



Big data

Rayan Chikhi
Institut Pasteur
Workshop on Genomics 2025



High expectations from last year (and the year before).
This won't be the greatest big data talk, just a tribute of a tribute

Founding members of biological big data

Early Eras of Bioinformatics, Representative Leaders

- » Generation -1: E.O. Wilson (compatibility aka perfect-phylogeny - 1965)
- » Generation 0: Margret Dayhoff, Russ Doolittle, Joe Felsenstein
- » Generation 1: Mike Waterman, David Sankoff (Era of algorithms, pre-data)
- » Generation 2: Gene Myers, Russ Altman, Richard Durbin, Sean Eddy

Dayhoff-Eck

- » Worked out the theoretical basis of "shotgun-sequencing" of protein (1970)
- » Published the first "Atlas of protein sequence and structure" (1966) with 65 sequences. Really the first comprehensive database in bioinformatics. Continued with several additional editions.

technologies to support advances in biology and medicine, most notably the creation of protein and nucleic acid databases and tools to interrogate the databases. She originated one of the first [substitution matrices](#), [point accepted mutations \(PAM\)](#). The [one-letter code](#) used for amino acids was developed by her, reflecting an attempt to reduce the size of the data files used to describe amino acid sequences in an era of punch-card computing.

Margaret Oakley Dayhoff



Born

Margaret Belle Oakley
March 11, 1925
[Philadelphia,](#)
[Pennsylvania](#)

Died

February 5, 1983

technologies to support advances in biology and medicine, most notably the creation of protein and nucleic acid databases and tools to interrogate the databases. She originated one of the first [substitution matrices](#), [point accepted mutations](#) (PAM). The [one-letter code](#) used for amino acids was developed by her, reflecting an attempt to reduce the size of the data files used to describe amino acid sequences in an era of punch-card computing.

Margaret Oakley Dayhoff



The first big data bioinformatician

Born Margaret Belle Oakley
March 11, 1925
[Philadelphia, Pennsylvania](#)

Died February 5, 1982

Professor of Chemistry and Chemical Biology at [Rutgers University](#) and a former director of the RCSB Protein Data Bank (one of the member organizations of the [Worldwide Protein Data Bank](#)). A structural biologist, her work includes structural analysis of protein-nucleic acid complexes, and the role of water in molecular interactions. She is also the founder and director of the Nucleic Acid Database, and led the Protein Structure Initiative Structural Genomics Knowledgebase.^{[1][2][3]}

Helen Berman



Helen Berman in 2008.

Born Helen Miriam Berman
1943 (age 81–82)
Chicago, Illinois

1972: single gene sequenced

2000: 1 high-quality human genome

2013: many low-quality human genomes

2021: 10 petabases of reads analyzed

2022: 1 million humans VCFs

2022: 50 high-quality human genomes

2024-: ?



Is big data just a *technical* matter?!

Information technologies scale exponentially

Sydney Brenner and Nathan Myhrvold, ~2005

		Base pairs
1995	Bacterium	2×10^6
2000/3	Mammal	3×10^9
2013	2500 humans	7.5×10^{12}
2021	~1M genomes	3×10^{15}

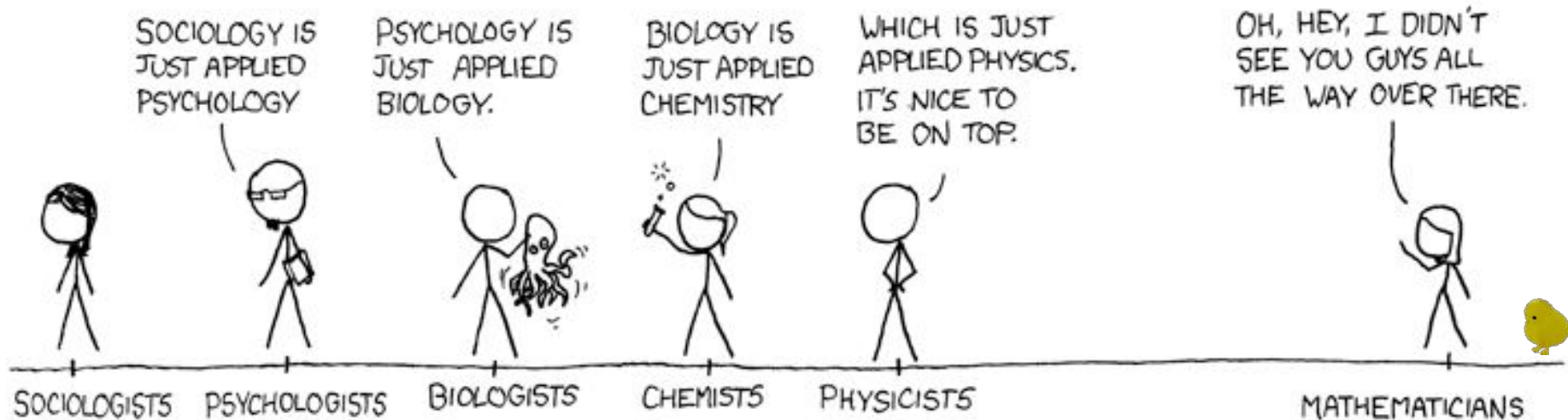
Cost drop from \$1/bp to \$10⁻⁷/bp

- Sustained increase in data at more than 2-fold per year over two decades
- Faster than Moore's law implies continual demand for computational improvements
- Interplay between
 - Analysis and understanding of gene function
 - Improved computational and mathematical methods
 - Evolutionary models

DNA sequence, genomes and computation together
Informatics is to biology what mathematics is to physics ?

*“Informatics is to biology,
what mathematics is to physics”*

Richard Durbin, RECOMB 2023 keynote



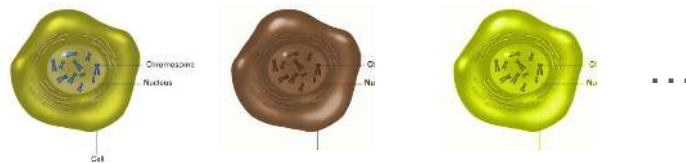
Informatics?

→
"purity"
←
"usefulness"

Types of genomic data

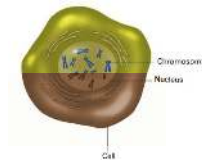
- **Raw sequencing data**

- Error-prone ($\sim 1\text{-}10\%$ per base)
- Abundant (petabytes)
- Contains inter-cell diversity



- **Reconstructed genomes**

- High quality ($< 0.001\%$ per base)
- Rare (gigabytes)
- Collapses inter-cell diversity

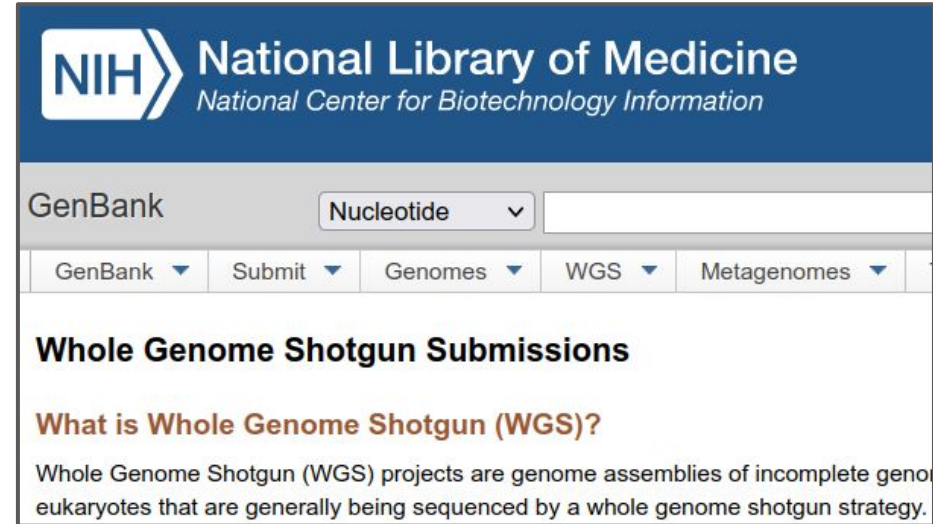


Big data in biology: NCBI GenBank & WGS



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

All sequences are *annotated*



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

Unannotated

How complete are those databases?

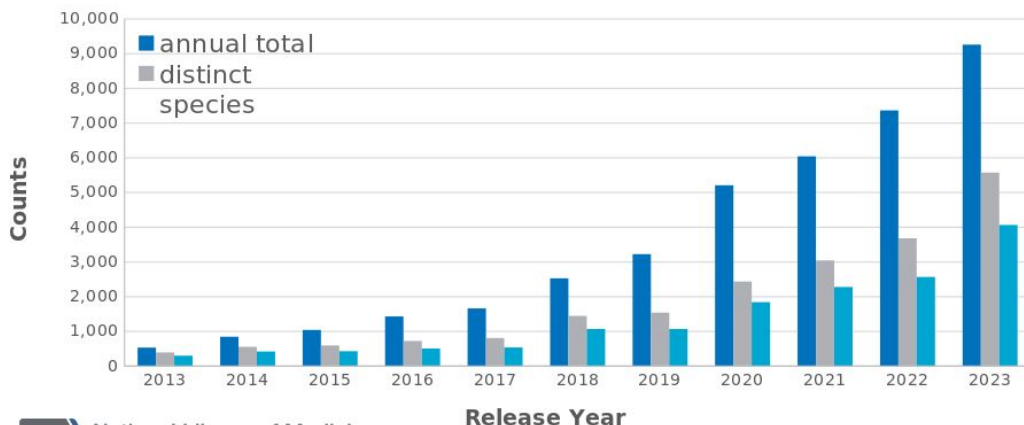
ALL EUKARYOTIC GENOMES (Cumulative: Dec 2023):

GenBank genomes (all): 36,593 (15,453 species)

GenBank (with annotation): 6,817 (3,801 species)

(Out of 8 million known species..)

Annual Growth in Sequenced Species and Genomes



NIH National Library of Medicine
National Center for Biotechnology Information

GenBank eukaryotic genome submissions (2021):

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

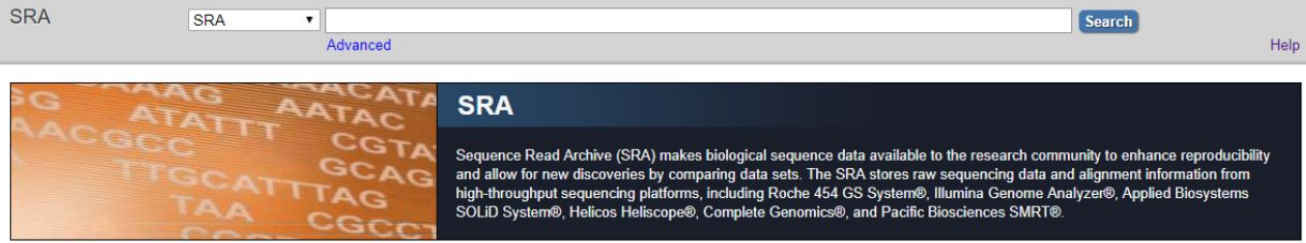
NCBI

NCBI SRA

All public
sequencing reads

Size: 50 Pbases
as of Dec 2023

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



Search results

Items: 1 to 20 of 19964

NextSeq 500 paired end sequencing (ERR3407135)

Metadata Analysis (alpha) Reads Download

☐ [NextSeq 500 paire](#)

1. 1 ILLUMINA (Illumina)
Accession: ERX34307

☐ [NextSeq 500 paire](#)

2. 1 ILLUMINA (Illumina)
Accession: ERX34307

☐ [NextSeq 500 paire](#)

INA (Illumina)
n: ERX34307

Filter: Find Filtered Download [What does it do?](#)

[What can the filter be applied to?](#)

< 1 1 346553 >

View: ☒ biological reads ☐ technical reads

Reads (separated)

1. ERR3407135.1 ERS3549882

name: NB551234.144.HL523AFX.1.11101.5421
member: default

2. ERR3407135.2 ERS3549882

name: NB551234.144.HL523AFX.1.11101.2248
member: default

3. ERR3407135.3 ERS3549882

name: NB551234.144.HL523AFX.1.11101.2566

>gnl|SRA|ERR3407135.1.1 NB551234.144:HL523AFX.1:11101.5421:1076 F (Biological)

ACCTGAGCGCGCAGCTCCAGTAAATCAAACGCGCGCGGAATTGGGATGTTCCATCAGT

TTCCAGGCGCGTTGCCCTGACGTCGCGACATGCGTAAGCTGCCAAATATCACGG

GTAAAGCGTGGTAAGCGCG

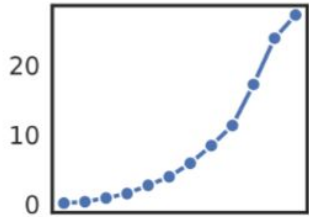
>gnl|SRA|ERR340713

ATCAACAACAGCGGGAA

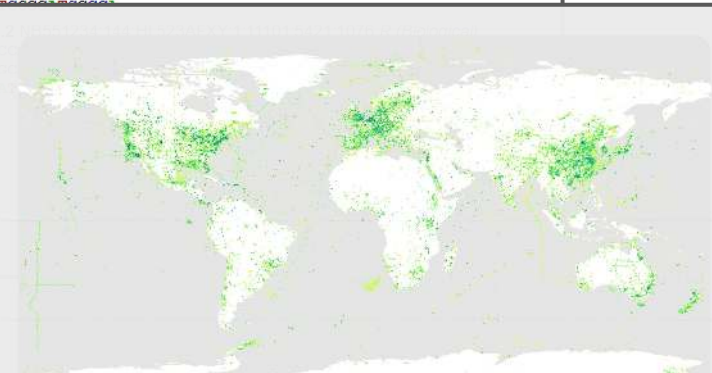
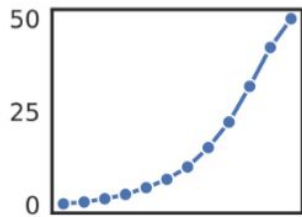
TACCGCAACCGCGACA

AAACCGCATCCGAAACG

accessions (millions)



size (petabytes)



Planetary DNA/RNA sequencing

11

Public sequence datasets

50 Pb



SRA

24 Tb

NCBI WGS (2023)

2.5 Tb

GenBank (2023)

283 GB

BLAST nt

Units

yotta [Y] $10^{24} = 1\,000\,000\,000\,000\,000\,000\,000\,000$

zetta [Z] $10^{21} = 1\,000\,000\,000\,000\,000\,000\,000\,000$

exa [E] $10^{18} = 1\,000\,000\,000\,000\,000\,000\,000$

peta [P] $10^{15} = 1\,000\,000\,000\,000\,000\,000$

tera [T] $10^{12} = 1\,000\,000\,000\,000\,000$

giga [G] $10^9 = 1\,000\,000\,000$

mega [M] $10^6 = 1\,000\,000$

kilo [k] $10^3 = 1\,000$

hecto [h] $10^2 = 100$

deca [da] $10^1 = 10$

(Youtube: 300 PB)

NCBI SRA database: 50 PB



Institut Pasteur: 10 PB



Your laptop: 0.001 PB



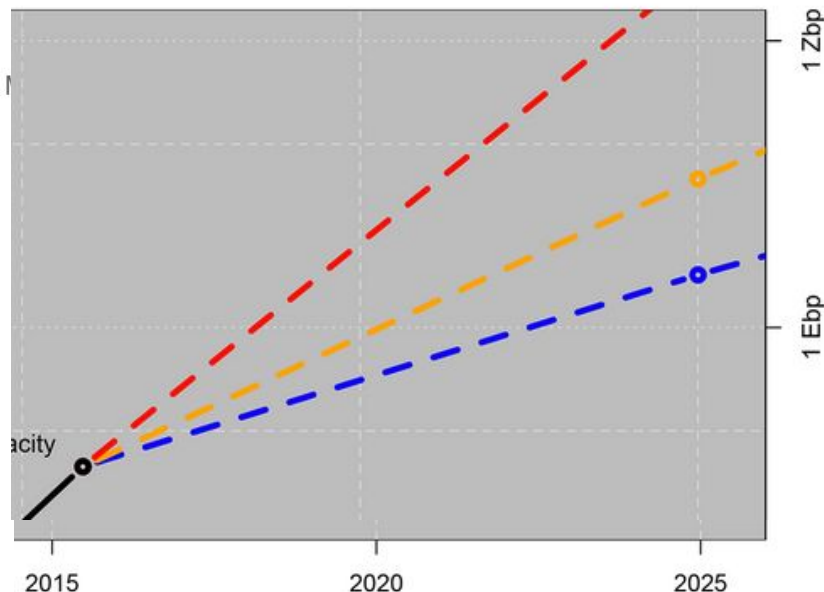


Big Data: Astronomical or Genomical?

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

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

- Projected 5 exabytes - 1 zettabyte of seq data in 2025
- Actual: 5 exabytes based # sequencers in world (total capacity: 45,000,000 human genomes per year)



<https://docs.google.com/spreadsheets/d/1GMMfhyLK0-q8XkIo3YxIWaZA5vVMuhU1kg41g4xLkXc/htmlview>

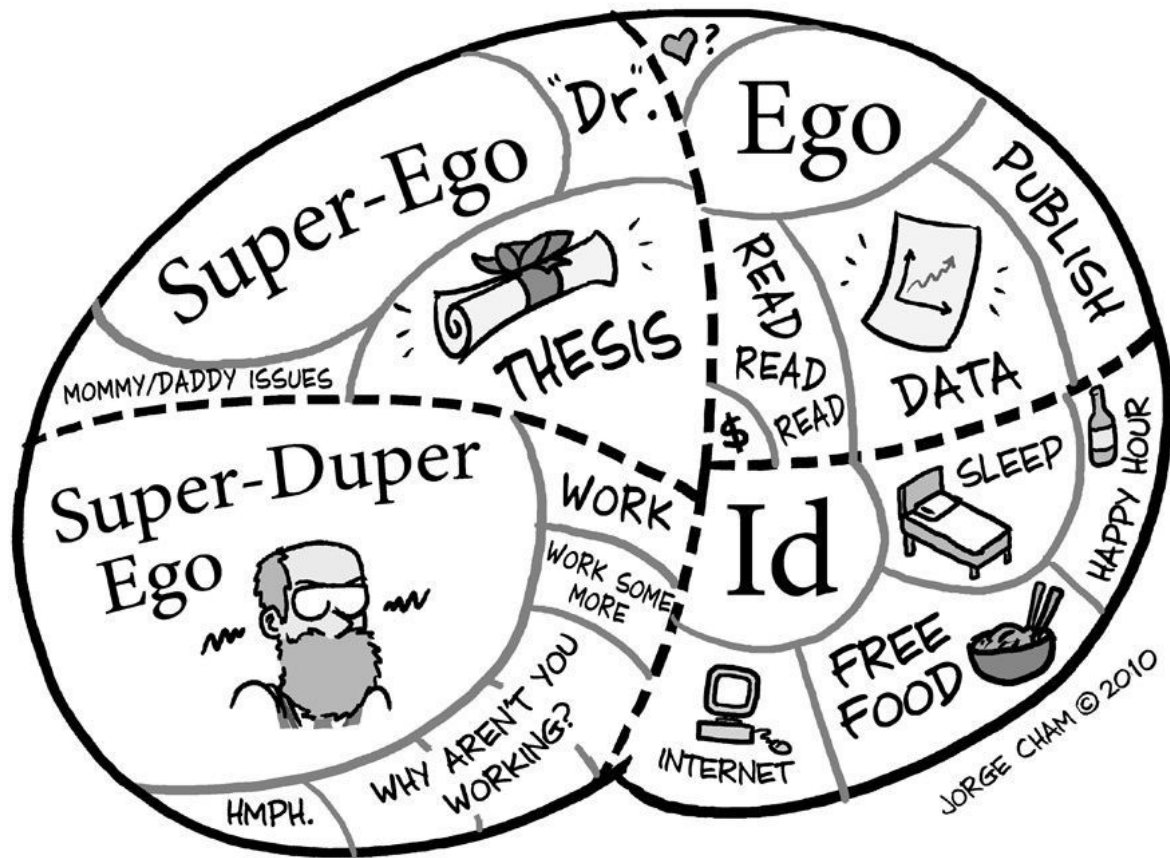
State of Data Archives (2025):



For the last 2 weeks, the Workshop on Genomics has given you access, & asked you use, an infinitely valuable resource and perhaps you did not even notice it.

I know what you're thinking

(because I've been there)



1st year PhD: *“Is my project any good?”*

2nd year PhD: *“What am I even doing?”*

3rd year PhD: *“I’d give anything to not write
this thesis”*

Postdoc:



*> No time to
learn new things*

This past week you have been using

- limitless^{*} computation
- &
- super fast^{*} access to data

^{*} but, limited by Guy





With big data and big computers, one could perform wonderful, ground-breaking genomics



... But how?

Part 2: Big Data Toolbox

Computation

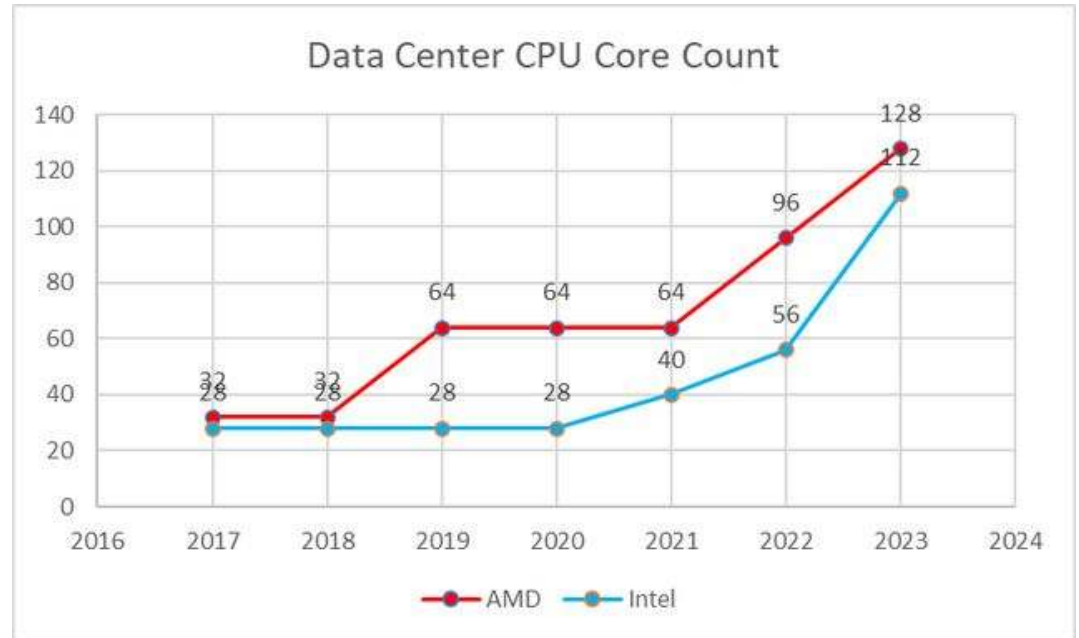
- Big computers, Cluster, Cloud
- Storage management
- Galaxy
- Knowledge of scaling limits
- Knowledge of cloud costs
- Parallel execution
- AI

Data mining

- Pebblescout, branchwater
- ORA
- deCOM
- SRA metadata

Future genomics, today?

'big data'
=
'small computers'



University Cluster

Acquire knowledge about it:

- Queues:
 - How many CPUs/RAM per job, what timelimit
 - Can your group access any ✨*special* queue✨
- Storage:
 - Your quota
 - Is “scratch” quota-free? Do files expire?



My scripts:

```
srun -q seqbio -p seqbio --mem 100G -c 10 --pty bash
```

Quickly allocates a terminal on any machine

```
squeue -o "%.18i %.9P %.8j %.8u %.2t %.10M %.6D %R cores:%c mem:%m cmd:%o " | grep seqbio
```

See what machines are currently being used

Storage management

- How to never run out of storage space:
 - Have 2 folders:
 - ~/archive
 - ~/scratch
 - Rules:
 - Archive = backed up command lines and final results
 - Scratch = fast, may be deleted at any time
 - Keep the list of files for both, somewhere
 - Keep a dummy 100 GB file ready to be deleted?
- Data compression
 - BAM => CRAM => delete it
 - FASTQ => gzip => delete it
 - VCF => BCF
 - GFF/GTF => don't annotate

Galaxy Project



Data Intensive *analysis* for everyone

- Versatile and reproducible workflows
- **Web** platform
- **Open source** under [Academic Free License](#)

- If you do not have a cluster
- ..or the will to install tools..
- Galaxy offers free computation on pre-installed workflows

Main Galaxy interface

The screenshot displays the Galaxy web interface in a browser window. The interface is divided into several sections:

- Tools Panel (Left):** A sidebar with a search bar and a list of tool categories including 'Get Data', 'GENERAL TEXT TOOLS', 'GENOMIC FILE MANIPULATION', 'COMMON GENOMICS TOOLS', and 'Interactive tools'.
- Main Content Area (Center):** Features a large banner for the 'James P. Taylor Foundation for Open Science' with a video thumbnail and a 'Learn More' button. Below the banner is a notification about SARS-CoV-2 data analysis and logos for partner institutions like PennState, Johns Hopkins, and Oregon Health & Science University.
- History Panel (Right):** A sidebar titled 'History' showing a list of recent datasets, including 'Galaxy 101 History' and '2: SNPs'.

Cloud

= A collection of computers owned by a single organization and accessible from the Internet



Part 2: Big Data Toolbox

Computation

- Big computers, Cluster, Cloud
- Storage management
- Galaxy
- Knowledge of scaling limits
- Knowledge of cloud costs
- Parallel execution
- AI

Data mining

- Pebblescout, branchwater
- ORA
- deCOM
- SRA metadata

Knowledge of scaling limits

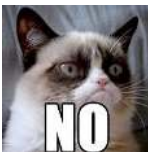
In order of difficulty:

1. **Estimate** how long an analysis will take
2. Reasons **why** some analyses are slower than expected
3. **How** to reduce that time

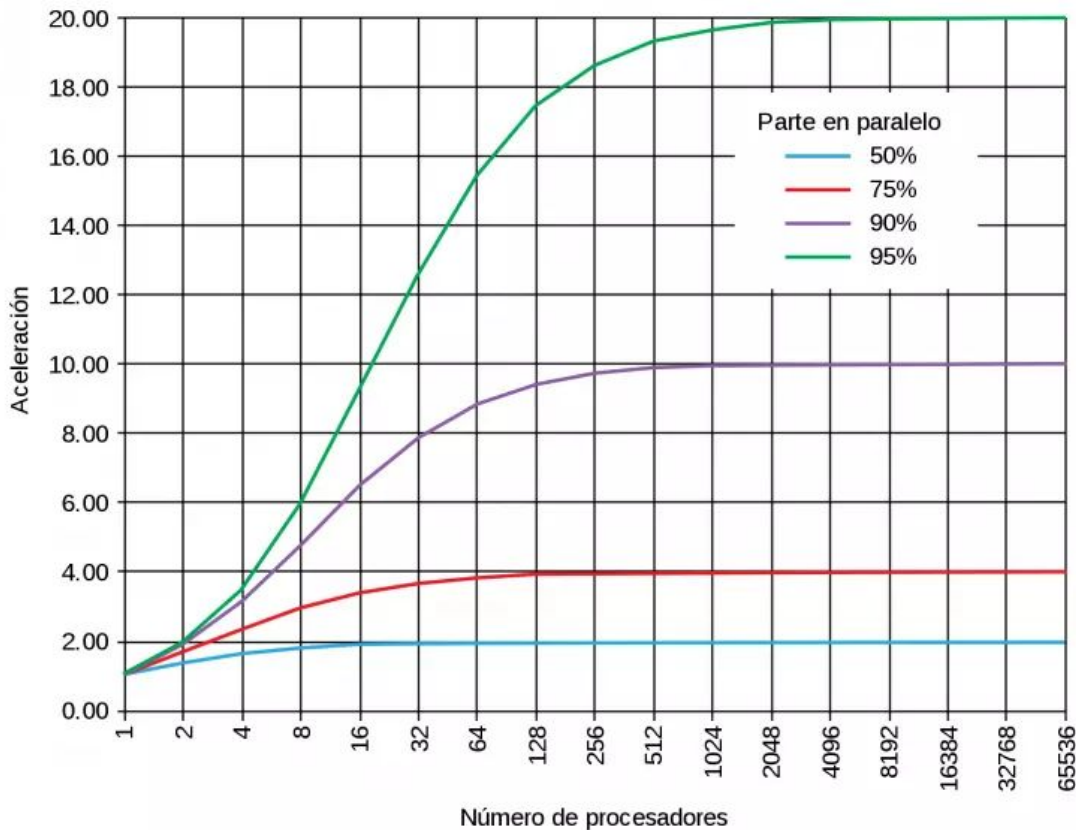
Do 200 CPUs (or threads) always go 200x faster?



Amdahl's law:



Ley de Amdahl

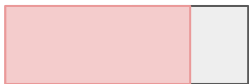


1 core

non-parallel
work

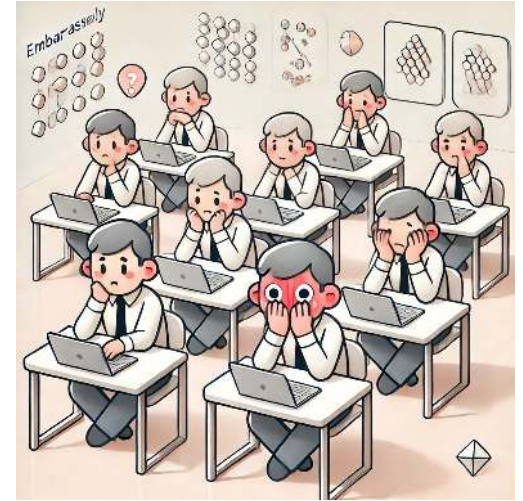
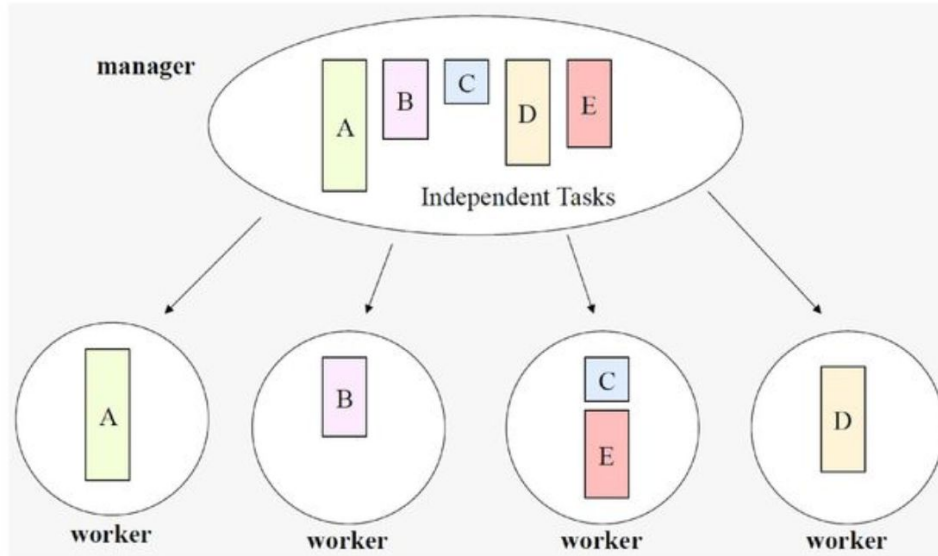
parallel work

8 cores



Except..

.. if you have an **embarrassingly parallel** problem.
i.e. composed of *independent tasks*



Examples of embarrassingly parallel problems

- Alignment of N different sequences to a reference genome
- Annotation of N different genomes
- Assembly of N different samples

Examples of problems NOT embarrassingly parallel :

- An entire bioinformatics pipeline (e.g. alignment->variant calling->annotation of variants)
- Assembly of a single sample
- Alignment of a single sequence

How to run things in parallel!

- Single machine, many threads
- Many machines, by hand
- GNU parallel
- bash tricks
- SLURM (cluster tools)
- Cloud infrastructure

GNU parallel



Allows to run the same task on multiple files, simultaneously.

To count number of lines across many FASTQ files:

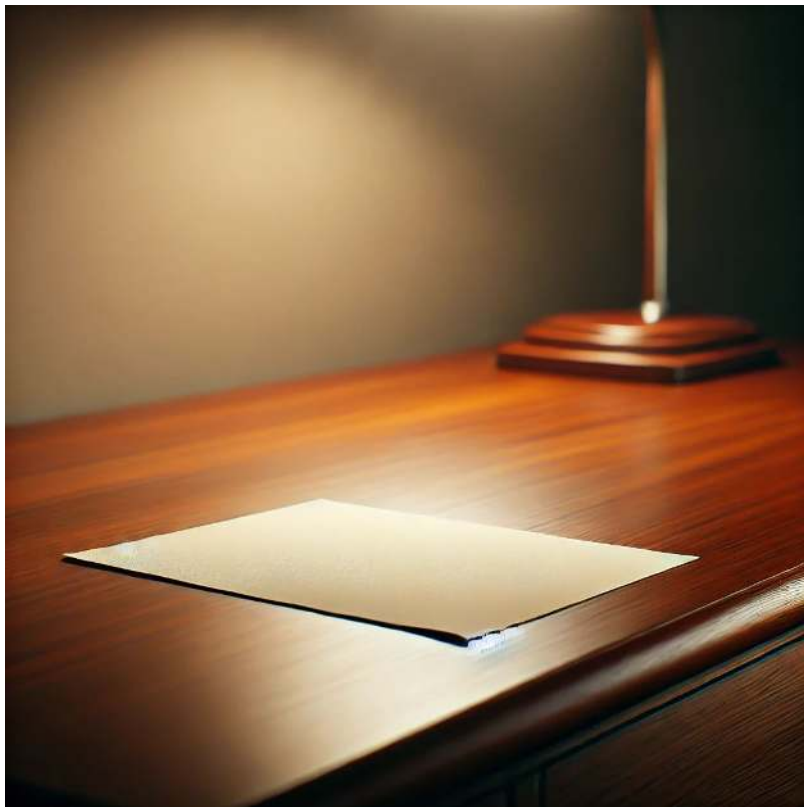
```
find . -name *.fastq | parallel -j10 "wc -l {} > {}.nb_lines"
```

To run many jobs defined by CSV data:

```
cat data.csv | parallel --colsep ',' "./myprogram {1} {2}"
```

(these are examples of embarrassingly parallel tasks)

Bash parallel tricks



Connect the dots from left to right

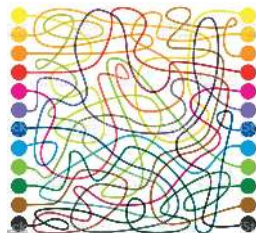
1) Access data from a SSD disk

2) Access data in memory

3) Access <http://www.evomics.org> in Australia

4) Human cell cycle

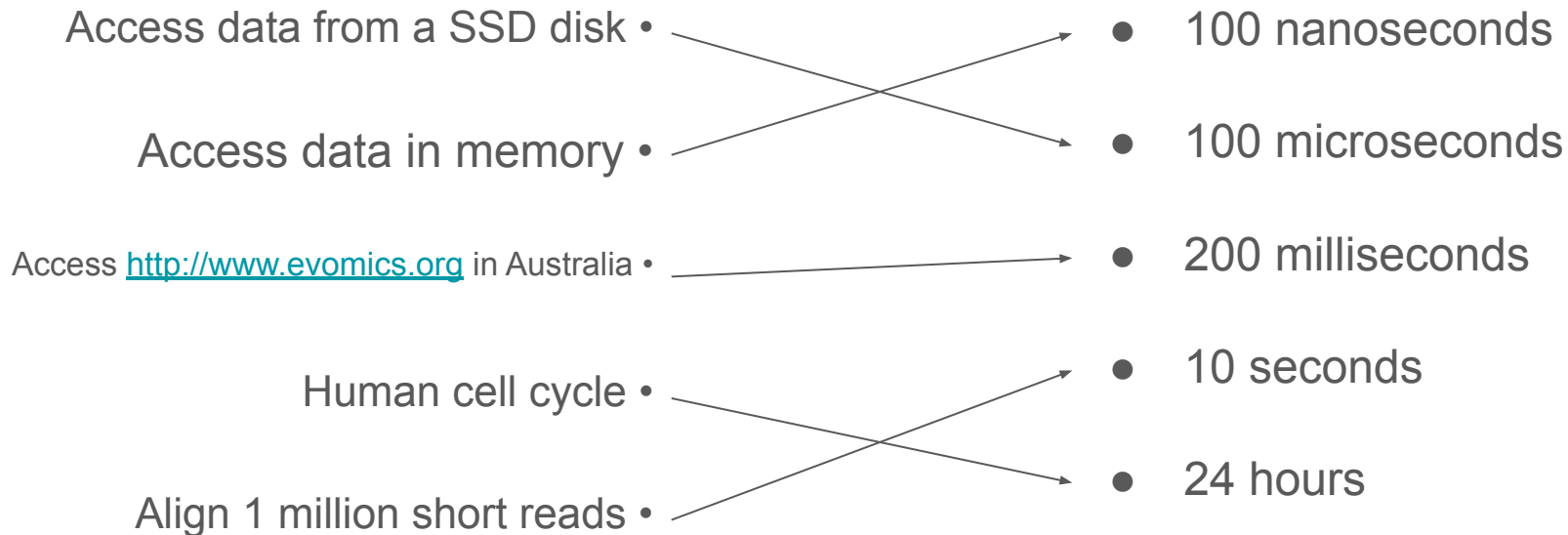
5) Align 1 million short reads



- 100 nanoseconds
- 100 microseconds
- 200 milliseconds
- 10 seconds
- 24 hours

n	nano	10^{-9}
μ	micro	10^{-6}
m	milli	10^{-3}

Connect the dots from left to right



n	nano	10^{-9}
μ	micro	10^{-6}
m	milli	10^{-3}

Knowledge of scaling limits

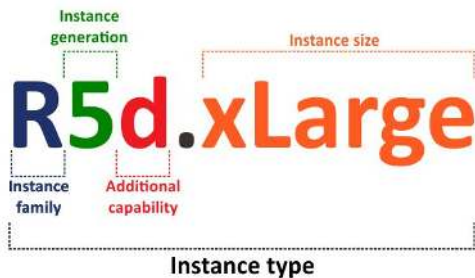
In order of difficulty:

1. **Estimate** how long an analysis will take
 - Look at performance table in tool paper
 - Try on smaller data and extrapolate
2. **Reasons why** some analyses are slower than expected
 - Limited number of CPUs
 - Limited RAM
 - Slow disk (HDD < Cluster network drives < SSD < NVMe)
3. **How** to reduce that time
 - Most analyses go fast enough on a big cloud/cluster and the right tools

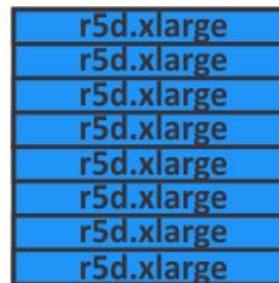
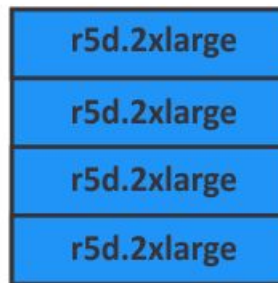
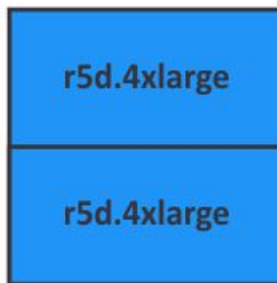
Knowledge of cloud costs

Your workshop instance: `t3a.large` : 2 CPU cores, 8 GB memory
15 cents per hour, 3\$/day

AWS EC2 instance naming



AWS EC2 instance sizes



💖 `c6a.48xlarge` 💖 : 192 cores, 384 GB mem, 7\$/hour

All costs: <https://instances.vantage.sh/>

Knowledge of cloud storage costs

EBS (instances hard drive): \$0.08/GB/month

S3 (“Dropbox”): \$0.023/GB/month

- If an instance is stopped: EBS costs occur
- If you create an instance snapshot: EBS costs occur too

How to avoid these costs? Terminate instances, delete snapshots, don't store too much on your S3

General scaling considerations

- **Alignment**

- Highly parallel, low memory, scales well with number of CPUs

- **Assembly**

- Moderately parallel, high memory, single big machine

- **Annotation**

- Don't! (jk), but moderately parallel. Single machine too

- **Phylogenomics**

- Can be made parallel (RAxML, Iq-Tree)



AI in bioinformatics

Sutskever @ NeurIPS'24

<https://x.com/TillLindeman90/status/1867764342172921901>

Pre-training as we know it will end

Compute is growing:

- Better hardware
- Better algorithms
- Larger clusters

Data is not growing:

- We have but one internet
- **The fossil fuel of AI**

Responses:



Derya Unutmaz, MD @DeryaTR_ · Dec 15

We may be running out of human-generated data on the internet, but there is vastly more data locked within biological systems! Just this one experiment generated a dataset with billions of tokens. We can easily generate trillions more of such data to train AI and solve biology!



Simona Cristea @simocristea · Dec 14

new human CD8+ T cell atlas of 1,151,678 cells from 961 samples, 68 studies & diseases. Grouped into 18 cell subtypes & w paired TCR info

+ a new VAE method scAtlasVAE for integrating cross-study atlas-level scRNAseq w cell subtype alignment & automatic cell subtype ...



Dhakshina @dhaksr · Dec 14

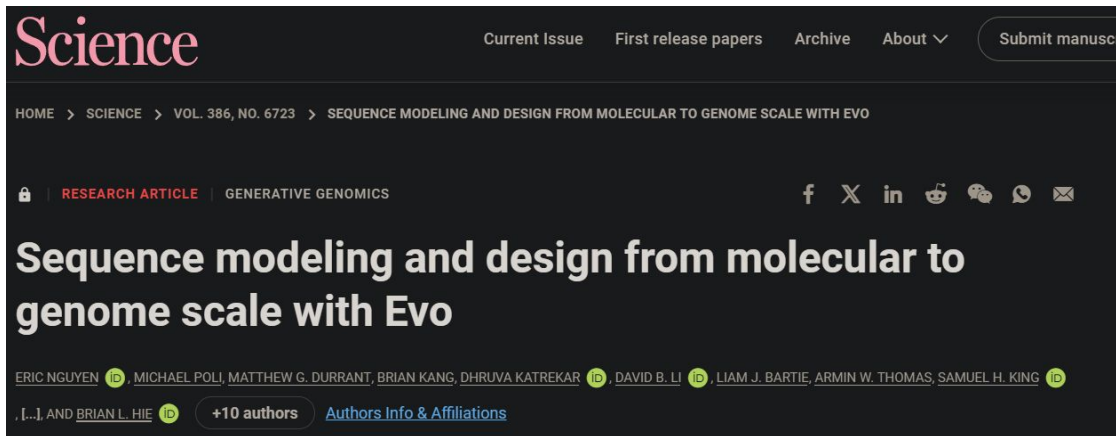
There is still so much industrial, enterprise, **bio data** that's not understood. Internet is only comp sci view of data. Sorry Ilya. I think Google brain (Dennis team) is in the right direction of fundamental research



Jason Wei @_jasonwei · Dec 13

Yall heard it from the man himself

Evo 7B foundation model



“Trained on 2.7 million prokaryotic and phage genomes”
(from GTDB, IMG/VRv4, IMG/PR)

“Excluding eukaryotic viruses”

(Is that a lot?)

-> **~10 TB** of genomic data

“Non-biological” data for training models

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

Common Crawl is a 501(c)(3) non-profit founded in 2007.

We make wholesale extraction, transformation and analysis of
open web data accessible to researchers.

[Overview](#)

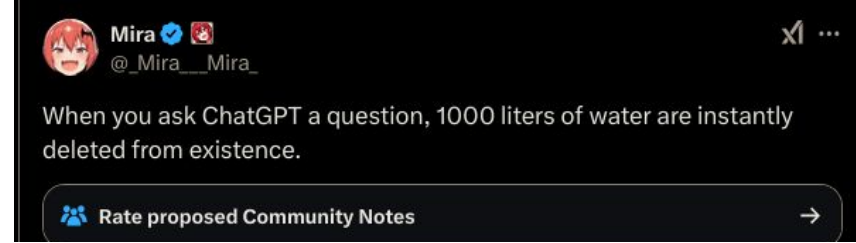


~400 terabytes of
uncompressed data

(updated every month,
history is kept)

We're making the “Common Crawl of biological data”

(Aside) AI and water use



“ChatGPT consumes half a litre of water for every 5-50 responses”

- How is the water used? Cooling systems
- Training vs inference
- Gpt3 vs Gpt4. Nowadays closer to 5ml per conversation

<https://www.seangoedecke.com/water-impact-of-ai/>

- “Water cost” of a hamburger: ~1000 litres

<https://www.weforum.org/stories/2019/02/this-is-how-much-water-is-in-your-burger/>
<https://pmc.ncbi.nlm.nih.gov/articles/PMC7442390/>

(Aside) AI and electricity use

- Training GPT-4: annual consumption of 6,500 homes

<https://www.weforum.org/stories/2024/07/generative-ai-energy-emissions/>

- Inference (queries):
 - Google query: 0.0003 kilowatt-hours
 - ChatGPT: 0.00289 kilowatt-hours (10x more)

<https://www.contrary.com/foundations-and-frontiers/ai-inference>

- New Nvidia chips 25x more energy efficient

<https://www.newscientist.com/article/2422928-nvidias-blackwell-ai-superchip-s-the-most-powerful-yet/>

Part 2: Big Data Toolbox

Computation

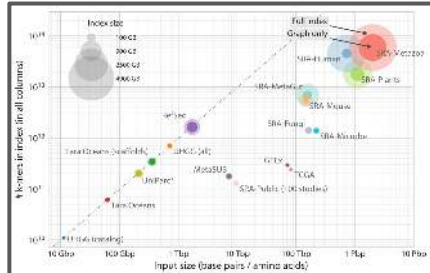
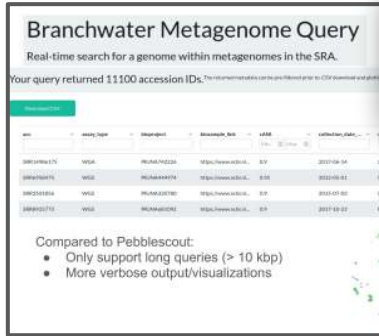
- Big computers, Cloud, Cluster
- Storage management
- Galaxy
- Knowledge of scaling limits
- Knowledge of cloud costs
- GNU parallel
- AI

Data mining

- Pebblescout, branchwater, ORA
- deCOM
- SRA metadata

Exploring metagenomes: Pebblescout, Branchwater, ORA

- Cutting-edge sequence database search tools
- Think BLAST, but the database is no longer “nr”; it’s all metagenomes.





Pebblescout pre-indexes nucleotide resources and searches them. The index contains at least one 25-mer from every 42-mer for all subjects in the database. Search has three modes: profile, summary, and detailed. Summary search ranks matching subjects using Pebblescout score. Search generates hashes from given user queries using the same scheme as used for indexing. This guarantees that every 42 bp match between the user query and any subject in the database is found.

Seven databases currently available are as follows:

1. **Metagenomic:** All metagenomic and metatranscriptomic runs released in public SRA before the end of 2021
2. **WGS:** All assemblies for the Whole Genome Shotgun sequencing projects available as of Feb 14, 2022
3. **RefSeq:** All assemblies available in the Reference Sequence collection as of April 22, 2022
4. **PH2HS_Runs:** Runs from Phase 3 of the 1000 Genomes project
5. **PH3HS_Biosample:** Runs from Phase 3 of the 1000 Genomes project where all runs for the same BioSample are considered as one subject
6. **Human RNAseq 2021:** All Human RNAseq runs released in public SRA in the year 2021
7. **Virus PacBio HiFi:** Viral samples sequenced with the PacBio SMRT technology defined in [PMC9528980](#)

[Documentation](#) provides additional information. A preprint for the [Pebblescout manuscript](#) is available at [biorxiv](#).

Please provide nucleotide queries, choose database and type of search to be performed, change parameters, as needed, and click View or Download. Please re-click View or Download if you change inputs.

Type FASTA Lines or GenBank Accessions Separated by Commas

Type FASTA lines here (sequence length must be at least 42 bases) or comma separated list or GenBank accessions



or Upload FASTA File

- All metagenomes, all assemblies (WGS), all human RNAseq, RefSeq
- Search for any sequence > 42 nt using k-mers (minimizers)

Pebblescout usage example



Collaborator needs all SRA samples with Wolbachia, to find new hosts



We did exactly this in our paper!

- (36 host species were known for Wolbachia)
 - Found by searching SRA metadata (2,545 runs)
- Pebblescout: searching for 3 genes (ftsZ, groE, wsp)
 - Found **16 more hosts** (35 runs)

Branchwater Metagenome Query

Real-time search for a genome within metagenomes in the SRA.

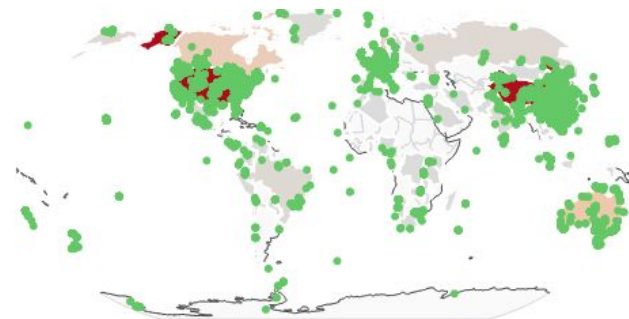
Your query returned 11100 accession IDs. The returned metadata can be pre-filtered prior to .CSV download and plotting with the table below. Your filtered table contains 11100 accession IDs

Download CSV

acc	assay_type	bioproject	biosample_link	cANI	collection_date_...	containment	geo_loc_name_c...	lat_lon	organism
				Min Max		Min Max			
SRR14986175	WGA	PRJNA742226	https://www.ncbi.nl...	0.9	2017-06-14	0.12	Germany	49.61,10.28	soil metagenome
SRR6958475	WGS	PRJNA444974	https://www.ncbi.nl...	0.95	2012-05-01	0.37	USA	33.5944,-109.1397	soil metagenome
SRR3501856	WGS	PRJNA320780	https://www.ncbi.nl...	0.9	2015-07-03	0.11	Singapore	1.33,103.75	activated sludge met...
SRR8925775	WGS	PRJNA681092	https://www.ncbi.nl...	0.9	2017-10-23	0.12	China	36.19,111.59	bioreactor metagen...

Compared to Pebblescout:

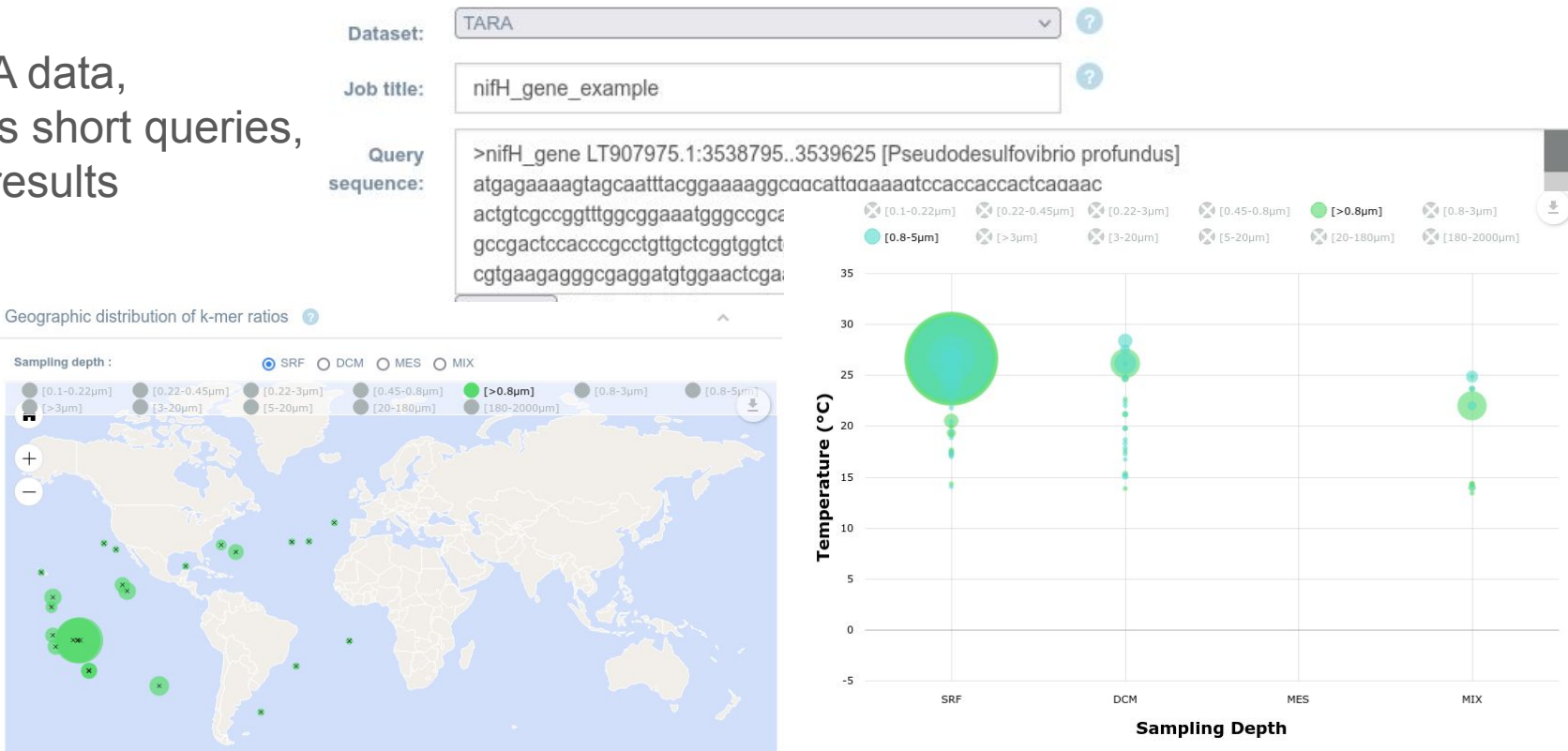
- Only support long queries (> 10 kbp)
- More verbose output/visualizations



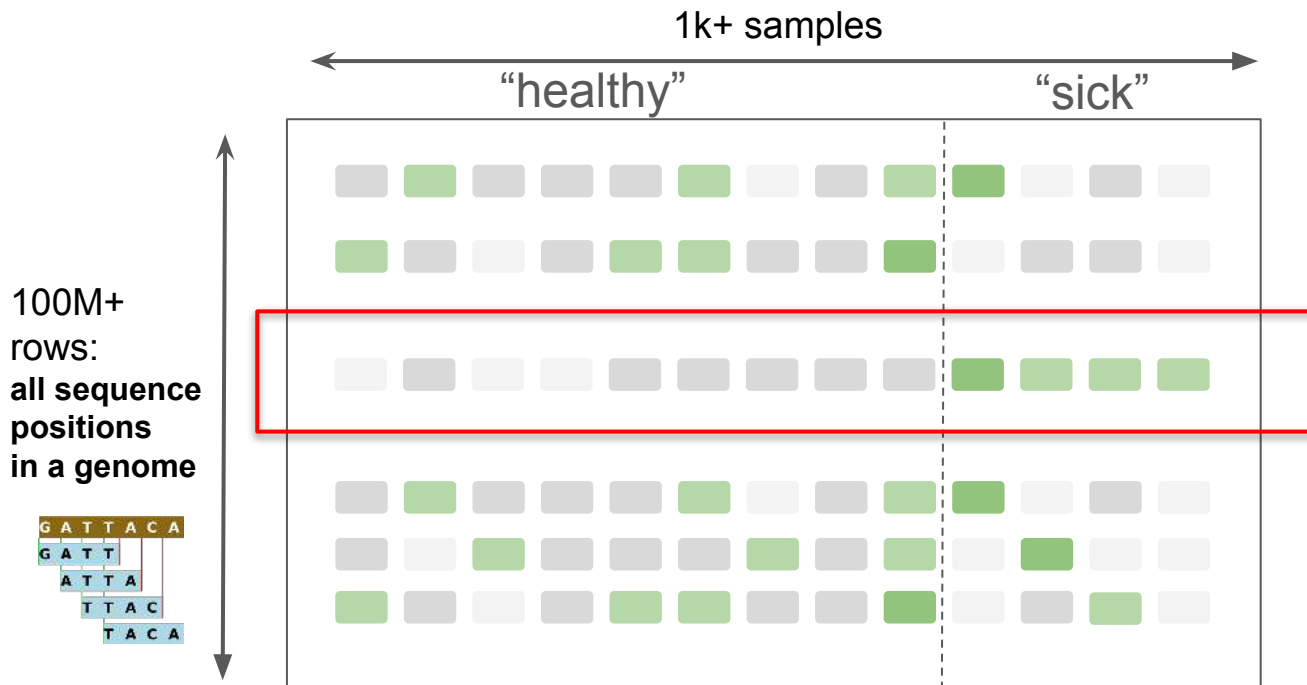


kmindex and ORA: indexing and real-time user-friendly queries in terabyte-sized complex genomic datasets
Lemane et al,
2023 (BioRxiv)
2024 (Nat Comp Biol)

All TARA data,
Supports short queries,
Instant results



Reference-free tools for detecting variation in large sequencing data cohorts



Methods:

1. New matrix construction algos
2. “Simple stats” on each row

Bioinformatics Advances 2022
Bioinformatics 2022
Nature Computational Science 2024
Nature Ecology & Evolution 2024
Bioinformatics 2024 to appear



From the G5:
T. Lemane (now GenoScope)
R. Vicedomini (now CNRS)
C. Duitama (**PRAIRIE** PhD student, now postdoc)

PR[A]IRIE
PRAIRIE is a French research network



Aschard Lab



Bourgeron Lab

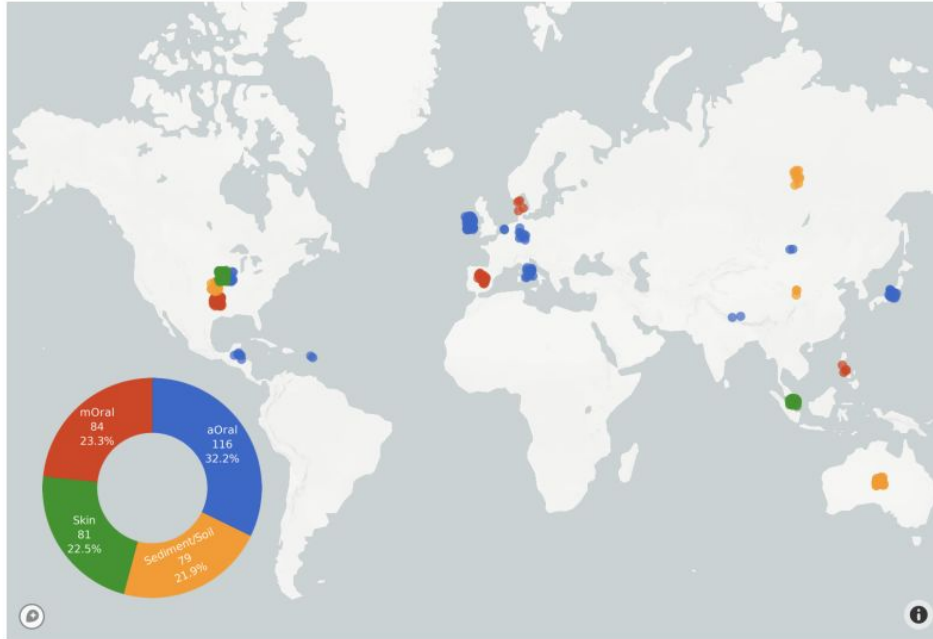


Quintana-Murci Lab





deCOM: integrating all ancient oral metagenomes



We gathered a collection of 360 samples (including contaminants and non contaminants) and obtained a k-mer matrix



	S1	S2	S3	S4	S5	S6	S7	S8
AATCG	1	0	0	0	0	1	1	0
GGGCT	0	0	0	0	0	1	1	0
TTCGA	0	0	1	1	0	1	1	0
AAACG	0	0	0	0	0	1	1	0
GGGCT	0	0	0	1	1	0	0	1
AATTT	0	0	0	0	0	0	0	0
ATCCC	0	0	1	0	0	0	0	0
GGGGT	1	1	1	1	1	1	1	1

New Results

[Follow this preprint](#)

deCOM: Similarity-based microbial source tracking of ancient oral samples using k-mer-based methods

Camila Duitama González, Riccardo Vicedomini, Téó Lemane, Nicolas Rascovan, Hugues Richard, Rayan Chikhi

doi: <https://doi.org/10.1101/2023.01.26.525439>

This article is a preprint and has not been certified by peer review [what does this mean?].

Wrapping up of Part 2: Big Data Toolbox

Computation

- Big computers, Cloud, Cluster
- Galaxy
- Storage management
- Knowledge of scaling limits
- Knowledge of cloud costs
- GNU parallel

Data mining

- Pebblescout, branchwater
- ORA
- deCOM
- SRA metadata





Part 3

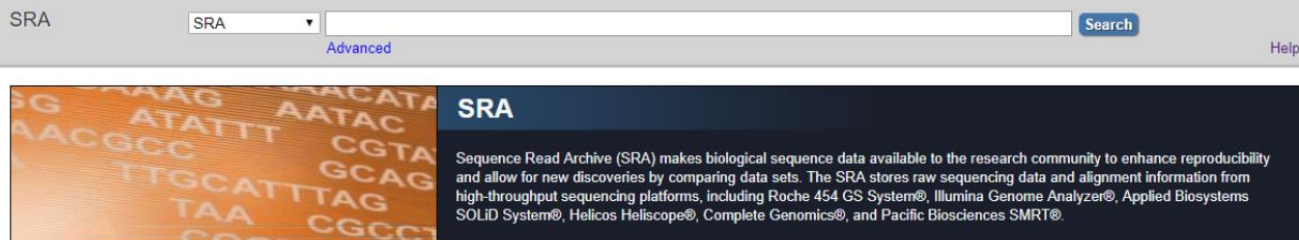
SRA-scale sequence exploration

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] $10^6 = 1\,000\,000$



Search results

Items: 1 to 20 of 19964

NextSeq 500 paired end sequencing (ERR3407135)

Metadata Analysis (alpha) Reads Download

☐ [NextSeq 500 paire](#)

1. 1 ILLUMINA (Illumina)
Accession: ERX34307

☐ [NextSeq 500 paire](#)

2. 1 ILLUMINA (Illumina)
Accession: ERX34307

☐ [NextSeq 500 paire](#)

INA (Illumina)
n: ERX34307

Filter: Find Filtered Download [What does it do?](#)

[What can the filter be applied to?](#)

< 1 1 346553 >

View: ☒ biological reads ☐ technical reads

Reads (separated)

1. ERR3407135.1 ERS3549882

name: NB551234.144.HL523AFX:1.11101.5421
member: default

2. ERR3407135.2 ERS3549882

name: NB551234.144.HL523AFX:1.11101.2248
member: default

3. ERR3407135.3 ERS3549882

name: NB551234.144.HL523AFX:1.11101.2566

>gnl|SRA|ERR3407135.1.1 NB551234.144:HL523AFX:1.11101.5421:1076 F (Biological)

ACCTGAGCGCGCAGCTCCAGTAAATCAAACGCGCGCGGAATTGGGATGTTCCATCAGT
TTCCAGGCGGTTTGCCCTGACGTCGCGACATGCGTAAGCTGCCAAATATCACGG

GTAAAGCTGGTAAGCGG

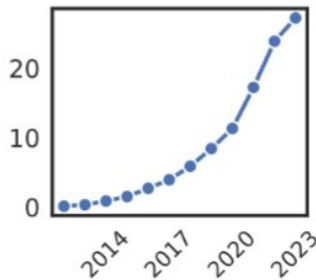
>gnl|SRA|ERR340713

ATCAACAACAGCGGGA

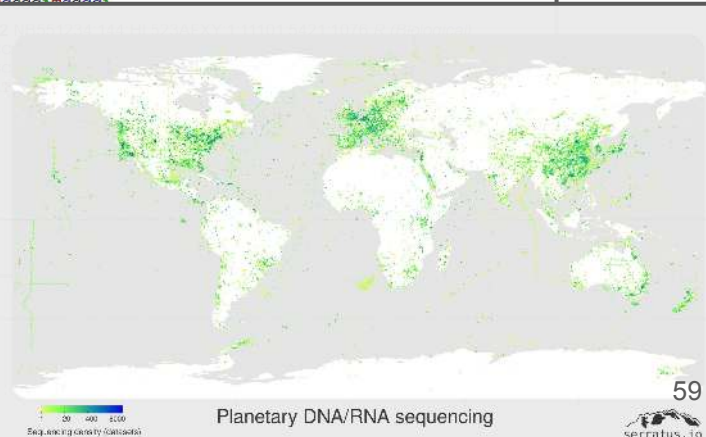
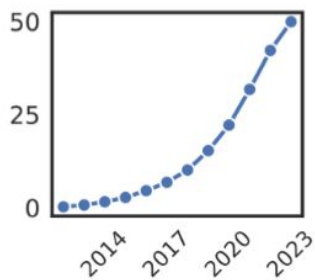
TACCGAAACCGGCACA

AAACCGCATCCGAAACG

accessions (millions)



size (petabytes)



What to do with the entire SRA?

Serratus: all public RNA-seqs analyzed for viral discovery



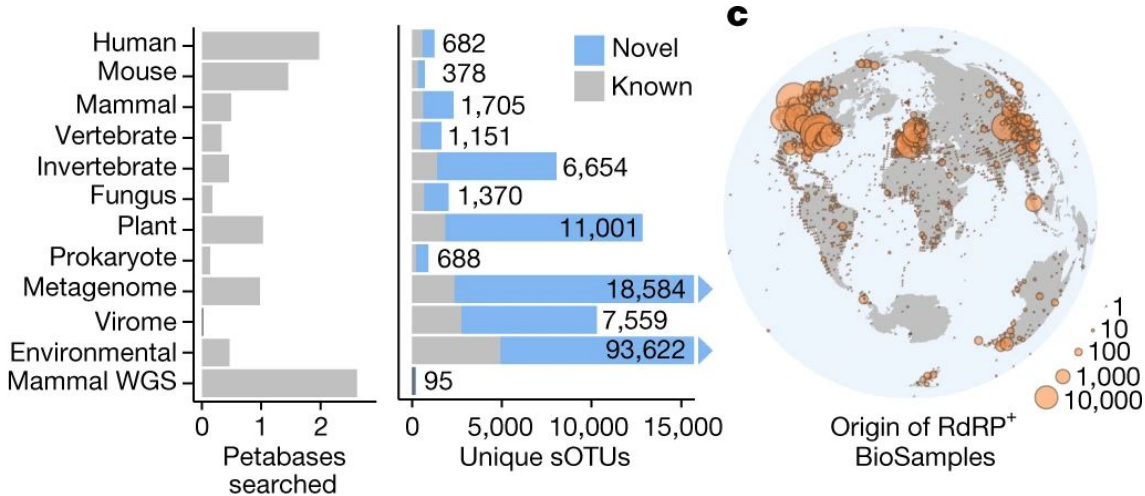
Discovered 130,000 new RNA viral species through large-scale read alignment, 9 new coronaviruses species.
One-off **cloud** analysis
(Edgar *et al*, Nature, 2022)

Some follow-ups to Serratus

Viral reactivation (Nature 2023)



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

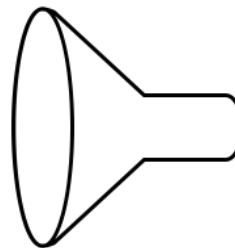


Obelisks



All RNA-seqs
pre-2020
(10 petabases)

**Serratus download &
align (bowtie2) to all
viral reference
genomes**

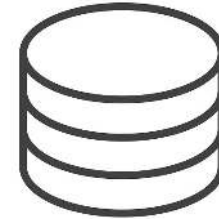
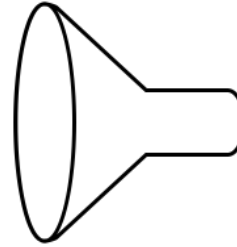


56,000 CoV+ samples
including 9 novel
coronavirus species
discovered



All RNA-seqs
pre-2020

**Serratus download &
sensitive align
(DIAMOND2)
to all known versions of
RNA virus universal gene**



**aligned reads
(.bam files)**
130k novel species
discovered

Toolbox used in Serratus

Part 2: Big Data Toolbox

Computation

- Big computers, Cloud
- Galaxy
- Knowledge of scaling limits
- Knowledge of cloud costs
- GNU parallel

Data mining

- Pebblescout, branchwater
- ORA
- deCOM
- SRA metadata

Didn't exist



Diving into SRA's data

What are SRA metadata?



[SRX8451857](#): Resequencing of *Vicugna vicugna* V_ss18

1 ILLUMINA (HiSeq X Ten) run: 111.2M spots, 33.4G bases, 11.8Gb downloads

Design: Resequencing

Submitted by: Universidad Austral de Chile

Study: Resequencing of Genomes of South American Camelids

[PRJNA612032](#) • [SRP265528](#) • [All experiments](#) • [All runs](#)

Sample: V_ss18

[SAMN14360346](#) • SRS6753932 • [All experiments](#) • [All runs](#)

Organism: [Vicugna vicugna mensalis](#)

Library:

Name: Vss18

Instrument: HiSeq X Ten

Strategy: WGS

Source: GENOMIC

Selection: RANDOM

Layout: PAIRED

Runs: 1 run, 111.2M spots, 33.4G bases, [11.8Gb](#)

Run	# of Spots	# of Bases	Size	Published
SRR11905265	111,191,160	33.4G	11.8Gb	2020-06-08

All of this



Accessing SRA metadata

~~0. NCBI website~~

1. NCBI FTP metadata

<https://trace.ncbi.nlm.nih.gov/Traces/index.html?view=mirroring>

2. SRA metadata on cloud SQL database
(AWS Athena, GCP BigQuery)

```
1 SELECT acc, mbases, mbytes, avgspotlen, librarylayout, instrument
2 FROM sra.metadata as s
3 WHERE consent = 'public' and avgspotlen >= 31
```

SQL Ln 1, Col 1

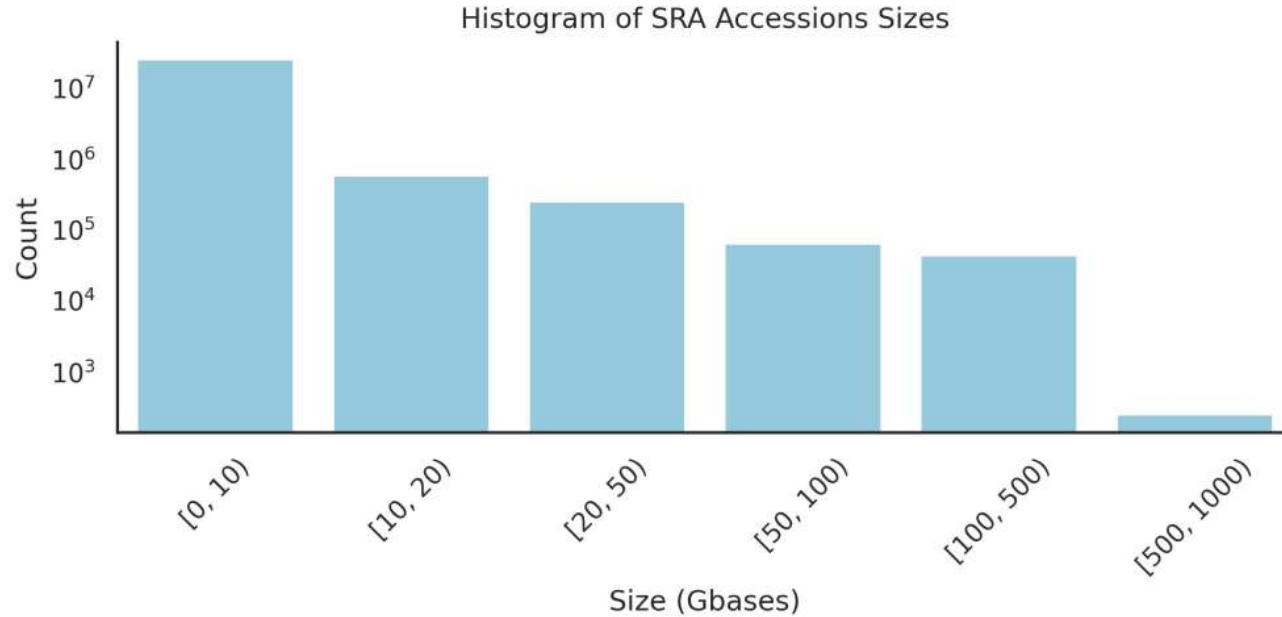
Run Explain [↗] Cancel Clear Create ▼

SRA metadata

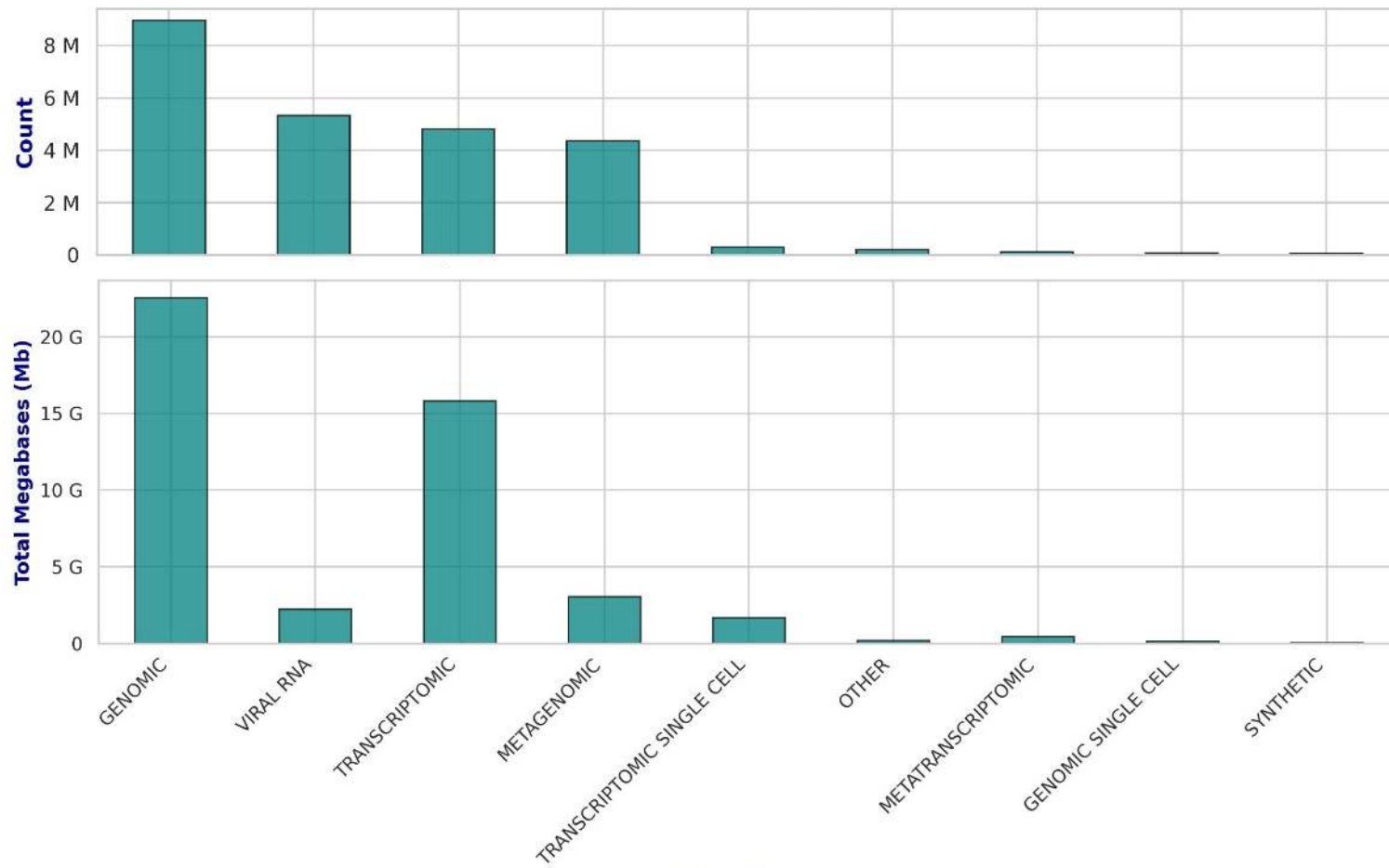
[-] tax_analysis	:
— acc	string :
— tax_id	int :
— rank	string :
— name	string :
— total_count	bigint :
— self_count	bigint :
— ilevel	int :
— ileft	int :
— iright	int :

[-] metadata	:	— organism	string :
— acc	string :	— sra_study	string :
— assay_type	string :	— releasedate	date :
— center_name	string :	— bioproject	string :
— consent	string :	— mbytes	int :
— experiment	string :	— loaddate	timestamp :
— sample_name	string :	— avgspotlen	int :
— instrument	string :	— mbases	int :
— librarylayout	string :	— insertsize	int :
— libraryselection	string :	— library_name	string :
— librarysource	string :	— biosamplemodel_sam	array<string> :
— platform	string :	— collection_date_sam	array<string> :
— sample_acc	string :	— geo_loc_name_country_calc	string :
— biosample	string :	— geo_loc_name_country_continent_calc	:

SRA accessions sizes (2023)



SRA accessions types (2023)



SRA taxonomy analysis

Method | Open Access | Published: 20 September 2021

STAT: a fast, scalable, MinHash-based k -mer tool to assess Sequence Read Archive next-generation sequence submissions

Kenneth S. Katz , Oleg Shutov, Richard Lapoint, Michael Kimelman, J. Rodney Brister & Christopher O'Sullivan

Genome Biology **22**, Article number: 270 (2021) | [Cite this article](#)

"we have processed more than 27.9 Peta base pairs from runs"

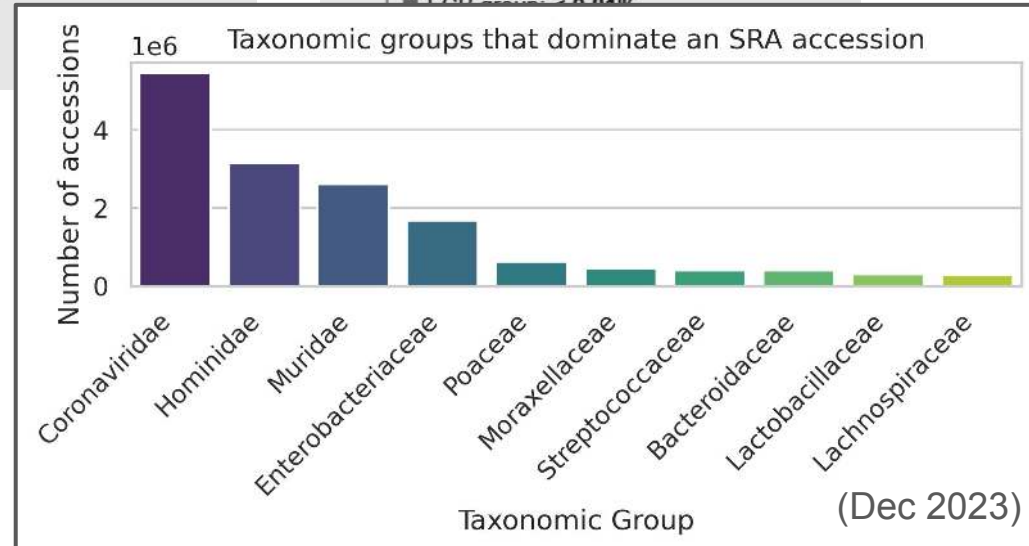
Example STAT output:

Taxonomy Analysis

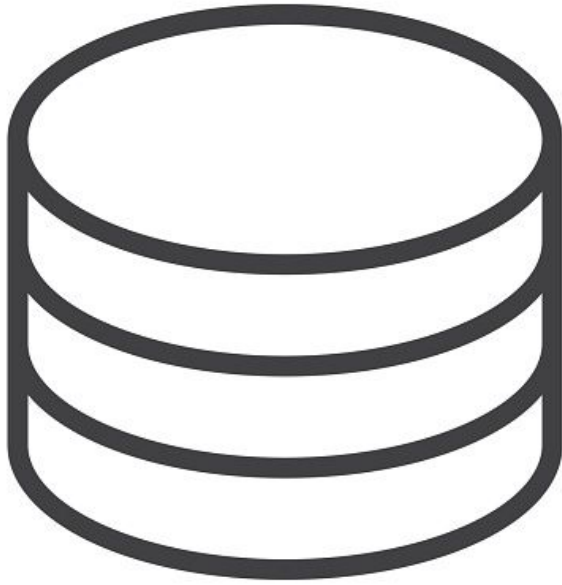
Unidentified reads: 40.04%

Identified reads: 59.96%

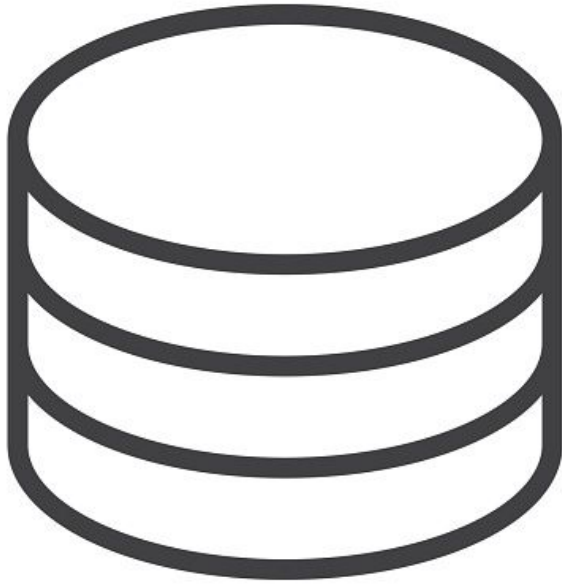
- ▢ Viruses: 50.55%
 - ▢ ssRNA viruses: 50.55%
 - ▢ Measles morbillivirus: 50.55%
 - ▢ dsDNA viruses, no RNA stage: < 0.01%
 - ▢ ssDNA viruses: < 0.01%
 - ▢ Ortervirales: < 0.01%
- ▢ cellular organisms: 9.4%
 - ▢ Bacteria: 6.44%
 - ▢ Proteobacteria: 1.76%
 - ▢ Terrabacteria group: 0.48%
 - ▢ FCB group: 0.01%



Can one analyze all of Life's genetic data?
(before Logan)



- How much time to download 40 petabytes at 200 MB/sec?



- How much time to download 40 petabytes at 200 MB/sec?

~ 6 years

How to analyze all of Life's genetic data?
(before Logan)
We can't

Serratus infrastructure

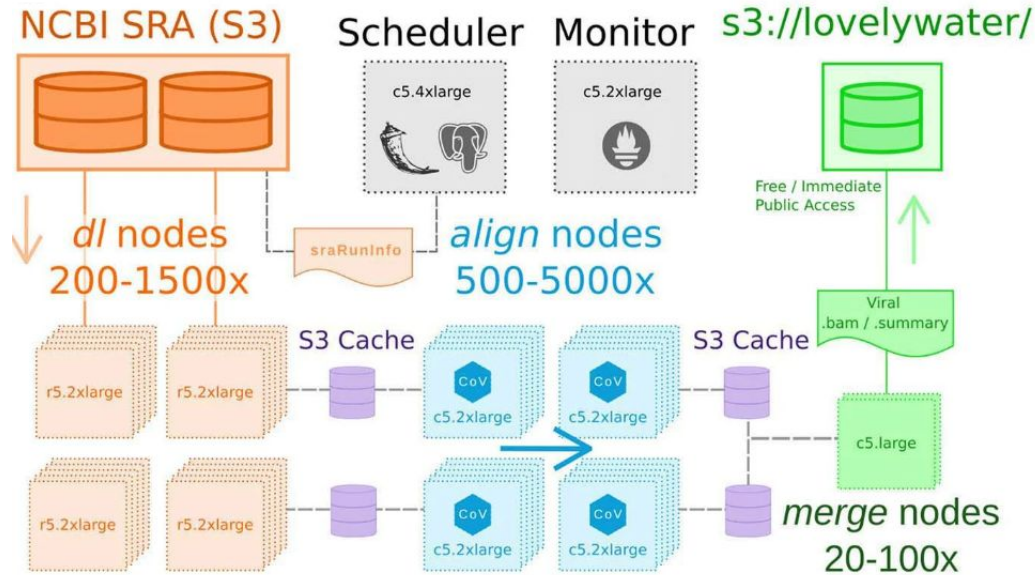


Fig: A. Babaian

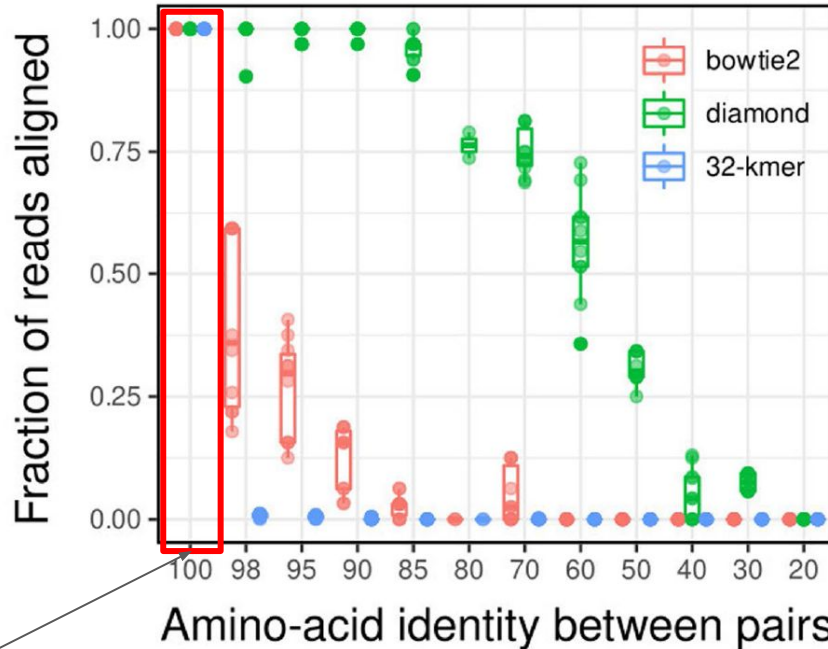


With this, we expanded the number of viruses species known by 10x!

How to analyze all of Life's genetic data?
(before Logan)

We can, with cloud-scale efforts

Alignment: high **speed** or high **sensitivity**, choose one



Credit: RC Edgar

Human reads alignment

SRA-scale alignment

State of the art (ordered by sensitivity/speed):

1. **Sourmash branchwater** (sketches)
 - Metagenomes, long sequences
2. **NCBI Pebblescout** (k-mers, no alignment)
 - Metagenomes, > 42 bp sequences
3. **Bowtie2, STAR** (k-mers, alignment)
 - Serratus1 (all RNAseqs)
 - Recount3 (750k human/mouse RNAseqs)
4. **DIAMOND** (AA-mers)
 - Serratus1.5 (all RNAseqs)
5. **HMMs?** (profile)



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 0.4 PB of contigs (high contiguity)
- It's done, finally

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

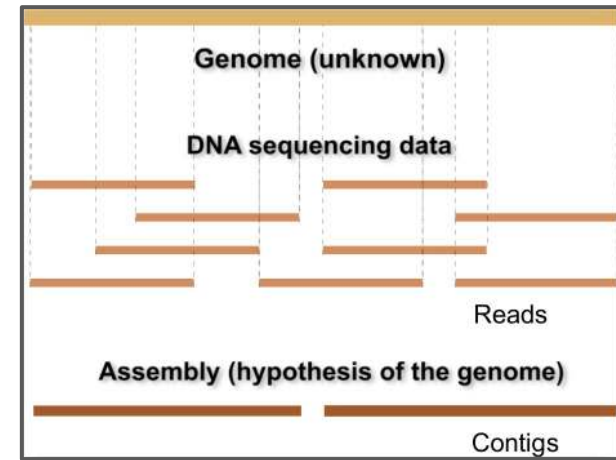
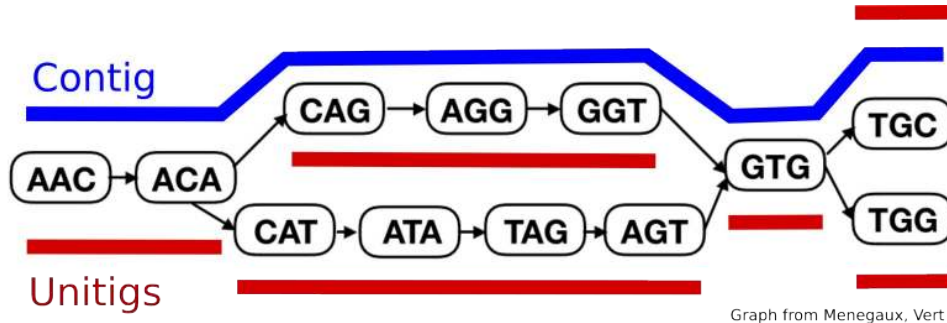
 Rayan Chikhi, Brice Raffestin,  Anton Korobeynikov,  Robert Edgar,  Artem Babaian

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



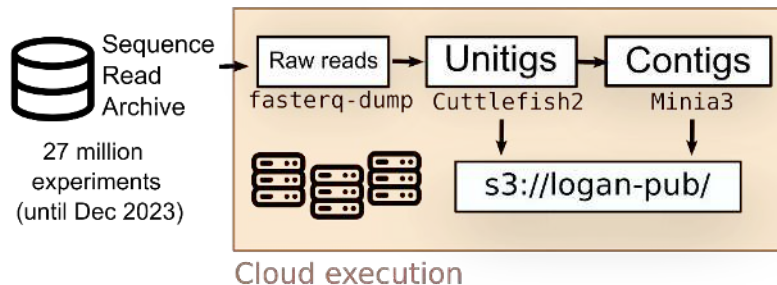
Why unitigs? they keep all variants (SNPs, indels, ..)

Contigs are consensus

Logan: project steps

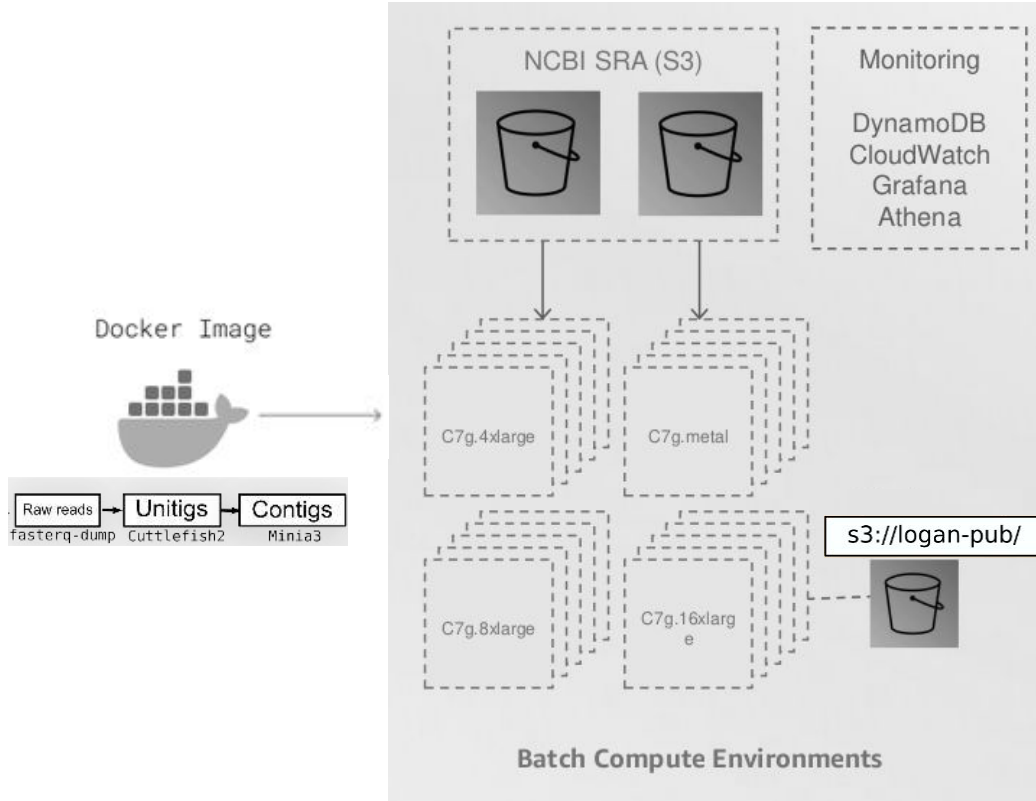
- **Step 1 (2024):** Download all of SRA, assemble each sample, host results publicly
[done]

30M CPU hours, 19 petabytes downloaded, 2 petabytes stored



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

Logan: infrastructure



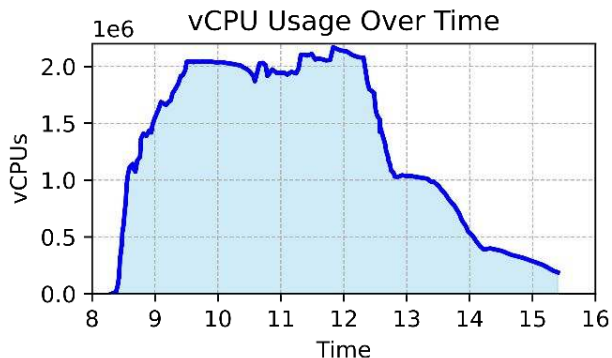
AWS services used:

Batch
S3
DynamoDB
Athena
CloudFormation
CloudWatch
Cost Explorer
Grafana

Logan: computation statistics

Global statistics

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



Run 6 statistics

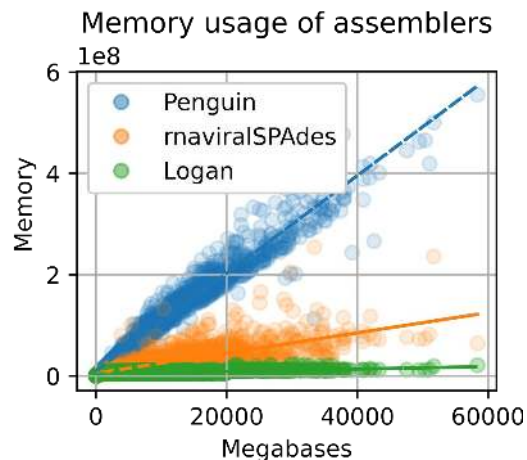
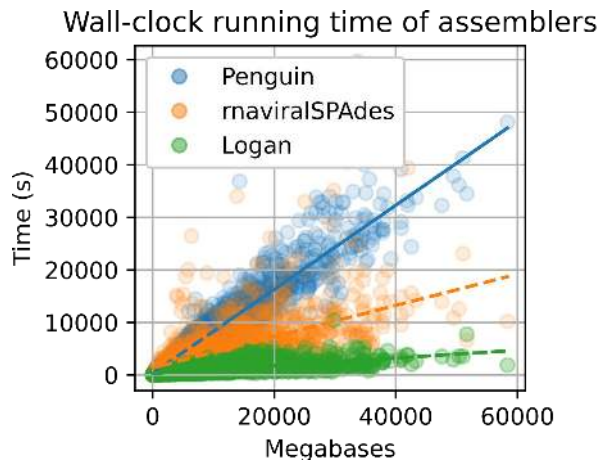
Input data	19.6 petabytes
Runtime*	7 hours
Peak Number of Instances	73,100
Peak Number of vCPUs	2.18 million
Peak Total EBS storage	52 petabytes

Many failures:

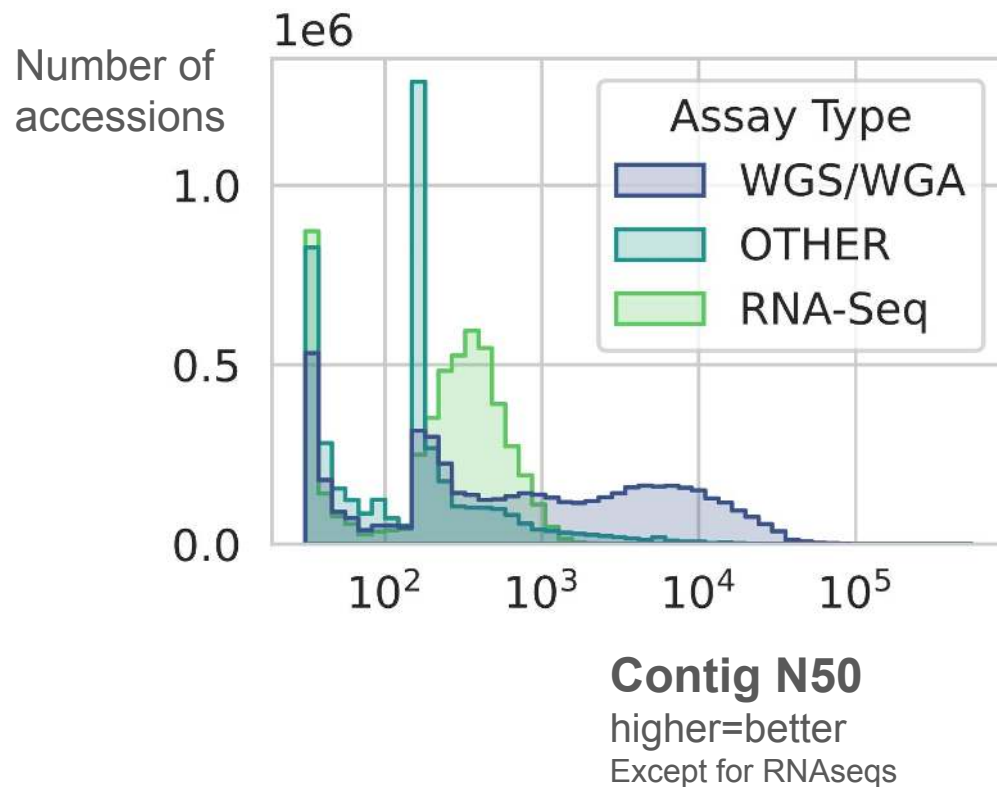
- Reached S3 write limits, learned the concept of “S3 prefixes”
- Reach DynamoDB write limits too
- `fasterq-dump` timeouts, turns out SRA aligned reads format (~15% of accessions) connects to internet

Why wasn't this done before?

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



Draft-level assembly contiguity



Algorithmic components used in Logan

- String algorithms (“minimizers” ($\sim$ =string attractors) 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)

Part of the algorithmic 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?

Accessing Logan

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

From anywhere, no account needed

Logan

2 petabases

GenBank

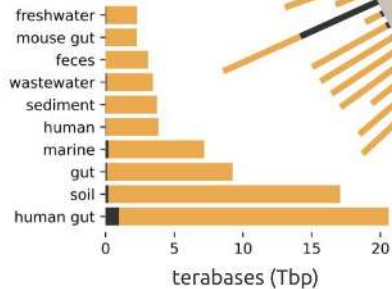
0.024 Pbp

BLAST 'nt'

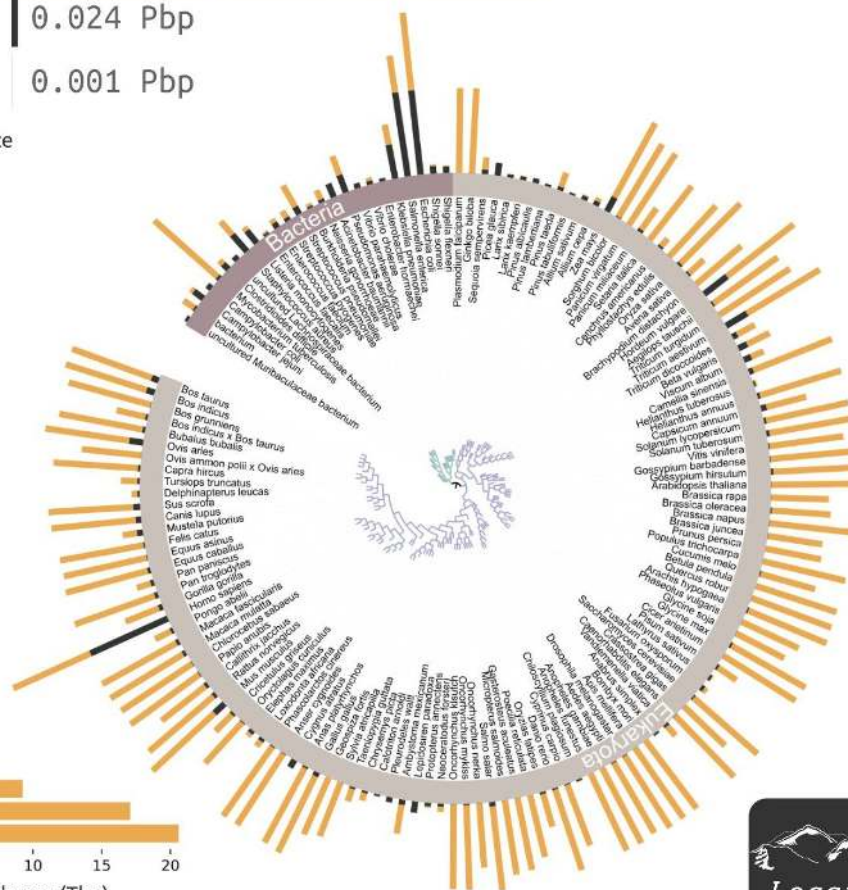
0.001 Pbp

Assembly Size

Metagenomes



terabases (Tbp)



A (draft-level) genome for all organisms

.. in fact, often more than one genome per species.

Reference to Olga's talk:
You now probably already
have a draft-level short-reads
genome for your species.



Logan “fun facts”

- Logan total computation: **30 hours**. Would have been ~1.5 years on local cluster.
- Just listing the S3 folder takes **~1 hour**
- Downloading all Logan contigs (385 TB) at 10 Gbits/s takes **3 days**
- Sequence alignment with DIAMOND (`--sensitive`) streaming all of Logan contigs takes **4 hours** on 60k cloud vCPUS (4k\$)

Logan Search

INPUT

text file session

Query sequence(s) *
Fasta/Fastq format

>Query
ACCGTAGCCTTAGAATTA

Load

NOTIFICATION

Email

Your email

CONFIGURATION

Groups

Threshold = 0.5

125 1.0

Submit Reset

Table Map Plot Contigs/Unligns Search (BETA) Help

lmer_coverage > 0.7 AND assay_type IN ('WGS','WGA')

ID	lmer_coverage	bioproject	biosample	bioproject_title	bioproject_description	sample_acc	assay_type	center_name	experim
ERR690905 (SRA OV)	1	PRJEB47927 (SRA OV)	SAMEA10271030 (SRA OV)	Chromosome-scale genome as...	Pelagophytes (Stramenopiles) a...	ERS7925717	WGS	GSC	ERX6037
ERR3487222 (SRA OV)	1	PRJEB34158 (SRA OV)	SAMEAS899549 (SRA OV)	Collection of Marine Eukaryote...	This project is part of the Marin...	ERS3588172	RNA-Seq	GSC	ERX3511
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	PRJNA517004 (SRA OV)	SAMN13301556 (SRA OV)	100 Algal genome project (ALG...	100 Algal genome project (ALG...	SR52763869	WGS	NEW YORK UNIVERSITY ABU D...	SRX104
ERR9764111 (SRA OV)	1	PRJEB47927 (SRA OV)	SAMEA14430949 (SRA OV)	Chromosome-scale genome as...	Pelagophytes (Stramenopiles) a...	ERS12037107	Hi-C	GSC	ERX931
ERR3487221 (SRA OV)	1	PRJEB34158 (SRA OV)	SAMEAS899549 (SRA OV)	Collection of Marine Eukaryote...	This project is part of the Marin...	ERS3588172	RNA-Seq	GSC	ERX3511
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
SRR1197240 (SRA OV)	1	PRJNA239089 (SRA OV)	SAMN01985059 (SRA OV)	Pelagomonas calceolata Ceno...	Pelagomonas calceolata geno...	SR5376831	WGS	ICVH	SRX4931
SRR1827860 (SRA OV)	1	PRJNA814250 (SRA OV)	SAMN26541151 (SRA OV)	Metagenomic time-series the ...	This study examines monthly d...	SR51225932	WGS	CLARK UNIVERSITY	SRX144
SRR1296779 (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
SR5851684 (SRA OV)	0.988	PRJNA193558 (SRA OV)	SAMN02144670 (SRA OV)	Pelagomonas calceolata strain...	Pelagomonas calceolata transcr...	SR5421523	RNA-Seq	ICVH	SRX278
ERR1726804 (SRA OV)	0.902	PRJEB4352 (SRA OV)	SAMEA2623204 (SRA OV)	EMG produced TPA metageno...	The Third Party Annotation (TP...	ERS493517	WGS	GSC	ERX179
ERR688425 (SRA OV)	0.972	PRJEB4352 (SRA OV)	SAMEA2623699 (SRA OV)	EMG produced TPA metageno...	The Third Party Annotation (TP...	ERS492708	WGS	GSC	ERX948
ERR660428 (SRA OV)	0.972	PRJEB4352 (SRA OV)	SAMEA2619943 (SRA OV)	EMG produced TPA metageno...	The Third Party Annotation (TP...	ERS488730	WGS	GSC	ERX948
SRR13206796 (SRA OV)	0.972	PRJNA960716 (SRA OV)	SAMN17237790 (SRA OV)	Pelagomonas calceolata CCMF...	Genome sequencing of Pelago...	SR57990976	WGS	DALHOUSIE UNIVERSITY	SRX909
SRR8790596 (SRA OV)	0.972	SAMN12637356 (SRA OV)	SAMN12637356 (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...	ERS490541	WGS	GSC	ERX556
ERR660454 (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	PRJEB609 (SRA OV)	SAMEA2623379 (SRA OV)	Metatranscriptome sequencin...	Metatranscriptome sequencin...	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...	ERS490507	WGS	GSC	ERX556
ERR660450 (SRA OV)	0.97	PRJEB4352 (SRA OV)	SAMEA2622936 (SRA OV)	EMG produced TPA metageno...	The Third Party Annotation (TP...	ERS493111	WGS	GSC	ERX947
SR50524774 (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	PRJEB609 (SRA OV)	SAMEA2620025 (SRA OV)	Metatranscriptome sequencin...	Metatranscriptome sequencin...	ERS488834	RNA-Seq	GSC	ERX178
ERR660441 (SRA OV)	0.97	PRJEB4352 (SRA OV)	SAMEA2622325 (SRA OV)	EMG produced TPA metageno...	The Third Party Annotation (TP...	ERS492154	WGS	GSC	ERX948
SR52584947 (SRA OV)	0.97	PRJNA1003508 (SRA OV)	SAMN36908814 (SRA OV)	Nitrite oxidizing bacteria in o...	Novel nitrite oxidizing bacteria...	SR518561338	WGS	PRINCETON UNIVERSITY	SRX213

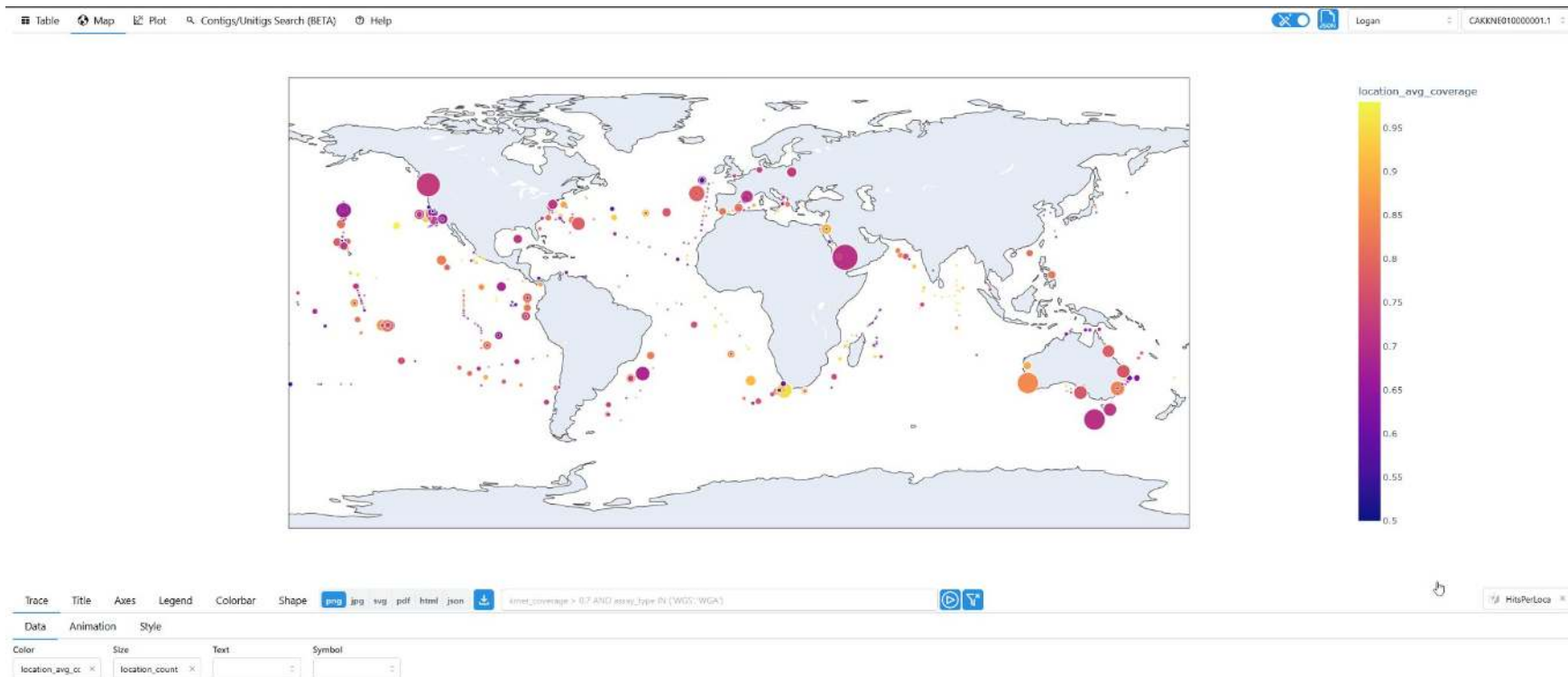
Page Size: 100 1 to 1

Export Remove filters Filter NaN

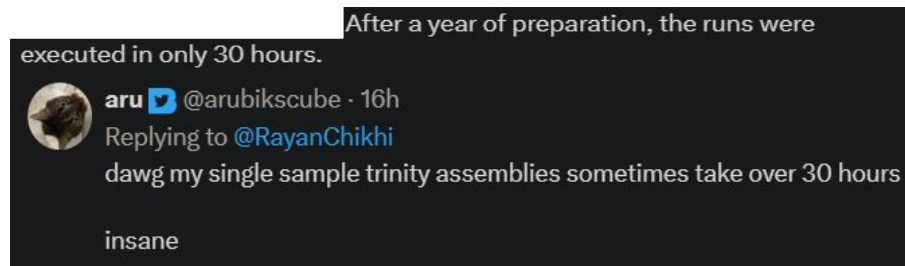
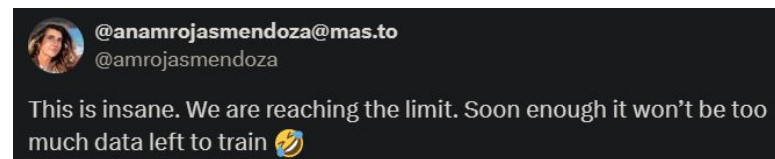
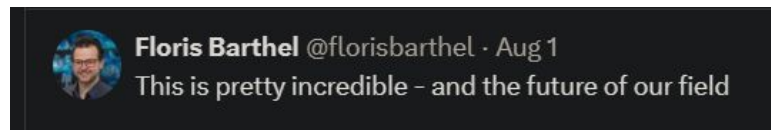
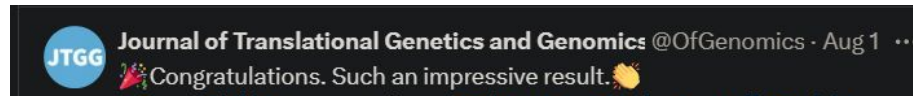
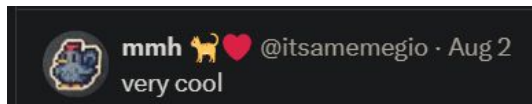
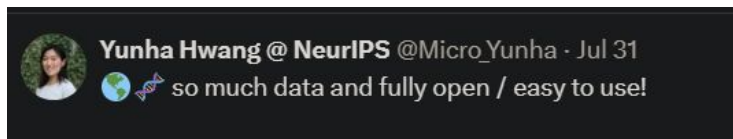
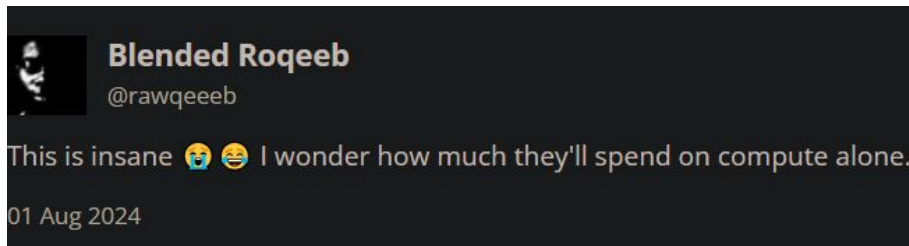
About your query (AI generated)

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.

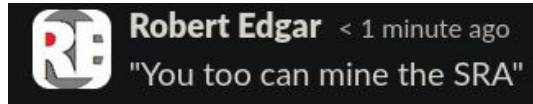
Logan Search



Logan reactions



Want to dive in Logan data ?



- We do whole-SRA high-sensitivity alignments regularly
 - Ask to include your sequence(s) in the next batch
- All Logan unitigs & contigs are public, but if you need assistance: contact me
- Logan-search.org service for high-identity alignments

Many planned analyses

- RNA viruses (Serratus group)
- Viroids (help wanted)
- K-mer indexing (Peterlongo/Lemane)
- Compression (Rouze/Limasset)
- Meta-data parsing and geographic/ecology explorer (help wanted)
- Bacteria/AMR (Sedlazeck lab)
- Improving genome assemblies (maybe)
- Eukaryotic barcodes (help wanted)
- SRA-scale protein clustering (Steinegger lab)
- SRA metadata in a LLM for textual queries (help wanted)



Call for collaborations

We have a very special moment right now to liberate all the data in the SRA. I'm asking for all of your help so that we can make this a landmark project from the community.



Can you do hands-on bioinformatics?

Contact rayan.chikhi@pasteur.fr and we'll add you to Logan/Serratus Slack

Also: Artem Babaian (Serratus PI/Logan co-PI) is looking for postdocs: <https://www.rnalab.ca/> *Laboratory for RNA-Based Lifeforms*

How can Logan be useful?

A “fun” experiment..

Pick an organism: Chicken



(From 2024 Workshop on Genomics - a perfectly sane year)

Pick a biological question: what's the genetic basis for its color?

Logan can get you all the data you need for any study.

- 1) For the purpose of the demo, we'll focus on one gene (MC1R)
- 2) Then we'll gather sequence data from chickens, isolate that gene, and look for variants associated to breed/color

Collecting chickens

How to retrieve many chicken sequences?



- 0) ~~BLAST~~ Not enough individuals in nt
- 1) ~~NCBI Peptide~~ Only has metagenomes
- 2) SRA metadata query
- 3) SRA taxonomy query

SRA metadata query 1: fail

SRA [Create alert](#) [Advanced](#)

Summary ▾ 20 per page ▾

Send to: ▾

Search results

Items: 1 to 20 of 235320

<< First < Prev Page of 11766 Next > Last >>

☐ [WGS of E.coli isolate](#)

1. 1 ILLUMINA (Illumina MiSeq) run: 8.2M spots, 2.5G bases, 1.5Gb downloads
Accession: SRX25244676

☐ [WGS of E.coli isolate](#)

2. 1 ILLUMINA (Illumina MiSeq) run: 9M spots, 2.7G bases, 1.7Gb downloads
Accession: SRX25244666

☐ [WGS of E.coli isolate](#)

3. 1 ILLUMINA (Illumina MiSeq) run: 13.5M spots, 4.1G bases, 2.6Gb downloads
Accession: SRX25244661

SRA metadata query 2: better

<https://www.ncbi.nlm.nih.gov/sra/?term='yellow+chicken'>

SRA Run Selector	Select	Runs	Bytes	Bases	Download	
	Total	324	1.23 Tb	3.43 T	Metadata	or Accession List

[https://www.ncbi.nlm.nih.gov/sra/SRX4478521\[accn\]](https://www.ncbi.nlm.nih.gov/sra/SRX4478521[accn])

SRX4478521: DNA-seq of Gallus gallus: Wuhua yellow chicken

1 ILLUMINA (HiSeq X Ten) run: 38M spots, 11G bases, 3.9Gb downloads

Getting sequencing data from the SRA (without Logan)

TL;DR: state of the art is `prefetch` + `fasterq-dump`

`prefetch`: downloads `.sra` file locally

`fasterq-dump`: transforms `.sra` to `.fastq` or `.fasta`

Example:

```
prefetch [accession] && fasterq-dump [accession].sra
```

Big data genomics:)

```
$ cat download_and_map_accession.sh
```

```
set -e  
accession=$1
```

```
aws s3 cp s3://sra-pub-run-odp/sra/$accession/$accession \  
    $accession.sra --no-sign-request
```

```
minimap2 -t20 -x sr mclr.fa <(fasterq-dump --fasta-unsorted $accession.sra) \  
    -o mapping/$accession.minimap2_output
```

```
rm -f $accession.sra
```

Parallelize processing:

```
cat accessions.txt | parallel -j 10 "./download_and_map_accession.sh {}"
```

Analyzing ~300 SRA samples (without Logan)

3 terabases from “yellow chicken” SRA accessions downloaded and mapped to MC1R

```
-rw-r--r--. 1 ec2-user ec2-user 154700 Jan 11 18:22 SRR11521907.minimap2_output
-rw-r--r--. 1 ec2-user ec2-user 174639 Jan 11 18:24 SRR11521908.minimap2_output
-rw-r--r--. 1 ec2-user ec2-user 150667 Jan 11 18:25 SRR11521909.minimap2_output
-rw-r--r--. 1 ec2-user ec2-user 135759 Jan 11 18:25 SRR11521910.minimap2_output
-rw-r--r--. 1 ec2-user ec2-user 194411 Jan 11 18:23 SRR11521911.minimap2_output
-rw-r--r--. 1 ec2-user ec2-user 149717 Jan 11 18:24 SRR11521912.minimap2_output
-rw-r--r--. 1 ec2-user ec2-user 149674 Jan 11 18:25 SRR11521913.minimap2_output
-rw-r--r--. 1 ec2-user ec2-user 204873 Jan 11 18:26 SRR11521914.minimap2_output
-rw-r--r--. 1 ec2-user ec2-user 180067 Jan 11 18:26 SRR11521915.minimap2_output
-rw-r--r--. 1 ec2-user ec2-user 139216 Jan 11 18:26 SRR11521916.minimap2_output
-rw-r--r--. 1 ec2-user ec2-user 113860 Jan 11 18:26 SRR11521917.minimap2_output
-rw-r--r--. 1 ec2-user ec2-user 157065 Jan 11 18:27 SRR11521918.minimap2_output
-rw-r--r--. 1 ec2-user ec2-user    6240 Jan 11 18:25 SRR11678145.minimap2_output
-rw-r--r--. 1 ec2-user ec2-user   11665 Jan 11 18:25 SRR11678146.minimap2_output
-rw-r--r--. 1 ec2-user ec2-user   15025 Jan 11 18:25 SRR11678147.minimap2_output
```

Took around 1.5 hours, on a 6\$/hour cloud machine

```
1:36:09elapsed 2026%CPU (0avgtext+0avgdata 1182952maxresident)k
```

Chicken pangenomics

- Constructed pangenome (de Bruijn) graph of MC1R from the “yellow chicken” accessions
- BLASTed a consensus gene to the graph



.. good, but this is only for one breed.



We need more data

Getting *all* SRA entries containing chicken reads: SRA taxonomy query through STAT

```
SELECT acc  
FROM "sra"."tax_analysis"  
WHERE name = 'Gallus gallus' AND total_count > 100000
```

Results (59,240)



With a little help from Logan

- Logan = 27 million SRA assemblies



- All of the **Results (59,240)** are now already assembled
 - Chicken data =
 - 4.3 terabases of contigs
 - **374 terabases** of reads 🤖 (= 1000GP twice)

Logan analysis

Cloud download of Logan accessions, mapping on the fly to MC1R:

```
minimap2 -x asm20 -t 8 -a mc1r.fa \
<(aws s3 cp s3://logan-pub/c/$accession.contigs.fa.zst - | zstdcat) \
| samtools view -hF4 - \
> mapping-logan/$accession.minimap2_output
```

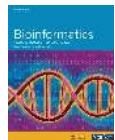
16 hours on a 4xlarge instance (16 vCPUs, 0.6\$/hour).

i.e. 124x more data for same \$'s than direct SRA download

11,072 MC1R genes pangenome (de Bruijn graph, k=31, BCALM2)



GWAS directly from sequences
(skips SNP detection):



JOURNAL ARTICLE

kmdiff, large-scale and user-friendly differential
k-mer analyses 

Téo Lemané, Rayan Chikhi, Pierre Peterlongo 



TGGGGGTCATCGCCGTGGACCGCTACATCG..



TGGGGGTCATCGCCGTGGACCGCTACAT**A**..

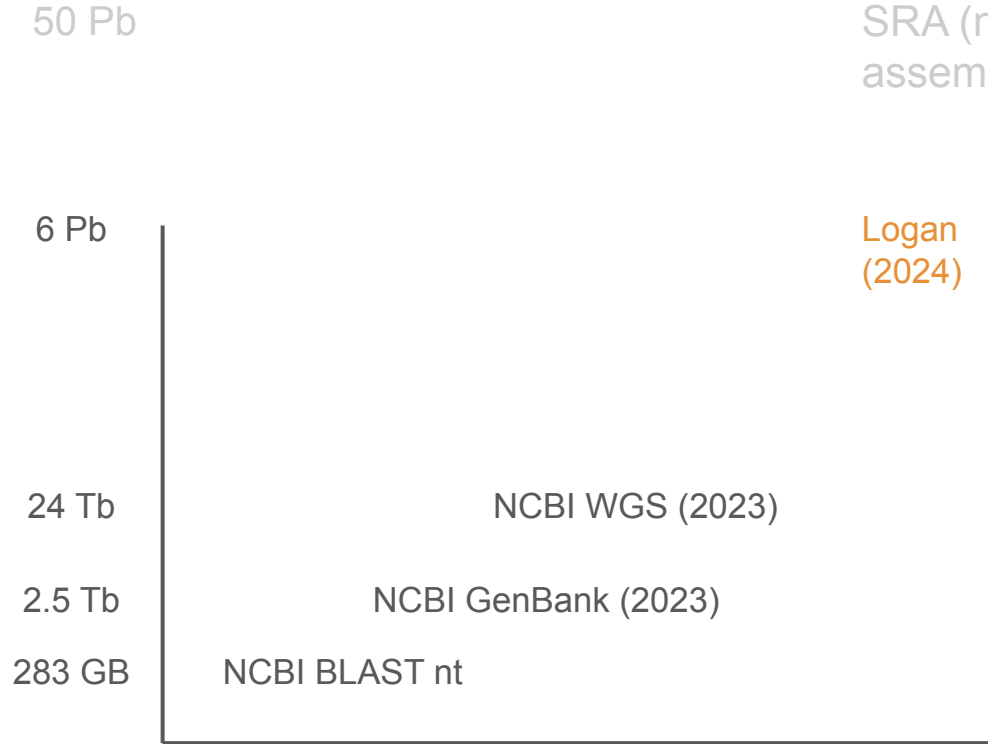
$p < 10^{-7}$

What just happened?

- Casually analyzed 59,000 SRA accessions for this talk
- 374 Terabases of reads, **0.7% of all public sequencing data**
- Downloaded assemblies and mapped to a reference gene in < 1 day on a **single** modest AWS **instance**
- Total analysis cost: 9\$

This enables any biological question to be investigated using all of the planet's sequencing data quickly, by anyone

Public sequence datasets



SRA (not
assembled)

Logan
(2024)



Last year..



Milos: still, you should present real biological results



me: we don't have any yet :(

Thankfully this year, things have changed :)

Logan x ???

(slides by A. Babaian)

Not ready for prime-time disclosure.
Stay tuned for Logan preprint update!

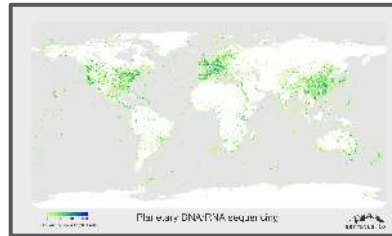
Conclusion

- **SRA-scale analyses now 100x more tractable**
- **Logan: all of Life's genomic data finally accessible**
- **Many biological discoveries to be made**
- **Better foundation models**

What Logan doesn't replace

Generation of new samples

**High-quality curated
genomes**





Outro

What we've seen today

- Some elements of big data bioinformatics
- Toolbox for Big Data
 - Cloud, parallelism, storage handling, knowledge of limitations, AI
- SRA primer
 - Mining metadata
 - Mining sequences
 - Aligning at scale
 - Serratus
- Logan
 - All of Life's genomic data, available

bigger data

big data



WE'RE-GONNA-NEED



A BIGGER INSTANCE TYPE

Sequence Bioinformatics



Lab members:

Francesco Andreace
Gaetan Benoit
Rayan Chikhi
Camila Duitama
Yoann Dufresne
Victor Levallois
Mélanie Ridet
Timothé Rouze
Yoshihiro Shibuya

Alumni:

Luc Blassel
Luca Denti
Mael Kerbiriou
Téo Lemane
Camille Marchet
Pierre Marijon
Riccardo Vicedomini

Support for ERC
+ Prairie + Pasteur:
Olivier Gascuel

Logan co-creators:

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

AWS support: Dorian Schaal,
Adrien Lainé



Dorian Schaal
Sales Representative, AWS



Adrien Lainé
Account Manager, AWS



Greg Autric
Solution Architect, AWS



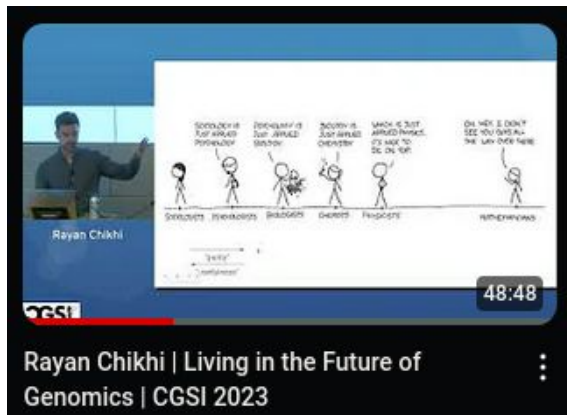
Brice Raffestin
DevOps, Institut Pasteur



Dr. Maxime Hugues
HPC Solution Architect, AWS



Nostalgic of this talk ? CGSI 2023 talk: Living in the future of genomics



6a.48xlarge:~\$ aws s3 cp s3://sra-pub-src-2/SRR11292120/m64062_190806_063919.fastq.1 --no-sign-request
Completed 4.6 GiB/39.1 GiB (278.0 MiB/s) with 1 file(s) remaining

Rethinking bioinformatics analyses using the cloud

96[| Tasks: 45, 263 thr | 192 running
97[| Load average: 84.83 25.05 8.99
98[| Uptime: 01:29:46
99[
100[
101[
102[
103[
104[
105[
106[
107[
108[
109[
110[

Need to Mapquik

Live demo of mapping human HiFi reads in ~seconds, using mapquik

17-JUL-23

123

Thank you for your
attention!

