Petabase-Scale Search and Containment Analysis with Fractional Sketches

...

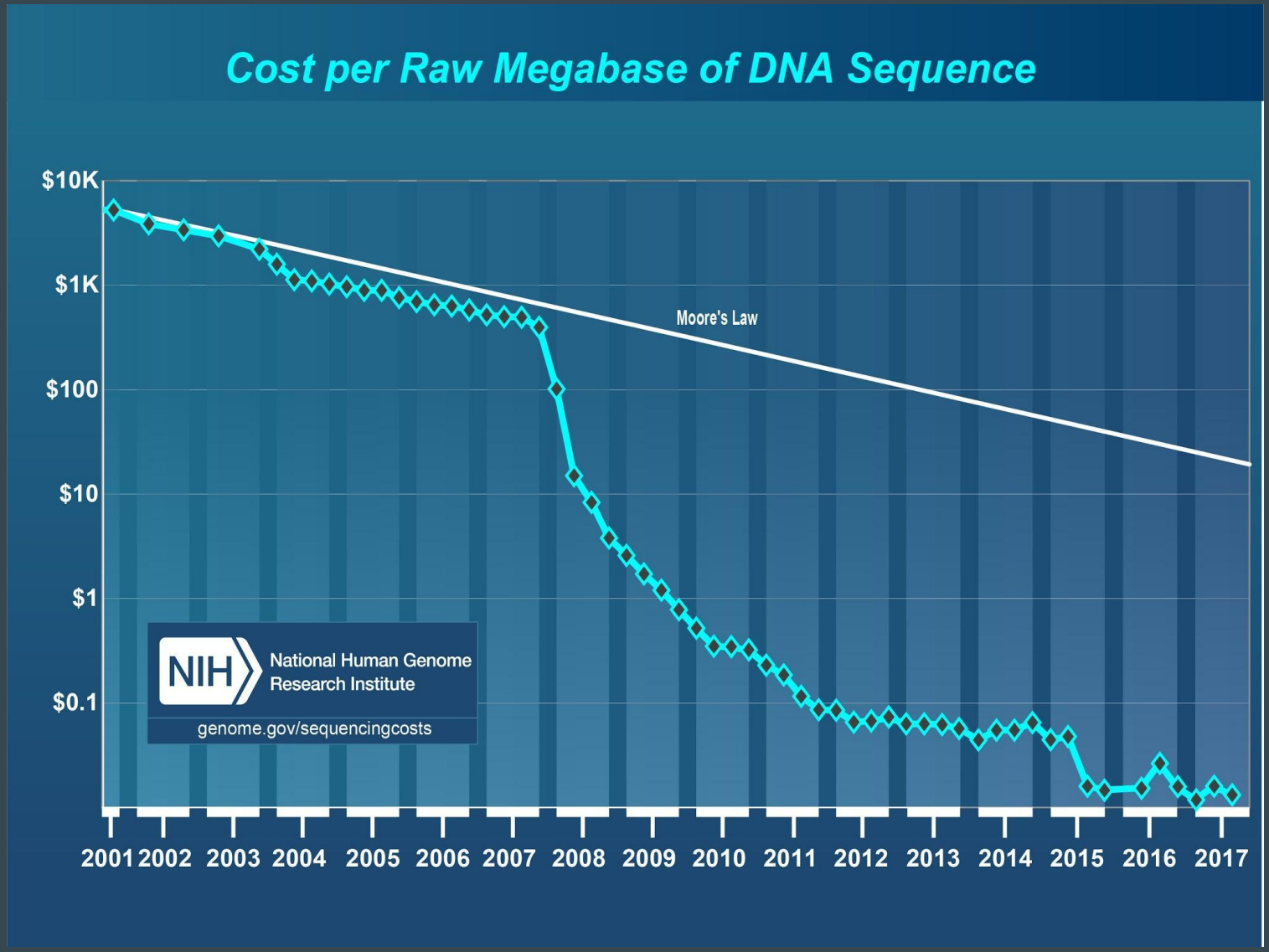
Luiz Irber
Computational Biologist @ 10x Genomics
CS PhD, UC Davis
2022/08/30

Acknowledgements

FracMinHash work is in collaboration with Mahmud Rahman Hera and David Koslicki (PSU).

Biogeography work is in collaboration with Jessica Lumian, Christy Grettenberger, and Dawn Sumner (UC Davis).

UCDAVIS VETERINARY MEDICINE



DNA Sequencing Costs

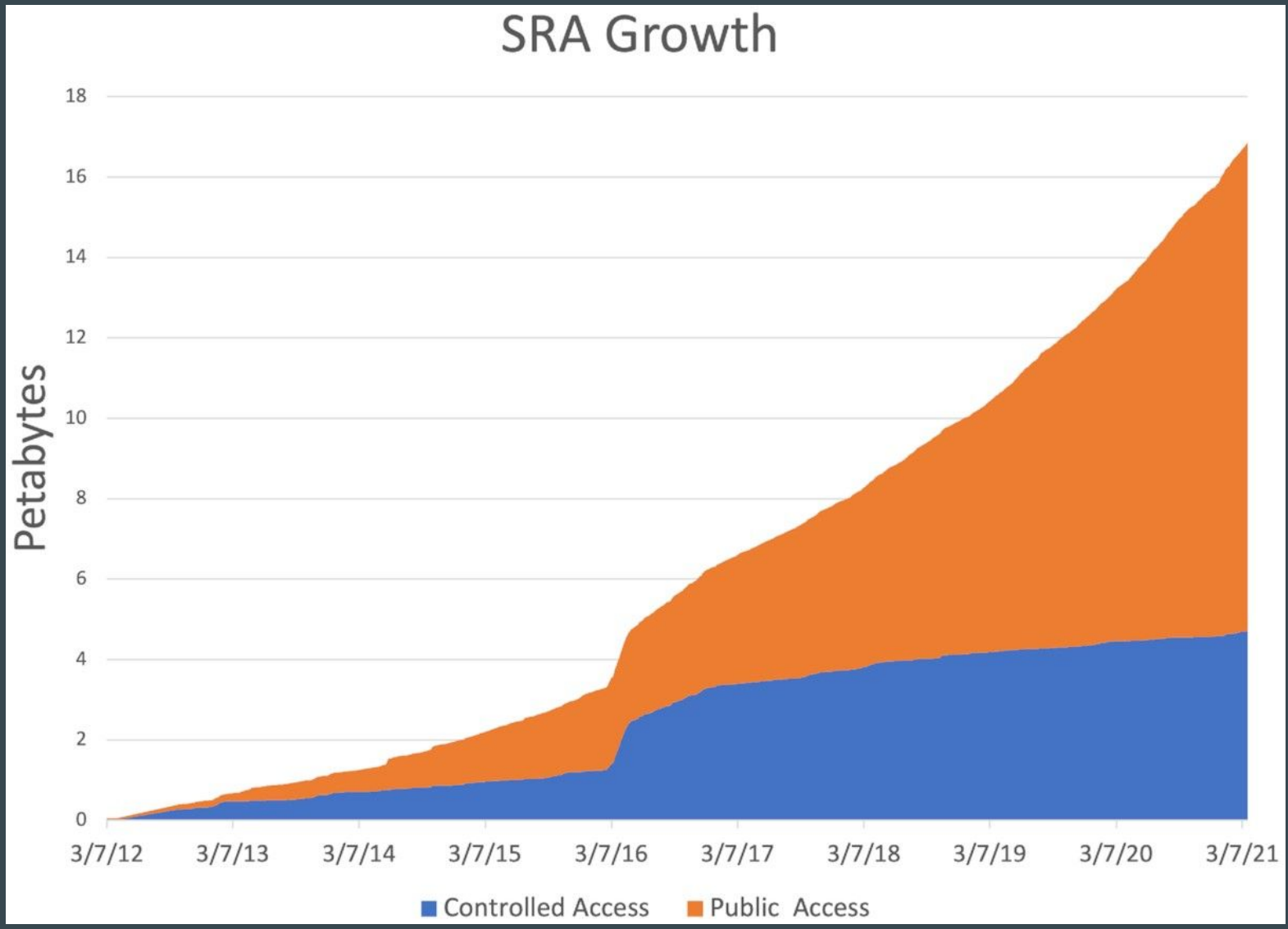
https://www.genome.gov/sequencingcostsdata/

Table 1. Growth of GenBank Divisions

Division	Description	Base pairs[a]	Annual increase[b]
VRL	Viruses	39 351 597 469	575.68%
UNA	Unannotated	4 421 782	550.93%
INV	Invertebrates	108 680 334 593	450.00%
ROD	Rodents	23 336 550 435	93.02%
PRI	Primates	15 165 437 356	72.97%
WGS	Whole genome shotgun data	13 888 187 863 722	57.08%
HTG	High-throughput genomic	27 800 219 072	0.07%
EST	Expressed sequence tags	43 324 455 796	0.05%
GSS	Genome survey sequences	26 380 049 011	0.01%
STS	Sequence tagged sites	640 923 137	0.00%
TOTAL	All GenBank sequences	15 309 209 714 374	54.79%

GenBank - Growth of the Database (2020 to 2021)

doi:10.1093/nar/gkab1135



The Sequence Read Archive: a decade more of explosive growth

doi:10.1093/nar/gkab1053

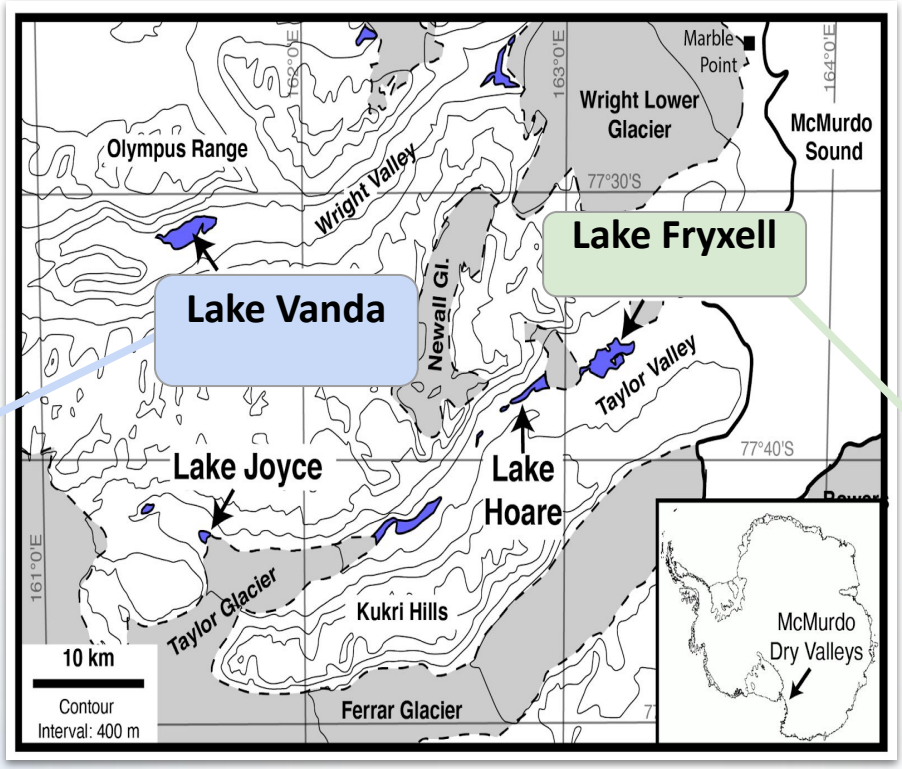
How do deal with the data deluge?

- Lightweight approaches for data analysis
  - Sketching and streaming
- Initial Exploration -> Select subset of all public data -> Deeper analysis
  - Not many people have 25PB available...
- What research questions can we rephrase?

Biogeography of Antarctic Cyanobacteria

Jessica Lumian, Christy Grettenberger, and Dawn Sumner

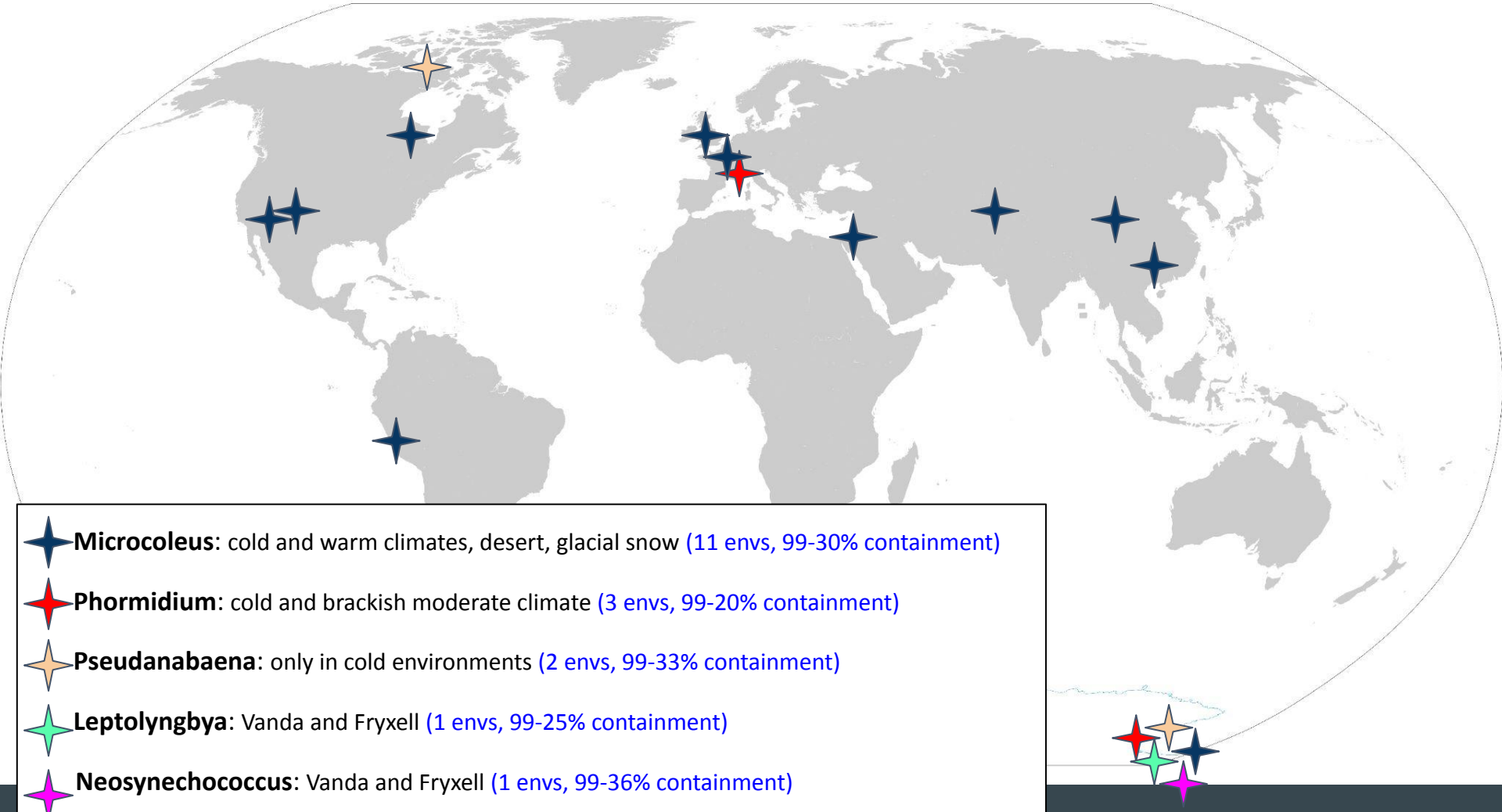
Leptolyngbya
Pseudanabaena
Neosynechococcus
Microcoleus



Phormidium pseudopriestleyi

Antarctic MAG Metrics from QUAST

Metric	*Leptolyngbya*	*Pseudanabaena*	*Microcoleus*	*Neosynechococcus*	*Phormidium pseudopriestleyi*
# of Contigs	654	505	671	552	678
Total length (bp)	5,729,854	2,764,673	6,072,304	4,887,811	5,965,908
GC content (%)	51.23	45.31	45.47	50.03	47.43
Completion (%)	92.57	74.2	88.88	92.39	91.73



Microcoleus: cold and warm climates, desert, glacial snow (11 envs, 99-30% containment)

Phormidium: cold and brackish moderate climate (3 envs, 99-20% containment)

Pseudanabaena: only in cold environments (2 envs, 99-33% containment)

Leptolyngbya: Vanda and Fryxell (1 envs, 99-25% containment)

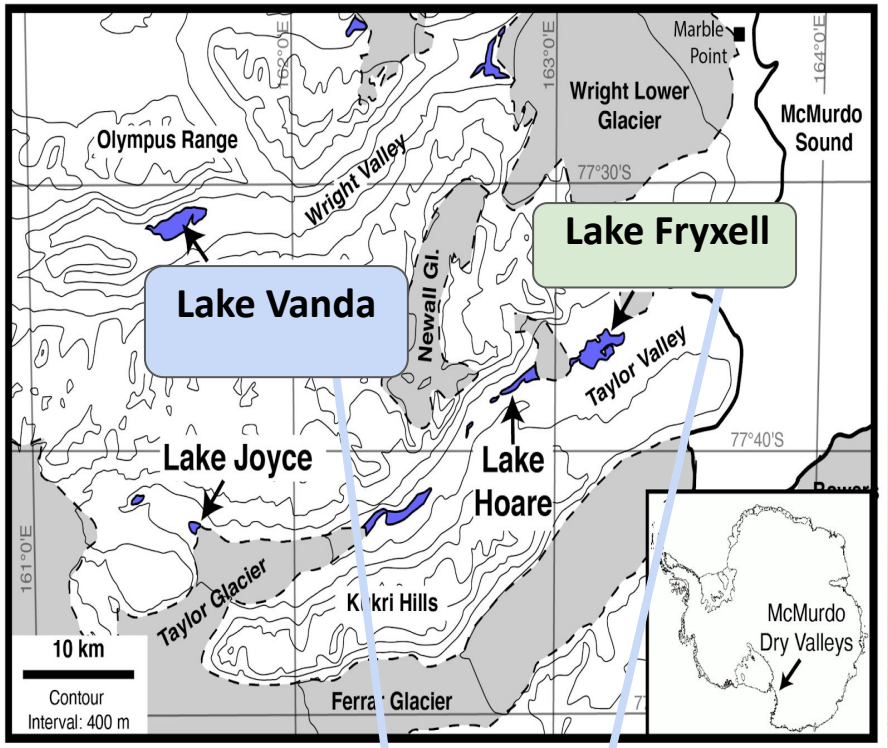
Neosynechococcus: Vanda and Fryxell (1 envs, 99-36% containment)

Biogeography of Antarctic Cyanobacteria

Jessica Lumian, Christy Grettenberger, and Dawn Sumner

MAGs in Predominantly Cold Environments

MAG	Environment	Number of Hits	Containment
Pseudanabaena	Lake Fryxell liftoff and glacier meltwater	6	99.49 % - 98.12%
	Dry Valley Sand Communities, Antarctica	1	37.45%
	Nunavut, Canada	3	33.54% - 26.16%
	Deception Island, Antarctica	3	18.31% - 16.43%
Neosynechococcus	Lake Fryxell liftoff and glacier meltwater	6	98.72 % - 25.16%
Leptolyngbya	Lake Fryxell liftoff and glacier meltwater	6	97.82% - 36.51%



- ➔ *Pseudanabaena*, *Neosynechococcus*, *Leptolyngbya* are in the cryosphere
- ➔ All Vanda MAGs were also found in Lake Fryxell but not from our samples!

Microcoleus is in a Variety of Environments

Environment	Number of Hits	Containment
Lake Fryxell, Antarctica liftoff and glacial meltwater	6	99.18% - 84.99%
Antarctic Desert Sand Communities	2	65.02% - 39.54%
Moab Soil Crust, USA	11	57.50% - 34.78%
Soil Crust, Ningxia, China	2	41.65% - 41.24%
Sonoran Desert, USA	1	41.10%
Pig Farm, UK	1	40.61%
Glacial snow, China	1	39.54%
Mine Tailing Pool, China	1	38.36%
Wastewater, Wisconsin, USA	1	37.04%
Puca Glacier, Peru	1	36.30%
Mediterranean Desert Community	3	35.71% - 34.30%
Arabidopsis community, Southwest Germany	1	33.57%

Polar

Non-Polar Desert

Glacial

Microcoleus is in many environments with different conditions

Phormidium pseudopriestleyi is in Harsh Environments

Environment	Number of Hits	Containment
Lake Fryxell, Antarctica mat	78	99.39% - 6.48%
Ace Lake, Antarctica (saline)	5	55.79% - 11.73%
Rauer Islands, Antarctica (saline)	3	22.92% - 21.16%
Lagoon, France (SL)	6	20.63% - 18.87%
Lagoon, France (EBD)	3	18.25% - 7.30%
Big Soda Lake, Nevada	3	17.71% - 8.48%
Antimicrobial treated sewage, Nairobi, Kenya	1	11.99%
Wetland Soil, Yanghu, China	1	10.37%

Saline

Polar

EBD Lagoon has sulfide and hydrocarbon pollution

Phormidium pseudopriestleyi grows slowly but survives challenging environments

mastiff: content-based SRA search

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

https://mastiff.sourmash.bio

- Disk-based sourmash index
- 486k+ SRA runs indexed

On 2,631 MAG queries

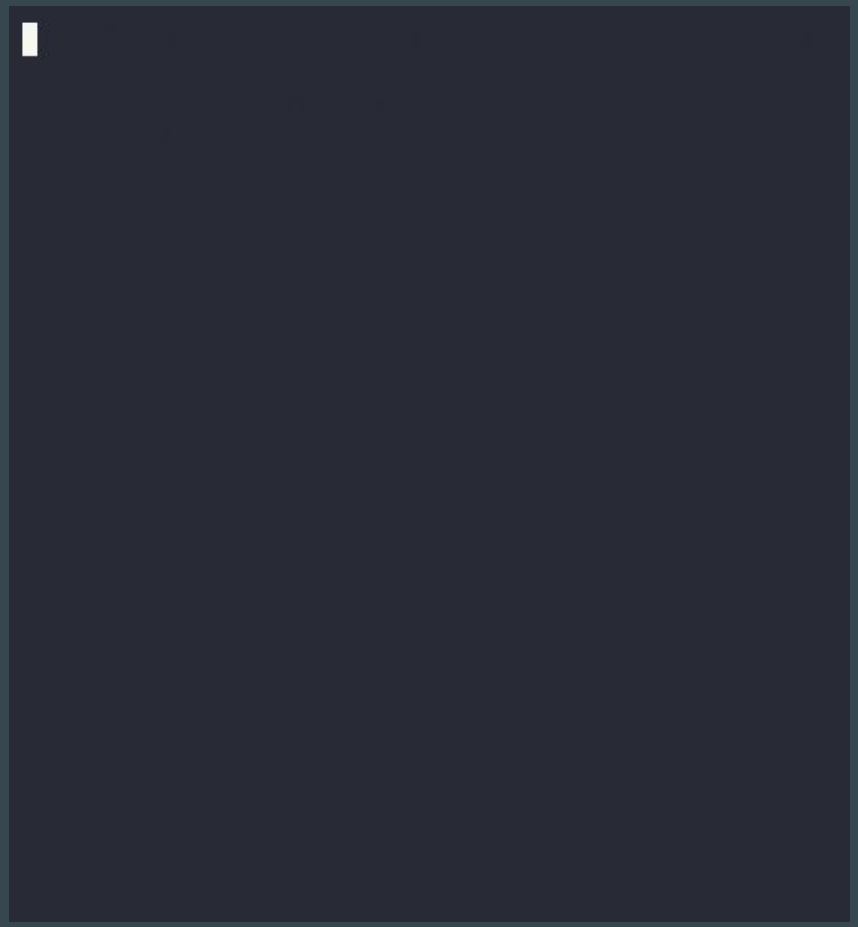
- 55 minutes, 24 cores, 540MB mem
- 10s/query

Returns SRA accession and *containment*

Sign up for more details!

Hands-On Petabyte Scale Sequence Search of SRA

Thursday, September 1st 8:00AM – 12:00PM



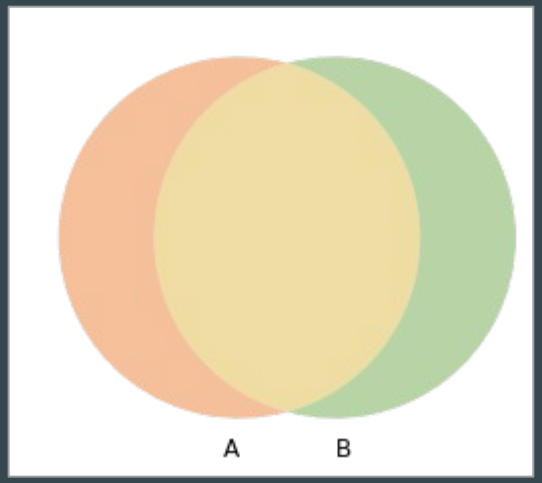
Containment and Similarity

$$C(A,B) = \frac{|A \cap B|}{|A|}$$

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

(Jaccard) Similarity

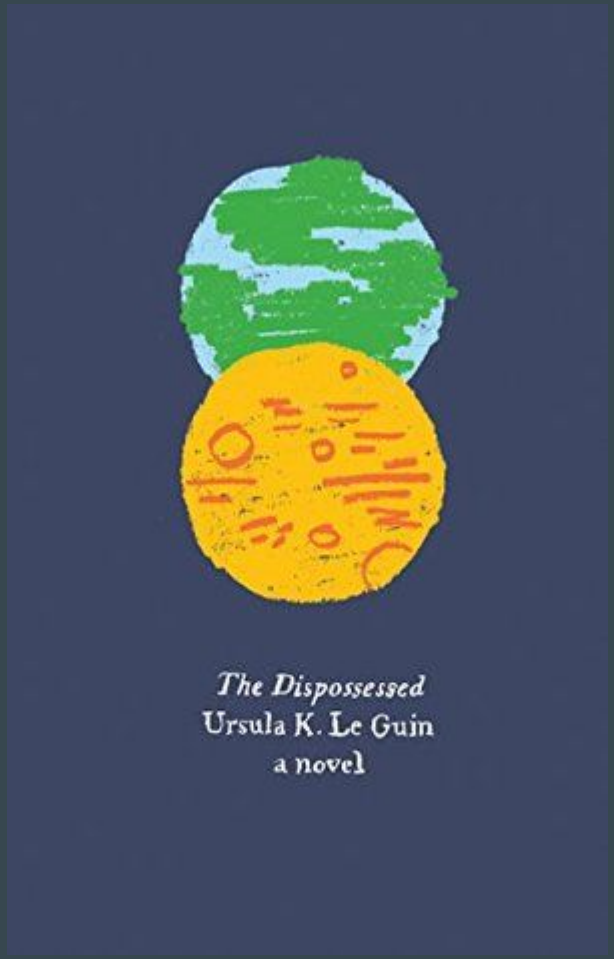
- How similar are two items to each other?
- Better when items about the same size
- Symmetrical: J(A, B) = J(B, A)

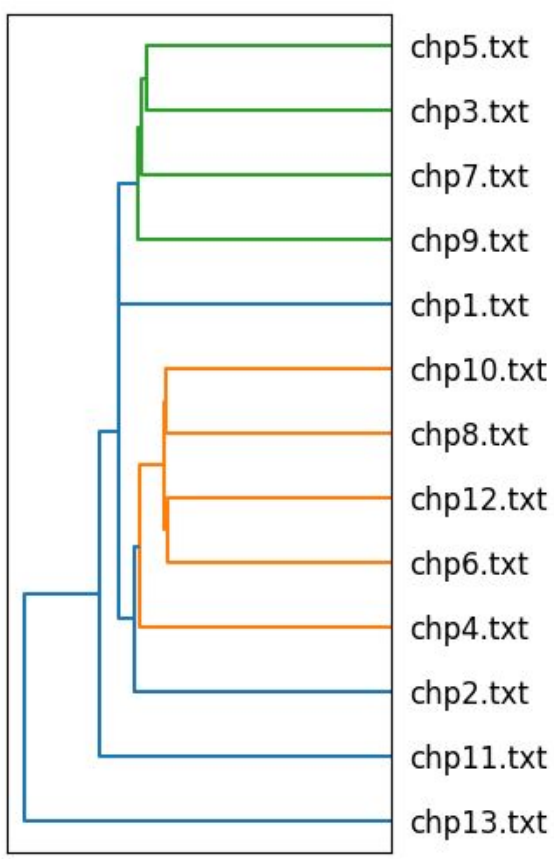


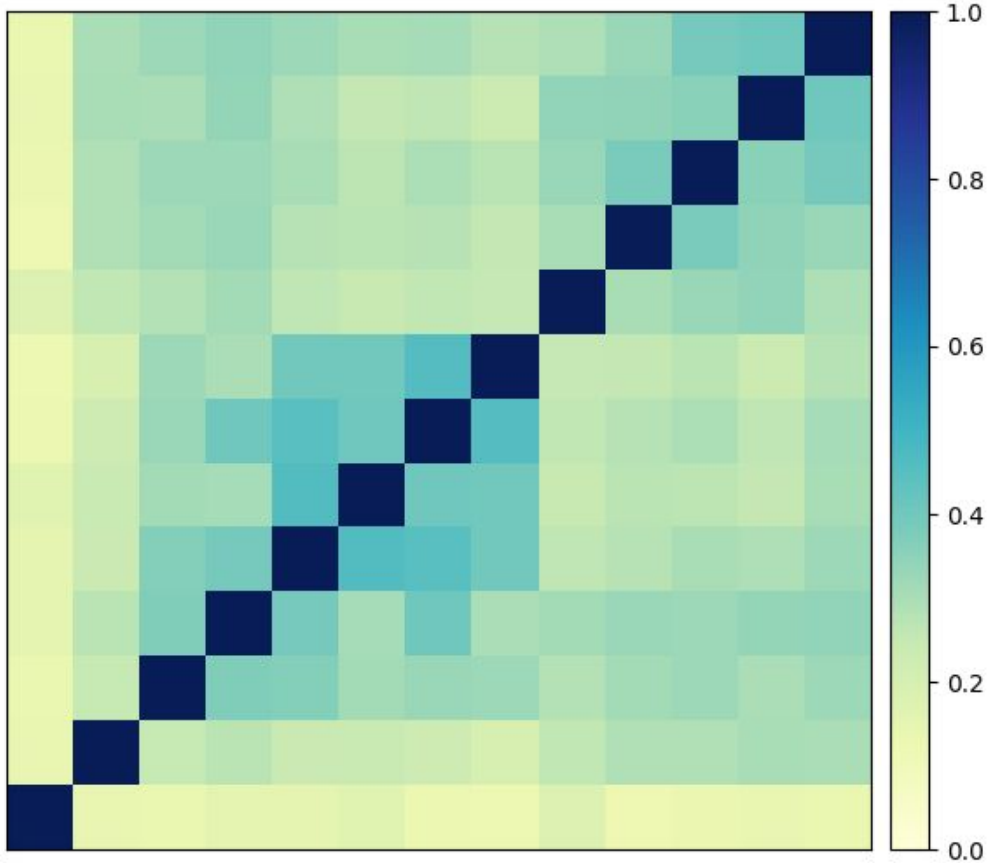
$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

A book example: The Dispossessed

- Structure: alternating chapters
- Two locations and timeframes:
  - Anarres/past
  - Urras/present
- First and last chapters are “space travel”
- Do the chapters cluster together?
- Processing chapters into sets
  - Tokenizing
  - Remove stop words
  - Stemming







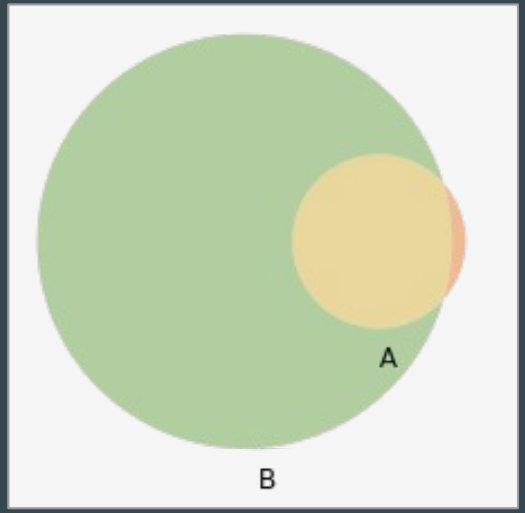
“The Dispossessed” chapters clustering

https://github.com/luizirber/2021-02-26-text-minhash

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Containment

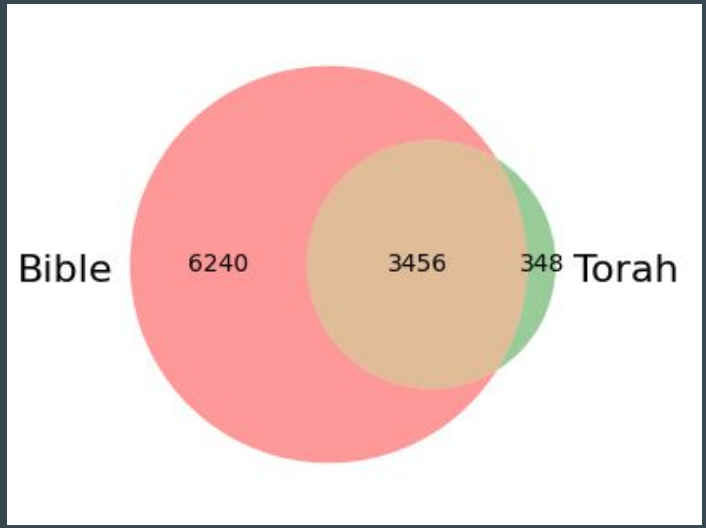
- How much of one is present in another?
- Better when items have different sizes
- Asymmetrical:
  - C(A, B) ≠ C(B, A)
  - A and B same size -> symmetrical



$$C(A,B) = \frac{|A \cap B|}{|A|}$$

A book example: Torah and Bible

- Torah: first 5 books of the Bible
- J(T, B) = 0.34
- C(B, T) = 0.35
- C(T, B) = 0.91

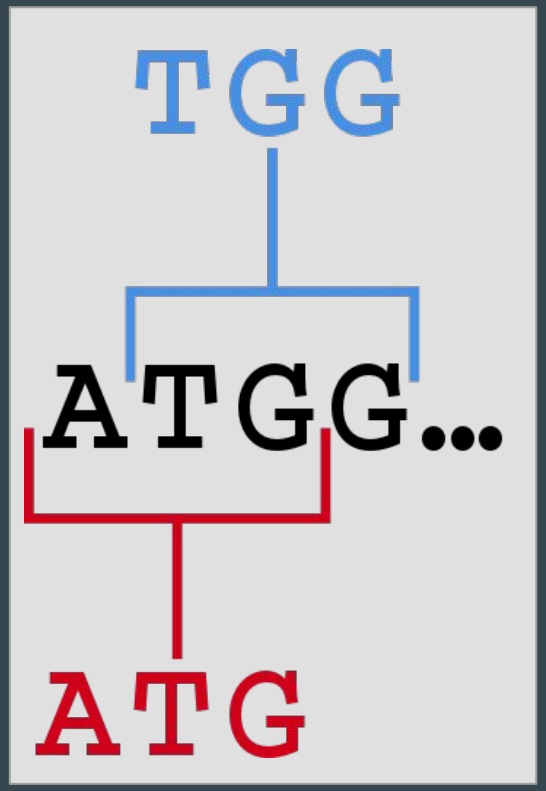


$$C(A,B) = \frac{|A \cap B|}{|A|}$$

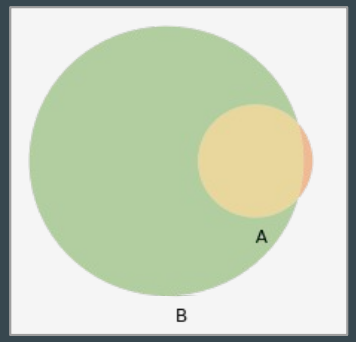
https://github.com/luizirber/2021-02-26-text-minhash

What are the elements for each item?

- Books: set of words
- Sequencing data: set of k-mers
- **Shingling**: the process of converting a dataset into elements of a set
- Conversion:
  - Words: tokenizing, remove stop words, stemming
  - Sequencing data: it's complicated.
  - Nucleotide k-mers as baseline



How do I choose?

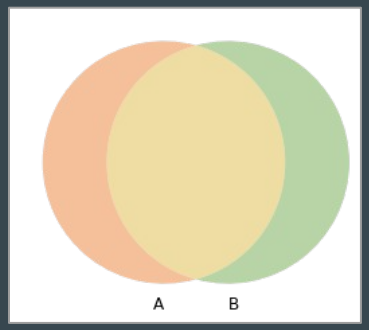


$$C(A,B) = \frac{|A \cap B|}{|A|}$$

- MAG/genome in Metagenome
- Contig in genome

Rule of thumb: |A| << |B|

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

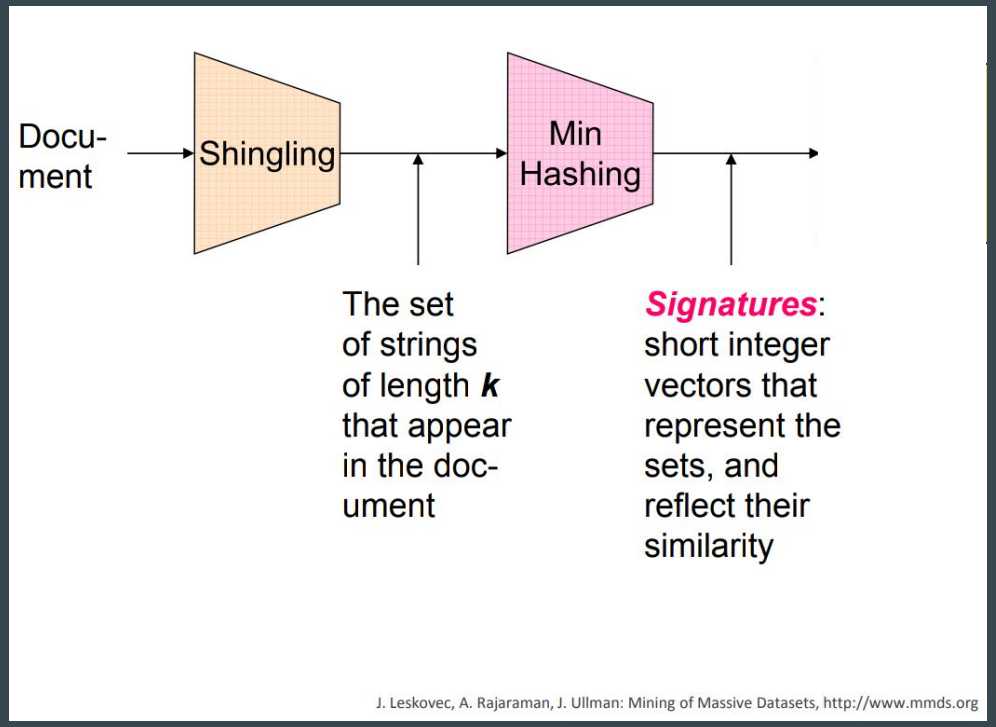


- Genomes
  - Same kingdom
- Chromosomes
- Metagenomes

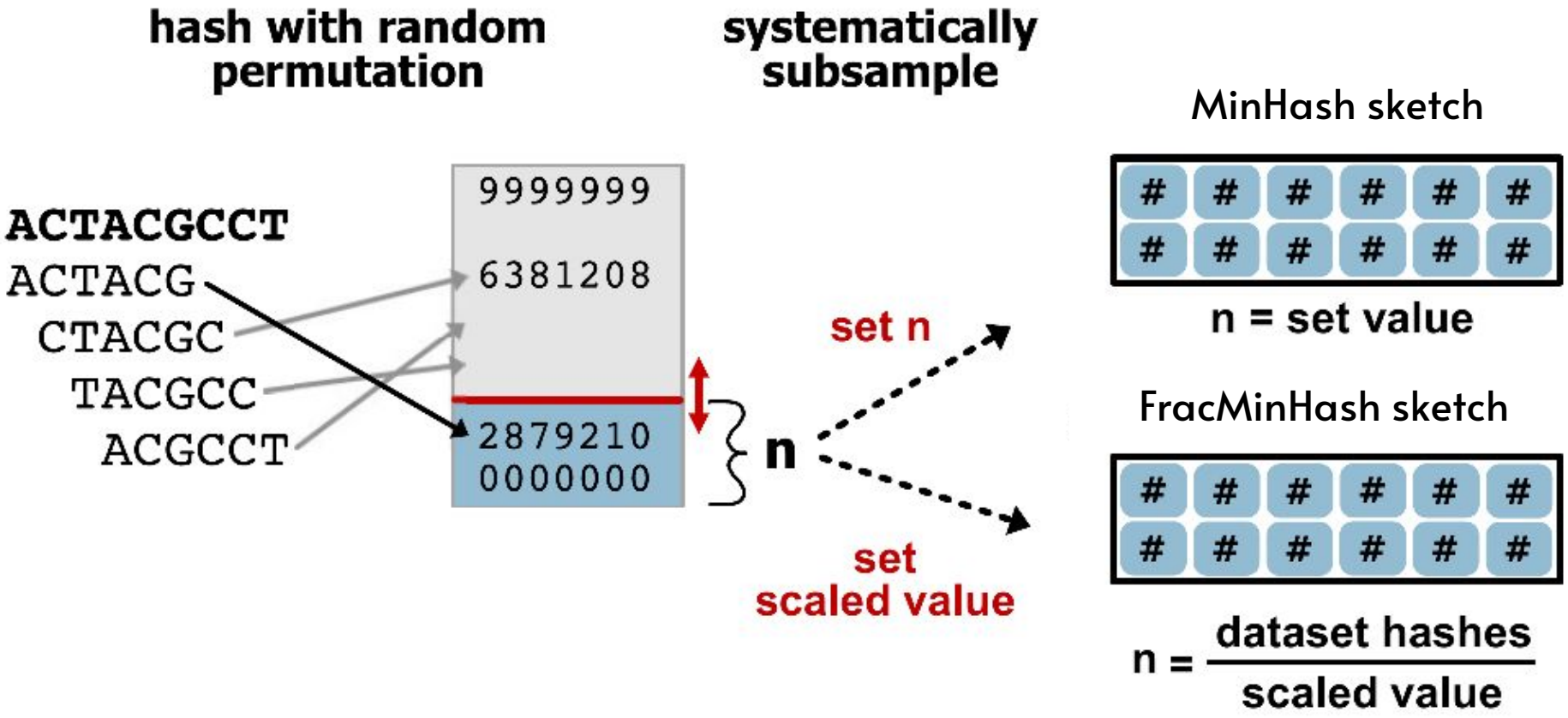
Rule of thumb: |A| ~ |B|

MinHash

- Comparing web pages (Altavista)
- Estimators
  - Resemblance (Jaccard similarity)
  - Containment
- MinHash: resemblance
  - Fixed size
- ModHash: resemblance and containment
  - Variable size
- Original focus on Resemblance
  - More computational-friendly



Broder 1997, doi:10.1109/SEQUEN.1997.666900



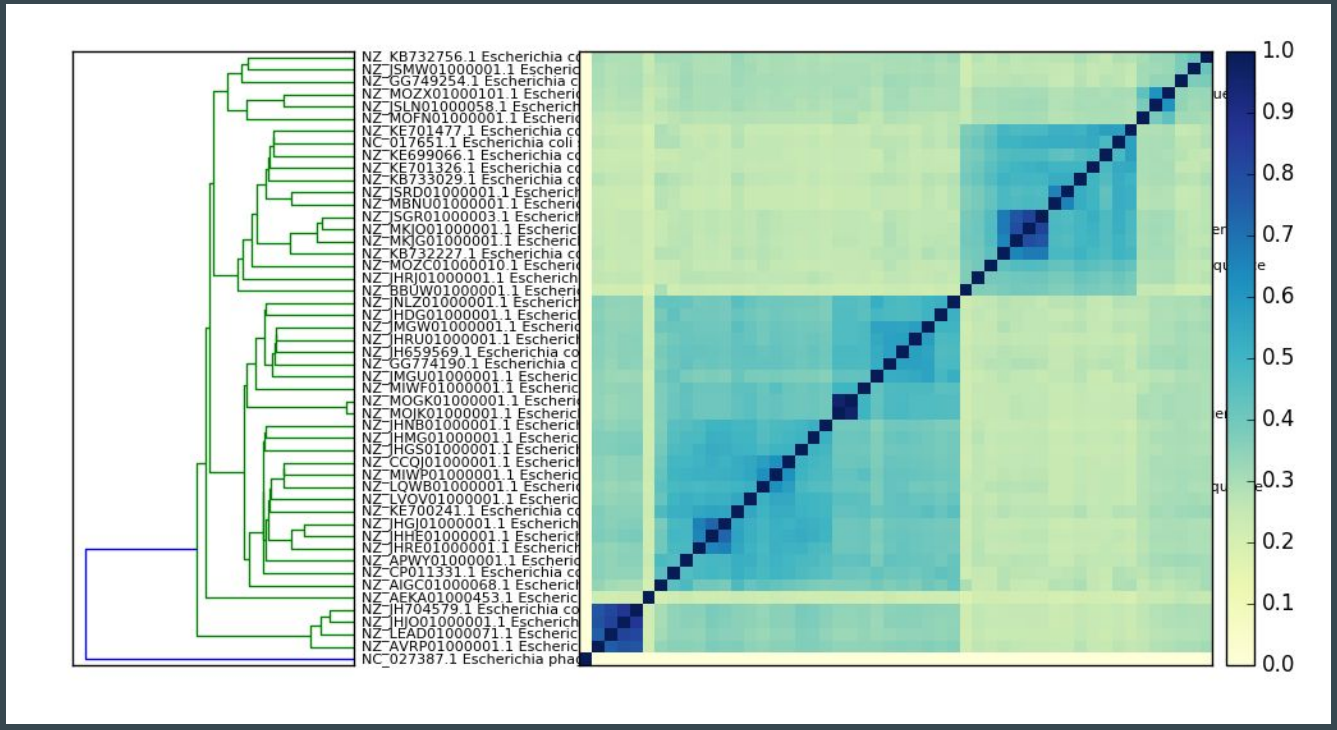
Genomic MinHash

Mash: fast genome and metagenome distance estimation using MinHash. Ondov et al., 2016
Large-scale sequence comparisons with sourmash. Pierce NT, Irber L, Reiter T et al., 2019

sourmash

https://sourmash.bio/

- Python library and CLI
  - Rust core
- Similarity and containment estimation between datasets
- Documentation
- Tutorials
- Tests



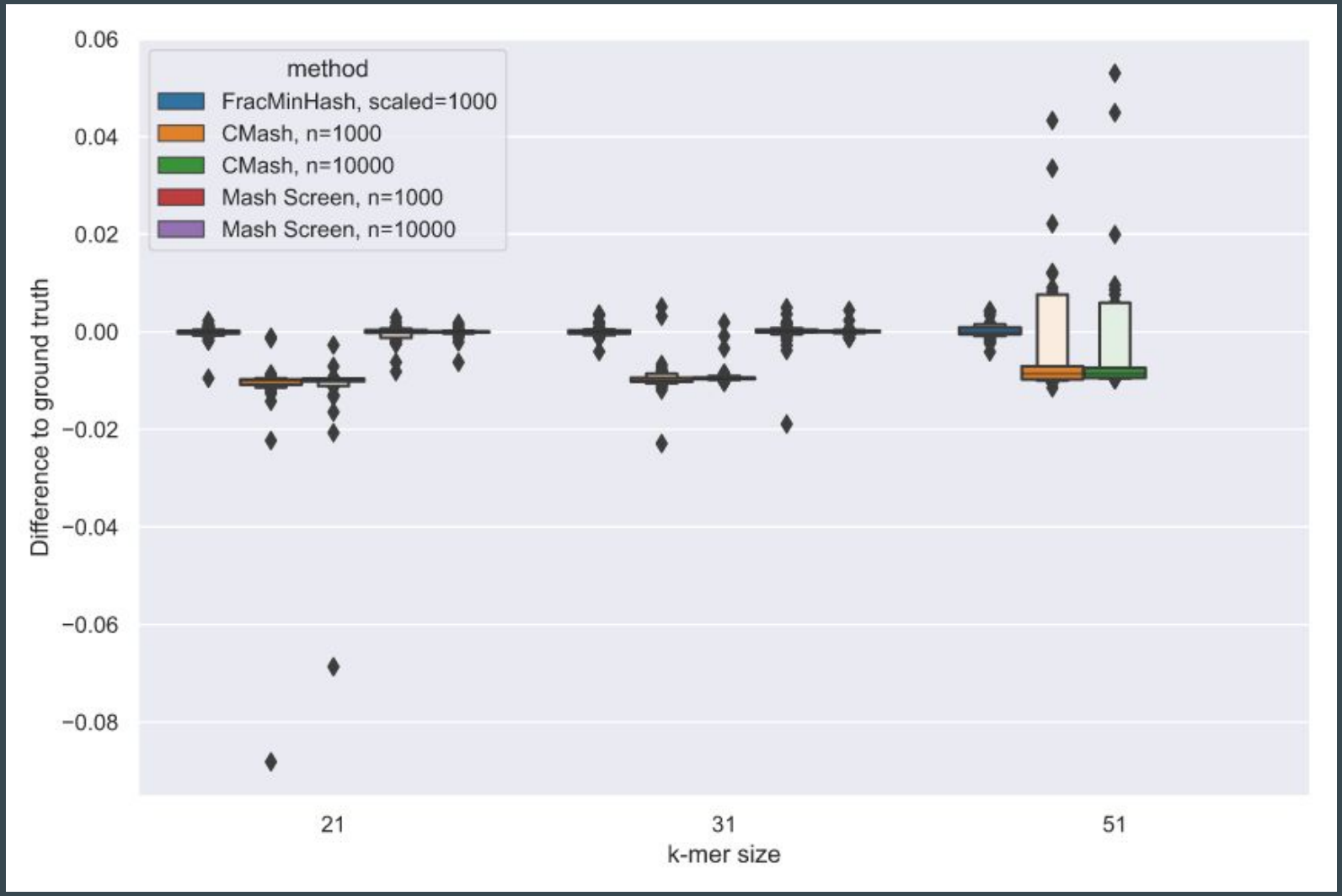
sourmash

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

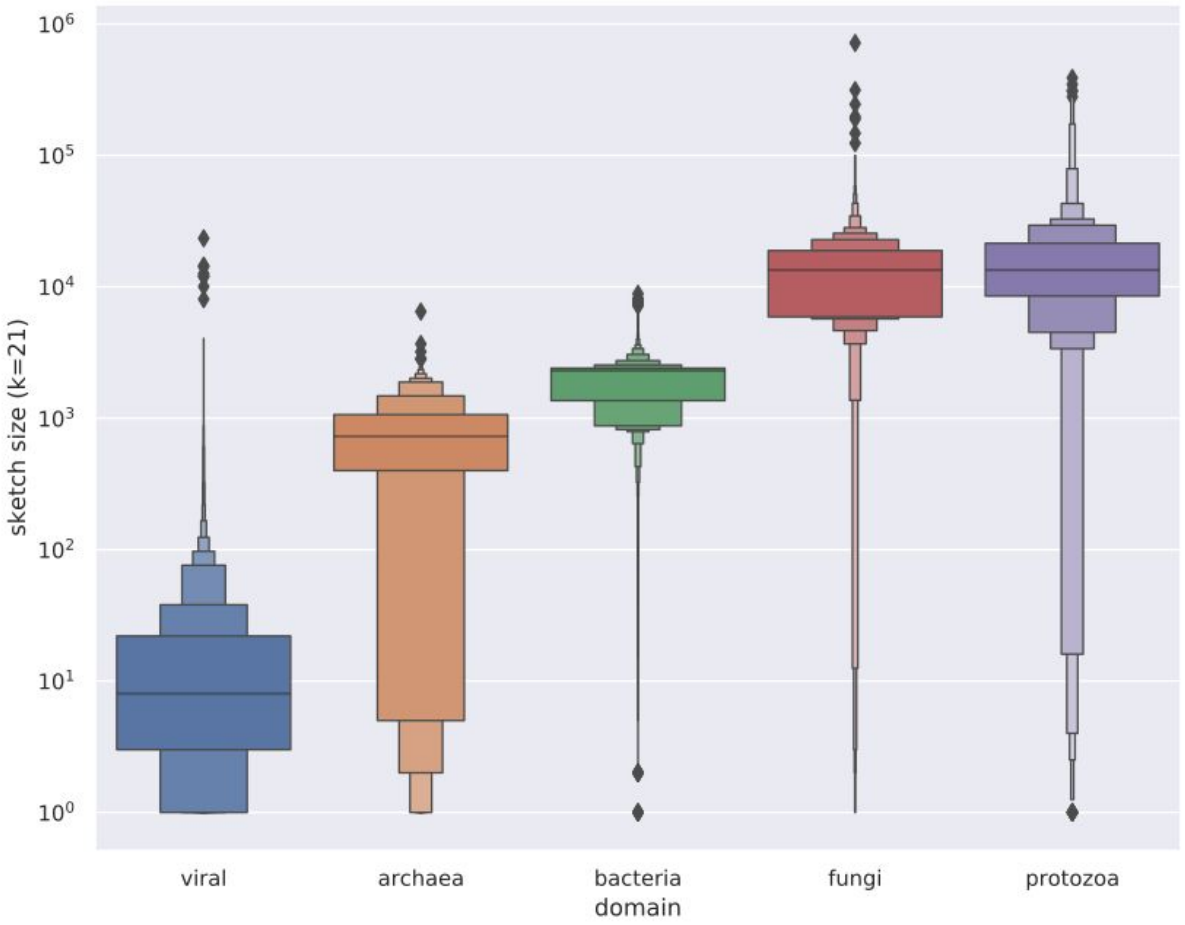
https://github.com/sourmash-bio/sourmash/graphs/contributors

Other methods implementing containment queries

- mash screen and Containment score
  - Need original dataset
- CMash and Containment MinHash
  - Use a Bloom Filter for reduced representation of original dataset
- sourmash: need only FracMinHash sketches of the original dataset



Lightweight compositional analysis of metagenomes with FracMinHash and minimum metagenome covers
doi:10.1101/2022.01.11.475838



FracMinHash sketch sizes for GenBank domains

scaled=1000

FracMinHash: efficient containment queries

- No need to access original data
- No auxiliary data structures required
- Allow other operations like subtraction and abundance filtering
- Tunable trade-off between precision and resource consumption
  - Lower scaled values = more precise with smaller queries (viruses, plasmids)
  - Higher scaled values = faster processing, less memory (metagenomes)

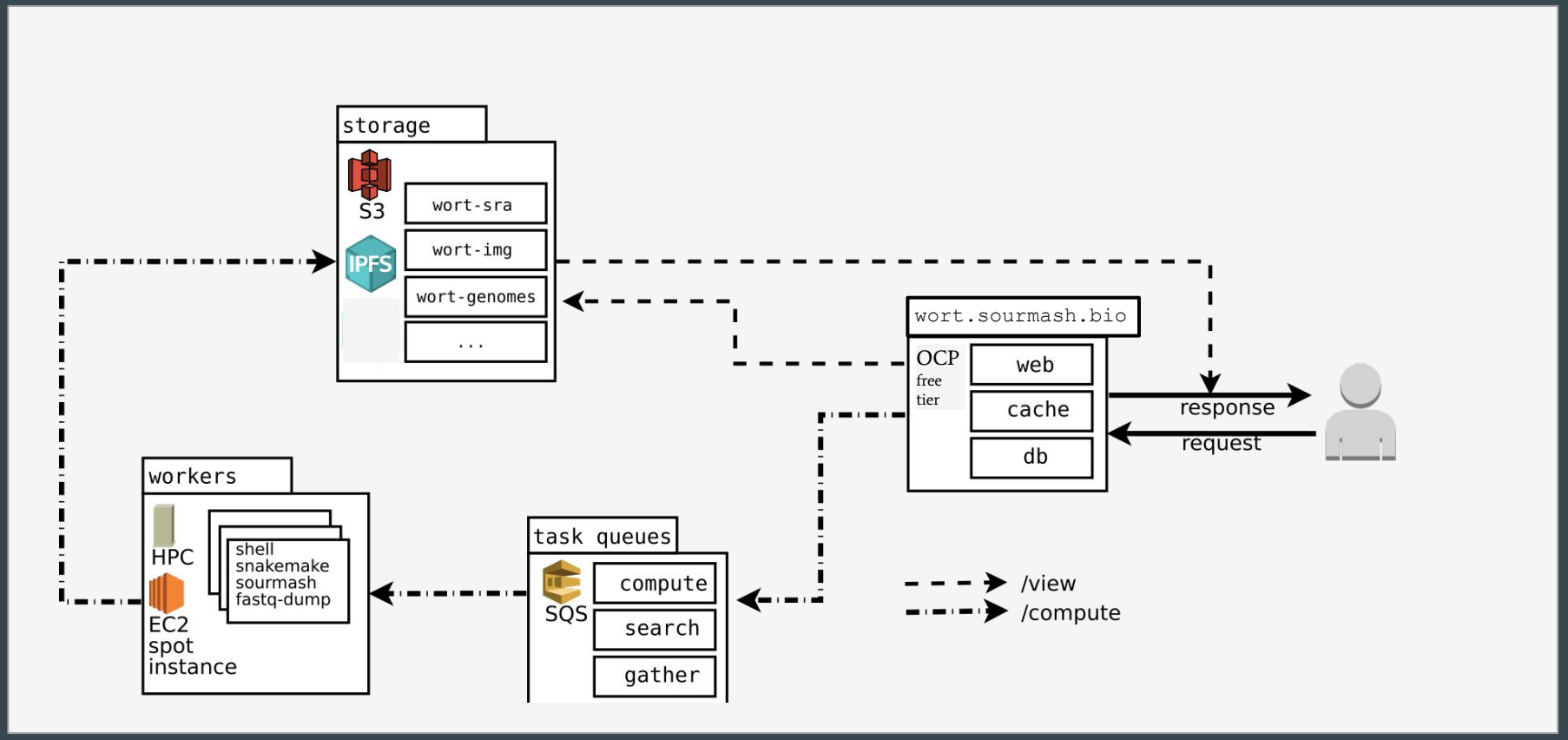
Distributed signature calculation with wort

You can have a second computer once you've shown you know how to use the first one.

Paul Barham

Workers of the world, unite!

Flora Tristan

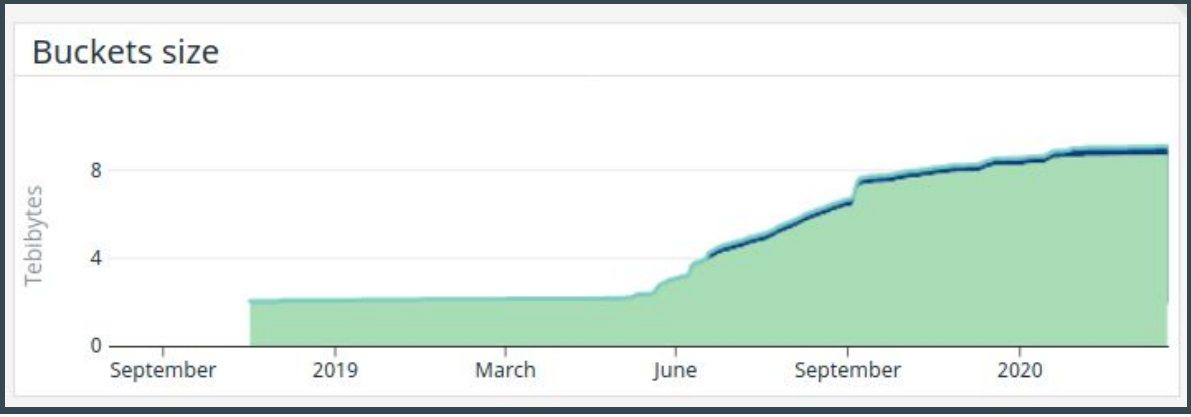


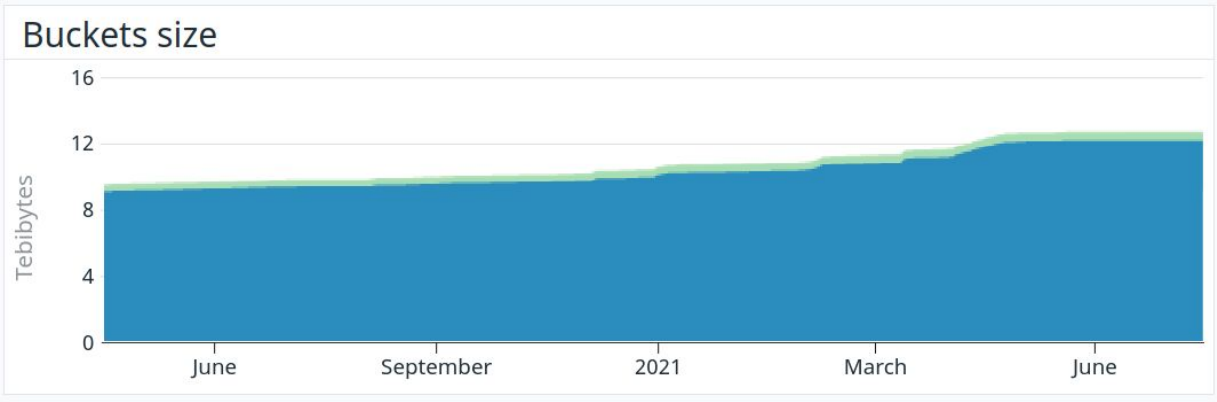
wort architecture

https://wort.sourmash.bio

Processed data

Total s3 buckets size	1d	Number of objects	1d
12.66 TiB		**4.68M**	





Single-genome

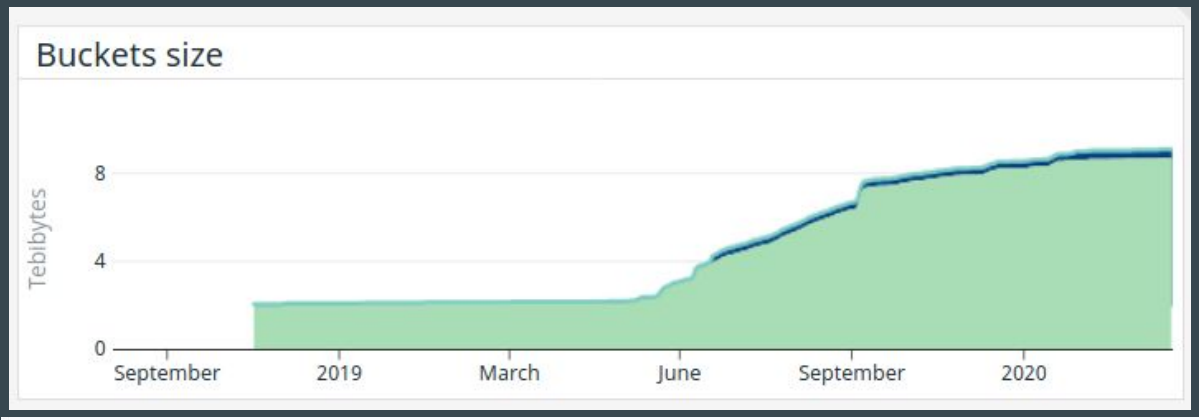
	all	missing	completed
count	1,932,292	32,669	1,899,623
mean	385	457	384
std	1,348	2,022	1,333
min	1	1	1
25%	86	103	86
50%	179	186	179
75%	304	306	304
max	209,497	209,497	135,737
sum	744,563,374	14,918,684	729,644,690

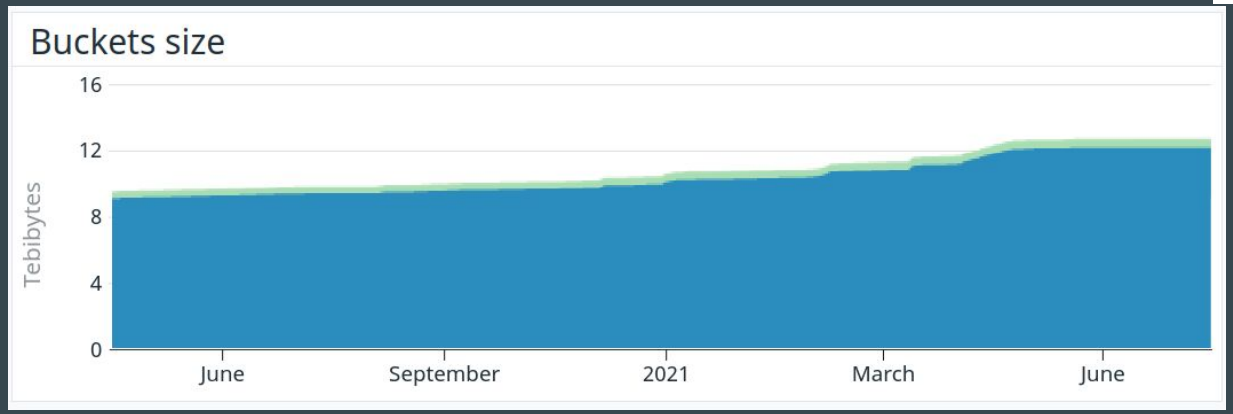
Metagenomes

	all	missing	completed
count	594,043	56,396	537,647
mean	932	646	962
std	2,856	1,420	2,965
min	1	1	1
25%	8	8	8
50%	72	151	66
75%	695	832	683
max	170,085	111,994	170,085
sum	553,407,224	36,440,618	516,966,606

Processed data

Total s3 buckets size	1d	Number of objects	1d
12.66 TiB		**4.68M**	





Single-genome

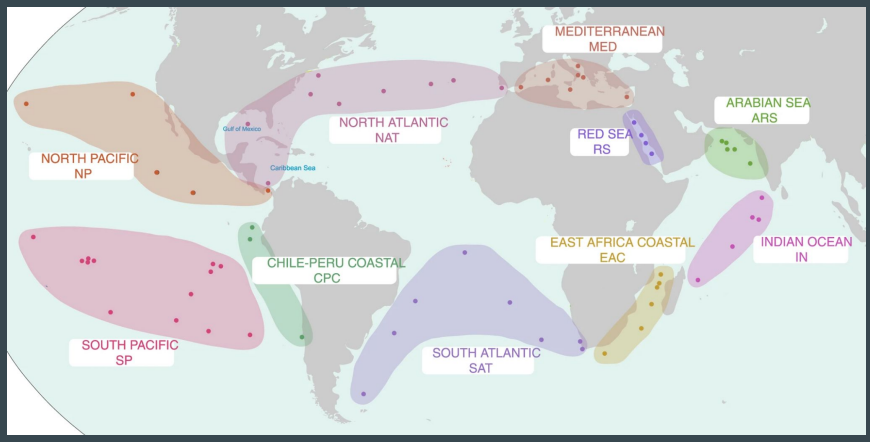
	all	missing	completed
count	1,932,292	32,669	1,899,623
mean	385	457	384
std	1,348	2,022	1,333
min	1	1	1
25%	86	103	86
50%	179	186	179
75%	304	306	304
max	209,497	209,497	135,737
sum	744,563,374	14,918,684	729,644,690

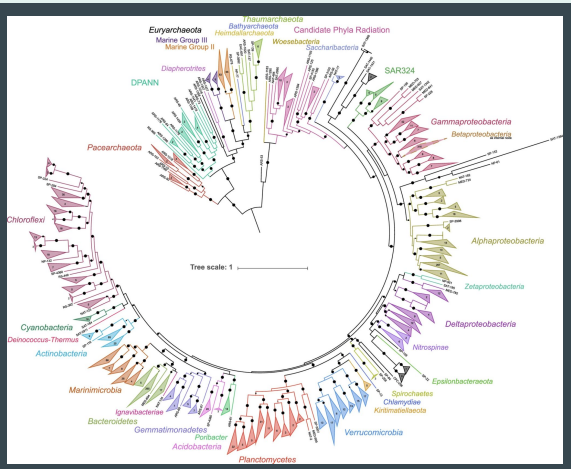
Metagenomes

	all	missing	completed
count	594,043	56,396	537,647
mean	932	646	962
std	2,856	1,420	2,965
min	1	1	1
25%	8	8	8
50%	72	151	66
75%	695	832	683
max	170,085	111,994	170,085
sum	553,407,224	36,440,618	516,966,606

Searching large public genomic databases

- 500k metagenomes
- 2,631 MAG queries from Tully et al (2017)
- 23,644 matches above 50% containment
- 6,398 unique SRA runs
- 11 hours, 32 cores, 12GB of memory
  - Load metagenomes in parallel, compare with queries
- High latency
  - time for 1 query ~ time for 2,631 queries





https://blog.luizirber.org/2020/07/22/mag-search/
https://blog.luizirber.org/2020/07/24/mag-results/

The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans
doi:10.1038/sdata.2017.203

mastiff: content-based SRA search

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

https://mastiff.sourmash.bio

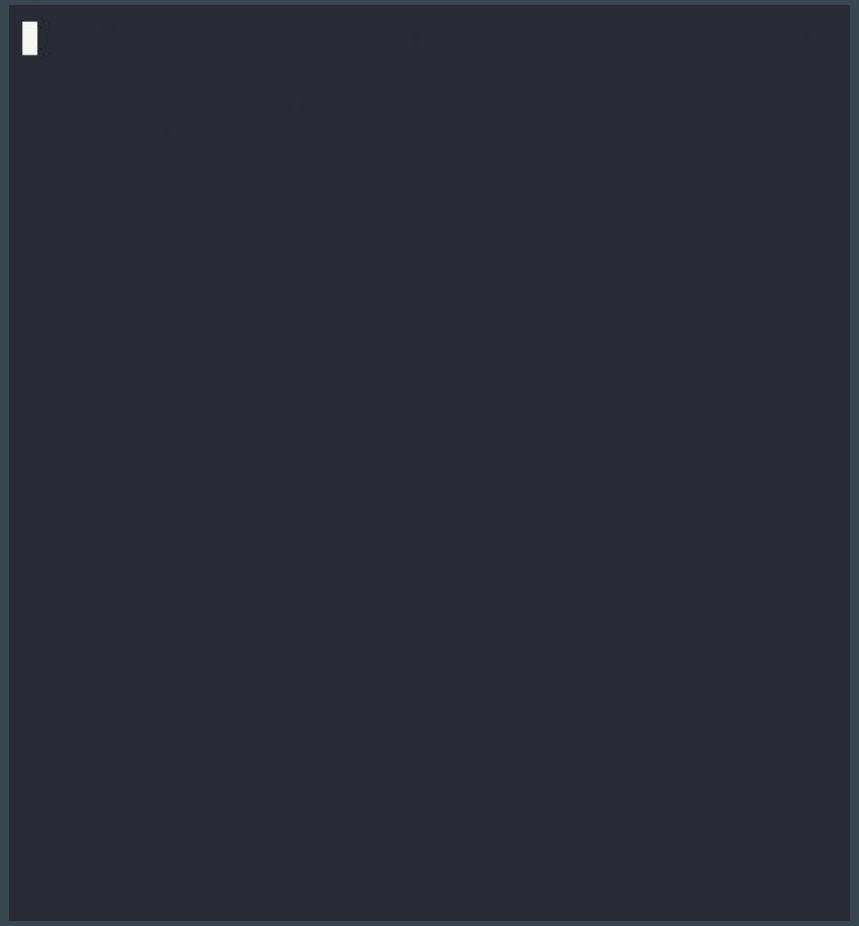
- Disk-based sourmash index
  - Inverted index
  - Based on rocksdb
- 486k+ SRA runs indexed

On 2,631 MAG queries from Tully et al

- 55 minutes, 24 cores, 540MB mem
- (Previous: 11 hours, 32 cores, 12GB of mem)
- 10s/query

Client available

- From raw data or signature



```
$ mastiff --sig -o matches.csv \
  <(curl -sL https://wort.sourmash.bio/v1/view/genomes/GCF_000195915.1)
```

Verifying results with alignment

Download original data, minimap2

- MAG: TOBG_NP-110 (North Pacific)
- Archaeal MAG that failed to be classified further than Phylum level (Euryarchaeota)
- 12 matches above 50% containment
- 5 from one study, SRP044185, with samples collected from a column of water in a station in Manzanillo, Mexico.
- 3 matches come from SRP003331, in the South Pacific ocean (in northern Chile).
- ERR3256923, also comes from the South Pacific.

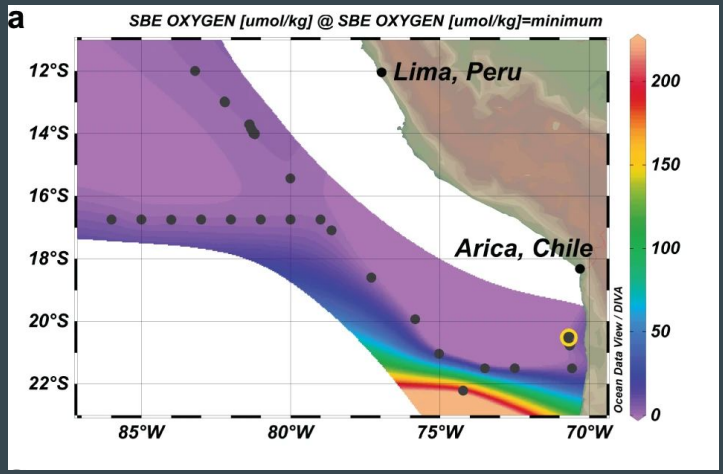
SRA Run ID	Containment search	Containment reads	Missed bp	%bp missed	Coverage	Reads mapped
SRR5868539	0.99	0.99	1,475	0.119	19.85	131,640
SRR1509798	0.98	0.98	7,771	0.628	18.53	101,640
SRR1509792	0.97	0.97	26,145	2.111	9.79	63,662
SRR5868540	0.91	0.91	52,686	4.255	4.27	24,983
SRR1509799	0.89	0.89	71,072	5.740	3.88	20,036
SRR070081	0.85	0.85	109,714	8.860	3.00	9,755
ERR3256923	0.81	0.81	116,961	9.446	4.09	32,295
SRR070083	0.79	0.79	202,038	16.316	2.04	6,115
SRR1509793	0.79	0.79	142,974	11.546	3.47	22,493
SRR304680	0.64	0.64	359,675	29.047	1.35	4,181
SRR070084	0.58	0.58	450,118	36.351	1.16	3,403
SRR1509794	0.56	0.56	427,130	34.495	1.60	9,828

Verifying results with alignment

Download original data, minimap2

- MAG: TOBG_NP-110 (North Pacific)
- Archaeal MAG that failed to be classified further than Phylum level (Euryarchaeota)
- 12 matches above 50% containment
- 5 from one study, SRP044185, with samples collected from a column of water in a station in Manzanillo, Mexico.
- 3 matches come from SRP003331, in the South Pacific ocean (in northern Chile).
- ERR3256923, also comes from the South Pacific.

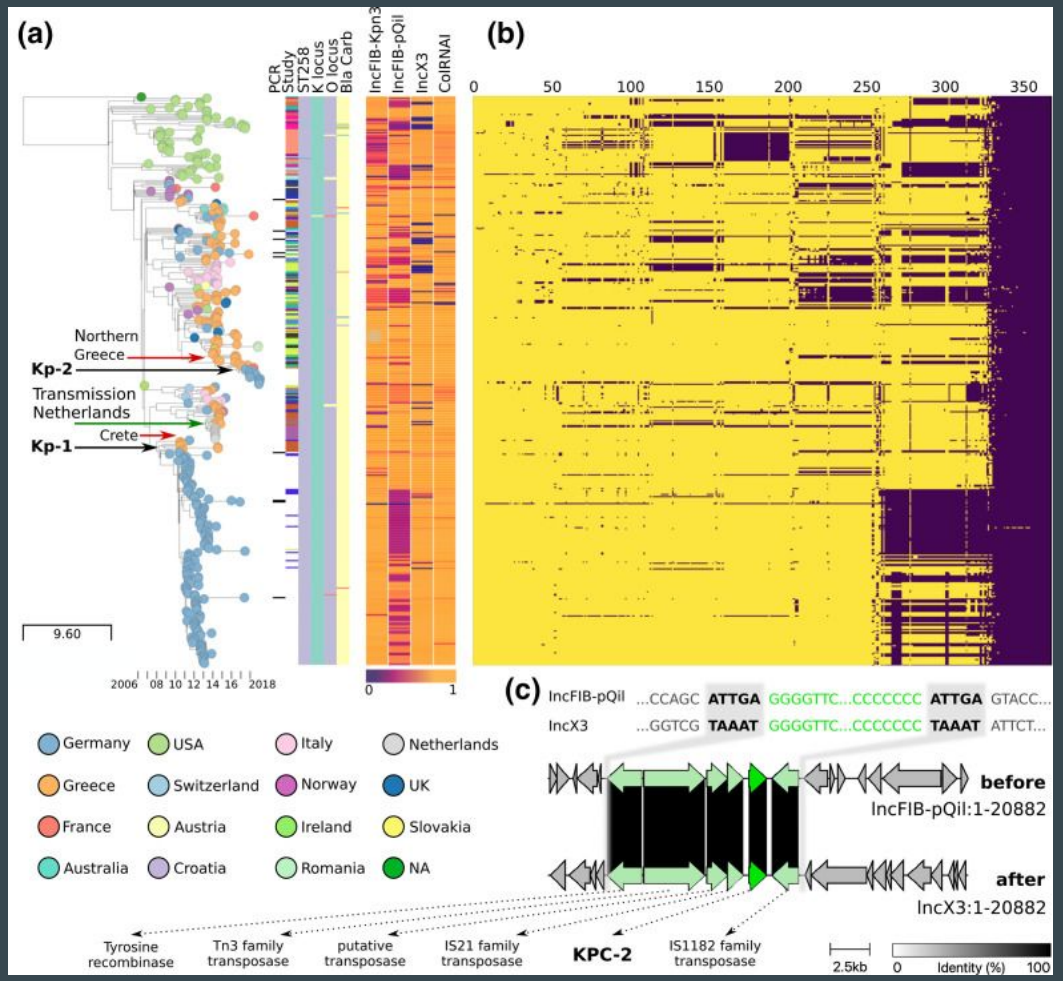
nitrite reductase, and N_2O reductase (Fig. 2). Two MAGs from the Tara Oceans metagenomes (Table S1) were identified as the same species as MG-II MAG-2. TOBG_NP-110 (ANI to MG-II MAG-2 = 99.8%) from the North Pacific encoded Nar and nitrate/nitrite transporters, and TOBG_SP-208 (ANI to MG-II MAG-2 = 99.6%) from the South Pacific also contained the same denitrification genes as MG-II MAG-2 (Table S2). In addition, two MG-II SAGs (AD-615-F09 and



Novel metagenome-assembled genomes involved in the nitrogen cycle from a Pacific oxygen minimum zone

Sun, X., Ward, B.B (2021). doi:10.1038/s43705-021-00030-2

“Furthermore, we searched the index Kp-1 isolate in a k-mer database of over 400000 metagenomic read datasets. We identified a single sample from an unpublished study of ICU patient colonization (NCBI, project ID PRJNA561398) where we could recover a closely related, metagenome-assembled Kp genome.”

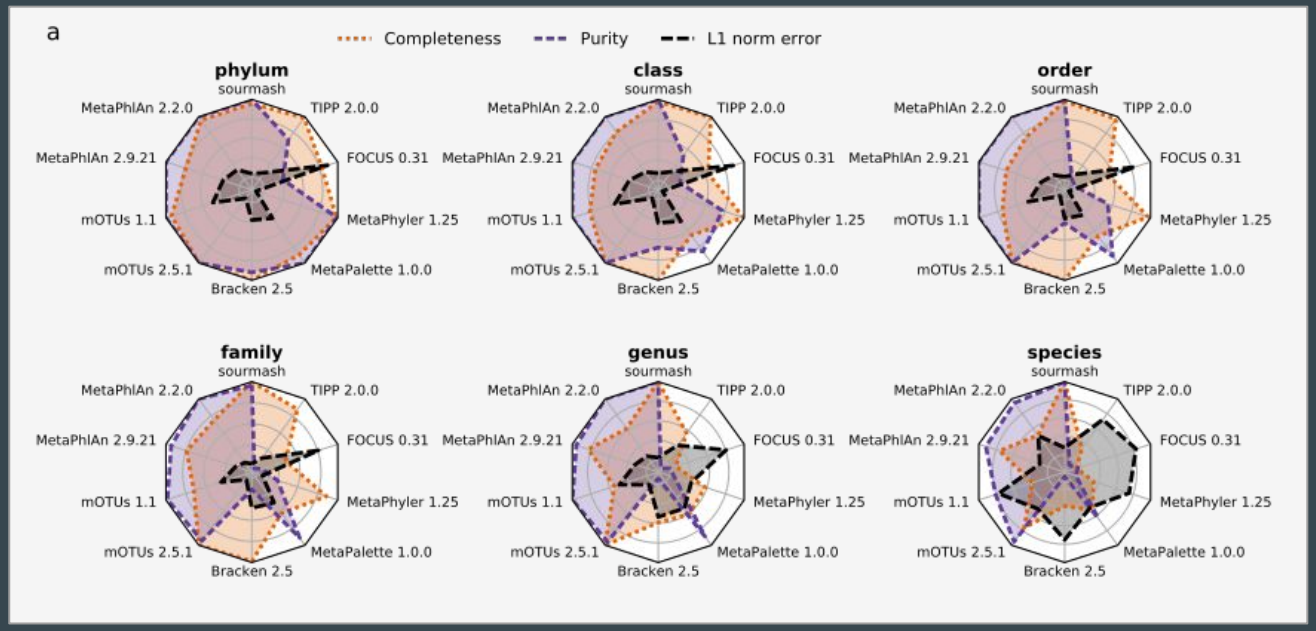


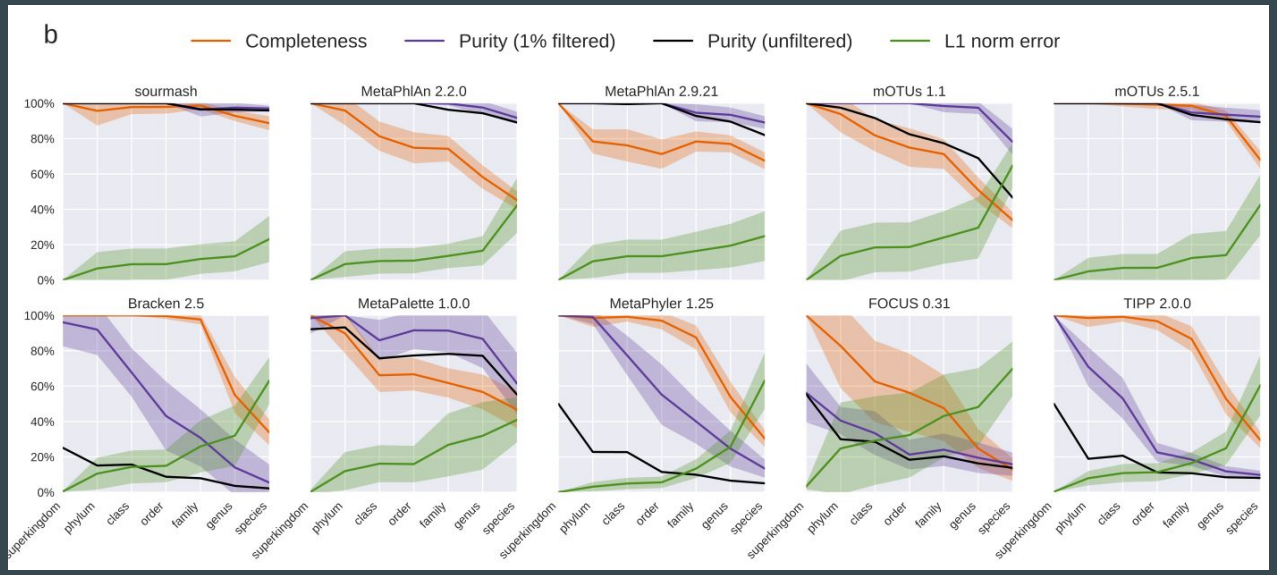
Context-aware genomic surveillance reveals hidden transmission of a carbapenemase-producing Klebsiella pneumoniae

10.1099/mgen.0.000741

gather for taxonomic profiling of metagenomes

- What are the species present in a sample?
- And what are their abundances?
- Current methods
  - K-mer composition (Kraken)
  - Alignment (Diamond-MEGAN)
  - Markers (mOTUs, MetaPhlAn)
- Usually assign taxonomy to reads/contigs, count assignments for each rank
  - Bottom-up
- Gather + FracMinHash: no reads/contigs available anymore
  - Iterative top-down
  - Greedy solution to Min-set cover





Lightweight compositional analysis of metagenomes with FracMinHash and minimum metagenome covers

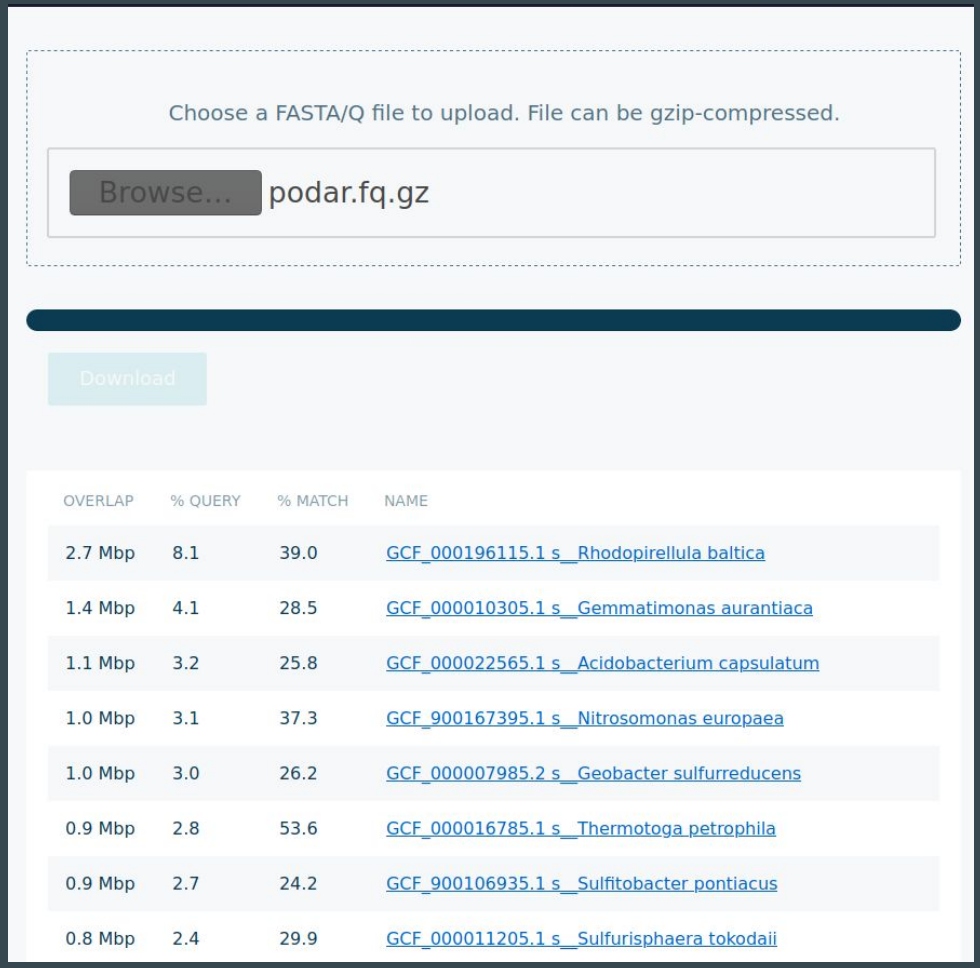
doi:10.1101/2022.01.11.475838

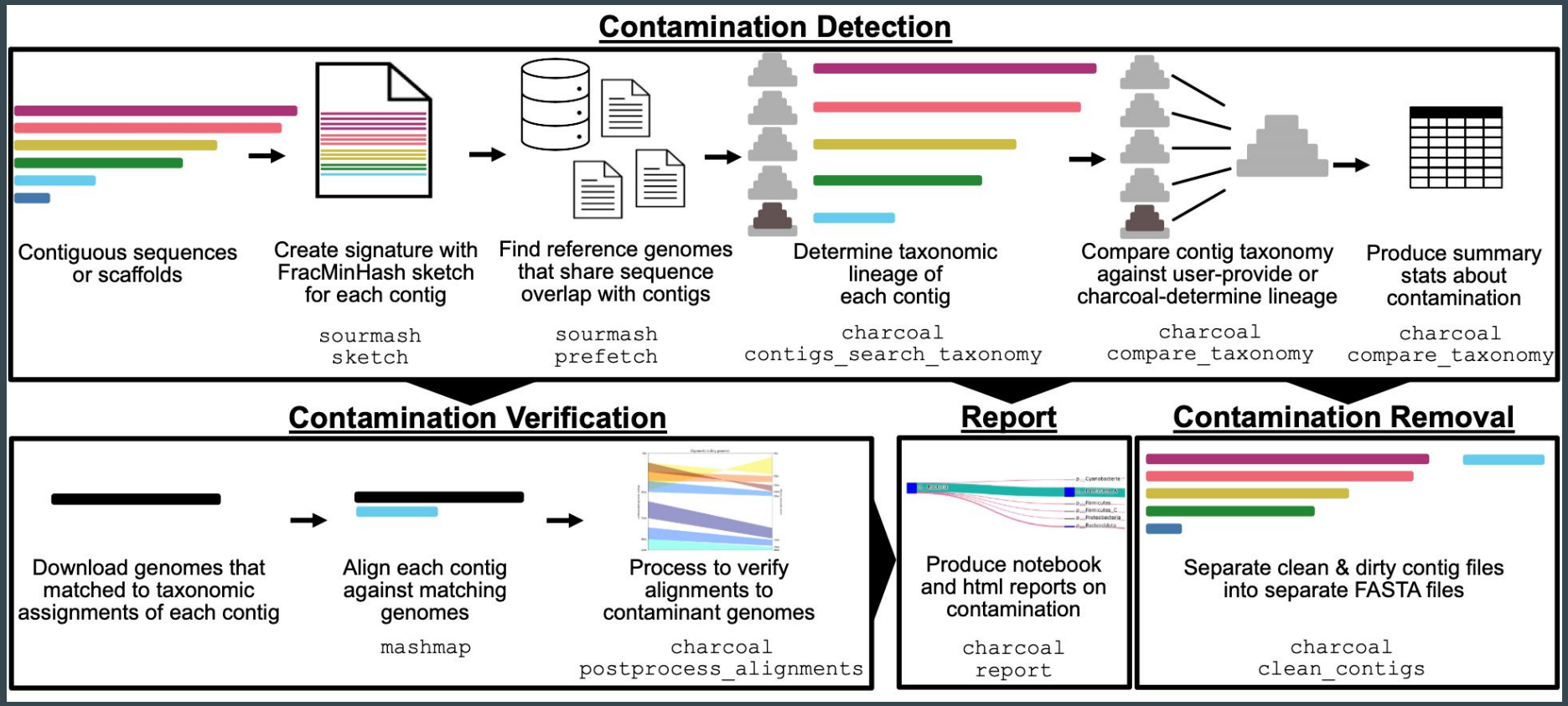
Greyhound: sketching in the browser

https://github.com/luizirber/2020-11-02-greyhound
https://github.com/dib-lab/sourmash/pull/1238

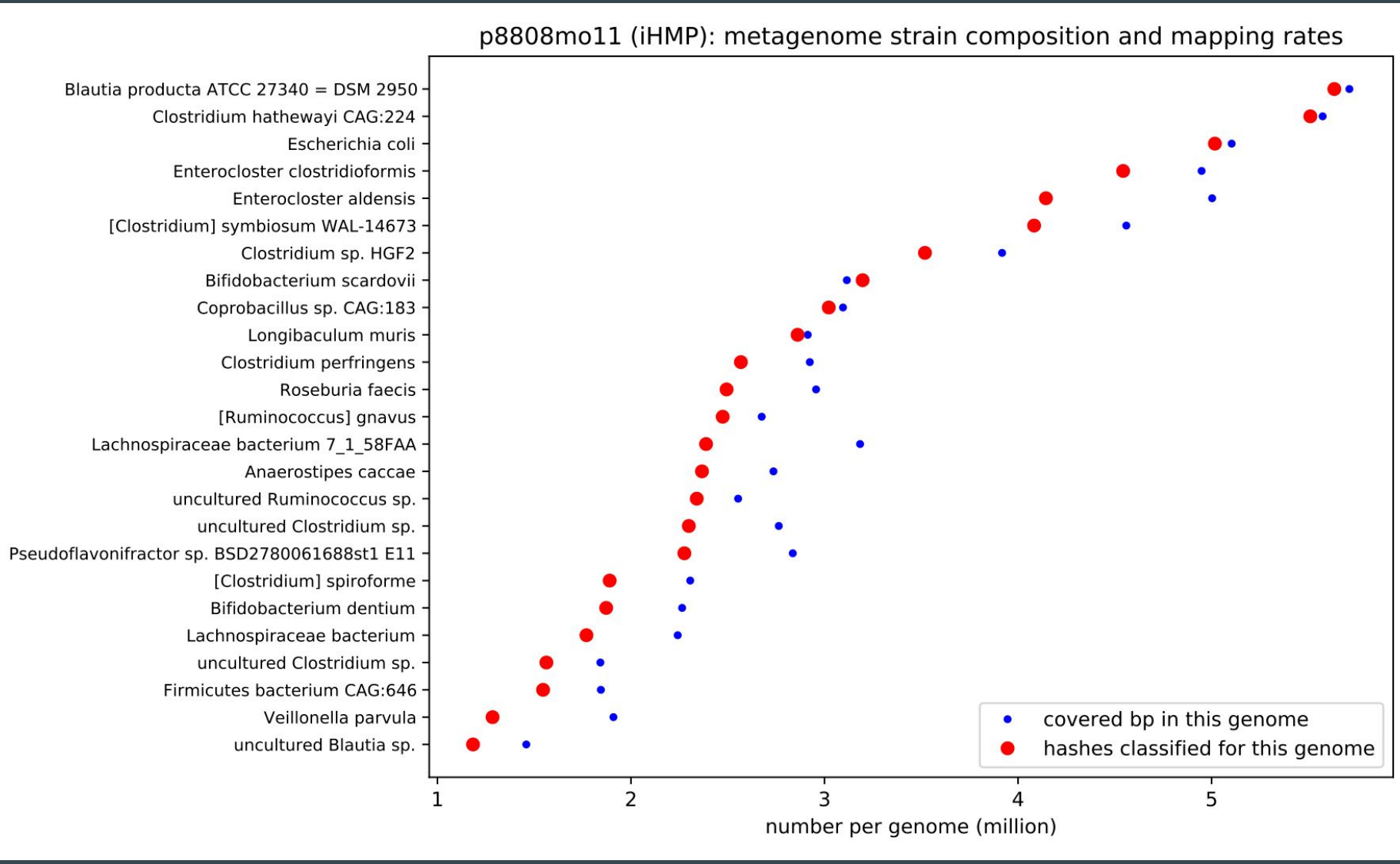
https://greyhound.sourmash.bio

- A full stack Rust project
- MinHash sketch calculated in the browser
  - Webassembly
- Frontend: Yew
  - Elm architecture, compiles to Wasm
- Web backend: Tide
- Bfx method: sourmash
  - RevIndex: A new in-memory index
  - Mapping hash -> color
  - Fast, but memory-intensive
- Demo:
  - 65,703 species clusters in GTDB rs207
  - ~7GB of RAM

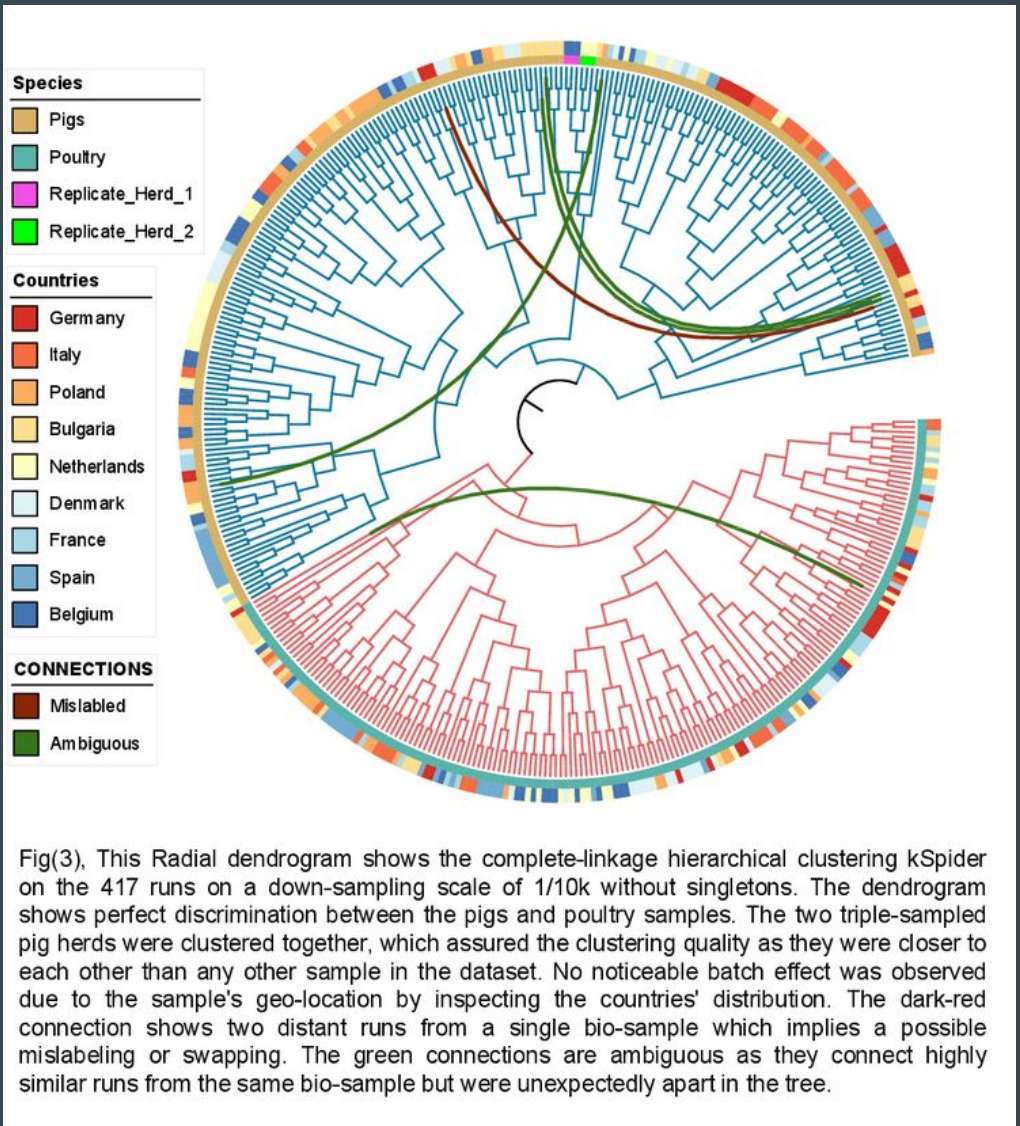



Contamination Detection
Contiguous sequences or scaffolds
Create signature with FracMinHash sketch for each contig
sourmash sketch
Find reference genomes that share sequence overlap with contigs
sourmash prefetch
Determine taxonomic lineage of each contig
charcoal contigs_search_taxonomy
Compare contig taxonomy against user-provide or charcoal-determine lineage
charcoal compare_taxonomy
Produce summary stats about contamination
charcoal compare_taxonomy
Contamination Verification
Download genomes that matched to taxonomic assignments of each contig
Align each contig against matching genomes
mashmap
Process to verify alignments to contaminant genomes
charcoal postprocess_alignments
Report
Produce notebook and html reports on contamination
charcoal report
Contamination Removal
Separate clean & dirty contig files into separate FASTA files
charcoal clean_contigs

Charcoal



genome-grist



Fig(3), This Radial dendrogram shows the complete-linkage hierarchical clustering kSpider on the 417 runs on a down-sampling scale of 1/10k without singletons. The dendrogram shows perfect discrimination between the pigs and poultry samples. The two triple-sampled pig herds were clustered together, which assured the clustering quality as they were closer to each other than any other sample in the dataset. No noticeable batch effect was observed due to the sample's geo-location by inspecting the countries' distribution. The dark-red connection shows two distant runs from a single bio-sample which implies a possible mislabeling or swapping. The green connections are ambiguous as they connect highly similar runs from the same bio-sample but were unexpectedly apart in the tree.

Fast Clustering of Hundreds of Microbiome Datasets

https://pag.confex.com/pag/xxix/meetingapp.cgi/Paper/45450
https://github.com/dib-lab/kSpider

Thanks!

luizirber.org

twitter.com/luizirber

github.com/luizirber

luiz@sourmash.bio

https://sourmash.bio

https://mastiff.sourmash.bio

https://greyhound.sourmash.bio

Lab for Data Intensive Biology

UCDAVIS VETERINARY MEDICINE