

Looking through **all** the genomic data ever published: the Logan project

Roland Faure¹, Anton Nekrutenko², Kelsey Beavers³, Wei Shen⁴, Rayan Chikhi¹

¹Institut Pasteur ²Pennsylvania State University ³Texas Advanced Computing Center ⁴Chongqing Medical University

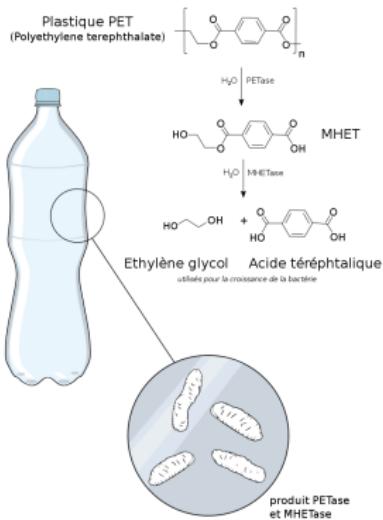
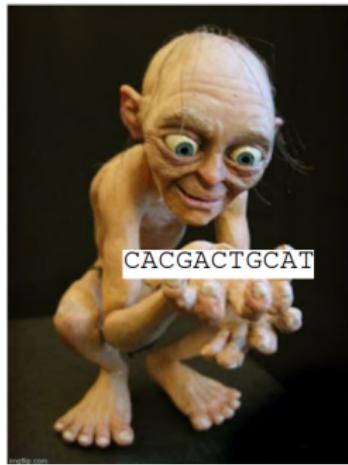
EAGS, February 2026

Slides available (CC-BY) at: rolandfaure.github.io

POV: you have a favorite sequence

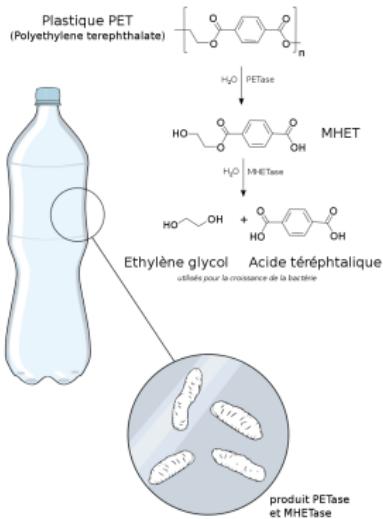


POV: you have a favorite sequence



- ▶ Wow, a PETase!!

POV: you have a favorite sequence



- ▶ Wow, a PETase!!
- ▶ Has this already been described?
- ▶ Is it found somewhere else? Where?
- ▶ I want to find other PETases!

Let's BLAST the PETase

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

NEWS

Mon, 21 Jul 2025
Here are a few highlights in our latest BLAST+ release:
[Download BLAST+ 2.17.0 now!](#) [More BLAST news...](#)

Web BLAST

Nucleotide BLAST nucleotide ▶ nucleotide

blast translated nucleotide ▶ protein

tblastn protein ▶ translated nucleotide

Protein BLAST protein ▶ protein

select all 50 sequences selected

GenBank Graphics Distance tree of results MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/> <i>Vibrio gazogenes</i> strain ATCC 43942 chromosome 1, complete sequence	<i>Vibrio gazogenes</i>	1618	1618	100%	0.0	100.00%	3471064	CP016835.1
<input checked="" type="checkbox"/> <i>Vibrio gazogenes</i> strain PB1 chromosome 1, complete sequence	<i>Vibrio gazogenes</i>	1416	1416	100%	0.0	94.98%	3516273	CP092587.1
<input checked="" type="checkbox"/> <i>Vibrio gazogenes</i> strain DSM 21264 chromosome 1, complete sequence	<i>Vibrio gazogenes</i>	1416	1416	100%	0.0	94.98%	3516262	CP151640.1
<input checked="" type="checkbox"/> <i>Vibrio gazogenes</i> ATCC 29988 DNA, chromosome 1, complete sequence	<i>Vibrio gazogenes</i>	1416	1416	100%	0.0	94.98%	3516247	AP024873.1
<input checked="" type="checkbox"/> <i>Vibrio spartinus</i> CECT 9026 DNA, chromosome 1, complete sequence	<i>Vibrio spartinus</i>	1321	1321	100%	0.0	92.64%	4003627	AP024807.1
<input checked="" type="checkbox"/> <i>Vibrio spartinus</i> strain 3.6 chromosome 1, complete sequence	<i>Vibrio spartinus</i>	1321	1321	100%	0.0	92.64%	3817659	CP060268.1

Let's BLAST the PETase

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

NEWS

Mon, 21 Jul 2025
Here are a few highlights in our latest BLAST+ release:
[Download BLAST+ 2.17.0 now!](#) [More BLAST news...](#)

Web BLAST

Nucleotide BLAST nucleotide ▶ nucleotide

blast translated nucleotide ▶ protein

tblastn protein ▶ translated nucleotide

Protein BLAST protein ▶ protein

select all 50 sequences selected

GenBank Graphics Distance tree of results MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident.	Acc. Len	Accession
<input checked="" type="checkbox"/> <i>Vibrio gazogenes</i> strain ATCC 43942 chromosome 1, complete sequence	<i>Vibrio gazogenes</i>	1618	1618	100%	0.0	100.00%	3471064	CP016835.1
<input checked="" type="checkbox"/> <i>Vibrio gazogenes</i> strain PB1 chromosome 1, complete sequence	<i>Vibrio gazogenes</i>	1416	1416	100%	0.0	94.98%	3516273	CP092587.1
<input checked="" type="checkbox"/> <i>Vibrio gazogenes</i> strain DSM 21264 chromosome 1, complete sequence	<i>Vibrio gazogenes</i>	1416	1416	100%	0.0	94.98%	3516262	CP151640.1
<input checked="" type="checkbox"/> <i>Vibrio gazogenes</i> ATCC 29988 DNA, chromosome 1, complete sequence	<i>Vibrio gazogenes</i>	1416	1416	100%	0.0	94.98%	3516247	AP024873.1
<input checked="" type="checkbox"/> <i>Vibrio spartinus</i> CECT 9026 DNA, chromosome 1, complete sequence	<i>Vibrio spartinus</i>	1321	1321	100%	0.0	92.64%	4003627	AP024807.1
<input checked="" type="checkbox"/> <i>Vibrio spartinus</i> strain 3.6 chromosome 1, complete sequence	<i>Vibrio spartinus</i>	1321	1321	100%	0.0	92.64%	3817659	CP060268.1

► Found 50 hits from a few different species

What is indexed in BLAST exactly?

In BLAST nr/nt



Well-assembled genomes

Non-redundant

NOT in BLAST nr/nt



Random pieces of sequenced genomes

Every version of every sequence

What is indexed in BLAST exactly?

In BLAST nr/nt



Well-assembled genomes

Non-redundant

0.1%

NOT in BLAST nr/nt



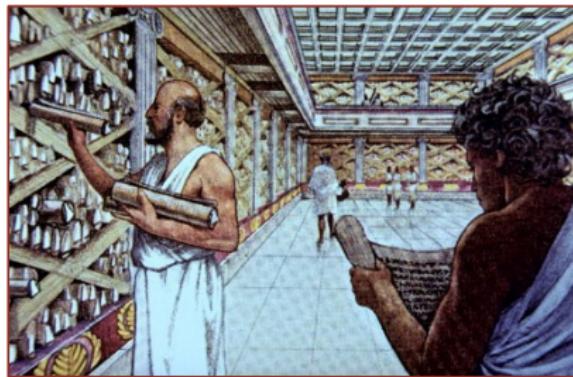
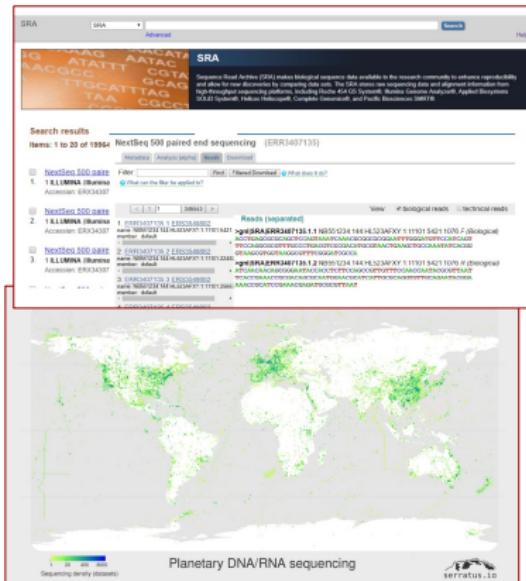
Random pieces of sequenced genomes

Every version of every sequence

99.9%

The biggest database: the SRA

SRA: All public sequencing reads, 80 PB of data



"Library of Alexandria" for genetics

Slide Credits: Rayan Chikhi

Let's BLAST against *everything*

- ▶ Let's align our sequence against all the reads of the SRA

Let's BLAST against *everything*

- ▶ Let's align our sequence against all the reads of the SRA
A thousand years later...



Let's BLAST against *everything*

- ▶ Let's align our sequence against all the reads of the SRA
A thousand years later...



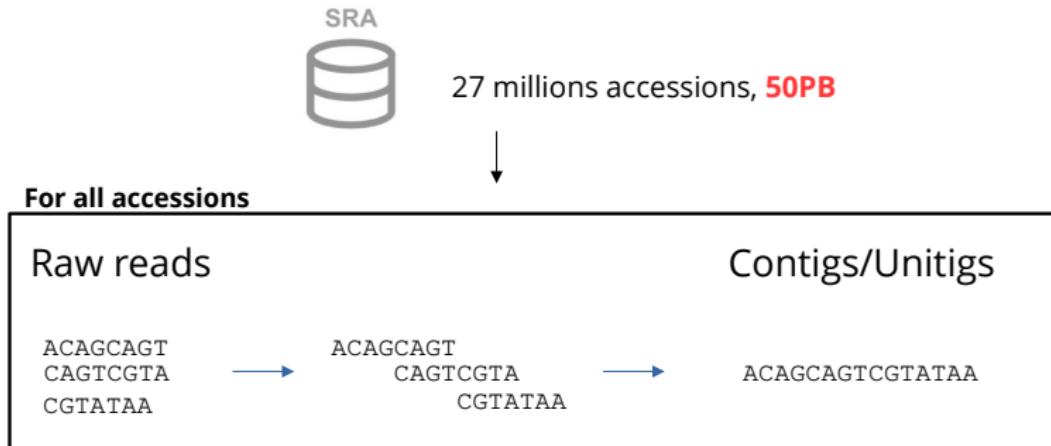
- ▶ Doing this is the primary goal of the **Logan project**

The Logan project: step 1



27 millions accessions, **50PB**

The Logan project: step 1



Let's BLAST against *everything* - second try

- ▶ Let's align our sequence against all the ~~reads of the SRA~~ contigs of Logan

Let's BLAST against *everything* - second try

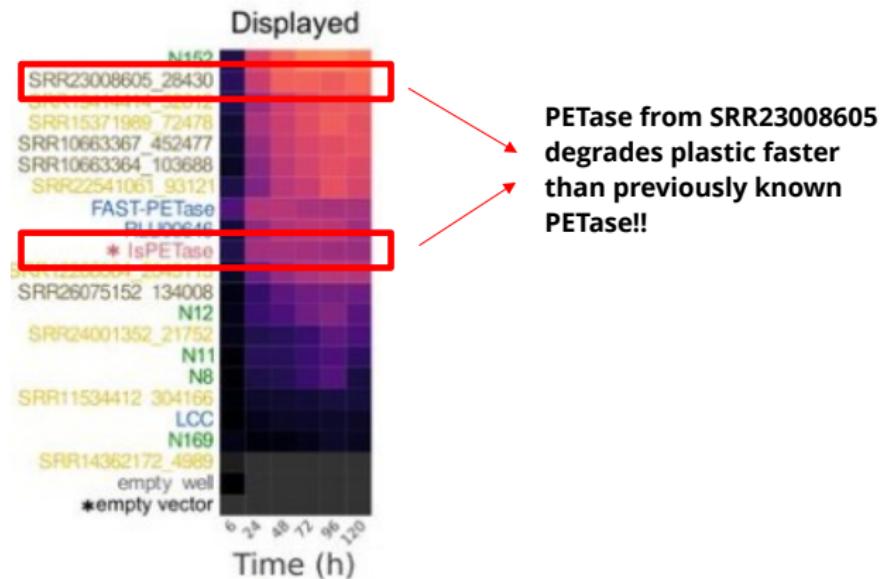
- ▶ Let's align our sequence against all the ~~reads of the SRA~~ contigs of Logan
- ▶ Doable!

Let's BLAST against *everything* - second try

- ▶ Let's align our sequence against all the ~~reads~~ of the SRA contigs of Logan
- ▶ Doable! *If you have 10 000\$*

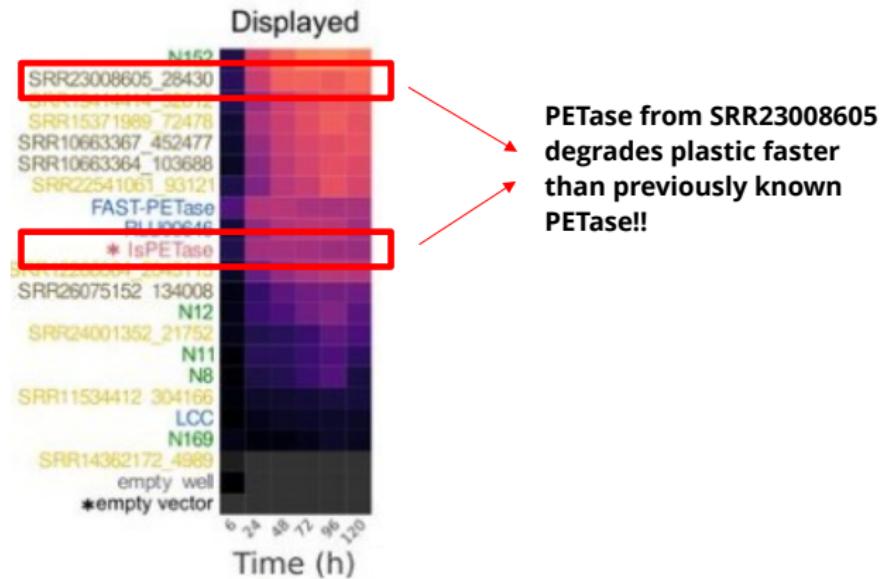


Finding new PETase in the SRA



Logan: Planetary-Scale Genome Assembly Surveys Life's Diversity, Chikhi et al., 2025

Finding new PETase in the SRA



- ▶ Let's try to browse for cheaper, shall we?

Browsing Logan: strategy 1 - Logan-search.org

kmviz v0.8.0

INPUT

text file session

Query sequence(s) *

Fasta/Fastq format

```
>Query
ACCGTAGCCTAGAATT
```

Load

NOTIFICATION

Email

CONFIGURATION

Groups

Threshold = 0.5

0.25 1.0

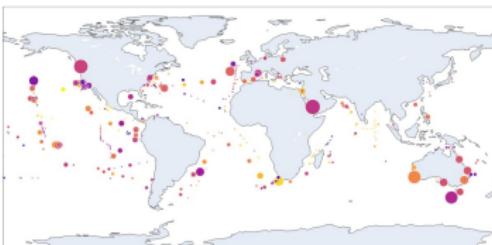
5 minutes
later



Table Map Plot BLAST-like alignment Help

kmer_coverage > 0.7 AND assay_type IN ('WGS','WGA')

ID	kmer_coverage	bioproject
SRR18278860 (SRA OV)	0.991	PRJNA814250 (SRA OV)
ERR3497222 (SRA OV)	0.991	PRJEB34158 (SRA OV)
ERR3497221 (SRA OV)	0.971	PRJEB34158 (SRA OV)
ERR6909055 (SRA OV)	0.971	PRJEB47927 (SRA OV)
SRR1296779 (SRA OV)	0.971	PRJNA248394 (SRA OV)
SRR1296780 (SRA OV)	0.971	PRJNA248394 (SRA OV)
SRR1296778 (SRA OV)	0.971	PRJNA248394 (SRA OV)



Let's BLAST against *everything* - third try

Table Map Plot BLAST-like alignment Help

kmer_coverage > 0.7 AND assay_type IN ('WGS','WGA')

ID	kmer_coverage	bioproject	biosample	bioproject_title	bioproject_description
SRR7154899 (SRA OV)	0.63	PRJNA465688 (SRA OV)	SAMN09092028 (SRA OV)	Coastal salt marsh microbial c...	Coastal salt marsh microbial c...
ERR1817120 (SRA OV)	0.295	PRJEB17713 (SRA OV)	SAMEA78796918 (SRA OV)	Whole genome sequencing of ...	Vibrio gazogenes is a member ...

- ▶ Found a hit in a coastal salt marsh microbial community!

Let's BLAST against *everything* - third try

Table Map Plot BLAST-like alignment Help

kmer_coverage > 0.7 AND assay_type IN ('WGS','WGA')

ID	kmer_coverage	bioproject	biosample	bioproject_title	bioproject_description
SRR7154899 (SRA OV)	0.63	PRJNA465688 (SRA OV)	SAMN09092028 (SRA OV)	Coastal salt marsh microbial c...	Coastal salt marsh microbial c...
ERR1817120 (SRA OV)	0.295	PRJEB17713 (SRA OV)	SAMEA78796918 (SRA OV)	Whole genome sequencing of ...	Vibrio gazogenes is a member ...

- ▶ Found a hit in a coastal salt marsh microbial community!
- ▶ Wait, only two hits !??

Let's BLAST against *everything* - third try

Table Map Plot BLAST-like alignment Help

kmer_coverage > 0.7 AND assay_type IN ('WGS','WGA')

ID	kmer_coverage	bioproject	biosample	bioproject_title	bioproject_description
SRR7154899 [SRA OV]	0.63	PRJNA465688 [SRA OV]	SAMN09092028 [SRA OV]	Coastal salt marsh microbial c...	Coastal salt marsh microbial c...
ERR1817120 [SRA OV]	0.295	PRJEB17713 [SRA OV]	SAMEA78796918 [SRA OV]	Whole genome sequencing of ...	Vibrio gazogenes is a member ...

- ▶ Found a hit in a coastal salt marsh microbial community!
- ▶ Wait, only two hits !??
- ▶ Relies on *exact* 31-mers matches → not very sensitive

Let's BLAST against *everything* - third try

Table Map Plot BLAST-like alignment Help

kmer_coverage > 0.7 AND assay_type IN ('WGS','WGA')

ID	kmer_coverage	bioproject	biosample	bioproject_title	bioproject_description
SRR7154899 [SRA OV]	0.63	PRJNA465688 [SRA OV]	SAMN09092028 [SRA OV]	Coastal salt marsh microbial c...	Coastal salt marsh microbial c...
ERR1817120 [SRA OV]	0.295	PRJEB17713 [SRA OV]	SAMEA78796918 [SRA OV]	Whole genome sequencing of ...	Vibrio gazogenes is a member ...

- ▶ Found a hit in a coastal salt marsh microbial community!
- ▶ Wait, only two hits !??
- ▶ Relies on *exact* 31-mers matches → not very sensitive
- ▶ *Logan: Planetary-Scale Genome Assembly Surveys Life's Diversity* - biorXiv

Browsing Logan: strategy 2 - LexicMap

Article | [Open access](#) | Published: 10 September 2025

Efficient sequence alignment against millions of prokaryotic genomes with LexicMap

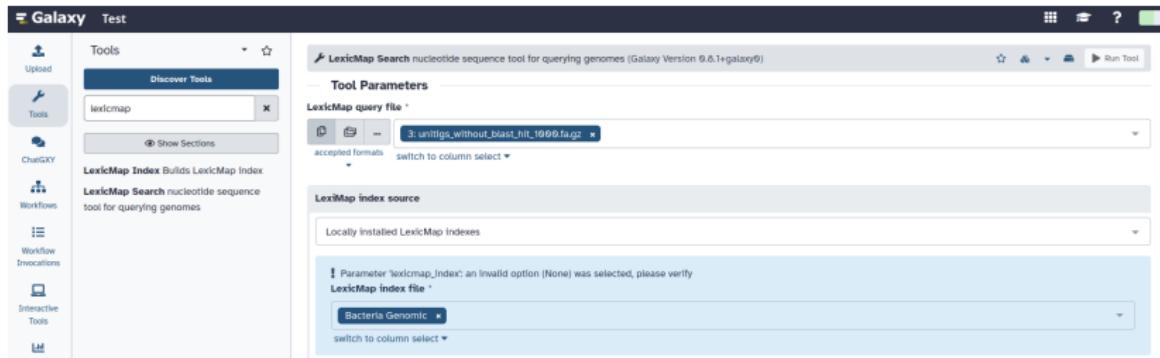
[Wei Shen](#) , [John A. Lees](#) & [Zamin Iqbal](#) 

Browsing Logan: strategy 2 - LexicMap

Article | [Open access](#) | Published: 10 September 2025

Efficient sequence alignment against millions of prokaryotic genomes with LexicMap

Wei Shen , John A. Lees & Zamin Iqbal 



The screenshot shows the Galaxy web interface with the 'Test' workspace selected. The left sidebar has 'Tools' selected. The main panel shows the 'LexicMap Search' tool configuration. The 'Tool Parameters' section has a 'LexicMap query file' set to '3:unitigs_without_blast_hit_1000.fasta' and an 'accepted formats' dropdown. The 'LexicMap index source' section has 'Locally Installed LexicMap Indexes' selected and a 'Bacteria Genomic' dropdown. A warning message 'Parameter 'lexicmap_index': an invalid option (None) was selected, please verify' is displayed above the index source dropdown.

- ▶ Few hours of computation to return results
- ▶ Soon available on Galaxy (usegalaxy.org)

Let's BLAST against *everything* - fourth try

- ▶ 200 results
- ▶ Detects sequences with $\geq 70\%$ similarity

Let's BLAST against *everything* - fourth try

- ▶ 200 results
- ▶ Detects sequences with $\geq 70\%$ similarity
- ▶ PETase found e.g. in "Fermented Xuecai"



Source: chillcrispbyxueci.substack.com

Limits of LexicMap

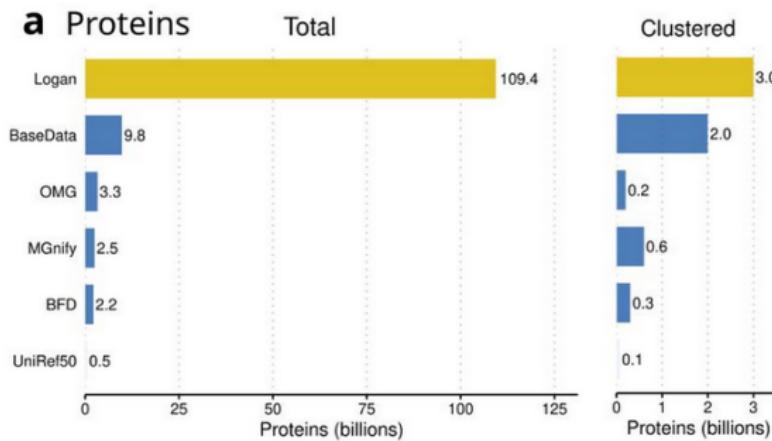
- ▶ Relatively slow
- ▶ Still less sensitive than BLAST or DIAMOND...

Browsing Logan: strategy 3 - **protein search**

- ▶ Let's focus only on the proteins (100x smaller)

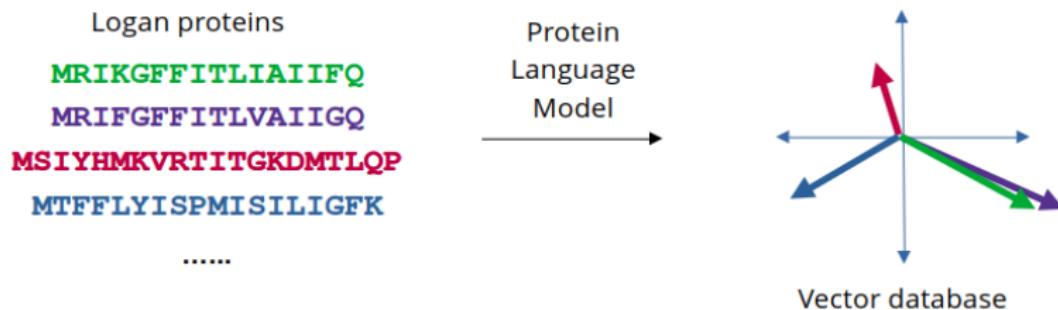
Browsing Logan: strategy 3 - protein search

- ▶ Let's focus only on the proteins (100x smaller)
- ▶ Detect proteins of Logan with Prodigal
- ▶ 100 billion proteins clustered in 3 billion clusters (accessible online)



Browsing Logan: strategy 3 - protein search

- ▶ With Protein Language Models, proteins can be transformed in vectors
- ▶ We can build a search engine on vectors (just like Google)



Let's look for the PETase

- ▶ One hour computation
- ▶ 300k results!
- ▶ Down to 20% amino-acid identity!

Query_10001	150	YHQIDEKEIGIVGYSQGGAGAYNTLEGKDGDK--FKTM- TVSG-----VTE SIGEKLHL PWIYDPSK VTI	213
Query_10002	98	QPEVDANRVALIGWALGGGVVVAEADDQRVK--AVVTCNAIGDGERS[5]DEQswsRLQD- DMVRDRPERARSGRSRTV	176
Query_10003	95	EK-TGNPRVGVVGYCAGGGLALM LAAQRP-----DAVKAVA	149
Query_10004	130	SAFVDPERIGVIGFSHGGWTLLDFLGPPAIHasATATDARDGLRSVV	198
Query_10005	151	YQKVDT EHI GIGSHS QGGGVGVFN AISEQPHSN--LYTCAVS LSP-----TQQD LAEALKIP--YDPT KTOI	212
Query_10006	99	DP-RCTGKVGIVGFCMGGGFTLLLAPRG-----IFDAAA	148
Query_10007	99	LEF VDPDRIGVLGV CAGGGY SVNAAMTEHRIKav GTVVGANI G-----[4]ENNpiqTLEAIGK QRTAEANGA EPMI INW	174
Query_10008	136	GKF ILSNKIAVIGHSMGGYTALALAGGI PWTQeaERVETSSD ARVKAI	202

Soon available on Galaxy

- ▶ Soon on Galaxy (usegalaxy.org)

The screenshot shows the Galaxy web interface. On the left is a sidebar with navigation links: Upload, Tools (selected), Workflows, Visualization, Histories, and Pages. The main content area shows a search bar with 'All Tools' and 'Discover Tools' buttons, and a search input field. Below the search is a text block: 'Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.' To the right of this text is a promotional banner for the 'GALAXY COMMUNITY CONFERENCE 2026' in Clermont-Ferrand, France, June 22-24, 2026. The banner features a colorful illustration of a city skyline with a cathedral and a hot air balloon. Below the banner is a link: 'Followed by CoFest: June 25-26, 2026'. To the right of the banner is a video thumbnail titled 'What is Galaxy? GalaxyProject' with a play button, and a 'Watch on YouTube' link.

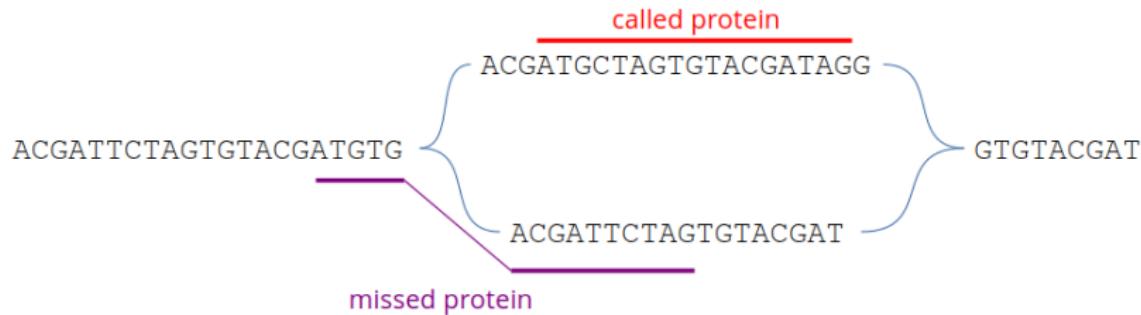
- ▶ Maybe also distributed and installable?? (a few TB of data)

Limits of this index

- ▶ Only full proteins

Limits of this index

- ▶ Only full proteins
- ▶ 90% proteins missing in the database because of the protein calling



Take-home messages

- ▶ It is possible to look for your sequence **in the SRA**

Take-home messages

- ▶ It is possible to look for your sequence **in the SRA**
- ▶ Already existing Logan-search.org, included in Logan preprint:
Logan: Planetary-Scale Genome Assembly Surveys Life's Diversity

Take-home messages

- ▶ It is possible to look for your sequence **in the SRA**
- ▶ Already existing **Logan-search.org**, included in Logan preprint:
Logan: Planetary-Scale Genome Assembly Surveys Life's Diversity
- ▶ Upcoming: **more sensitive searches**, online on Galaxy

Take-home messages

- ▶ It is possible to look for your sequence **in the SRA**
- ▶ Already existing **Logan-search.org**, included in Logan preprint:
Logan: Planetary-Scale Genome Assembly Surveys Life's Diversity
- ▶ Upcoming: **more sensitive searches**, online on Galaxy
- ▶ We are looking for applications, **contact us!**



Acknowledgments

- ▶ Rayan Chikhi
- ▶ Logan team: Téo Lemane, Pierre Peterlongo, Artem Babaian & others
- ▶ Galaxy team: Anton Nekrutenko, Björn Grüning, Patrik Smeds, Nate Coraor
- ▶ TACC team: Kelsey Beavers, Felix Zuo
- ▶ Wei Shen



Rayan Chikhi

