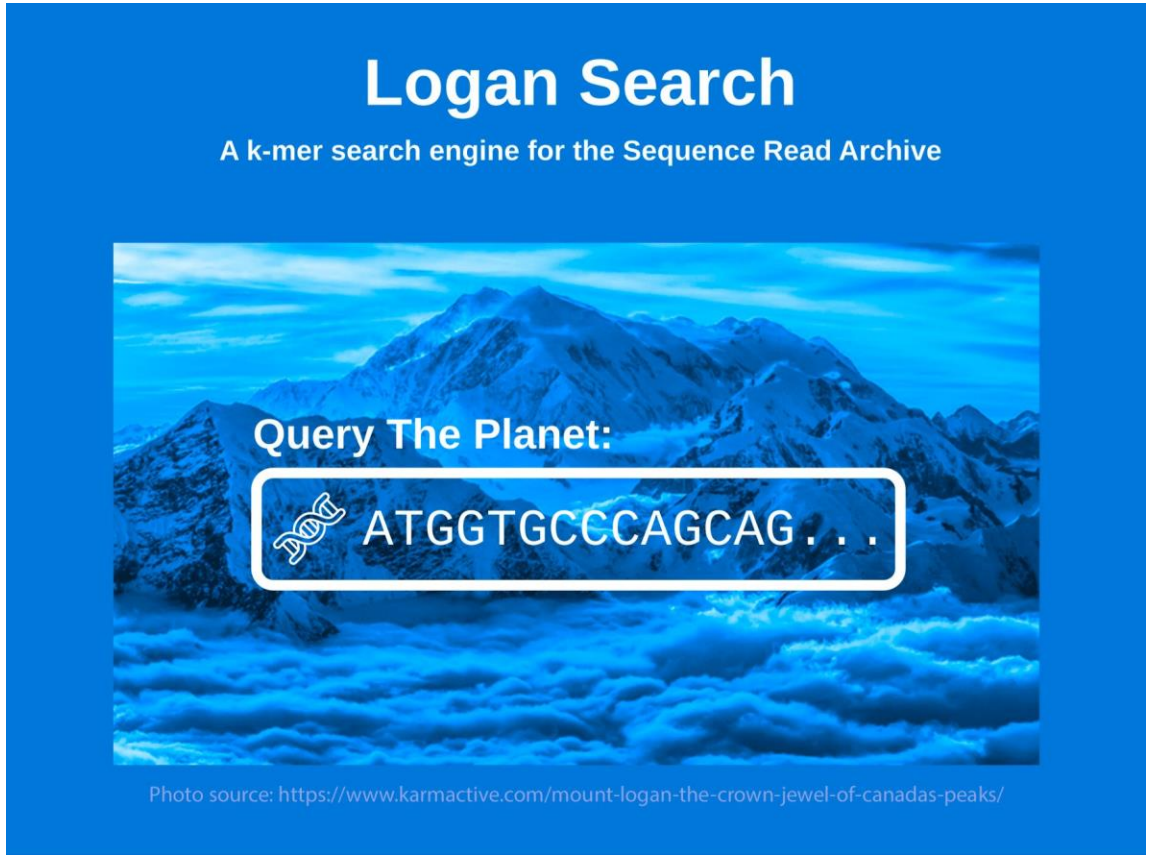
15:15	15:30	Index the planet: index SRA unitigs with kmindex. Results, joys and sorrows *Pierre Peterlongo, Inria, France*



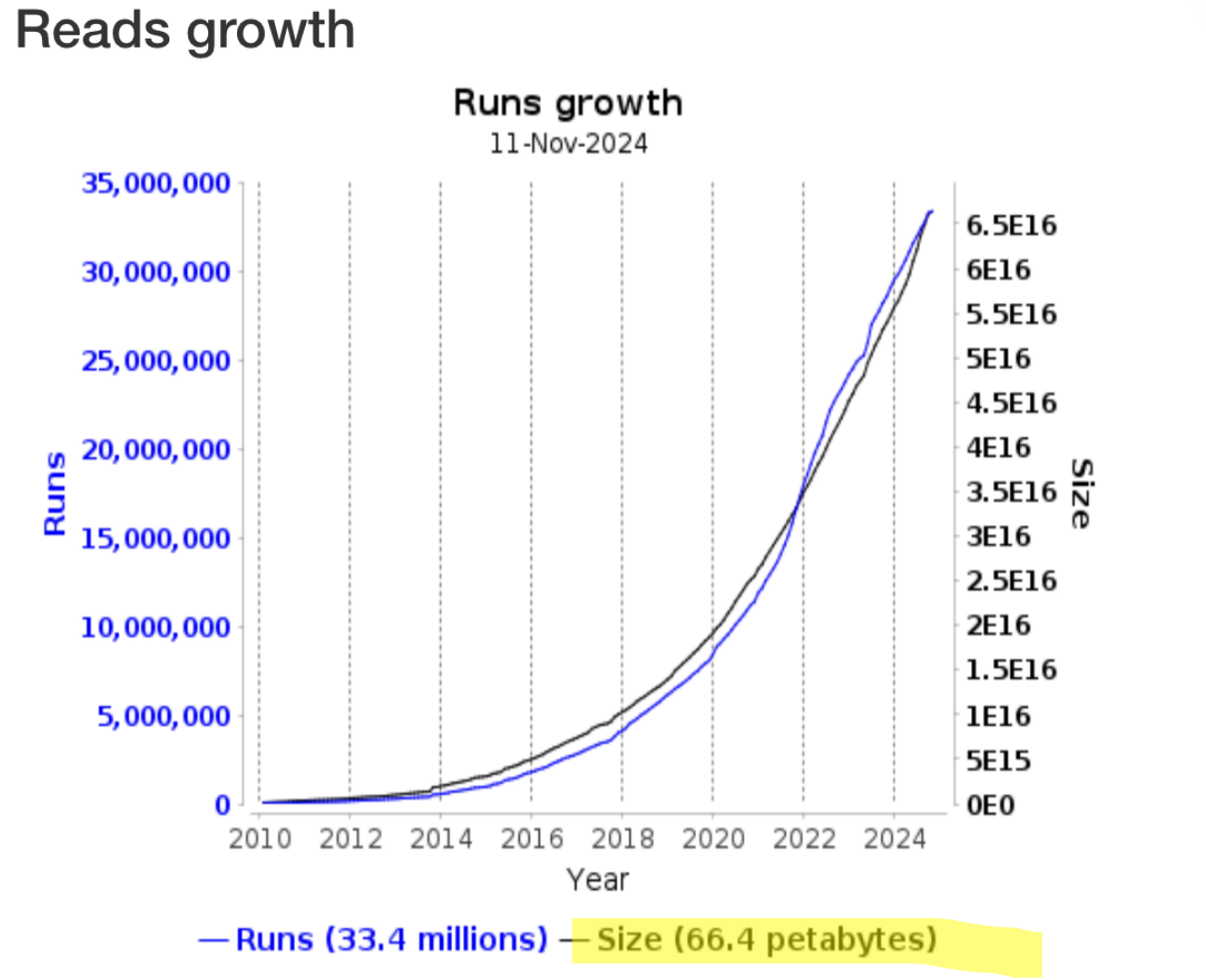
15:15 15:30 Index the planet: index SRA unitigs with kmindex. Results, joys and sorrows
Pierre Peterlongo, Inria, France

Logan Search: index and query SRA using kmindex

Pierre Peterlongo & Téo Lemane
Collabs: Rayan Chikhi, Artem Babaian, Luke Pereira,

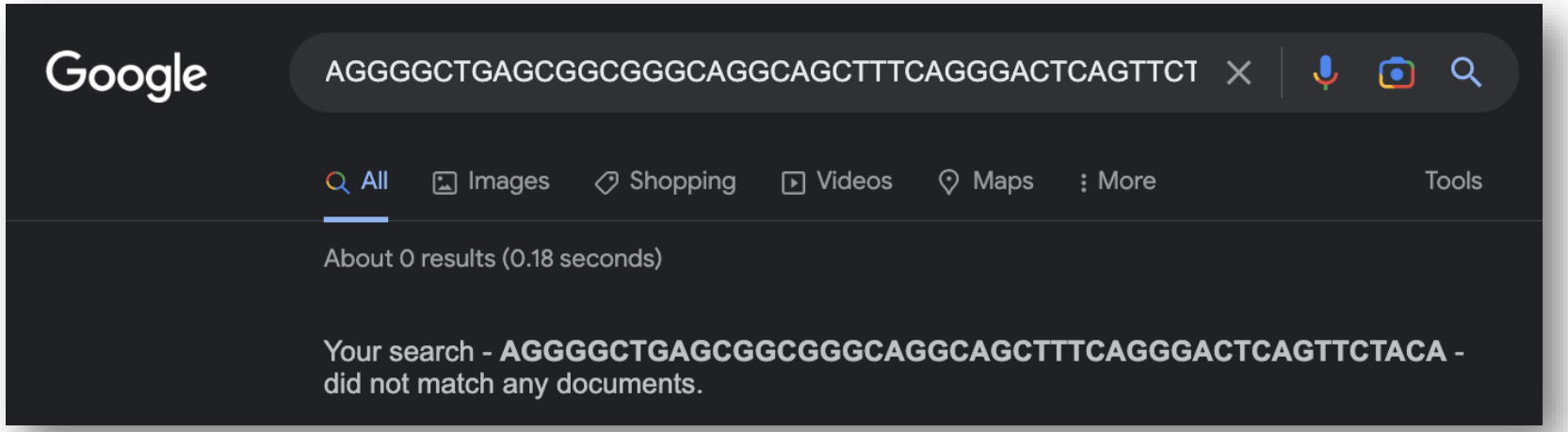
Genome Informatics – November 2024

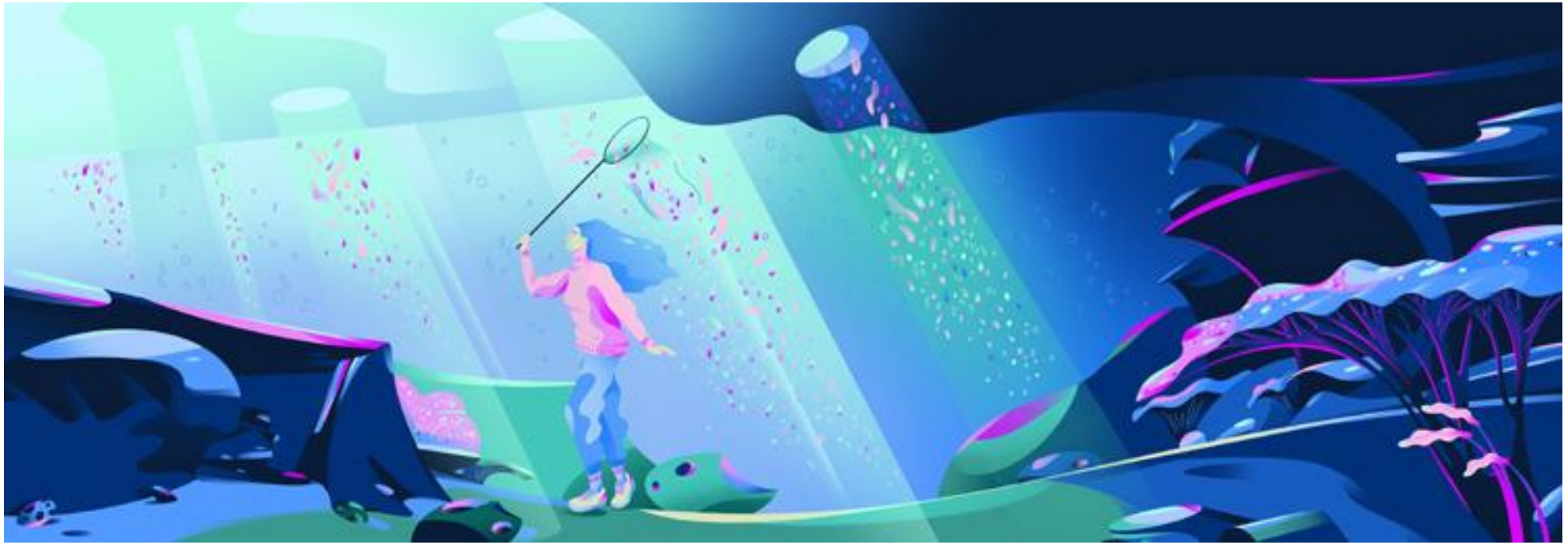
We sit on a treasure



https://www.ebi.ac.uk/ena/browser/about/statistics

We sit on a treasure, but…





@Manon Sauzara

kmindex

Téo Lemane

Lemane, T., Lezzoche, N., Lecubin, J., Pelletier, E., Lescot, M., Chikhi, R., & Peterlongo, P. (2024). “Indexing and real-time user-friendly queries in terabyte-sized complex genomic datasets with kmindex and ORA.”
Nature Computational Science, 4(2), 104-109.

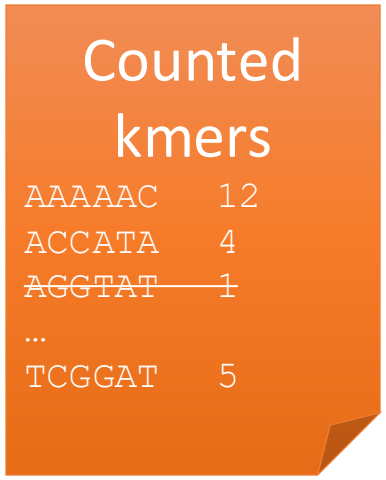
Indexing: conceptual view

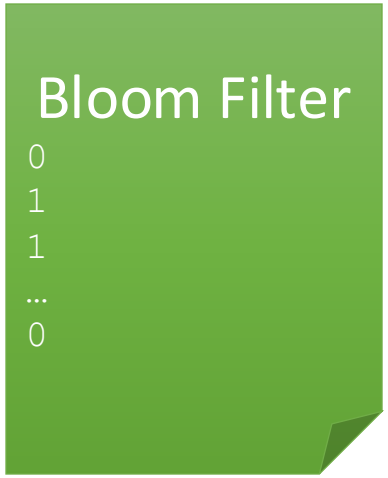
One read set:

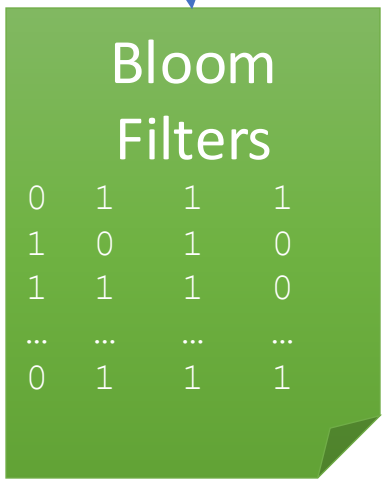
- Extract & count **kmers**
- Filter kmers
- Generate a bloom filter

***N* read sets**:

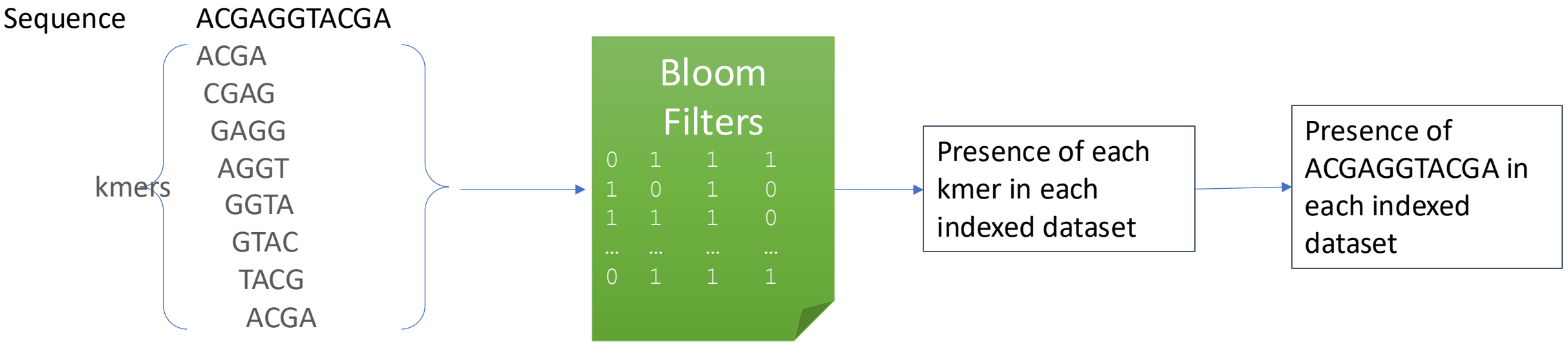
- Create *N* bloom filters
- This is the index







Querying: conceptual view



Logan-Search

Logan

~50 petabases
of raw reads (SRA)

385 terabytes contigs
s3://logan-pub/c/

2.18 petabytes unitigs
s3://logan-pub/u/

27 million samples
2 million billions k-mers

Chikhi, R., Raffestin, B., Korobeynikov, A., Edgar, R. C., & Babaian, A. (2024).
Logan: Planetary-Scale Genome Assembly Surveys Life's Diversity.
bioRxiv, 2024-07.

Logan + kmindex = Logan Search

~50 petabases
of raw reads (SRA)

385 terabytes contigs
s3://logan-pub/c/

2.18 petabytes unitigs
s3://logan-pub/u/

27 million samples
2 million billions k-mers

kmindex

2 centuries cpu time

Chikhi, R., Raffestin, B., Korobeynikov, A., Edgar, R. C., & Babaian, A. (2024).
Logan: Planetary-Scale Genome Assembly Surveys Life's Diversity.
bioRxiv, 2024-07.

Logan + kmindex = Logan Search

~50 petabases
of raw reads (SRA)

385 terabytes contigs
s3://logan-pub/c/

2.18 petabytes unitigs
s3://logan-pub/u/

27 million samples
2 million billions k-mers

kmindex

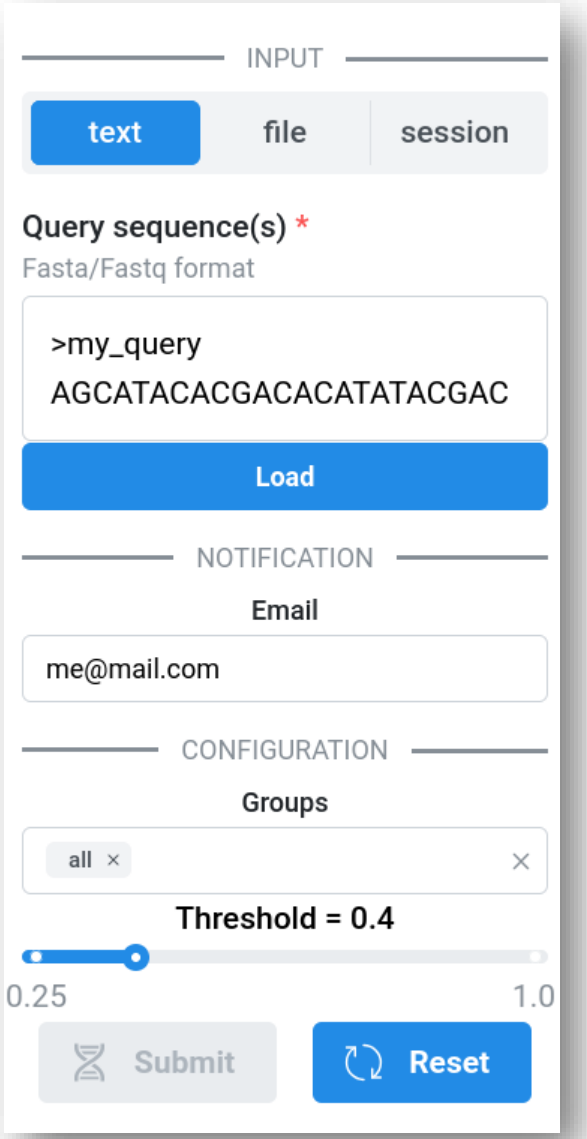
2 centuries cpu time

109 sub-indexes:

Genomic Transcript. MetaG MetaT SingleCell …	×	Human Mice Viruses Mamals Bact. …

Chikhi, R., Raffestin, B., Korobeynikov, A., Edgar, R. C., & Babaian, A. (2024). Logan: Planetary-Scale Genome Assembly Surveys Life's Diversity. bioRxiv, 2024-07.

Logan-search: perform a query

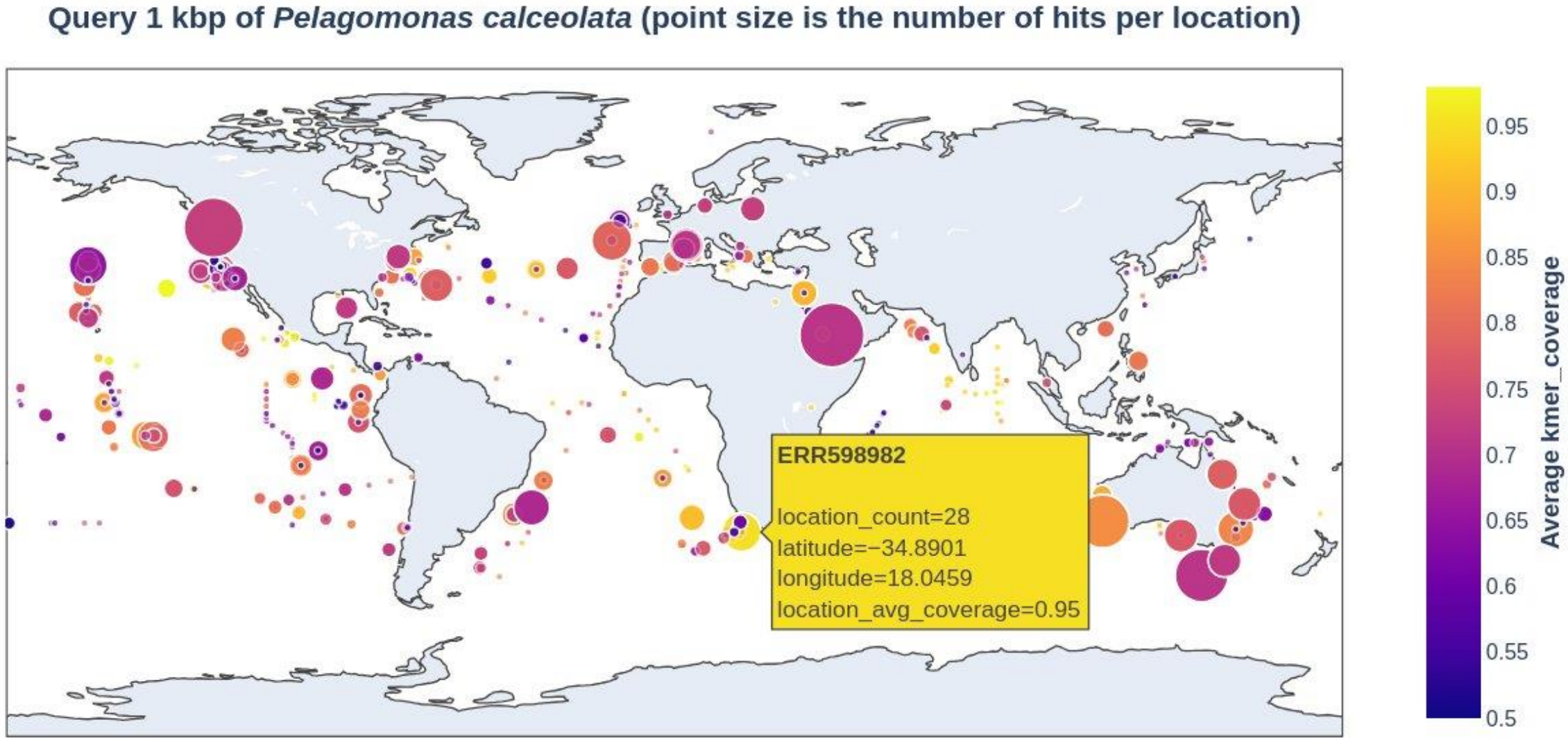


Logan-search answer: accession list

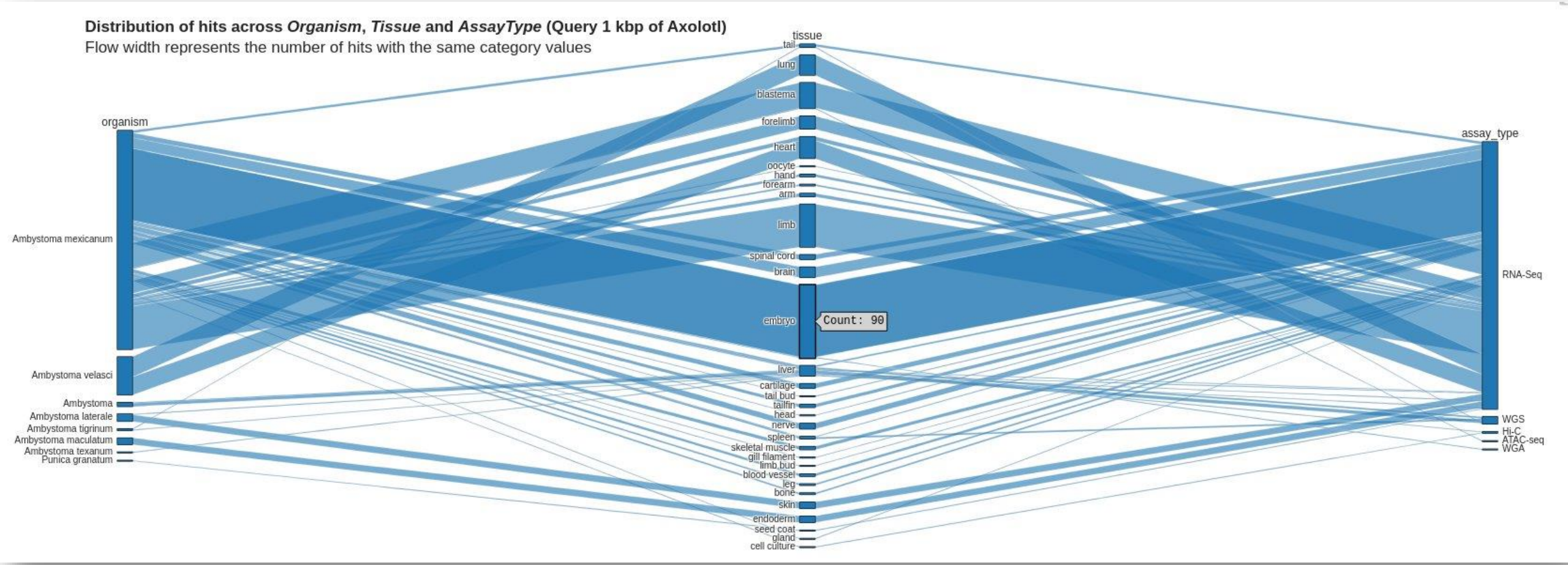
ID ↓	Similarity	Bioproject	Biosample
SRR9860238 (SRA|OV)	0.723	PRJNA556735 (SRA|OV)	SAMN12384881 (SRA|OV)
SRR9860233 (SRA|OV)	0.689	PRJNA556735 (SRA|OV)	SAMN12384871 (SRA|OV)
SRR9860232 (SRA|OV)	0.727	PRJNA556735 (SRA|OV)	SAMN12384871 (SRA|OV)
SRR9860231 (SRA|OV)	0.727	PRJNA556735 (SRA|OV)	SAMN12384871 (SRA|OV)
SRR9860230 (SRA|OV)	0.727	PRJNA556735 (SRA|OV)	SAMN12384871 (SRA|OV)
SRR9860229 (SRA|OV)	0.525	PRJNA556735 (SRA|OV)	SAMN12384871 (SRA|OV)
SRR9860228 (SRA|OV)	0.655	PRJNA556735 (SRA|OV)	SAMN12384871 (SRA|OV)
SRR9860227 (SRA|OV)	0.727	PRJNA556735 (SRA|OV)	SAMN12384882 (SRA|OV)
SRR9860226 (SRA|OV)	0.727	PRJNA556735 (SRA|OV)	SAMN12384882 (SRA|OV)
SRR9860223 (SRA|OV)	0.761	PRJNA556735 (SRA|OV)	SAMN12384872 (SRA|OV)
SRR9860222 (SRA|OV)	0.681	PRJNA556735 (SRA|OV)	SAMN12384872 (SRA|OV)
SRR9860221 (SRA|OV)	0.702	PRJNA556735 (SRA|OV)	SAMN12384872 (SRA|OV)
SRR9860219 (SRA|OV)	0.857	PRJNA556735 (SRA|OV)	SAMN12384872 (SRA|OV)
SRR9860218 (SRA|OV)	0.857	PRJNA556735 (SRA|OV)	SAMN12384872 (SRA|OV)
SRR9860217 (SRA|OV)	0.857	PRJNA556735 (SRA|OV)	SAMN12384872 (SRA|OV)
SRR9860216 (SRA|OV)	0.828	PRJNA556735 (SRA|OV)	SAMN12384872 (SRA|OV)

Page Size: 100 1 to 100 of 111 Page 1 of 2

Logan-search answer: geographical



Logan-search answer: exploit metadata



Logan-search answer:

Retrieve contigs or unitigs matching the query

SRR6322938 × unitigs contigs **Search**

Unitig ID	Length	Kmer Found	Unitig Kmer C...	Abundance	Unitig
SRR6322938_134...	114	84	1	14.2	AGCGTGAGAAGT...
SRR6322938_182...	301	119	0.44	15.9	GCAGCTAATCAG...
SRR6322938_801...	61	31	1	8	AAGTGTTTGTATA...
SRR6322938_821...	61	3	0.1	6.1	GCAGGAACACAA...

Anthony Baire, Pierre Marijon, Francesco Andreace and Pierre Peterlongo (2024).
Back to sequences: Find the origin of k-mers.
Journal of Open Source Software, 9(101), 7066, https://doi.org/10.21105/joss.07066

Existing alternatives

	Indexed	Metric	Answer	Metadata	Query time	Sub-index selection
Logan-search	~48.2PB	Approx. kmers Exact kmers	Accession id matched sequence*	Visualization + analyse	5mn / 15mn**	All or any group(s)
PebbleScout [1]	~3.7PB	Exact minimizers	Accession id	Brief summary	Instant	One at a time
Metagraph [2]	~2.2PB	Exact kmers	ids + graph align	Taxonomy	Instant	One at a time

* As a second step, finding contig or unitig sharing kmers with the query.

** Whether or not virtual machines are running

[1] Shiryev, S. A., & Agarwala, R. (2024). Indexing and searching petabase-scale nucleotide resources. Nature Methods, 1-9.
[2] Karasikov, M., Mustafa, H., Danciu, D., Zimmermann, M., Barber, C., Rätsch, G., & Kahles, A. (2024). Indexing all life's known biological sequences. BioRxiv.

Thanks!

https://logan-search.org/

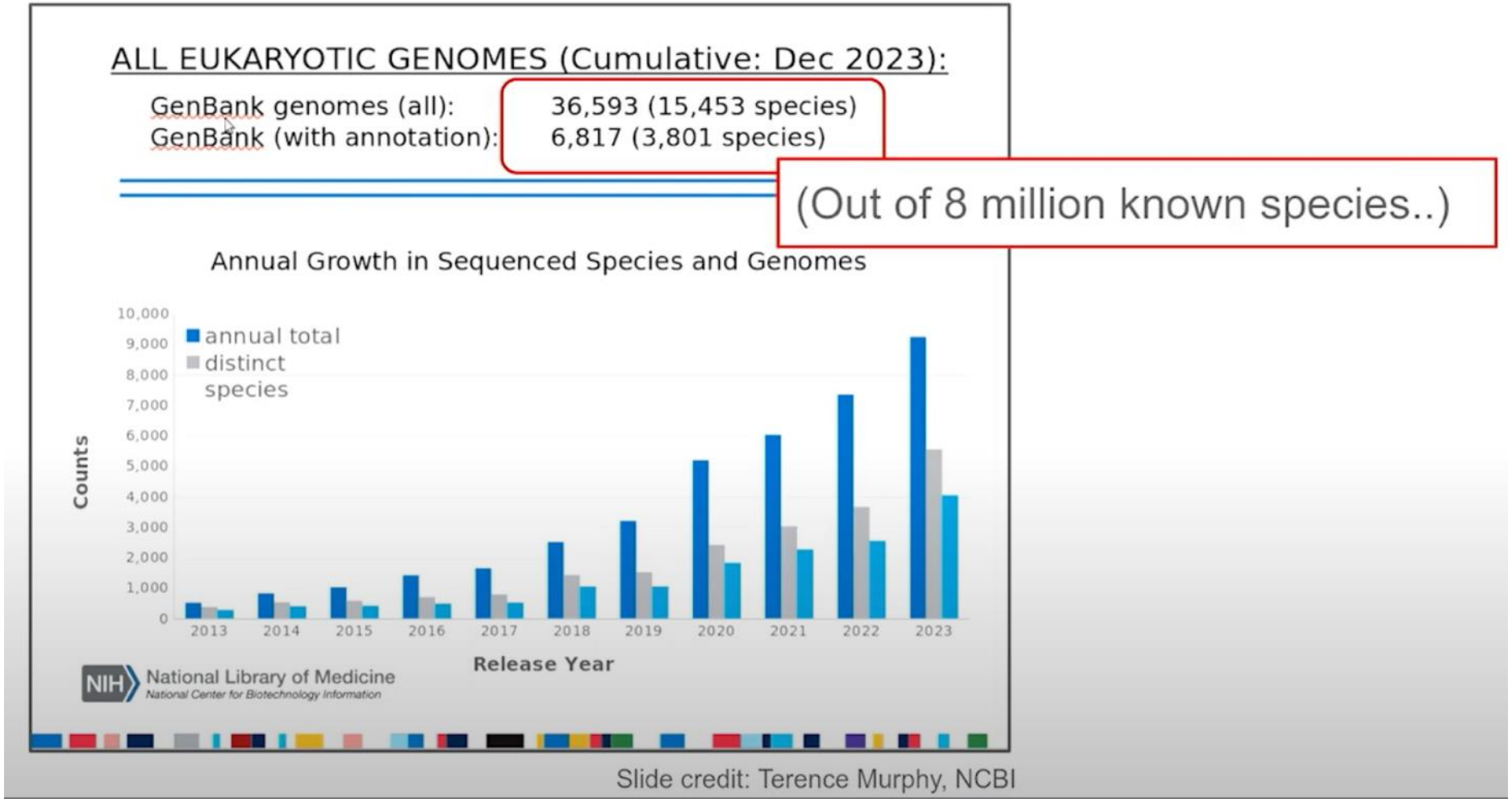
Documentation

- Logan: https://github.com/IndexThePlanet/Logan
- kmindex: https://tlemane.github.io/kmindex/
- kmviz: https://tlemane.github.io/kmviz/

Logan Search Issues

- https://github.com/IndexThePlanet/LoganSearch

We have genomes?



@Rayan Chikhi CGSI -- https://www.youtube.com/watch?v=H-8L_hFWXCQ

Use blast?

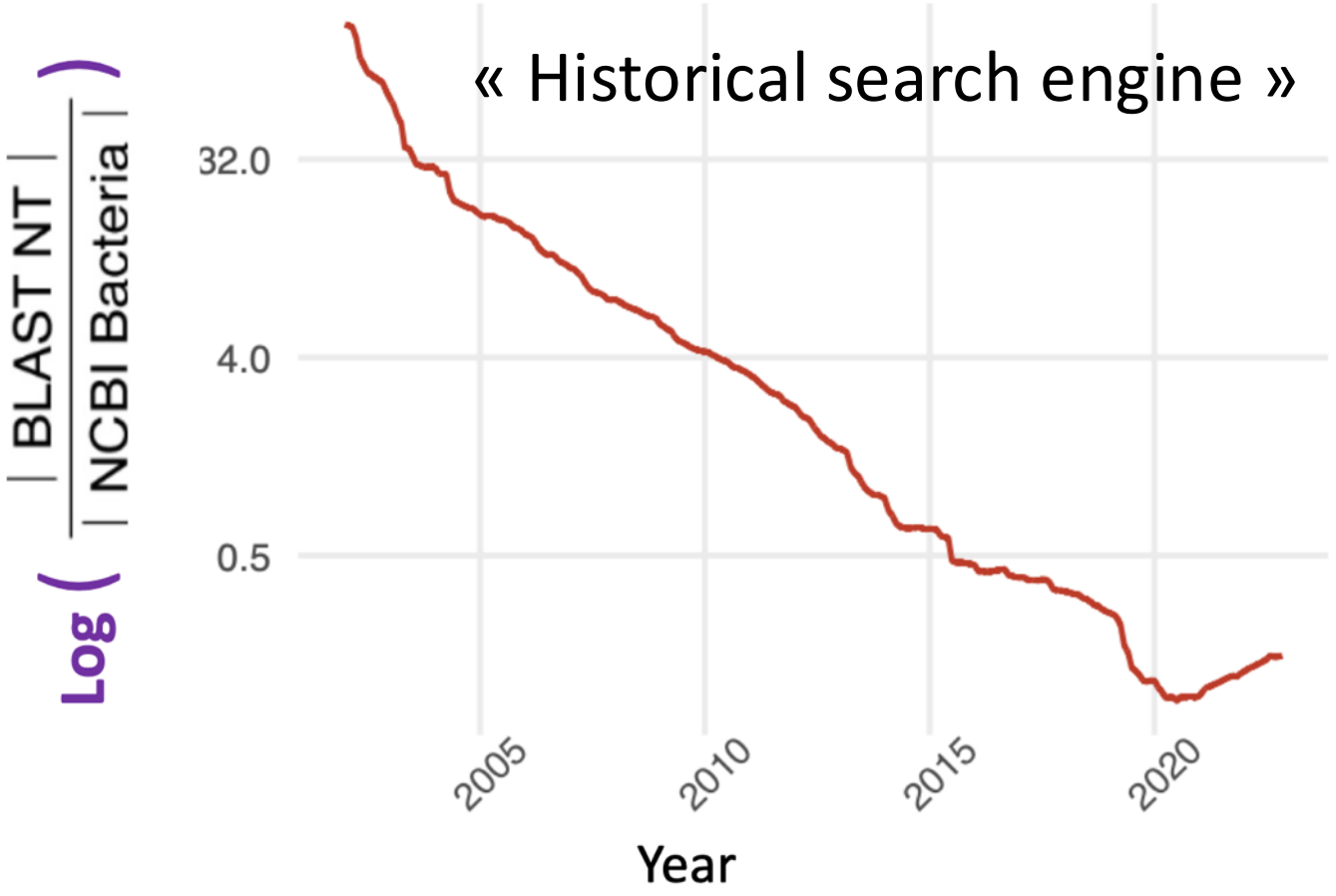
Basic local alignment search tool

SF Altschul, W Gish, W Miller, EW Myers… - Journal of molecular …, 1990 - Elsevier

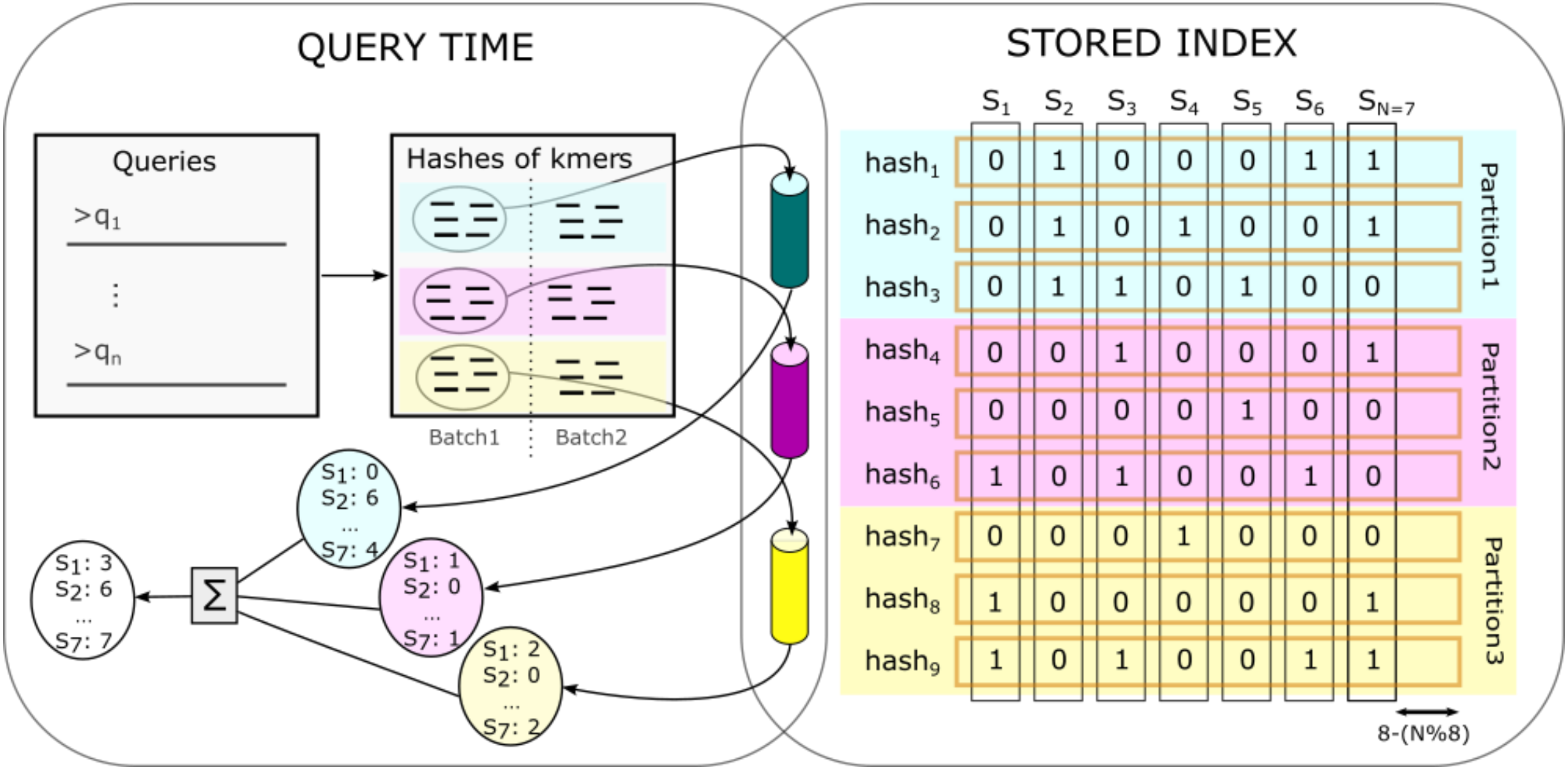
… A new approach to rapid **sequence** comparison, **basic local alignment search tool** (BLAST), directly approximates **alignments** that optimize a measure of **local** similarity, the maximal …

☆ Save Cite Cited by 111045 Related articles All 43 versions

« Historical search engine »



Břinda et al, *bioRxiv*, 2023


QUERY TIME
Queries
>q1
⋮
>qn
Hashes of kmers
Batch1
Batch2
S1: 0
S2: 6
...
S7: 4
S1: 1
S2: 0
...
S7: 1
S1: 2
S2: 0
...
S7: 2
Σ
S1: 3
S2: 6
...
S7: 7
STORED INDEX
S1 S2 S3 S4 S5 S6 SN=7
hash1 0 1 0 0 0 1 1
hash2 0 1 0 1 0 0 1
hash3 0 1 1 0 1 0 0
hash4 0 0 1 0 0 0 1
hash5 0 0 0 0 1 0 0
hash6 1 0 1 0 0 1 0
hash7 0 0 0 1 0 0 0
hash8 1 0 0 0 0 0 1
hash9 1 0 1 0 0 1 1
Partition1
Partition2
Partition3
8-(N%8)

Logan +

~50 petabases
of raw reads (SRA)

385 terabytes contigs
s3://logan-pub/c/

2.18 petabytes unitigs
s3://logan-pub/u/

27 million samples
- 2 million billions k-mers

kmindex

2 centuries cpu time

109 sub-indexes:

Genomic Transcript. MetaG MetaT SingleCell …	×	Human Mice Viruses Mamals Bact. …

Chikhi, R., Raffestin, B., Korobeynikov, A., Edgar, R. C., & Babaian, A. (2024). Logan: Planetary-Scale Genome Assembly Surveys Life's Diversity. bioRxiv, 2024-07.