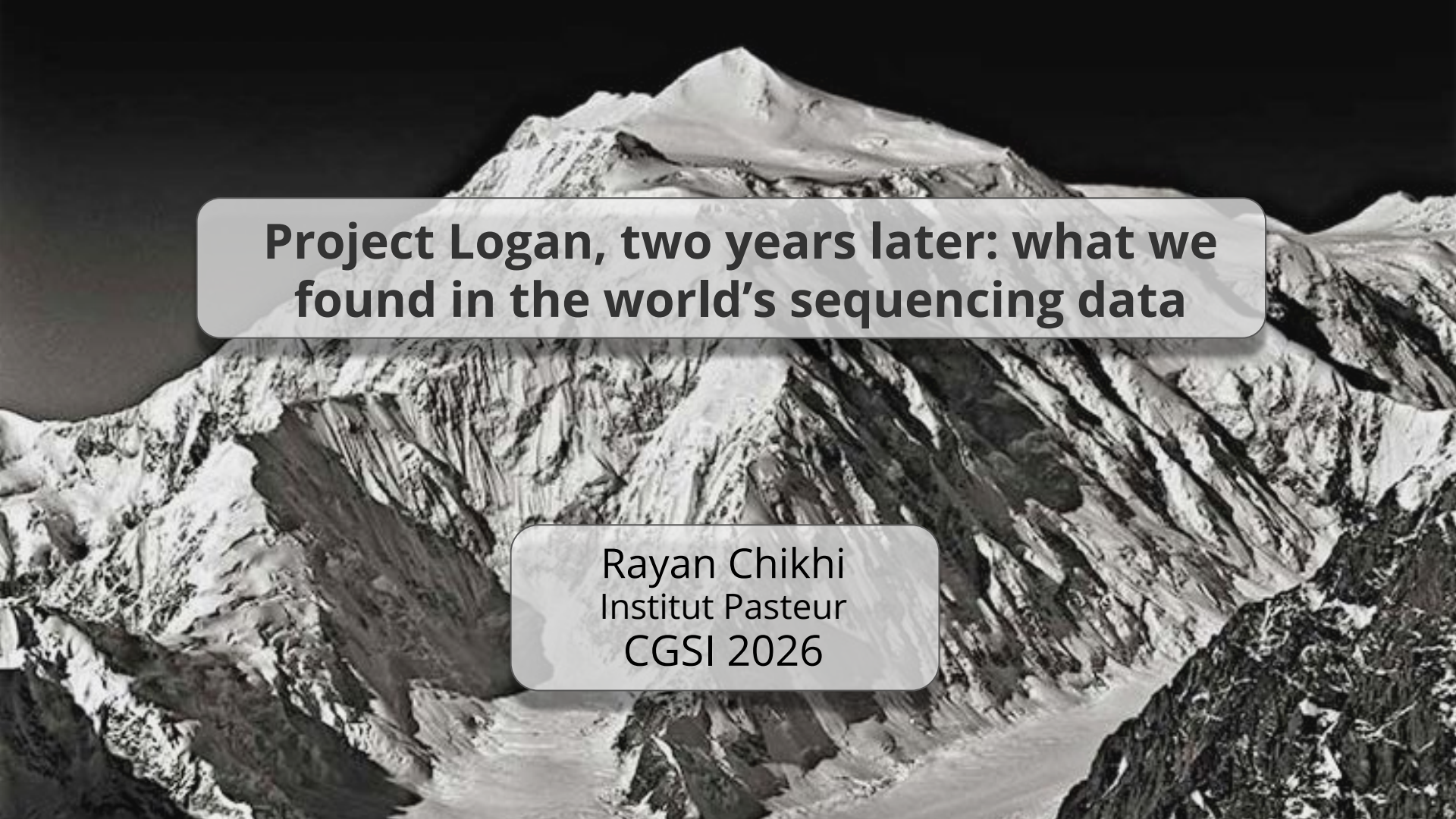


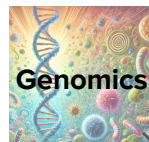


Project Logan, two years later: what we found in the world's sequencing data

Rayan Chikhi
Institut Pasteur
CGSI 2026

Sequence

Bioinformatics



@ INSTITUT
pasteur



AEGIS

PR[AI]RIE

PaRis Artificial Intelligence Research Institute



Serratus & Logan collectives

Artem Babaian, UofT
Robert C. Edgar, Independent
Anton Korobeynikov, Independent
Martin Steinegger, SNU
Caleb Lareau, MSK
Pierre Peterlongo, Inria
Anton Nekrutenko, Penn State

Téo Lemane, Genoscope
Raphael Loll-Krippleber, UofT
Mercè Montoliu-Nerin, Sanger
Antonio Camargo, U Sao Paulo
Paul Medvedev, Penn State
David Koslicki, Penn State
...and many more



5 EXABYTES*

of sequence data

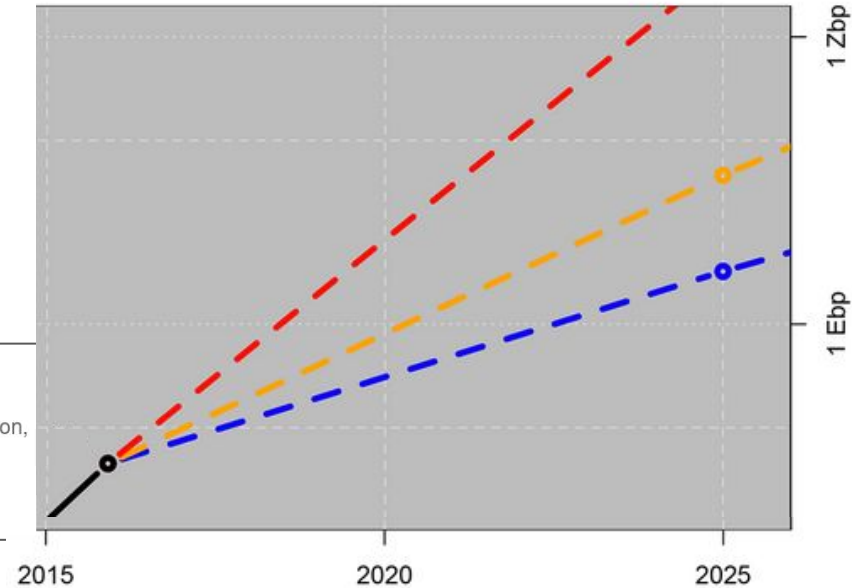
= 5 billion gigabytes

= 50,000,000 sequenced human genomes

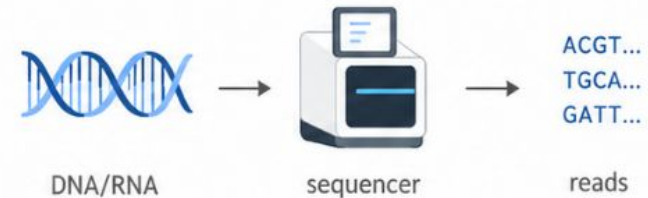
Big Data: Astronomical or Genomical?

Zachary D. Stephens, Skylar Y. Lee, Faraz Faghri, Roy H. Campbell, Chengxiang Zhai, Miles J. Efron, Michael C. Schatz , Saurabh Sinha , Gene E. Robinson 

Published: July 7, 2015 • <https://doi.org/10.1371/journal.pbio.1002195>



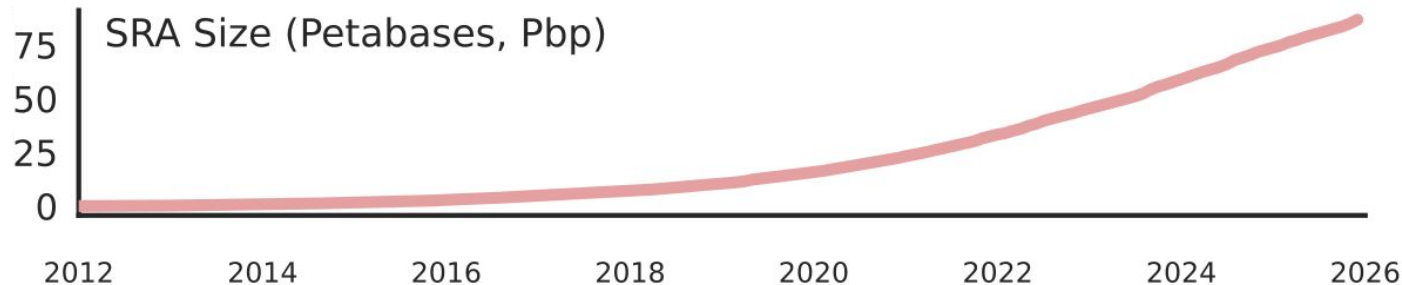
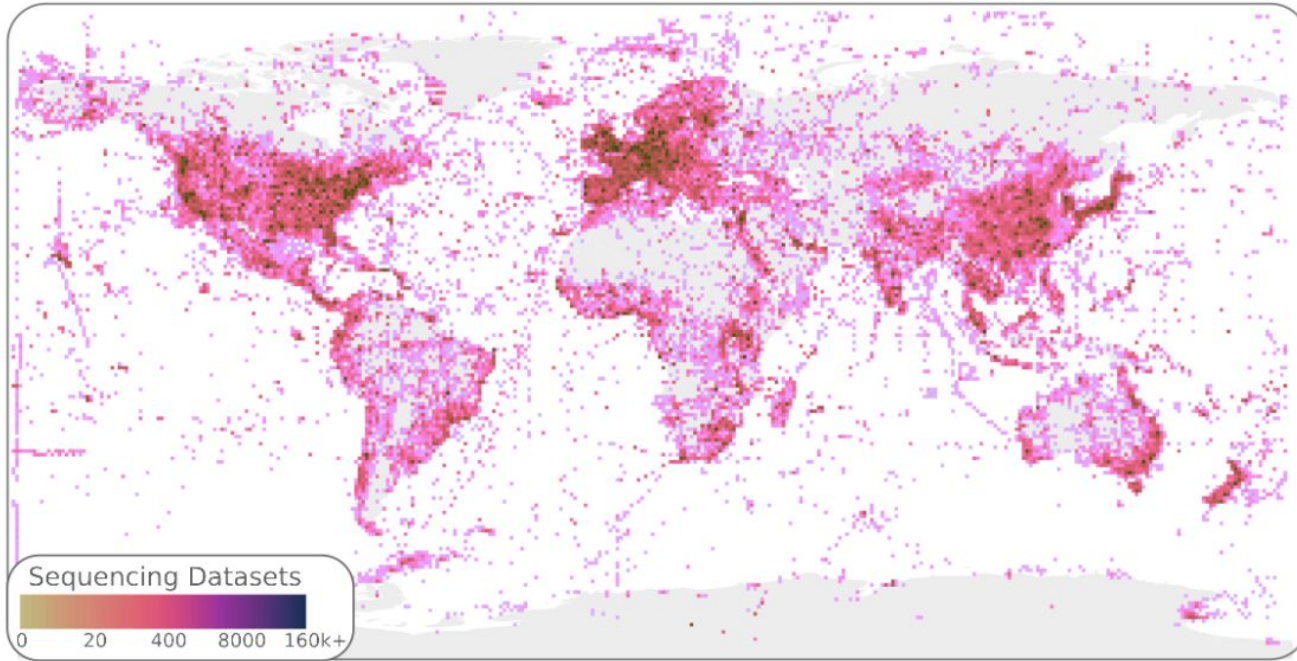
- 2015 projection for 2025: 5 exabytes - 1 zettabyte
- **\$150 billion in direct sequencing costs (!)**
= the cost of building the International Space Station



* Based on # sequencers in the world, <https://tinyurl.com/seqcapacity> from @albertvilella. Slide idea: G. Rizk

Sequence Read Archive @ NCBI

DNA and RNA
sequencing data



~100 petabases

~\$5 billion

in direct seq. costs

Fully open

SRA: Open Science at its best



A. Babaian
(Serratus, Logan)

*"Earth's genetic biodiversity is the **shared heritage of all living organisms**, and as scientists we are **responsible for liberating and sharing this heritage with everyone. Freely, and openly.**"*



About INSDC Global Participation Technical Specifications Announcements Contact Us

INSDC Policy

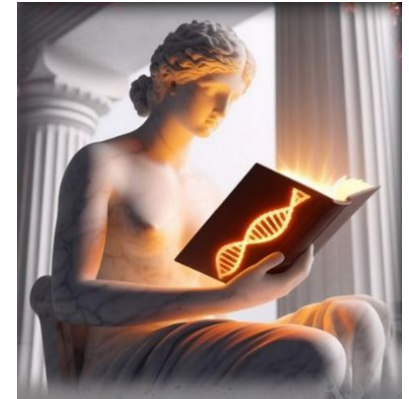
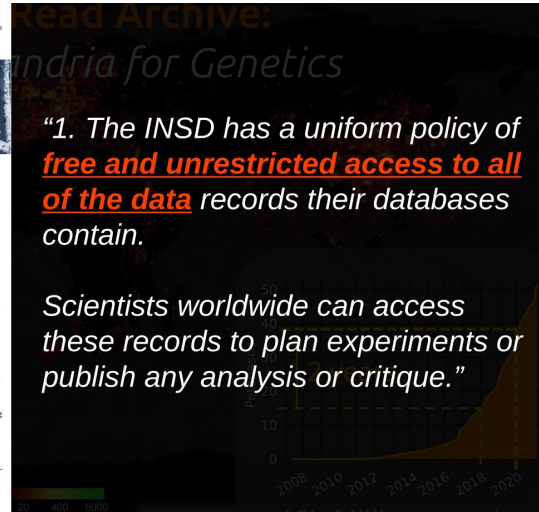
Soren Brunak, Antoine Danchin, Masahira Hattori, Haruki Nakamura, Kazuo Shinozaki, Tara Matisse, Daphne Preuss (2002)

Nucleotide Sequence Database Policies

Science 298 (5597): 1333 15 Nov 2002

1. The INSDC has a uniform policy of free and unrestricted access to all of the data records their databases contain. Scientists worldwide can access these records to plan experiments or publish any analysis or critique. Appropriate credit is given by citing the original submission, following the practices of scientists utilizing published scientific literature.
2. The INSDC will not attach statements to records that restrict access to the data, limit the use of the information in these records, or prohibit certain types of publications based on these records. Specifically, no use restrictions or licensing requirements will be included in any sequence data records, and no restrictions or licensing fees will be placed on the redistribution or use of the database by any party.
3. All database records submitted to the INSDC will remain permanently accessible as part of the scientific record. Corrections of errors and updates of the records by authors are welcome and erroneous records may be removed from the next database release, but all will remain permanently accessible by accession number.
4. Submitters are advised that the information displayed on the Web sites maintained by the INSDC is fully disclosed to the public. It is the responsibility of the submitters to ascertain that they have the right to submit the data.
5. Beyond limited editorial control and some internal integrity checks (for example, proper use of INSDC formats and translation of coding regions specified in CDS entries are verified), the quality and accuracy of the record are the responsibility of the submitting author, not of the database. The databases will work with submitters and users of the database to achieve the best quality resource possible.

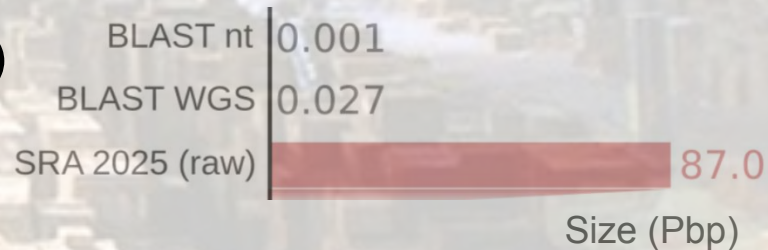
About INSDC Global Participation
Technical Specifications Announcements
Contact Us



"Library of Alexandria"
for genetics

SRA: an unfathomable / unanalyzable pile of data (until recently)

- NCBI **GenBank**, RefSeq:
 - 1000x smaller than SRA (~terabases)
 - searchable by BLAST
- **UK BioBank**: summarizable by VCFs
- **SRA**:
 - no BLAST, no VCF, just sequences and vibes
 - ~5 million human samples, ~7 million metagenomes

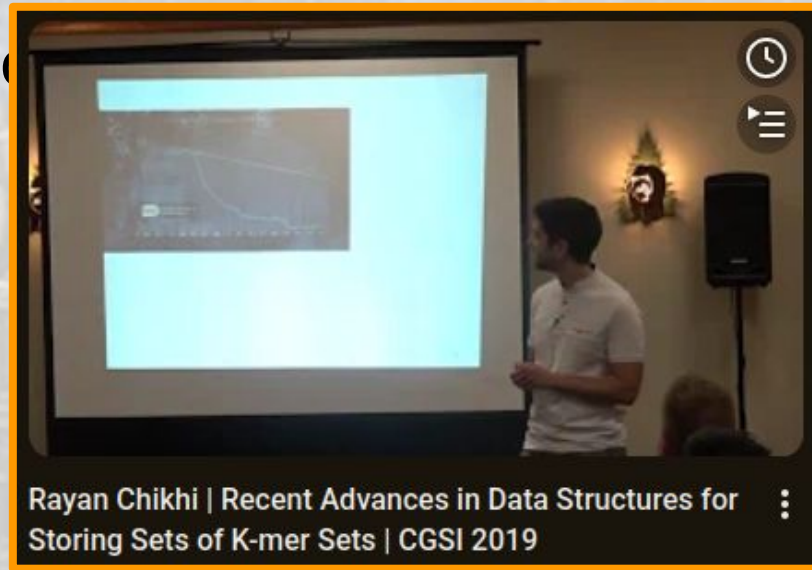


Enter **k-mer** algorithms:

- Represent sequences through their **length-k substrings**
- Better ways to **align** / **assemble** / **index** sequencing data
- Store each k-mer in **O(1)** space; quantify RNA-seq in **minutes**, ..
- Index **100,000's** of **bacterial genomes**, **1,000's** of **metagenomes**
- Success stories: Jellyfish, SPAdes, MEGAHIT, Kraken, MetaGraph, ..

SRA: an unfathomable

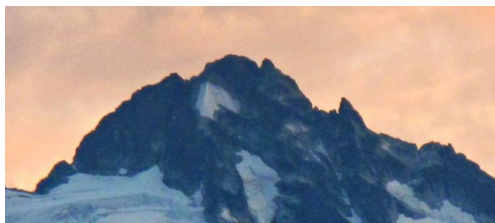
of



Enter **k-mer algorithms**:

- Represent sequences through their **length-k substrings**
- Better ways to **align** / **assemble** / **index** sequencing data
- Store each k-mer in **O(1)** space; quantify RNA-seq in **minutes**, ..
- Index **100,000's** of **bacterial genomes**, **1,000's** of **metagenomes**
- Success stories: Jellyfish, SPAdes, MEGAHIT, Kraken, MetaGraph, ..
- Still fell short of handling **SRA-scale data**

CGSI 2022 talk: **Serratus**: all public RNA-seqs analyzed for viral discovery



Discovered 130,000 new RNA viral species

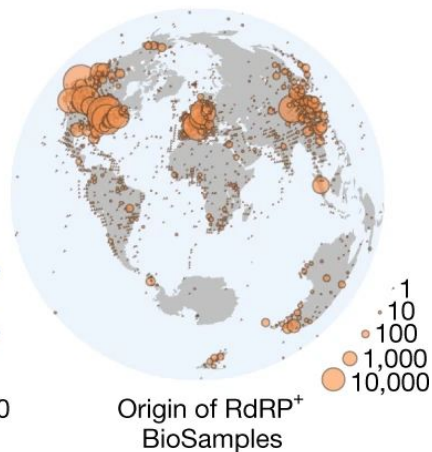
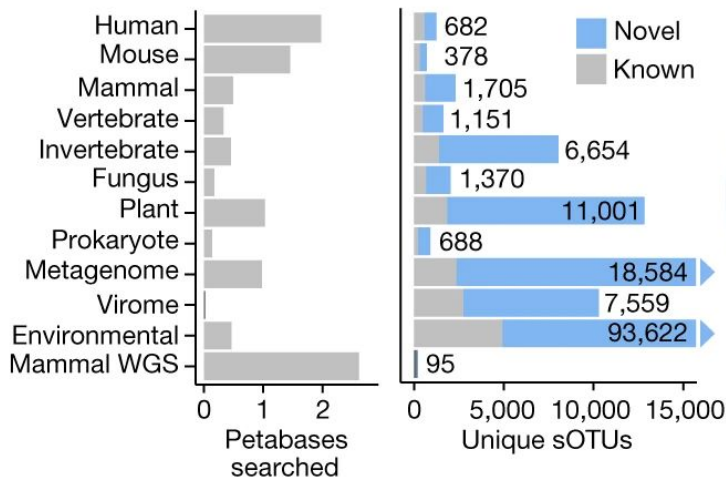
First meta-analysis of the entire SRA (RNA-seqs)

Cloud computation, using **off-the-shelf sequence aligners**

Edgar *et al*, Nature, 2022



Rayan Chikhi | Recent progress
towards petabase-scale genomics



SRA: an unfathomable / unanalyzable pile of data (until recently)

- NCBI GenBank
 - 1000x s
 - searcha
- UK BioBank
- SRA:
 - no BLA
 - ~5 millio

Enter k-mer algo

- Increasingly
- Consider ex
- Stored k-me
- Indexed 100
- Success stories: Jellyfish, SPAdes, MEGAHIT, Kraken, MetaGraph, ..
- Still fell short of handling SRA-scale data



Rich Sutton's ***Bitter lesson*** (2019) on AI

general approaches
with high
computational cost
(e.g. AlphaGo)

>

domain-specific
understanding

(e.g. human-designed strategies)

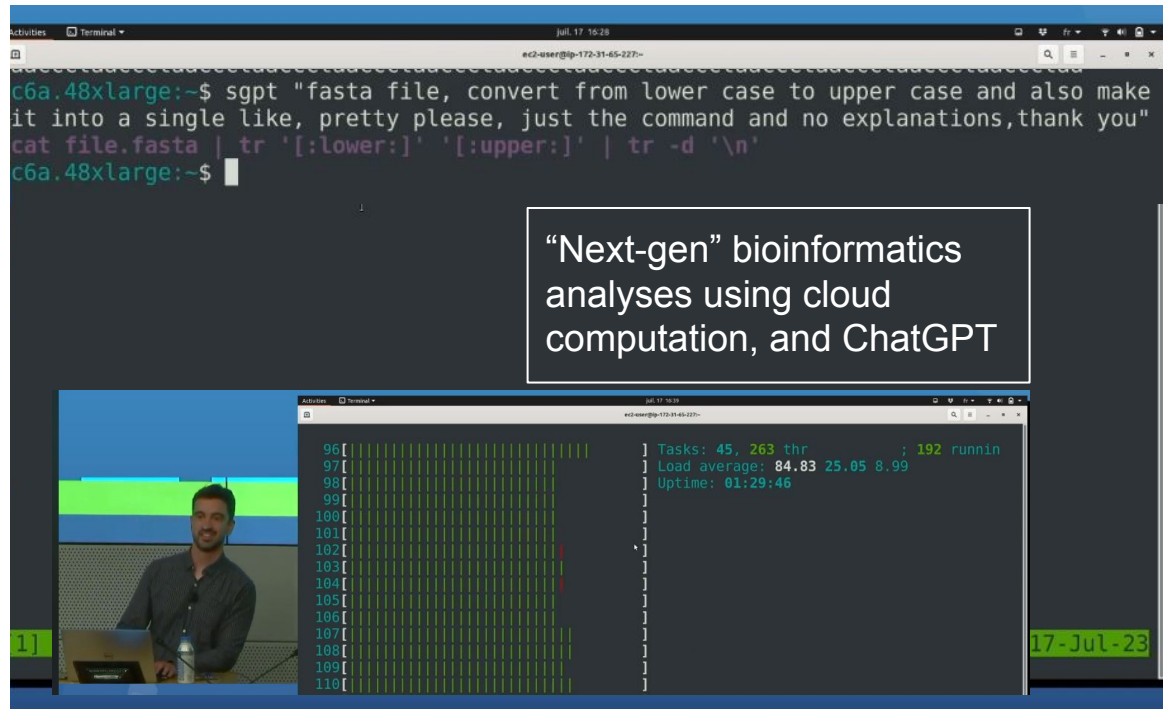
Computational biology

general methods
operating over
peta-scale raw data

>?

specialized tools built
around small,
well-curated datasets

CGSI 2023 talk: Living in the future of genomics



Logan:

The public SRA assembled into longer, searchable, compressed sequences



The screenshot shows a video player interface. The main content is a presentation slide with a mountain background. The slide text reads: "Project Logan: Assembling all public sequencing data", "Rayan Chikhi", "Institut Pasteur", and "CGSI 2024". On the left side of the video player, there is a small inset video of a man, Rayan Chikhi, speaking. The video player has a blue header bar with a clock icon and a menu icon. The bottom of the video player shows a progress bar and the time "35:14". Below the video player, there is a black bar with white text: "Rayan Chikhi | Project Logan Assembling all public sequencing data | CGSI 2024".

Project Logan:
Assembling all public
sequencing data

Rayan Chikhi
Institut Pasteur
CGSI 2024

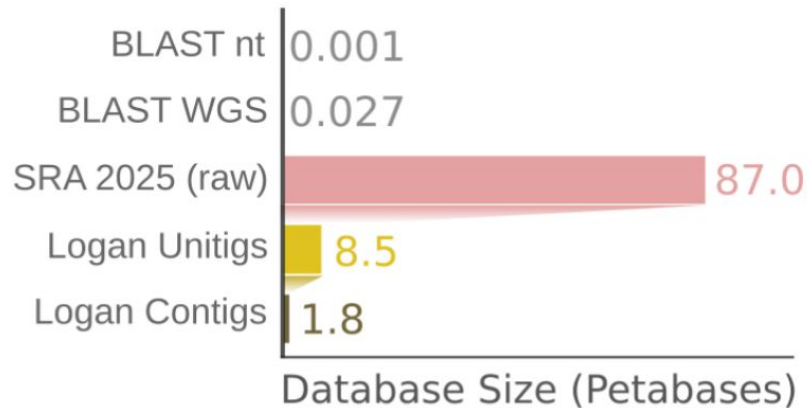
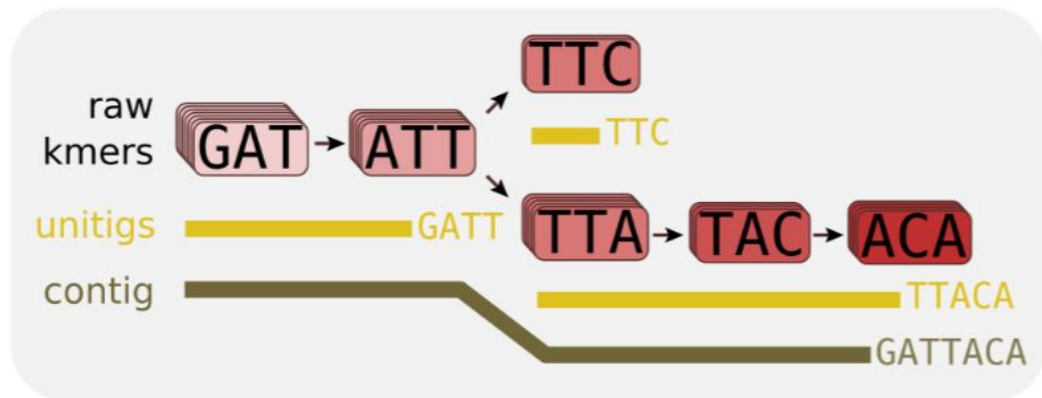
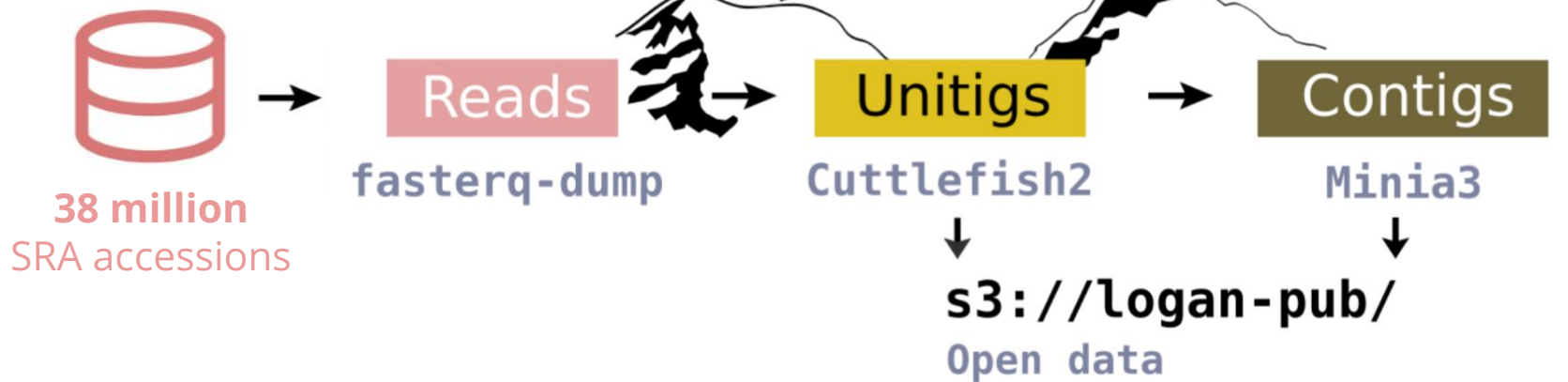
Rayan Chikhi

CGSI

35:14

Rayan Chikhi | Project Logan Assembling all public
sequencing data | CGSI 2024

Logan:



AutoScaling Groups In-Service Capacity

vCPUs

2.00M

1.50M

1.00M

500k

0

05:00 06:00 07:00 08:00 09:00 10:00 11:00 12:00 13:00 14:00 15:00

6 Days: 2.18 Million concurrent CPUs

~100 tonnes CO₂, ~half a one-way transatlantic flight

A resource comparable to Common Crawl

Common Crawl
maintains a **free, open**
repository of web crawl
data that can be used by
anyone.

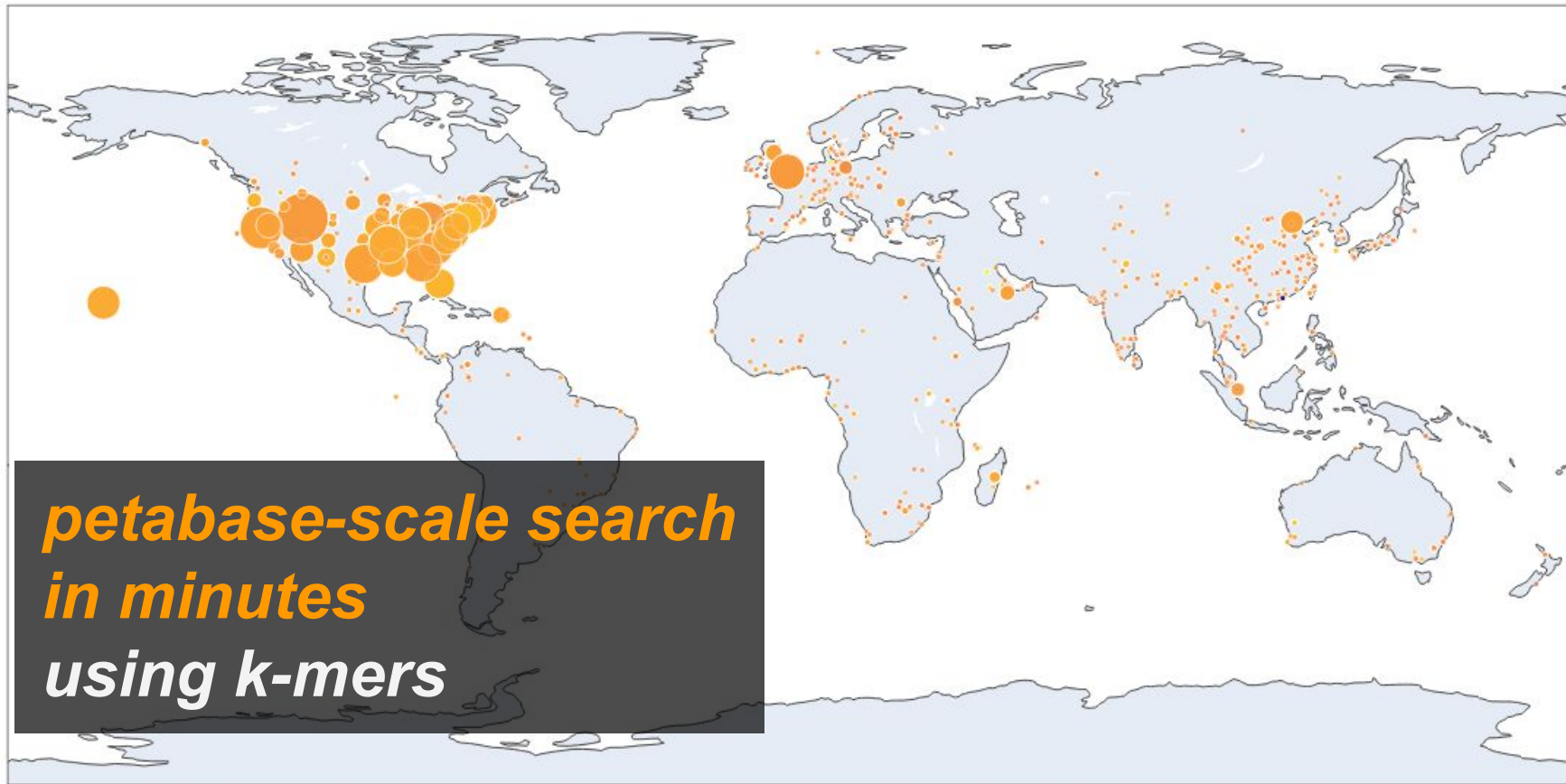
HTML pages, images, PDFs,
JavaScript, etc..

Full historical archive:
>10 petabytes



Logan-Search.org

Query:
***Klebsiella* AMR**



How Logan-Search works

- 1) Paste a query DNA sequence,
e.g. ACTAGTAGAGTAGATGAGGACATAGATGAGCACATGATGAGAGACACATTG
- 2) Search **23 million SRA accessions** in minutes
- 3) Obtain:
 - matching SRA accessions and metadata
 - geographic / host / biome context
 - per-accession BLAST-like alignments

Built by **Teo Lemane** and
Pierre Peterlongo, INRIA Rennes



Logan Search

INPUT

text

file

session

Query sequence(s) *
Fasta/Fastq format

>Query
ACCGTAGCCTTAGAATTA

Load

NOTIFICATION

Email

Your email

CONFIGURATION

Groups

Threshold = 0.5

0.251.0

Submit

Reset

Logan-Search results interface

ID	kmer_coverage	ANI_estimat...	p_value	e_value	bioproject	biosample	bioproject_title	bioproject_description
GCA_918797485.2 (GENOME)	1	1	0.000000e+0	0.000000e+0	PRJEB47927 (SRA OV)	SAMEA10271030 (SRA OV)	Chromosome-scale genome as...	Pelagophytes (Stramenopile)
SRR1197259 (SRA OV)	1	1	0.000000e+0	0.000000e+0	PRJNA239089 (SRA OV)	SAMN01985059 (SRA OV)	Pelagomonas calceolata Geno...	Pelagomonas calceolata gen
SRR851684 (SRA OV)	1	1	0.000000e+0	0.000000e+0	PRJNA193556 (SRA OV)	SAMN02144620 (SRA OV)	Pelagomonas calceolata strain...	Pelagomonas calceolata tran
SRR851686 (SRA OV)	1	1	0.000000e+0	0.000000e+0	PRJNA193556 (SRA OV)	SAMN02144622 (SRA OV)	Pelagomonas calceolata strain...	Pelagomonas calceolata tran
ERR1110185 (SRA OV)	1	1	0.000000e+0	0.000000e+0	PRJEB11660 (SRA OV)	SAMEA3646037 (SRA OV)	Marine Microbes pilot study	Pilot study for the Australian
SRR5918662 (SRA OV)	1	1	0.000000e+0	0.000000e+0	PRJNA385736 (SRA OV)	SAMN07483521 (SRA OV)	Marine metagenomes Metage...	Marine amplicons from Austr
ERR3497222 (SRA OV)	1	1	0.000000e+0	0.000000e+0	PRJEB34158 (SRA OV)	SAMEA5899549 (SRA OV)	Collection of Marine Eukaryot...	This project is part of the Ma
ERR9764111 (SRA OV)	1	1	0.000000e+0	0.000000e+0	PRJEB47927 (SRA OV)	SAMEA14430949 (SRA OV)	Chromosome-scale genome as...	Pelagophytes (Stramenopile)
ERR6909055 (SRA OV)	1	1	0.000000e+0	0.000000e+0	PRJEB47927 (SRA OV)	SAMEA10271030 (SRA OV)	Chromosome-scale genome as...	Pelagophytes (Stramenopile)
ERR3497221 (SRA OV)	1	1	0.000000e+0	0.000000e+0	PRJEB34158 (SRA OV)	SAMEA5899549 (SRA OV)	Collection of Marine Eukaryot...	This project is part of the Ma
SRR18278860 (SRA OV)	1	1	0.000000e+0	0.000000e+0	PRJNA814250 (SRA OV)	SAMN26541151 (SRA OV)	Metagenomic time-series the ...	This study examines monthly
SRR1296778 (SRA OV)	1	1	0.000000e+0	0.000000e+0	PRJNA248394 (SRA OV)	SAMN02740026 (SRA OV)	Marine Microbial Eukaryote Tr...	The Marine Microbial Eukary
SRR14100031 (SRA OV)	1	1	0.000000e+0	0.000000e+0	PRJNA517804 (SRA OV)	SAMN13381556 (SRA OV)	100 Algal genome project (AL...	100 Algal genome project (A
SRR1296779 (SRA OV)	1	1	0.000000e+0	0.000000e+0	PRJNA248394 (SRA OV)	SAMN02740027 (SRA OV)	Marine Microbial Eukaryote Tr...	The Marine Microbial Eukary
SRR1296780 (SRA OV)	1	1	0.000000e+0	0.000000e+0	PRJNA248394 (SRA OV)	SAMN02740028 (SRA OV)	Marine Microbial Eukaryote Tr...	The Marine Microbial Eukary
Page Size: 100 1 to 100 of 1,939 < > Page 1 of 20 < >								

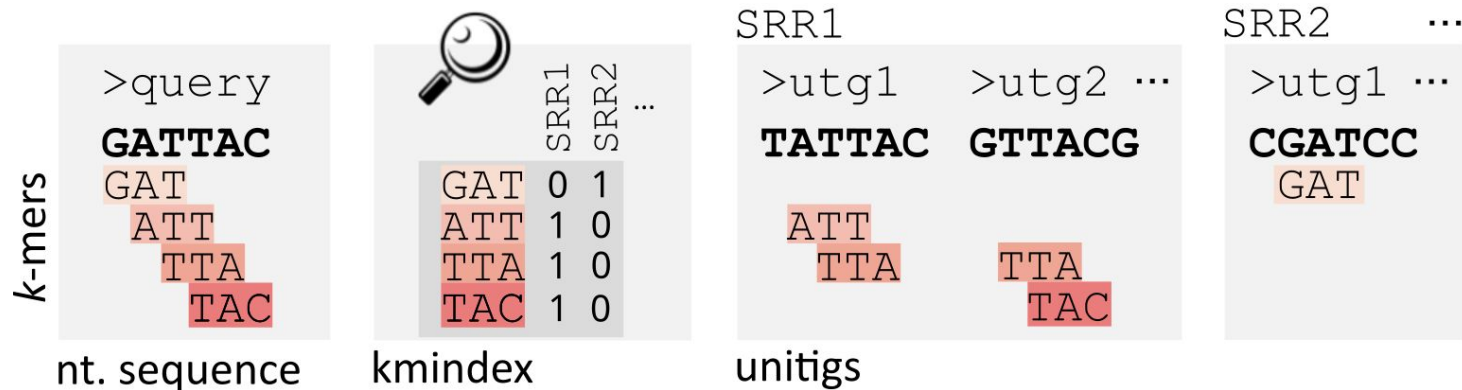
 Remove filters

About your query (AI generated)

The query sequence seems to be linked primarily to *Pelagomonas calceolata*, a type of microalgae categorized within the pelagophytes (Stramenopiles). Pelagophytes are diverse microalgae, known for their role in coastal species causative of harmful algal blooms, and are also abundant in oceanic ecosystems. This organism appears to be involved in various

Under the hood of Logan-Search

Uses a **Bloom filter matrix of k-mers** Lemane et al 2024



Stored on disk, **row-wise**, for **fast retrieval** per k-mer

1.1 petabytes for the Bloom Filters

2.15 quadrillion k-mers inserted

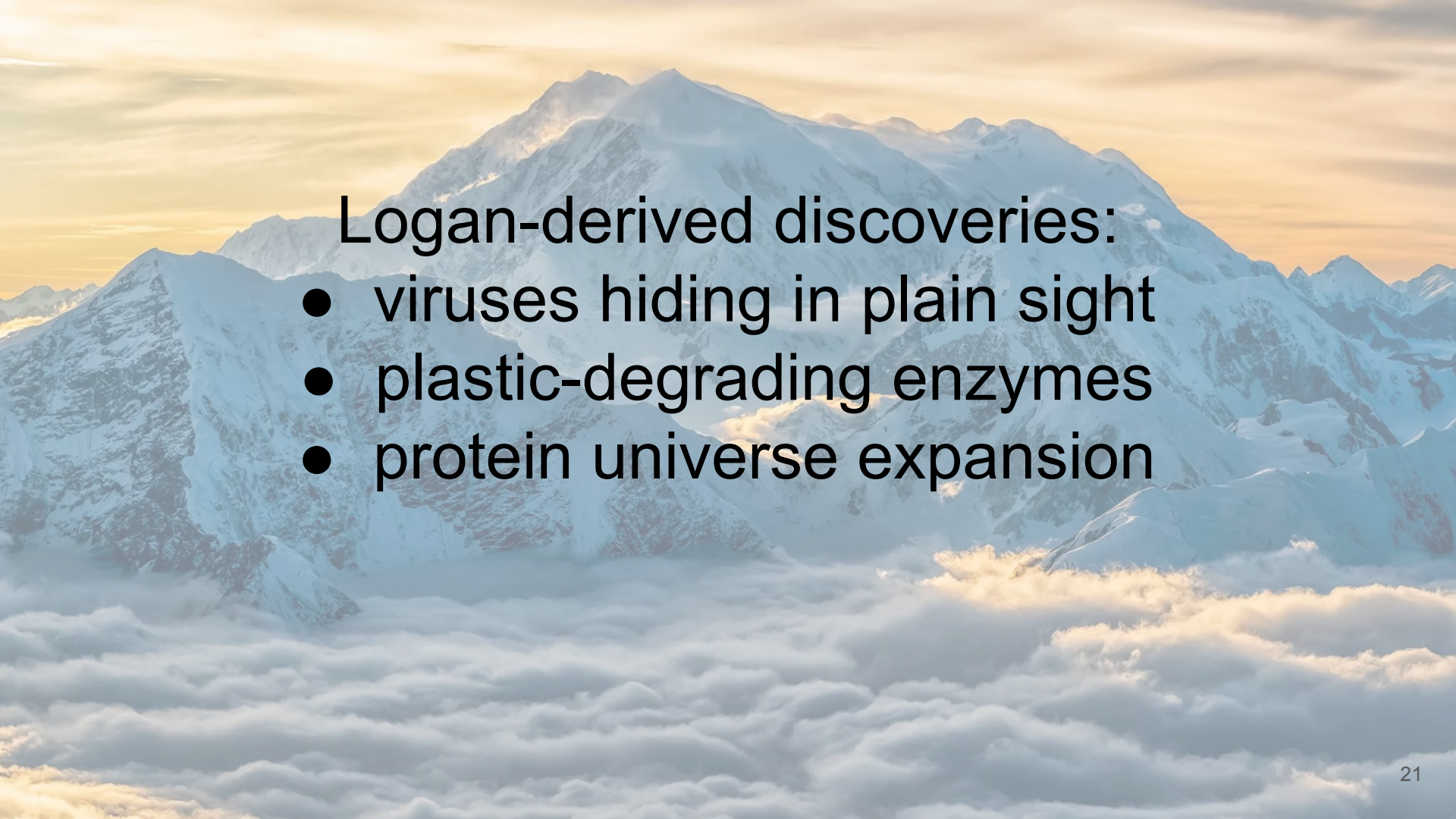
Many algorithmic challenges remain

At Logan scale, we still cannot efficiently:

- **Compress** Logan unitigs (Rouzé et al 2025, scales up to 1000s of human genomes)
- Increase the **contiguity** of Logan assemblies (need to redownload the reads)
- **Index** Logan without spending exorbitant time or space

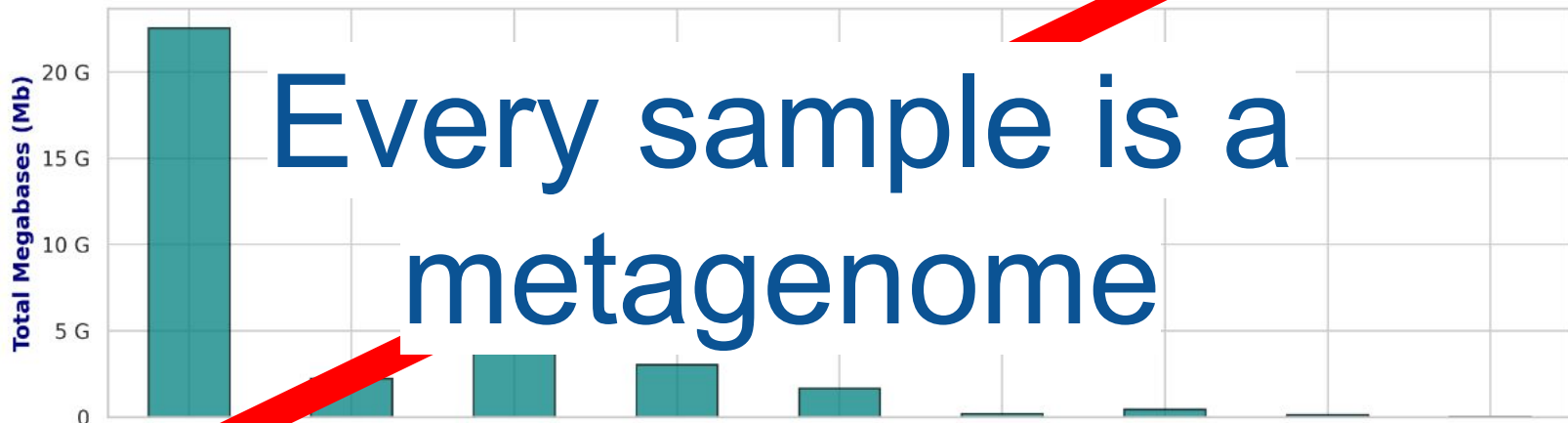
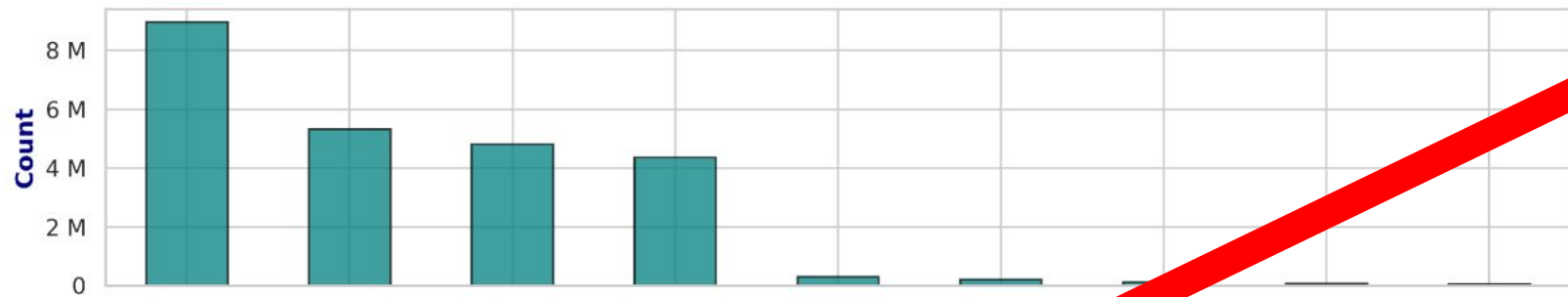
But.. these are pré-Logan problems:

- **GenBank WGS** (10x larger than RefSeq) is also **unindexed, uncompressed**
- (WGS is nearly impossible to download: millions of files on a FTP server)
- Why not start here?



Logan-derived discoveries:

- viruses hiding in plain sight
- plastic-degrading enzymes
- protein universe expansion



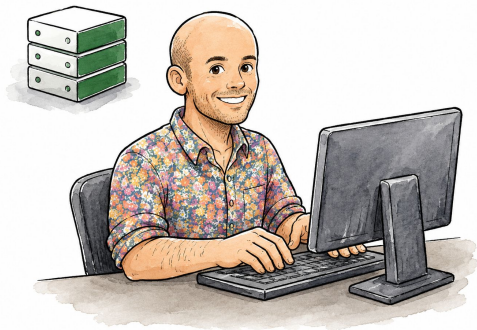
Every sample is a
metagenome

Petabase-scale search of human viruses

using Serratus

Caleb Lareau et al., Nature 2023

Sloan-Kettering Cancer Center



129 human viruses screened
across all public RNA-seqs

nature

Explore content ▾

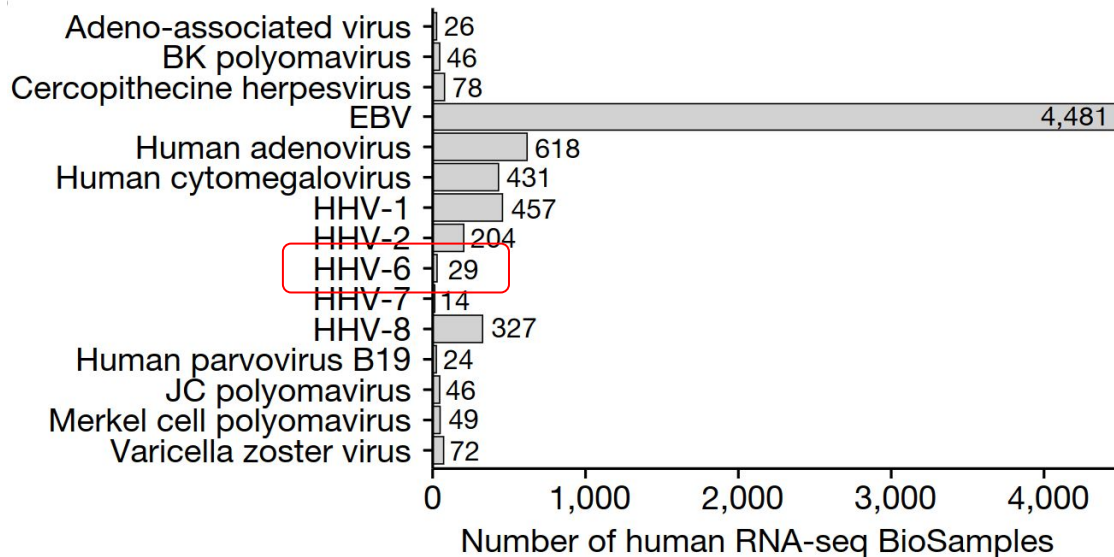
About the journal ▾

Publish with us ▾

[nature](#) > [articles](#) > article

Article | Published: 08 November 2023

Latent human herpesvirus 6 is reactivated in CAR T cells



From petabase-scale search to HHV-6 detected in CAR T cells

Discovery made using Serratus
Caleb Lareau et al., Nature 2023

CAR-T datasets

No. of cells	No. of reads	No. of HHV-6B UMIs	No. of HHV-6 ⁺ cells (No. of HHV-6 super-expressors)
137,326	8,972,716,245	0	0 (0)
602,577	15,403,016,654	919	7 (4)
798,618	66,076,393,465	374	6 (3)
31,953	3,086,842,082	789	15 (6)
599,926	15,680,467,351	0	0 (0)
1,449,385	53,872,337,003	6	0 (0)

HHV-6 expression in large-scale scRNA datasets

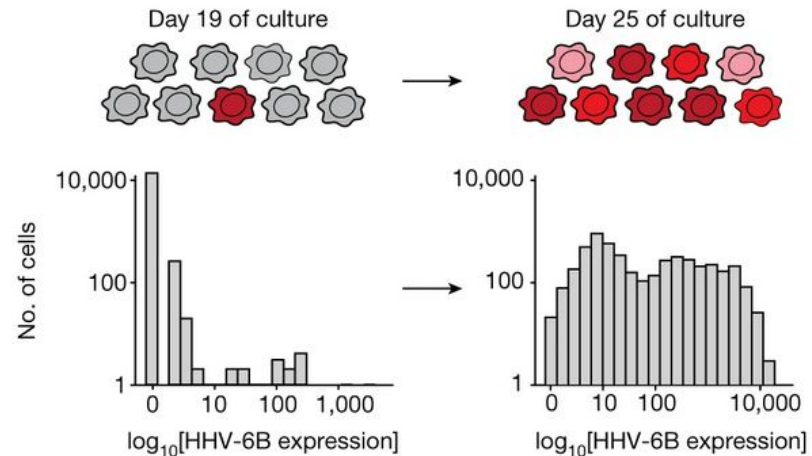
nature

Explore content ▾ About the journal ▾ Publish with us ▾

[nature](#) > [articles](#) > article

Article | Published: 08 November 2023

Latent human herpesvirus 6 is reactivated in CAR T cells



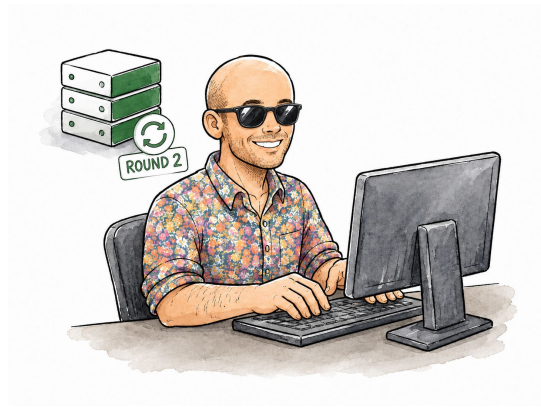
A small number of cells are super-expressors of HHV-6

Logan-Search extends HHV-6 discovery to Tumor-Infiltrating Lymphocyte therapy

Discovery made using Logan-search
Caleb Lareau, Sloan-Kettering Cancer Center

Presented in the Logan preprint

1 Logan: Planetary-Scale Genome Assembly Surveys Life's Diversity
2 Rayan Chikhi¹, Téo Lemane², Raphaël Loll-Krippelbein^{3,4}, Mercè Montoliu-Nerin⁵, Brice
3 Raffestin¹, Antonio Pedro Camargo^{6,7}, Carson J. Miller⁸, Mateus Bernabe Fiamenghi⁷, Daniel
4 Paiva Agostinho⁹, Sina Majidian¹⁰, Greg Autric¹¹, Maxime Hugues¹², Junkyoung Lee¹³, Roland
5 Faure¹, Kristen D. Curry¹, Jorge A. Moura de Sousa¹, Eduardo P. C. Rocha¹, David Koslicki¹⁴,
6 Paul Medvedev¹⁴, Purav Gupta^{4,15}, Jessica Shen^{4,15}, Alejandro Morales-Tapia^{4,15}, Kate Sihuta^{3,4},
7 Peter J. Roy^{3,4}, Grant W. Brown^{3,4}, Robert C. Edgar¹⁶, Anton Korobeynikov¹⁶, Martin
8 Steinegger¹², Caleb A. Lareau¹⁷, Pierre Peterlongo¹⁸, and Artem Babaian^{4,15}

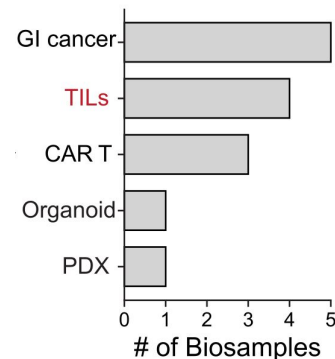


HHV-6 searched again in human RNA-seqs,
this time with Logan



Logan-Search

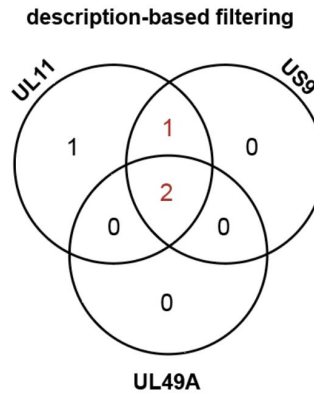
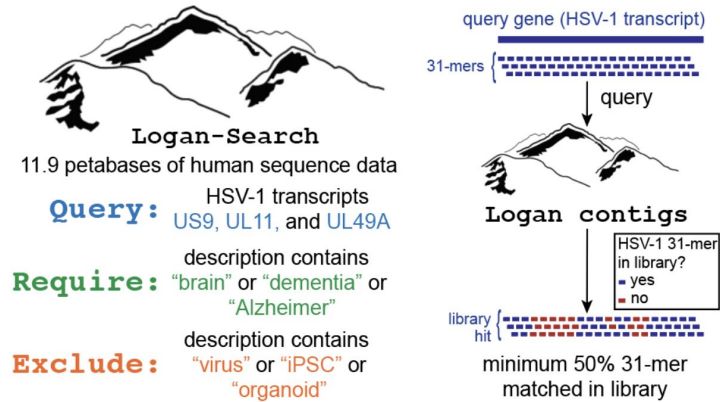
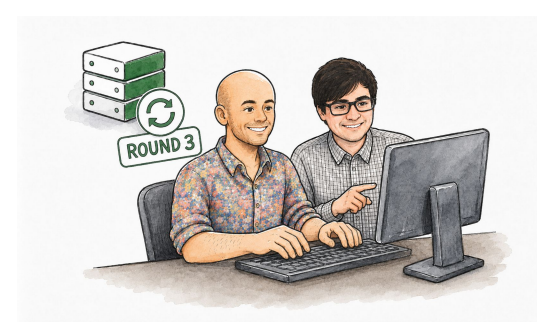
Query: HHV-6B transcripts
U83 and U91
Filter: assay_type == RNA-Seq
organism == Homo sapiens



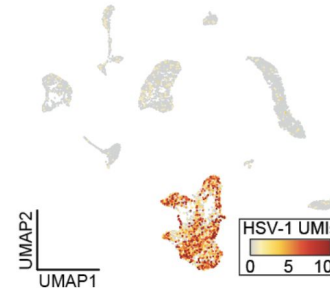
HHV-6 detected in TIL infusion products

Could hidden HSV-1 reactivation contribute to neurodegeneration?

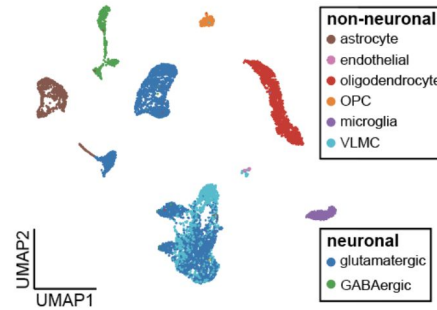
Recent preprint (Lareau lab, May 1st 2026)



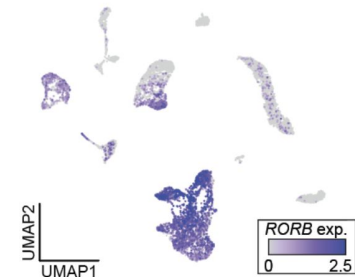
viral gene expression



cell type annotation



RORB expression



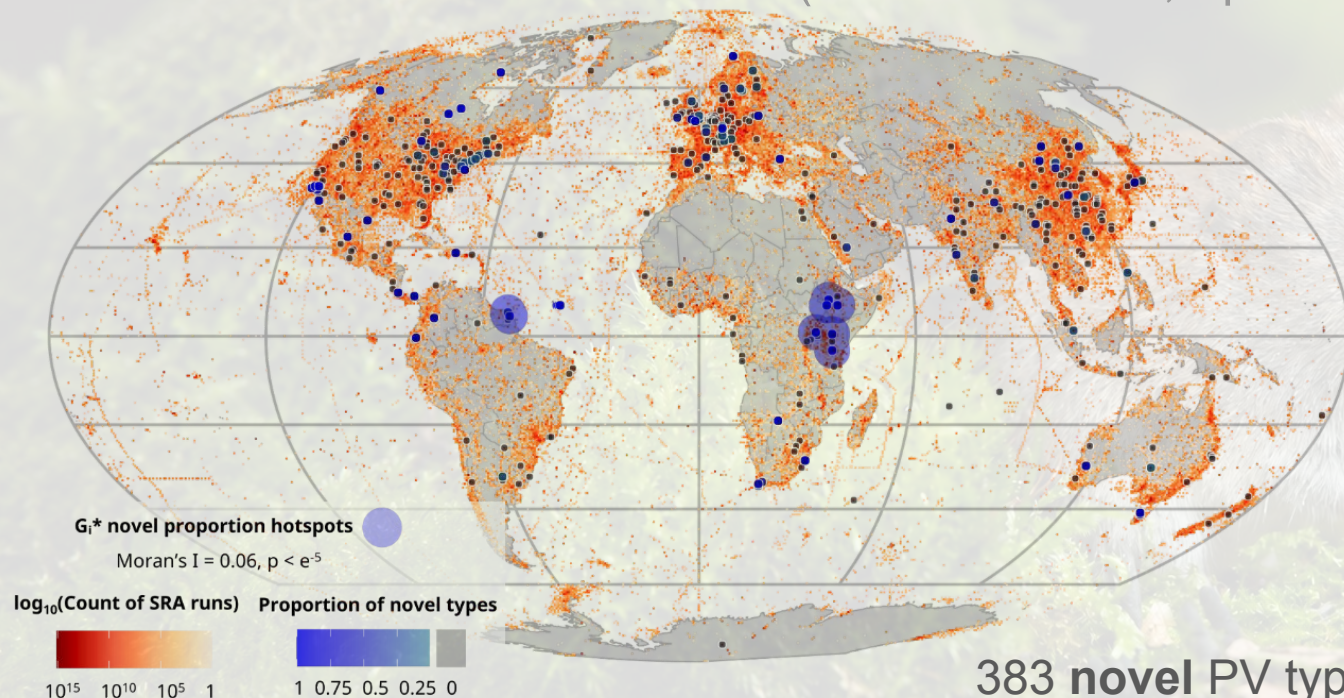
n=1 HSV-1 expression event localized to RORB+ human neurons

Missed by host-only alignment: HSV-1 is absent from the human reference

Gutierrez et al., bioRxiv 2026

Shout out to: SRA-scale papillomavirus discovery

(Jess Shen *et al*, April 2026 preprint)

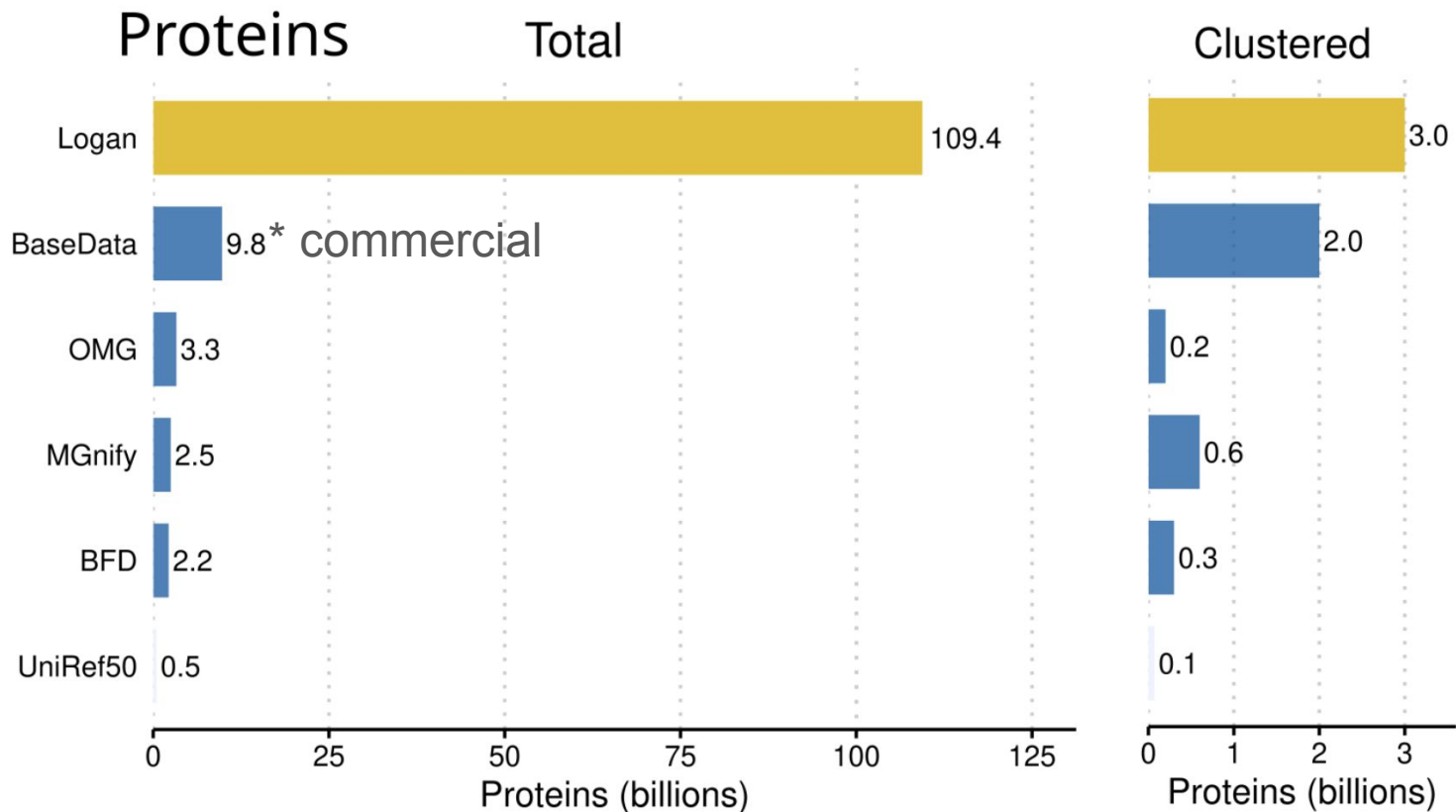


383 novel PV types
(992 previously known)

New hosts: rhinoceros, voles, grey foxes



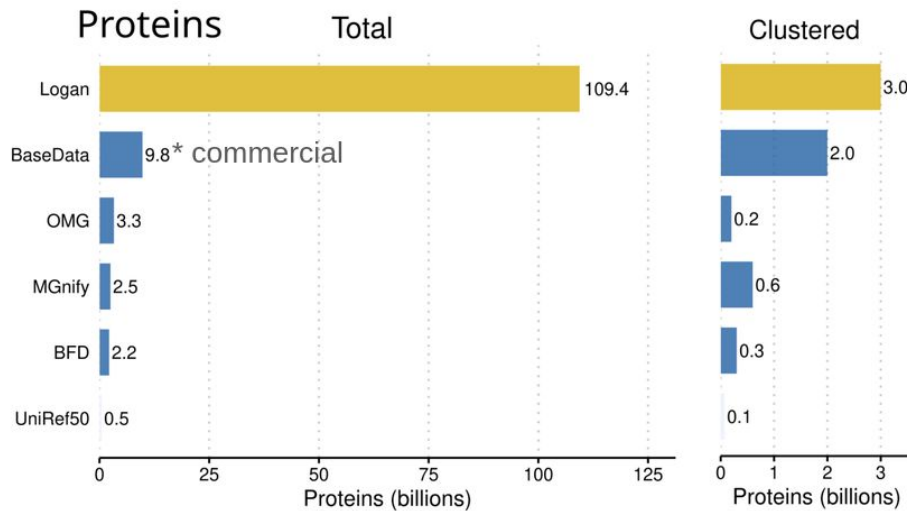
Expanding the Protein Universe



Expanding the Protein Universe

- Extracted ORFs from Logan v1.1 contigs with Prodigal
- MMseqs2 clusters
 - 90% identity
 - **50% identity**
- **71M** 'human-associated'
- **3.0B** 'nonhuman-associated'

Soon searchable via embeddings
(deep homology search) in **Galaxy**



Analyses by:

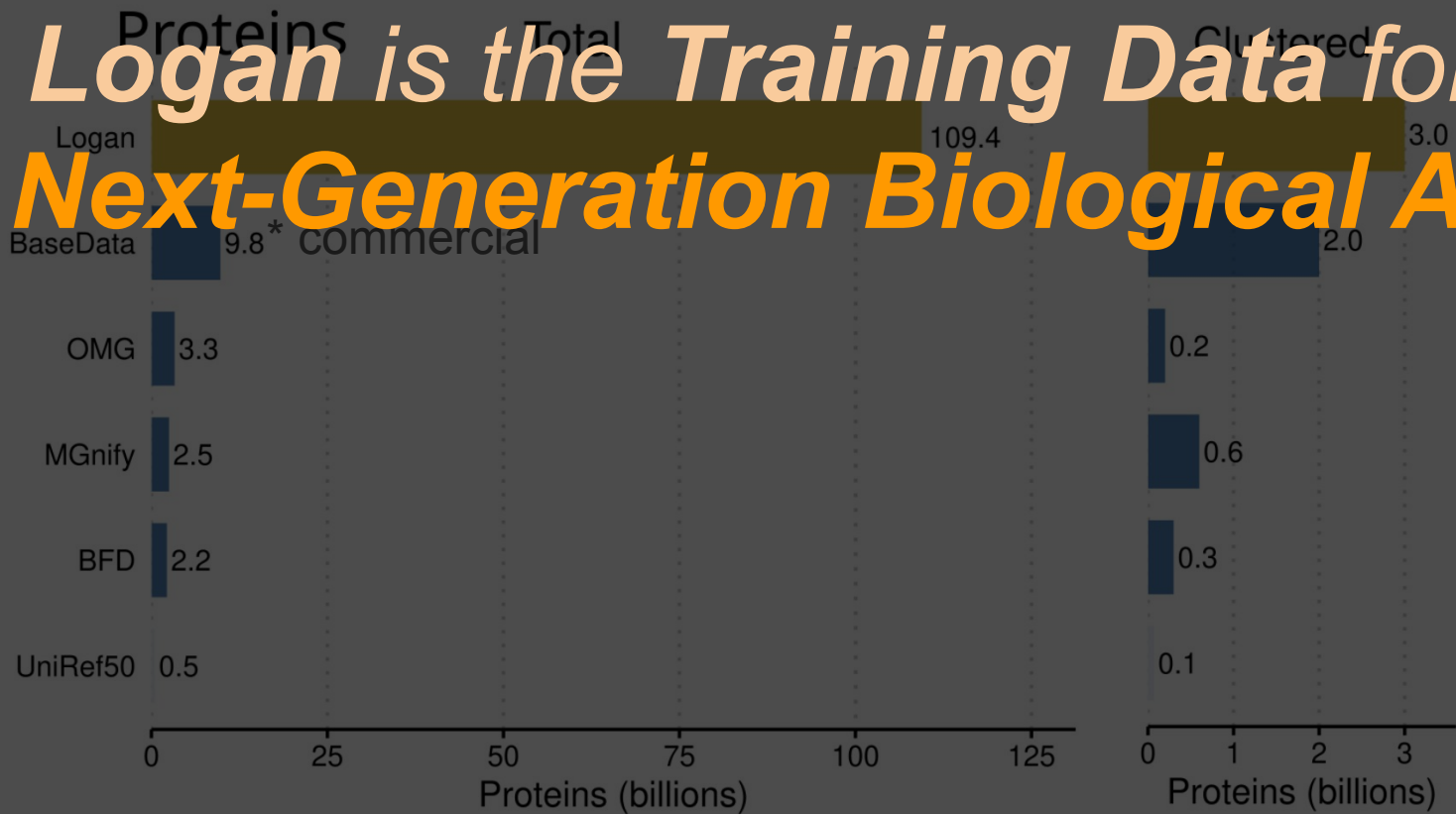
Martin Steinegger, SNU, Korea

Joey Lee, SNU, Korea

Roland Faure, Institut Pasteur

Expanding the Protein Universe

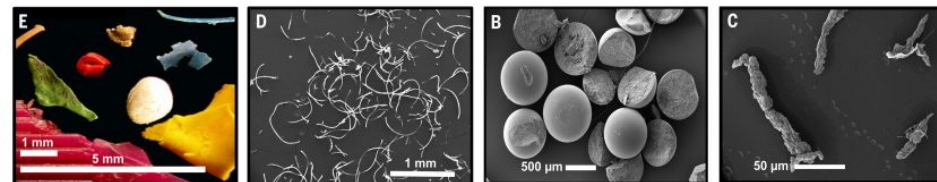
*Logan is the Training Data for
Next-Generation Biological AI*



Microplastics are infiltrating brain tissue, studies show: 'There's nowhere left untouched'



0.5% of Human Brains are microplastic by weight



MP < 5mm

Microplastics in the human body

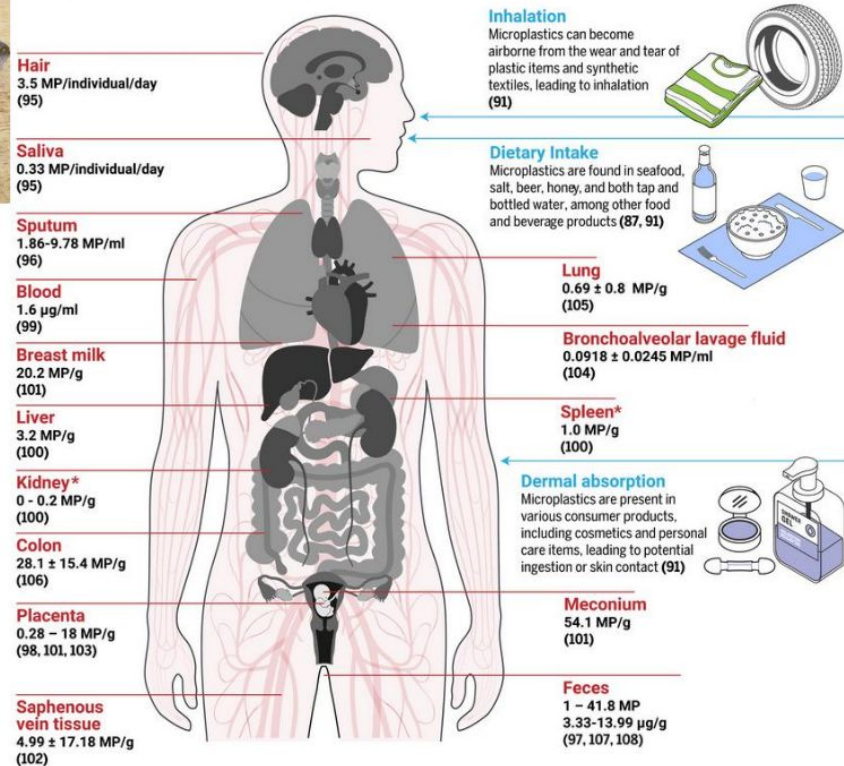
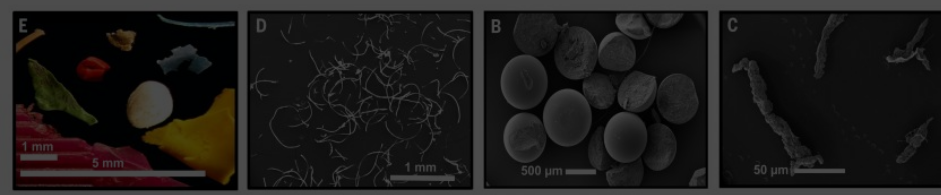


Fig. 4. Locations in the human body where microplastics have been reported.

Microplastics are infiltrating brain tissue, studies show: 'There's nowhere left untouched'

Logan Protein Search: 213 Plastic-degrading enzymes 1 Million in NCBI 50+ Million in Logan

0.5% of Human Brains are
microplastic by weight



MP < 5mm

Microplastics in the human body

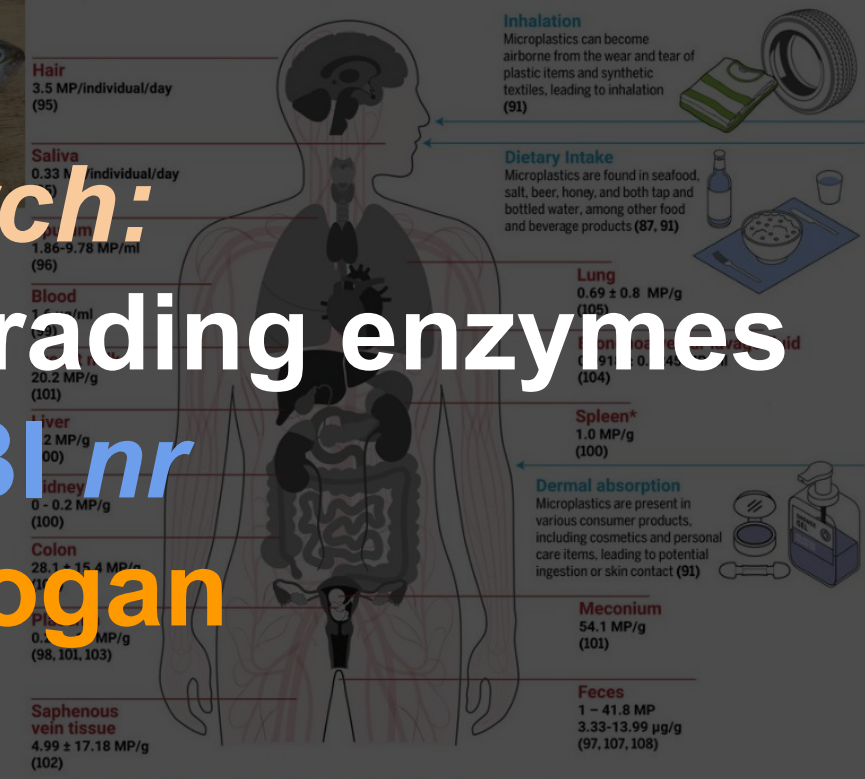


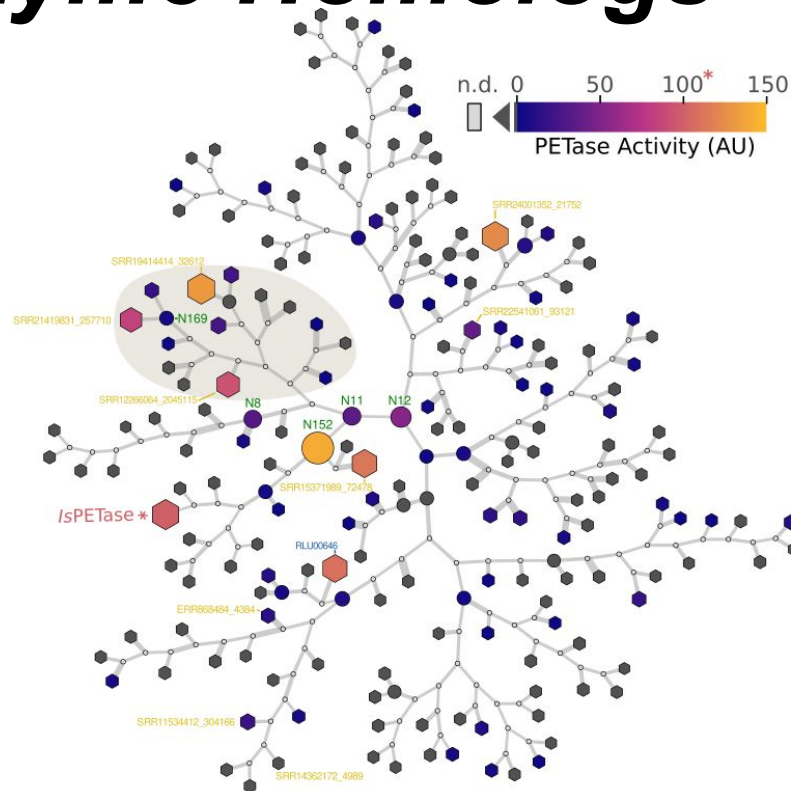
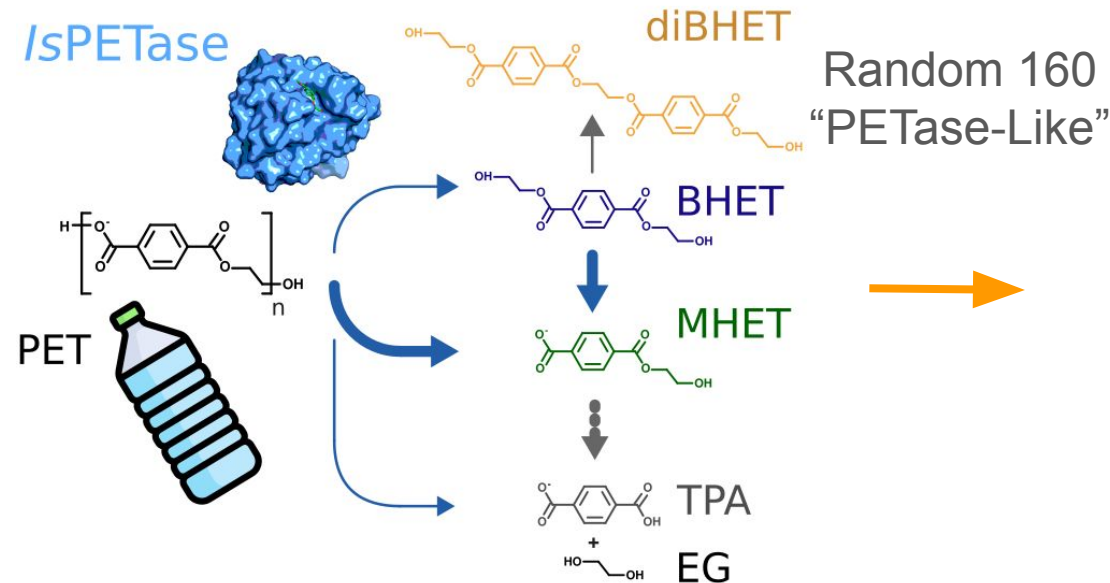
Fig. 4. Locations in the human body where microplastics have been reported.

Discovering +50M

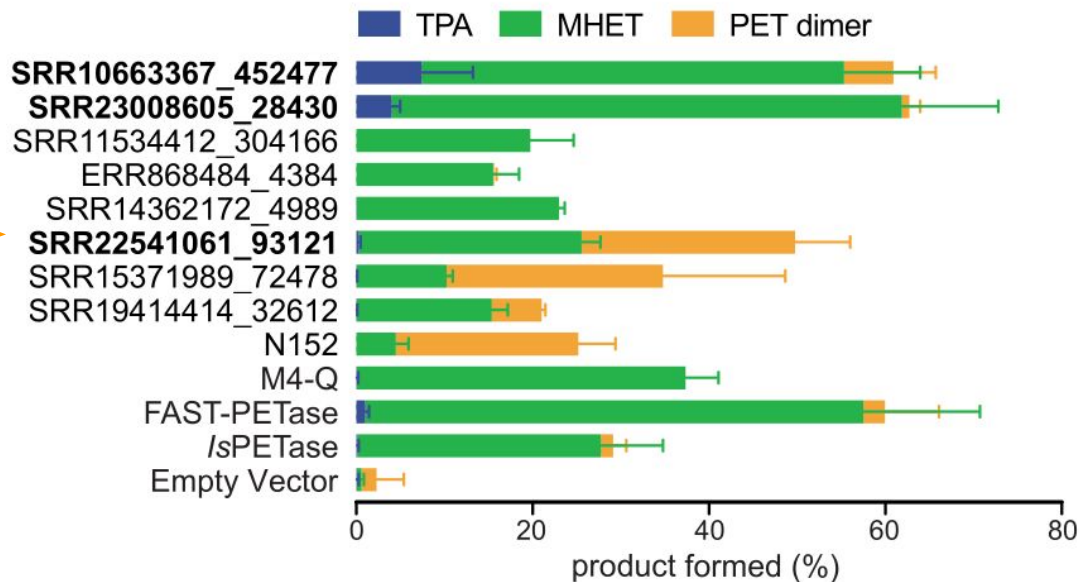
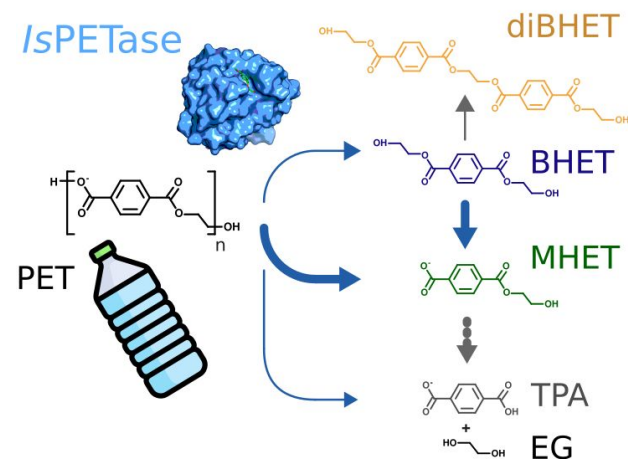
A Babaian's group
R Loll-Krippelber

Plastic-Degrading Enzyme Homologs

*Is*PETase



Found Natural and Ultra-Efficient Plastic-Degrading Enzymes



The MYSTERY *of the* OBELISKS

Uncovering hidden biological entities with **Logan**



Originally discovered by **Vanya Zheludev** (Stanford)
in 100-human microbiome project:

→ **1 species**

NCBI Pebblescout k-mer search:

→ **2 species**

Serratus:

→ **7,000 species**

Obelisks:

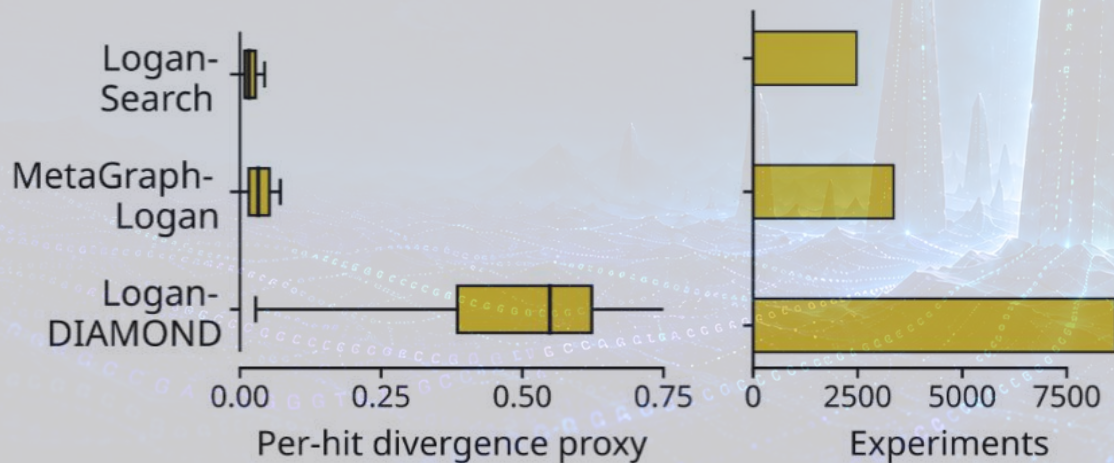


- 1 kb circular RNA elements, rod-like structure, 1-2 proteins
- Found in ~50% of human oral microbiomes
- Mostly everything unknown about their biology

Zheludev IN, Edgar RC, Lopez-Galiano MJ, De la Peña M, Babaian A, Bhatt AS, Fire AZ. Viroid-like colonists of human microbiomes. *Cell* (2024)

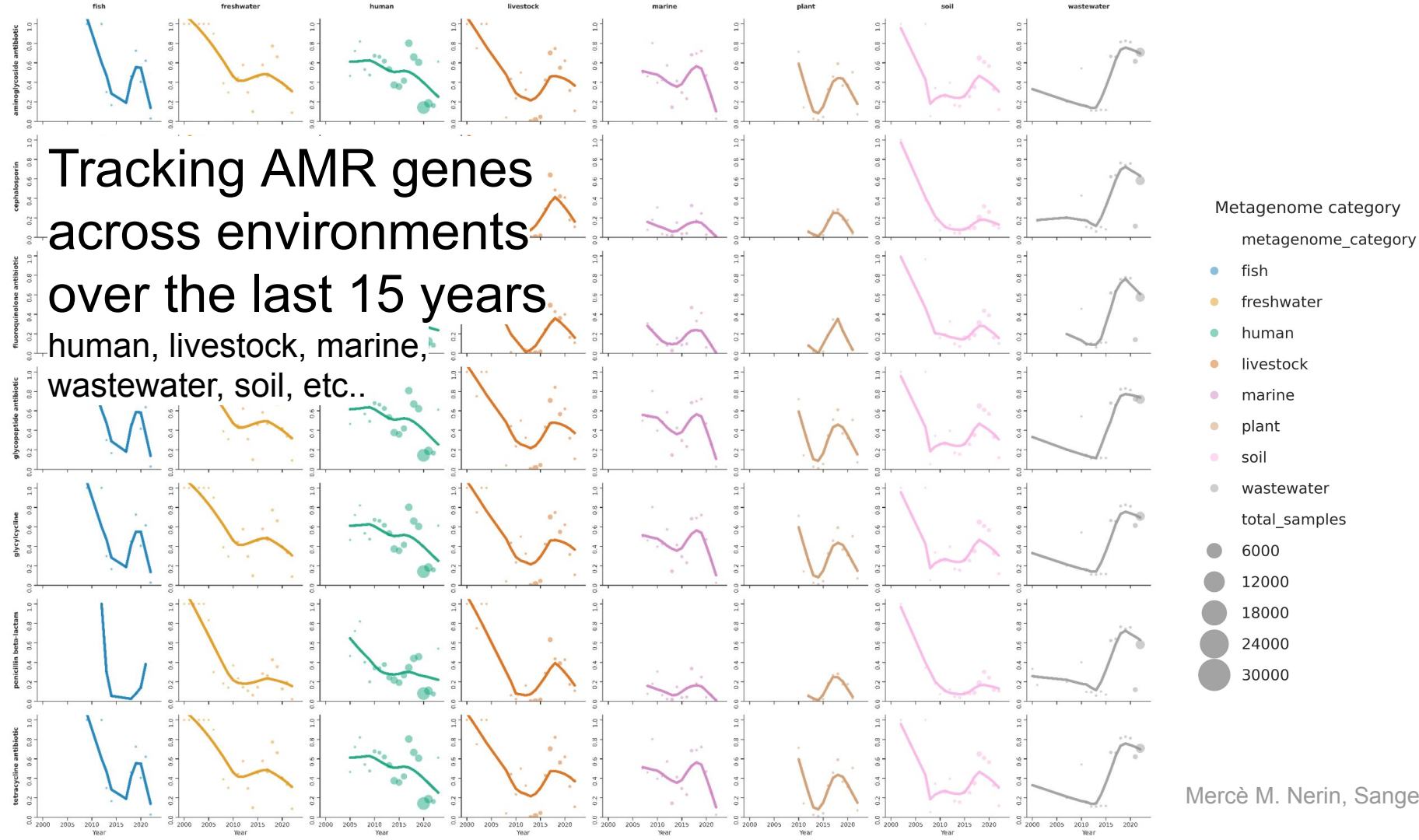
Logan: > 3x as many Obelisk species

Translated protein search with Logan-DIAMOND recovers substantially more divergent Obelisk homologs than nucleotide k-mer search



Tracking AMR genes across environments over the last 15 years

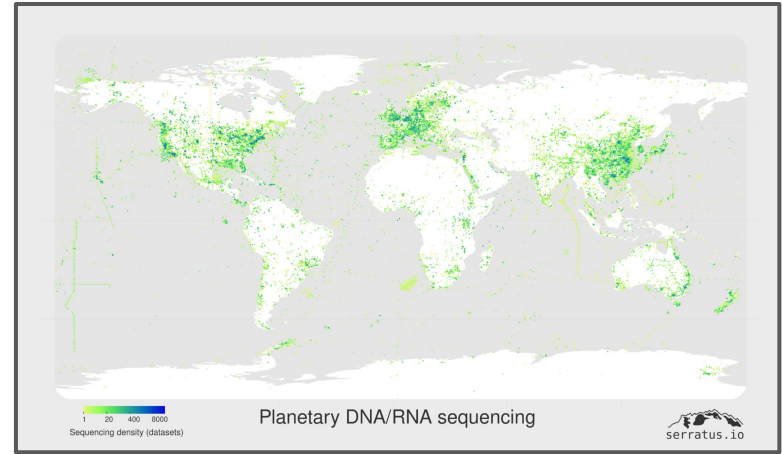
human, livestock, marine,
wastewater, soil, etc..



What Logan doesn't replace

Generation of new samples

High-quality assemblies



Sequence Bioinformatics

INSTITUT
Pasteur



Lab members:

Francesco Andreace
Gaetan Benoit
Rayan Chikhi
Kristen Curry
Roland Faure
Raphaël Malak
Mélanie Ridel
Timothé Rouze
Yoshihiro Shibuya

Alumni:

Luc Blassel
Luca Denti
Yoann Dufresne
Camila Duitama
Mael Kerbiriou
Téo Lemane
Camille Marchet
Pierre Marijon
Riccardo Vicedomini

Mentorship support:

Eduardo Rocha
Olivier Gascuel

Logan collaborators:

Artem Babaian, UofT
Brice Raffestin, IP
Greg Autric, AWS
Maxime Hugues, AWS
Anton Korobeynikov, IND
Robert Edgar, IND
Anton Nekrutenko, PSU
+ 25 Logan paper co-authors

Stéphane Fournier, DSI
AWS support: Dorian Schaal



Dorian Schaal
Sales Representative, AWS



Adrien Lainé
Account Manager, AWS



Greg Autric
Solution Architect, AWS



Brice Raffestin
DevOps, Institut Pasteur



Dr. Maxime Hugues
HPC Solution Architect, AWS