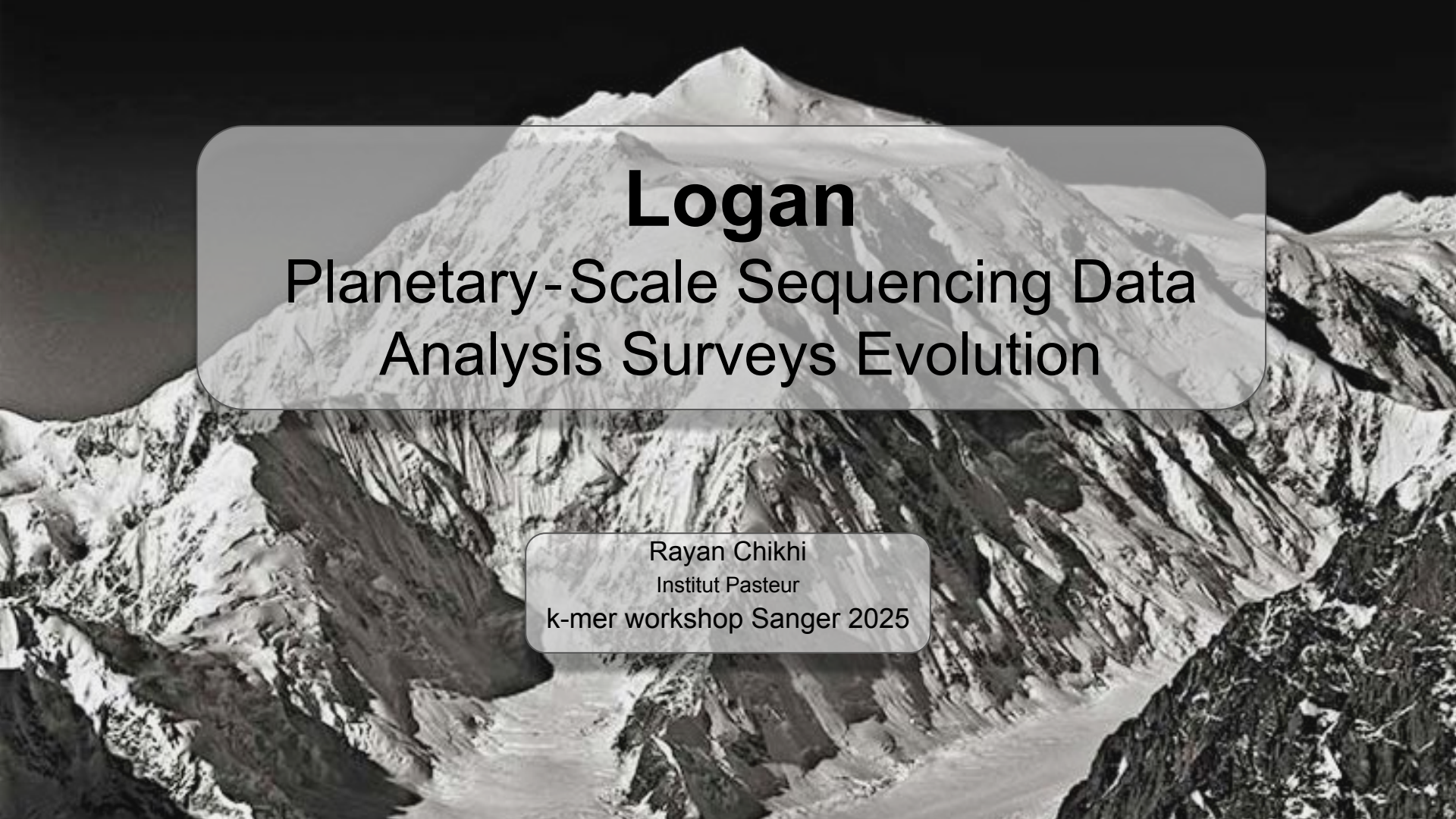


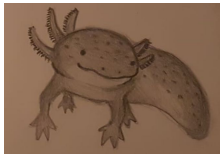
# Logan

## 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

INSTITUT PASTEUR PR[AI]RIE PaRis Artificial Intelligence Research InstitutE cnrs erc

# 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



# de Bruijn graph

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

```
ACTG
 CTGC
  TGCC
```

# de Bruijn graph: redundancy

What happens if we add redundancy?

```
ACTG
ACTG
 CTGC
 CTGC
 CTGC
  TGCC
  TGCC
```

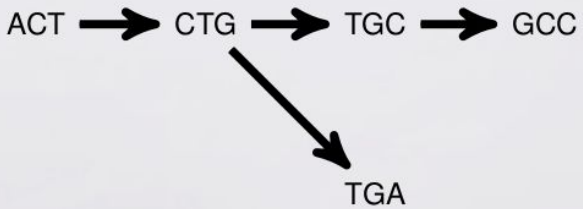
dBG, $k = 3$:

# de Bruijn graph: errors

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

```
ACTG
 CTGC
 CTGA
  TGCC
```

dBG, $k = 3$:

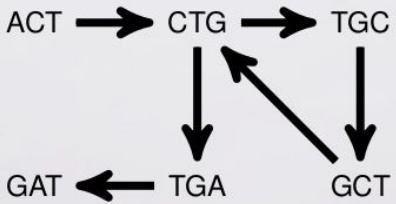


# de Bruijn graph: repeats

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

```
ACTG
 CTGC
  TGCT
   GCTG
    CTGA
     TGAT
```

dBG, $k = 3$:

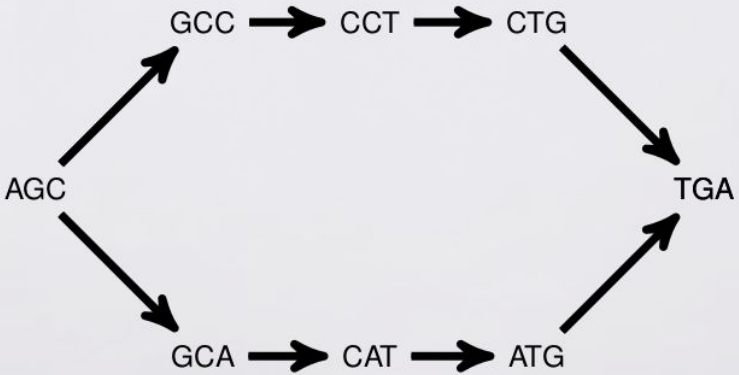


# de Bruijn graph: SNPs

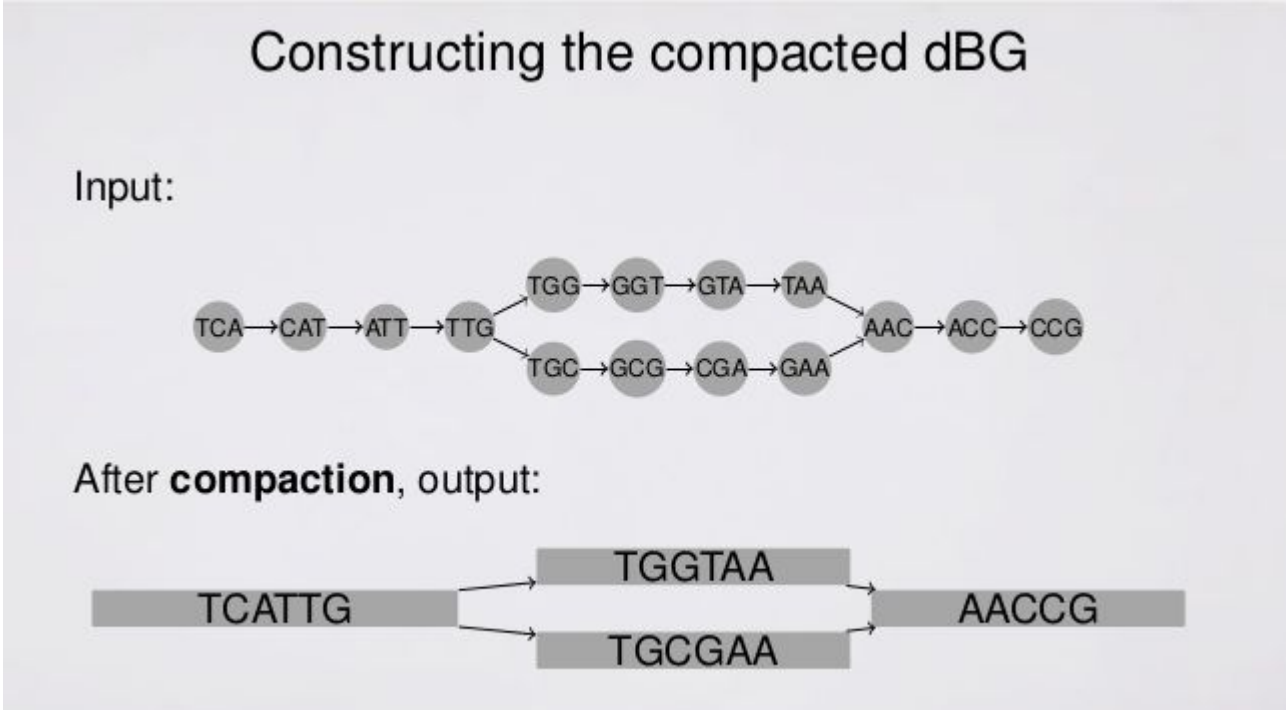
SNPs can be directly “found” in the graph.

```
AGCCTGA
AGCATGA
```

dBG, $k = 3$:

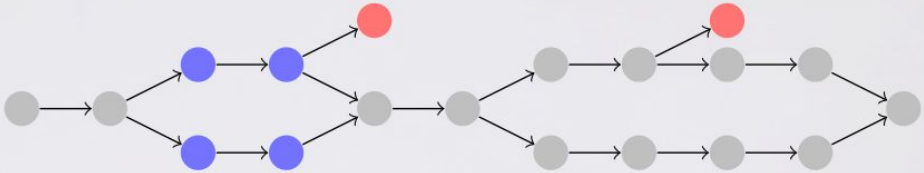


**Unitigs** = compacted de Bruijn graph

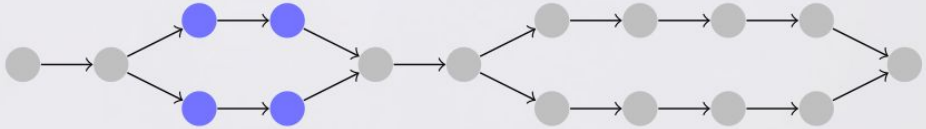


# Short read assemblers

1) de Bruijn **graph** construction



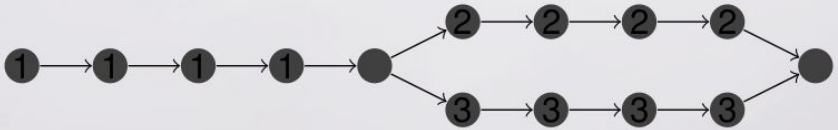
2) Likely sequencing errors are removed.



3) Variations (e.g. SNPs, similar repetitions) are removed.

→ **Collapses strains**

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

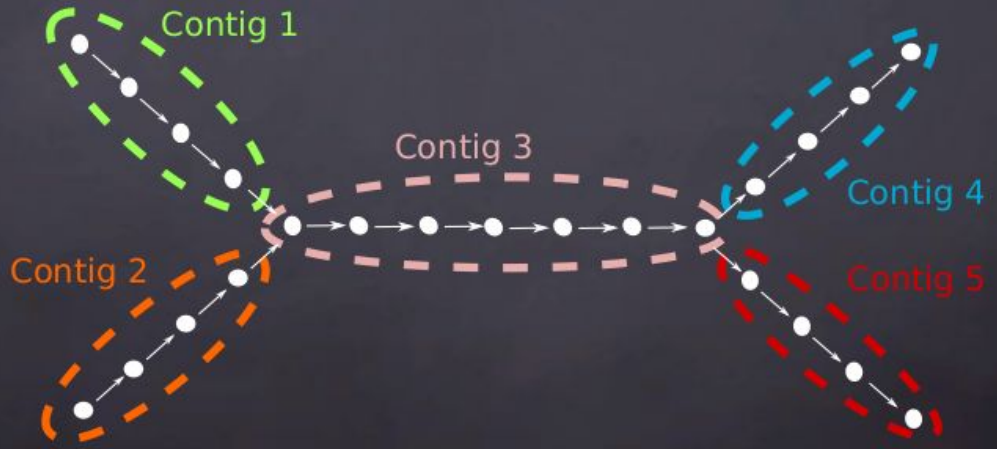



5) Extra steps: repeat-resolving, scaffolding

# CONTIGS CONSTRUCTION

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

*simple path:* all internal nodes have a single in/out edge.
*node-disjoint:* two different paths cannot share a node.



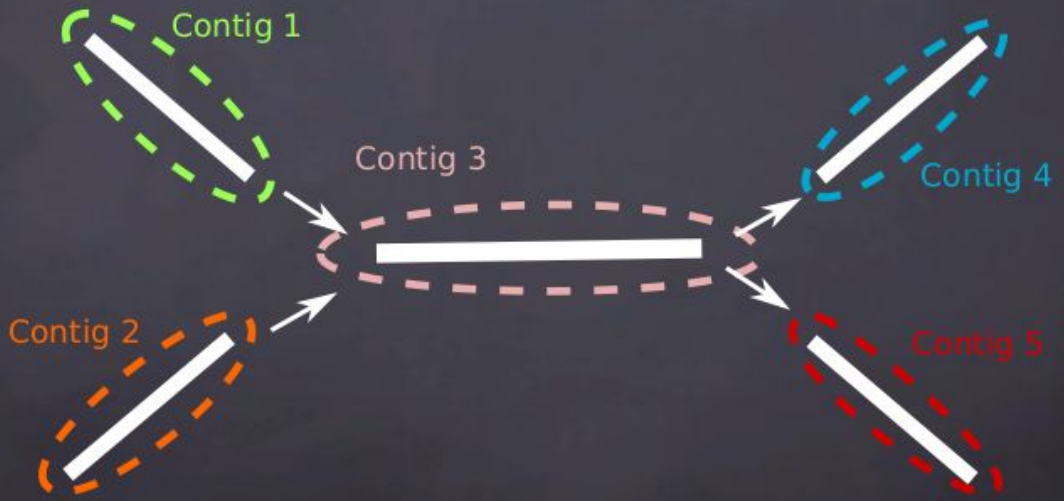

* of the simplified assembly graph

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

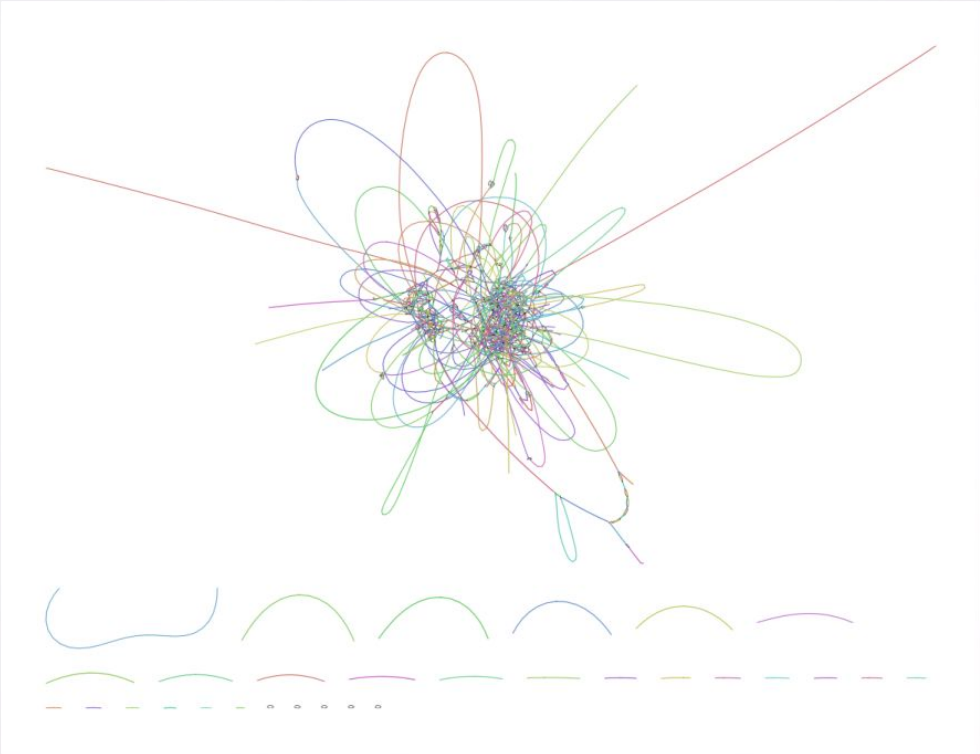
# CONTIGS GRAPH

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

- nodes = contigs
- edges = overlaps between contigs

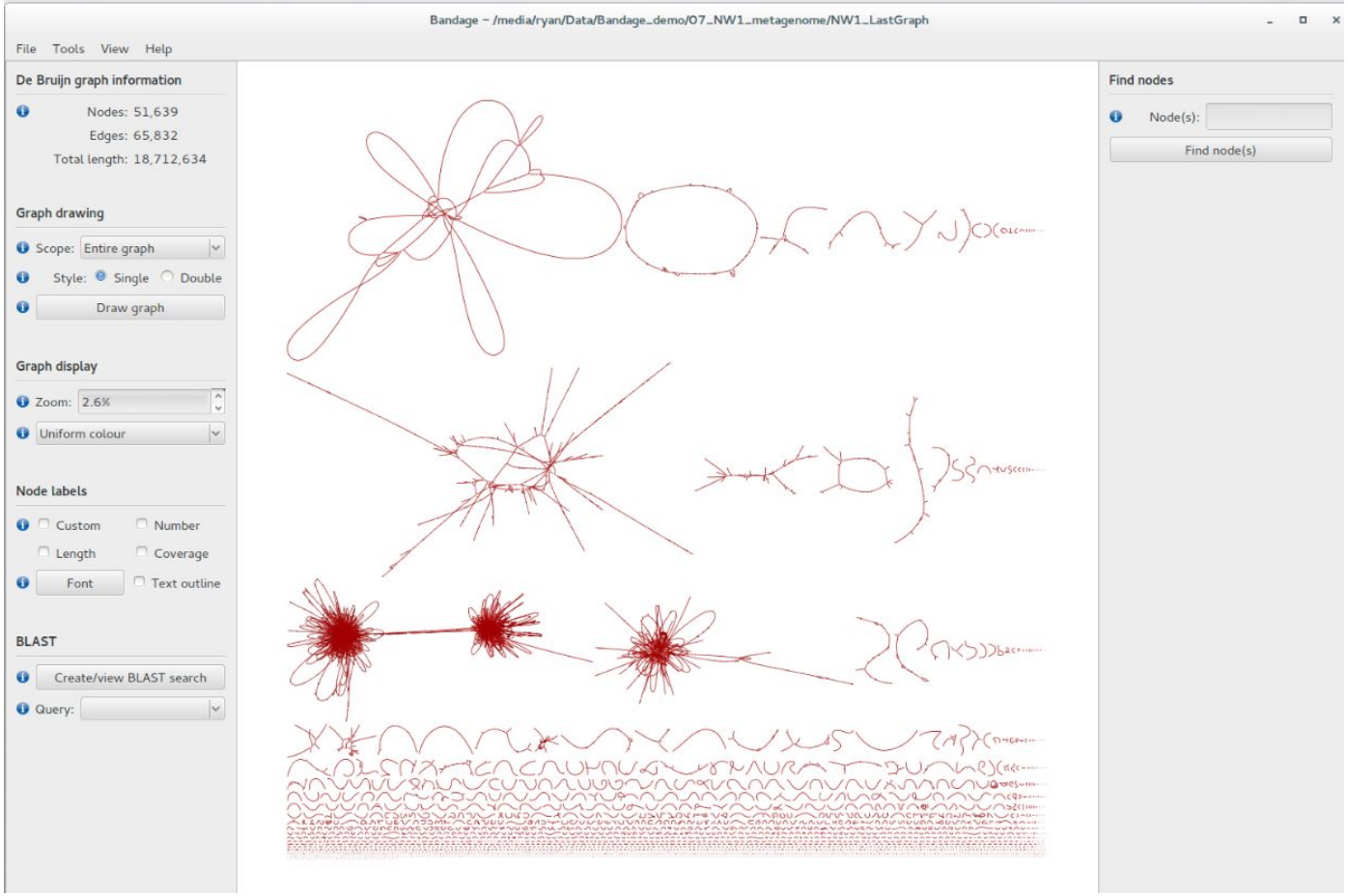


# Actual (simplified) de Bruijn graph

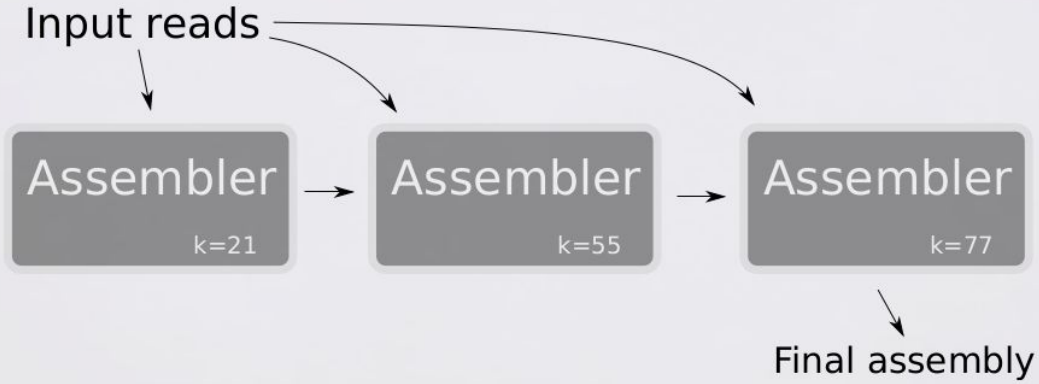


chr14:20Mbp-20.5Mbp GAGE PE reads, **SPAdes** 3.8 k=31: 1k nodes

# Assembly graph visualization: Bandage

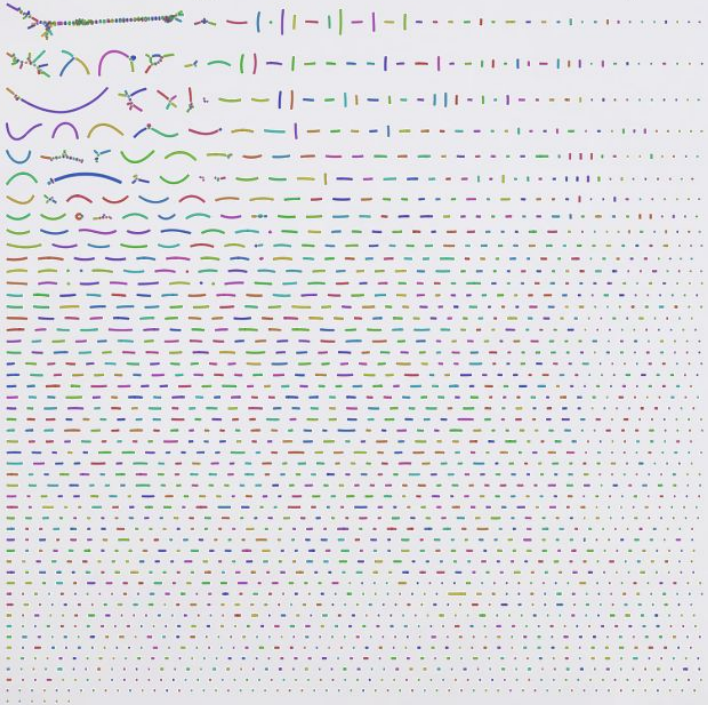


# Multi-k




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

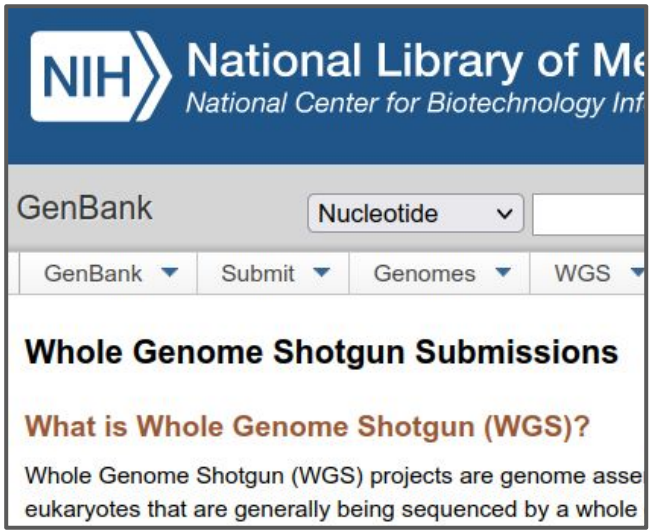
multi k

# Wrap-up assembly part

# Big data in biology: NCBI GenBank & WGS


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

incl. RefSeq
All sequences are *annotated*




**Type:** genome assemblies
**Size:** 16 TB (2022)

*Unannotated*

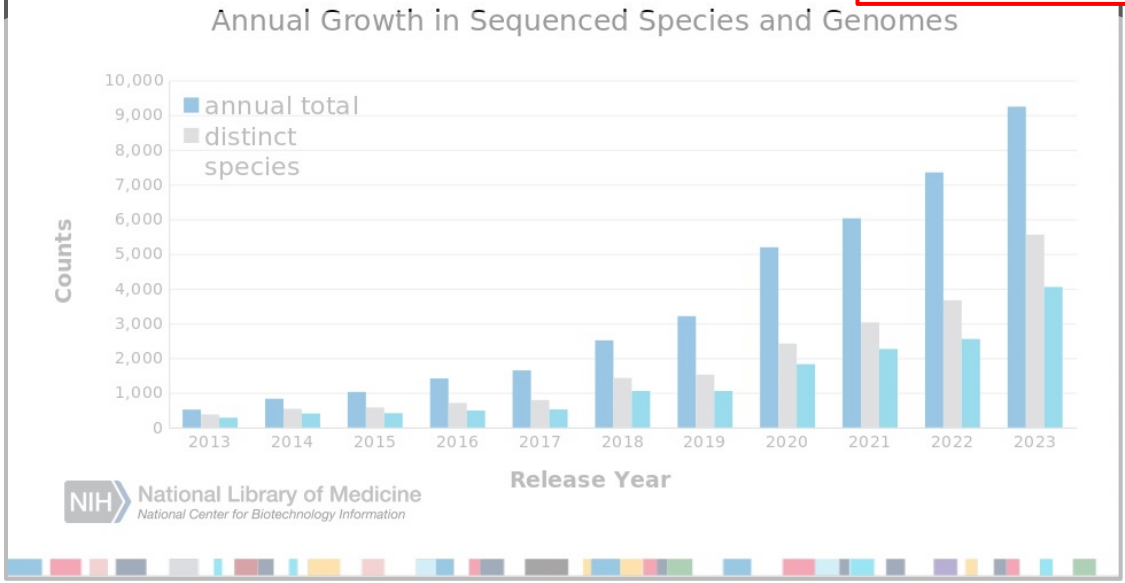
Also big “restricted” data:
UK Biobank, TCGA, ..

# 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

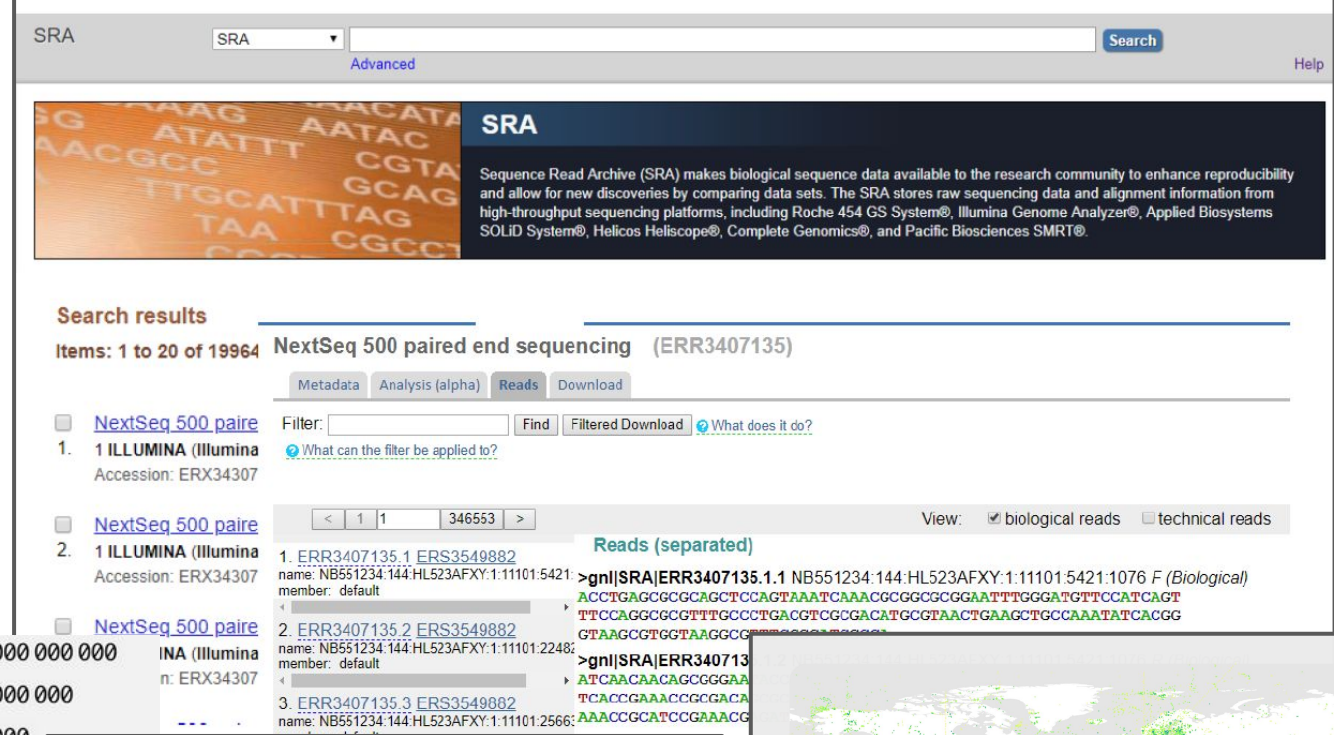
Slide credit: Terence Murphy, NCBI

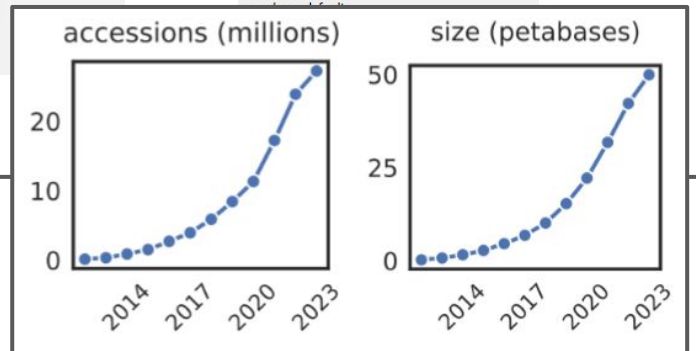
# NCBI SRA

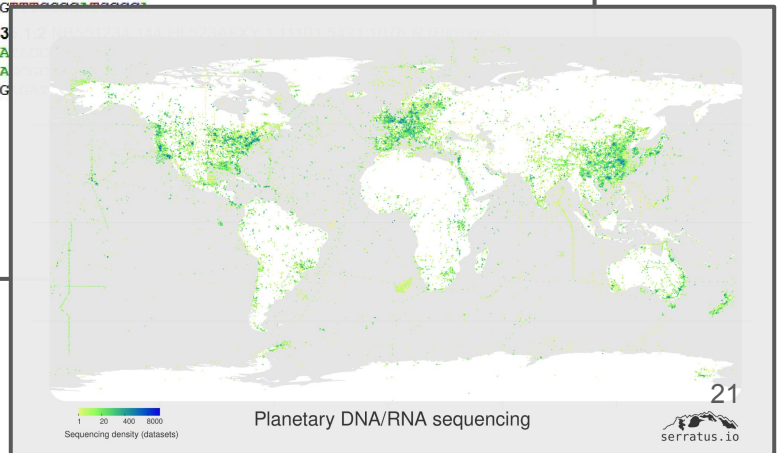
All public sequencing reads

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

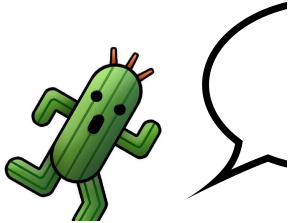
| | | |
|---|---|---|
| peta | [P] | $10^{15}$ = 1 000 000 000 000 000 |
| tera | [T] | $10^{12}$ = 1 000 000 000 000 |
| giga | [G] | $10^{9}$ = 1 000 000 000 |
| mega | [M] | $10^{6}$ = 1 000 000 |



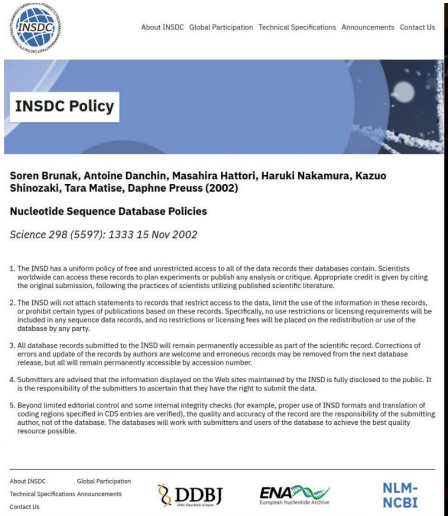



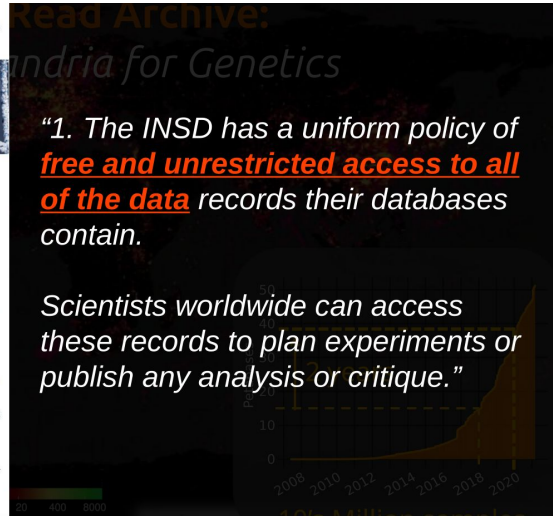
# SRA: Open Science at Its Best

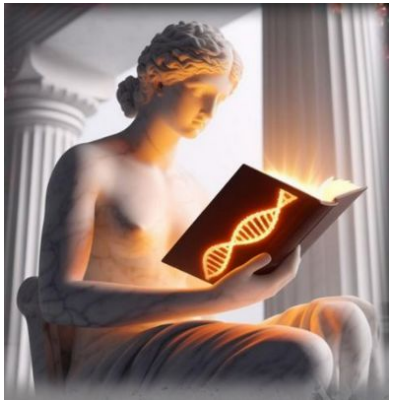


*"Earth's genetic biodiversity is the **shared heritage of all living organisms**, and as scientists we are **responsible for liberating and sharing** this heritage with everyone. Freely, and openly."*

A. Babaian
(Serratus, Logan)



"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

# SRA search tools

## Human/animal DNA/RNA

MetaGraph (Karasikov et al, 2024 preprint)

No online search

## 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)

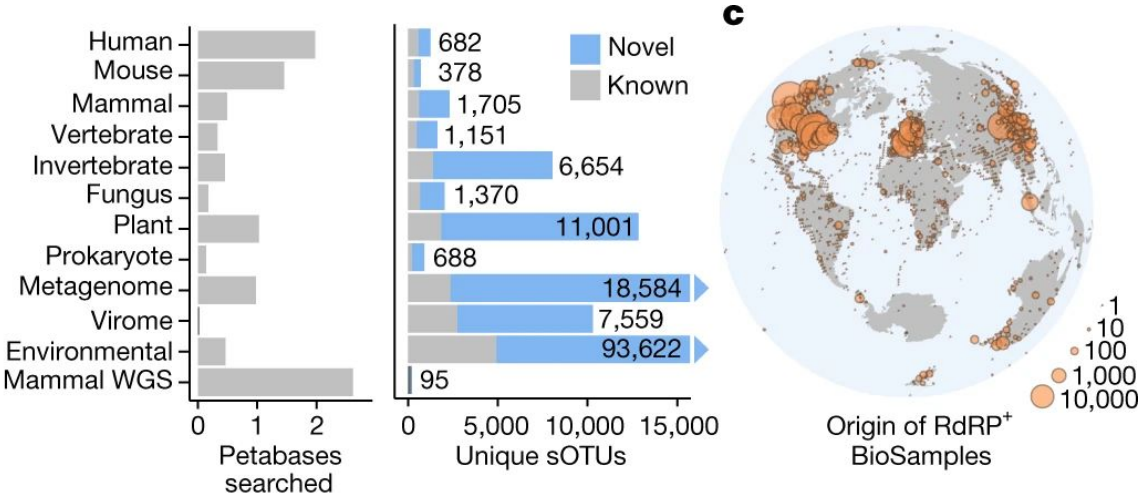
kmindex (Lemane et al, 2013)

PebbleScout (Shiryev et al, 2024)

Limited search features

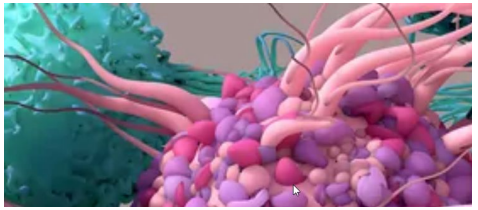
# **Serratus:** all public RNA-seqs analyzed for viral discovery

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 o viruses and viroids.

# State of Sequence Data Archives (2025):


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**

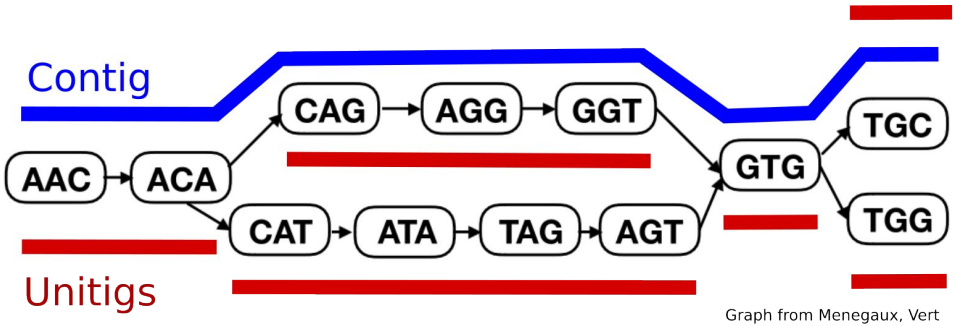
Rayan Chikhi, Brice Raffestin, Anton Korobeynikov, Robert Edgar, Artem Babaian

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

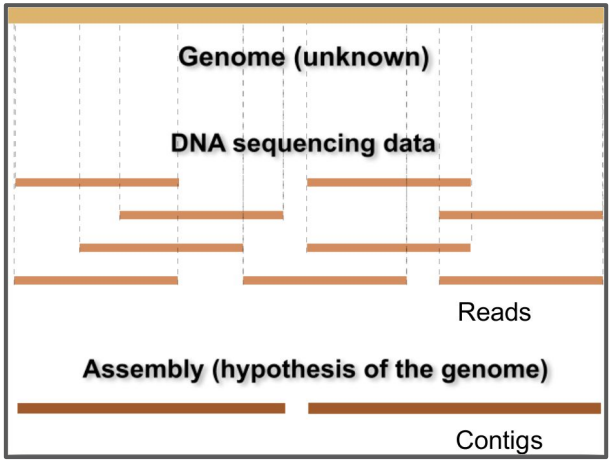
# Unitigs? Contigs?

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

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




Graph from Menegaux, Vert



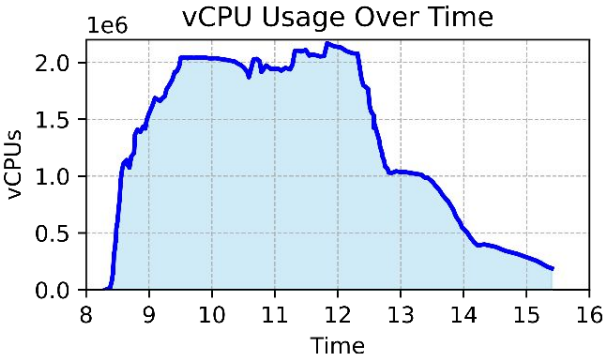

Why unitigs? they keep all variants (SNPs, indels, ..)

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 |

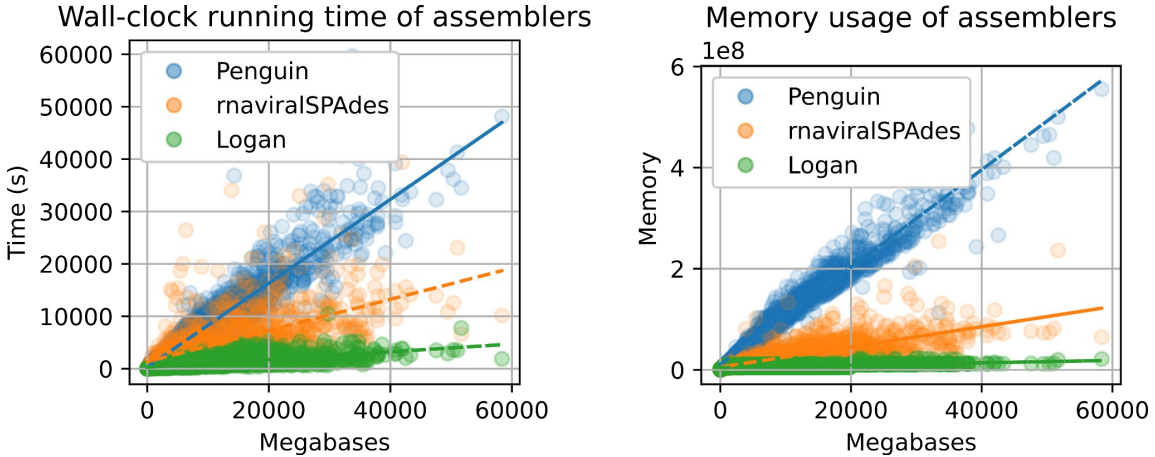



**Run 6 statistics**

| | |
|---|---|
| **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

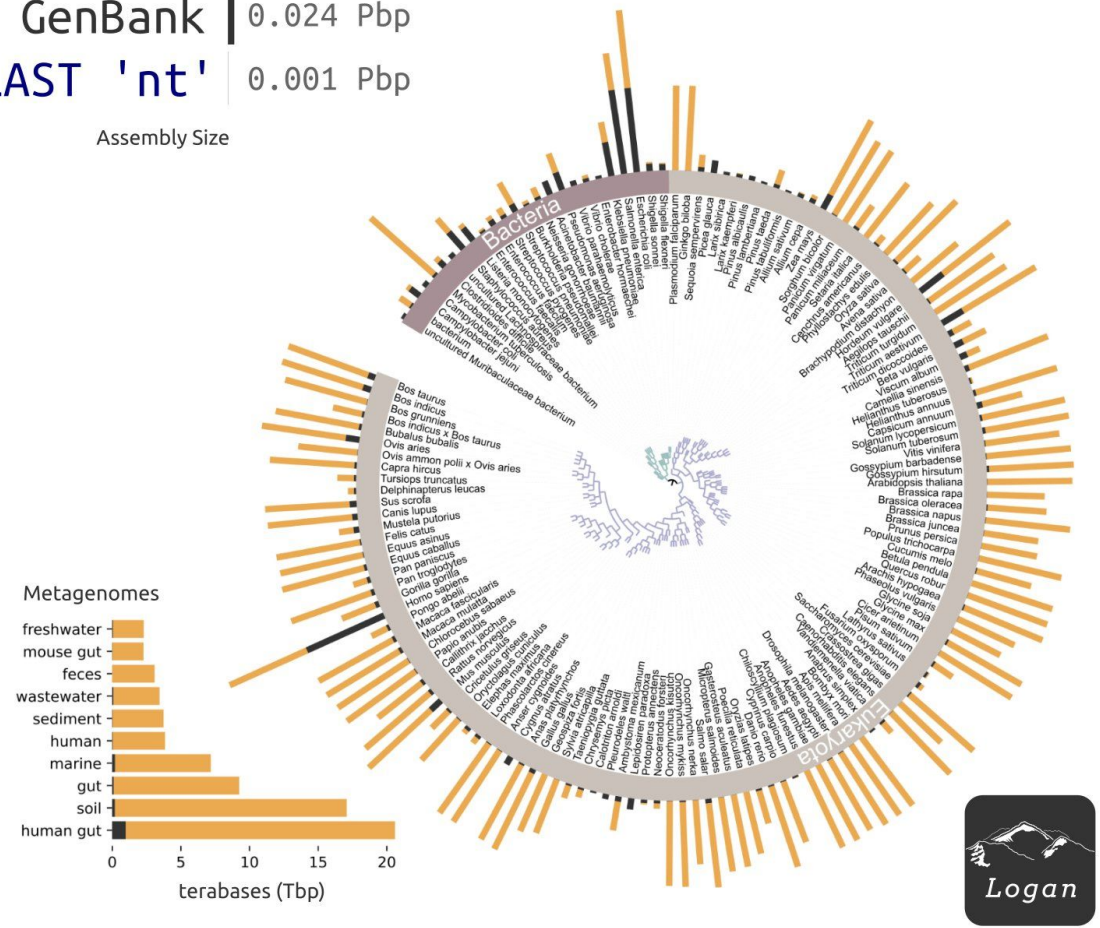
- String algorithms (minimizers in KMC inside cuttlefish2)
- Parallel efficient algorithms (cuttlefish2)
- Minimum perfect hashing (BBHash inside cuttlefish2, Minia)
- Large (billions+ nodes) graph manipulation (Minia)
- Compression (zstd in f2sz)

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?


Logan data

| | |
|---|---|
| Logan | 2 petabases |
| GenBank | 0.024 Pbp |
| BLAST 'nt' | 0.001 Pbp |



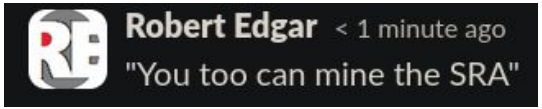
# Accessing Logan

```
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
- But if you need assistance: contact me
- 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


Tools around Logan

# Logan Search: how to use

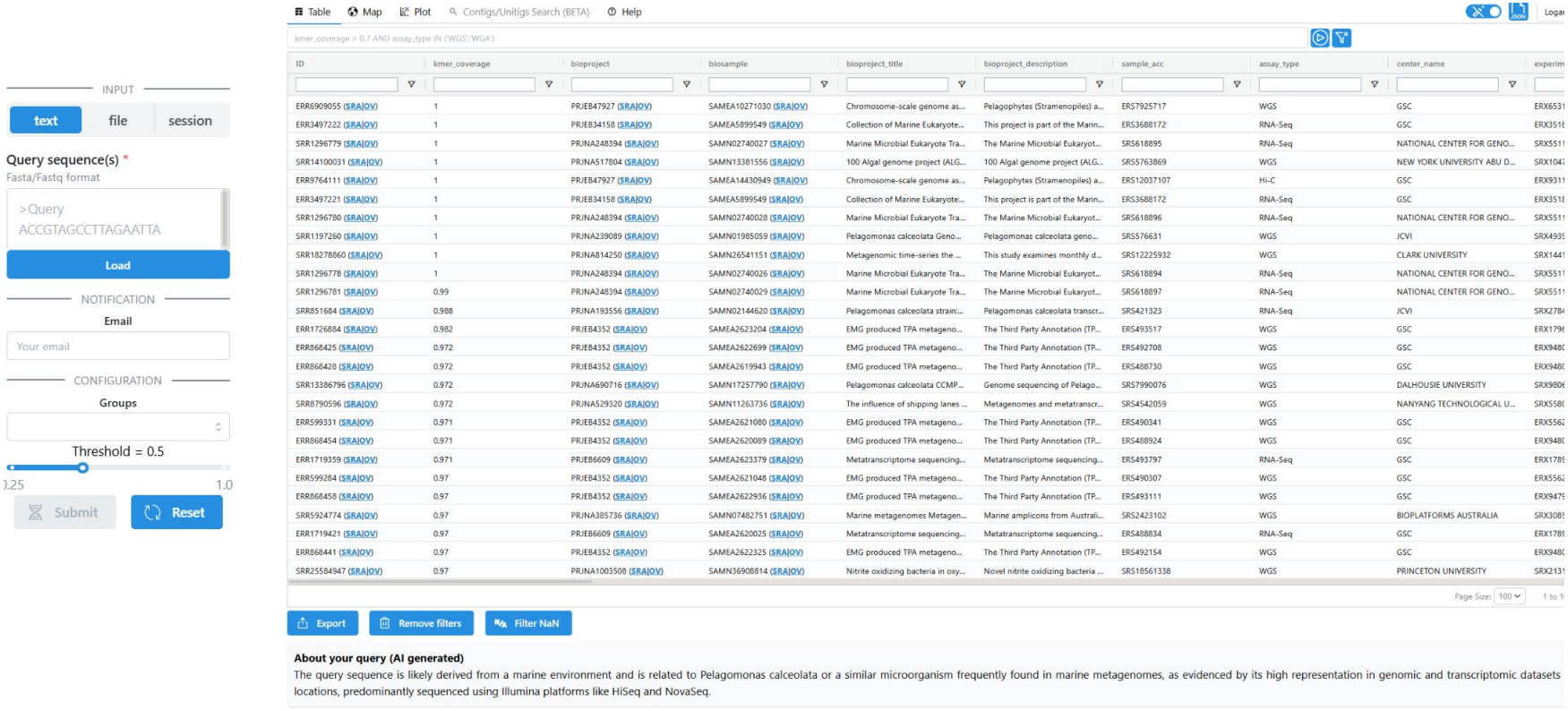
1) Have a query DNA sequence

e.g. `ACTAGTAGAGTAGATGAGGACATAGATGAGCACATGATGAGAGACACATTG`

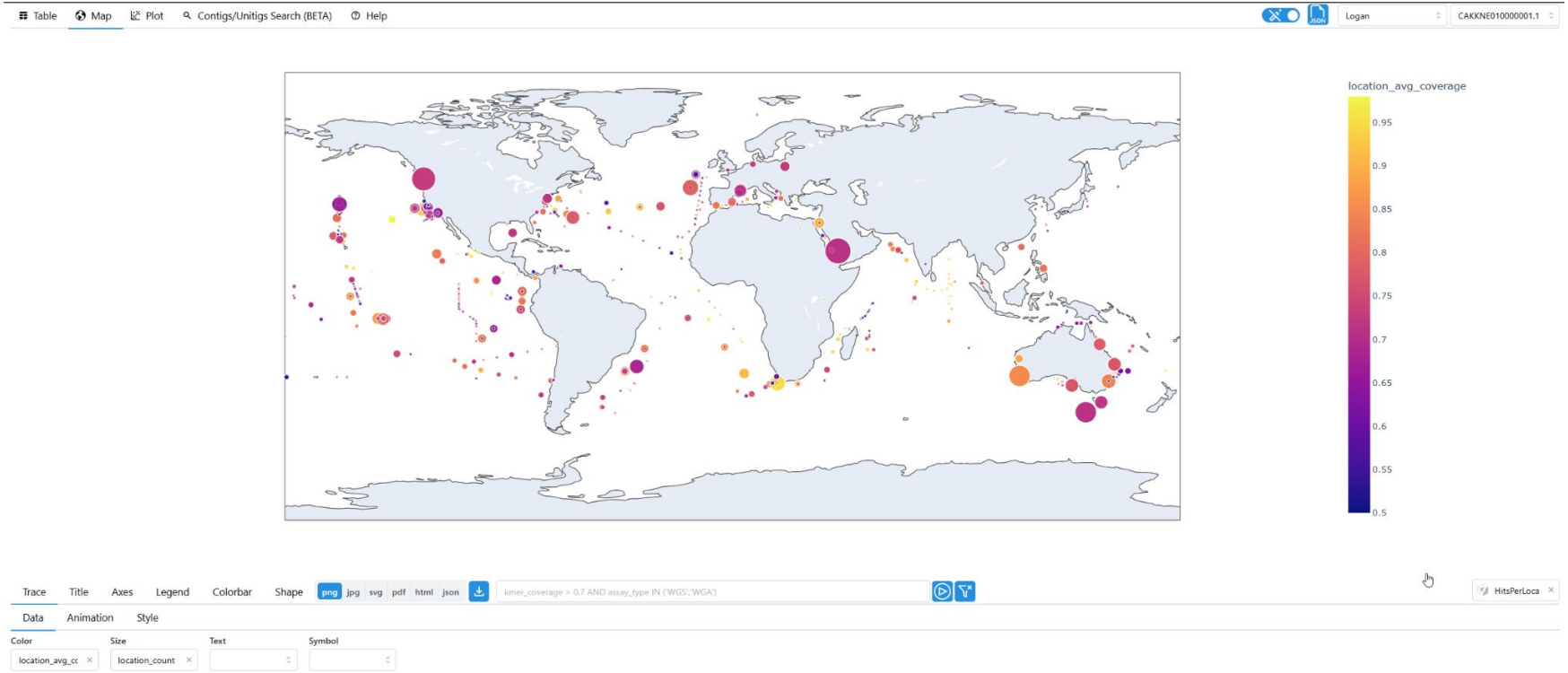
2) Query it in **Logan Search**, wait 2 minutes

3) Explore where this sequence is present across the world/SRA

# Logan Search



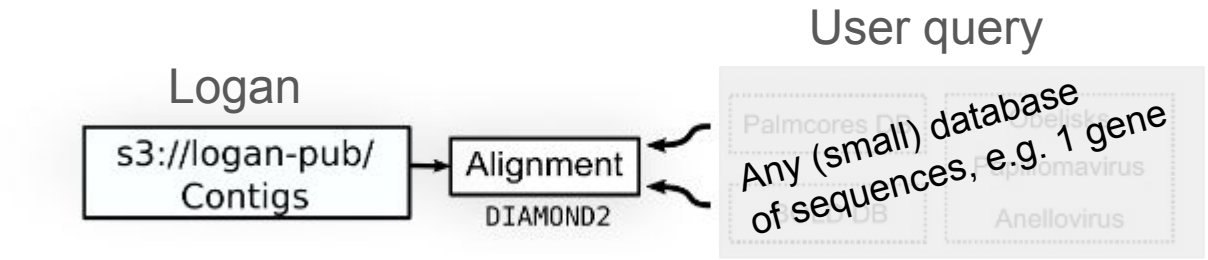
# Logan Search



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


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

# Our typical analysis workflow

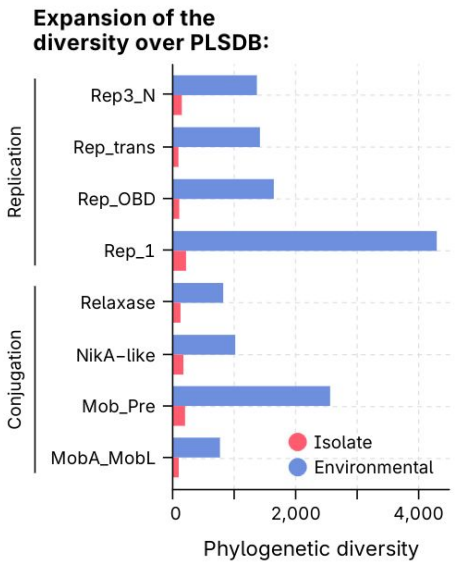



Results: collection of homologs to the query db, found across the entire SRA

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

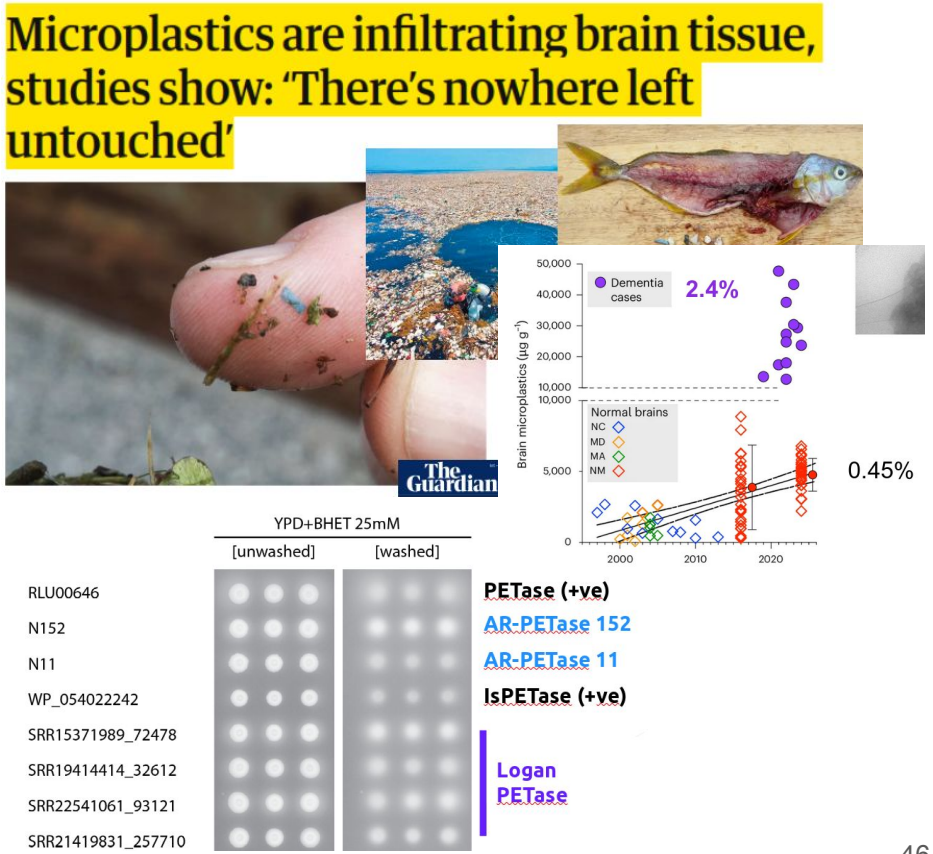
Analyses by:
Martin Steinegger, SNU, Korea
Joey Lee, SNU, Korea

# 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






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$)

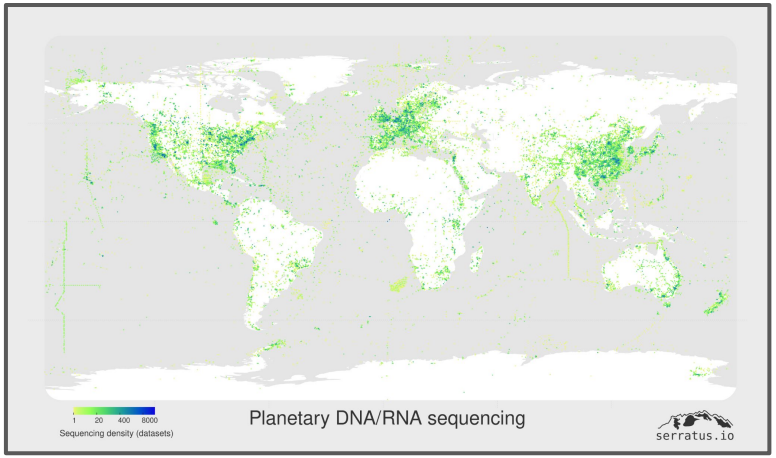
# Conclusion

- **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

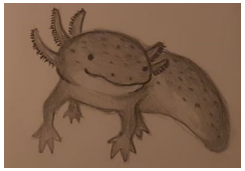



**Generation of new samples**

**High-quality assemblies**

# Sequence Bioinformatics

INSTITUT PASTEUR




Lab members:

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

Alumni:

Luc Blassel
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

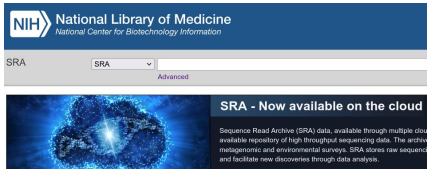
AWS support: Dorian Schaal

cnrs PR[AI]RIE PaRis Artificial Intelligence Research InstitutE erc

# Logan

Funded by ERC CoG "*IndexThePlanet*"

*"A public dataset of all DNA/RNA assembled sequences on Earth (SRA)"*

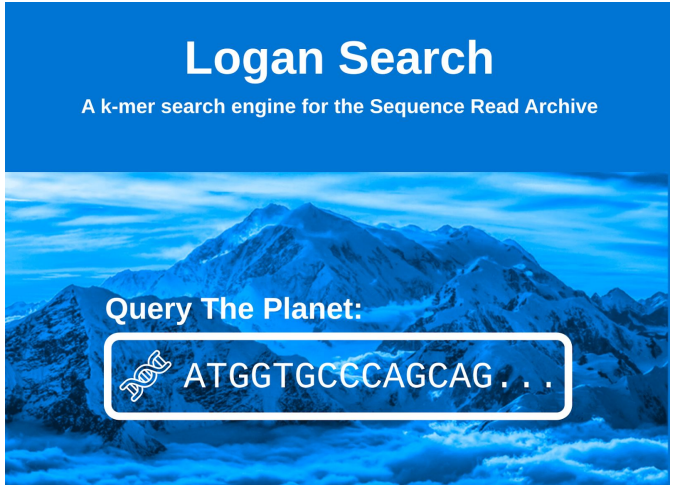


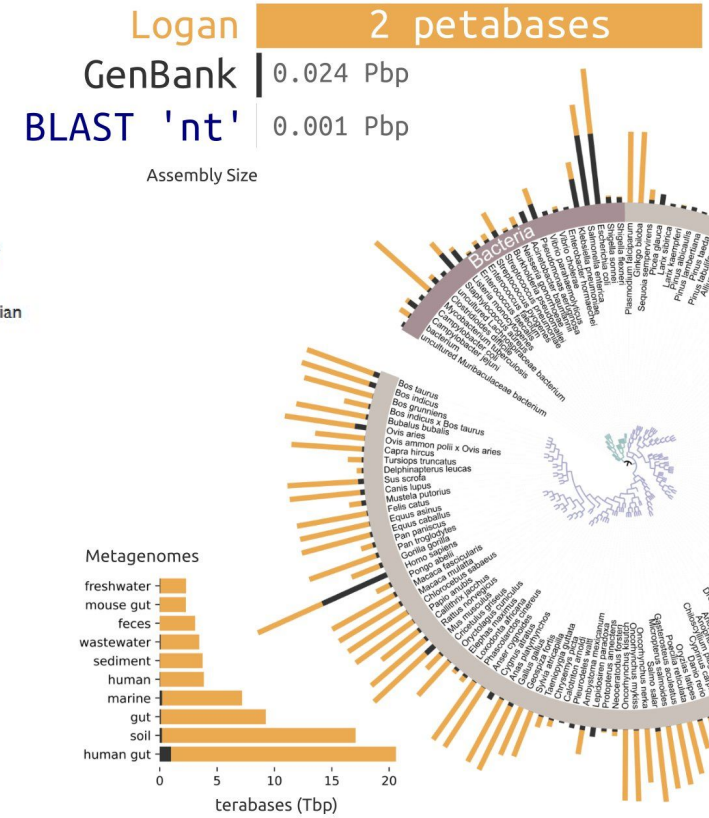

SRA = $10B of direct seq costs

**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




Logan

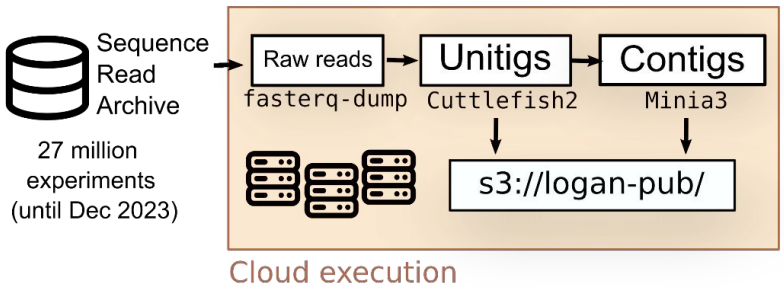
# Future directions

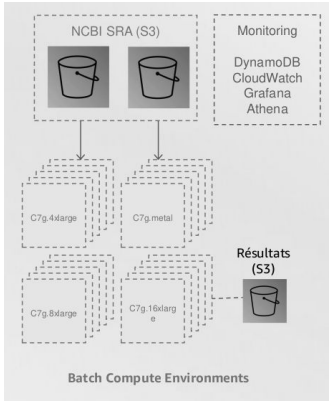
- LexicMap
- Training embeddings on protein clusters for search
- Integration with Galaxy
- Publish the paper :)
- Updates to Logan? maybe, for 1-2 years of new sequencing data

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




Batch
S3
DynamoDB
Athena
CloudFormation
CloudWatch
Cost Explorer
Grafana

- **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