

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

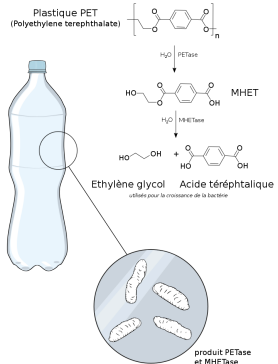
EAGS, February 2026

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

POV: you have a favorite sequence

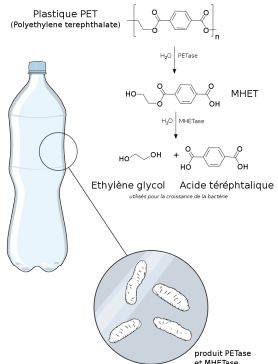
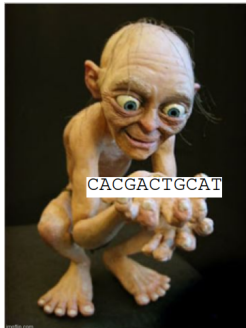


POV: you have a favorite sequence



► 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](#)

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
<input checked="" type="checkbox"/>	Vibrio gazogenes strain ATCC 43942 chromosome 1, complete sequence	Vibrio gazogenes	1618	1618	100%	0.0	100.00%	3471064	CP018835.1
<input checked="" type="checkbox"/>	Vibrio gazogenes strain PB1 chromosome 1, complete sequence	Vibrio gazogenes	1416	1416	100%	0.0	94.98%	3516273	CP092587.1
<input checked="" type="checkbox"/>	Vibrio gazogenes strain DSM 21264 chromosome 1, complete sequence	Vibrio gazogenes	1416	1416	100%	0.0	94.98%	3516282	CP151640.1
<input checked="" type="checkbox"/>	Vibrio gazogenes ATCC 29988 DNA, chromosome 1, complete sequence	Vibrio gazogenes	1416	1416	100%	0.0	94.98%	3516247	AP024873.1
<input checked="" type="checkbox"/>	Vibrio spartinae CECT 9026 DNA, chromosome 1, complete sequence	Vibrio spartinae	1321	1321	100%	0.0	92.64%	4003627	AP024907.1
<input checked="" type="checkbox"/>	Vibrio spartinae strain 3.6 chromosome 1, complete sequence	Vibrio spartinae	1321	1321	100%	0.0	92.64%	3817959	CP046268.1


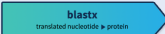

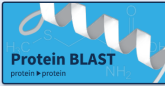
Let's BLAST the PETase

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](#)

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

☒ 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"/>	Vibrio gazogenes strain ATCC 43942 chromosome 1, complete sequence	Vibrio gazogenes	1618	1618	100%	0.0	100.00%	3471064	CP018835.1
<input checked="" type="checkbox"/>	Vibrio gazogenes strain PB1 chromosome 1, complete sequence	Vibrio gazogenes	1416	1416	100%	0.0	94.98%	3516273	CP092587.1
<input checked="" type="checkbox"/>	Vibrio gazogenes strain DSM 21264 chromosome 1, complete sequence	Vibrio gazogenes	1416	1416	100%	0.0	94.98%	3516282	CP151640.1
<input checked="" type="checkbox"/>	Vibrio gazogenes ATCC 29988 DNA, chromosome 1, complete sequence	Vibrio gazogenes	1416	1416	100%	0.0	94.98%	3516247	AP024873.1
<input checked="" type="checkbox"/>	Vibrio spartinae CECT 9026 DNA, chromosome 1, complete sequence	Vibrio spartinae	1321	1321	100%	0.0	92.64%	4003627	AP024907.1
<input checked="" type="checkbox"/>	Vibrio spartinae strain 3.6 chromosome 1, complete sequence	Vibrio spartinae	1321	1321	100%	0.0	92.64%	3817059	CP046268.1

► 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



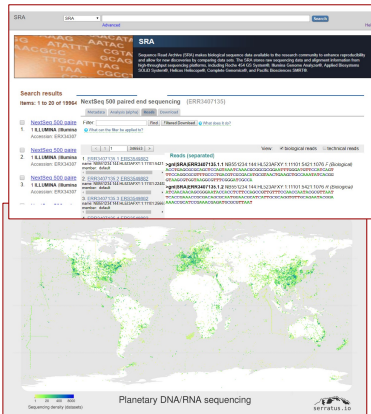
Random pieces of sequenced genomes

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...



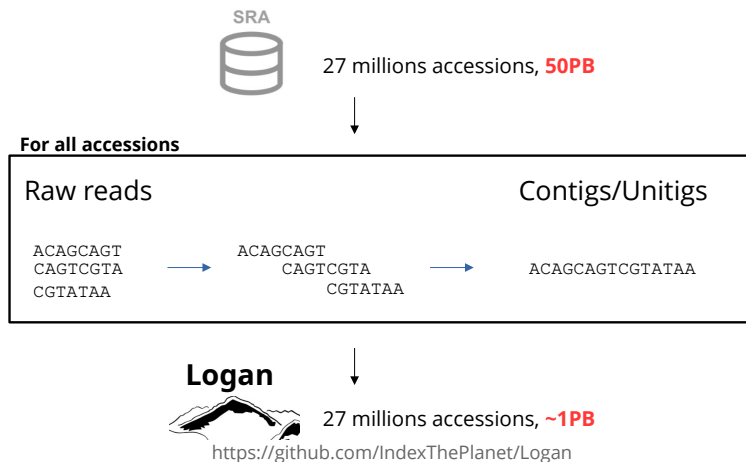
- ▶ Doing this is the primary goal of the **Logan project**

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

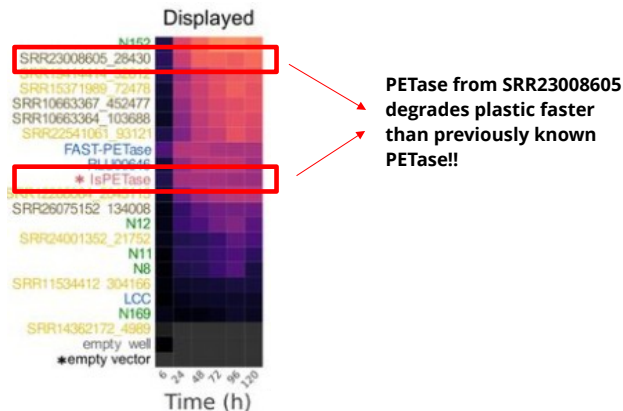
- ▶ Let's align our sequence against all the ~~reads of the SRA~~ contigs of Logan
- ▶ Doable!

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\$*

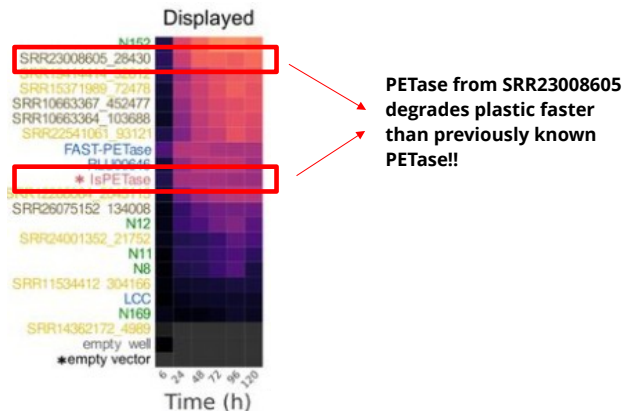


Finding new PETase in the SRA



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

Finding new PETase in the SRA



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
ACCGTAGCCTTAGAATTA

Load

NOTIFICATION

Email

Your email

CONFIGURATION

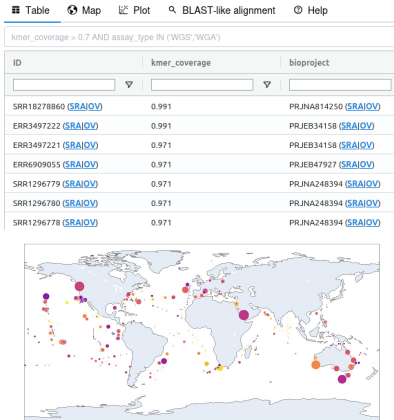
Groups

Threshold = 0.5

.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
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
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
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
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
<input type="text"/> ▼	<input type="text"/> ▼	<input type="text"/> ▼	<input type="text"/> ▼	<input type="text"/> ▼	<input type="text"/> ▼
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
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
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 displays the Galaxy web interface. On the left is a sidebar with navigation icons for Upload, Tools, ChatGPT, Workflows, Workflow Invocations, and Interactive Tools. The main panel is titled 'Tools' and shows a search for 'lexicmap'. Below the search bar, there are two tool descriptions: 'LexicMap Index Builds LexicMap Index' and 'LexicMap Search nucleotide sequence tool for querying genomes'. The 'LexicMap Search' tool is selected, and its configuration page is shown. The page title is 'LexicMap Search nucleotide sequence tool for querying genomes (Galaxy Version 0.0.1+galaxy0)'. Under 'Tool Parameters', the 'LexicMap query file' is set to '3. sniffls_without_bias_nit_1900.fa.gz'. Below this, there is a section for 'LexicMap index source' with a dropdown menu set to 'Locally installed LexicMap indexes'. A message box indicates a parameter error: 'Parameter 'lexicmap_index': an invalid option [None] was selected, please verify LexicMap index file'. Below the message, the 'LexicMap index file' is set to 'Bacteria Genomic'.

- ▶ 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

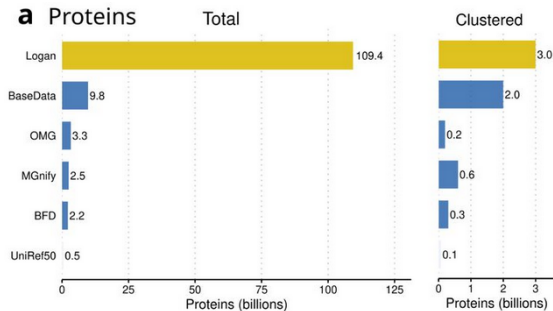
- ▶ Relatively slow
- ▶ Still less sensitive than BLAST or DIAMOND...

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

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

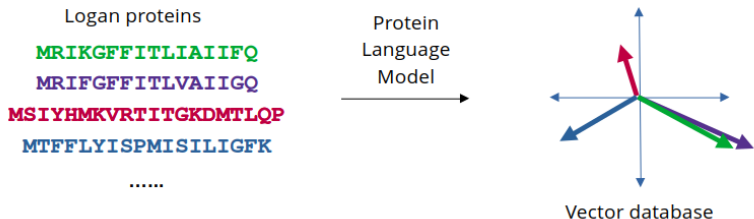
Browsing Logan: strategy 3 - protein search

- ▶ 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!

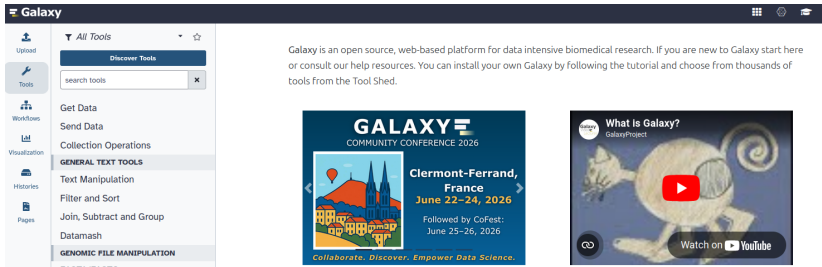
```

Query_10001 150 YHQIDEKEIGIVGYSQGGAGAYNTLEGKDGDK---FKTM-VTVSG--- -----VTESIGEKLHPWVIYDPSKVTI 213
Query_10002 98 QPEVDANRVALIGWALGGGVVVAEAAADDQRVK--AVVTCNAIGDGERS[5]DEQswsRLQD-DMVRDRPERARSGRSRTV 176
Query_10003 95 EK-TGNPRVGVVGYCAGGGLALMLAAQRP-----DAVKAVA PYY-----GLIPWPGAQPDWAAIDA 149
Query_10004 130 SAFVDPERIGVIGFSHGWTLLDFLGPPAIHasATATDARDGLRSVV AVYpycGADVQAGLGKWP-----ADV 198
Query_10005 151 YQKVDTHEHIGISGHSQGGVGVFNAISEQPHSN---LYTCAVSLSP--- -----TQQDLAEALKIP---YDPTKTQI 212
Query_10006 99 DP-RCTGKVGIVGFCMGGGFTLLAPRG-----IFDAAA PNY-----GVLP---RDLSALSSSC 148
Query_10007 99 LEFVDPDRIGVLGVCGGGYSVNAAMTEHRIKavGTVVGANIG----[4]ENNpiqTLEAIGKQRTAEEANGAEPMIINW 174
Query_10008 136 GKFILSNKIAVIGHSMGGYTALALAGGIPTWQeaERVETSSDARVKAI -----VLMAPGAGWFMNSLSKVTI 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 'Tools' section is expanded, showing a search bar and a list of tool categories: Get Data, Send Data, Collection Operations, GENERAL TEXT TOOLS (highlighted), Text Manipulation, Filter and Sort, Join, Subtract and Group, Datamash, and GENOMIC FILE MANIPULATION. The main content area on the right contains a text block stating: '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.' Below this text are two video thumbnails. The first is for the 'GALAXY COMMUNITY CONFERENCE 2026' in Clermont-Ferrand, France, from June 22-24, 2026, followed by CoFest on June 25-26, 2026. The second is a video titled 'What is Galaxy?' from the GalaxyProject channel, featuring a cartoon rabbit and a play button icon.

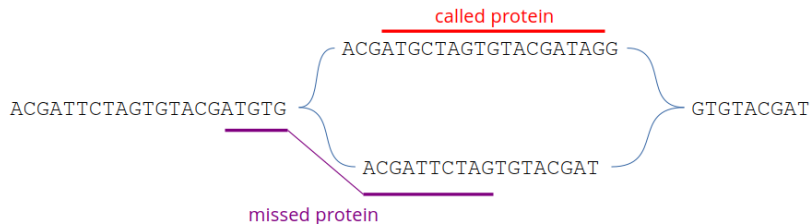
- Maybe also distributed and installable?? (a few TB of data)

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

