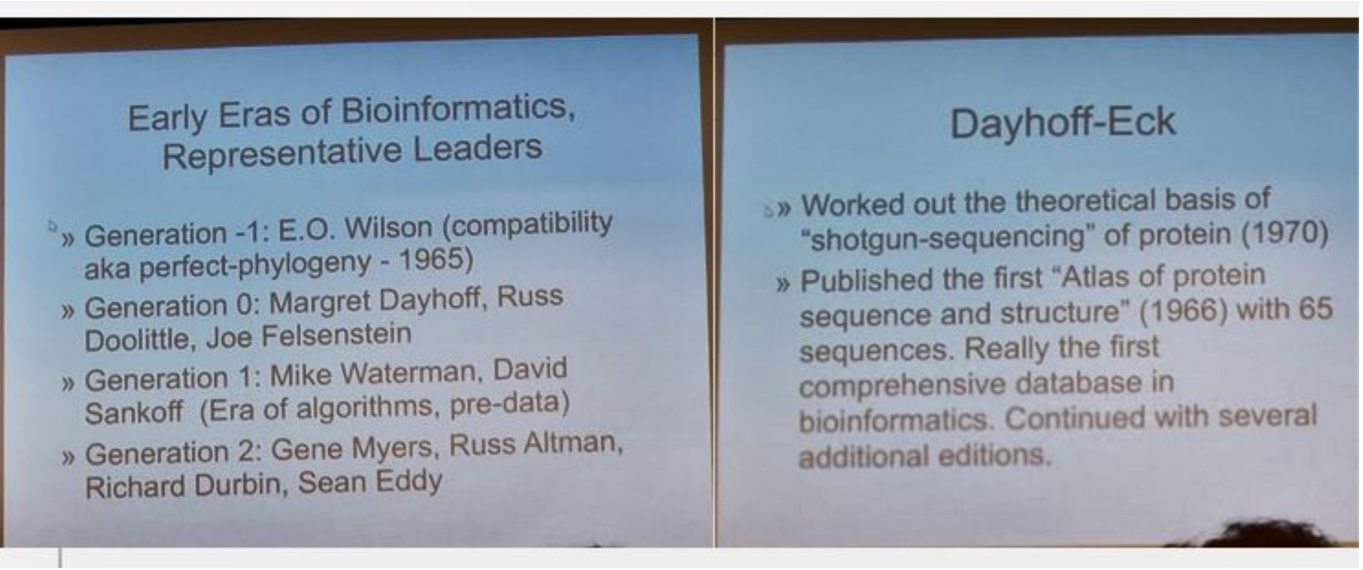
Big data

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



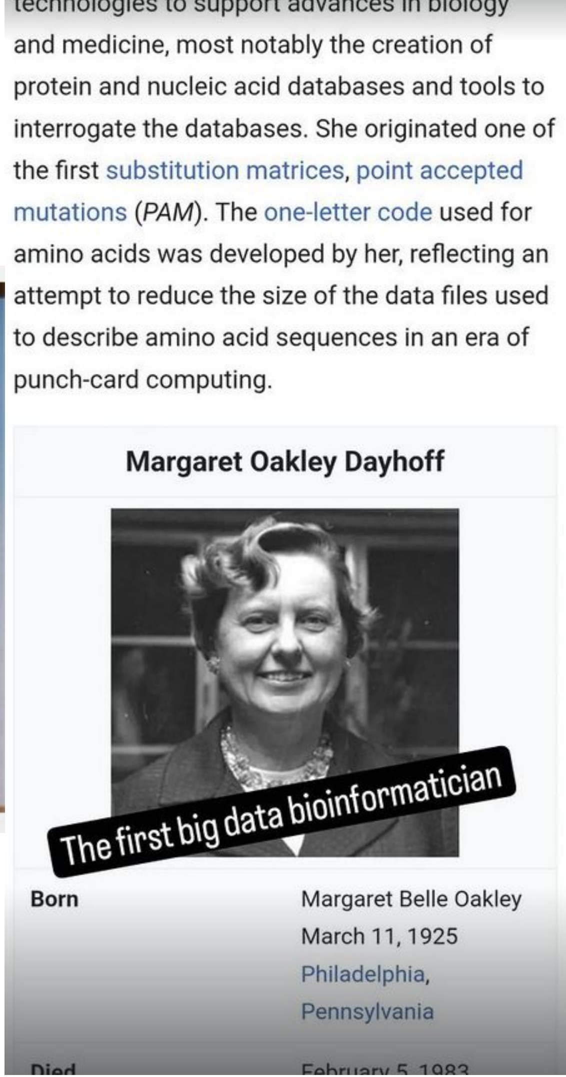
Early Eras of Bioinformatics, Representative Leaders

- 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.

Slides: Dan Gusfield

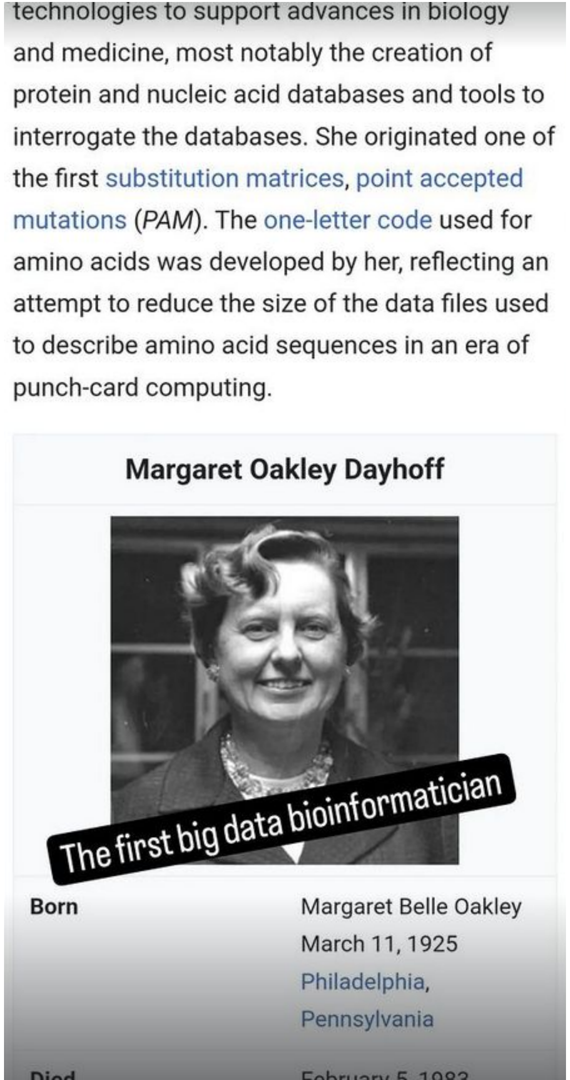


technologies to support advances in biology and medicine, most notably the creation of protein and nucleic acid databases and tools to interrogate the databases. She originated one of the first substitution matrices, point accepted mutations (*PAM*). The one-letter code used for amino acids was developed 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, 1983

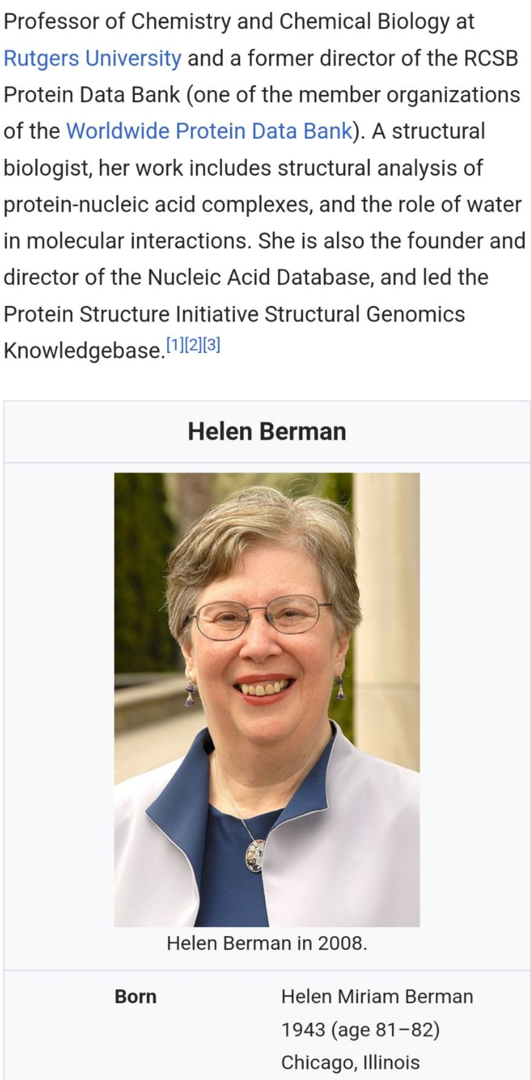


technologies to support advances in biology and medicine, most notably the creation of protein and nucleic acid databases and tools to interrogate the databases. She originated one of the first substitution matrices, point accepted mutations (*PAM*). The one-letter code used for amino acids was developed 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, 1983



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

Big data is the natural flow of biology

1972: single gene sequenced

2000: 1 high-quality human genome

2013: many low-quality human genomes

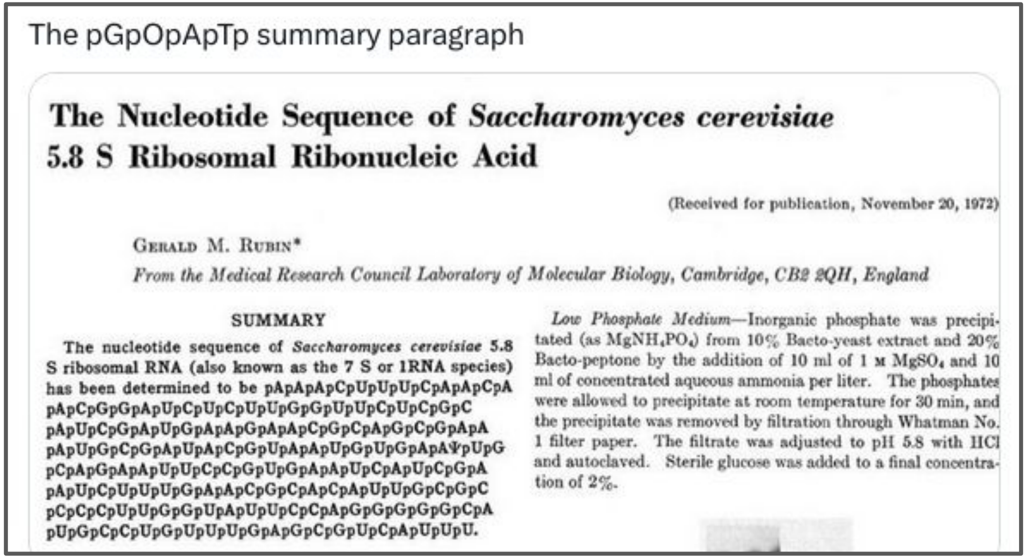
2021: 10 petabases of reads analyzed

2022: 1 million humans VCFs

2022: 50 high-quality human genomes

2024–: ?

Data size



The pGpOpApTp summary paragraph

The Nucleotide Sequence of *Saccharomyces cerevisiae* 5.8 S Ribosomal Ribonucleic Acid

(Received for publication, November 20, 1972)

Gerald M. Rubin*

From the Medical Research Council Laboratory of Molecular Biology, Cambridge, CB2 2QH, England

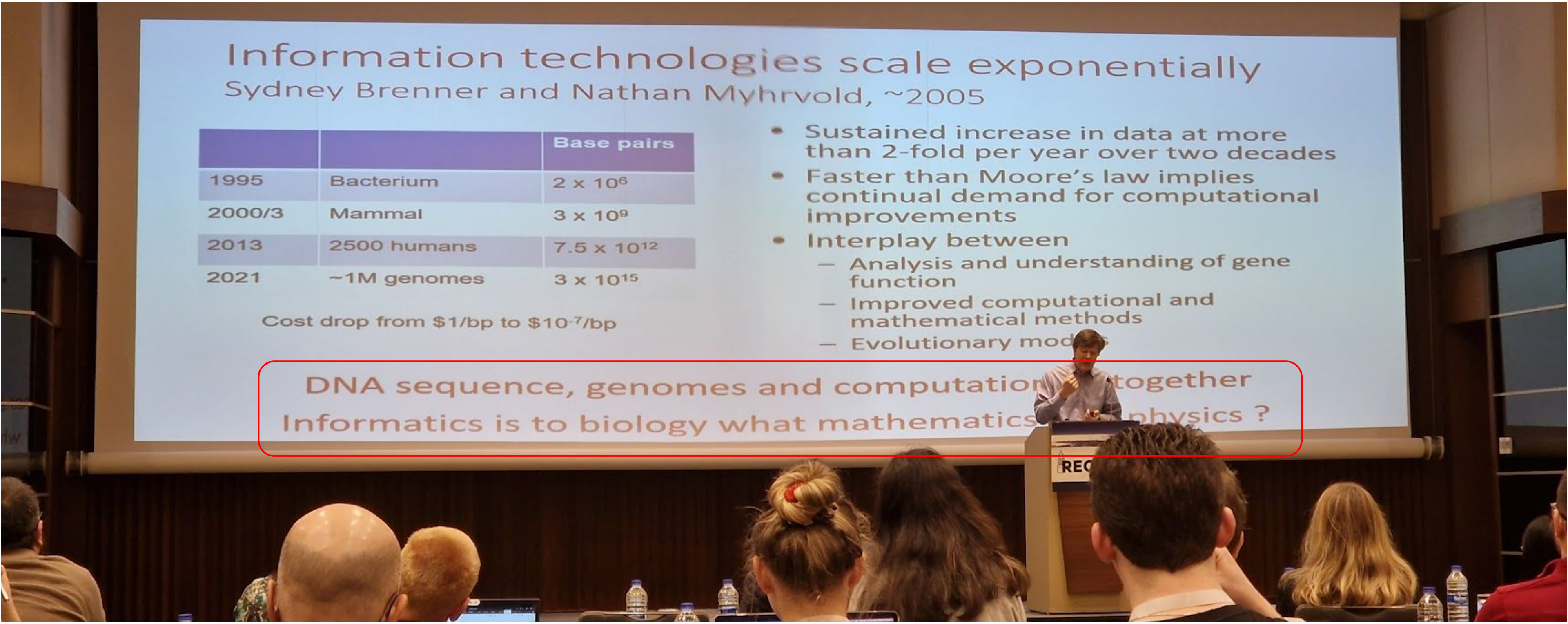
SUMMARY

The nucleotide sequence of *Saccharomyces cerevisiae* 5.8 S ribosomal RNA (also known as the 7 S or 1RNA species) has been determined to be pApApApCpUpUpUpCpApApCpA pApCpGpGpApUpCpUpCpUpUpGpGpUpUpCpUpCpGpC pApUpCpGpApUpGpApApGpApApCpGpCpApGpCpGpApA pApUpGpCpGpApUpApCpGpUpApApUpGpUpGpApAΨpUpG pCpApGpApApUpUpCpCpGpUpGpApApUpCpApUpCpGpA pApUpCpUpUpUpGpApApCpGpCpApCpApUpUpGpCpGpC pCpCpCpUpUpGpGpUpApUpUpCpCpApGpGpGpGpGpCpA pUpGpCpCpUpGpUpUpUpGpApGpCpGpUpCpApUpUpU.

Low Phosphate Medium—Inorganic phosphate was precipitated (as $MgNH_4PO_4$) from 10% Bacto-yeast extract and 20% Bacto-peptone by the addition of 10 ml of 1 M $MgSO_4$ and 10 ml of concentrated aqueous ammonia per liter. The phosphates were allowed to precipitate at room temperature for 30 min, and the precipitate was removed by filtration through Whatman No. 1 filter paper. The filtrate was adjusted to pH 5.8 with HCl and autoclaved. Sterile glucose was added to a final concentration of 2%.

Credit: @SynBio1

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

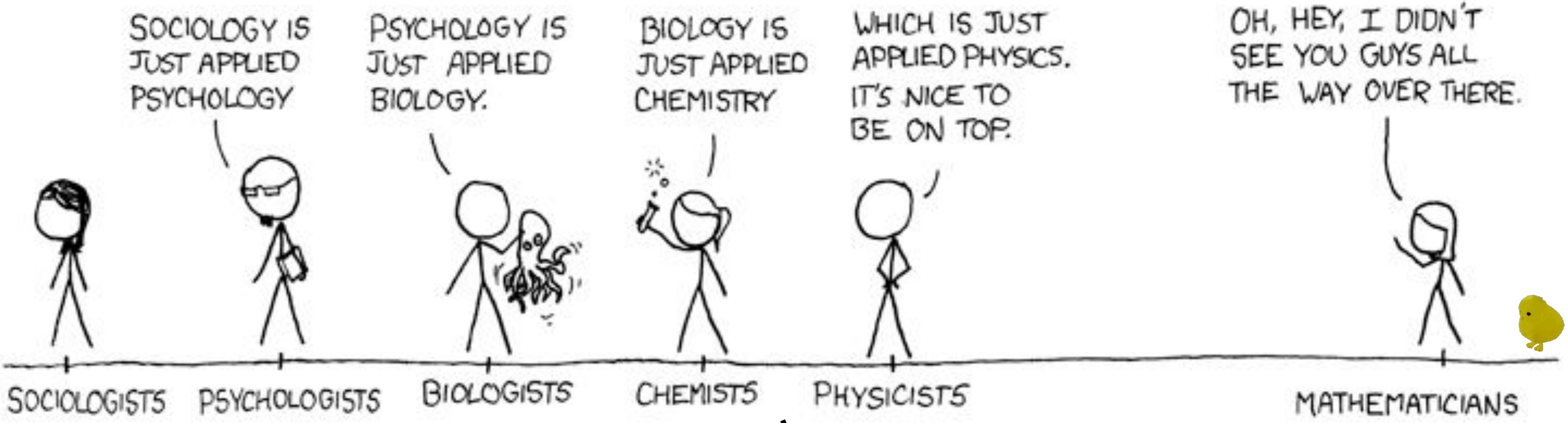
- Sustained increase in data at more than 2-fold per year over two decades
- Faster than Moore's law implies continual demand for computational improvements
- Interplay between
  - Analysis and understanding of gene function
  - Improved computational and mathematical methods
  - Evolutionary models

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


SOCIOLOGY IS JUST APPLIED PSYCHOLOGY
PSYCHOLOGY IS JUST APPLIED BIOLOGY.
BIOLOGY IS JUST APPLIED CHEMISTRY
WHICH IS JUST APPLIED PHYSICS. IT'S NICE TO BE ON TOP.
OH, HEY, I DIDN'T SEE YOU GUYS ALL THE WAY OVER THERE.
SOCIOLOGISTS
PSYCHOLOGISTS
BIOLOGISTS
CHEMISTS
PHYSICISTS
MATHEMATICIANS

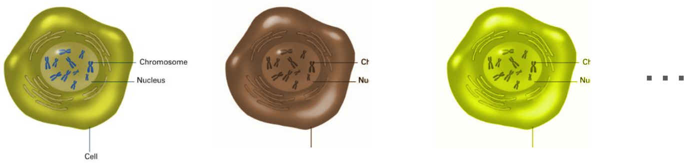
Informatics?

“purity”

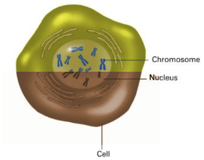
“usefulness”

Types of genomic data

- **Raw sequencing data**
  - Error-prone (~1-10% per base)
  - Abundant (petabytes)
  - Contains inter-cell diversity



- **Reconstructed genomes**
  - High quality (<0.001% per base)
  - Rare (gigabytes)
  - Collapses inter-cell diversity

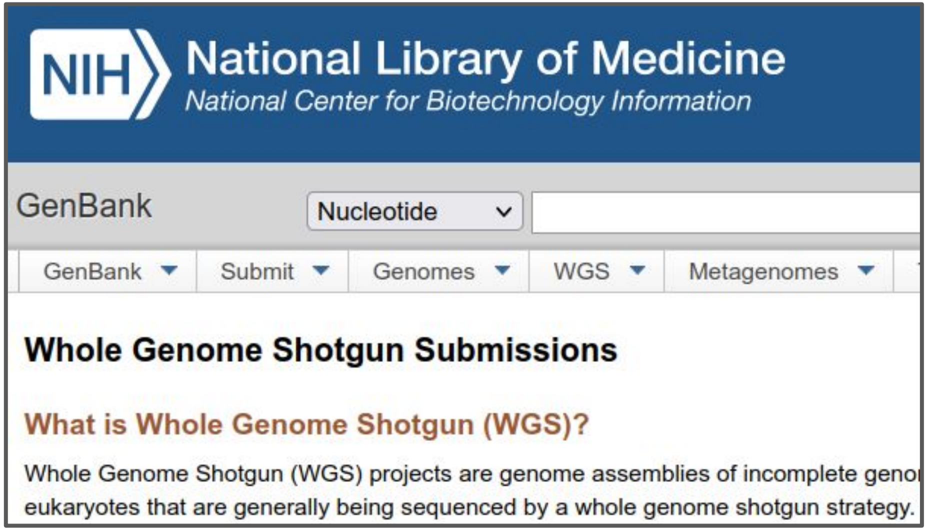


Big data in biology: NCBI GenBank & WGS

GenBank Overview

What is GenBank?

GenBank® is the NIH genetic sequence database,


NIH
National Library of Medicine
National Center for Biotechnology Information
GenBank
Nucleotide
GenBank
Submit
Genomes
WGS
Metagenomes

Whole Genome Shotgun Submissions

What is Whole Genome Shotgun (WGS)?

Whole Genome Shotgun (WGS) projects are genome assemblies of incomplete geno
eukaryotes that are generally being sequenced by a whole genome shotgun strategy.

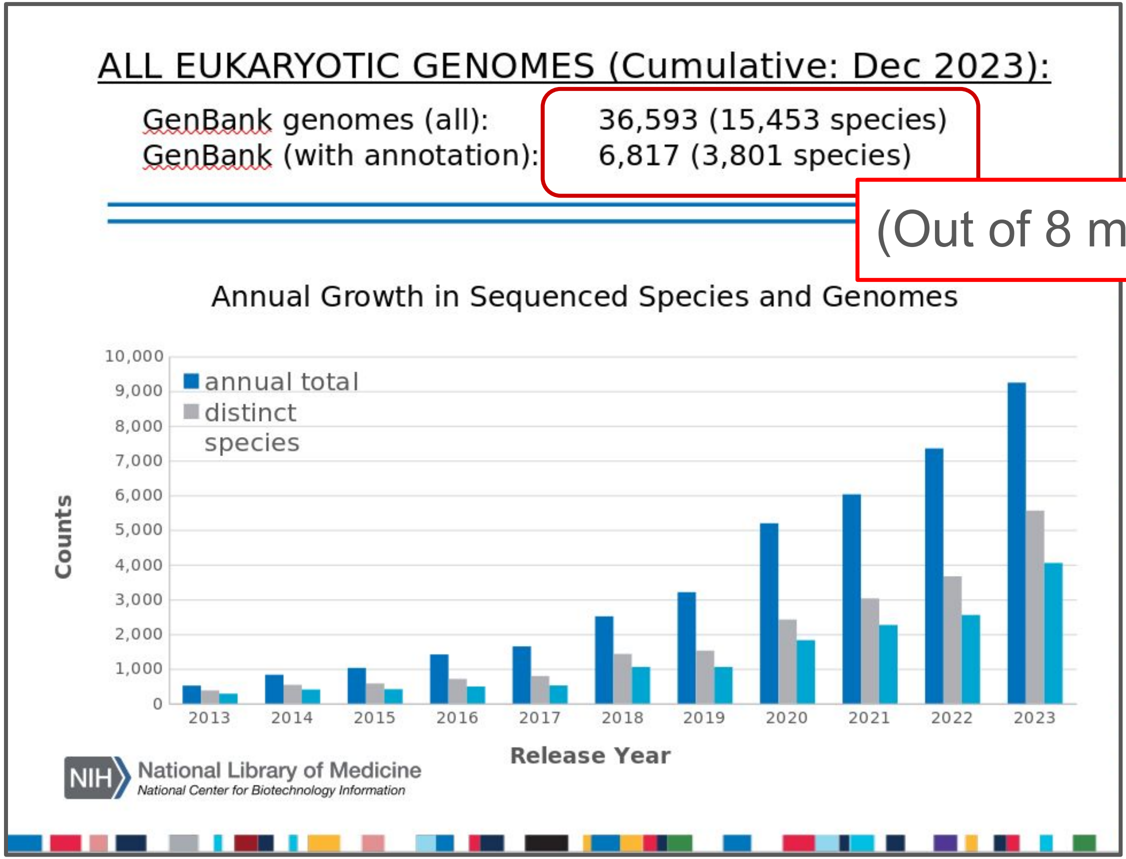
Type: genome assemblies of >500,000 species
Size: 1.2 terabytes (TB) (2022)

All sequences are *annotated*

Type: genome assemblies
Size: 16 TB (2022)

Unannotated

How complete are those databases?



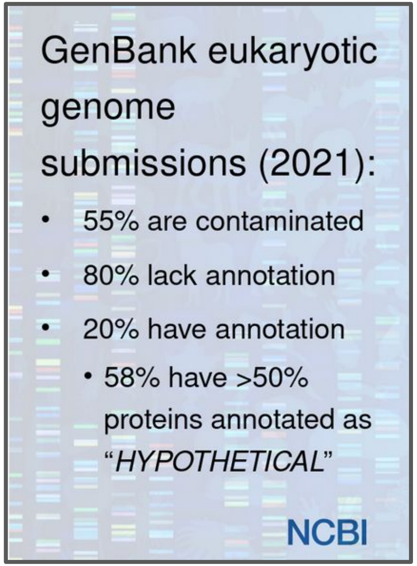
ALL EUKARYOTIC GENOMES (Cumulative: Dec 2023):

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

Annual Growth in Sequenced Species and Genomes

NIH National Library of Medicine
National Center for Biotechnology Information

(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*"

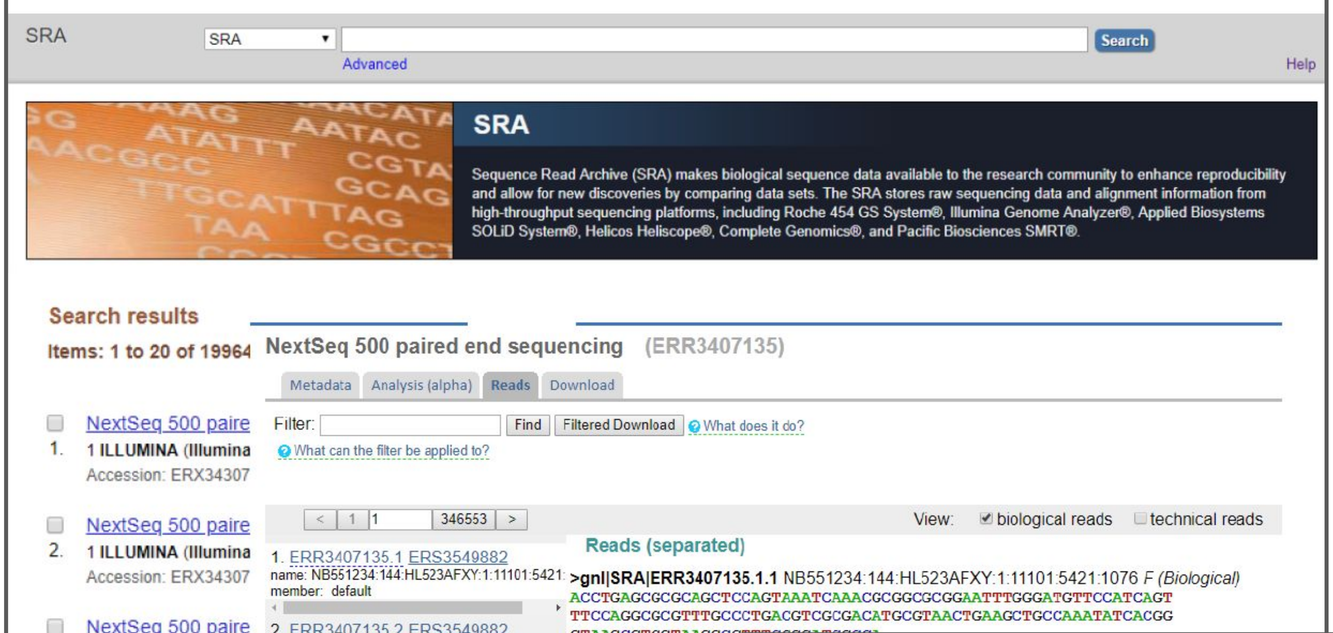
NCBI

Slide credit: Terence Murphy, NCBI

NCBI SRA

All public sequencing reads

Size: 50 Pbases as of Dec 2023



SRA

Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®.

Search results

Items: 1 to 20 of 19964

□ 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

NextSeq 500 paired end sequencing (ERR3407135)

1. ERR3407135.1 ERS3549882
name: NB551234:144:HL523AFXY:1:11101:5421:
member: default

2. ERR3407135.2 ERS3549882
name: NB551234:144:HL523AFXY:1:11101:22482
member: default

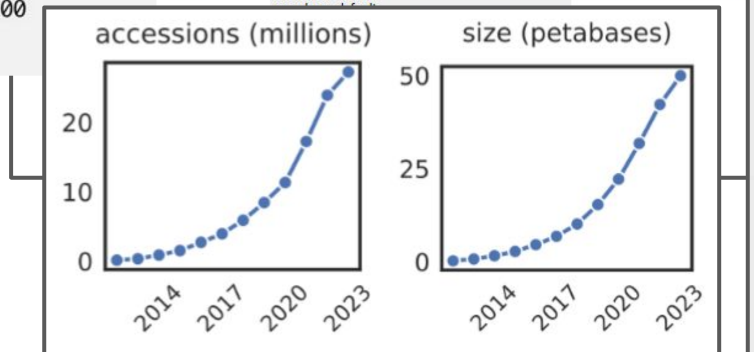
3. ERR3407135.3 ERS3549882
name: NB551234:144:HL523AFXY:1:11101:2566:

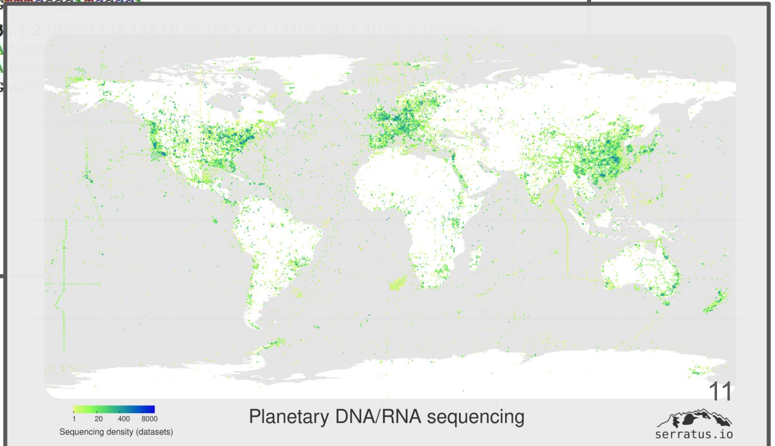
Reads (separated)

>gnl|SRA|ERR3407135.1.1 NB551234:144:HL523AFXY:1:11101:5421:1076 *F (Biological)*
ACCTGAGCGCGCAGCTCCAGTAAATCAAACGCGGCGCGGAATTTGGGATGTTCCATCAGT
TTCCAGGCGCGTTTGCCCTGACGTCGCGACATGCGTAACTGAAGCTGCCAAATATCACGG
GTAAGCGTGGTAAGGCG

>gnl|SRA|ERR340713
ATCAACAACAGCGGGAA
TCACCGAAACCGCGACA
AAACCGCATCCGAAACG

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



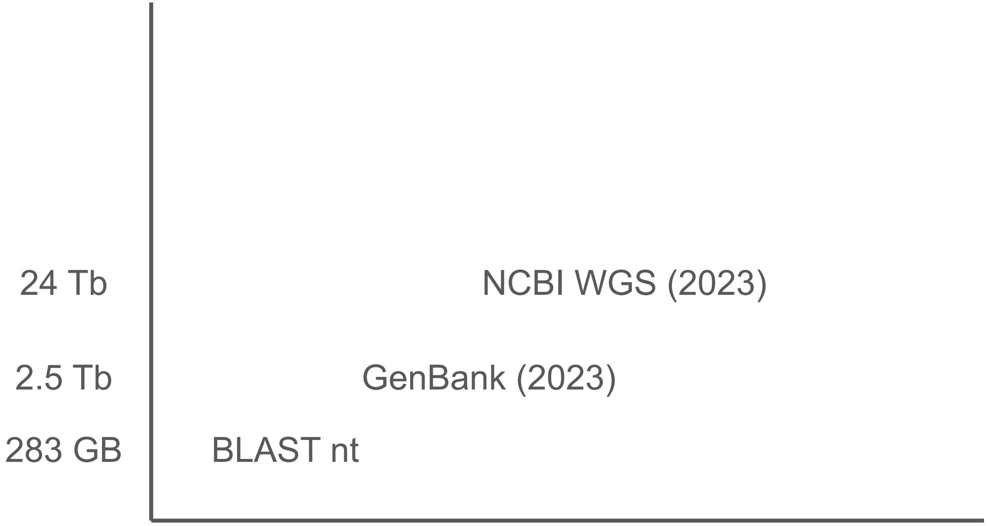


Planetary DNA/RNA sequencing

serratus.io

Public sequence datasets

50 Pb SRA

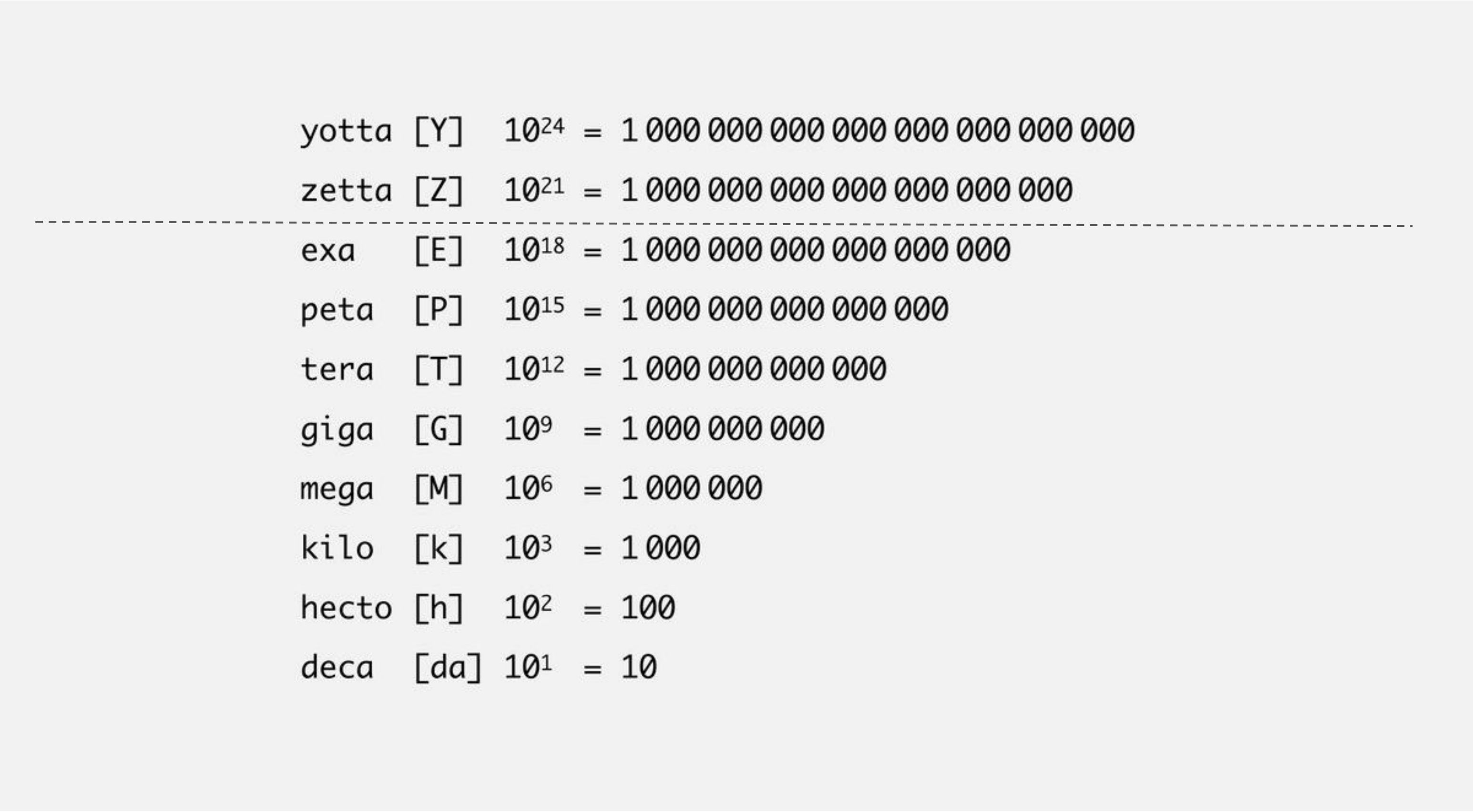


24 Tb NCBI WGS (2023)

2.5 Tb GenBank (2023)

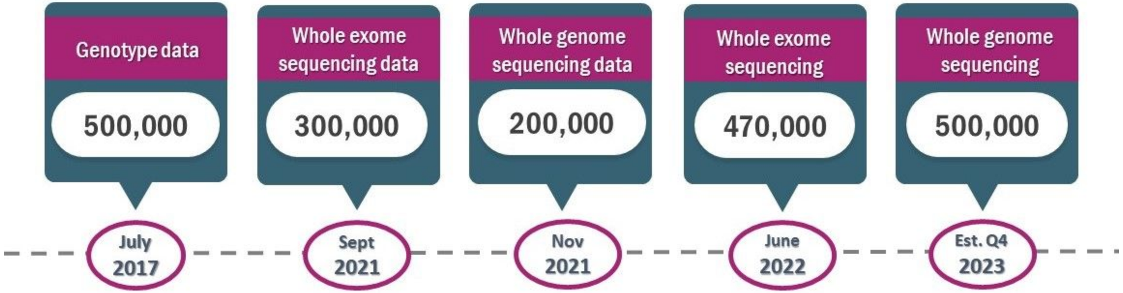
283 GB BLAST nt

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
exa	[E]	10^{18}	=	1 000 000 000 000 000 000
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
kilo	[k]	10^{3}	=	1 000
hecto	[h]	10^{2}	=	100
deca	[da]	10^{1}	=	10

UK Biobank



Genotype data 500,000 July 2017

Whole exome sequencing data 300,000 Sept 2021

Whole genome sequencing data 200,000 Nov 2021

Whole exome sequencing 470,000 June 2022

Whole genome sequencing 500,000 Est. Q4 2023

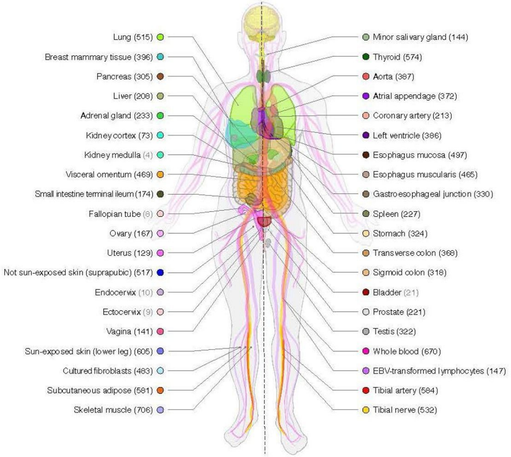
Size: 25+ PB

source: https://twitter.com/uk_biobank/status/1578023831578427393

Type: reads*

* but many use just the SNPs

GTEx



Size: 150 TB

from: https://www.genomeweb.com/informatics/anvil-platform-makes-popular-nhgri-gtex-database-free-download

Type: reads*

* but many use just the expression data

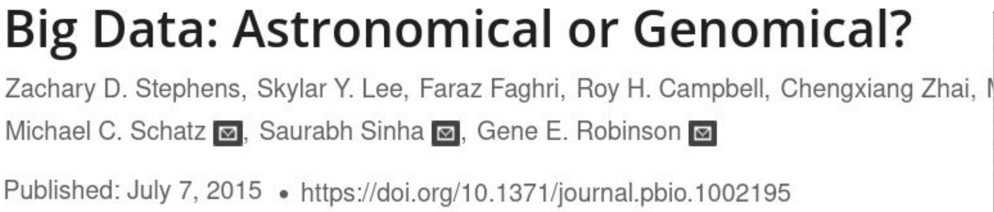
(Youtube: 300 PB)

NCBI SRA database : 50 PB

Institut Pasteur: 10 PB

Your laptop: 0.001 PB

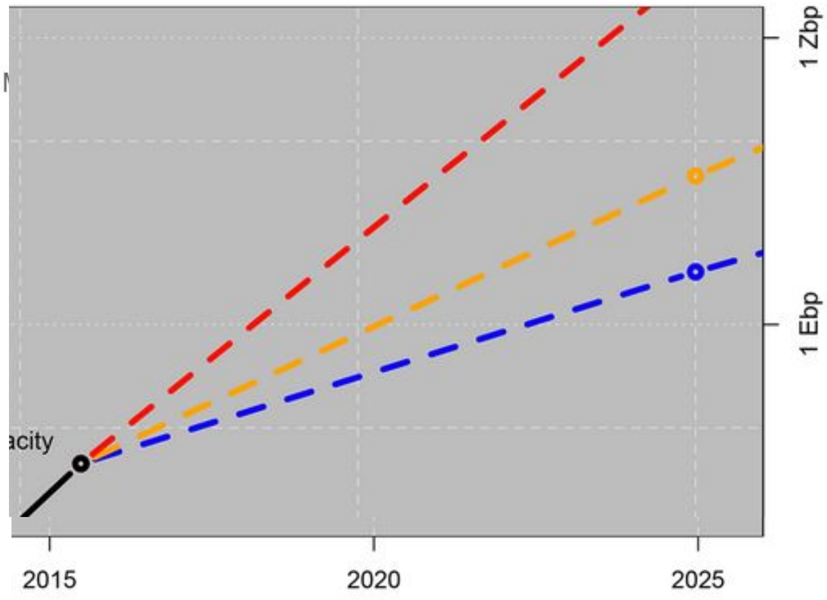
PLOS Biology



Big Data: Astronomical or Genomical?

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

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



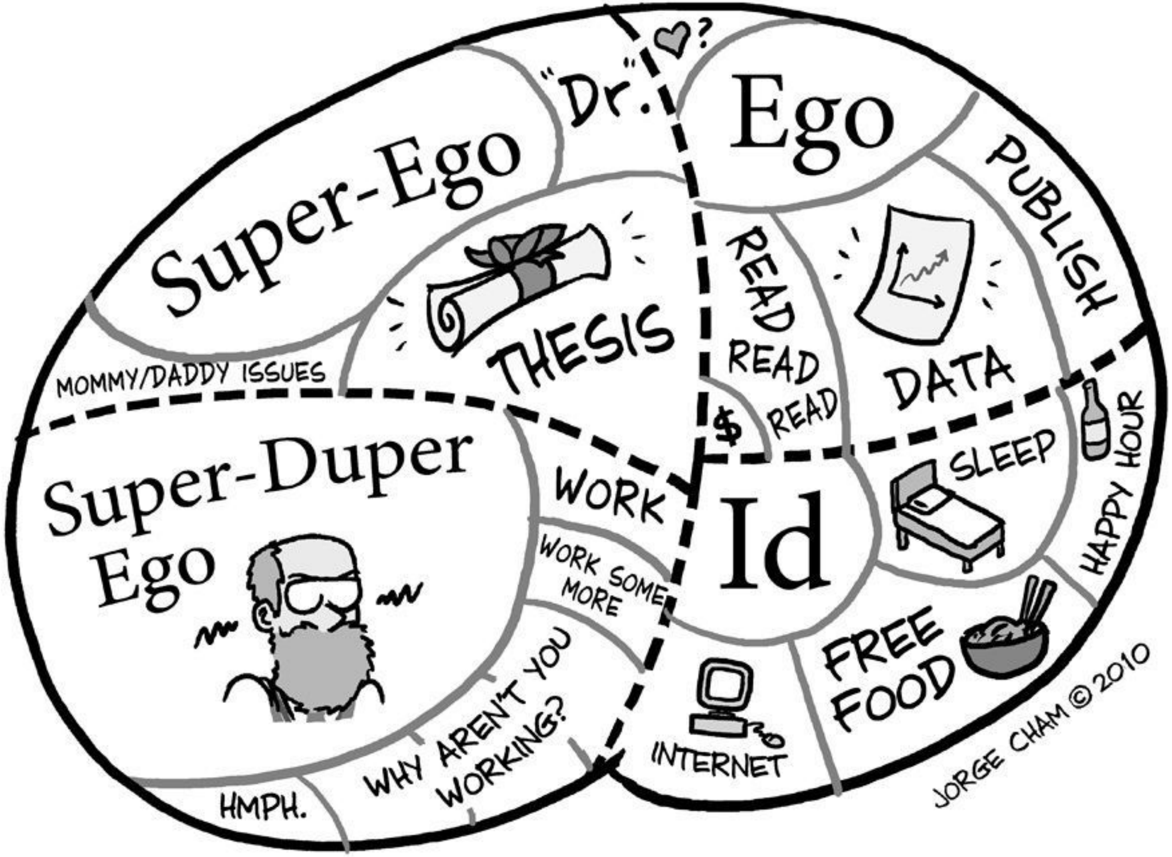
- Projected 5 exabytes - 1 zettabyte of seq data in 2025
- Actual: 5 exabytes based # sequencers in world (total capacity: 45,000,000 human genomes per year)

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

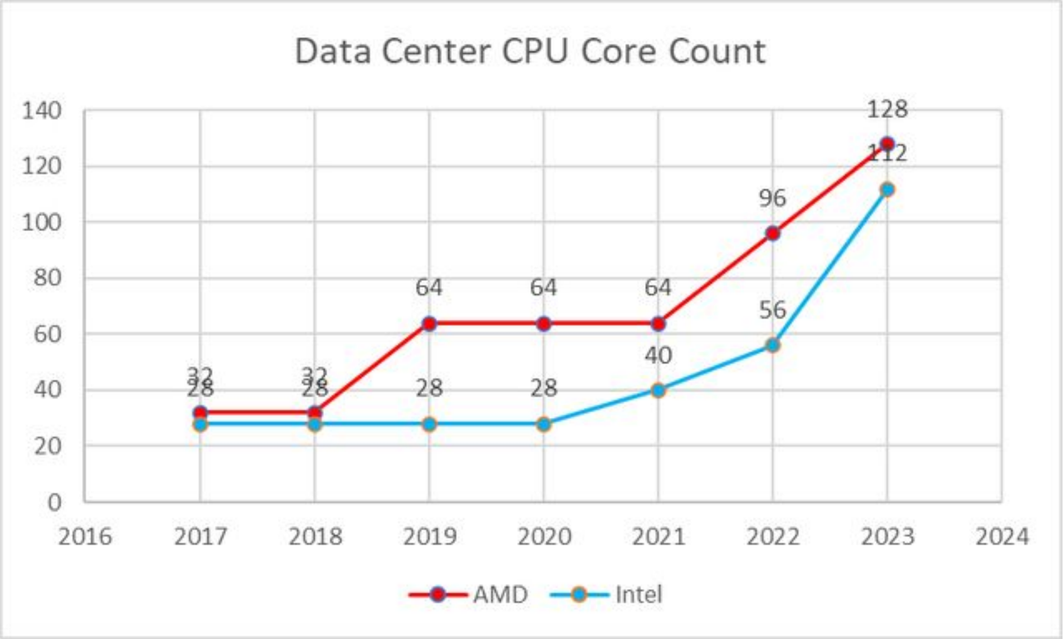
- Big computers, Cluster, Cloud
- Storage management
- Galaxy
- Knowledge of scaling limits
- Knowledge of cloud costs
- Parallel execution
- AI

Data mining

- Pebblescout, branchwater
- ORA
- deCOM
- SRA metadata

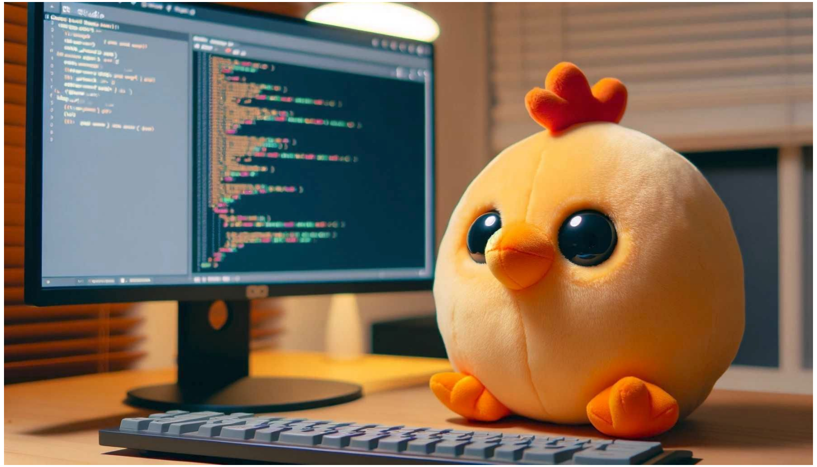
Future genomics, today?

‘big data’
=
‘small computers’



source
https://seekingalpha.com/article/4468119-advanced-micro-devices-amd-server-roadmap-not-strong-enough

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

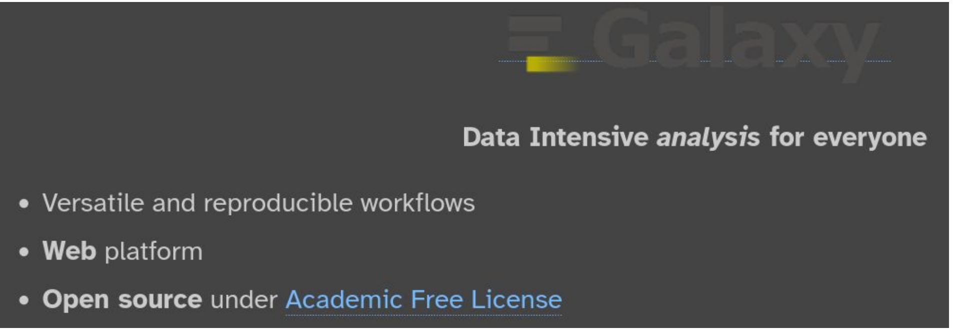
```
squeue -o "%.18i %.9P %.8j %.8u %.2t %.10M %.6D %R cores:%c mem:%m cmd:%o " | grep seqbio
```

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`

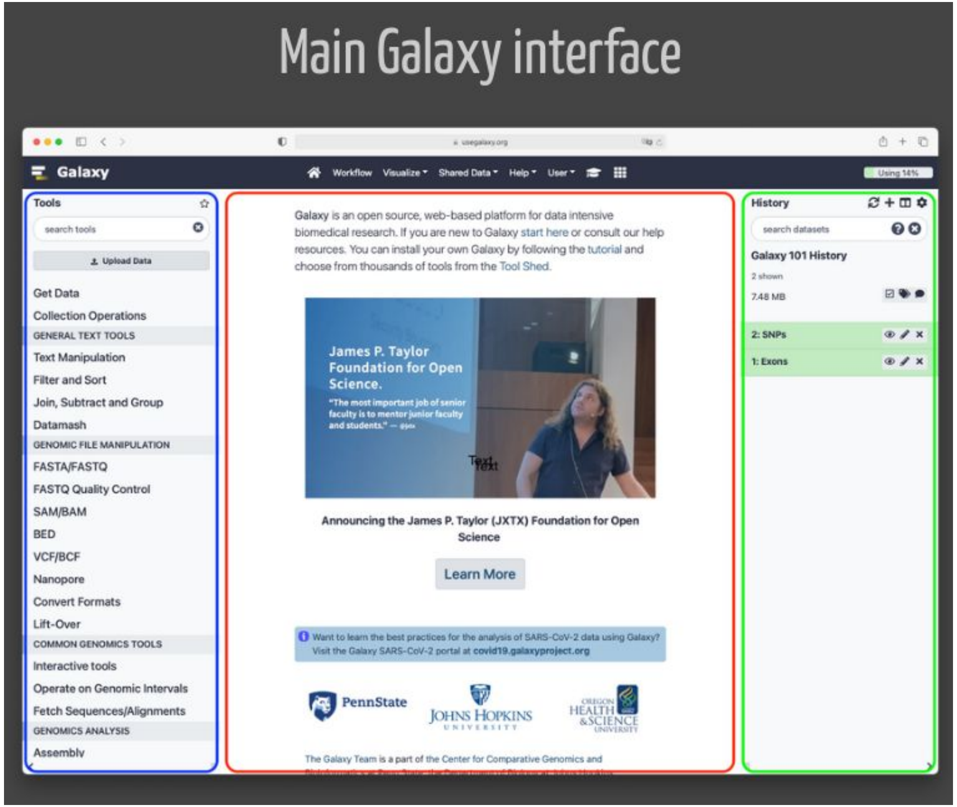
Galaxy Project



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

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.

Announcing the James P. Taylor (JXTX) Foundation for Open Science

Want to learn the best practices for the analysis of SARS-CoV-2 data using Galaxy? Visit the Galaxy SARS-CoV-2 portal at **covid19.galaxyproject.org**

The Galaxy Team is a part of the Center for Comparative Genomics and

Cloud

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

LES DATA CENTERS DANS LE MONDE :
LES DATA CENTERS EN FRANCHE-COMTÉ :

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

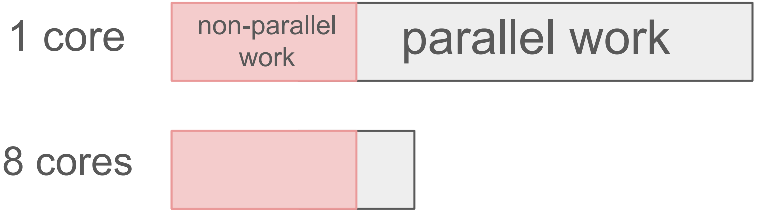
Knowledge of scaling limits

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:

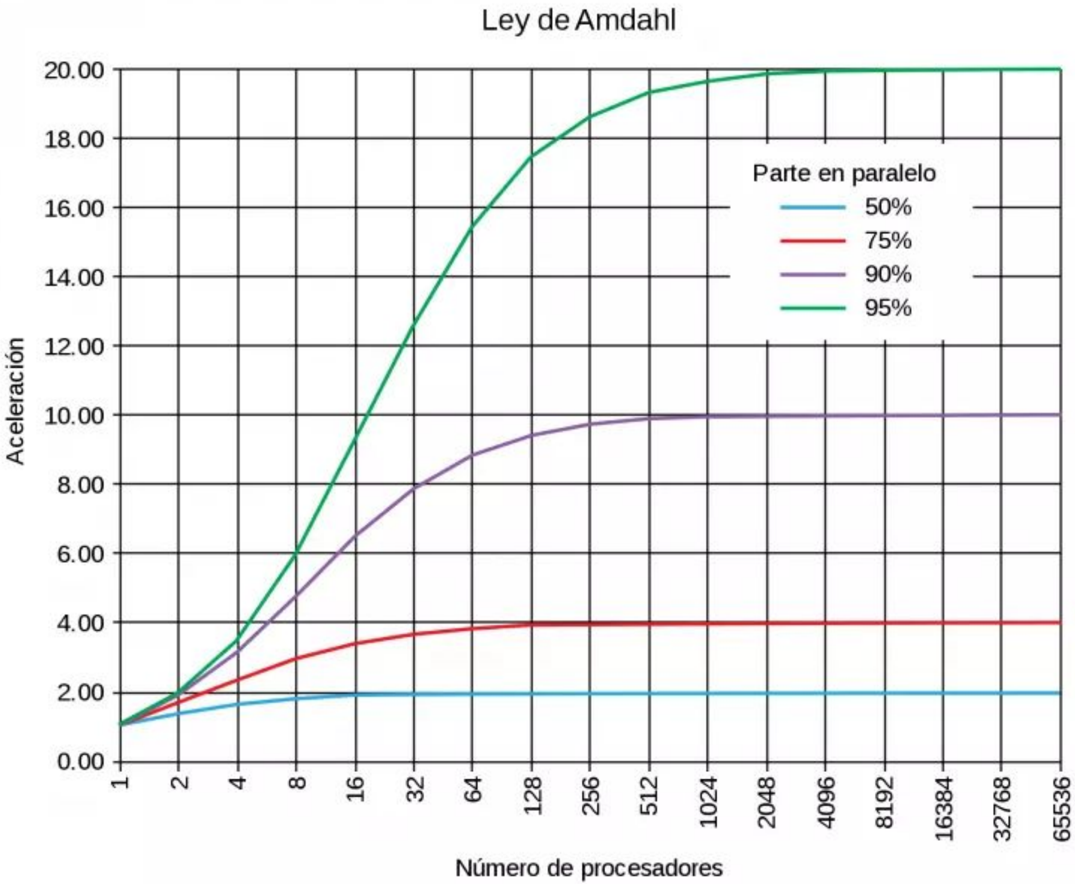


1 core

non-parallel work

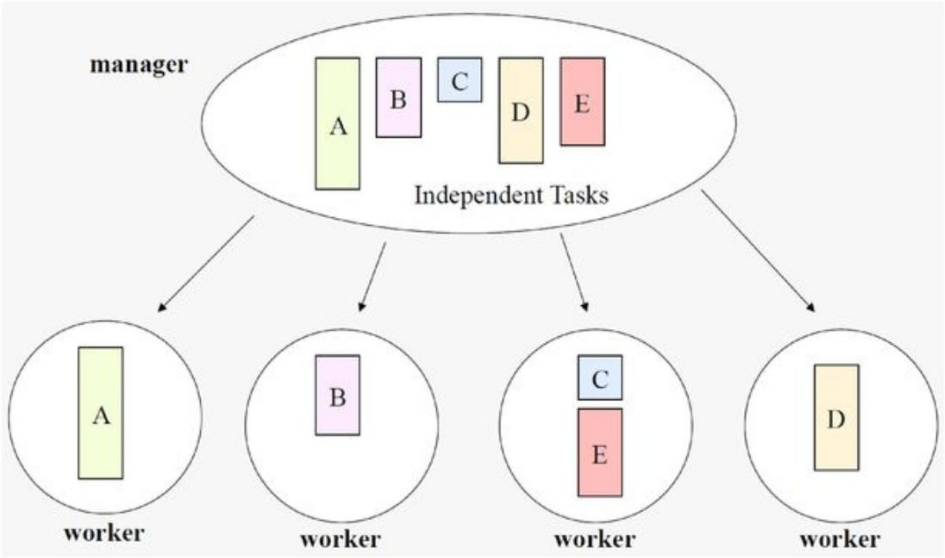
parallel work

8 cores

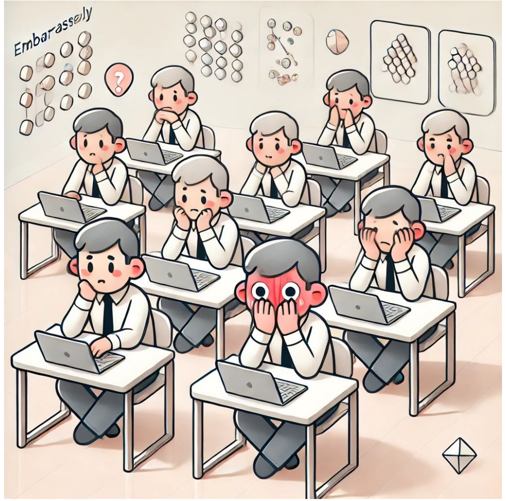


Except..

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



https://slideplayer.com/slide/12789171/



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

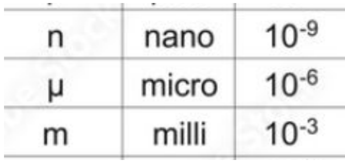
2) Access data in memory

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

4) Human cell cycle

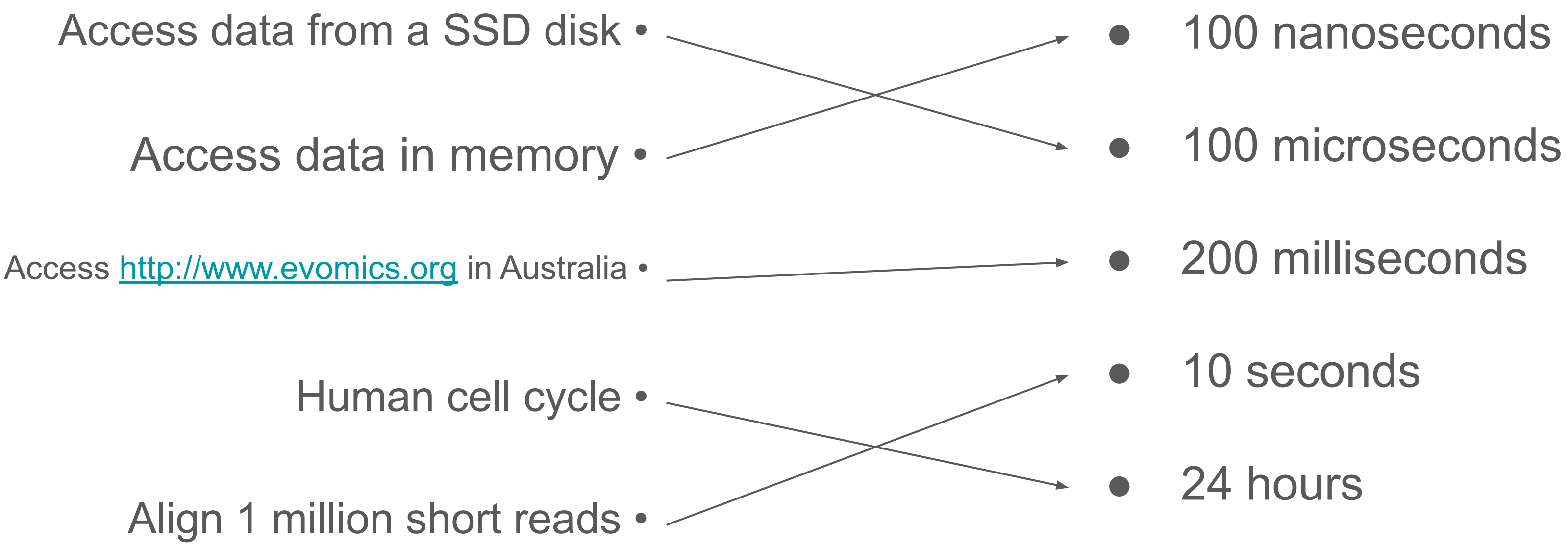
5) Align 1 million short reads

- 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



Access data from a SSD disk •

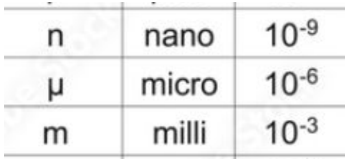
Access data in memory •

Access http://www.evomics.org in Australia •

Human cell cycle •

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}

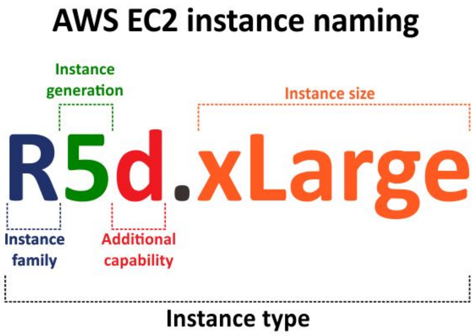
Knowledge of scaling limits

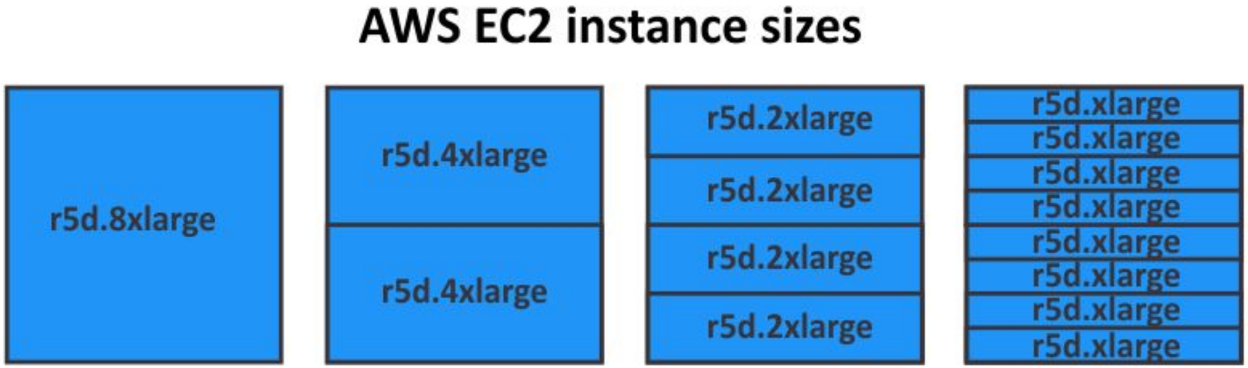
In order of difficulty:

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





💕 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

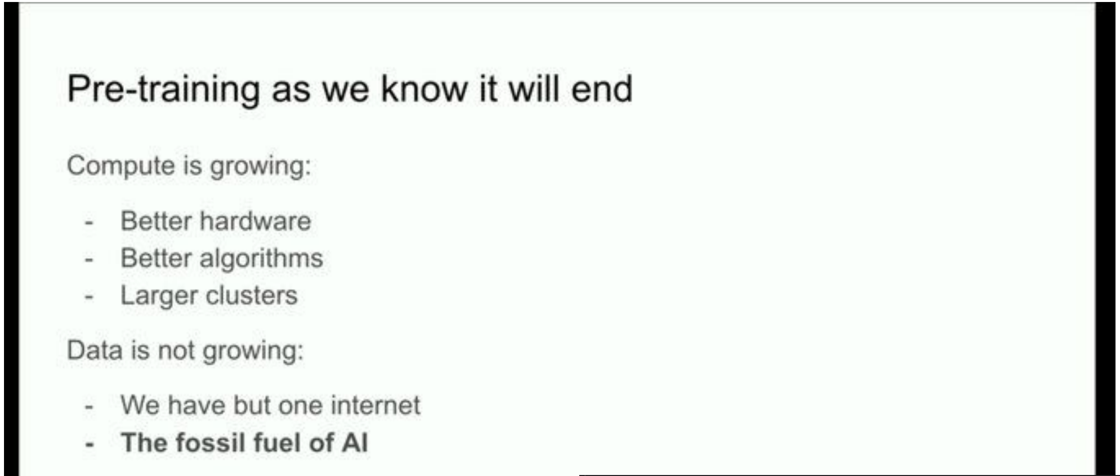
- **Alignment**
  - Highly parallel, low memory, scales well with number of CPUs
- **Assembly**
  - Moderately parallel, high memory, single big machine
- **Annotation**
  - Don’t! (jk), but moderately parallel. Single machine too
- **Phylogenomics**
  - Can be made parallel (RAxML, Iq-Tree)

Can you draw me an image that contains a group of clocks showing the time of 3 minutes past 12

AI in bioinformatics

Sutskever @ NeurIPS'24

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



Pre-training as we know it will end

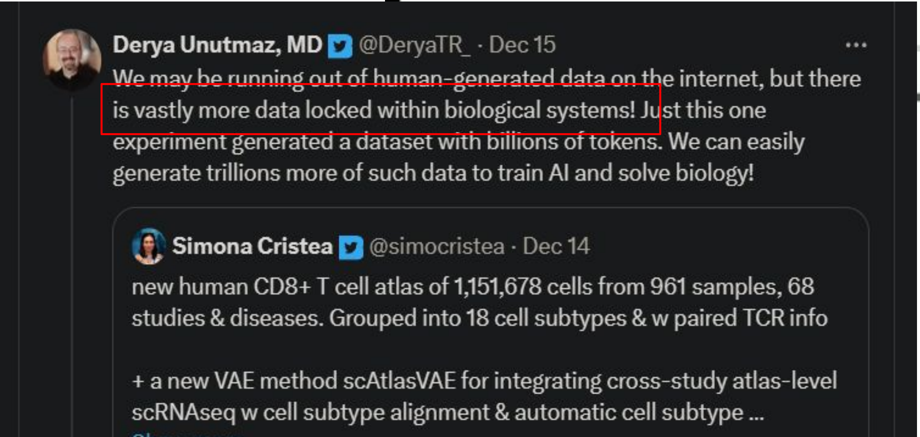
Compute is growing:

- Better hardware
- Better algorithms
- Larger clusters

Data is not growing:

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

Responses:



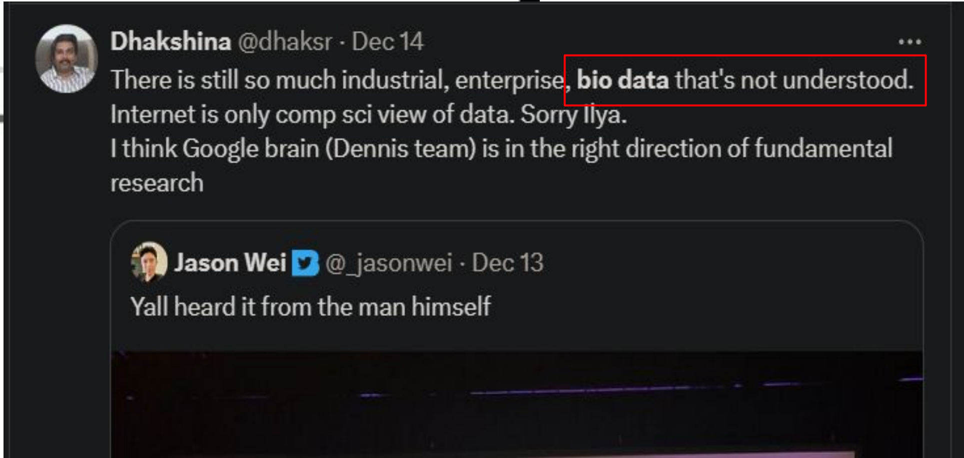
Derya Unutmaz, MD @DeryaTR_ · Dec 15

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 ...



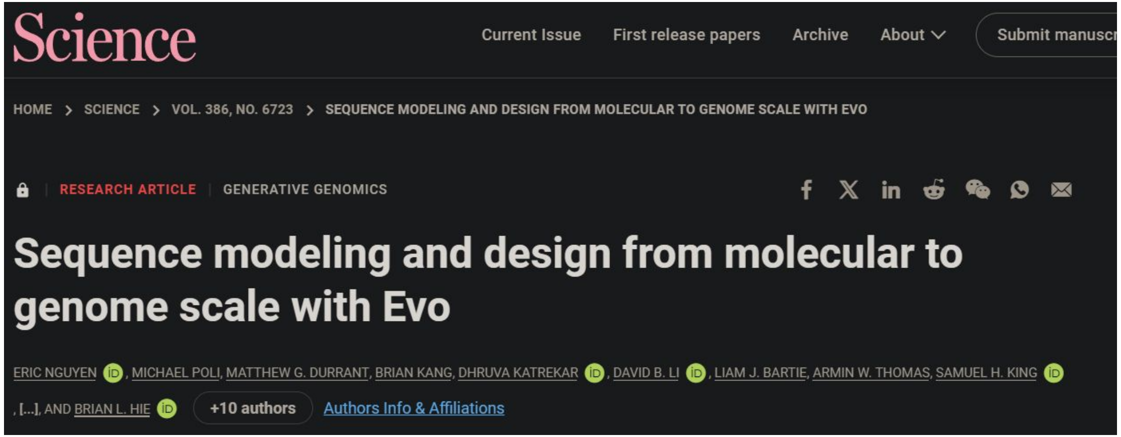
Dhakshina @dhaksr · Dec 14

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



HOME > SCIENCE > VOL. 386, NO. 6723 > SEQUENCE MODELING AND DESIGN FROM MOLECULAR TO GENOME SCALE WITH EVO

RESEARCH ARTICLE | GENERATIVE GENOMICS

Sequence modeling and design from molecular to genome scale with Evo

ERIC NGUYEN, MICHAEL POLI, MATTHEW G. DURRANT, BRIAN KANG, DHRUVA KATREKAR, DAVID B. LI, LIAM J. BARTIE, ARMIN W. THOMAS, SAMUEL H. KING, [...], AND BRIAN L. HIE +10 authors Authors Info & Affiliations

“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.

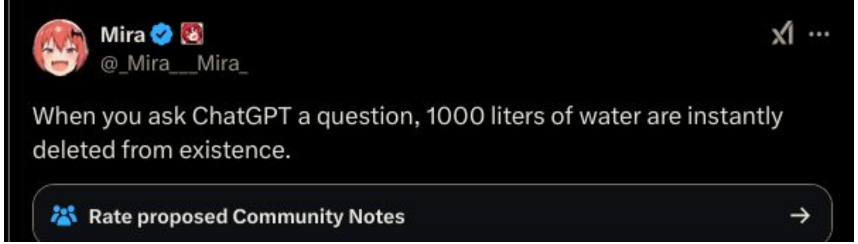
We make wholesale extraction, transformation and analysis of open web data accessible to researchers.

~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



Mira @_Mira___Mira_

When you ask ChatGPT a question, 1000 liters of water are instantly deleted from existence.

“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

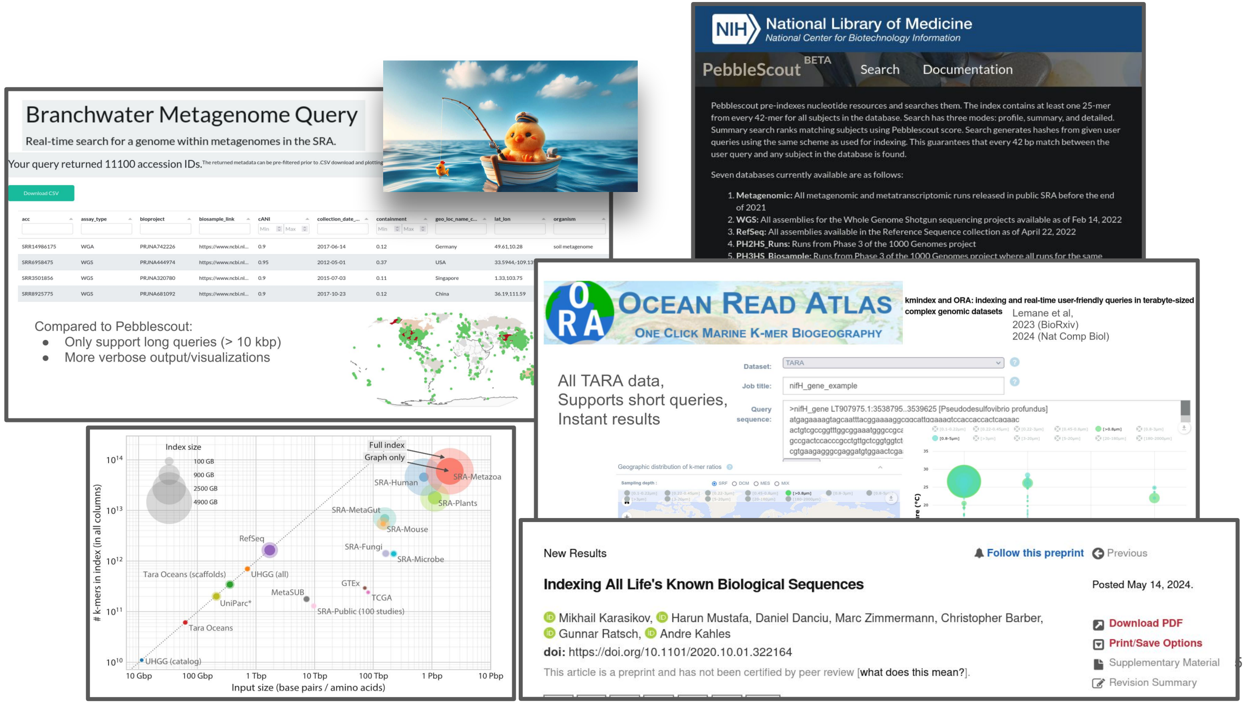
- Big computers, Cloud, Cluster
- Storage management
- Galaxy
- Knowledge of scaling limits
- Knowledge of cloud costs
- GNU parallel
- AI

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.



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

acc	assay_type	bioproject	biosample_link	cANI	collection_date_...	containment	geo_loc_name_c...	lat_lon	organism
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,-109.13	
SRR3501856	WGS	PRJNA320780	https://www.ncbi.nl...	0.9	2015-07-03	0.11	Singapore	1.33,103.75	
SRR8925775	WGS	PRJNA681092	https://www.ncbi.nl...	0.9	2017-10-23	0.12	China	36.19,111.59	

Compared to Pebblescout:

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

NIH National Library of Medicine
National Center for Biotechnology Information

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

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

Dataset: TARA

Job title: nifH_gene_example

Query sequence:

```
>nifH_gene LT907975.1:3538795..3539625 [Pseudodesulfovibrio profundus]
atgagaaaagtagcaatttacggaaaaggcggcattggaaaatccaccaccactcagaac
actgtcgccggtttggcggaaatgggccgca
gccgactccacccgcctgttgctcggtggtct
cgtgaagagggcgaggatgtggaactcga
```

New Results

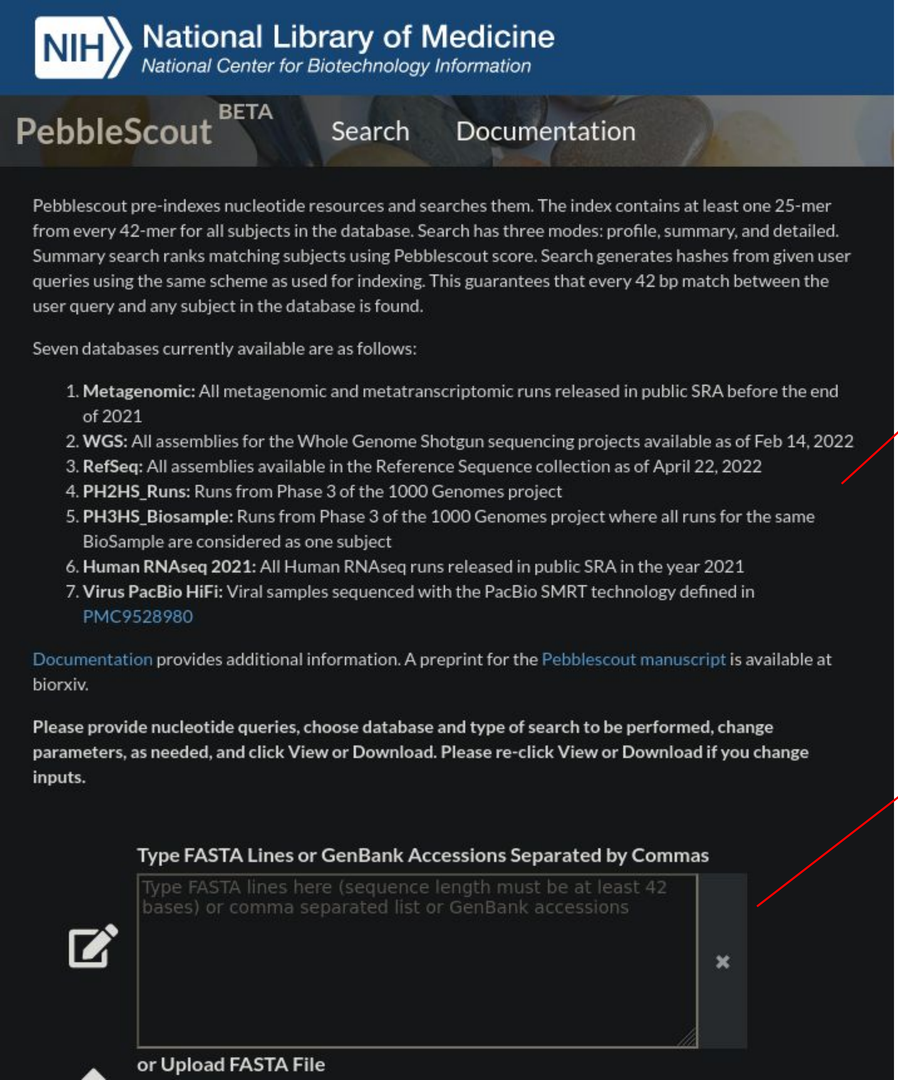
Indexing All Life's Known Biological Sequences

Posted May 14, 2024.

Mikhail Karasikov, Harun Mustafa, Daniel Danciu, Marc Zimmermann, Christopher Barber, Gunnar Ratsch, Andre Kahles

doi: https://doi.org/10.1101/2020.10.01.322164

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


NIH National Library of Medicine
National Center for Biotechnology Information
PebbleScout BETA
Search
Documentation

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)

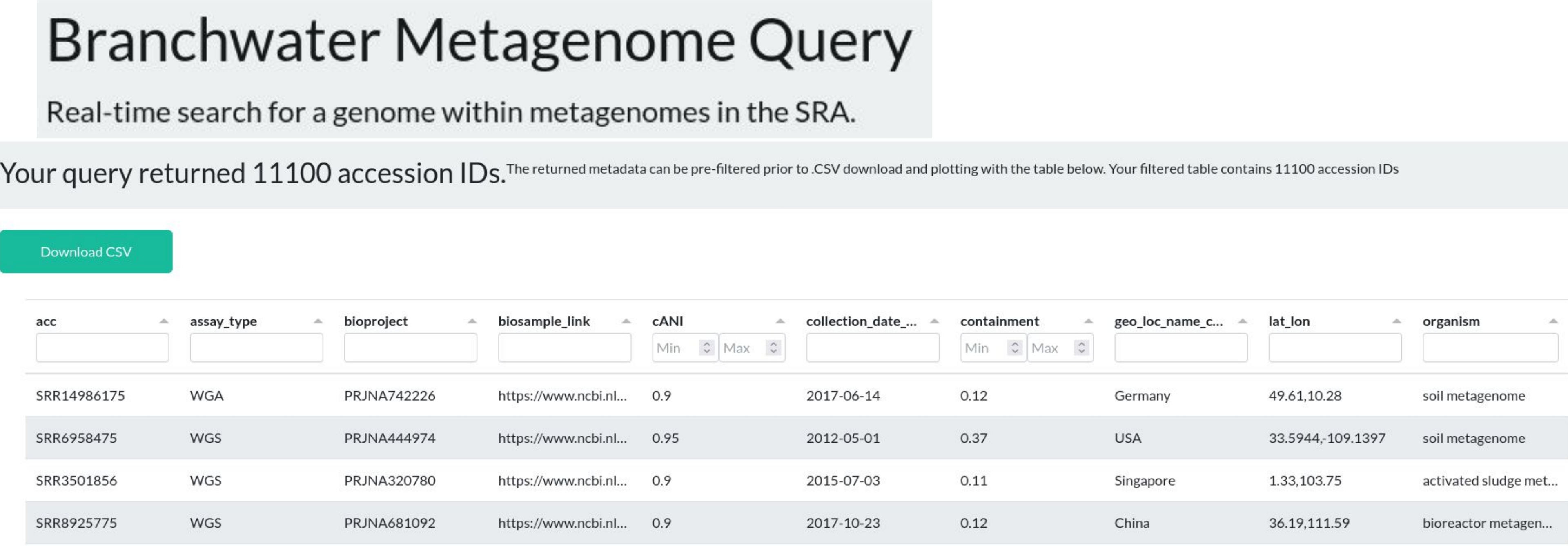
Pebblescout usage example

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

PebbleScout BETA

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

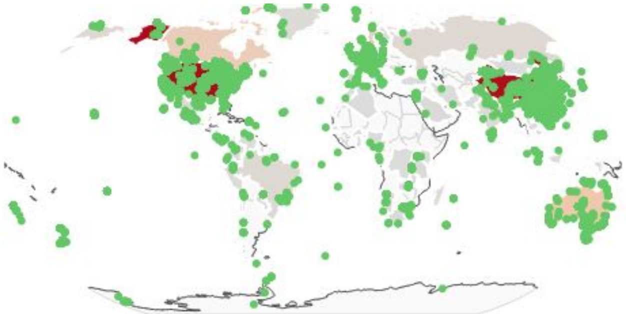
Real-time search for a genome within metagenomes in the SRA.

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

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



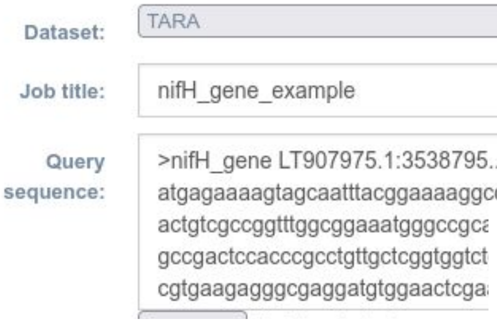
OCEAN READ ATLAS

ONE CLICK MARINE K-MER BIOGEOGRAPHY

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

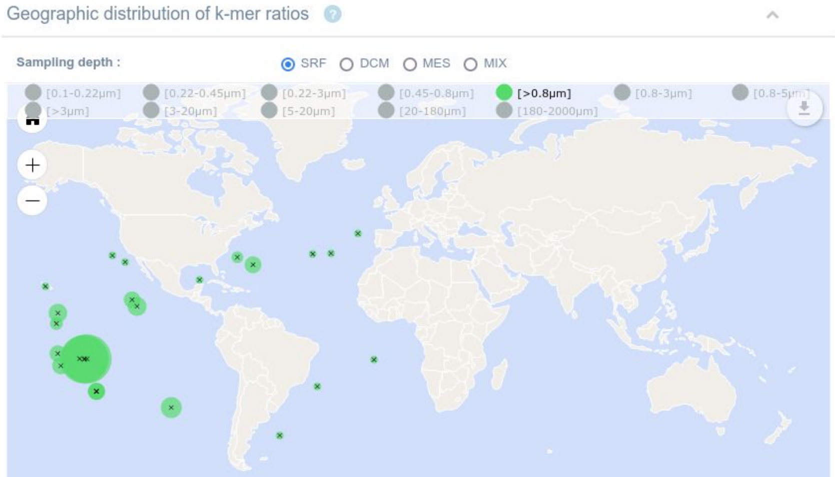


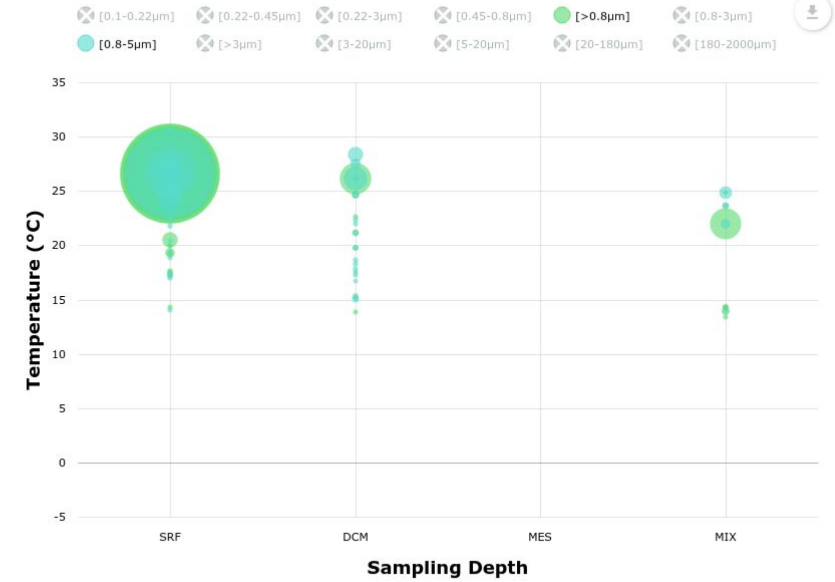
Dataset: TARA

Job title: nifH_gene_example

Query sequence:

```
>nifH_gene LT907975.1:3538795..3539625 [Pseudodesulfovibrio profundus]
atgagaaaagtagcaatttacggaaaaggcggcattggaaaatccaccaccactcagaac
actgtcgccggtttggcggaaatgggccgca
gccgactccacccgcctgttgctcggtggtct
cgtgaagagggcgaggatgtggaactcgaa
```



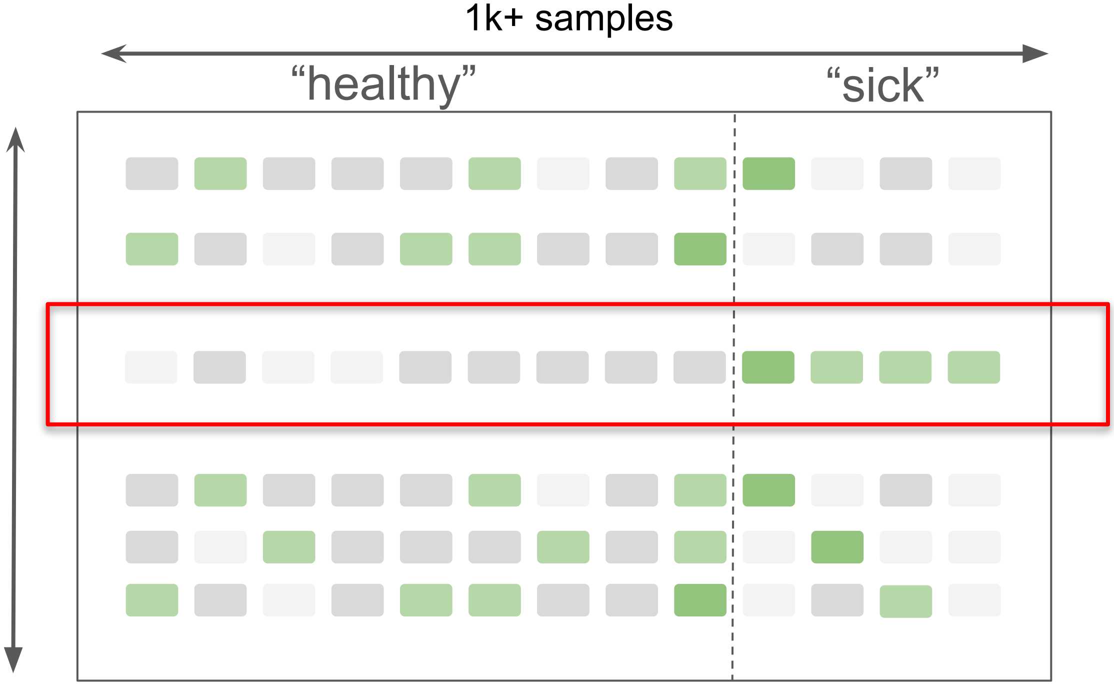


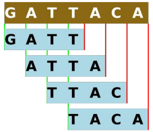
Reference-free tools for detecting variation in large sequencing data cohorts

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

PR[AI]RIE
PaRis Artificial Intelligence Research InstitutE

100M+ rows:
all sequence positions in a genome

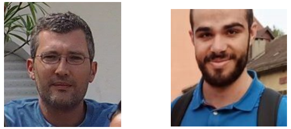




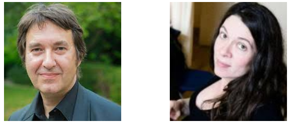
Methods:

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



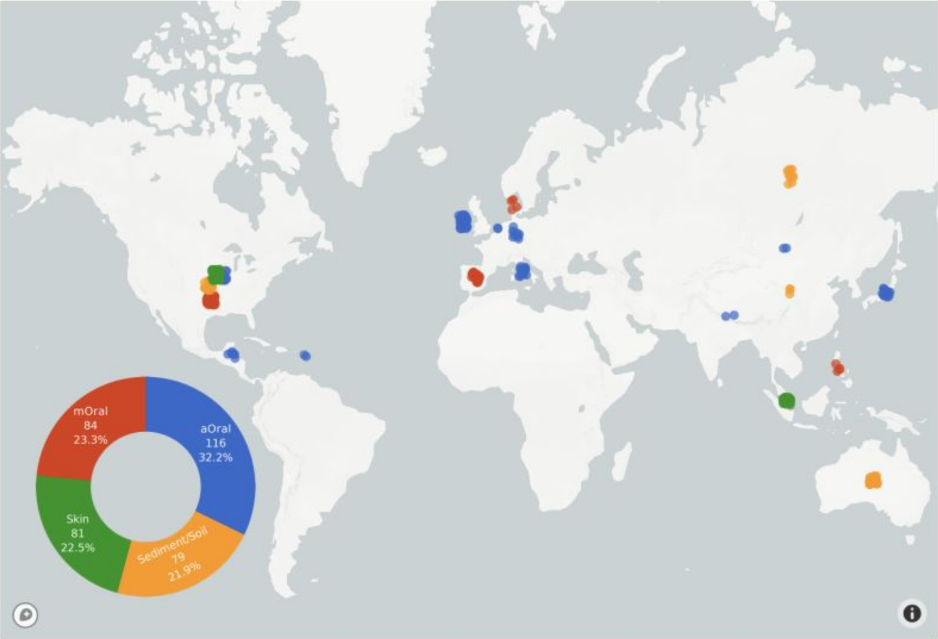
Aschard Lab



Bourgeron Lab

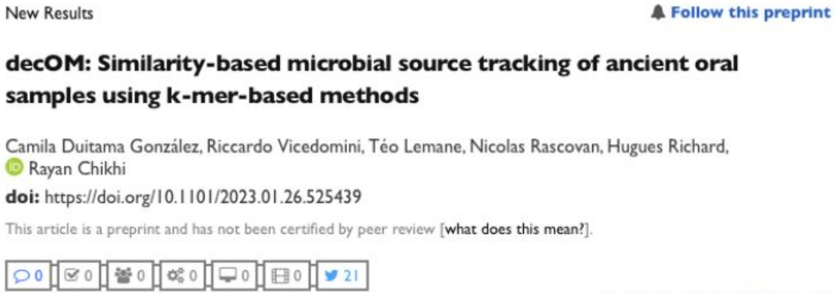
Quintana-Murci Lab

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, Téo 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

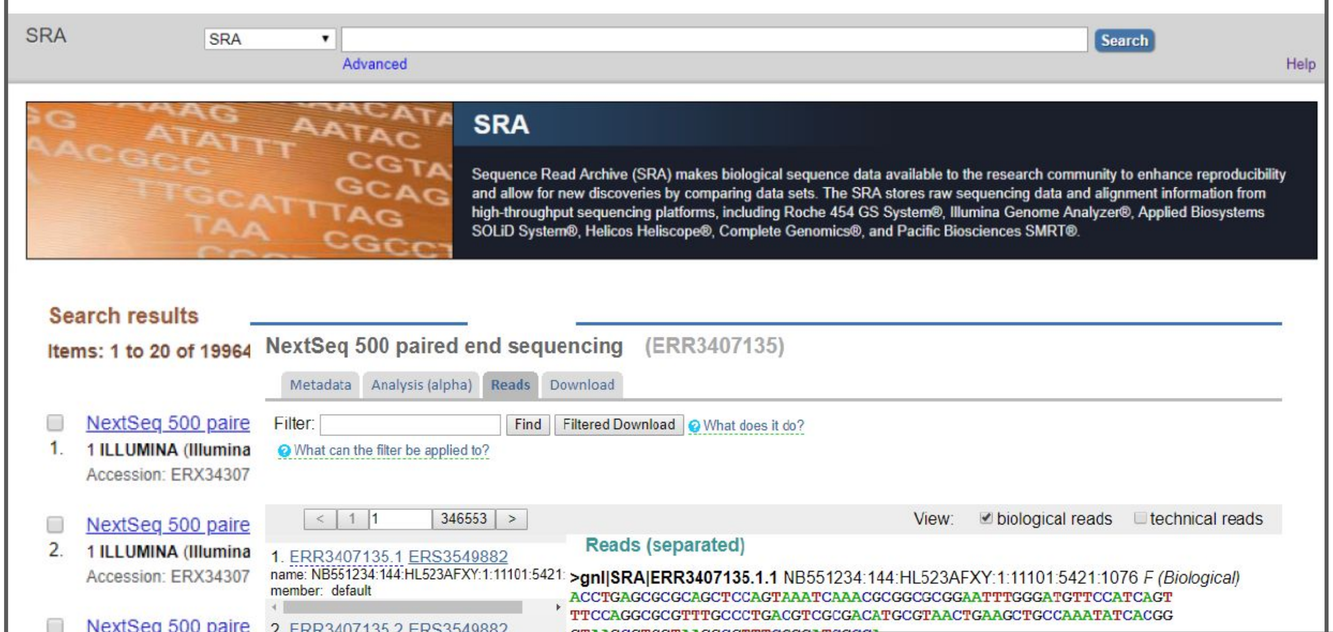
SRA-scale sequence exploration

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


SRA
SRA ▾
Search
Advanced
Help
AAAG ATATTT AACGCC TTGCATTTAG TAA AACATA AATAC CGTA GCAG CGCC

SRA

Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®.

Search results

Items: 1 to 20 of 19964

□ 1. NextSeq 500 paire
1 ILLUMINA (Illumina
Accession: ERX34307

□ 2. NextSeq 500 paire
1 ILLUMINA (Illumina
Accession: ERX34307

□ NextSeq 500 paire
INA (Illumina
n: ERX34307

NextSeq 500 paired end sequencing (ERR3407135)

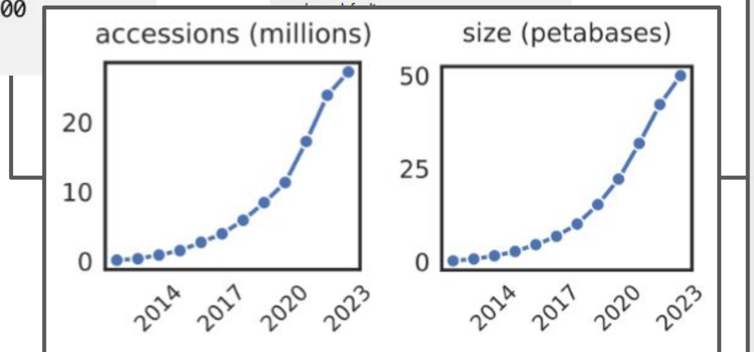
1. ERR3407135.1 ERS3549882
name: NB551234:144:HL523AFXY:1:11101:5421:
member: default

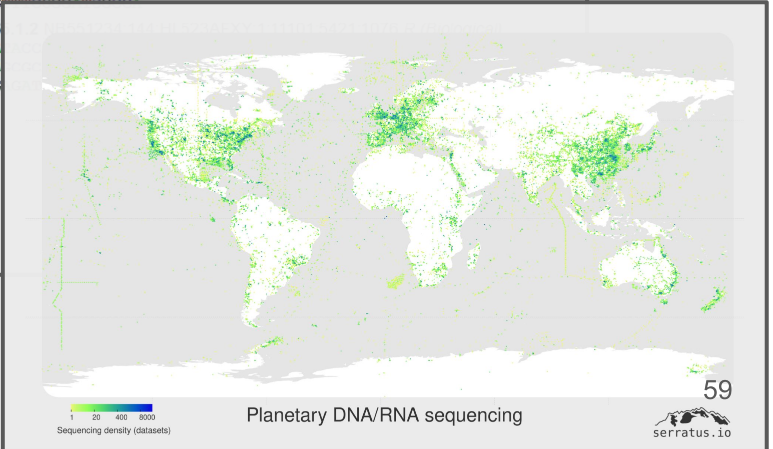
2. ERR3407135.2 ERS3549882
name: NB551234:144:HL523AFXY:1:11101:22482
member: default

3. ERR3407135.3 ERS3549882
name: NB551234:144:HL523AFXY:1:11101:2566:

Reads (separated)

```
>gnl|SRA|ERR3407135.1.1 NB551234:144:HL523AFXY:1:11101:5421:1076 F (Biological)
ACCTGAGCGCGCAGCTCCAGTAAATCAAACGCGGCGCGGAATTTGGGATGTTCCATCAGT
TTCCAGGCGCGTTTGCCCTGACGTCGCGACATGCGTAACTGAAGCTGCCAAATATCACGG
GTAAGCGTGGTAAGGCG
>gnl|SRA|ERR340713
ATCAACAACAGCGGGAA
TCACCGAAACCGCGACA
AAACCGCATCCGAAACG
```





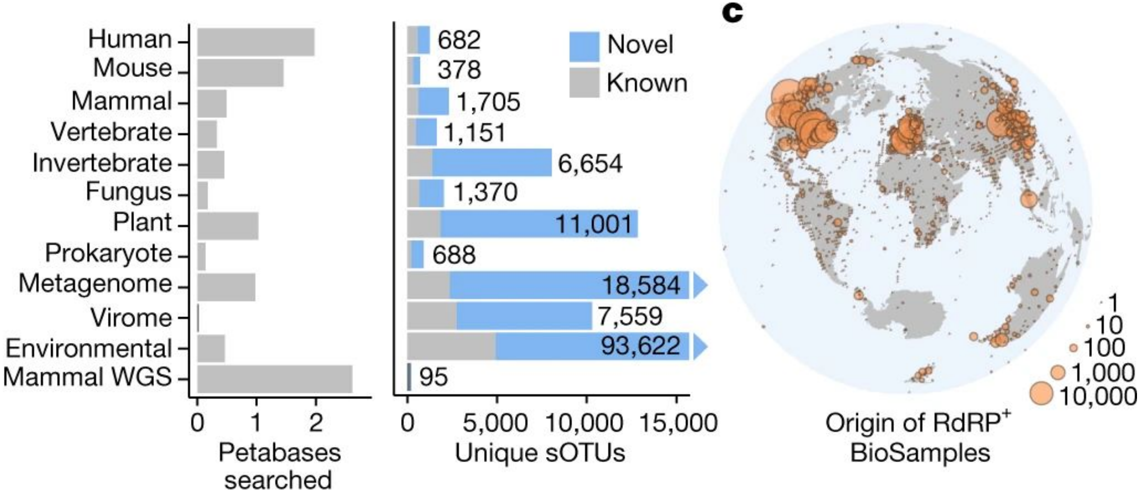
Planetary DNA/RNA sequencing

serratus.io

What to do with the entire SRA?

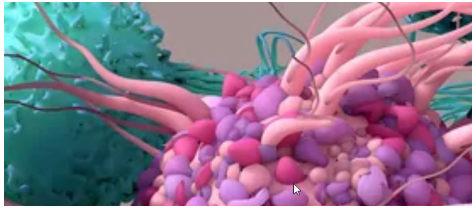
Serratus: all public RNA-seqs analyzed for viral discovery

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)



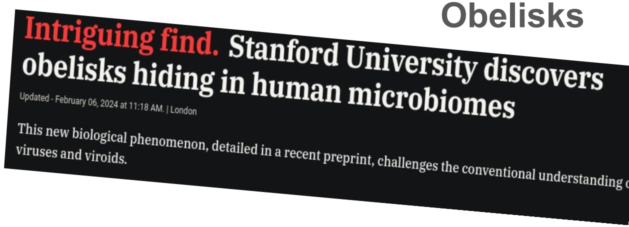
Some follow-ups to Serratus

Viral reactivation (Nature 2023)



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

Obelisks

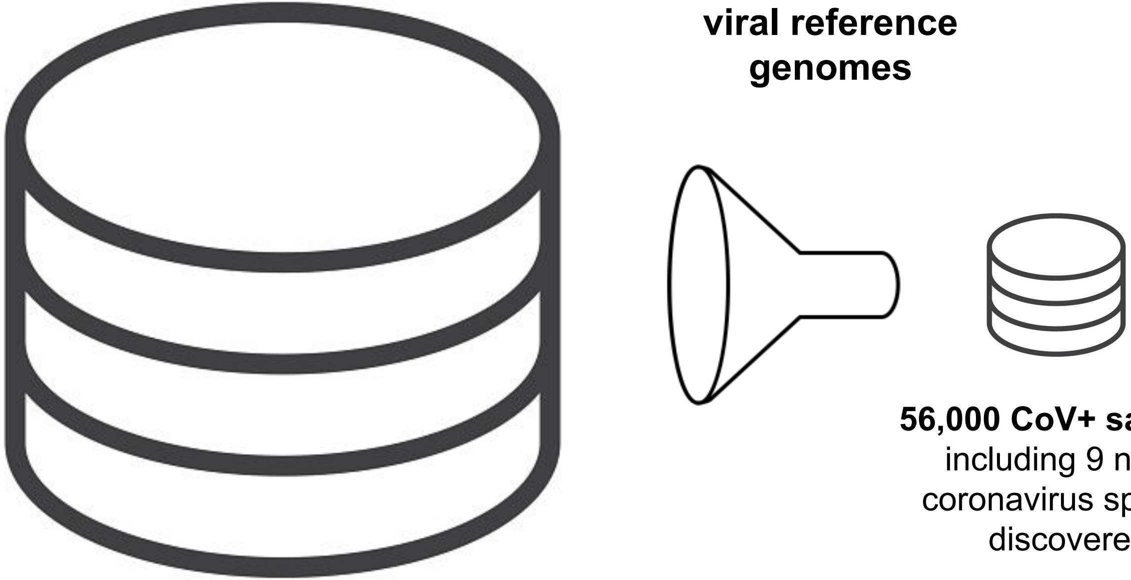


Intriguing find. **Stanford University discovers obelisks hiding in human microbiomes**

Updated - February 06, 2024 at 11:18 AM. | London

This new biological phenomenon, detailed in a recent preprint, challenges the conventional understanding viruses and viroids.

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

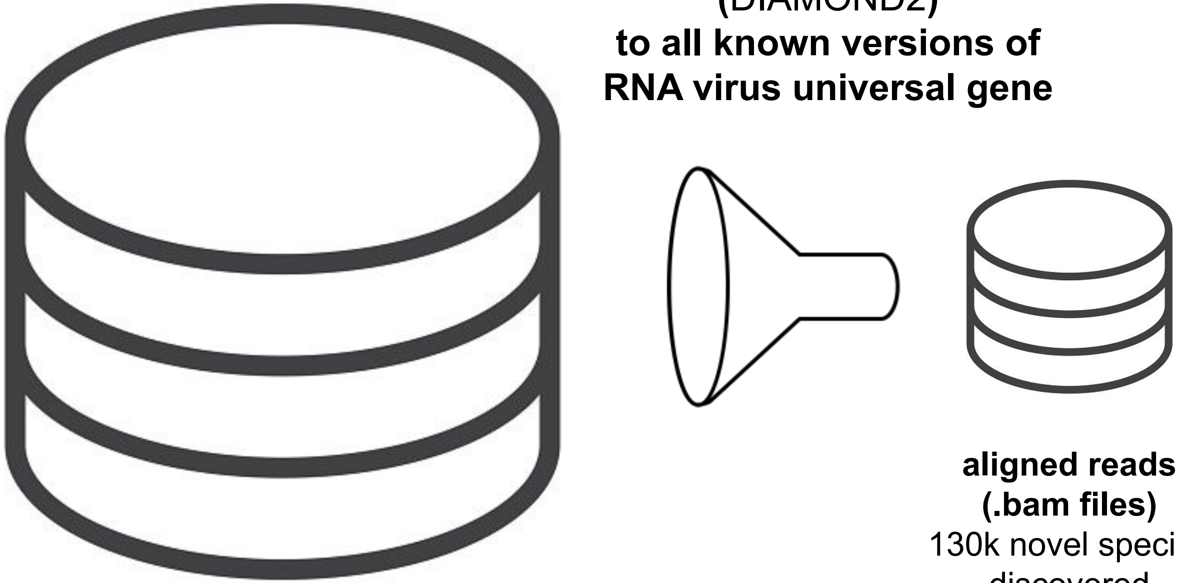


56,000 CoV+ samples including 9 novel coronavirus species discovered

All RNA-seqs pre-2020

(10 petabases)

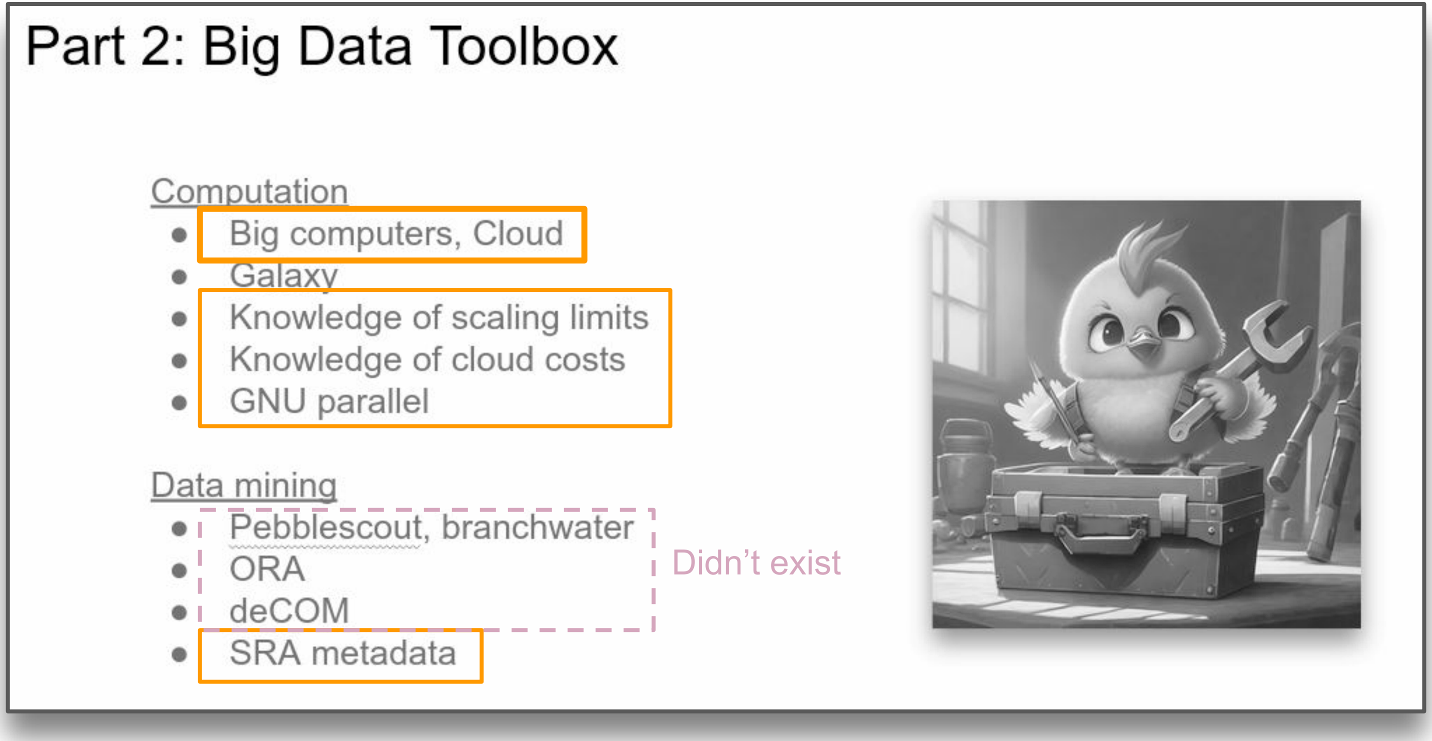
Serratus download &
sensitive align
(DIAMOND2**)**
to all known versions of
RNA virus universal gene



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

All RNA-seqs
pre-2020

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

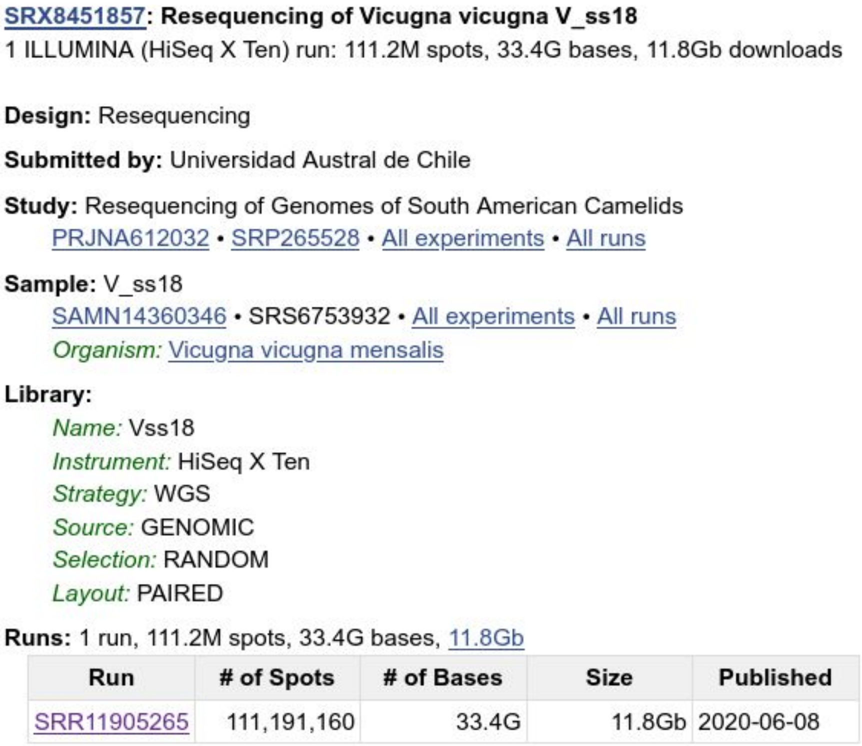
- Pebblescout, branchwater
- ORA
- deCOM
- SRA metadata

Didn’t exist

Diving into SRA’s data

What are SRA metadata?

All of this →



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

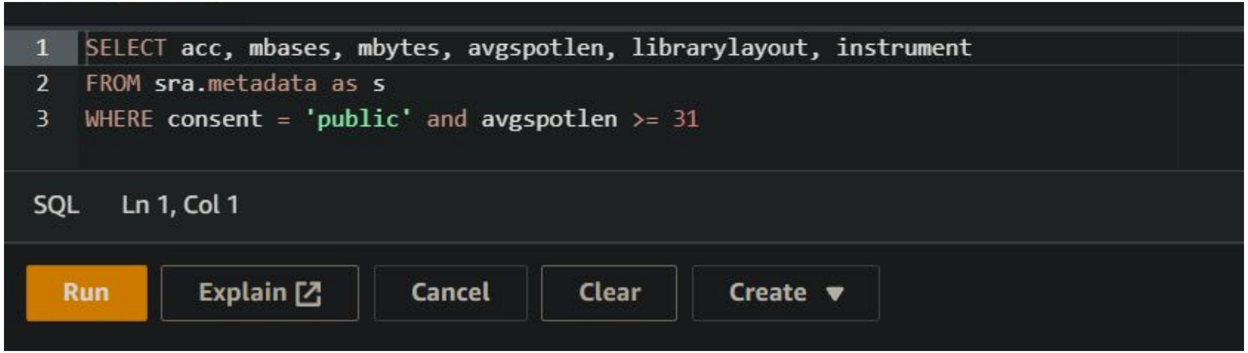
Accessing SRA metadata

~~0. NCBI website~~

1. NCBI FTP metadata

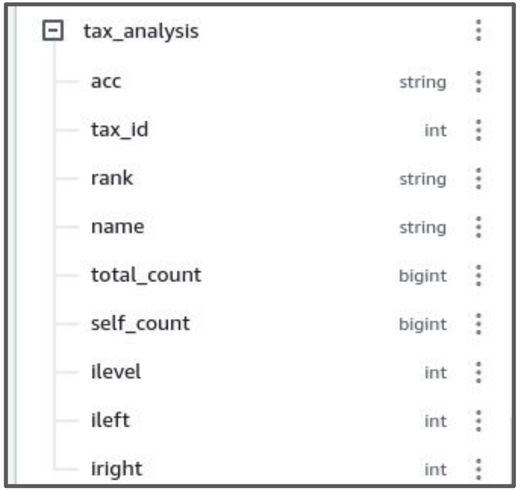
https://trace.ncbi.nlm.nih.gov/Traces/index.html?view=mirroring

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

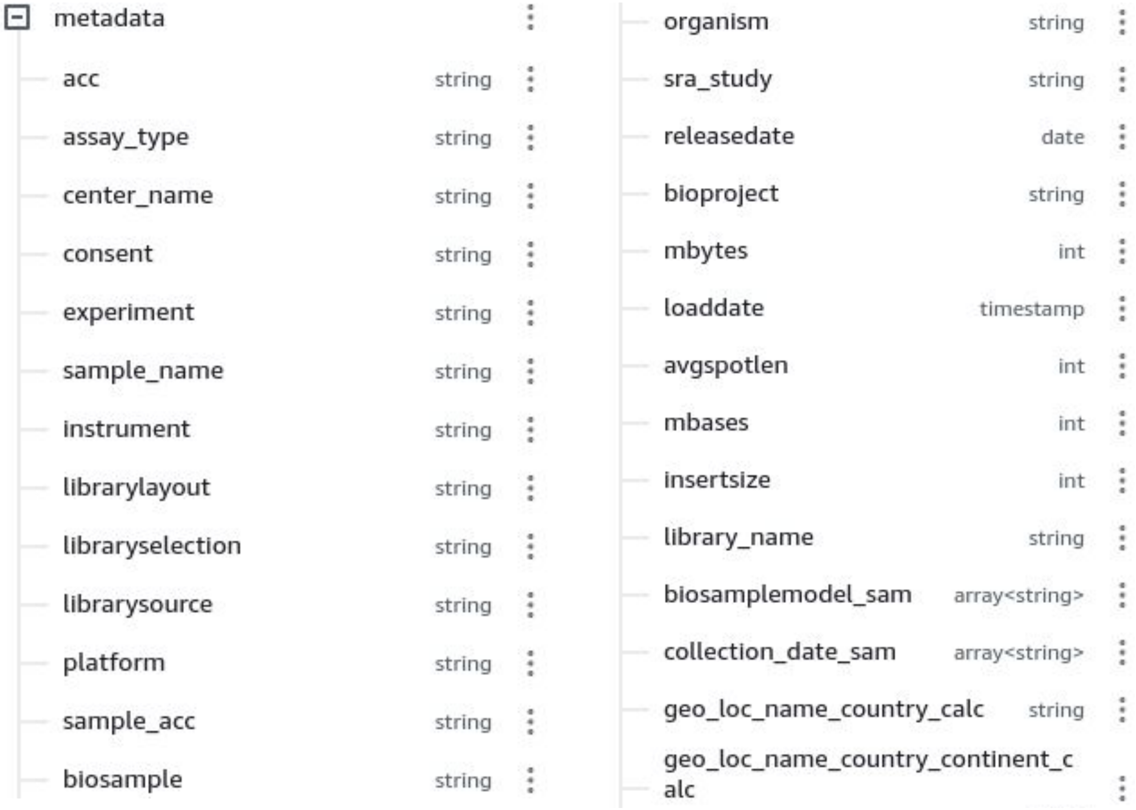


```sql
SELECT acc, mbases, mbytes, avgspotlen, librarylayout, instrument
FROM sra.metadata as s
WHERE consent = 'public' and avgspotlen >= 31
```

SRA metadata

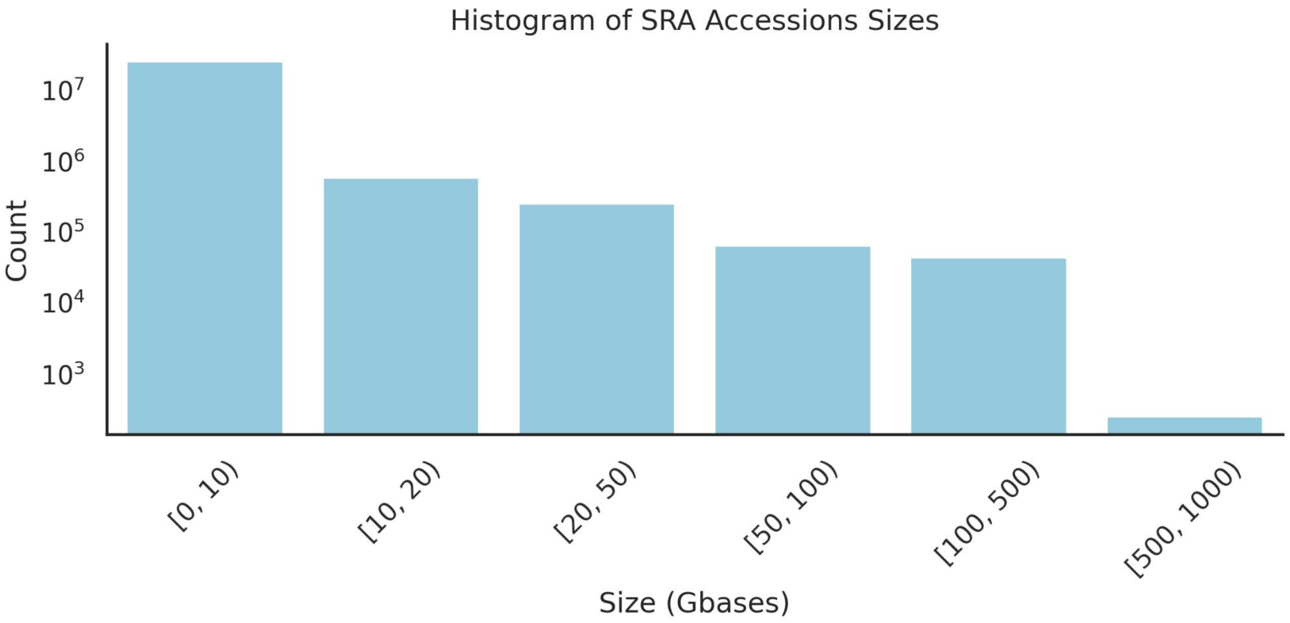


tax_analysis	
acc	string
tax_id	int
rank	string
name	string
total_count	bigint
self_count	bigint
ilevel	int
ileft	int
iright	int

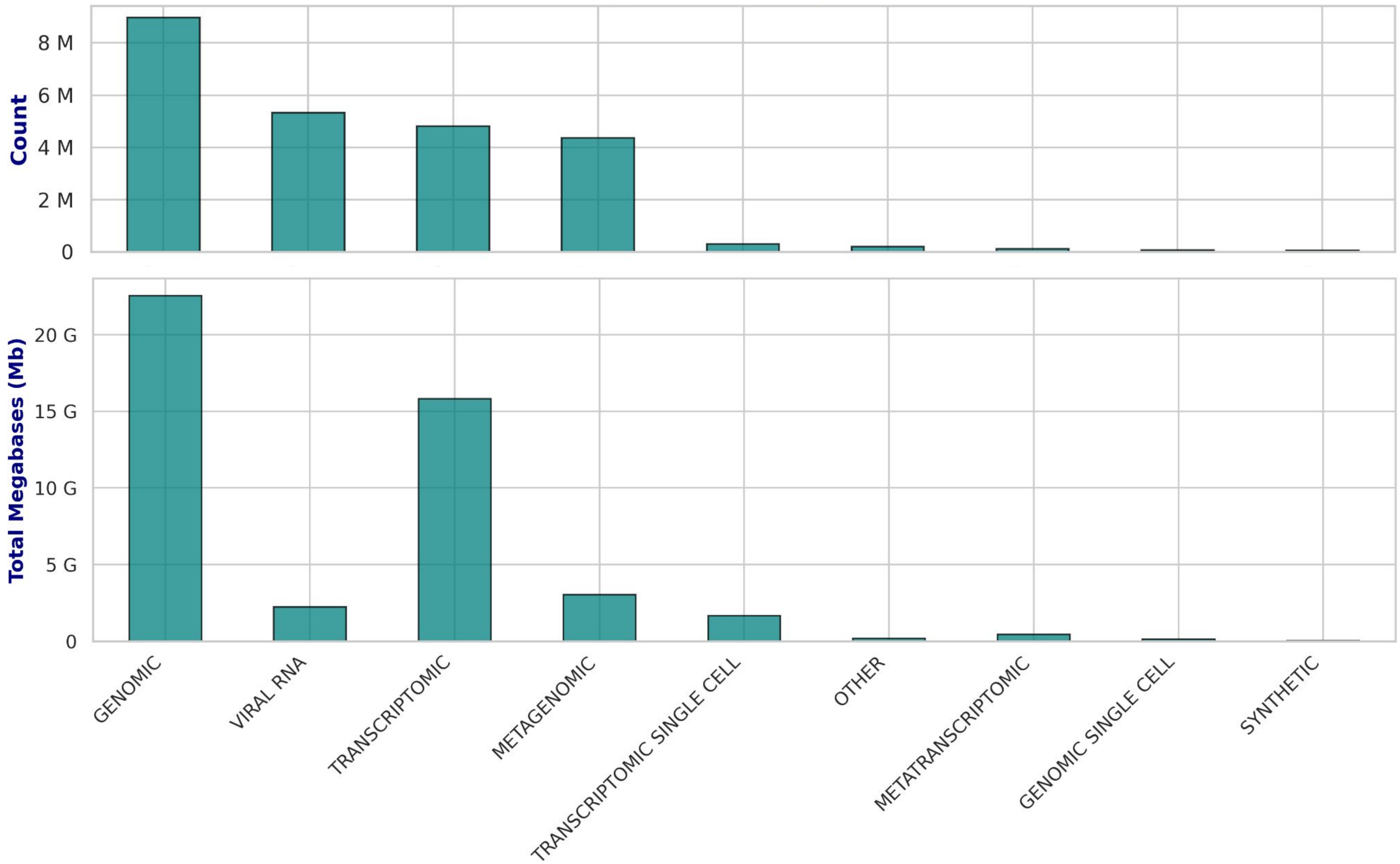


metadata	
acc	string
assay_type	string
center_name	string
consent	string
experiment	string
sample_name	string
instrument	string
librarylayout	string
libraryselection	string
librarysource	string
platform	string
sample_acc	string
biosample	string
organism	string
sra_study	string
releasedate	date
bioproject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

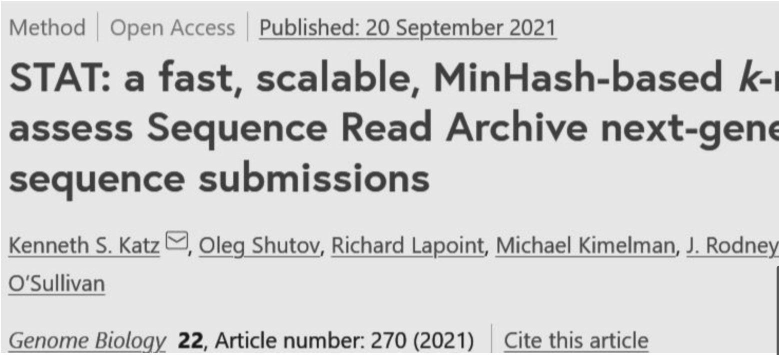
SRA accessions sizes (2023)



SRA accessions types (2023)



SRA taxonomy analysis



Method | Open Access | Published: 20 September 2021

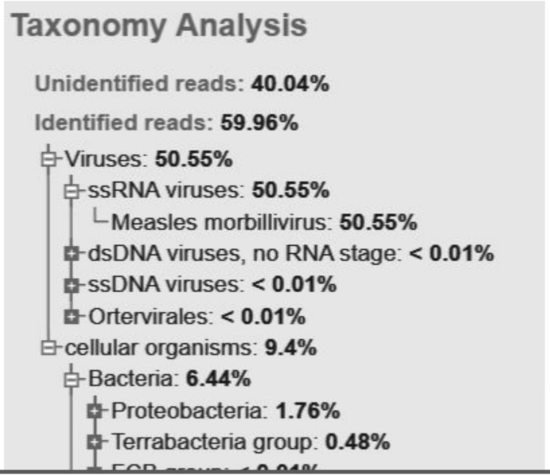
STAT: a fast, scalable, MinHash-based *k*-mer tool to assess Sequence Read Archive next-generation sequence submissions

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:

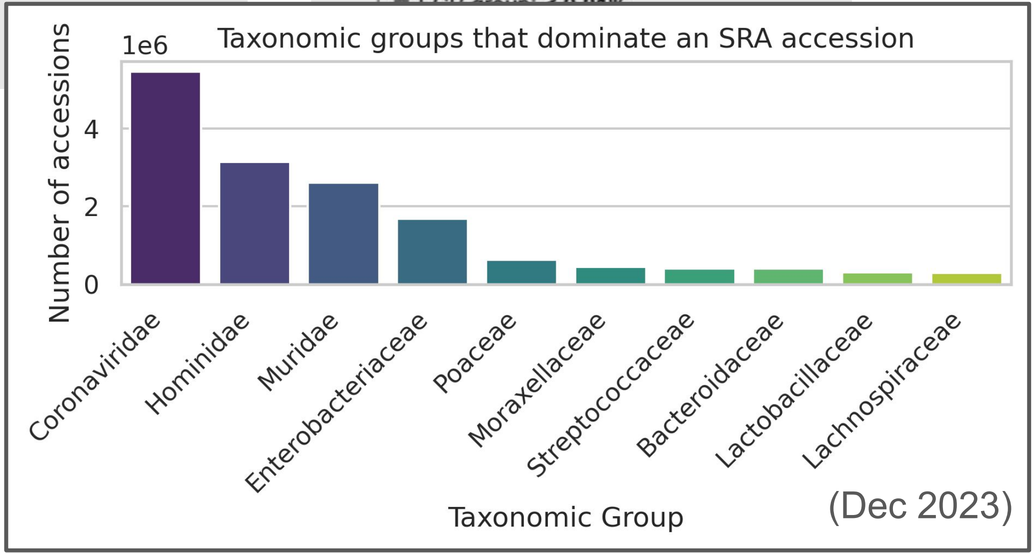


Taxonomy Analysis

Unidentified reads: **40.04%**

Identified reads: **59.96%**

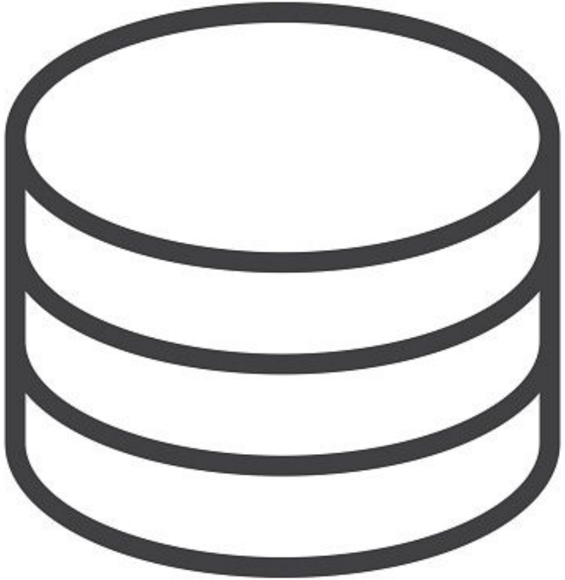
```
⊟Viruses: 50.55%
  ⊟ssRNA viruses: 50.55%
    └Measles morbillivirus: 50.55%
  ⊞dsDNA viruses, no RNA stage: < 0.01%
  ⊞ssDNA viruses: < 0.01%
  ⊞Ortervirales: < 0.01%
⊟cellular organisms: 9.4%
  ⊟Bacteria: 6.44%
    ⊞Proteobacteria: 1.76%
    ⊞Terrabacteria group: 0.48%
    ⊞FCB group: < 0.01%
```



(Dec 2023)

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

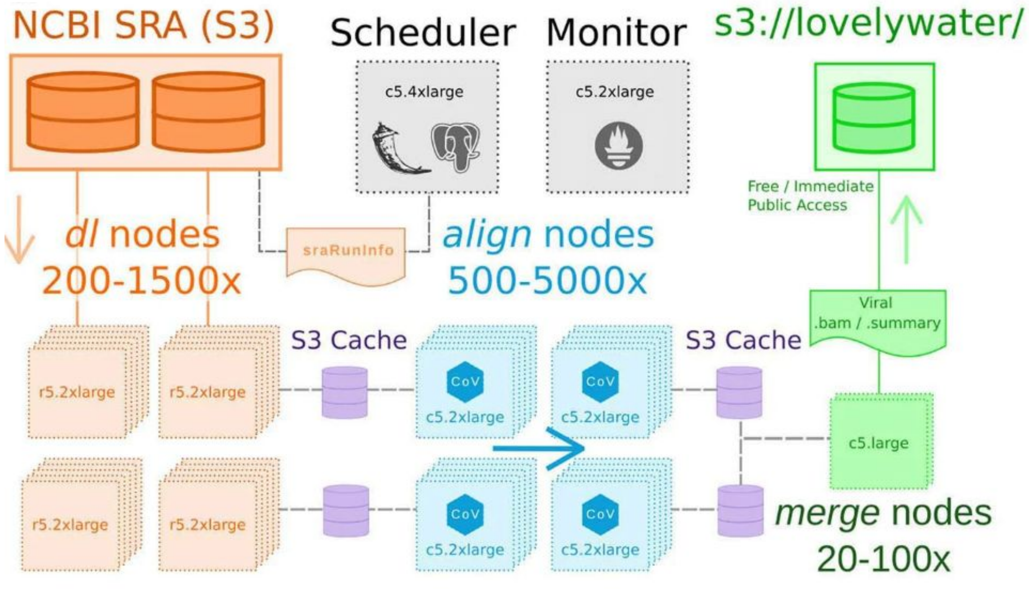


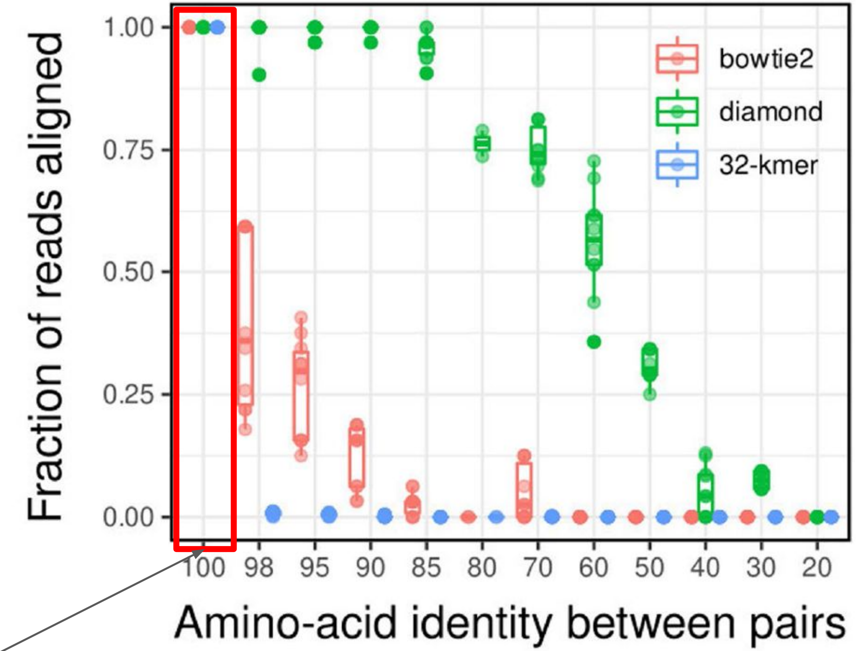
Fig: A. Babaian

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

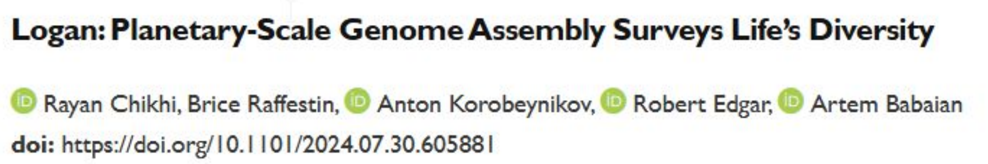
State of the art (ordered by sensitivity/speed):

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

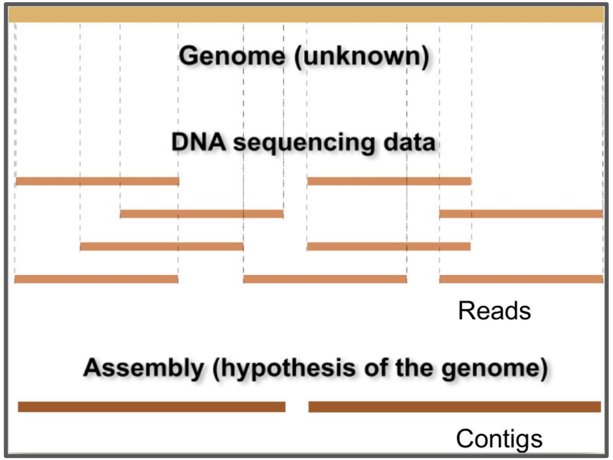


Logan: Planetary-Scale Genome Assembly Surveys Life’s Diversity

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

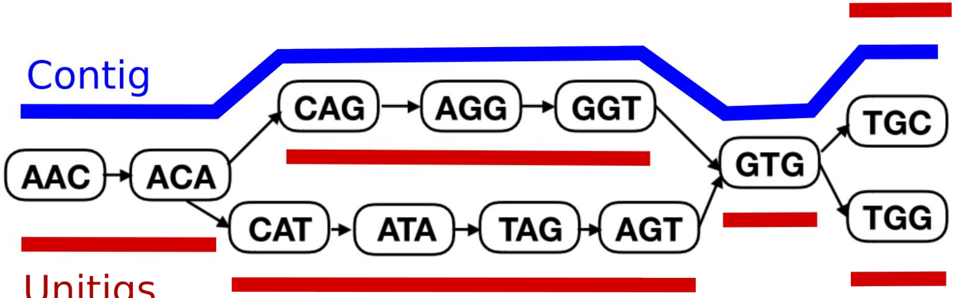
doi: https://doi.org/10.1101/2024.07.30.605881

Unitigs? Contigs?



Contigs: typical output of genome assembly methods

Unitig: simple path in the de Bruijn graph



Graph from Menegaux, Vert

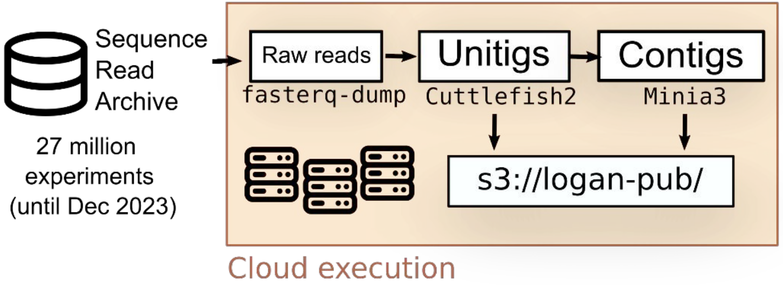
Why unitigs? they keep all variants (SNPs, indels, ..)

Contigs are consensuses

Logan: project steps

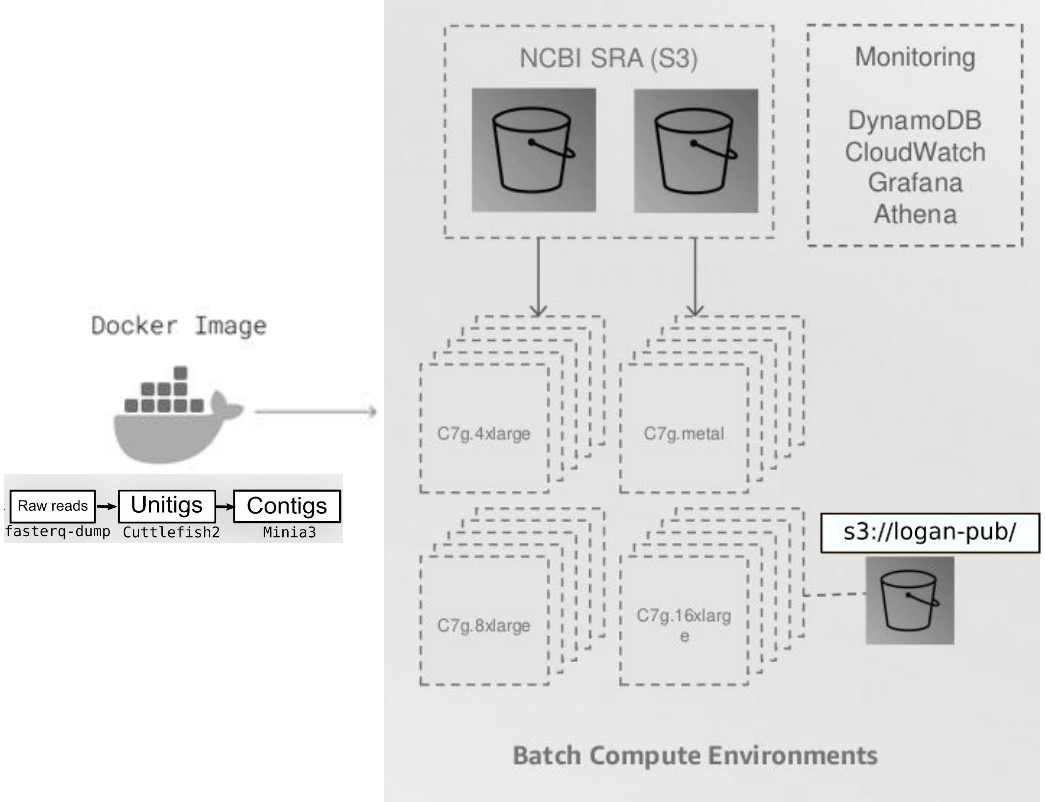
- **Step 1 (2024):** Download all of SRA, assemble each sample, host results publicly [done]

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

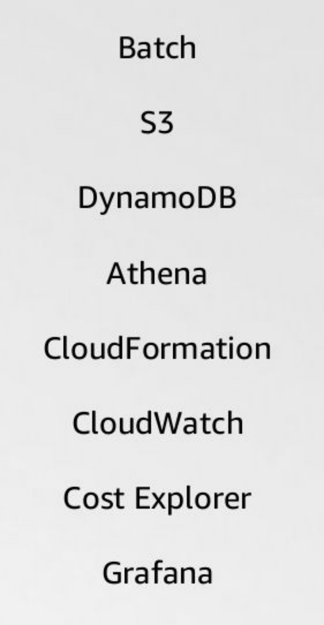


- **Step 2 (2025):** Index assemblies, create a search engine (“searching YouTube”) [done] https://logan-search.org/

Logan: infrastructure



AWS services used:



Batch

S3

DynamoDB

Athena

CloudFormation

CloudWatch

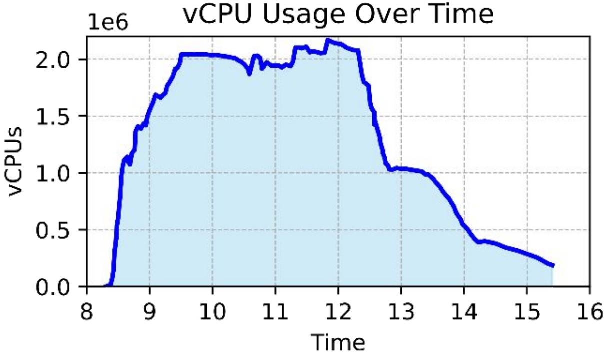
Cost Explorer

Grafana

Logan: computation statistics

Global statistics

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



Run 6 statistics

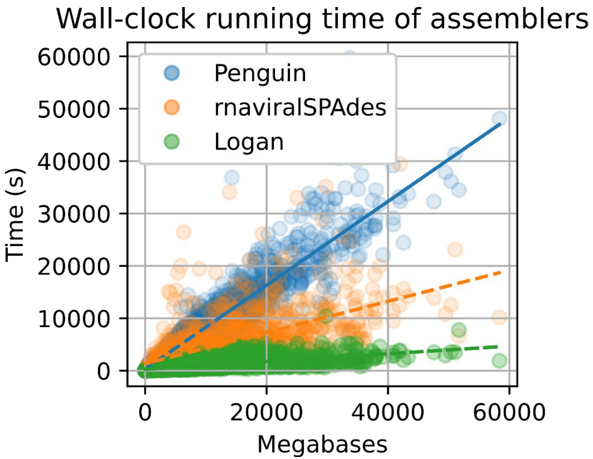
Input data	19.6 petabases
Runtime*	7 hours
Peak Number of Instances	73,100
Peak Number of vCPUs	2.18 million
Peak Total EBS storage	52 petabytes

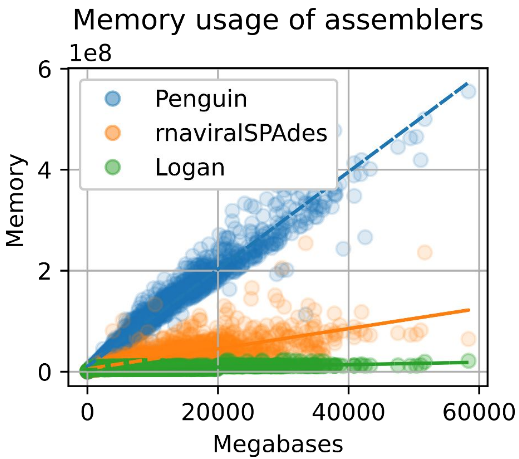
Many failures:

- 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?

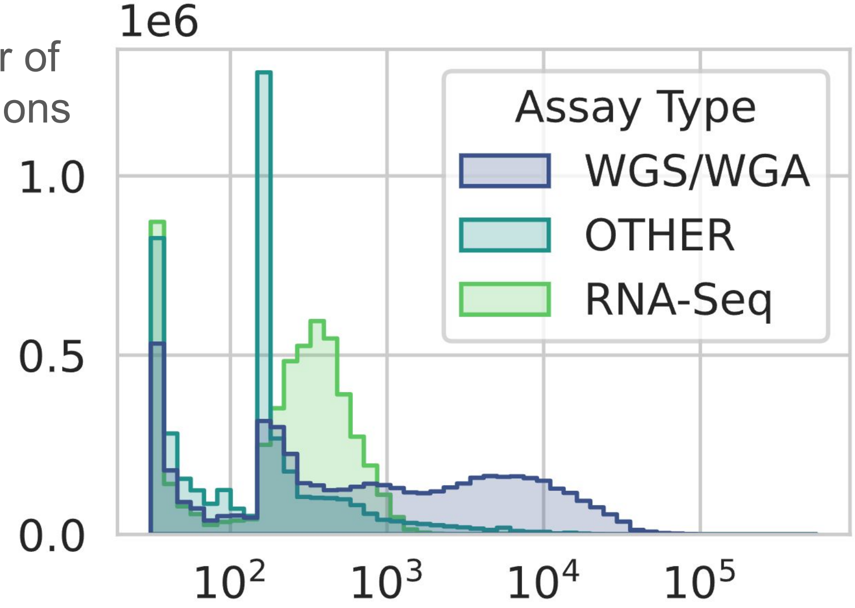
- **Genome assembly is compute- and memory-intensive, usually.**
- We used a simple pipeline of **highly optimized components**:
  - Reads → counted kmers → de Bruijn graph → unitigs
  - Unitigs → simplification of graph → contigs
- Speeding up each step took **decades of bioinformatics research**





Draft-level assembly contiguity

Number of accessions



Contig N50

higher=better

Except for RNAseqs

Algorithmic components used in Logan

- String algorithms (“minimizers” (~=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)

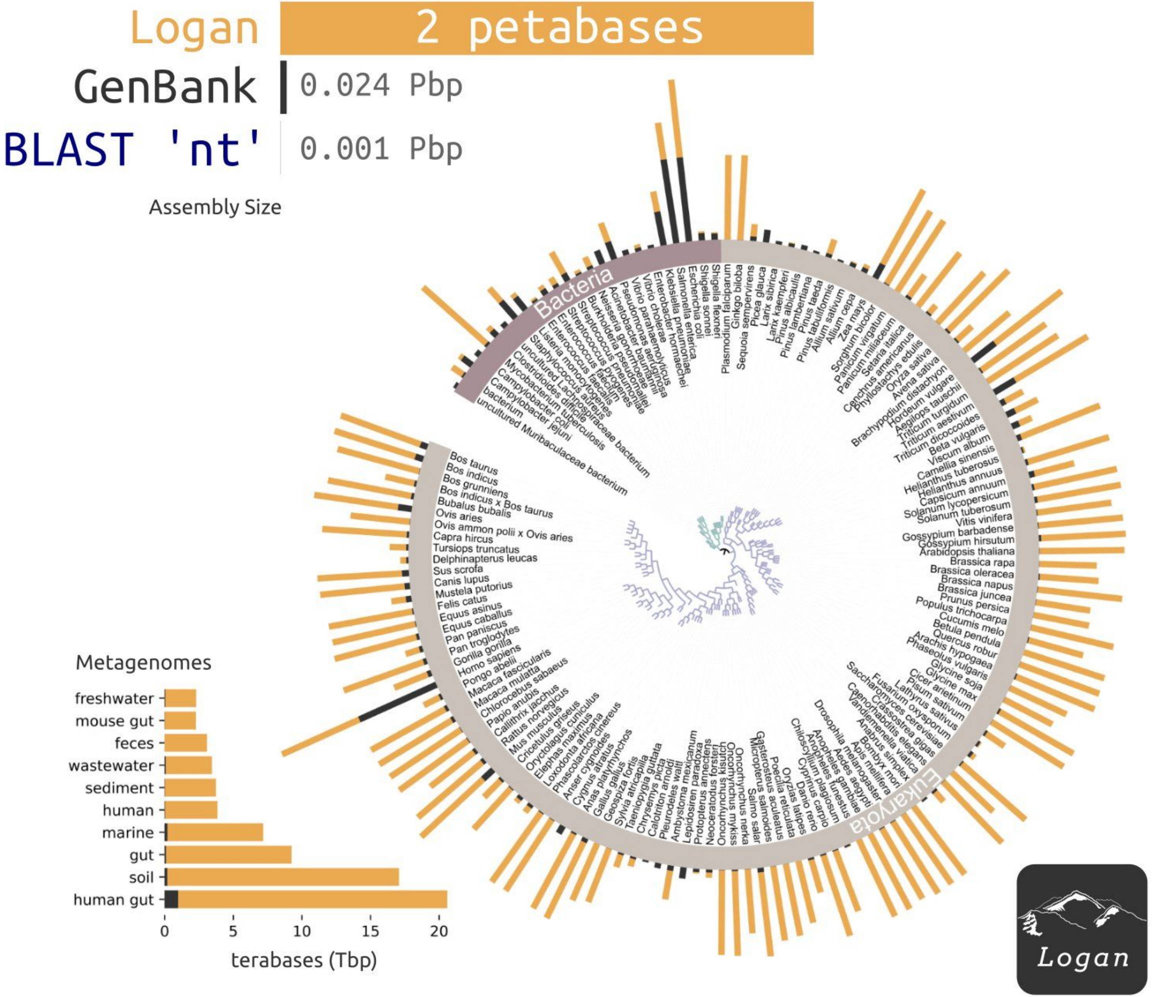
Part of the algorithmic story: R. Chikhi, ***A tale of optimizing the space taken by de Bruijn graphs***, Computability in Europe (2021) [PDF]

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
Bacteria
uncultured Muribaculaceae bacterium
bacterium
Campylobacter jejuni
Campylobacter coli
Mycobacterium tuberculosis
Clostridioides difficile
uncultured Lachnospiraceae bacterium
Staphylococcus aureus
Listeria monocytogenes
Enterococcus faecalis
Enterococcus faecium
Streptococcus pyogenes
Streptococcus pneumoniae
Burkholderia pseudomallei
Neisseria gonorrhoeae
Acinetobacter baumannii
Pseudomonas aeruginosa
Vibrio parahaemolyticus
Vibrio cholerae
Enterobacter hormaechei
Klebsiella pneumoniae
Salmonella enterica
Escherichia coli
Shigella sonnei
Shigella flexneri
Plasmodium falciparum
Ginkgo biloba
Sequoia sempervirens
Picea glauca
Larix sibirica
Larix kaempferi
Pinus albicaulis
Pinus lambertiana
Pinus taeda
Pinus tabuliformis
Allium sativum
Allium cepa
Zea mays
Sorghum bicolor
Panicum virgatum
Panicum miliaceum
Setaria italica
Cenchrus americanus
Phyllostachys edulis
Oryza sativa
Avena sativa
Brachypodium distachyon
Hordeum vulgare
Aegilops tauschii
Triticum turgidum
Triticum aestivum
Triticum dicoccoides
Beta vulgaris
Viscum album
Camellia sinensis
Helianthus tuberosus
Helianthus annuus
Capsicum annuum
Solanum lycopersicum
Solanum tuberosum
Vitis vinifera
Gossypium barbadense
Gossypium hirsutum
Arabidopsis thaliana
Brassica rapa
Brassica oleracea
Brassica napus
Brassica juncea
Prunus persica
Populus trichocarpa
Cucumis melo
Betula pendula
Quercus robur
Arachis hypogaea
Phaseolus vulgaris
Glycine soja
Glycine max
Cicer arietinum
Pisum sativum
Lathyrus sativus
Fusarium oxysporum
Saccharomyces cerevisiae
Crassostrea gigas
Caenorhabditis elegans
Vandiemenella viatica
Anabrus simplex
Bombyx mori
Apis mellifera
Drosophila melanogaster
Aedes aegypti
Anopheles gambiae
Anopheles funestus
Chiloscyllium plagiosum
Cyprinus carpio
Danio rerio
Oryzias latipes
Poecilia reticulata
Gasterosteus aculeatus
Micropterus salmoides
Salmo salar
Oncorhynchus nerka
Oncorhynchus mykiss
Oncorhynchus kisutch
Neoceratodus forsteri
Protopterus annectens
Lepidosiren paradoxa
Ambystoma mexicanum
Pleurodeles waltl
Calotriton arnoldi
Chrysemys picta
Taeniopygia guttata
Sylvia atricapilla
Geospiza fortis
Gallus gallus
Anas platyrhynchos
Cygnus atratus
Anser cygnoides
Phascolarctos cinereus
Loxodonta africana
Elephas maximus
Oryctolagus cuniculus
Cricetulus griseus
Mus musculus
Rattus norvegicus
Callithrix jacchus
Papio anubis
Chlorocebus sabaeus
Macaca mulatta
Macaca fascicularis
Pongo abelii
Homo sapiens
Gorilla gorilla
Pan troglodytes
Pan paniscus
Equus caballus
Equus asinus
Felis catus
Mustela putorius
Canis lupus
Sus scrofa
Delphinapterus leucas
Tursiops truncatus
Capra hircus
Ovis ammon polii x Ovis aries
Ovis aries
Bubalus bubalis
Bos indicus x Bos taurus
Bos grunniens
Bos indicus
Bos taurus
Eukaryota
Metagenomes
freshwater
mouse gut
feces
wastewater
sediment
human
marine
gut
soil
human gut
0
5
10
15
20
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.

Reference genome sequencing (RefGen)

- Only closely related taxa can be used for alignment
- Always done with long reads
- Enable any kind of downstream genetic analysis
- Generating a chromosome-scale reference genome is a life-time investment. But it is costly.
- Sit tight, it is going to be a long one

Logan “fun facts”

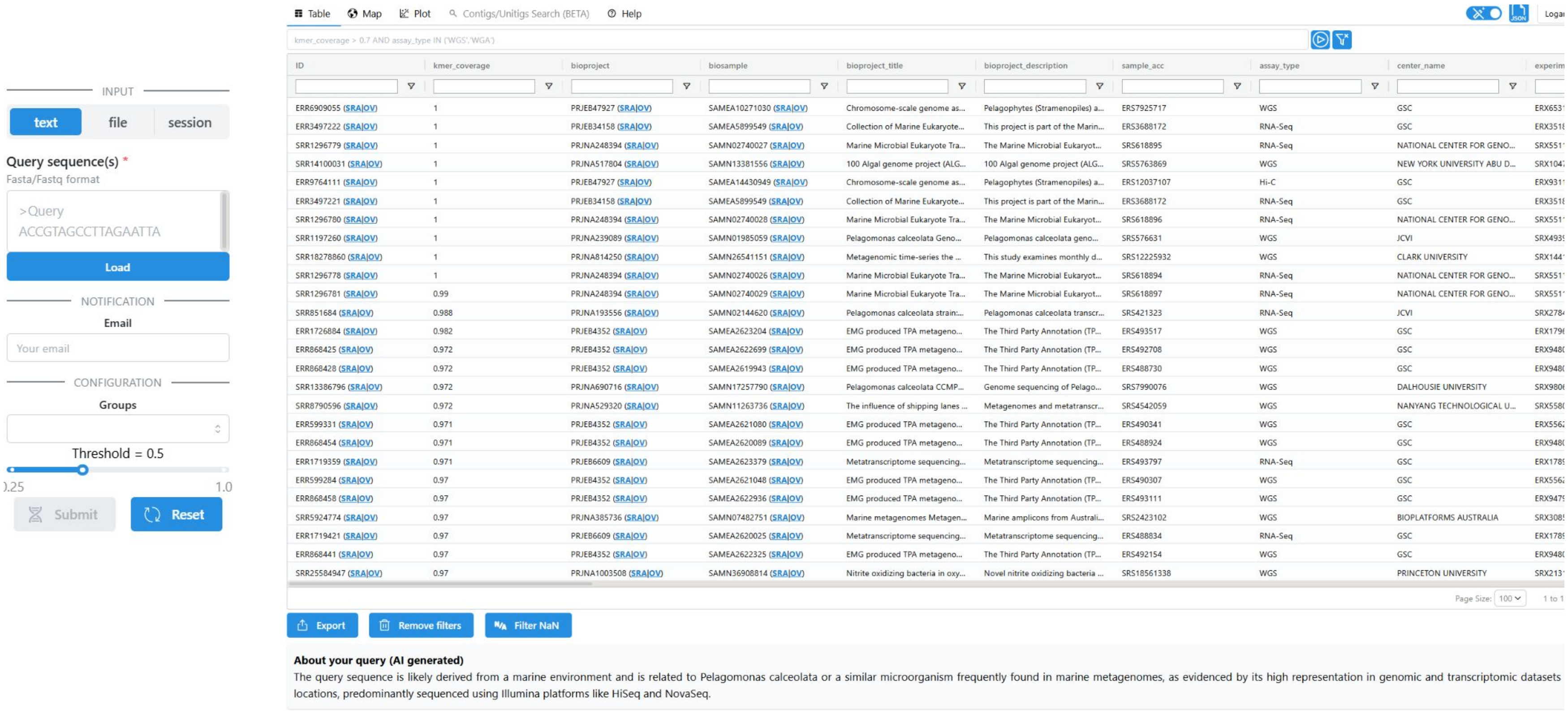
- Logan total computation: **30 hours**. Would have been ~1.5 years on local cluster.

- 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



Query sequence(s) *
Fasta/Fastq format

```
>Query
ACCGTAGCCTTAGAATTA
```

Email

Your email

Groups

Threshold = 0.5

).25 1.0

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

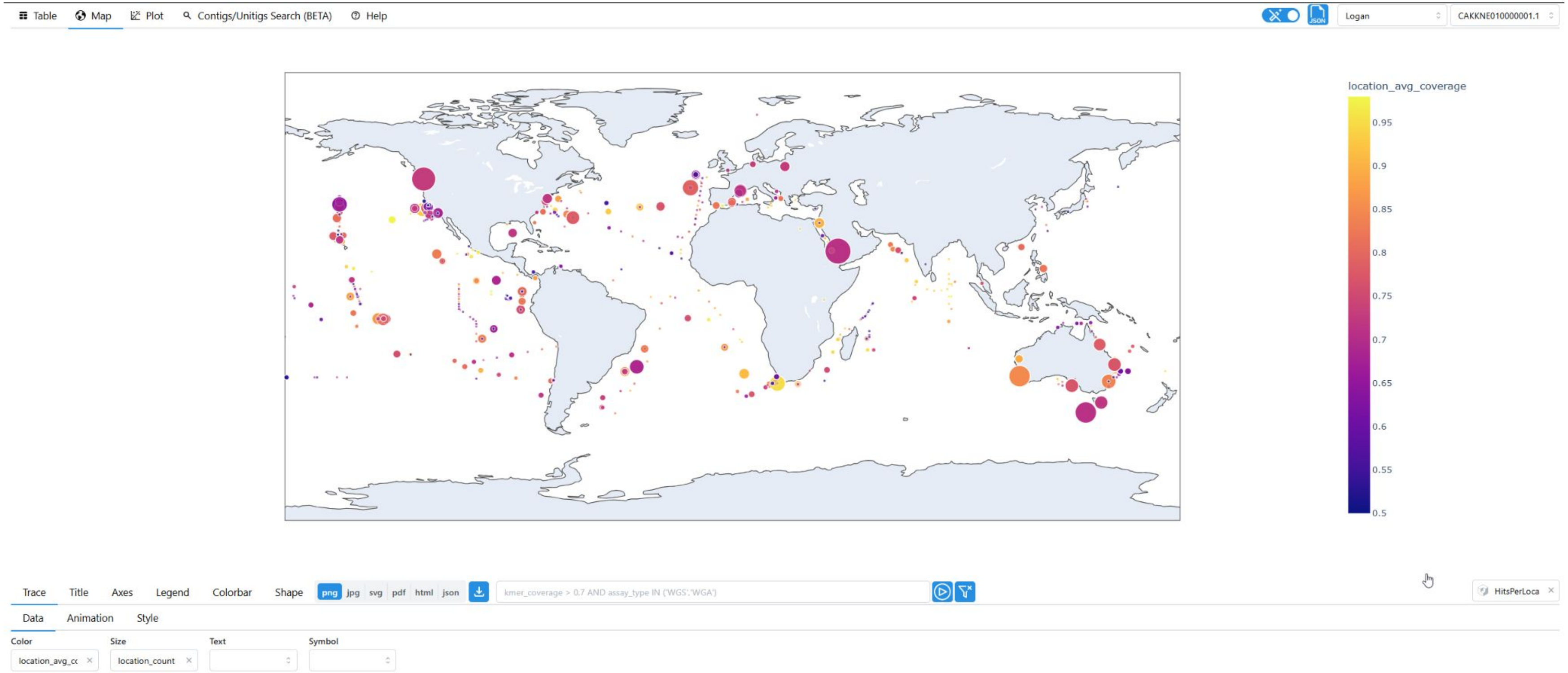
ID	kmer_coverage	bioproject	biosample	bioproject_title	bioproject_description	sample_acc	assay_type	center_name	experim
ERR6909055 (SRA|OV)	1	PRJEB47927 (SRA|OV)	SAMEA10271030 (SRA|OV)	Chromosome-scale genome as...	Pelagophytes (Stramenopiles) a...	ERS7925717	WGS	GSC	ERX6531
ERR3497222 (SRA|OV)	1	PRJEB34158 (SRA|OV)	SAMEA5899549 (SRA|OV)	Collection of Marine Eukaryote...	This project is part of the Marin...	ERS3688172	RNA-Seq	GSC	ERX3518
SRR1296779 (SRA|OV)	1	PRJNA248394 (SRA|OV)	SAMN02740027 (SRA|OV)	Marine Microbial Eukaryote Tra...	The Marine Microbial Eukaryot...	SRS618895	RNA-Seq	NATIONAL CENTER FOR GENO...	SRX5511
SRR14100031 (SRA|OV)	1	PRJNA517804 (SRA|OV)	SAMN13381556 (SRA|OV)	100 Algal genome project (ALG...	100 Algal genome project (ALG...	SRS5763869	WGS	NEW YORK UNIVERSITY ABU D...	SRX1047
ERR9764111 (SRA|OV)	1	PRJEB47927 (SRA|OV)	SAMEA14430949 (SRA|OV)	Chromosome-scale genome as...	Pelagophytes (Stramenopiles) a...	ERS12037107	Hi-C	GSC	ERX9311
ERR3497221 (SRA|OV)	1	PRJEB34158 (SRA|OV)	SAMEA5899549 (SRA|OV)	Collection of Marine Eukaryote...	This project is part of the Marin...	ERS3688172	RNA-Seq	GSC	ERX3518
SRR1296780 (SRA|OV)	1	PRJNA248394 (SRA|OV)	SAMN02740028 (SRA|OV)	Marine Microbial Eukaryote Tra...	The Marine Microbial Eukaryot...	SRS618896	RNA-Seq	NATIONAL CENTER FOR GENO...	SRX5511
SRR1197260 (SRA|OV)	1	PRJNA239089 (SRA|OV)	SAMN01985059 (SRA|OV)	Pelagomonas calceolata Geno...	Pelagomonas calceolata geno...	SRS576631	WGS	JCVI	SRX4939
SRR18278860 (SRA|OV)	1	PRJNA814250 (SRA|OV)	SAMN26541151 (SRA|OV)	Metagenomic time-series the ...	This study examines monthly d...	SRS12225932	WGS	CLARK UNIVERSITY	SRX1441
SRR1296778 (SRA|OV)	1	PRJNA248394 (SRA|OV)	SAMN02740026 (SRA|OV)	Marine Microbial Eukaryote Tra...	The Marine Microbial Eukaryot...	SRS618894	RNA-Seq	NATIONAL CENTER FOR GENO...	SRX5511
SRR1296781 (SRA|OV)	0.99	PRJNA248394 (SRA|OV)	SAMN02740029 (SRA|OV)	Marine Microbial Eukaryote Tra...	The Marine Microbial Eukaryot...	SRS618897	RNA-Seq	NATIONAL CENTER FOR GENO...	SRX5511
SRR851684 (SRA|OV)	0.988	PRJNA193556 (SRA|OV)	SAMN02144620 (SRA|OV)	Pelagomonas calceolata strain:...	Pelagomonas calceolata transcr...	SRS421323	RNA-Seq	JCVI	SRX2784
ERR1726884 (SRA|OV)	0.982	PRJEB4352 (SRA|OV)	SAMEA2623204 (SRA|OV)	EMG produced TPA metageno...	The Third Party Annotation (TP...	ERS493517	WGS	GSC	ERX1796
ERR868425 (SRA|OV)	0.972	PRJEB4352 (SRA|OV)	SAMEA2622699 (SRA|OV)	EMG produced TPA metageno...	The Third Party Annotation (TP...	ERS492708	WGS	GSC	ERX9480
ERR868428 (SRA|OV)	0.972	PRJEB4352 (SRA|OV)	SAMEA2619943 (SRA|OV)	EMG produced TPA metageno...	The Third Party Annotation (TP...	ERS488730	WGS	GSC	ERX9480
SRR13386796 (SRA|OV)	0.972	PRJNA690716 (SRA|OV)	SAMN17257790 (SRA|OV)	Pelagomonas calceolata CCMP...	Genome sequencing of Pelago...	SRS7990076	WGS	DALHOUSIE UNIVERSITY	SRX9806
SRR8790596 (SRA|OV)	0.972	PRJNA529320 (SRA|OV)	SAMN11263736 (SRA|OV)	The influence of shipping lanes ...	Metagenomes and metatranscr...	SRS4542059	WGS	NANYANG TECHNOLOGICAL U...	SRX5580
ERR599331 (SRA|OV)	0.971	PRJEB4352 (SRA|OV)	SAMEA2621080 (SRA|OV)	EMG produced TPA metageno...	The Third Party Annotation (TP...	ERS490341	WGS	GSC	ERX5562
ERR868454 (SRA|OV)	0.971	PRJEB4352 (SRA|OV)	SAMEA2620089 (SRA|OV)	EMG produced TPA metageno...	The Third Party Annotation (TP...	ERS488924	WGS	GSC	ERX9480
ERR1719359 (SRA|OV)	0.971	PRJEB6609 (SRA|OV)	SAMEA2623379 (SRA|OV)	Metatranscriptome sequencing...	Metatranscriptome sequencing...	ERS493797	RNA-Seq	GSC	ERX1789
ERR599284 (SRA|OV)	0.97	PRJEB4352 (SRA|OV)	SAMEA2621048 (SRA|OV)	EMG produced TPA metageno...	The Third Party Annotation (TP...	ERS490307	WGS	GSC	ERX5562
ERR868458 (SRA|OV)	0.97	PRJEB4352 (SRA|OV)	SAMEA2622936 (SRA|OV)	EMG produced TPA metageno...	The Third Party Annotation (TP...	ERS493111	WGS	GSC	ERX9479
SRR5924774 (SRA|OV)	0.97	PRJNA385736 (SRA|OV)	SAMN07482751 (SRA|OV)	Marine metagenomes Metagen...	Marine amplicons from Australi...	SRS2423102	WGS	BIOPLATFORMS AUSTRALIA	SRX3085
ERR1719421 (SRA|OV)	0.97	PRJEB6609 (SRA|OV)	SAMEA2620025 (SRA|OV)	Metatranscriptome sequencing...	Metatranscriptome sequencing...	ERS488834	RNA-Seq	GSC	ERX1789
ERR868441 (SRA|OV)	0.97	PRJEB4352 (SRA|OV)	SAMEA2622325 (SRA|OV)	EMG produced TPA metageno...	The Third Party Annotation (TP...	ERS492154	WGS	GSC	ERX9480
SRR25584947 (SRA|OV)	0.97	PRJNA1003508 (SRA|OV)	SAMN36908814 (SRA|OV)	Nitrite oxidizing bacteria in oxy...	Novel nitrite oxidizing bacteria ...	SRS18561338	WGS	PRINCETON UNIVERSITY	SRX2131

Page Size: 100 1 to 1

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



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

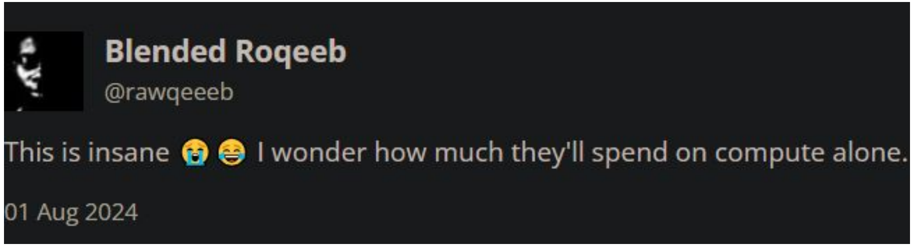
Color: location_avg_cc

Size: location_count

Text:

Symbol:

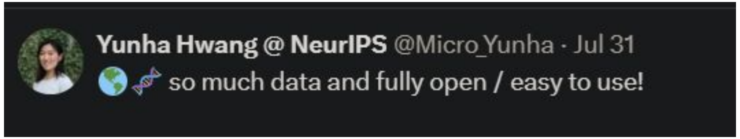
Logan reactions



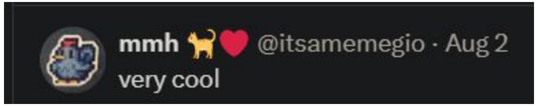
Blended Roqeeb
@rawqeeeb

This is insane 😭 😂 I wonder how much they'll spend on compute alone.

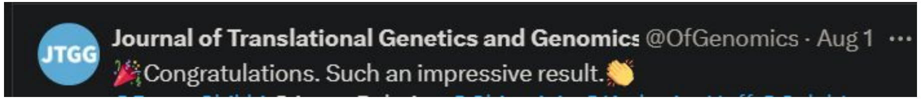
01 Aug 2024



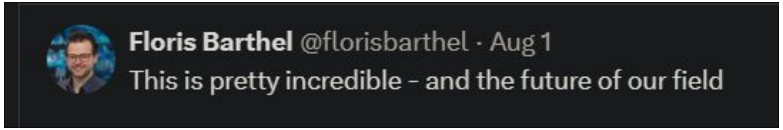
Yunha Hwang @ NeurIPS @Micro_Yunha · Jul 31
🌎🧬 so much data and fully open / easy to use!



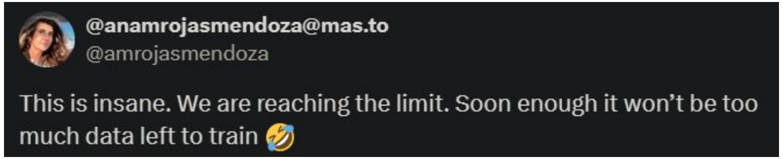
mmh 🐈❤️ @itsamemegio · Aug 2
very cool



Journal of Translational Genetics and Genomics @OfGenomics · Aug 1 ···
🎉Congratulations. Such an impressive result.👏

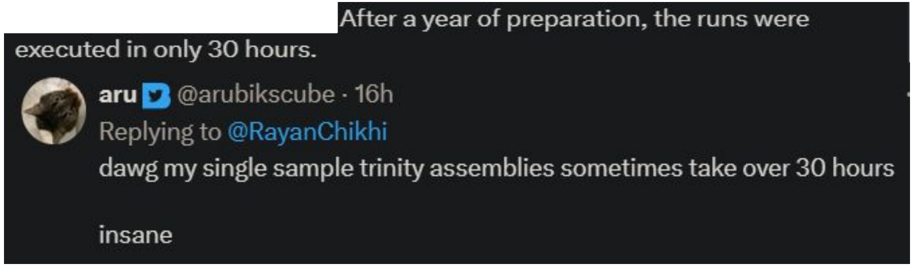


Floris Barthel @florisbarthel · Aug 1
This is pretty incredible – and the future of our field



@anamrojasmendoza@mas.to
@amrojasmendoza

This is insane. We are reaching the limit. Soon enough it won't be too much data left to train 🤣

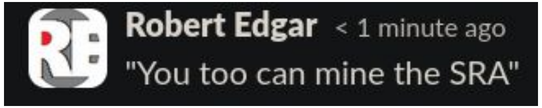


After a year of preparation, the runs were executed in only 30 hours.

aru @arubikscube · 16h
Replying to @RayanChikhi
dawg my single sample trinity assemblies sometimes take over 30 hours

insane

Want to dive in Logan data ?



Robert Edgar < 1 minute ago

"You too can mine the SRA"

- 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.

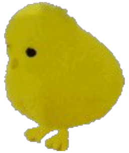
Can you do hands-on bioinformatics?
Contact rayan.chikhi@pasteur.fr and we'll add you to Logan/Serratus Slack

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?

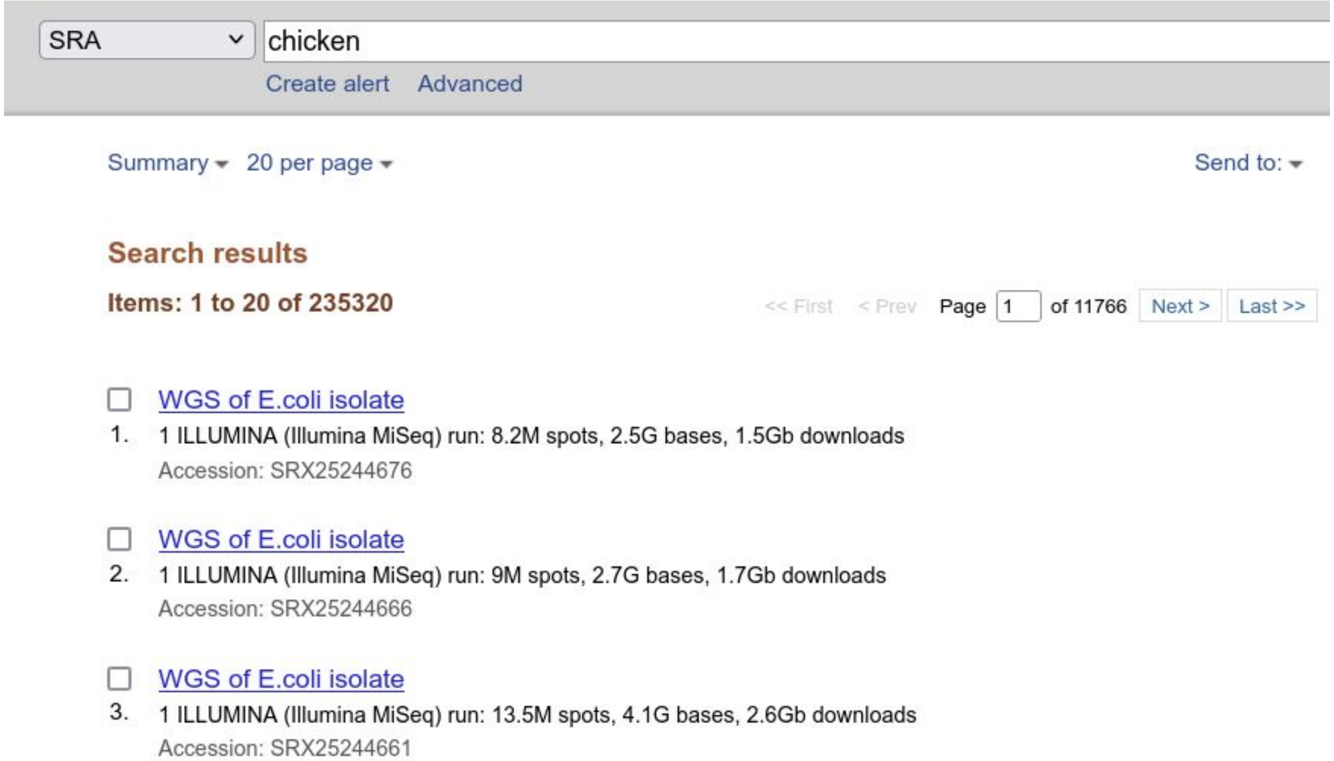
~~0) BLAST~~ Not enough individuals in nt

~~1) NCBI Pebblescout~~ Only has metagenomes

2) SRA metadata query

3) SRA taxonomy query

SRA metadata query 1: fail



chicken

Search results

Items: 1 to 20 of 235320

Page 1 of 11766

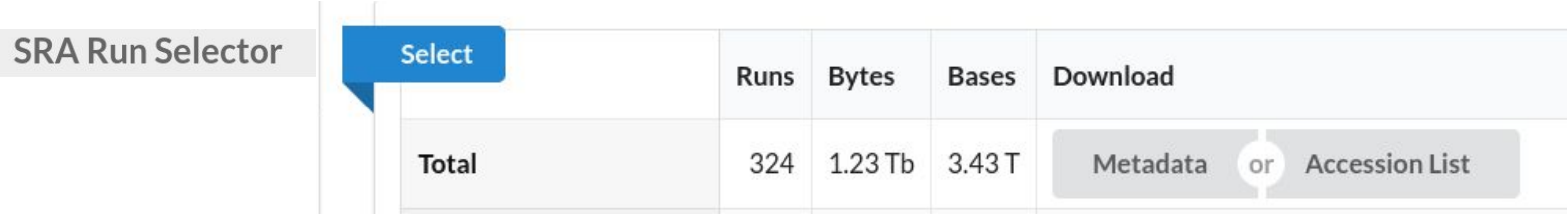
□ WGS of E.coli isolate
1. 1 ILLUMINA (Illumina MiSeq) run: 8.2M spots, 2.5G bases, 1.5Gb downloads
Accession: SRX25244676

□ 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

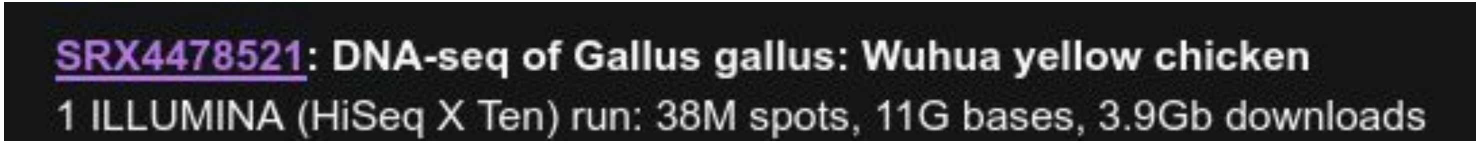
https://www.ncbi.nlm.nih.gov/sra/?term=”yellow+chicken”



SRA Run Selector

	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]



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:)

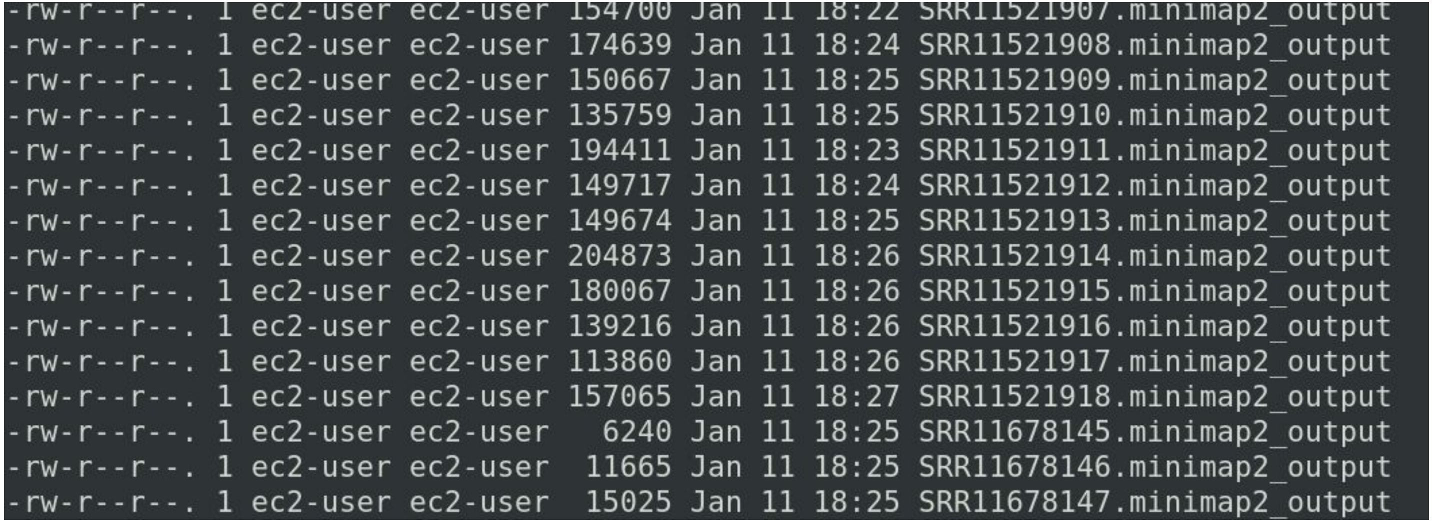
```
$ cat download_and_map_accession.sh

set -e
accession=$1

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

minimap2 -t20 -x sr mc1r.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



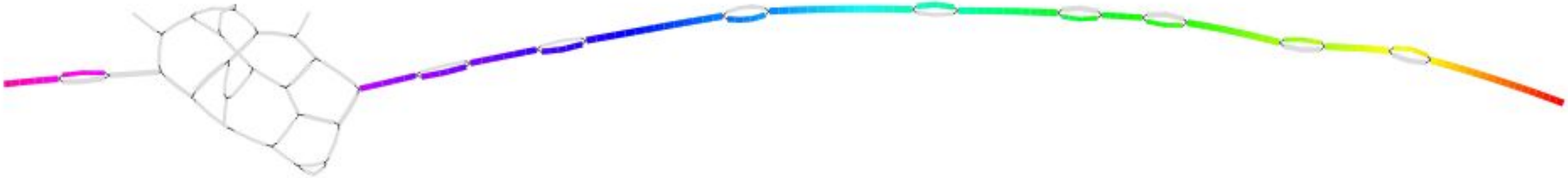
```
-rw-r--r--. 1 ec2-user ec2-user 154700 Jan 11 18:22 SRR11521907.minimap2_output
-rw-r--r--. 1 ec2-user ec2-user 174639 Jan 11 18:24 SRR11521908.minimap2_output
-rw-r--r--. 1 ec2-user ec2-user 150667 Jan 11 18:25 SRR11521909.minimap2_output
-rw-r--r--. 1 ec2-user ec2-user 135759 Jan 11 18:25 SRR11521910.minimap2_output
-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

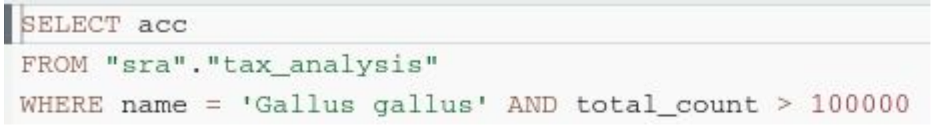
- Constructed pangenome (de Bruijn) graph of MC1R from the “yellow chicken” accessions
- BLASTed a consensus gene to the graph



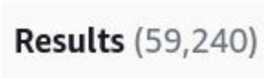
.. good, but this is only for one breed.

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)

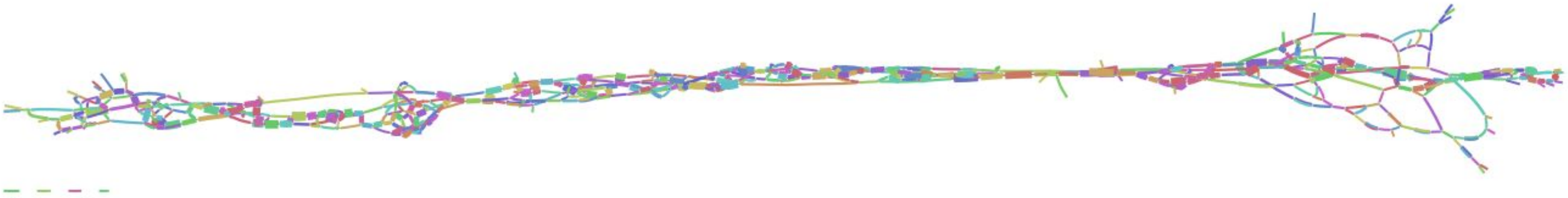
Logan analysis

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

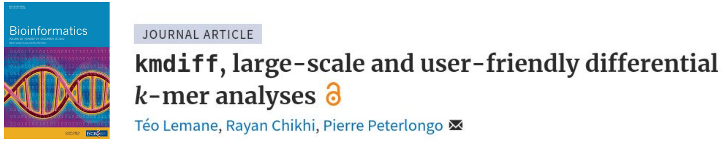
```
minimap2 -x asm20 -t 8 -a mc1r.fa                                  \

<(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 Lemane, Rayan Chikhi, Pierre Peterlongo

TGGGGGTCATCGCCGTGGACCGCTACATCG..

TGGGGGTCATCGCCGTGGACCGCTACATC**A**..

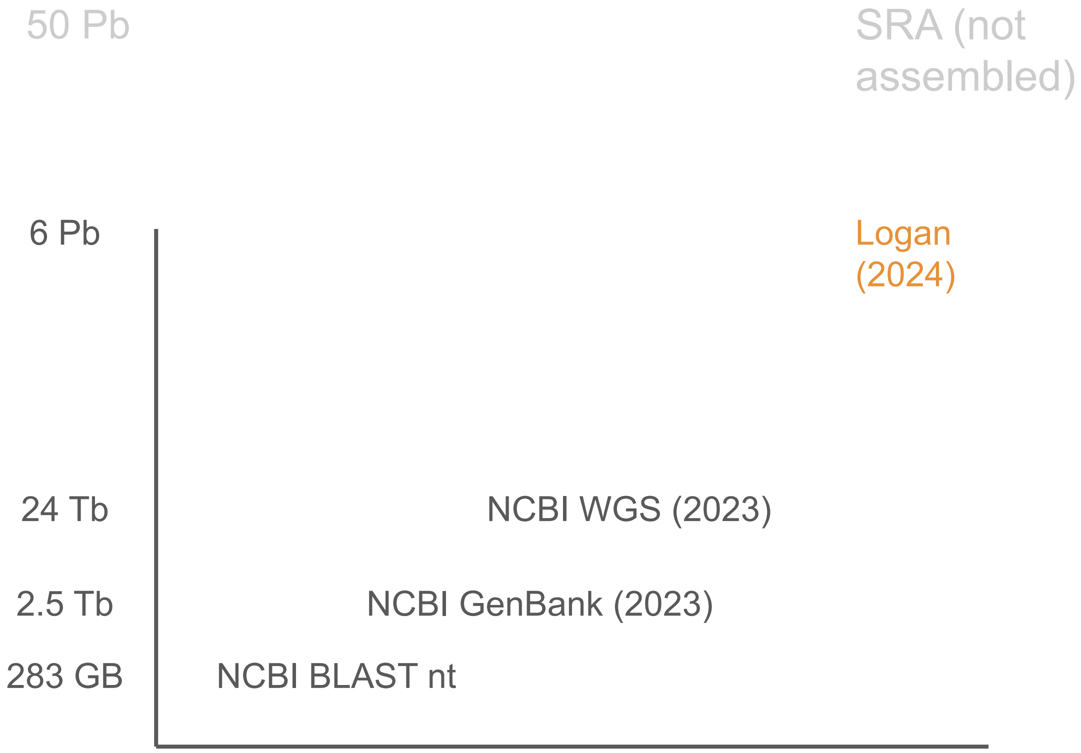
$p<10^{-7}$

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



50 Pb

SRA (not assembled)

6 Pb

Logan (2024)

24 Tb

NCBI WGS (2023)

2.5 Tb

NCBI GenBank (2023)

283 GB

NCBI BLAST nt

Meme credit: A. Babaian

Martin Steinegger

:D we are too haha

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

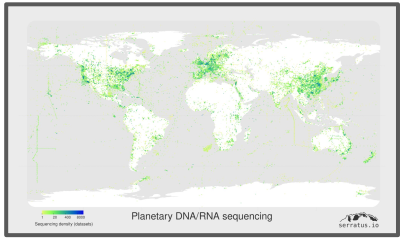
- **SRA-scale analyses now 100x more tractable**
- **Logan: all of Life’s genomic data finally accessible**
- **Many biological discoveries to be made**
- **Better foundation models**

What Logan doesn’t replace

Generation of new samples

High-quality curated genomes

Earth's heart of iron begins to yield its secrets p. 18
Microglia in chronic pain recovery and relapse pp. 33 & 86
Particle acceleration in a nova explosion p. 77
Science
$15
1 APRIL 2022
SPECIAL ISSUE
science.org
AAAS
FILLING THE GAPS
Closing in on a complete human genome p. 42



Outro

What we’ve seen today

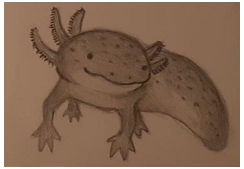
- Some elements of big data bioinformatics
- Toolbox for Big Data
  - Cloud, parallelism, storage handling, knowledge of limitations, AI
- SRA primer
  - Mining metadata
  - Mining sequences
  - Aligning at scale
  - Serratus
- Logan
  - All of Life’s genomic data, available

bigger data
NCC-1701-C
big data

WE'RE GONNA NEED
A BIGGER INSTANCE TYPE

Sequence Bioinformatics

INSTITUT PASTEUR



Lab members:

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

Alumni:

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

Logan co-creators:

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

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

Support for ERC
+ Prairie + Pasteur:
Olivier Gascuel

AWS support: Dorian Schaal,
Adrien Lainé

Dorian Schaal
Sales Representative, AWS

Adrien Lainé
Account Manager, AWS

Greg Autric
Solution Architect, AWS

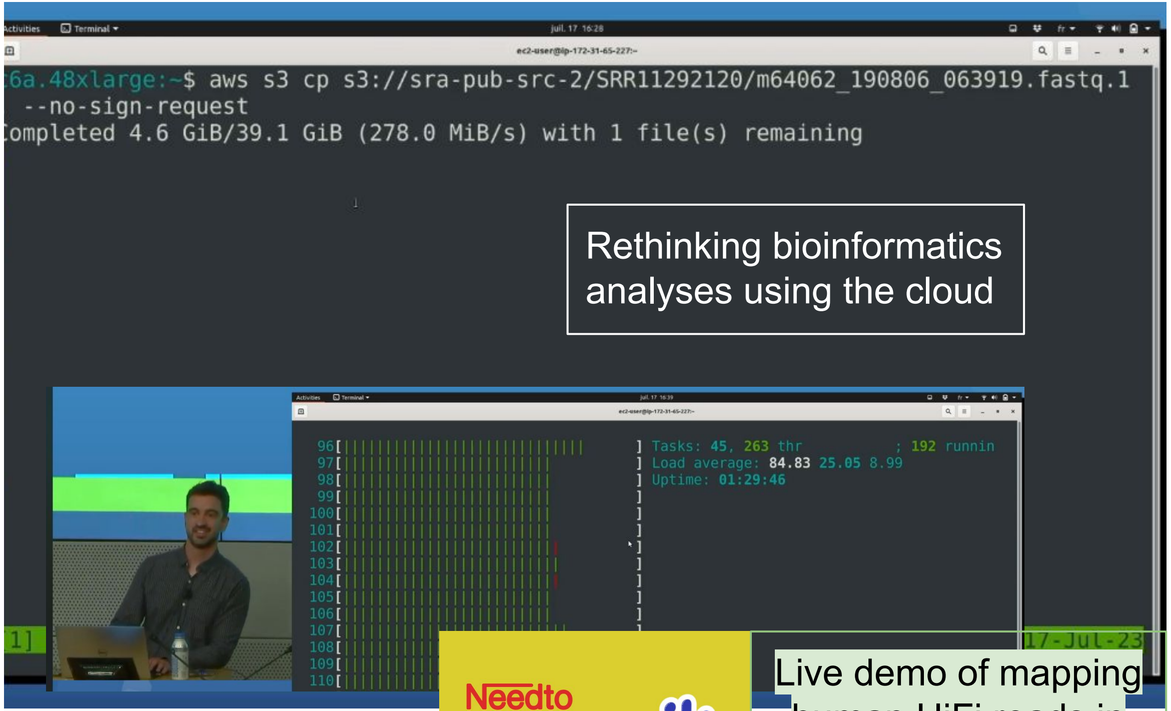
Brice Raffestin
DevOps, Institut Pasteur

Dr. Maxime Hugues
HPC Solution Architect, AWS

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

```
@SRR11606871.1 1
length=4250
CCGGGATGTGCTTGCC
TTTCGGCACCATGTAC
CTGGATGCCAAAGAAC
ACGGTGCCGTTATCCT
TACCGCTCATGAAGTC
ACGGGGCTGA
```

Rayan Chikhi | Living in the Future of Genomics | CGSI 2023



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

```
 96[||||||||||||||||||||||||||||||     ] Tasks: 45, 263 thr      ; 192 runnin
 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[|||||||||
```

Rethinking bioinformatics analyses using the cloud

Needto Mapquik

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

Thank you for your attention!