

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

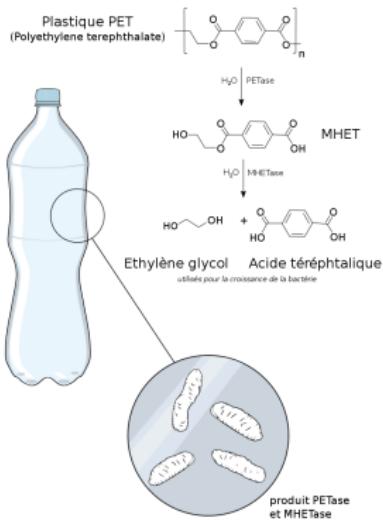
EAGS, February 2026

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

POV: you have a favorite sequence

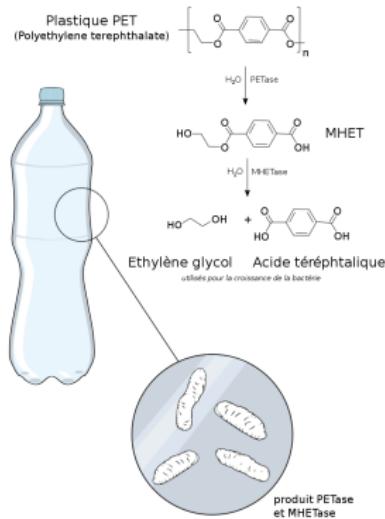


POV: you have a favorite sequence



- ▶ 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

blastx 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
Vibrio gazogenes strain ATCC 43942 chromosome 1, complete sequence	Vibrio gazogenes	1618	1618	100%	0.0	100.00%	3471064	CP016835.1
Vibrio gazogenes strain PB1 chromosome 1, complete sequence	Vibrio gazogenes	1416	1416	100%	0.0	94.98%	3516273	CP092587.1
Vibrio gazogenes strain DSM 21264 chromosome 1, complete sequence	Vibrio gazogenes	1416	1416	100%	0.0	94.98%	3516262	CP151640.1
Vibrio gazogenes ATCC 29988 DNA, chromosome 1, complete sequence	Vibrio gazogenes	1416	1416	100%	0.0	94.98%	3516247	AP024873.1
Vibrio spartinus CECT 9026 DNA, chromosome 1, complete sequence	Vibrio spartinus	1321	1321	100%	0.0	92.64%	4003627	AP024807.1
Vibrio spartinus strain 3.6 chromosome 1, complete sequence	Vibrio spartinus	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

blastx 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
Vibrio gazogenes strain ATCC 43942 chromosome 1, complete sequence	<i>Vibrio gazogenes</i>	1618	1618	100%	0.0	100.00%	3471064	CP016835.1
Vibrio gazogenes strain PB1 chromosome 1, complete sequence	<i>Vibrio gazogenes</i>	1416	1416	100%	0.0	94.98%	3516273	CP092587.1
Vibrio gazogenes strain DSM 21264 chromosome 1, complete sequence	<i>Vibrio gazogenes</i>	1416	1416	100%	0.0	94.98%	3516262	CP151640.1
Vibrio gazogenes ATCC 29988 DNA, chromosome 1, complete sequence	<i>Vibrio gazogenes</i>	1416	1416	100%	0.0	94.98%	3516247	AP024873.1
Vibrio spartinus CECT 9026 DNA, chromosome 1, complete sequence	<i>Vibrio spartinus</i>	1321	1321	100%	0.0	92.64%	4003627	AP024807.1
Vibrio spartinus 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



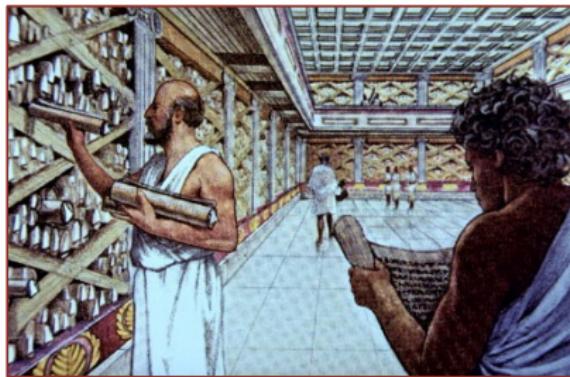
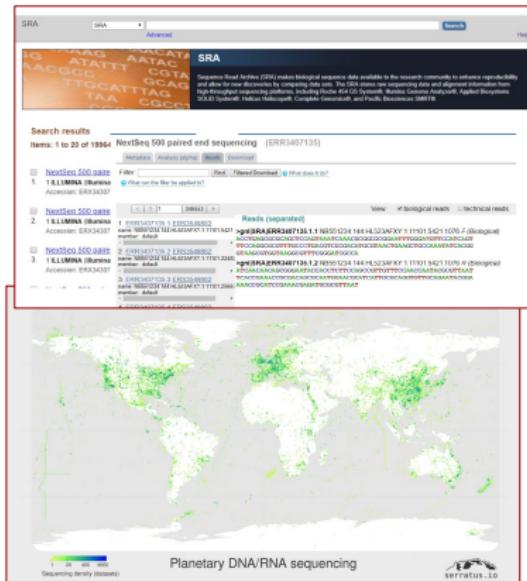
Random pieces of sequenced genomes

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

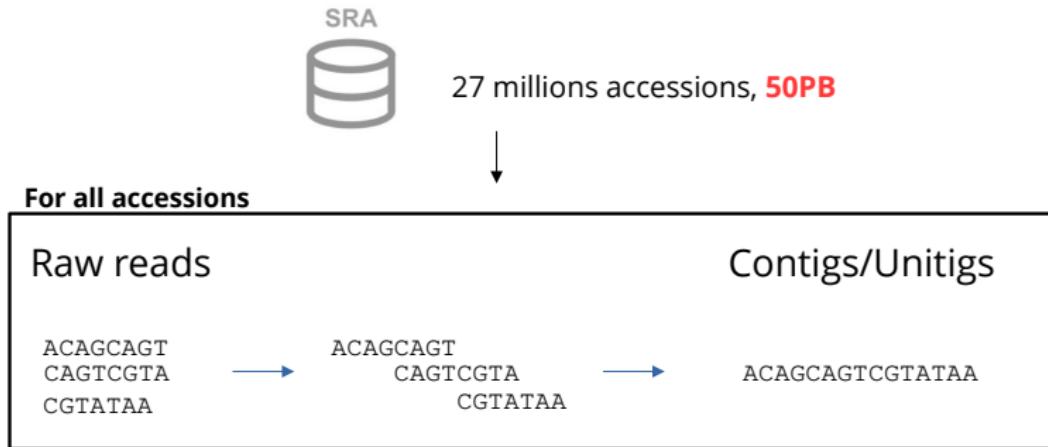
The Logan project: step 1



SRA

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

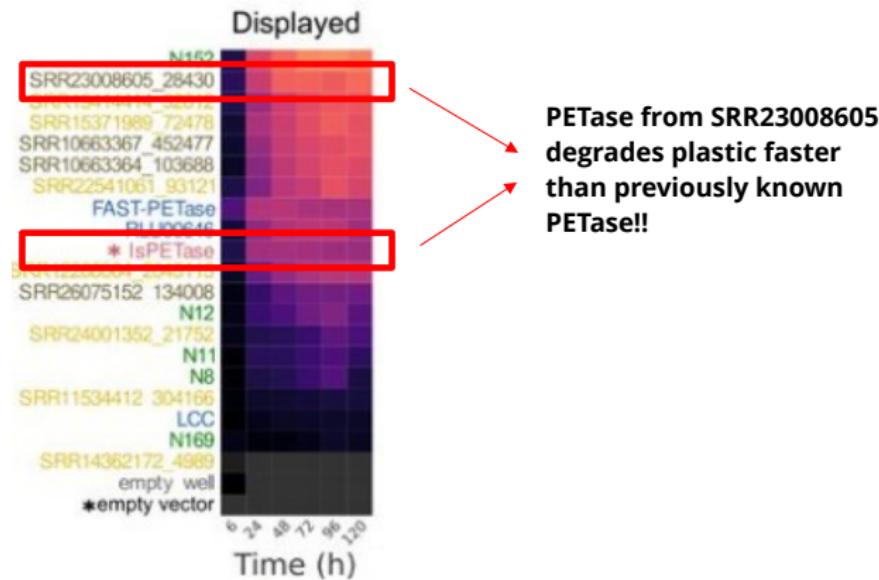
- ▶ Let's align our sequence against all the ~~reads of the SRA~~ contigs of Logan
- ▶ Doable!

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\$*

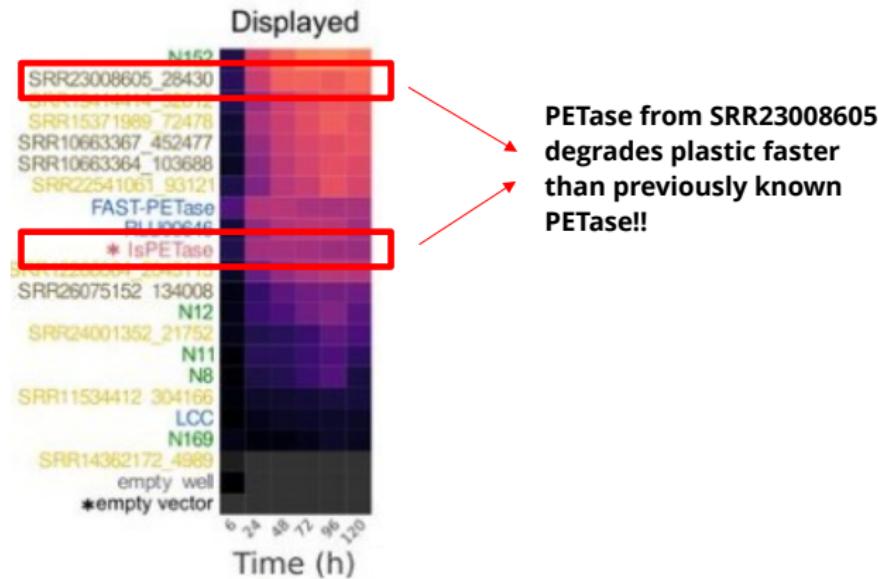


Finding new PETase in the SRA



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

Finding new PETase in the SRA



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

- ▶ 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
ACCGTAGCCTAGAATTAA
```

Load

NOTIFICATION

Email

CONFIGURATION

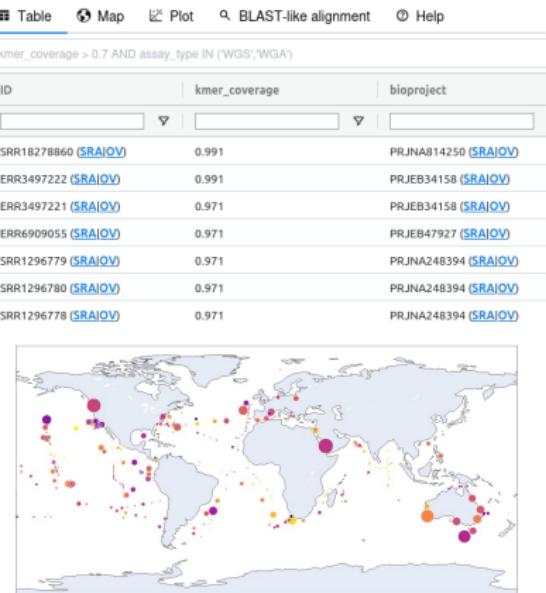
Groups

Threshold = 0.5

0.25 1.0

Submit Reset

5 minutes
later



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 title "Test". On the left, there's a sidebar with links for Upload, Tools (highlighted), Chat/GYX, Workflows, Workflow Invocations, and Interactive Tools. The main area shows the "Tools" section with "Discover Tools" and a search bar containing "lexicmap". Below this, the "LexicMap Index" tool is listed, which "Builds LexicMap Index". The "LexicMap Search" tool is also listed, described as "nucleotide sequence tool for querying genomes". The "Tool Parameters" section for "LexicMap Search" is open, showing the "Tool Parameters" tab. It includes fields for "LexicMap query file" (containing "3:unitigs_without_blast_hit_1000.fasta"), "accepted formats" (with "fa" selected), and "LexicMap index source" (set to "Locally Installed LexicMap Indexes"). A warning message states: "Parameter 'lexicmap_index' is an invalid option (None) was selected, please verify LexicMap index file". A dropdown menu for "Bacteria Genomic" is shown, with "switch to column select" below it.

- ▶ 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

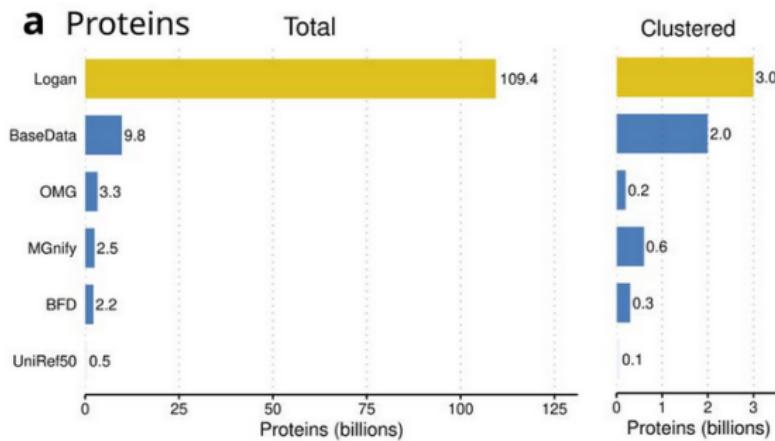
- ▶ Relatively slow
- ▶ Still less sensitive than BLAST or DIAMOND...

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

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

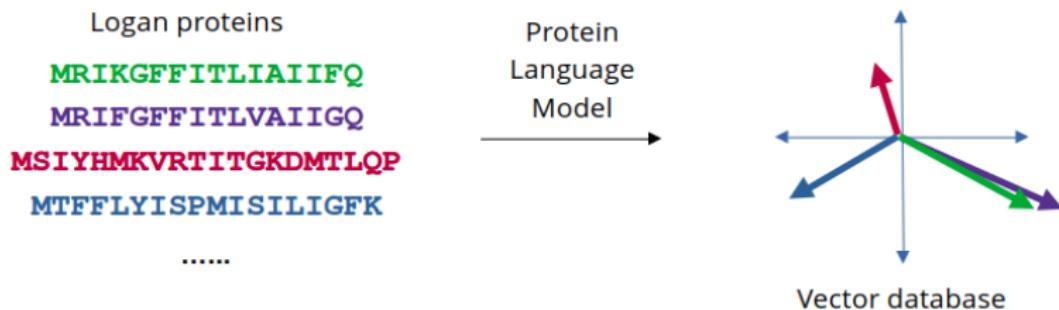
Browsing Logan: strategy 3 - protein search

- ▶ 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-VTVSG--	-----VTE SIGEKLHL PVWIYDPSK VTI	213
Query_10002	98	QPEVDANRVALIGWALGGGVVVAAADDQRVK--AVVTCNAIGDGERS[5]DEQswsRLQD-DMVDRPERARSGRSRTV	-----GLIPWPGAQPDWAAIDA	176
Query_10003	95	EK-TGNPRVGVVGYCAGGGLALMLAAQRP-----DAVKAVA	PYY-----GLIPWPGAQPDWAAIDA	149
Query_10004	130	SAFVDPERIGVIGFSHGGWTLLDFLGPPAIHasATATDARDGLRSVV	AVYpycGADVQAGLKWP-----ADV	198
Query_10005	151	YQKVDTETIGISGHSQGGVGVFN AISEQPHSN--LYTCAVS LSP--	-----TQQD LAEALKIP--YDPTKTQI	212
Query_10006	99	DP-RCTGKVGIVGFCMGGGTLLLAPRG-----IFDAAA	PNY-----GVLP---RDLSALSSC	148
Query_10007	99	LEFVDPDRIGVLGVCAGGGYSVNAA MTEHRIKavGT VVGANI G-----[4]ENNpiqTLEAIGKQRTAEANGAEPMIINW	-----VLMAPGAGWF MNLSK VTI	174
Query_10008	136	GKFILSNKIAVIGHSMGGYTALALAGGI PWTQeaERVETSSD ARVKAI	-----VLMAPGAGWF MNLSK VTI	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, Pages, and Help. The main area displays a search bar with 'All Tools' and 'Discover Tools' buttons, and a search input field containing 'search tools'. Below the search is a descriptive 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 are two promotional banners. The first is for the 'GALAXY COMMUNITY CONFERENCE 2026' in Clermont-Ferrand, France, from June 22-24, 2026. It features a colorful illustration of a city skyline with a cathedral and a hot air balloon. The second is a video thumbnail titled 'What is Galaxy?' by the Galaxy Project, with a play button and a 'Watch on YouTube' link.

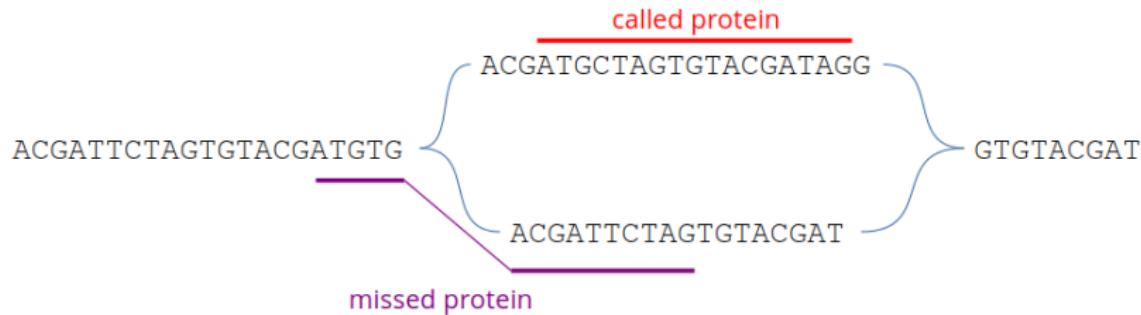
- ▶ Maybe also distributed and installable?? (a few TB of data)

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

