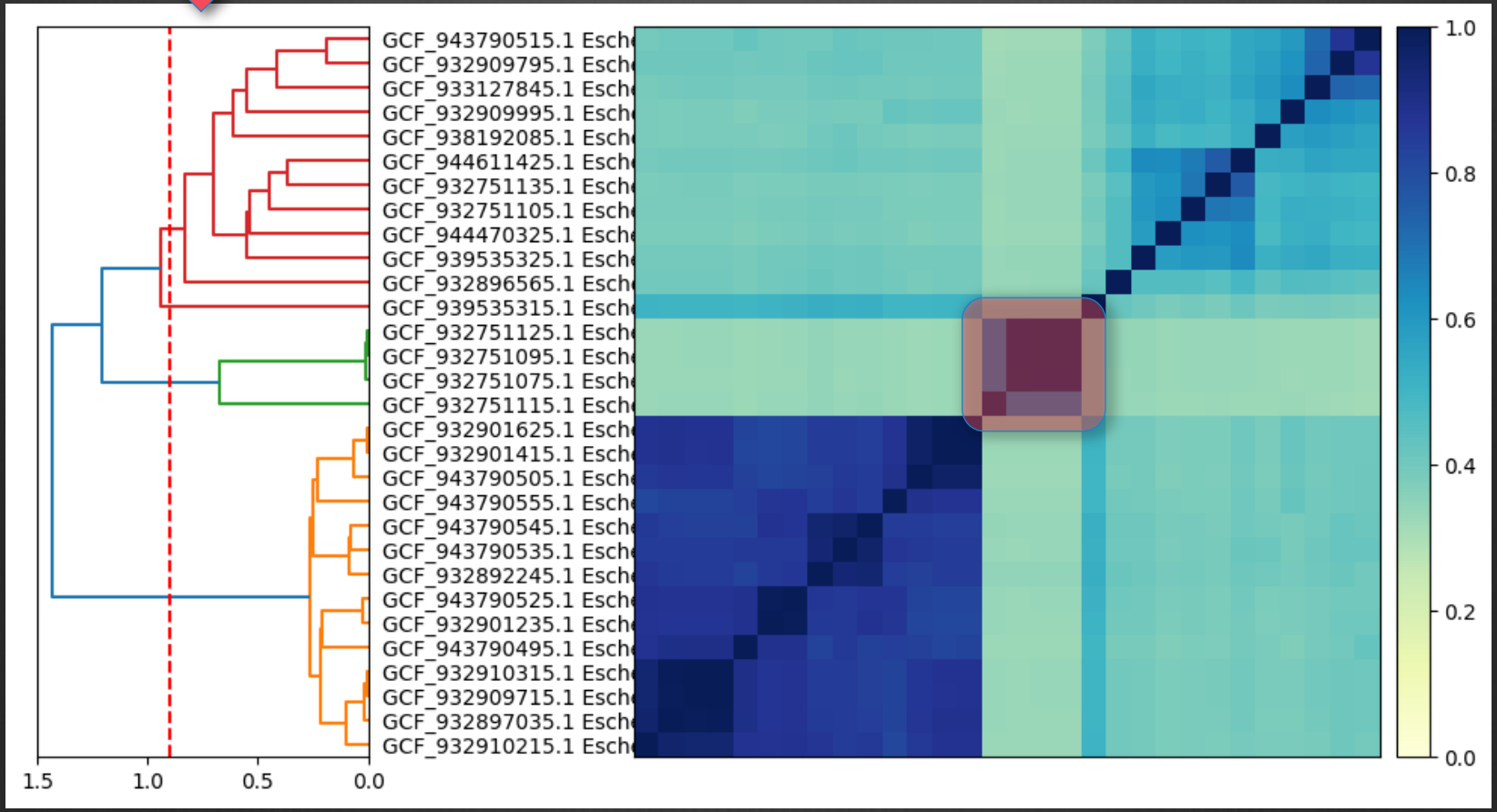
k=21

Reveal finer clade structure with Jaccard comparisons.



k=21

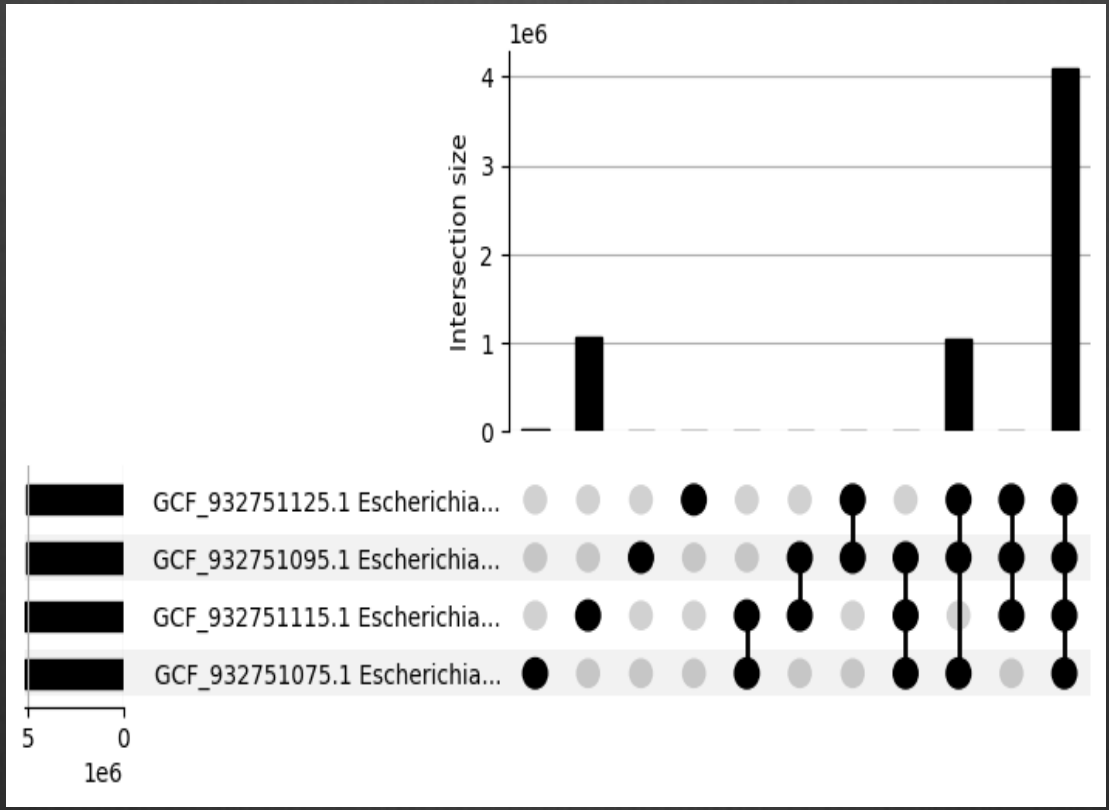
Extract clusters



k=21, cut=0.8

Examine cluster details

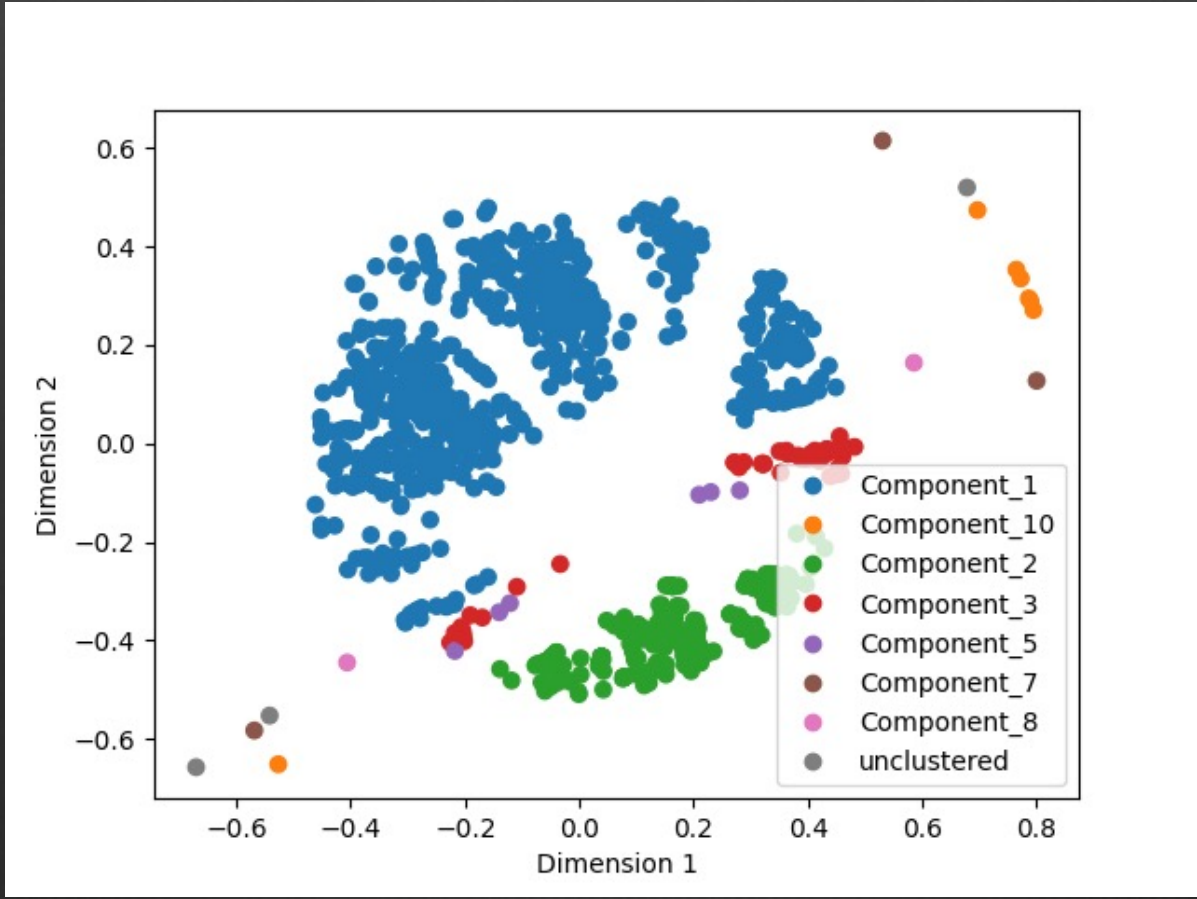
(pangenomic content overlap)



k=21

Work at (much) large scales.

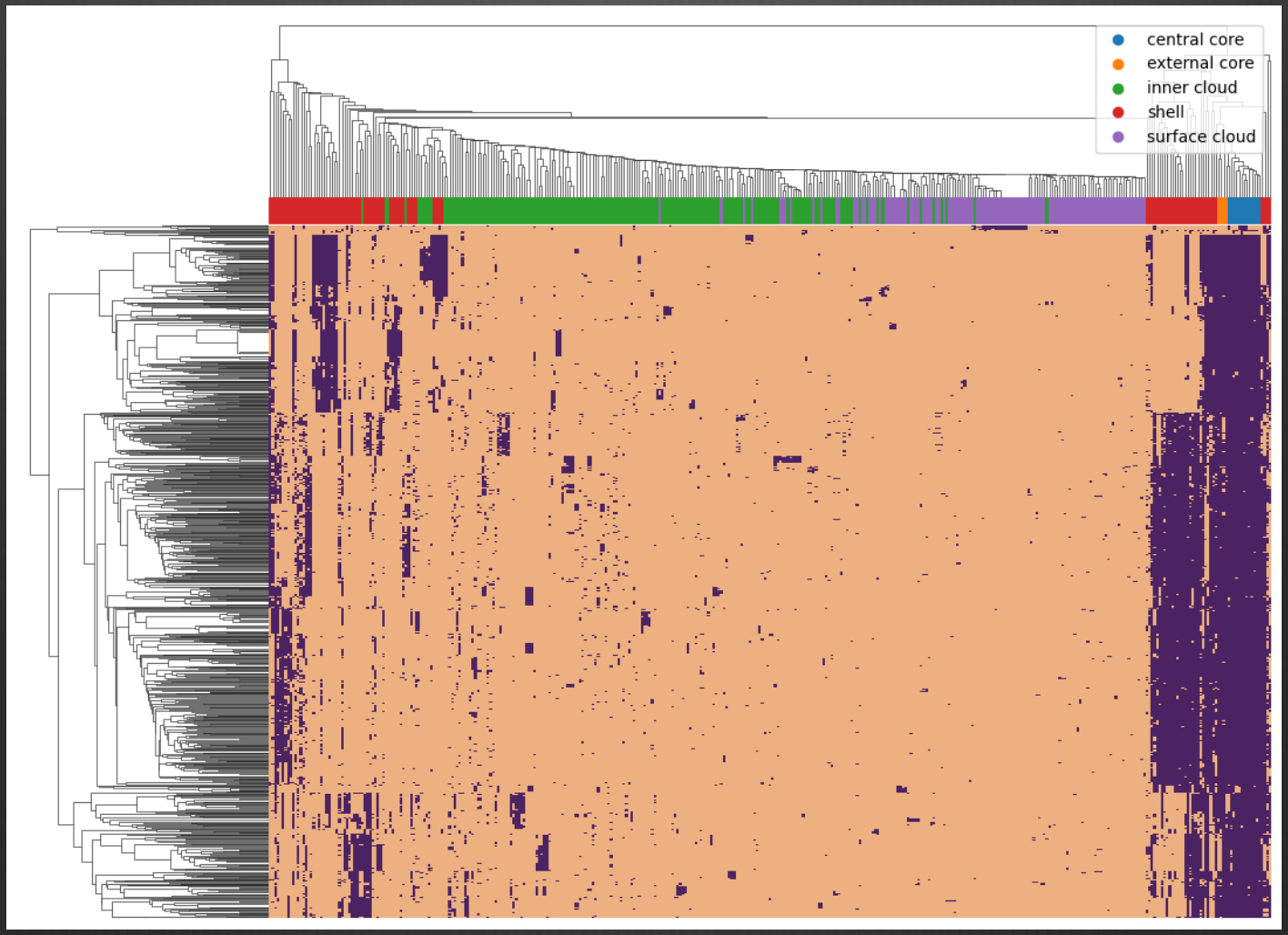
MDS plot



1000 *Escherichia*, k=21

Examine pangenomic structure

K-mer content (k=21)



Genomes (1000 *Escherichia*)

w/Colton Baumler and Anneliek ter Horst

Sourmash works with public *and* private collections of genomes.

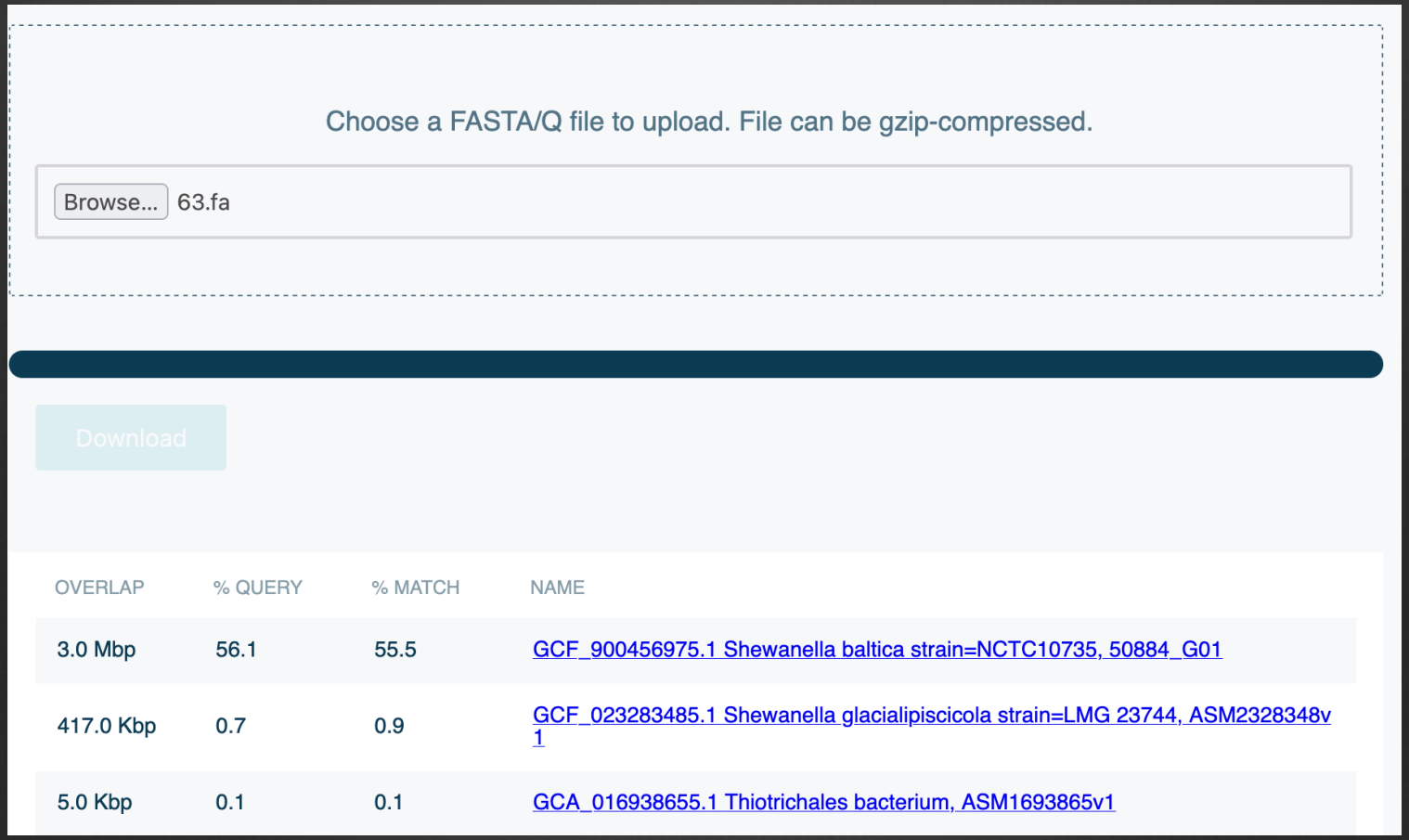
- There is no requirement for specific accessions, identifiers, etc.
- We also support multiple taxonomies, including NCBI and GTDB.
- We provide GTDB and Genbank indices.
- All of the sourmash commands support "n+1" mode where databases can be augmented by new genomes without rebuilding indices.

Sourmash also supports a kind of "differential privacy".

- Our sketching technique is *lossy* and *irreversible*, so sourmash databases can be made available for pre-publication and/or private data sets.
- This enables **genomic** search of private or unpublished culture collections without disclosure of full sequence; can tune resolution as desired. & includes metagenomics.
- (May also help with Nagoya Protocol issues?)
- Federated search is also possible, although not yet mature.

I'd love to talk more about the challenges and opportunities here!!

Real-time search on the Web



OVERLAP	% QUERY	% MATCH	NAME
3.0 Mbp	56.1	55.5	GCF_900456975.1 Shewanella baltica strain=NCTC10735, 50884_G01
417.0 Kbp	0.7	0.9	GCF_023283485.1 Shewanella glacialipiscicola strain=LMG 23744, ASM2328348v1
5.0 Kbp	0.1	0.1	GCA_016938655.1 Thiotrichales bacterium, ASM1693865v1

http://greyhound.sourmash.bio/

Real-time classification/search online using the sourmash library.

ACCESS MICROBIOLOGY
an open research platform

Volume 4, Issue 5

Meeting Report | Open Access

genomeRxiv: a microbial whole-genome database and diagnostic marker design resource for classification, identification, and data sharing

Leighton Pritchard[1], Parul Sharma[2], Reza Mazloom[2], Tessa Pierce[3], Luiz Irber[3], Bailey Harrington[1], Lenwood Heath[2], C Titus Brown[3] and Boris Vinatzer[2]

View Affiliations

Published: 27 May 2022 | https://doi.org/10.1099/acmi.ac2021.po0165

JOURNAL ARTICLE

Mibianto: ultra-efficient online microbiome analysis through *k*-mer based metagenomics

Pascal Hirsch, Leidy-Alejandra G Molano, Annika Engel, Jens Zentgraf, Sven Rahmann, Matthias Hannig, Rolf Müller, Fabian Kern, Andreas Keller ✉, Georges P Schmartz

Author Notes

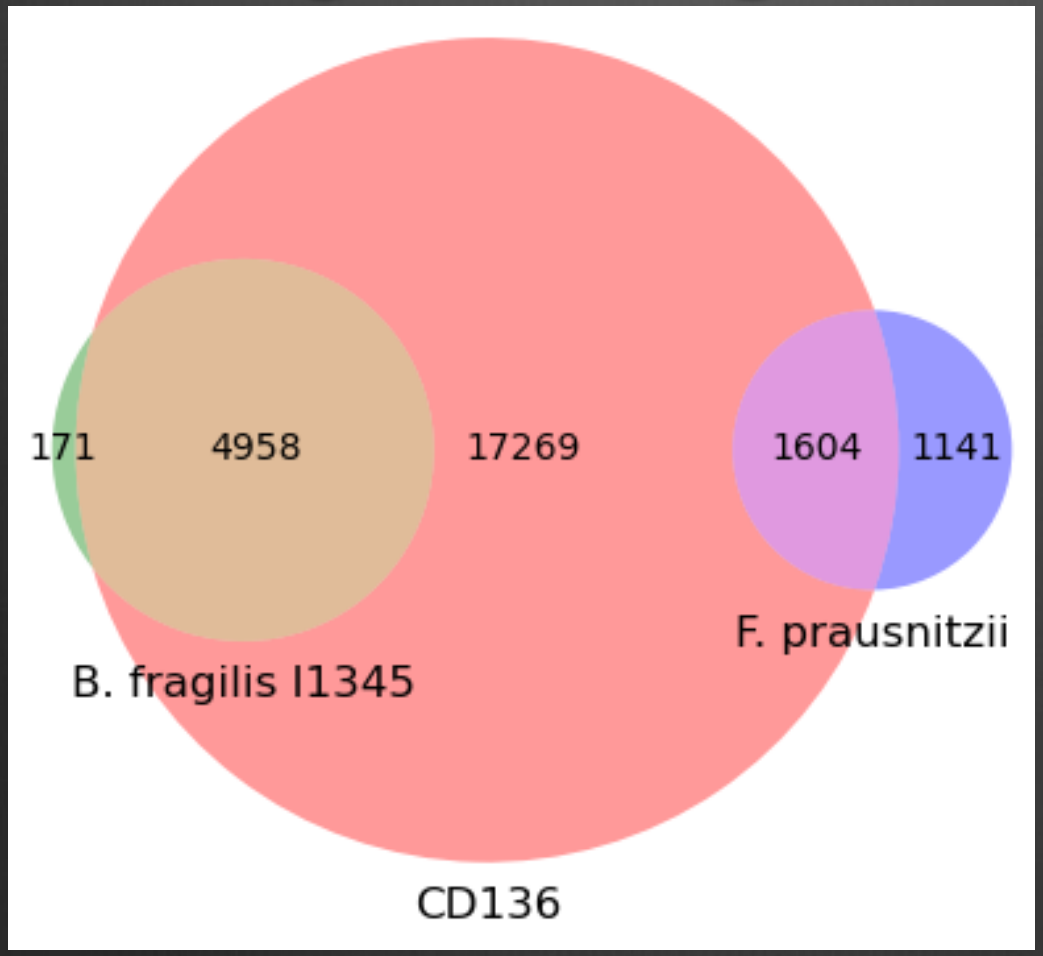
Nucleic Acids Research, gkae364, https://doi.org/10.1093/nar/gkae364

Published: 08 May 2024 **Article history** ▾

Metagenomics

- My core research interest is in applying these techniques to *shotgun metagenomes.*
- Metagenomes are rich & diverse, and hence much larger and more computationally challenging than microbial genome samples.
- We face many exciting challenges here!
  - Reference databases are extremely large;
  - We want strain-resolved analyses of metagenomes based on all available genomes;
  - Typically metagenomes do not closely match reference genomes.

Basic compositional analysis of a human gut metagenome, CD 136



Note: this *cannot* be done with MinHash, which only permits Jaccard, not overlap.

https://ngs-docs.github.io/2024-pig-paradigm-workshop/single-metagenomes-taxonomy/

=> Exhaustive decomposition of metagenomes

```
overlap     p_query p_match avg_abund
---------   ------- ------- ---------
5.0 Mbp       27.5%   96.7%       7.3    GCF_000598785.2 Bacteroides fragilis...
3.6 Mbp       10.3%   64.2%       3.8    GCF_009678525.1 Parabacteroides dist...
3.2 Mbp        4.6%   59.0%       1.9    GCF_015550345.1 Bacteroides uniformi...
1.6 Mbp       30.7%   58.4%      25.3    GCA_023708525.1 Faecalibacterium pra...
3.6 Mbp        1.2%    8.0%       3.8    GCF_009024595.1 Parabacteroides dist...
1.6 Mbp        5.6%   10.2%      25.1    GCF_017377615.1 Faecalibacterium sp....
2.6 Mbp        0.5%    3.3%       3.7    GCF_015548395.1 Parabacteroides dist...
1.6 Mbp        2.3%    5.1%      24.6    GCA_905199165.1 Faecalibacterium pra...
3.5 Mbp        0.2%    1.9%       2.8    GCA_009678725.1 Parabacteroides dist...
2.1 Mbp        0.1%    1.5%       1.9    GCF_009020375.1 Bacteroides uniformi...
3.3 Mbp        0.2%    1.5%       3.6    GCF_015552355.1 Parabacteroides dist...
1.6 Mbp        1.4%    2.4%      25.2    GCF_000166035.1 Faecalibacterium cf....
4.0 Mbp        0.3%    1.1%       7.3    GCF_009024655.1 Bacteroides fragilis...
1.6 Mbp        1.0%    1.9%      25.3    GCA_019425405.1 Faecalibacterium sp....
found less than 50.0 kbp in common. => exiting

found 14 matches total;
the recovered matches hit 86.0% of the abundance-weighted query.
the recovered matches hit 61.7% of the query k-mers (unweighted).
```

Can use any combination of public or private genomes.

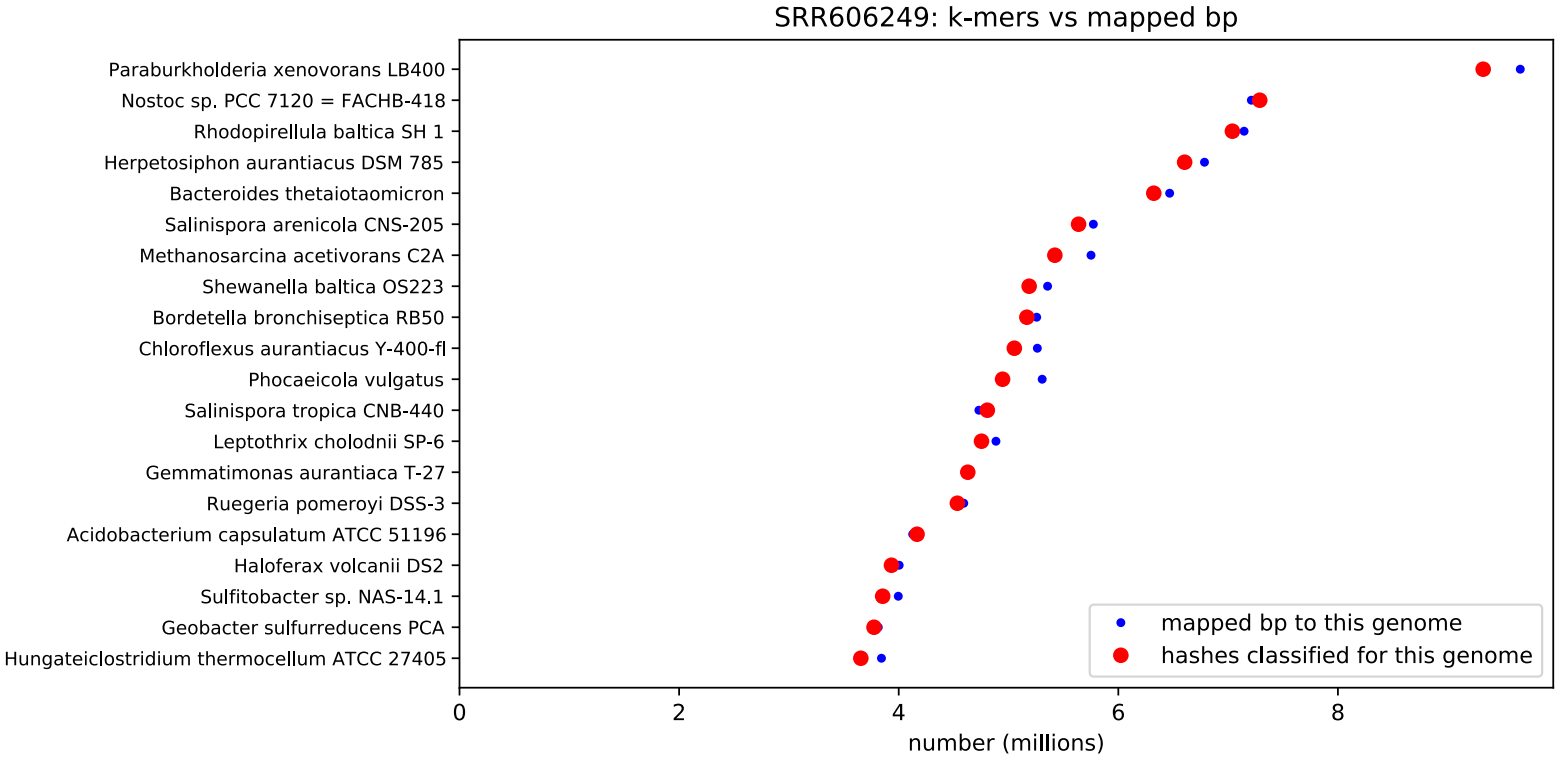
=> Taxonomic analysis

sample name	proportion	cANI	lineage
CD136	44.9%	98.0%	d__Bacteria;p__Bacteroidota;c__Bacteroidia
CD136	41.0%	92.5%	d__Bacteria;p__Bacillota_A;c__Clostridia
CD136	14.1%	-	unclassified

sourmash gather => sourmash tax

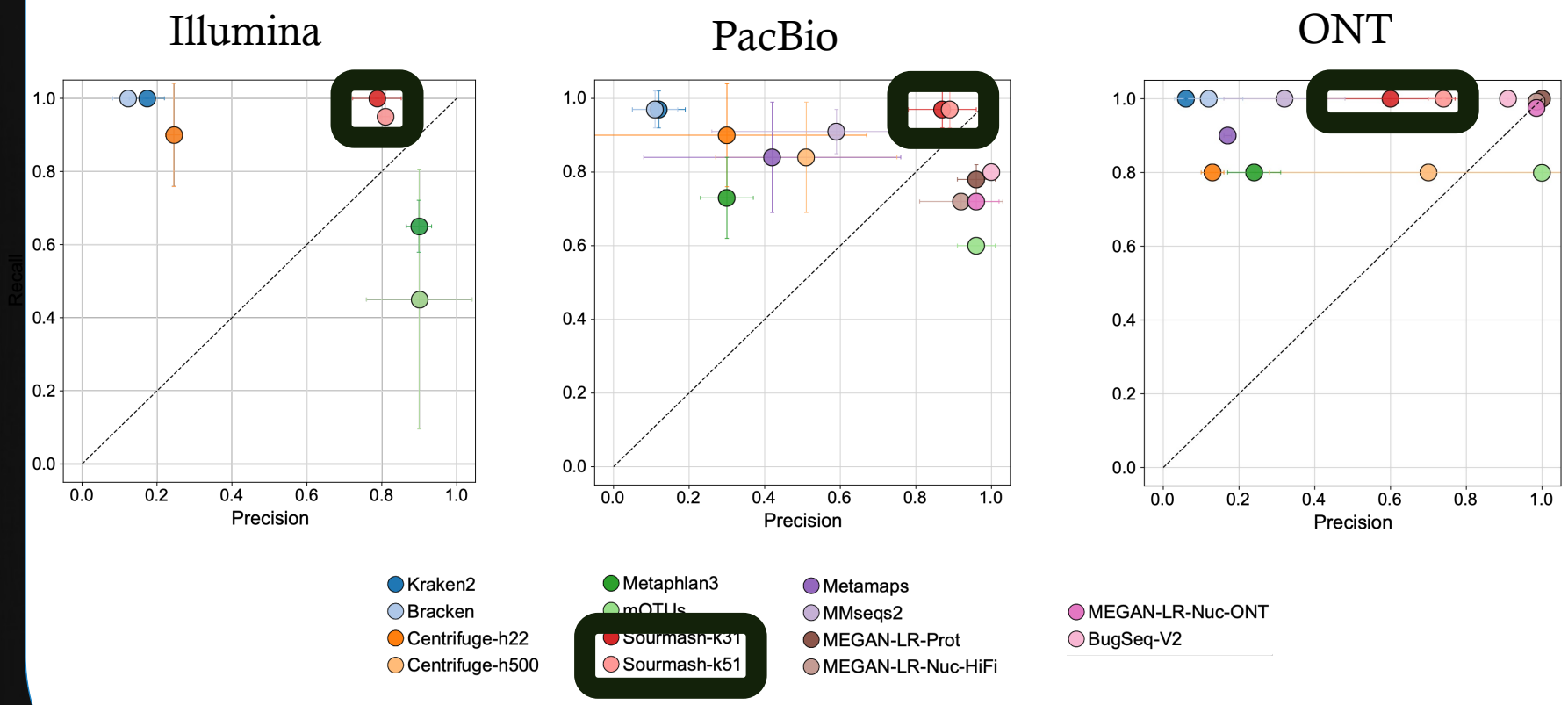
”Free taxonomy” approach – use NCBI, GTDB, LINS, ICTV, ??

Mapping metagenome reads to genomes closely matches k-mer estimates.



Irber et al., bioRxiv; 10.1101/2022.01.11.475838v2

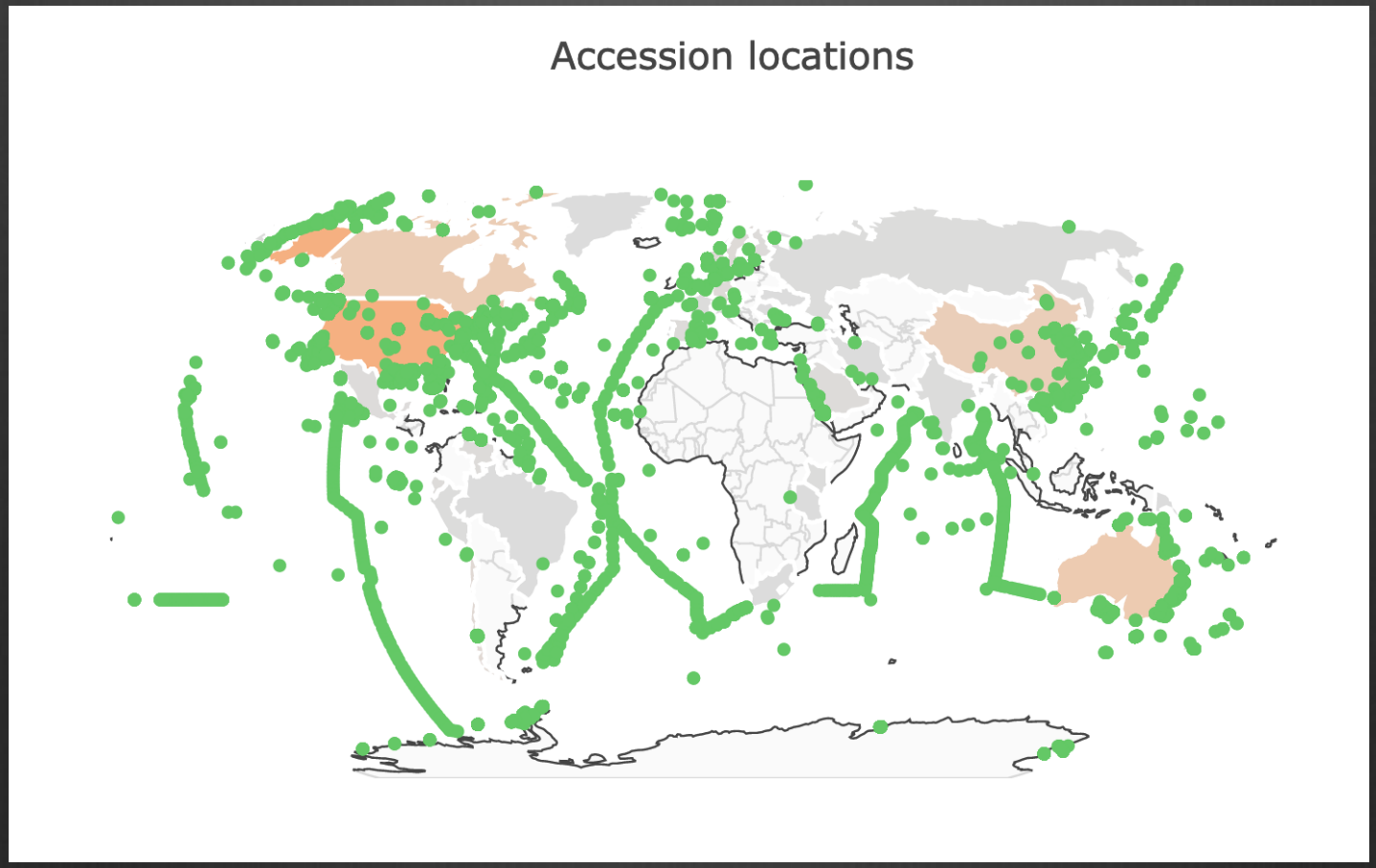
sourmash performs ***very*** well for species-level taxonomic assignment.



Portik et al., 2022

Where in the world is my genome??

(real-time search of 1m+ public metagenomes)



SAR11

A collaborative effort between JGI, UC-Davis, and the USDA ARS.

JGI JOINT GENOME INSTITUTE

UCDAVIS UNIVERSITY OF CALIFORNIA

USDA

+Suzanne Fleischmann & Adam Rivers

https://branchwater.jgi.doe.gov/

Concluding thoughts

- Sourmash is a k-mer multitool that supports a wide variety of basic k-mer analyses for assembled and unassembled data.
- Initially designed for metagenomics, but turns out to be useful beyond microbial data sets.
- Ability to scale ~1000x provides some interesting opportunities, too!
- As our data sets continue to grow, rapid and reliable approaches for exploring them are increasingly valuable!

Thanks for listening!

Please contact me at ctbrown@ucdavis.edu!

References

sourmash.readthedocs.org/ – sourmash docs

Lightweight compositional analysis of metagenomes with FracMinHash and minimum metagenome covers, Irber et al., 2022 (bioRxiv).

Evaluation of taxonomic classification and profiling methods for long-read shotgun metagenomic sequencing datasets, Portik et al., 2022.

Biogeographic distribution of five Antarctic cyanobacteria using large-scale k-mer searching with sourmash branchwater, Lumian et al., 2024.

branchwater.jgi.doe.gov/ - online metagenome search.

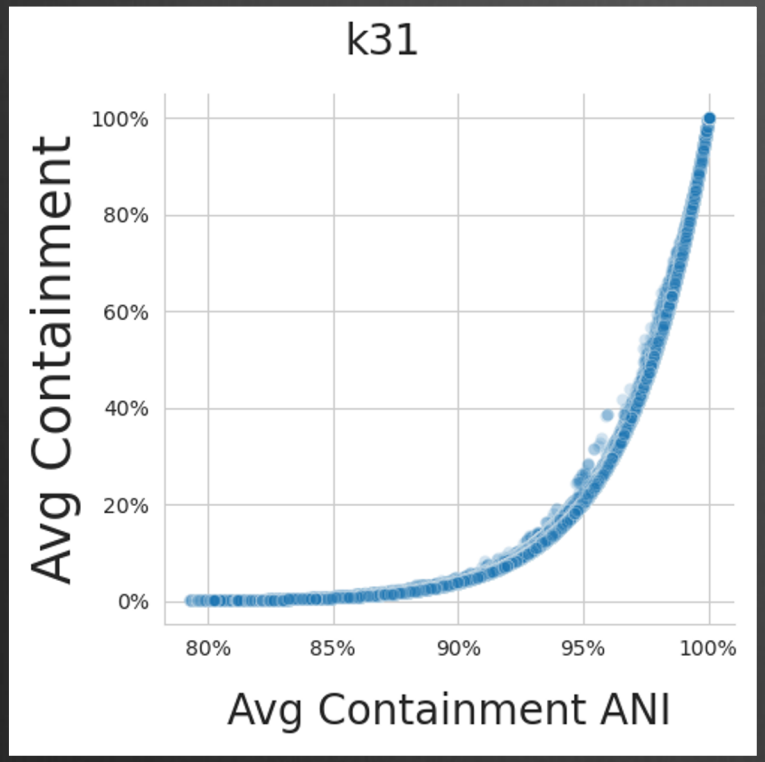
greyhound.sourmash.bio/ - online meta/genome analysis.

We take a k-mer-based approach:

```
[12]: build_kmers('ATGGACCAGATATAGGGAGAGCCAGGTAGGACA', 21)

[12]: ['ATGGACCAGATATAGGGAGAG',
       'TGGACCAGATATAGGGAGAGC',
       'GGACCAGATATAGGGAGAGCC',
       'GACCAGATATAGGGAGAGCCA',
       'ACCAGATATAGGGAGAGCCAG',
       'CCAGATATAGGGAGAGCCAGG',
       'CAGATATAGGGAGAGCCAGGT',
       'AGATATAGGGAGAGCCAGGTA',
       'GATATAGGGAGAGCCAGGTAG',
       'ATATAGGGAGAGCCAGGTAGG',
       'TATAGGGAGAGCCAGGTAGGA',
       'ATAGGGAGAGCCAGGTAGGAC',
       'TAGGGAGAGCCAGGTAGGACA']
```

K-mer containment can be converted to/from Average Nucleotide Identity.



Estimating the average containment to ANI relationship for k = 31

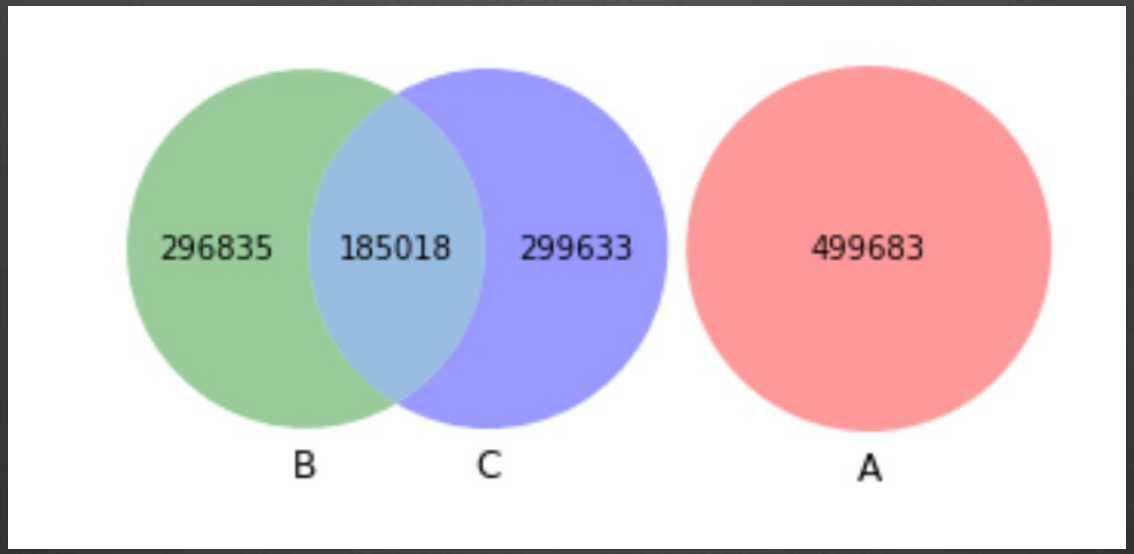
K-mer containment is something we can calculate with our tool quickly and easily from *unassembled* sequence data.

Average Nucleotide Identity can usually only be calculated from assemblies.

But we can convert between alignment free and alignment-based measures!

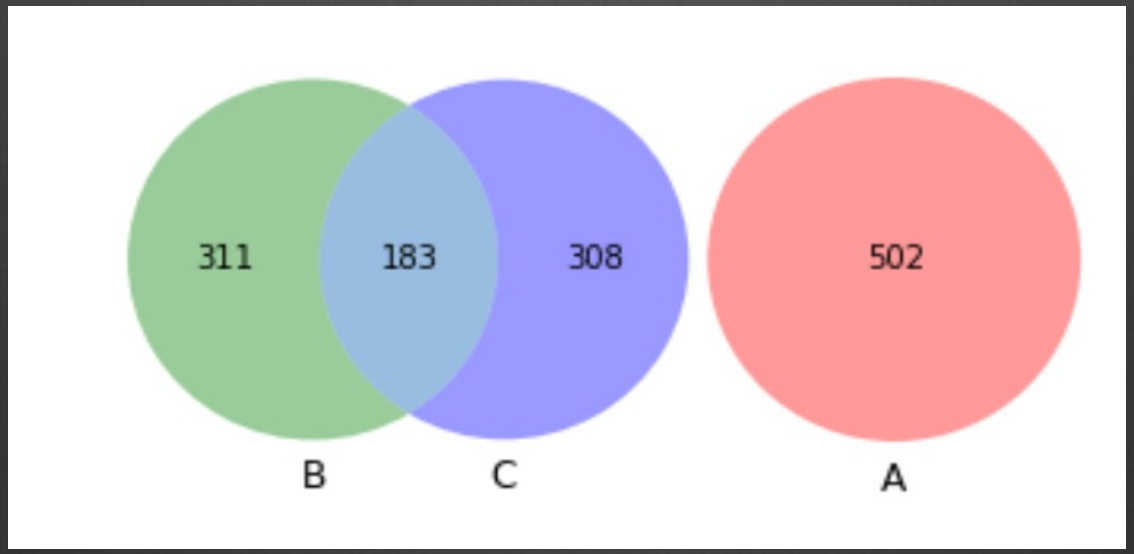
Work by Tessa Pierce-Ward, David Koslicki, and Mahmud Rahman. Rahman Hera et al., 2022, https://doi.org/10.1101/2022.01.11.475870

Genome comparisons with k-mer sets



Jaccard similarity between B & C: 0.237

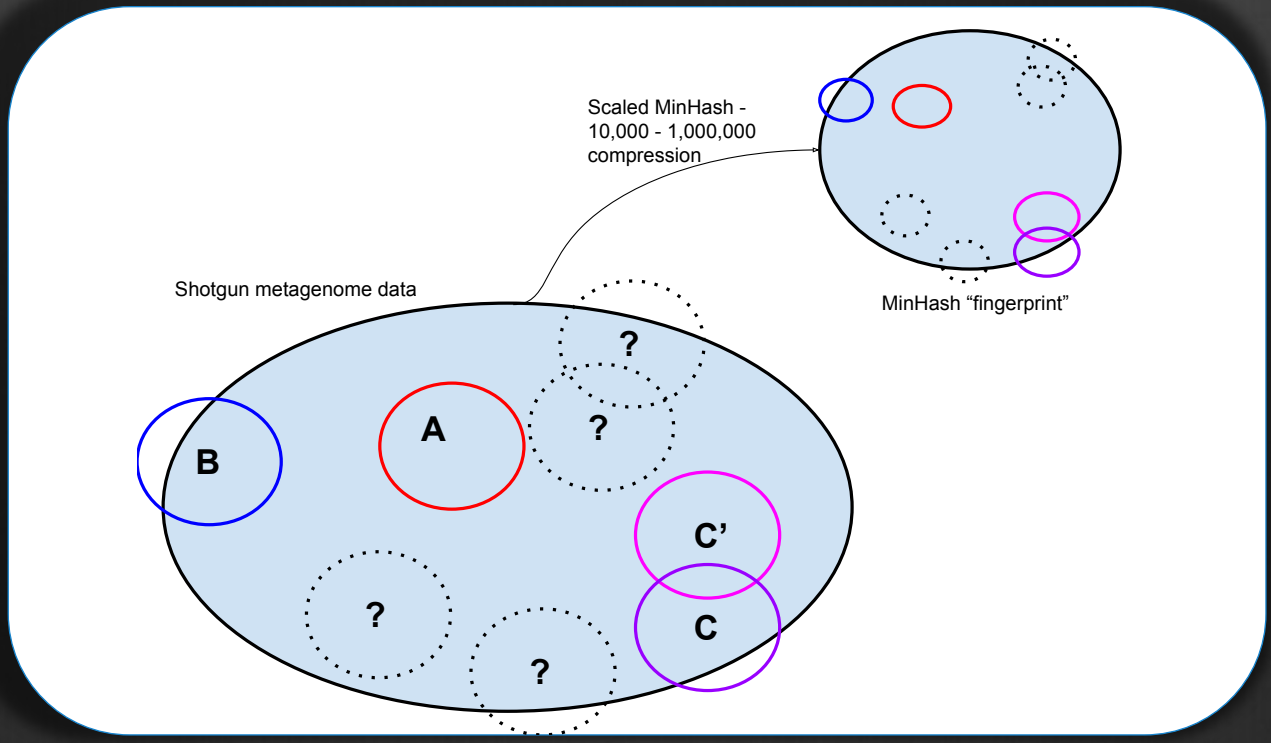
Genome comparisons with k-mer sketches (*compressed)*



Estimated Jaccard similarity between B & C: 0.228 (vs 0.237)

Alternatively ;) -

FracMinHash "shrinks" k-mer collections while still permitting many analyses.



https://f1000research.com/articles/8-1006

FracMinHash dramatically decreases data set size while still allowing search & comparisons

- A compression factor of 1000 is pretty good and brings even large metagenomic data sets into "laptop" range!
- You can do many operations on the sketches w/o looking at to original data
- 1.1m Genbank Bacterial genomes => 37 GB of sketches.
- Searching for matches to a genome against 60k bacterial genomes takes about 5 seconds and 139 MB of RAM.
- 85205 x 85205 (7.26 billion) comparisons in < 4 hours with 16 threads (and 4.42 GB RAM).

Features of FracMinHash

$$\mathbf{FRAC}_s(W) = \{ h(w) \leq \frac{H}{s} \mid \forall w \in W \}$$

- Tunable: adjust H/s (fraction of k-mers kept).
  - A s=scaled factor of 1000 => 1000-fold compression
- Can repeatedly downsample sketch to smaller fractions.
- Streaming compatible – no k-mer is every *removed*.
- Can take unions, intersections, and subtractions.
- Multiplicity (abundance) of k-mers can also be tracked.

Irber et al., bioRxiv; 10.1101/2022.01.11.475838v2

And then we added a nice front-end…

http://branchwater.sourmash.bio/

Suzanne Fleischmann (USDA), Tessa Pierce-Ward (UCD), Adam Rivers (USDA)

(Interactive demo if we have time!)

What data is being searched with branchwater?

Approximately 1m public data sets of type "metagenome" in the NCBI Sequence Read Archive.

We are searching ***unassembled*** data, so major biases will be sequencing and PCR biases, as well as limited depth of sequencing, but *not* limited by bioinformatics processing or known references.

All sketches together are > 10 TB.

This is a lossy compressed representation of 500 trillion distinct 31-mers across all metagenomes.

It represents 15-20 *petabytes* of original data.

What data is being searched with branchwater?

submitted as	distinct data sets
human gut metagenome	162187
metagenome	57048
gut metagenome	47244
human metagenome	36438
soil metagenome	35323
mouse gut metagenome	26482
human skin metagenome	25700
Homo sapiens	21020
marine metagenome	14400
human oral metagenome	14235

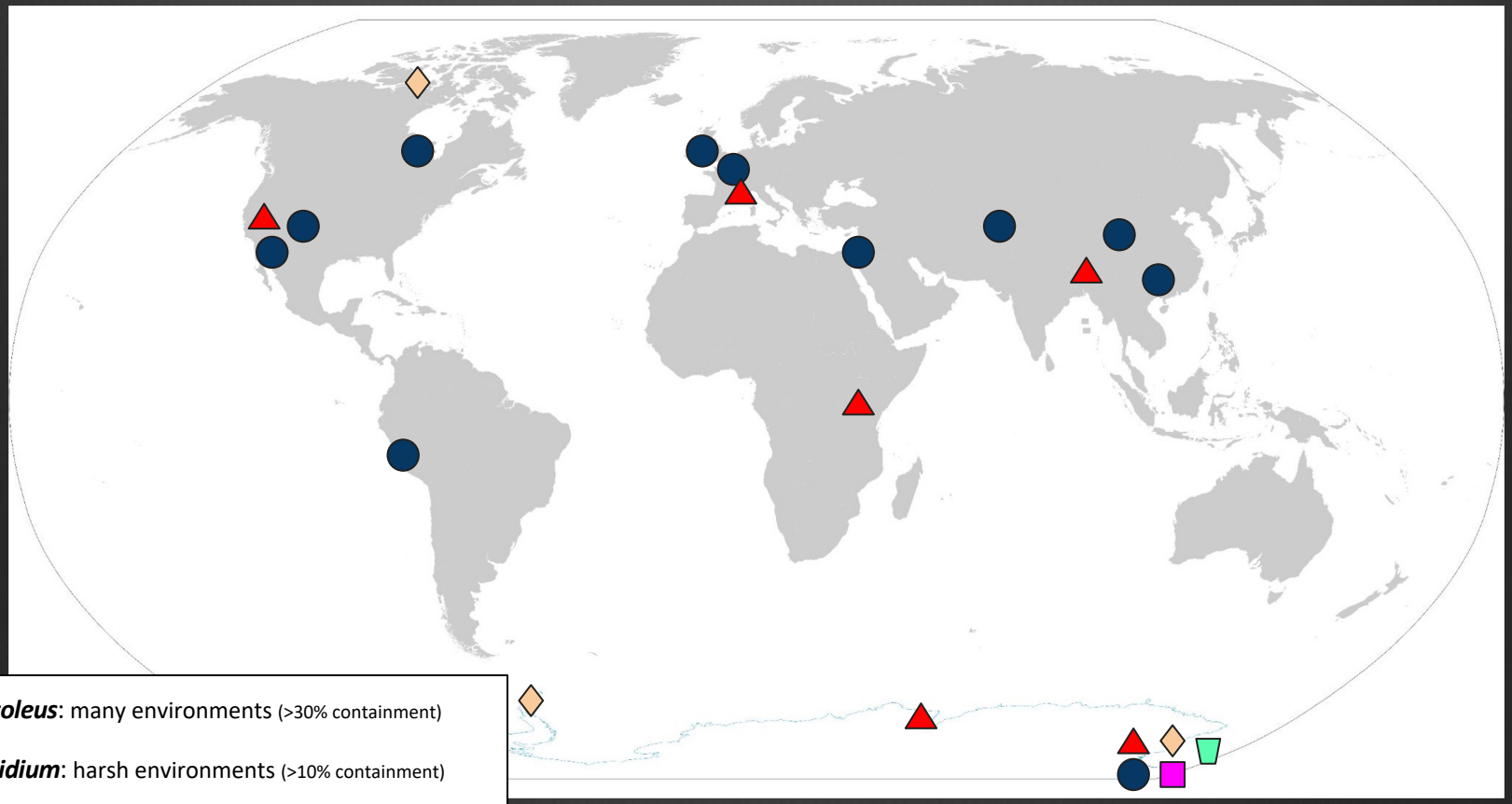
Nov 2022 breakdown ~800,000 public data sets of type "metagenome" in the NCBI Sequence Read Archive.

We are searching ***unassembled*** data, so major biases will be sequencing and PCR biases, as well as limited depth of sequencing, but *not* limited by bioinformatics processing or known references.

Vignette:

Where in the World are Antarctic Cyanobacteria?

(potential sampling locations!)



Microcoleus: many environments (>30% containment)

Phormidium: harsh environments (>10% containment)

Pseudanabaena: cold environments (>10% containment)

Leptolyngbya: Vanda and Fryxell (>10% containment)

(>10%
containment)

Lumian et al., 2024

What's a legitimate match?

Mapping-based validation confirms branchwater is high specificity.

Sample	31-mer containment	% genome mapped to	Effective coverage
SRR5468150 Mat lift-off from Lake Fryxell, Antarctica	99.18%	99.35%	125.84
SRR6266358 Polar Desert Sand Communities, Antarctica	65.02%	88.34%	93.34
SRR5247052 Sonoran Desert, Colorado Plateau, USA	41.10%	73.08%	180.87
SRR7428116 Les Salins du Lion Bird Reserve, France	20.63%	60.33%	11.68
SRR10186387 Salar del Huasco salt flat, Chile	8.98%	24.24%	3.64
SRR6262267 Human Gut	7.61%	25.27%	2.06

Genome matches

% genome mapped to is "detection" - how much of unweighted genome is in the sample?

Effective coverage is **weighted** abundance of detected genome.

Lumian et al., 2024

What's a legitimate match?

Mapping-based validation confirms branchwater is high specificity.

Sample	31-mer containment	% genome mapped to	Effective coverage
SRR5468150 Mat lift-off from Lake Fryxell, Antarctica	99.18%	99.35%	125.84
SRR6266358 Polar Desert Sand Communities, Antarctica	65.02%	88.34%	93.34
SRR5247052 Sonoran Desert, Colorado Plateau, USA	41.10%	73.08%	180.87
SRR7428116 Les Salins du Lion Bird Reserve, France	20.63%	60.33%	11.68
SRR10186387 Salar del Huasco salt flat, Chile	8.98%	24.24%	3.64
SRR6262267 Human Gut	7.61%	25.27%	2.06

Genome matches

% genome mapped to is "detection" - how much of unweighted genome is in the sample?

Effective coverage is **weighted** abundance of detected genome.

Lumian et al., 2024

What's a legitimate match?

Mapping-based validation confirms branchwater is high specificity.

Sample	31-mer containment	% genome mapped to	Effective coverage
SRR5468150 Mat lift-off from Lake Fryxell, Antarctica	99.18%	99.35%	125.84
SRR6266358 Polar Desert Sand Communities, Antarctica	65.02%	88.34%	93.34
SRR5247052 Sonoran Desert, Colorado Plateau, USA	41.10%	73.08%	180.87
SRR7428116 Les Salins du Lion Bird Reserve, France	20.63%	60.33%	11.68
SRR10186387 Salar del Huasco salt flat, Chile	8.98%	24.24%	3.64
SRR6262267 Human Gut	7.61%	25.27%	2.06

Genome matches

% genome mapped to is "detection" - how much of unweighted genome is in the sample?

Effective coverage is **weighted** abundance of detected genome.

Lumian et al., 2024

Vignette: the "Snipe" project.

Large scale sequence comparison (dog)

Preliminary Results (Mo Abuelanin)

1) **100-fold faster, 100-fold less memory, and 7,000-fold less disk space.**
2) Detected **duplicate** datasets.
3) Accurately estimated **depth**, **coverage**, and detected contamination.

Single Sample Benchmark: SRX9565374 | 778 Million Reads | 117.5 Giga base pair | 40X WGS

Classical Sequence Alignment

Step (32 cores)	Time	Memory	Disk
Alignment	4:22 hrs	11GB	182 GB BAM
Sorting	38 mins	29GB	18 GB BAM
Qualimap	33 mins	8.5GB	1.6 MB Report

5 Hours, 32 cores & 29 GB RAM utilizing 200GB of Disk Space

Our Alignment-Free Methods

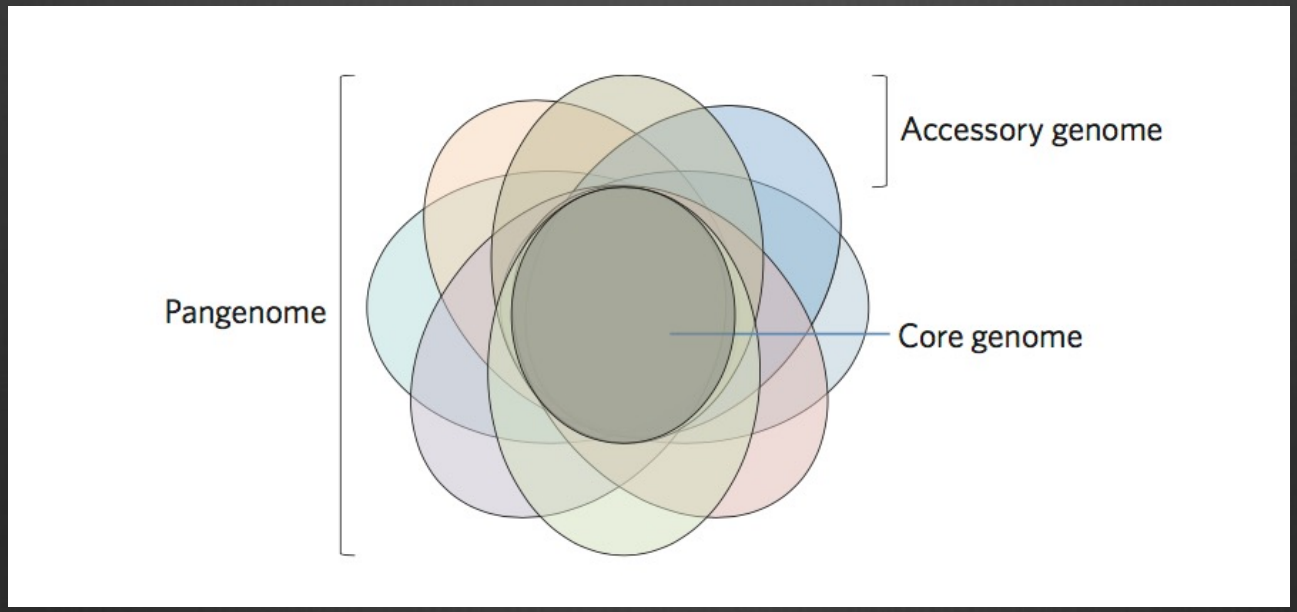
Step (1 core)	Time	Memory	Disk
Sketching	97 minutes	300MB	24 MB Sig
Assessment	<1 minute	200 MB	5 MB Report

1.6 Hours, 1 core & 300 MB RAM
utilizing **29 MB** of Disk Space

100x faster, 100x less RAM, 7000x less disk space than alignment based techniques.

Pangenomics and pan-meta-genomics

Most *species* have significant *strain* variation.



McInerny et al., DOI: 10.1038/nmicrobiol.2