

# Logan

## Planetary-Scale Sequencing Data Analysis Surveys Evolution

Rayan Chikhi  
Institut Pasteur  
k-mer workshop Sanger 2025

# Sequence Bioinformatics

@ Institut Pasteur



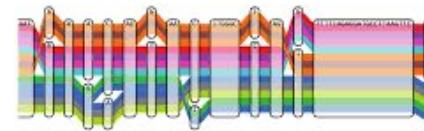
Genomes &  
metagenomes  
assembly



Algorithms and  
data structures  
on k-mers



Sequence  
search in very  
large datasets



Pangenomics

assembly graphs, contigs, unitigs

# de Bruijn graphs

A **de Bruijn** graph for a fixed integer  $k$ :

1. **Nodes** = all  $k$ -mers in the reads
2. **Edges** = all exact overlaps of length exactly  $(k - 1)$  between  $k$ -mers

Example for  $k = 3$  and a single read:

ACTG

ACT → CTG

Evomics course



# de Bruijn graph

Example for many reads and still  $k = 3$ .

ACTG

CTGC

TGCC

ACT → CTG → TGC → GCC

# de Bruijn graph: redundancy

What happens if we add redundancy?

ACTG

ACTG

CTGC

CTGC

CTGC

TGCC

TGCC

dBG,  $k = 3$ :

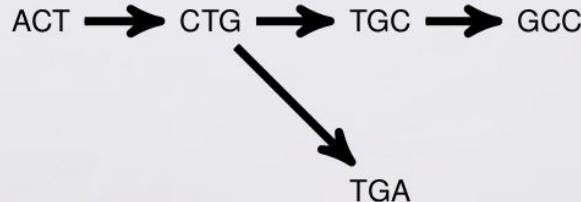


# de Bruijn graph: errors

How a sequencing error (at the end of a read) impacts the de Bruijn graph?

ACTG  
CTGC  
CTGA  
TGCC

dBG,  $k = 3$ :



# de Bruijn graph: repeats

What is the effect of a small repeat on the graph?

ACTG

CTGC

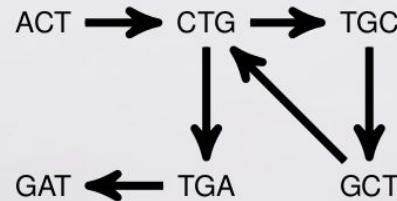
TGCT

GCTG

CTGA

TGAT

dBG,  $k = 3$ :



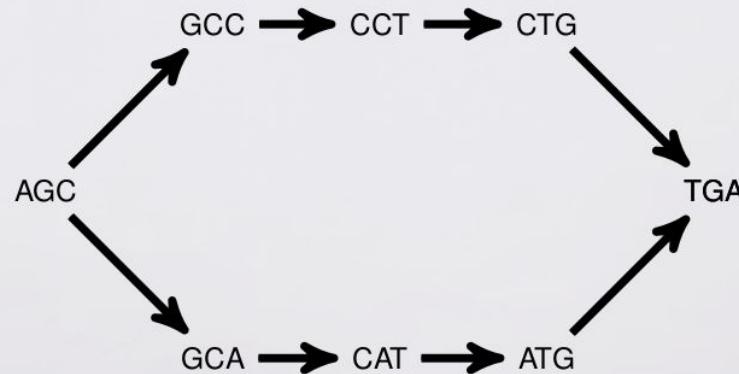
# de Bruijn graph: SNPs

SNPs can be directly “found” in the graph.

AGC~~C~~TGA

AGC~~A~~TGA

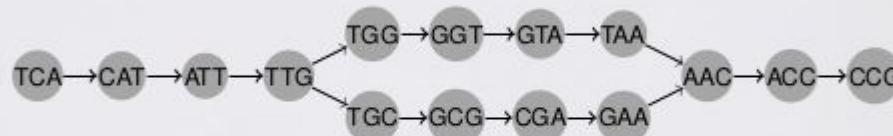
DBG,  $k = 3$ :



**Unitigs** = compacted de Bruijn graph

## Constructing the compacted dBG

Input:

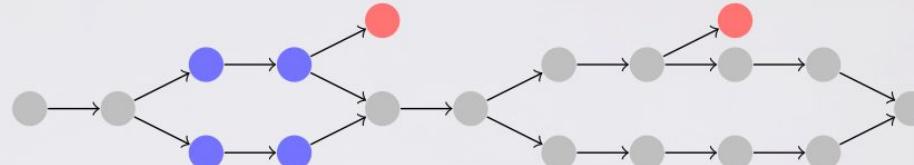


After **compaction**, output:

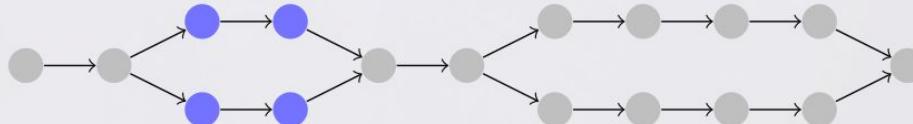


# Short read assemblers

- 1) de Bruijn **graph** construction



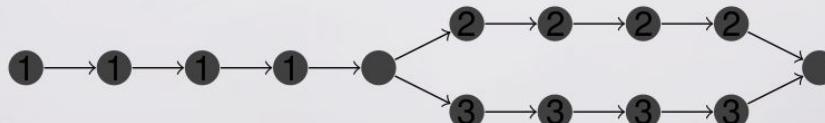
- 2) Likely **sequencing errors** are removed.



- 3) Variations (e.g. SNPs, similar repetitions) are removed.  
→ **Collapses strains**

→ **Collapses strains**

- 4) **Simple paths** (i.e. contigs) are returned.



- 5) Extra steps: repeat-resolving, scaffolding

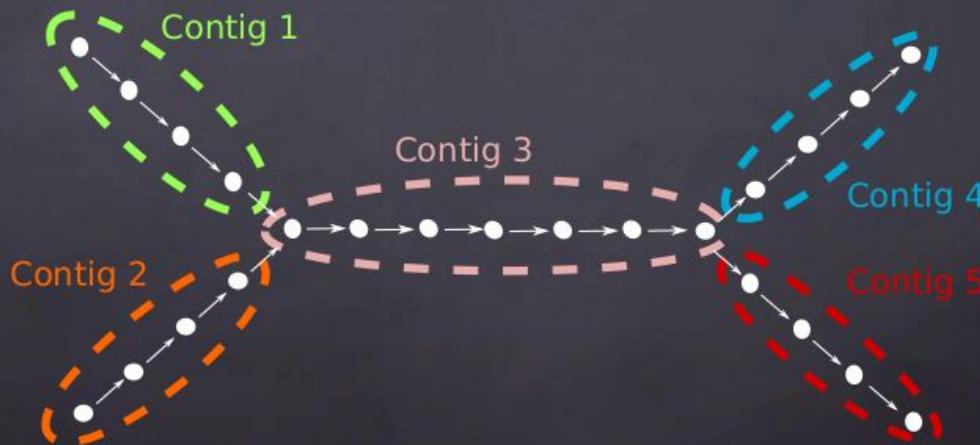
# CONTIGS CONSTRUCTION

**Contigs** are *node-disjoint* **simple paths**.

\* of the simplified assembly graph

*simple path*: all internal nodes have a single in/out edge.

*node-disjoint*: two different paths cannot share a node.

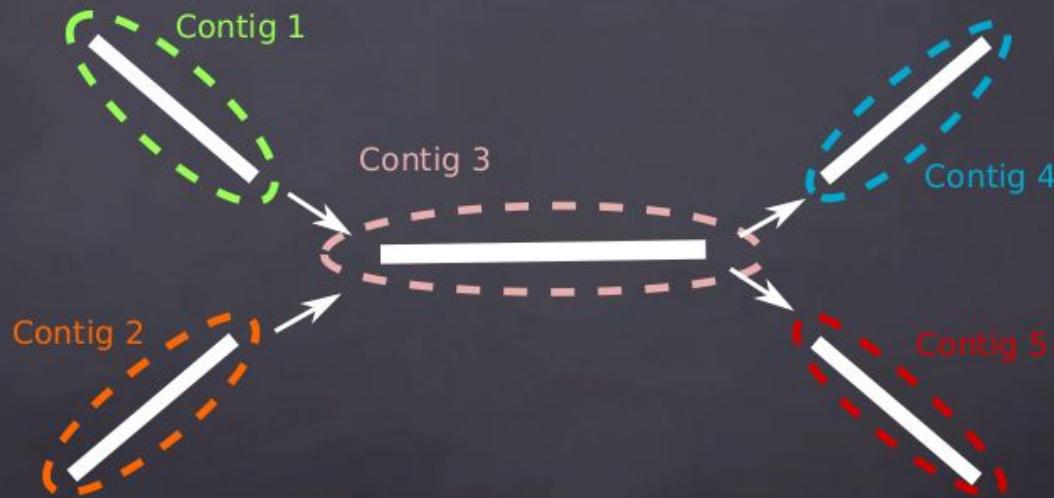


if you didn't simplify the graph, those paths are called **unitigs**

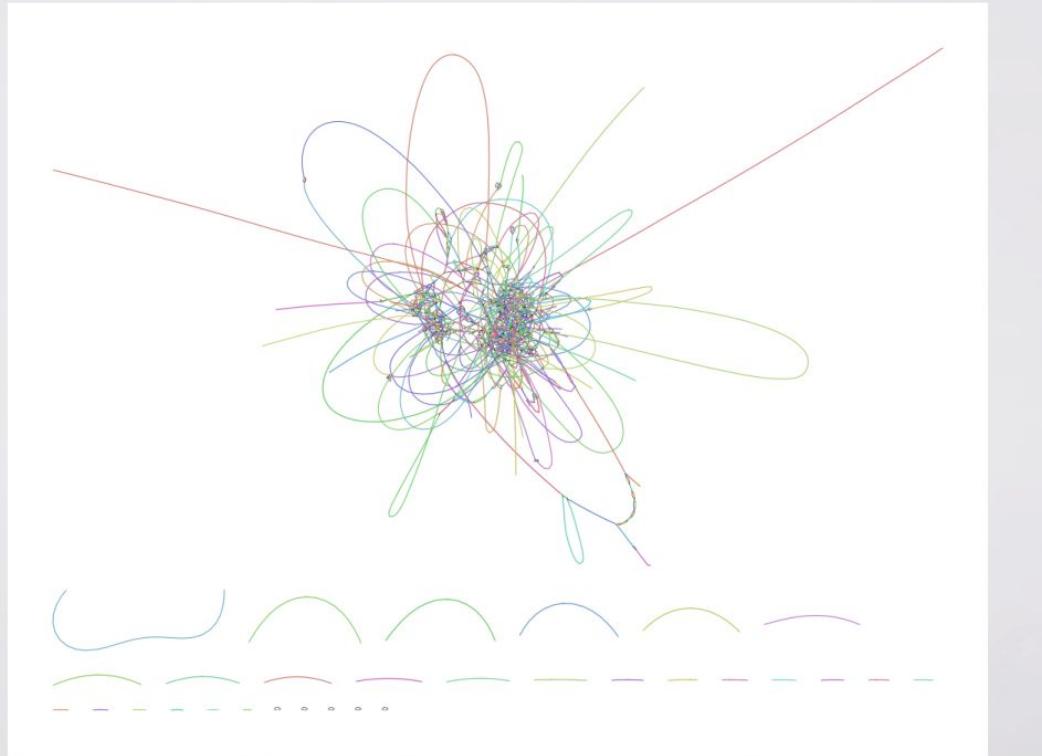
# CONTIGS GRAPH

The result of an assembly is a **contig graph**:

- nodes = contigs
- edges = overlaps between contigs



# Actual (simplified) de Bruijn graph



chr14:20Mbp-20.5Mbp GAGE PE reads, **SPAdes** 3.8 k=31: 1k nodes

# Assembly graph visualization: Bandage

Bandage ~ /media/ryan/Data/Bandage\_demo/O7\_NW1\_metagenome/NW1\_LastGraph

File Tools View Help

De Bruijn graph information

Nodes: 51,639  
Edges: 65,832  
Total length: 18,712,634

Graph drawing

Scope: Entire graph  
Style: Single  Double

Graph display

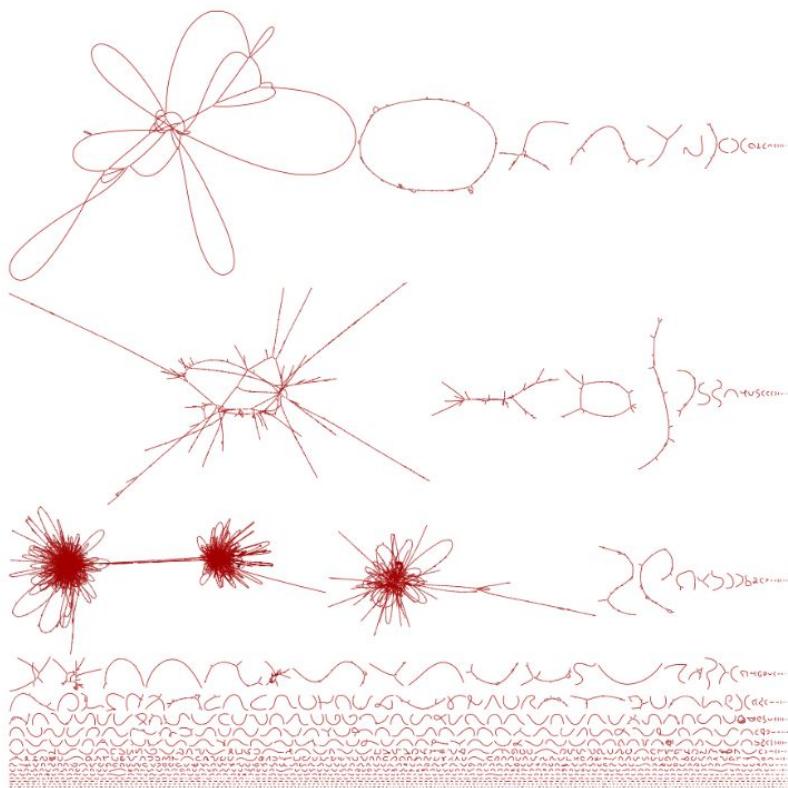
Zoom: 2.6%  
Uniform colour

Node labels

Custom  Number   
Length  Coverage

BLAST

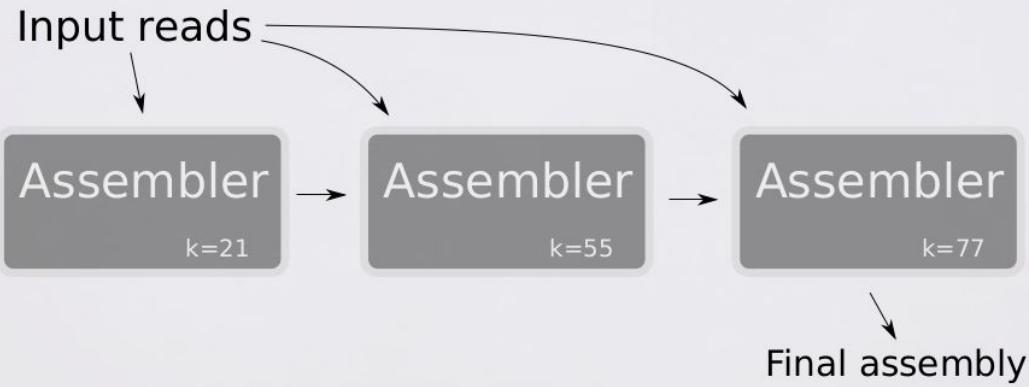
Create/view BLAST search  
Query:



Find nodes

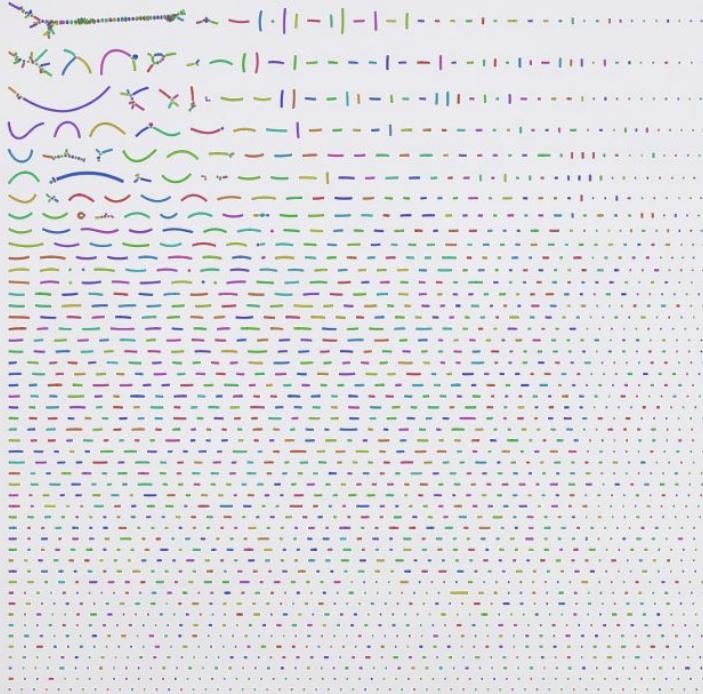
Node(s):

# Multi-k



In principle, **better** than single-k assembly.

*Salmonella* genome, Velvet assembly



$k = 91$  (too high, but shown for comparison)

single k

*Salmonella* genome, SPAdes assembly



final k:

$k = 99$

multi k

# Wrap-up assembly part

# Big data in biology: NCBI GenBank & WGS

The screenshot shows the NCBI GenBank homepage. At the top, there's a blue header with the NIH logo and "National Library o" followed by "National Center for Biotechnol". Below the header, a navigation bar has "GenBank" selected. A dropdown menu shows "Nucleotide" and other options like "Protein" and "Sample". Below the navigation bar, there are buttons for "GenBank", "Submit", "Genomes", and "WGS". The main content area is titled "GenBank Overview" and contains a section titled "What is GenBank?". It describes GenBank as "the NIH genetic sequence database".

**Type:** genome assemblies, >500,000 species  
**Size:** 1.2 terabytes ([2022](#))

incl. RefSeq  
All sequences are *annotated*

The screenshot shows the NCBI Whole Genome Shotgun Submissions page. The header is identical to the GenBank page. The navigation bar includes "WGS" in addition to "GenBank", "Submit", and "Genomes". The main content area is titled "Whole Genome Shotgun Submissions" and contains a section titled "What is Whole Genome Shotgun (WGS)?". It defines WGS as projects for eukaryotes being sequenced by a whole genome shotgun approach.

**Type:** genome assemblies  
**Size:** 16 TB ([2022](#))

*Unannotated*

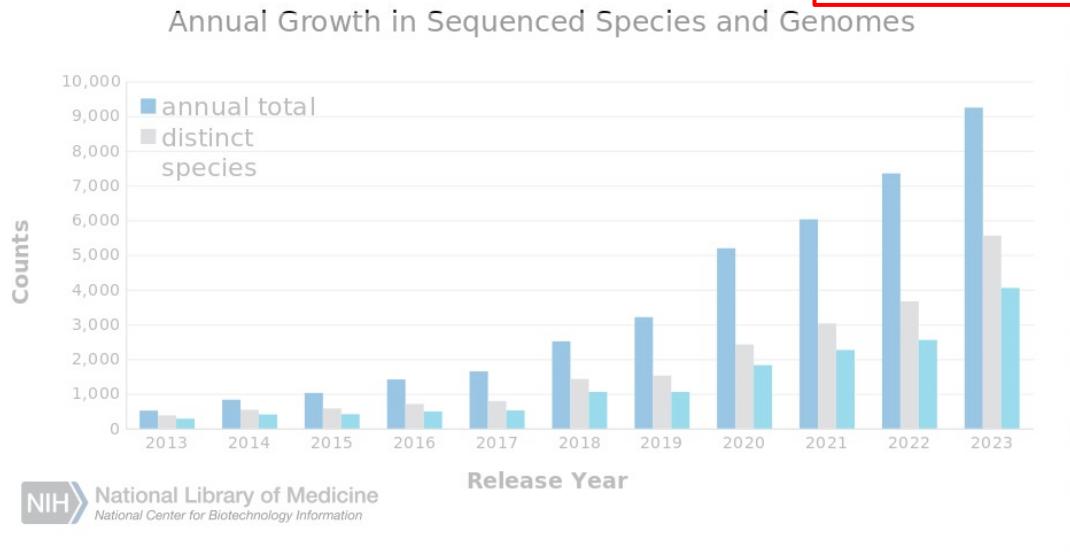
Also big “restricted” data:  
UK Biobank, TCGA, ..

# How complete are those databases?

## ALL EUKARYOTIC GENOMES (Cumulative: Dec 2023):

GenBank genomes (all): 36,593 (15,453 species)  
GenBank (with annotation): 6,817 (3,801 species)

(Out of 8 million known species..)



GenBank eukaryotic genome submissions (2021):

- 55% are contaminated
- 80% lack annotation
- 20% have annotation
  - 58% have >50% proteins annotated as "HYPOTHETICAL"

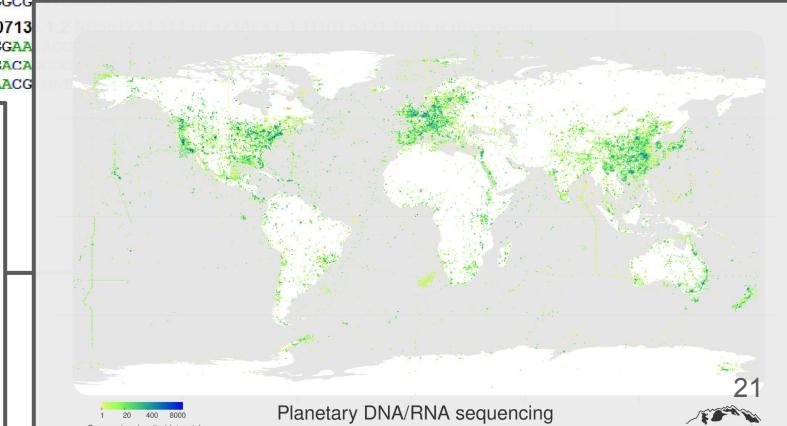
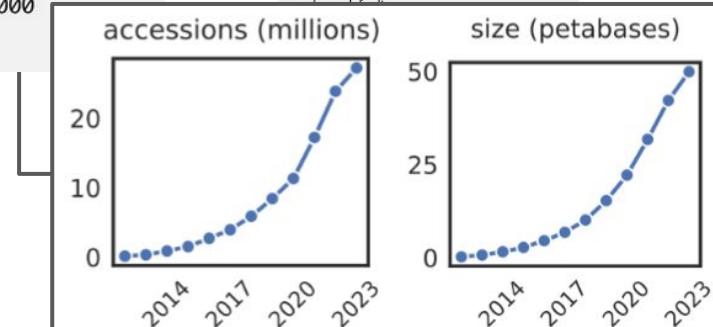
NCBI

NCBI  
SRA

## All public sequencing reads

**Size:** 50 Pbases  
as of Dec 2023

peta	[P]	$10^{15} = 1\,000\,000\,000\,000\,000$	INA (Illumina n: ERX34307)
tera	[T]	$10^{12} = 1\,000\,000\,000\,000$	-----
giga	[G]	$10^9 = 1\,000\,000\,000$	-----
mega	[M]	$10^6 = 1\,000\,000$	accessions



# 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

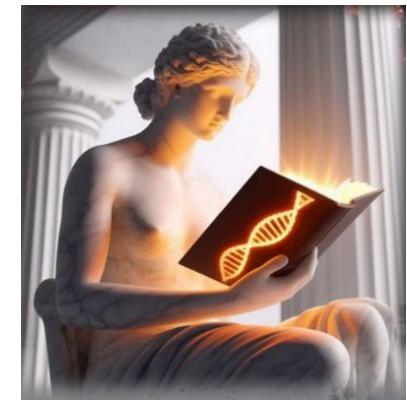
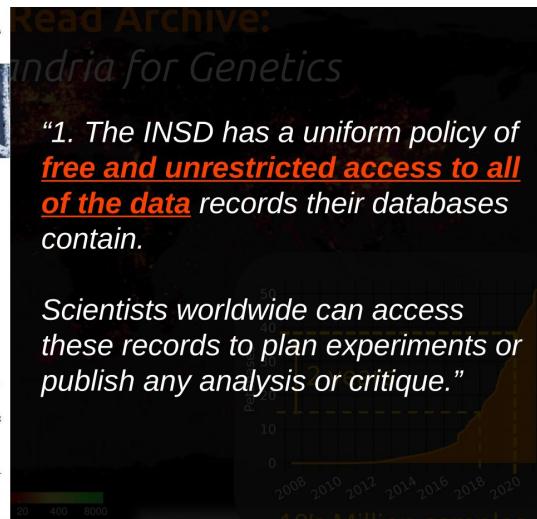


#### Nucleotide Sequence Database Policies

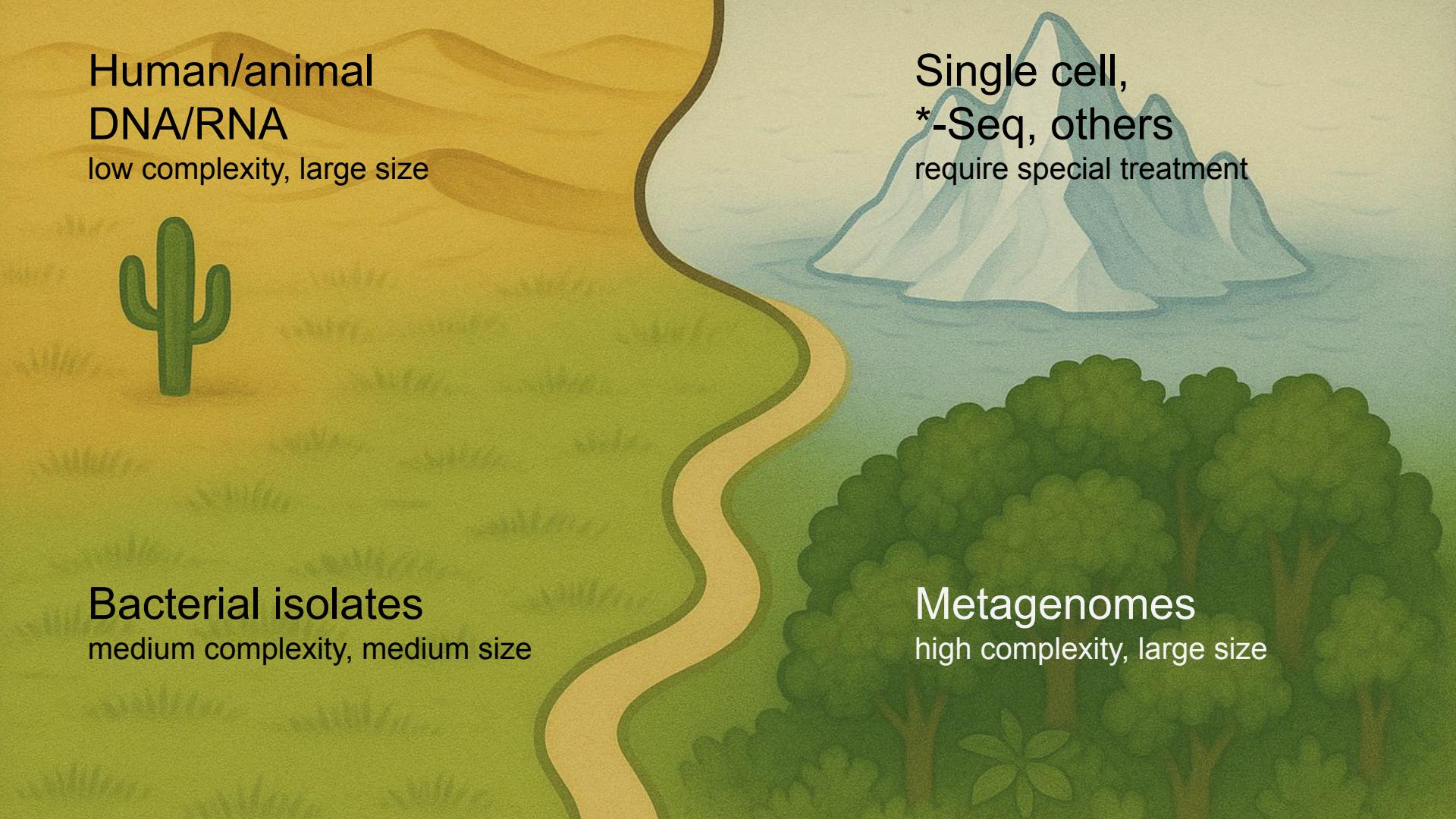
Science 298 (5597): 1333 15 Nov 2002

1. The INSD 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 submitters, following the practices of scientists utilizing other scientific literature.
2. The INSD 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 data.
3. All database records submitted to the INSD will remain permanently accessible as part of the scientific record. Corrections of errors and update 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 to check the information displayed on the Web sites maintained by the INSD to ensure that the public. It is the responsibility of the submitters to ascertain that they have the right to submit the data.
5. Beyond standard editorial control and some internal integrity checks (for example, proper use of INSD formats and translation of data as required), INSD entries are verified; the quality and accuracy of the record are the responsibility of the submitting author, not 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

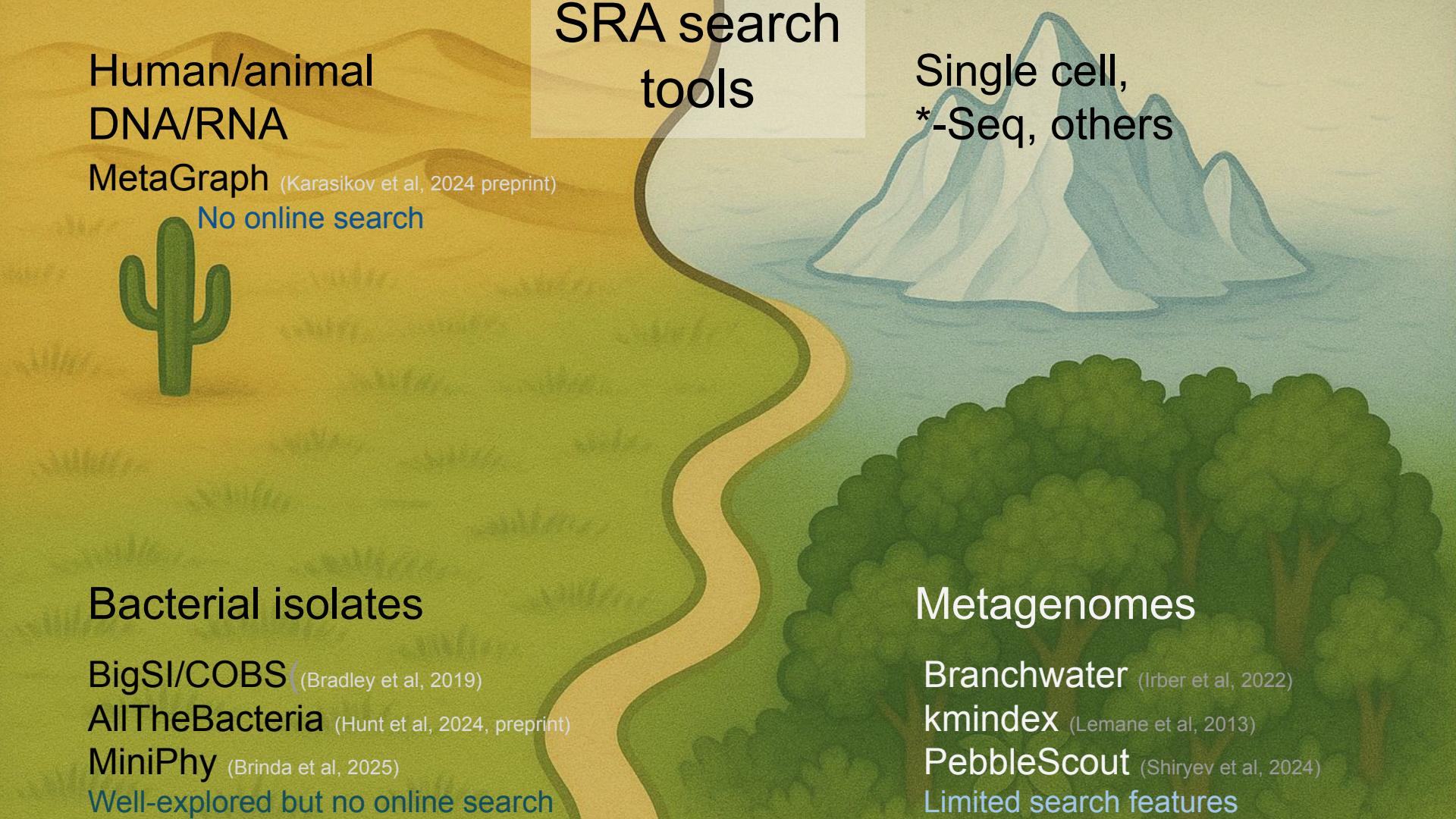


**Human/animal  
DNA/RNA**  
low complexity, large size

**Single cell,  
\*-Seq, others**  
require special treatment

**Bacterial isolates**  
medium complexity, medium size

**Metagenomes**  
high complexity, large size



Human/animal  
DNA/RNA

MetaGraph (Karasikov et al, 2024 preprint)  
No online search



## SRA search tools

Single cell,  
\*-Seq, others

Bacterial isolates

BigSI/COBS (Bradley et al, 2019)  
AllTheBacteria (Hunt et al, 2024, preprint)  
MiniPhy (Brinda et al, 2025)  
Well-explored but no online search

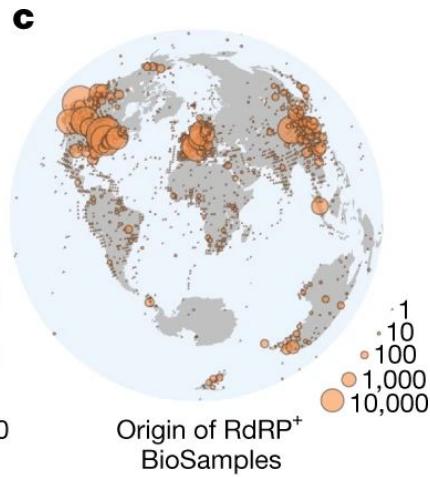
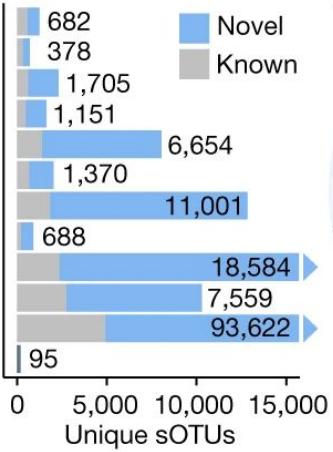
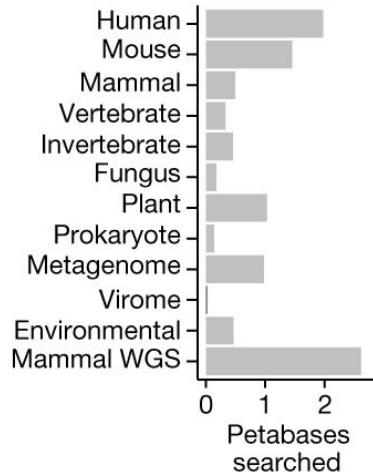
Metagenomes

Branchwater (Irber et al, 2022)  
kmindex (Lemane et al, 2013)  
PebbleScout (Shiryev et al, 2024)  
Limited search features

# Serratus: all public RNA-seqs analyzed for viral discovery



Discovered 130,000 new RNA viral species through large-scale read alignment, 9 new coronaviruses species.  
(Edgar et al, Nature, 2022)



Some follow-ups to Serratus

Viral reactivation (Nature 2023)



Discovered HHV-6 reactivation in CAR-T cells.  
Independent use of Serratus data

Obelisks

**Intriguing find.** Stanford University discovers obelisks hiding in human microbiomes  
Updated - February 06, 2024 at 11:18 AM | London

This new biological phenomenon, detailed in a recent preprint, challenges the conventional understanding

State of Sequence Data Archives (2025):



A wide-angle photograph of a vast mountain range under a dramatic sky. The mountains are heavily covered in snow, with deep shadows and bright highlights from the setting sun. In the foreground, a thick layer of white clouds covers the base of the mountains. The text "Logan" is centered in the middle ground, partially obscured by the clouds.

Logan

# Logan: Outline

- **Reconstructed all genomes in the entire SRA**
- (At draft-level quality, but still..)
- 50 petabases of reads were downloaded & assembled on AWS cloud
- Results are hosted on S3 with no egress charges (AWS Open Data)
- Publicly available: <https://github.com/IndexThePlanet/Logan>
- 2 PB of unitigs (high accuracy) and 0.3 PB of contigs (high contiguity)
- It's done, finally

**Logan: Planetary-Scale Genome Assembly Surveys Life's Diversity**

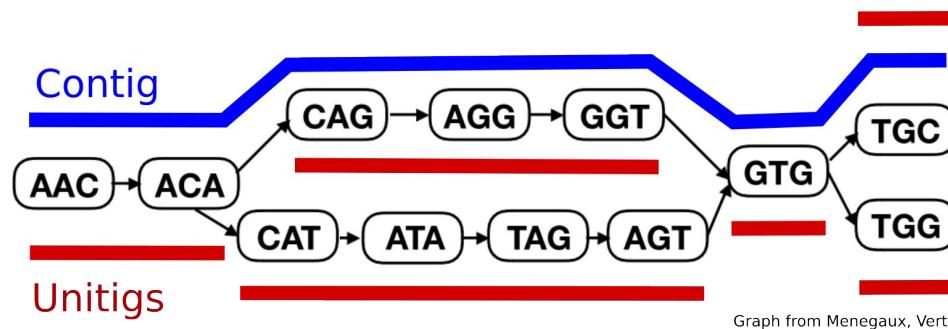
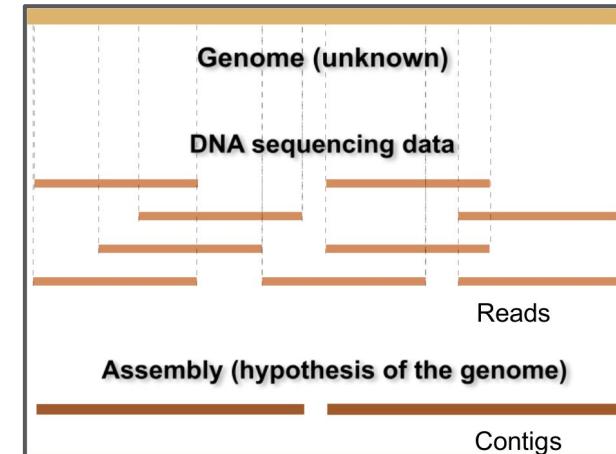
 Rayan Chikhi,  Brice Raffestin,  Anton Korobeynikov,  Robert Edgar,  Artem Babaian

**doi:** <https://doi.org/10.1101/2024.07.30.605881>

# Unitigs? Contigs?

**Contigs**: typical output of genome assembly methods

**Unitig**: simple path in the de Bruijn graph



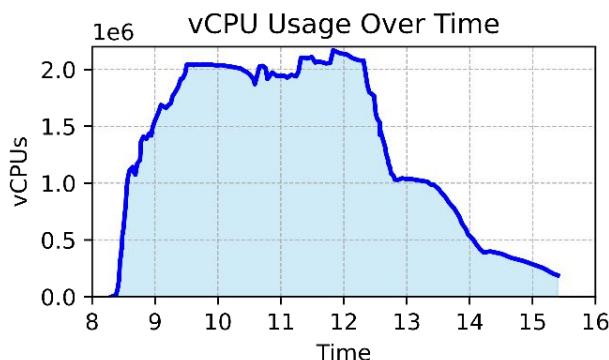
Why unitigs? they keep all variants (SNPs, indels, ...)

Contigs are consensus sequences

# Logan: computation statistics

## Global statistics

Input SRA Accessions	27 million
Input SRA size	50 petabases
Total CPU Hours	~30 million
Number of Runs	6
Total Runtime	30 hours

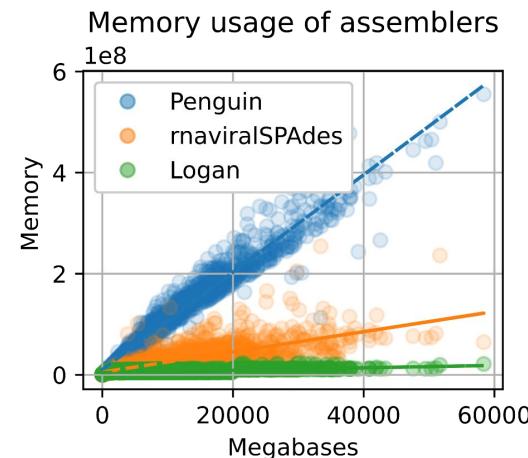
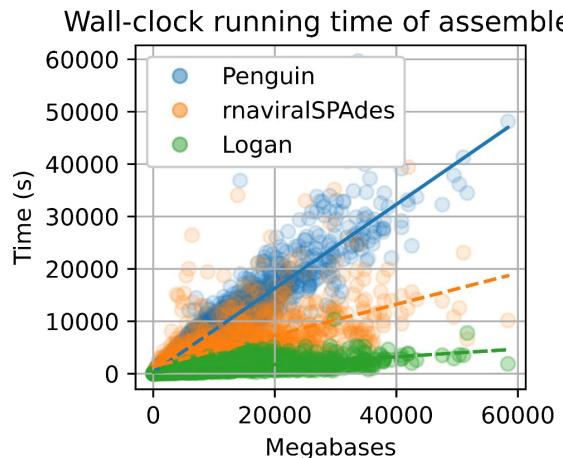


## Run 6 statistics

Input data	19.6 petabases
Runtime*	7 hours
Peak Number of Instances	73,100
Peak Number of vCPUs	2.18 million
Peak Total EBS storage	52 petabytes

# Why wasn't this done before?

- **Genome assembly is compute- and memory-intensive, usually.**
- We used a simple pipeline of **highly optimized components**:
  - Reads → counted kmers → de Bruijn graph → unitigs
  - Unitigs → simplification of graph → contigs
- Speeding up each step took **decades of bioinformatics research**



# Algorithmic components used in Logan

- String algorithms (minimizers in KMC inside cuttlefish2)
- Parallel efficient algorithms (cuttlefish2)
- Minimum perfect hashing (BBHash inside cuttlefish2, Minia)
- Large (billions+ nodes) graph manipulation (Minia)
- Compression (zstd in f2sz)

Some of the methods story: R. Chikhi, ***A tale of optimizing the space taken by de Bruijn graphs***, Computability in Europe (2021) [[PDF](#)]

Flavor: how to store 3 billion 31-length DNA strings in < 10 GB RAM with O(1) queries?

A wide-angle photograph of a mountain range under a dramatic sky. The mountains are covered in white snow, with deep shadows and bright highlights from the setting sun. In the foreground, a thick layer of white clouds covers the base of the mountains. The sky above is a mix of warm orange and yellow hues, transitioning into cooler blues and purples at the top.

Logan data

# Logan

# 2 petabases

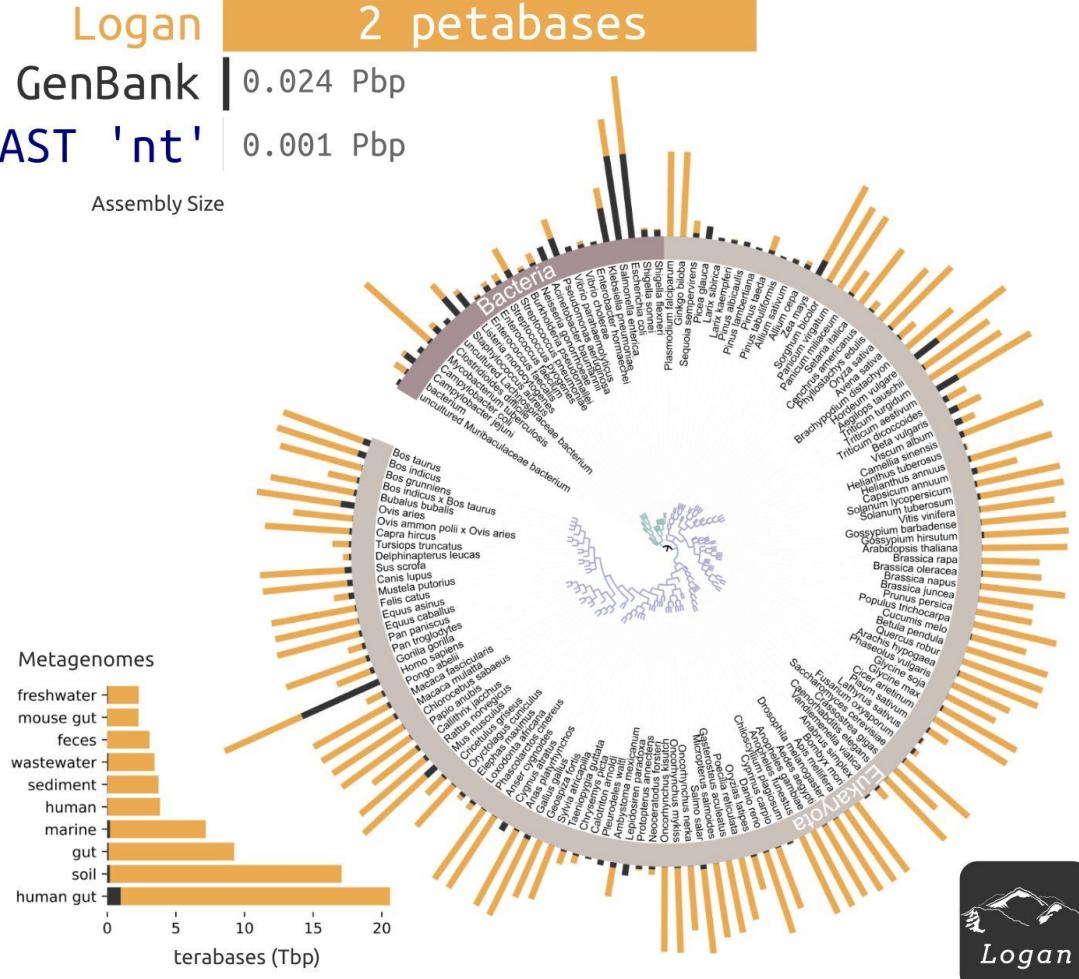
GenBank

0.024 Pbp

BLAST 'nt'

0.001 Pbp

Assembly Size



Logan

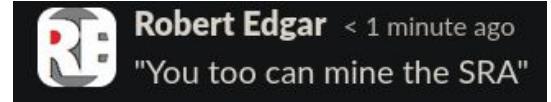
# Accessing Logan

```
aws s3 cp s3://logan-pub/c/[acc]/[acc].contigs.fa.zstd .
```

From anywhere, no account needed

(Many tutorials on Logan's Github)

# Want to dive in Logan data ?



- All Logan unitigs & contigs are public, and tutorials are available at [github.com/IndexThePlanet/Logan](https://github.com/IndexThePlanet/Logan)
- But if you need assistance: contact me
- [Logan-search.org](https://Logan-search.org) service for high-identity alignment search
- We do whole-Logan high-sensitivity alignments regularly
  - Can include your sequence(s) in the next batch

A wide-angle photograph of a mountain range, likely the Himalayas, during sunset or sunrise. The mountains are heavily covered in snow and ice, with deep shadows in the valleys. The sky is a warm, golden-yellow color, and the tops of the mountains are illuminated. In the foreground, there is a thick layer of white, fluffy clouds.

# Tools around Logan

# Logan Search: how to use

- 1) Have a query DNA sequence

e.g. ACTAGTAGAGTAGATGAGGGACATAGATGAGCACATGATGAGAGACACATTG

- 2) Query it in **Logan Search**, wait 2 minutes
- 3) Explore where this sequence is present across the world/SRA

# Logan Search

**INPUT**

**text** file session

**Query sequence(s) \***  
Fasta/Fastq format  
>Query  
ACCGTAGCCTTACAATTAA

**Load**

**NOTIFICATION**

Email  
Your email

**CONFIGURATION**

Groups

Threshold = 0.5

0.25 1.0

**Submit** **Reset**

**Table** **Map** **Plot** **Contigs/Unitigs Search (BETA)** **Help**

kmer\_coverage > 0.7 AND assay\_type IN (WGS,WGA)

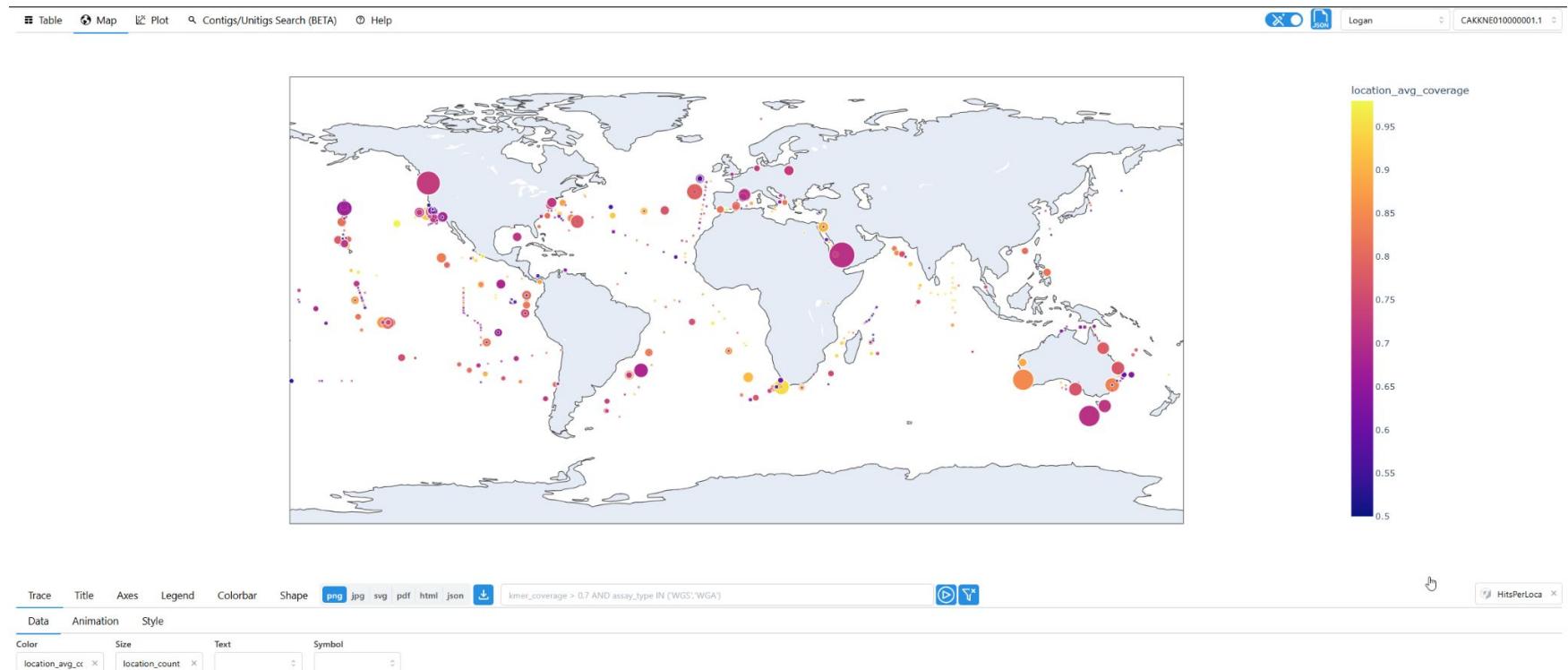
ID	kmer_coverage	bioproject	biosample	bioproject_title	bioproject_description	sample_acc	assay_type	center_name	experim...
ERR6909055 (SRA)[OV]	1	PRJEB847927 (SRA)[OV]	SAMEA10271030 (SRA)[OV]	Chromosome-scale genome as...	Pelagophytes (Stramenopiles) a...	ERS7925717	WGS	GSC	ERX6531
ERR3497222 (SRA)[OV]	1	PRJEB843158 (SRA)[OV]	SAMEA5899549 (SRA)[OV]	Collection of Marine Eukaryote...	This project is part of the Marin...	ERS3688172	RNA-Seq	GSC	ERX351
SRR1296779 (SRA)[OV]	1	PRJNA248394 (SRA)[OV]	SAMN02740027 (SRA)[OV]	Marine Microbial Eukaryote Tra...	The Marine Microbial Eukaryot...	SR5618895	RNA-Seq	NATIONAL CENTER FOR GENO...	SRX551
SRR14100031 (SRA)[OV]	1	PRJNA517804 (SRA)[OV]	SAMN1381556 (SRA)[OV]	100 Algal genome project (ALG...	100 Algal genome project (ALG...	SR5763869	WGS	NEW YORK UNIVERSITY ABU D...	SRX1041
ERR9764111 (SRA)[OV]	1	PRJEB847927 (SRA)[OV]	SAMEA14430949 (SRA)[OV]	Chromosome-scale genome as...	Pelagophytes (Stramenopiles) a...	ERS12037107	Hi-C	GSC	ERX9311
ERR3497221 (SRA)[OV]	1	PRJEB843158 (SRA)[OV]	SAMEA5899549 (SRA)[OV]	Collection of Marine Eukaryote...	This project is part of the Marin...	ERS3688172	RNA-Seq	GSC	ERX351
SRR1296780 (SRA)[OV]	1	PRJNA248394 (SRA)[OV]	SAMN02740028 (SRA)[OV]	Marine Microbial Eukaryote Tra...	The Marine Microbial Eukaryot...	SR5618896	RNA-Seq	NATIONAL CENTER FOR GENO...	SRX551
SRR1197260 (SRA)[OV]	1	PRJNA239089 (SRA)[OV]	SAMN01985059 (SRA)[OV]	Pelagomonas calceolata Geno...	Pelagomonas calceolata geno...	SR576631	WGS	JCVI	SRX4931
SRR1827860 (SRA)[OV]	1	PRJNA814250 (SRA)[OV]	SAMN02644151 (SRA)[OV]	Metagenomic time-series the ...	This study examines monthly th...	SR512225932	WGS	CLARK UNIVERSITY	SRX1441
SRR1296778 (SRA)[OV]	1	PRJNA248394 (SRA)[OV]	SAMN02740026 (SRA)[OV]	Marine Microbial Eukaryote Tra...	The Marine Microbial Eukaryot...	SR5618894	RNA-Seq	NATIONAL CENTER FOR GENO...	SRX551
SRR1296781 (SRA)[OV]	0.99	PRJNA248394 (SRA)[OV]	SAMN02740029 (SRA)[OV]	Marine Microbial Eukaryote Tra...	The Marine Microbial Eukaryot...	SR5618897	RNA-Seq	NATIONAL CENTER FOR GENO...	SRX551
SRR815684 (SRA)[OV]	0.988	PRJNA193556 (SRA)[OV]	SAMN02144620 (SRA)[OV]	Pelagomonas calceolata strain...	Pelagomonas calceolata transcr...	SR5421323	RNA-Seq	JCVI	SRX278
ERR1726884 (SRA)[OV]	0.982	PRJEB4352 (SRA)[OV]	SAMEA2623204 (SRA)[OV]	EMG produced TPA metageno...	The Third Party Annotation (TP...	ERS493517	WGS	GSC	ERX179
ERR868425 (SRA)[OV]	0.972	PRJEB4352 (SRA)[OV]	SAMEA2622699 (SRA)[OV]	EMG produced TPA metageno...	The Third Party Annotation (TP...	ERS492708	WGS	GSC	ERX948
ERR868426 (SRA)[OV]	0.972	PRJEB4352 (SRA)[OV]	SAMEA2619943 (SRA)[OV]	EMG produced TPA metageno...	The Third Party Annotation (TP...	ERS488730	WGS	GSC	ERX948
SRR13386796 (SRA)[OV]	0.972	PRJNA690716 (SRA)[OV]	SAMN17257790 (SRA)[OV]	Pelagomonas calceolata CCMP...	Genome sequencing of Pelago...	SR5799076	WGS	DAHOUISIE UNIVERSITY	SRX980
SRR8705096 (SRA)[OV]	0.972	PRJNA529320 (SRA)[OV]	SAMN11263736 (SRA)[OV]	The influence of shipping lanes ...	Metagenomes and metatranscr...	SR54542059	WGS	NANYANG TECHNOLOGICAL U...	SRX558
ERR599331 (SRA)[OV]	0.971	PRJEB4352 (SRA)[OV]	SAMEA2621080 (SRA)[OV]	EMG produced TPA metageno...	The Third Party Annotation (TP...	ERS490341	WGS	GSC	ERX556
ERR868454 (SRA)[OV]	0.971	PRJEB4352 (SRA)[OV]	SAMEA2620089 (SRA)[OV]	EMG produced TPA metageno...	The Third Party Annotation (TP...	ERS488924	WGS	GSC	ERX948
ERR1719359 (SRA)[OV]	0.971	PRJEB6609 (SRA)[OV]	SAMEA2623379 (SRA)[OV]	Metatranscriptome sequencing...	Metatranscriptome sequencing...	ERS493797	RNA-Seq	GSC	ERX178
ERR599284 (SRA)[OV]	0.97	PRJEB4352 (SRA)[OV]	SAMEA2621048 (SRA)[OV]	EMG produced TPA metageno...	The Third Party Annotation (TP...	ERS490307	WGS	GSC	ERX556
ERR868458 (SRA)[OV]	0.97	PRJEB4352 (SRA)[OV]	SAMEA2622936 (SRA)[OV]	EMG produced TPA metageno...	The Third Party Annotation (TP...	ERS493111	WGS	GSC	ERX947
SRR5924774 (SRA)[OV]	0.97	PRJNA385736 (SRA)[OV]	SAMN07482751 (SRA)[OV]	Marine metagenomes Metagen...	Marine amplicons from Austral...	SR52423102	WGS	BIOPLATFORMS AUSTRALIA	SRX308
ERR1719421 (SRA)[OV]	0.97	PRJEB6609 (SRA)[OV]	SAMEA2620025 (SRA)[OV]	Metatranscriptome sequencing...	Metatranscriptome sequencing...	ERS488834	RNA-Seq	GSC	ERX178
ERR868441 (SRA)[OV]	0.97	PRJEB4352 (SRA)[OV]	SAMEA2622325 (SRA)[OV]	EMG produced TPA metageno...	The Third Party Annotation (TP...	ERS492154	WGS	GSC	ERX948
SRR25584947 (SRA)[OV]	0.97	PRJNA1003508 (SRA)[OV]	SAMN36908814 (SRA)[OV]	Nitrite oxidizing bacteria in oxy...	Novel nitrite oxidizing bacteria ...	SR51856138	WGS	PRINCETON UNIVERSITY	SRX213

Page Size: 100

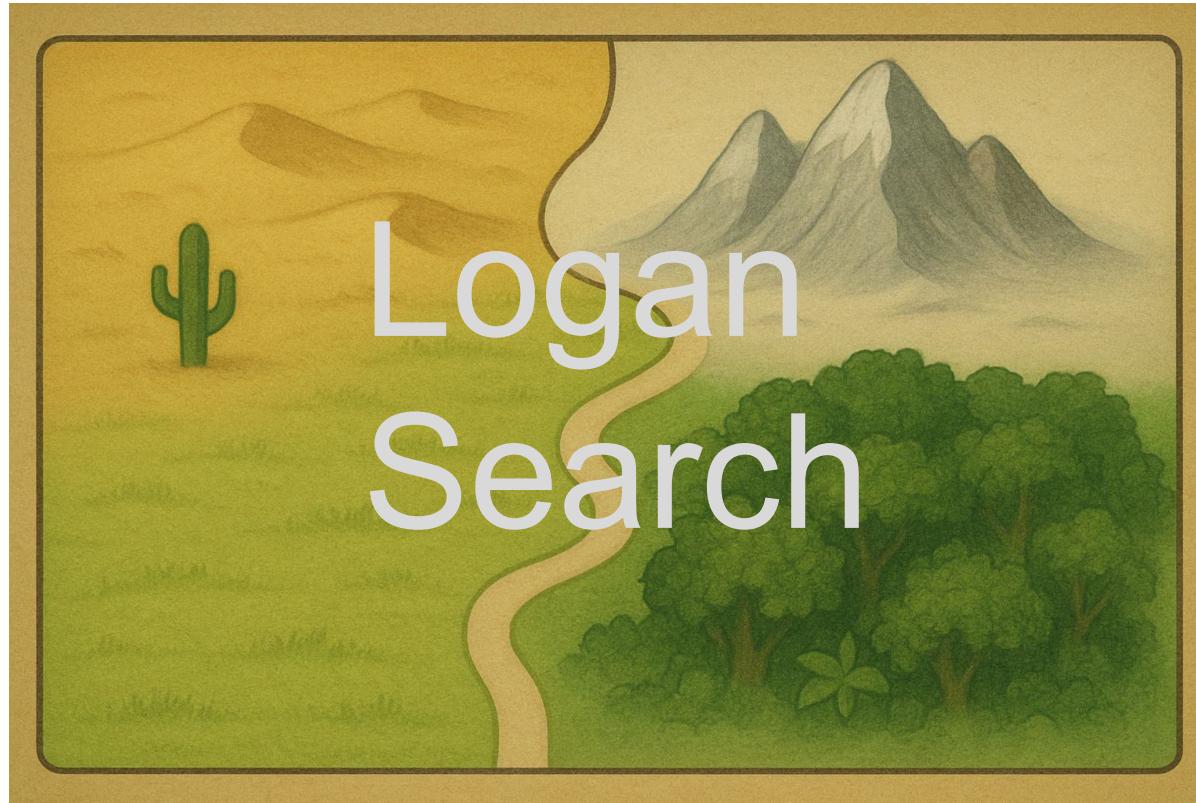
**Export** **Remove filters** **Filter NaN**

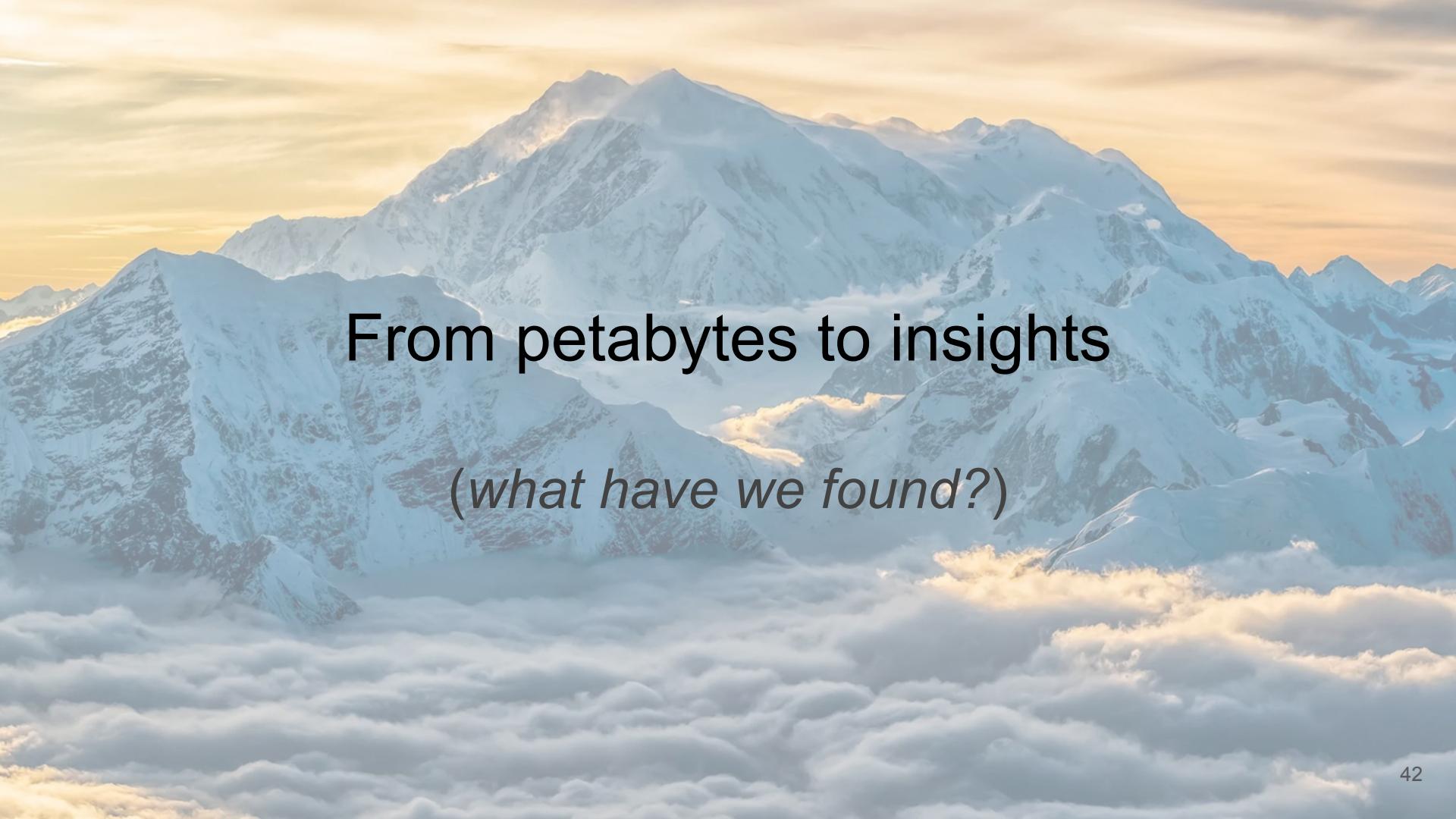
**About your query (AI generated)**  
The query sequence is likely derived from a marine environment and is related to Pelagomonas calceolata or a similar microorganism frequently found in marine metagenomes, as evidenced by its high representation in genomic and transcriptomic datasets locations, predominantly sequenced using Illumina platforms like HiSeq and NovaSeq.

# Logan Search



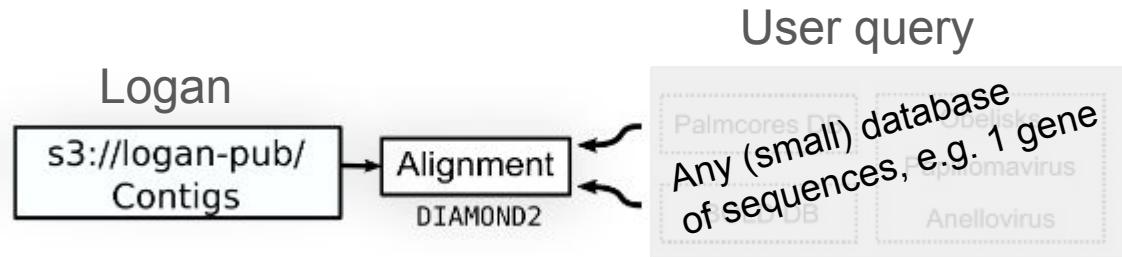
Logan Search is a k-mer index of all of SRA



A wide-angle photograph of a mountain range, likely the Himalayas, during sunset or sunrise. The mountains are heavily covered in snow, with deep shadows in the valleys and bright highlights on the peaks. In the foreground, a dense layer of white clouds covers the base of the mountains. The sky above is a warm, golden-yellow color, transitioning to a darker blue at the top.

From petabytes to insights  
*(what have we found?)*

# Our typical analysis workflow

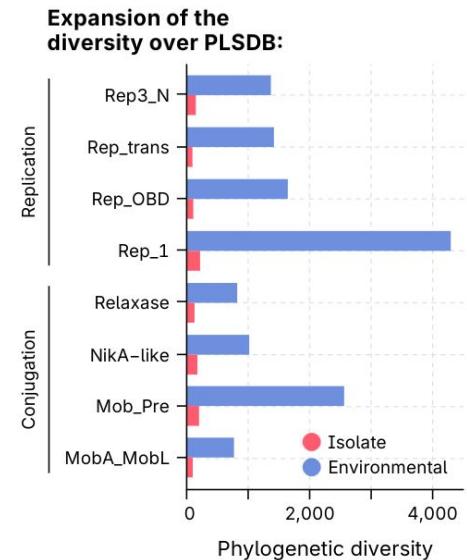


Results: collection of homologs to the query db, found across the entire SRA

This workflow can search for new *things* in genomes that were not found before

# Expansion of genetic diversities

- Obelisks
  - circular RNA, viroids, ~1kbp, mostly everything unknown about them
- Plasmids
  - Identification from Logan contigs, clustering, host assignment, ..
- P4 satellites
  - Bacterial mobile genetic elements, order of magnitude more
- AMR geography



Analyses by:  
Purav Gupta, UofT, Canada  
Marcos de la Pena, CSIC, Spain  
Jess Chen, UofT, Canada  
Antonio Camargo, JGI, USA  
Stephen Nayfach, JGI, USA  
Mateus Fiamenghi, JGI, USA  
Kristen Curry, Institut Pasteur, FR  
Eduardo Rocha, Institut Pasteur, FR  
Daniel Agustinho, BCM, USA  
Sina Majidian, BCM, USA  
Mercè Montoliu Nerin, Sanger, UK  
..and others

# Expanding the protein universe

- Predicted all proteins from Logan contigs using Prodigal
- Clustered using MMseqs2 (90% identity)

Obtained:

- **71M** ‘human-associated’ protein clusters
- **3.0B** ‘nonhuman-associated’ protein clusters

UniRef90 (UniProt clustered at 90% identity): **204M** clusters

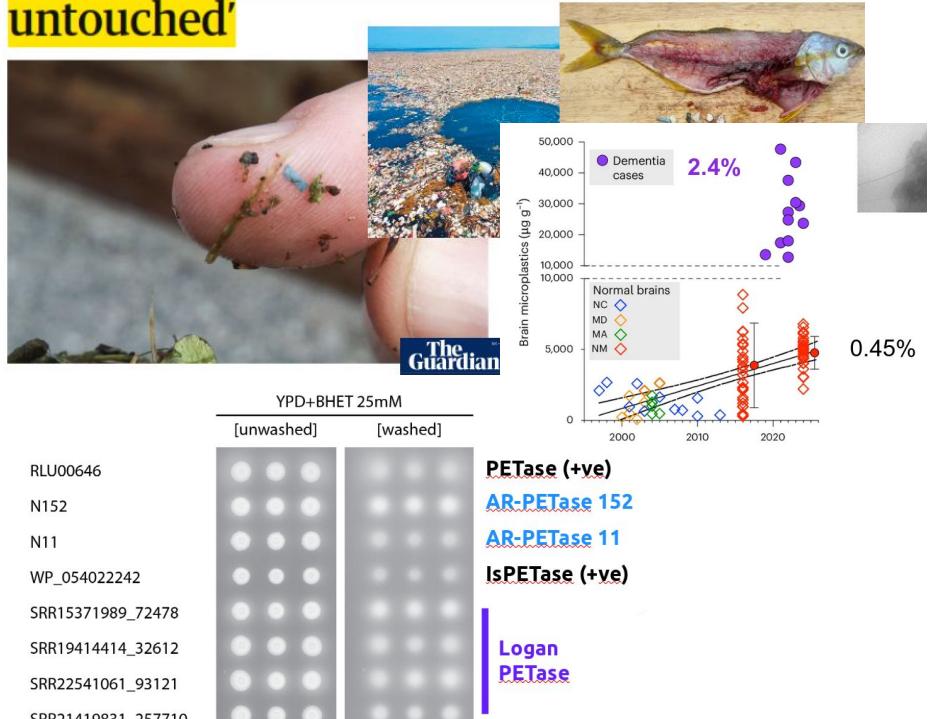
Analyses by:  
Martin Steinegger, SNU, Korea  
Joey Lee, SNU, Korea

# Logan discovers better plastic-eating enzymes

- **Microplastics crisis**
  - In our food, in our brain, in the oceans
- **Plastic-eating enzymes**

Originally discovered in a rubbish dump in Japan (IsPETase)
  - Promises to solve plastic pollution worldwide
- We found **thousands of natural homologs** of PETase in Logan data
- Logan PETases show better activity than state-of-the-art engineered ones

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



Ongoing work by RNAlab, UofT (please do not Tweet)

## Logan “fun facts”

- Logan computation would have taken ~1.5 years on local cluster.
- Just listing the S3 folder takes **~1 hour**
- Downloading all Logan contigs (385 TB) at 1 Gbits/s takes **30 days**
- Sequence alignment with DIAMOND2 (--sensitive) streaming all of Logan contigs takes **4 hours** on 60k cloud vCPUS (4k\$)

# Conclusion

- **SRA-scale analyses now 100x more tractable**
- **Logan: all of Life's genomic data at your fingertips**

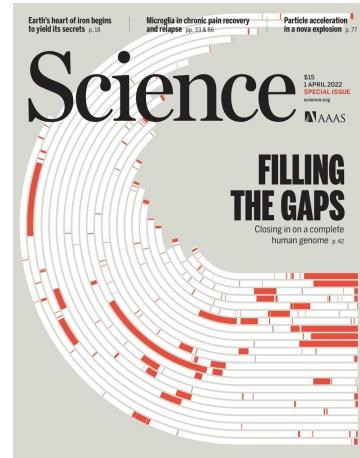
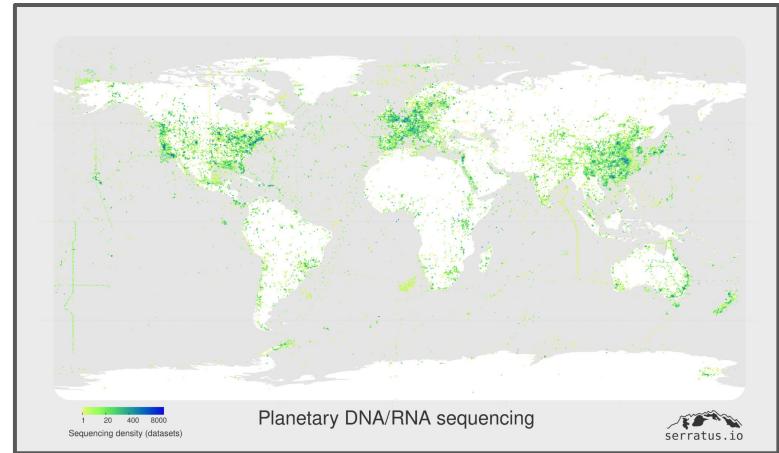
Technically:

- Easy data access (.fasta.zst format instead of .sra)
- K-mers pre-counted, mean abundance per unitig, assembly graphs provided

# What Logan doesn't replace

Generation of new samples

High-quality assemblies



# Sequence Bioinformatics



## Lab members:

Francesco Andreace  
Gaetan Benoit  
Rayan Chikhi  
Kristen Curry  
Yoann Dufresne  
Victor Levallois  
Roland Faure  
Mélanie Ridel  
Timothé Rouze  
Yoshihiro Shibuya



## Alumni:

Luc Bassel  
Luca Denti  
Camila Duitama  
Mael Kerbiriou  
Téo Lemane  
Camille Marchet  
Pierre Marijon  
Riccardo Vicedomini

Mentorship support:  
Eduardo Rocha  
Olivier Gascuel

## Logan co-creators:

Artem Babaian, UofT  
Brice Raffestin, IP  
Greg Autric, AWS  
Maxime Hugues, AWS  
Anton Korobeynikov, IND  
Robert Edgar, IND

- + too many current Logan collaborators to list

AWS support: Dorian Schaal



PR[AI]RIE  
PaRis Artificial Intelligence Research Institute



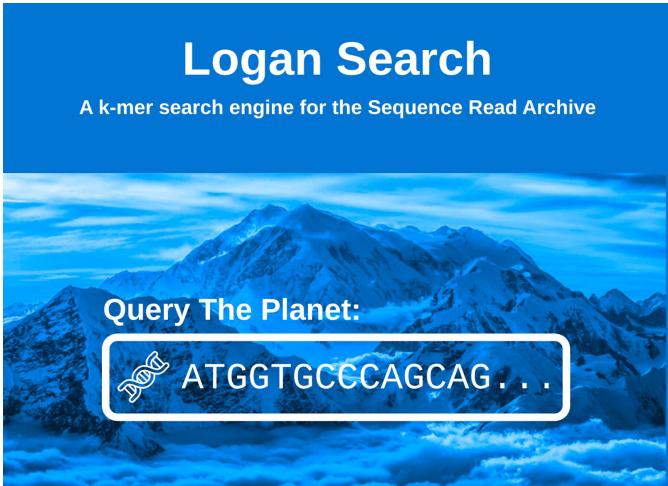
# Logan

Funded by ERC CoG “*IndexThePlanet*”

“A public dataset of all DNA/RNA assembled sequences on Earth (SRA)”

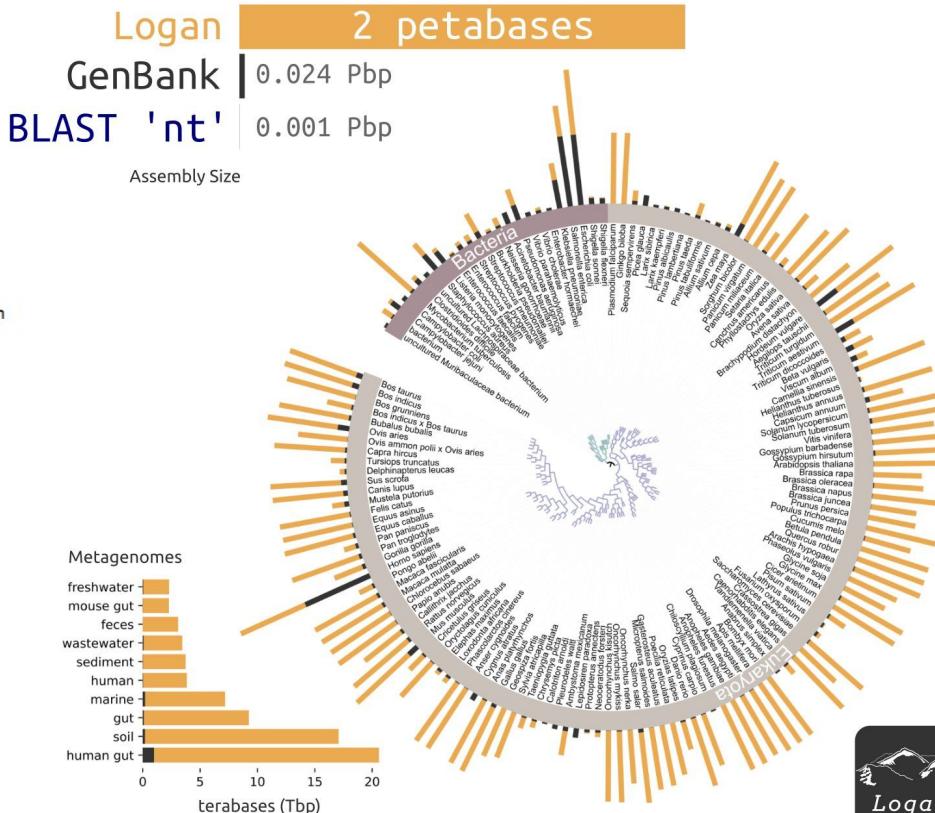
Logan: Planetary-Scale Genome Assembly Surveys Life’s Diversity

✉ Rayan Chikhi, Brice Raffestin, ⚡ Anton Korobeynikov, ⚡ Robert Edgar, ⚡ Artem Babaian  
doi: <https://doi.org/10.1101/2024.07.30.605881>



This screenshot shows the National Library of Medicine's SRA (Sequence Read Archive) website. At the top, it says "SRA - Now available on the cloud". Below that is a brief description: "Sequence Read Archive (SRA) data, available through multiple cloud-based services, provides a central repository of high throughput sequencing data. The archive supports metagenomic and environmental surveys. SRA stores raw sequence data and facilitates new discoveries through data analysis." There is also a link to "Advanced" search options.

SRA = \$10B of direct seq costs



# Future directions

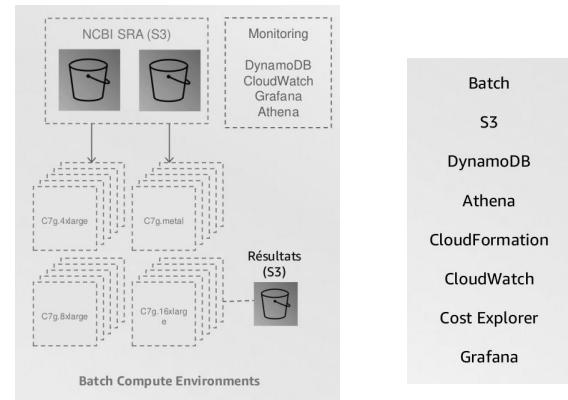
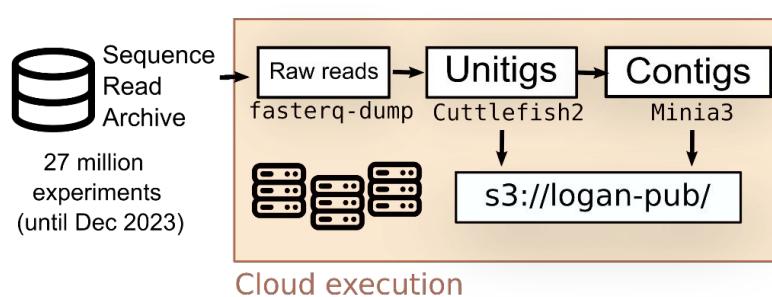
- LexicMap
- Training embeddings on protein clusters for search
- Integration with Galaxy
- Publish the paper :)
- Updates to Logan? maybe, for 1-2 years of new sequencing data

In collaboration with:

Anton Nekrutenko  
Wei Shen  
Zamin Iqbal  
Roland Faure  
and others

# Logan: project steps

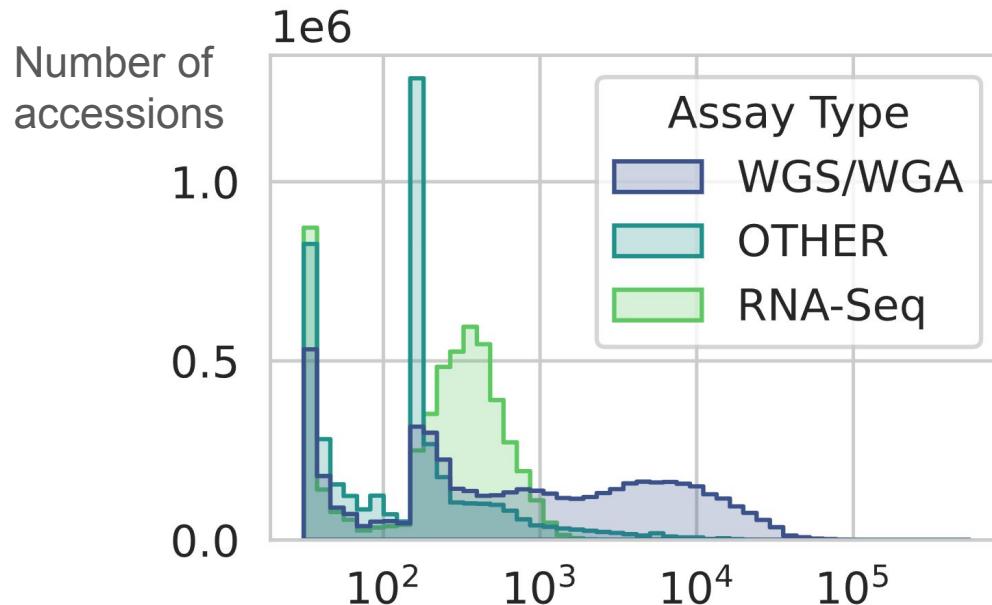
- **Step 1 (early 2024):** Download all of SRA, assemble each sample, host results publicly [done] 19 petabytes downloaded, 2 petabytes stored



- **Summer 2024:** preprint released
- **Step 2 (2025):** Index assemblies, create a search engine (“searching YouTube”) [partly done] <https://logan-search.org/>

# Assembly contiguity: draft level

& got slightly improved  
in v1.1 Logan contigs



**Contig N50**

higher=better

Except for RNAseqs