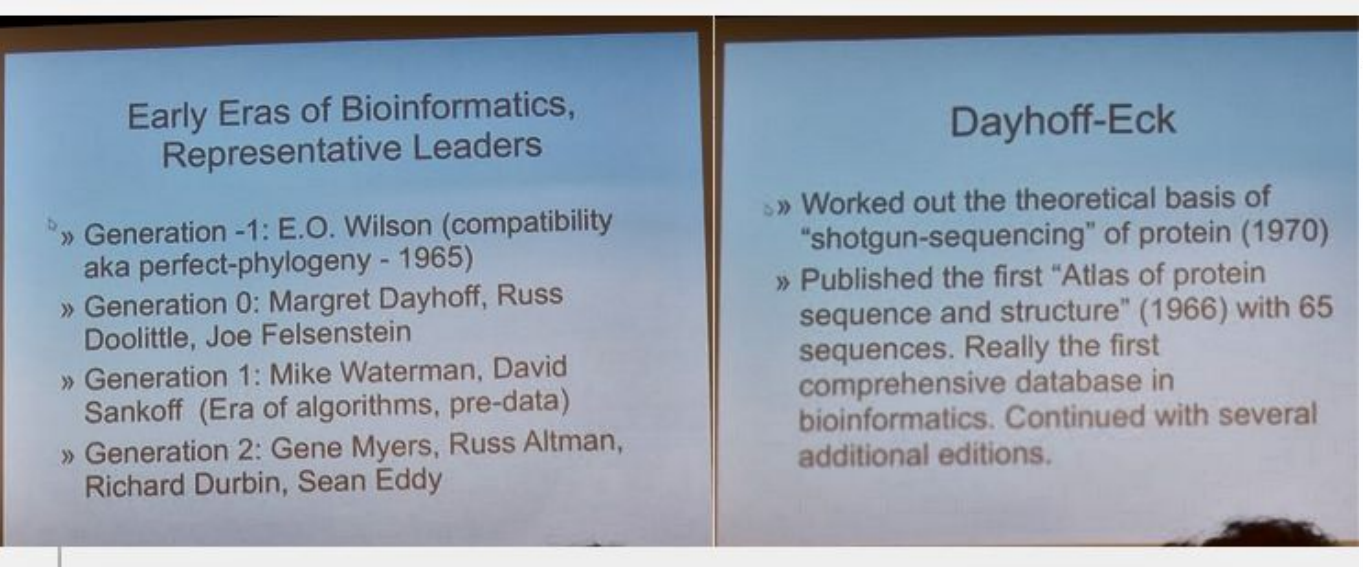
Project Logan: Assembling all public sequencing data

Rayan Chikhi
Institut Pasteur
CGSI 2024

Founding members of biological big data



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.

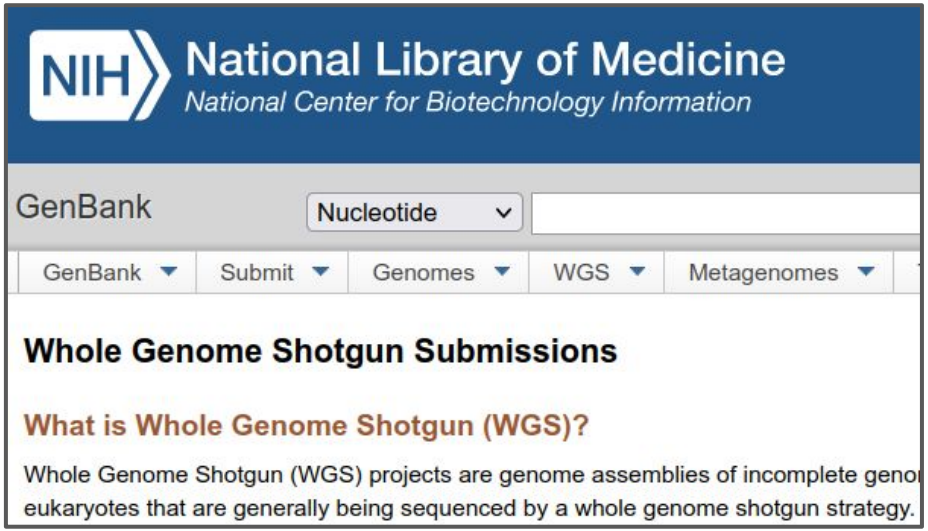
Margaret Oakley Dayhoff

Born	Margaret Belle Oakley March 11, 1925 Philadelphia, Pennsylvania
Died	February 5, 1983

Big data in biology: NCBI GenBank & WGS

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

All sequences are *annotated*



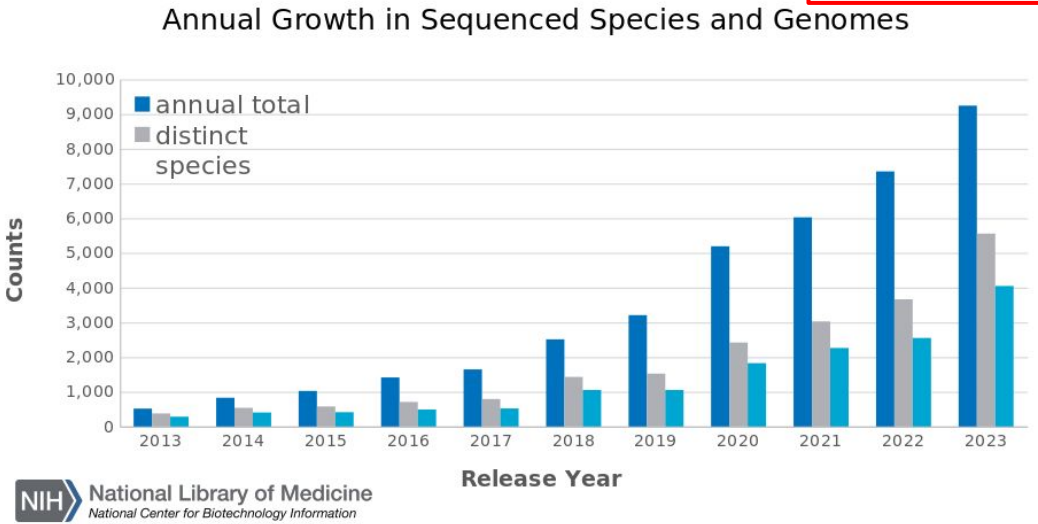
Type: genome assemblies
Size: 16 TB (2022)

Unannotated

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..)



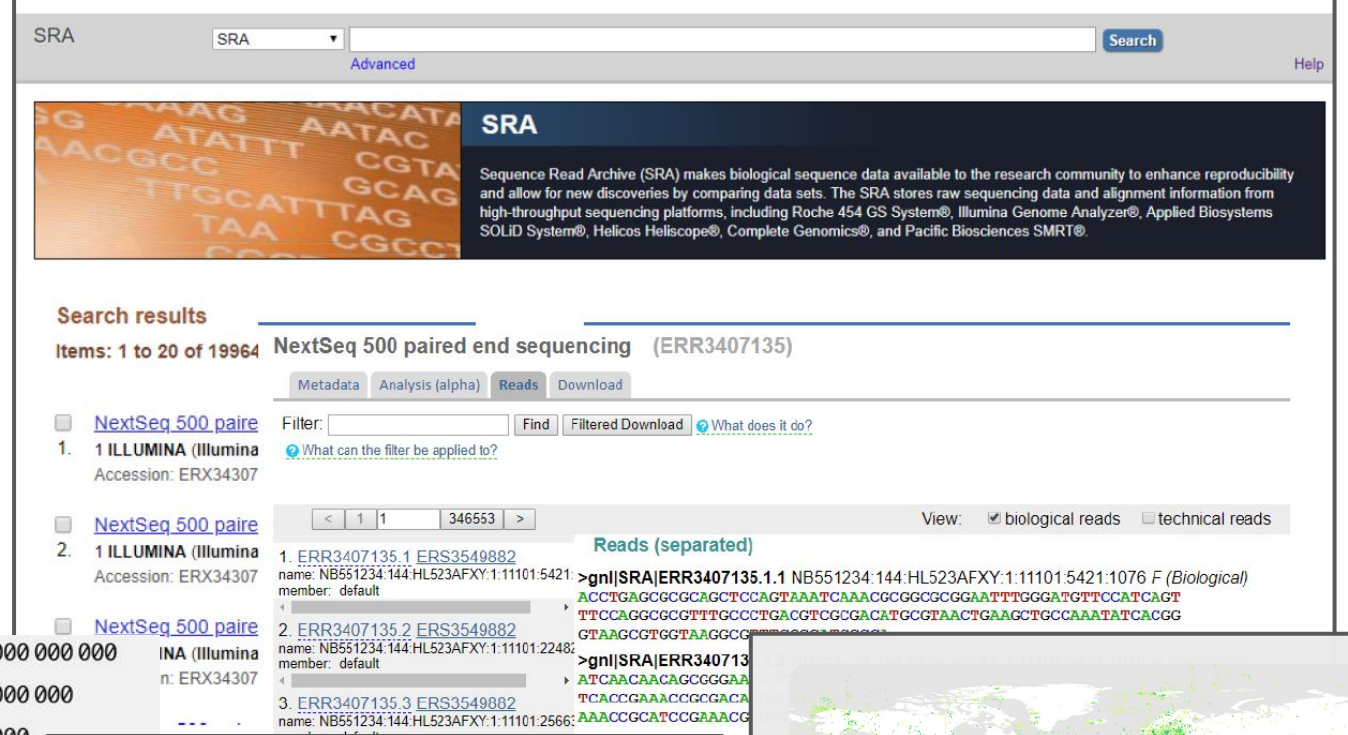
Slide credit: Terence Murphy, NCBI

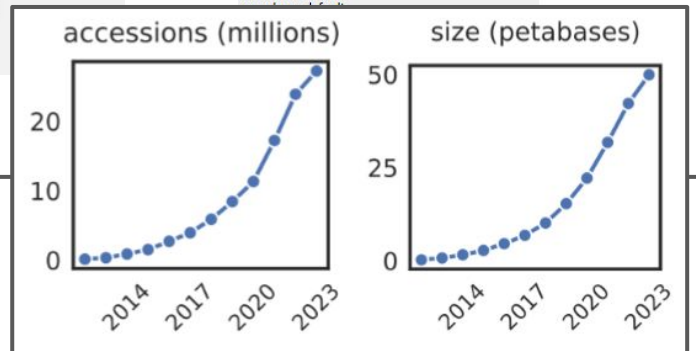
NCBI SRA

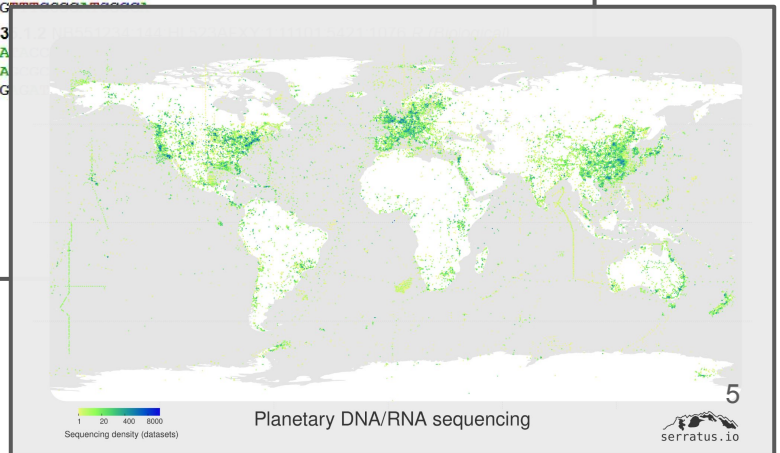
All public sequencing reads

Size: 50 Pbases as of Dec 2023

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



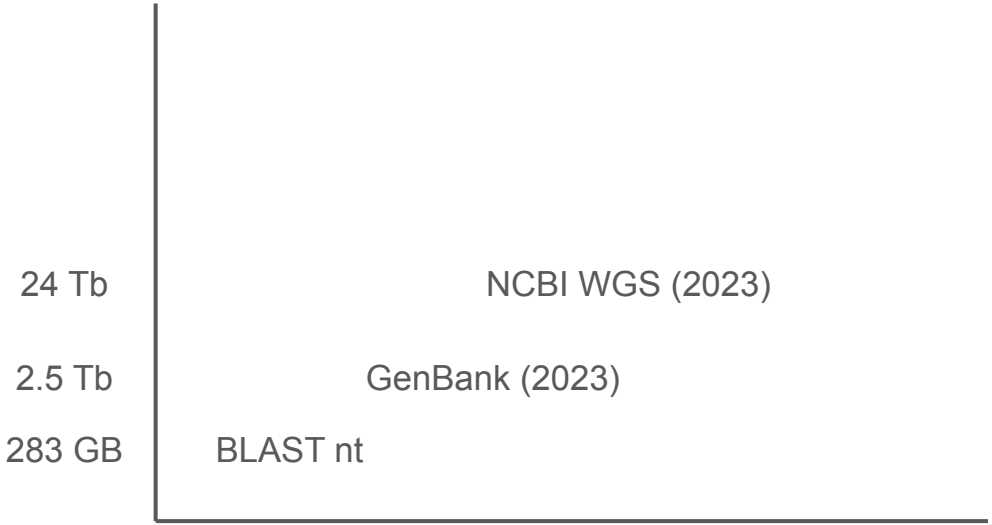




Public sequence datasets

50 Pb

SRA

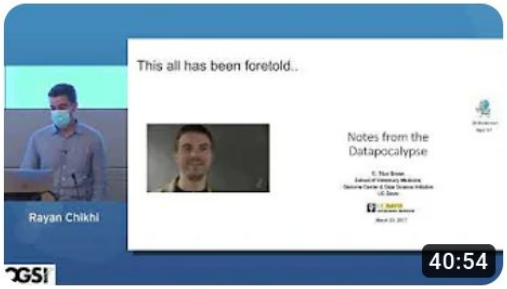


What can be done with the entire SRA?

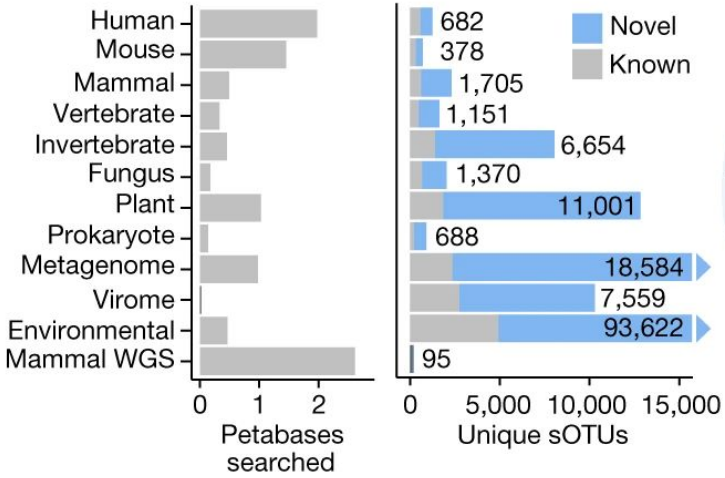
CGSI 2022 talk: **Serratus:** all public RNA-seqs analyzed for viral discovery

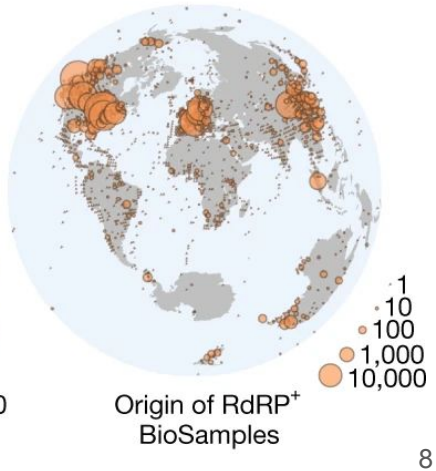
Discovered 130,000 new RNA viral species through large-scale read alignment, 9 new coronaviruses species. One-off **cloud** analysis (Edgar *et al*, Nature, 2022)

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

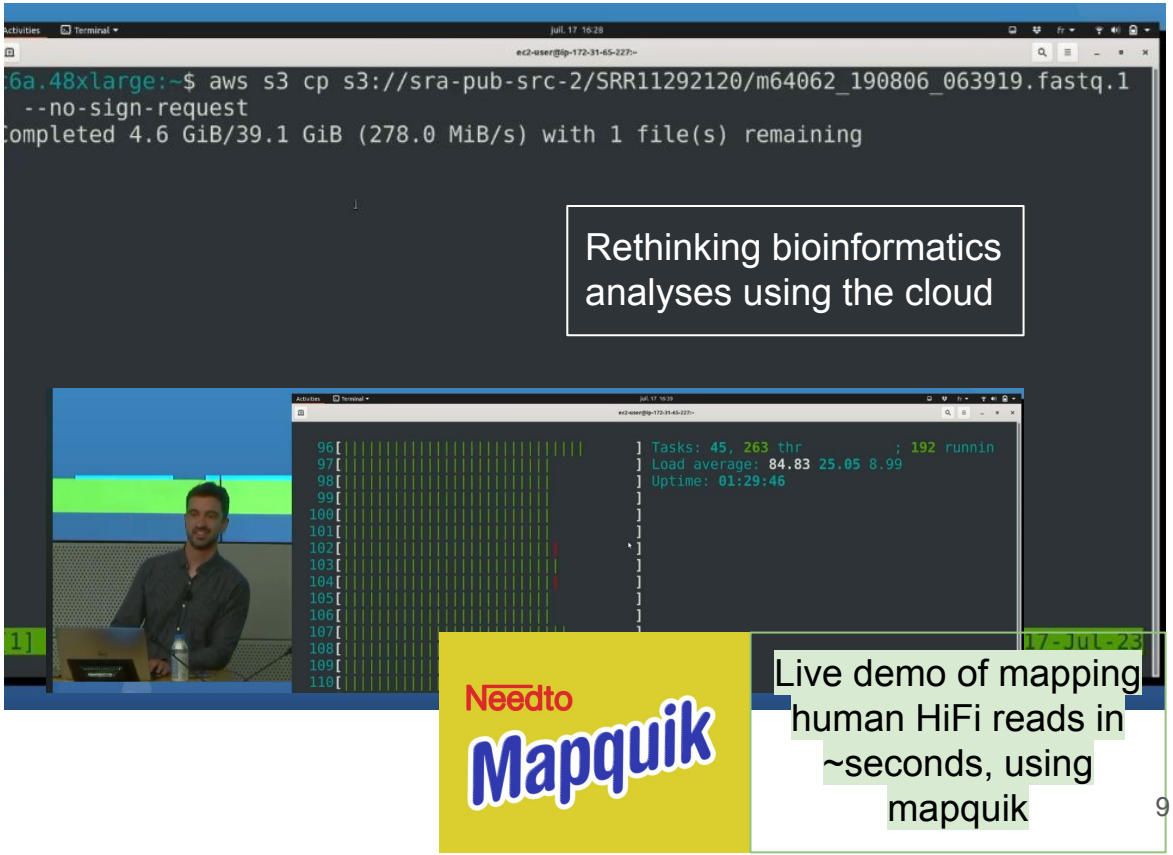


Rayan Chikhi | Recent progress towards petabase-scale genomics





CGSI 2023 talk: **Living in the future of genomics**



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

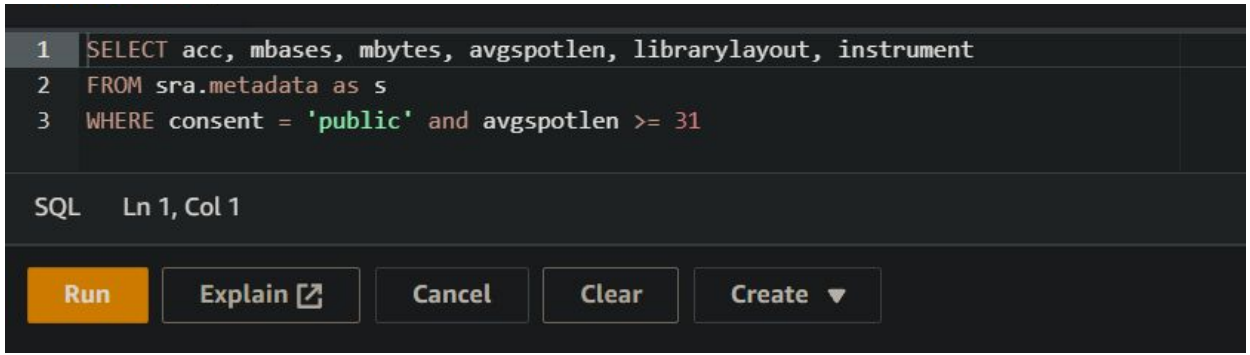
Accessing SRA metadata

~~0. NCBI website~~

1. NCBI FTP metadata

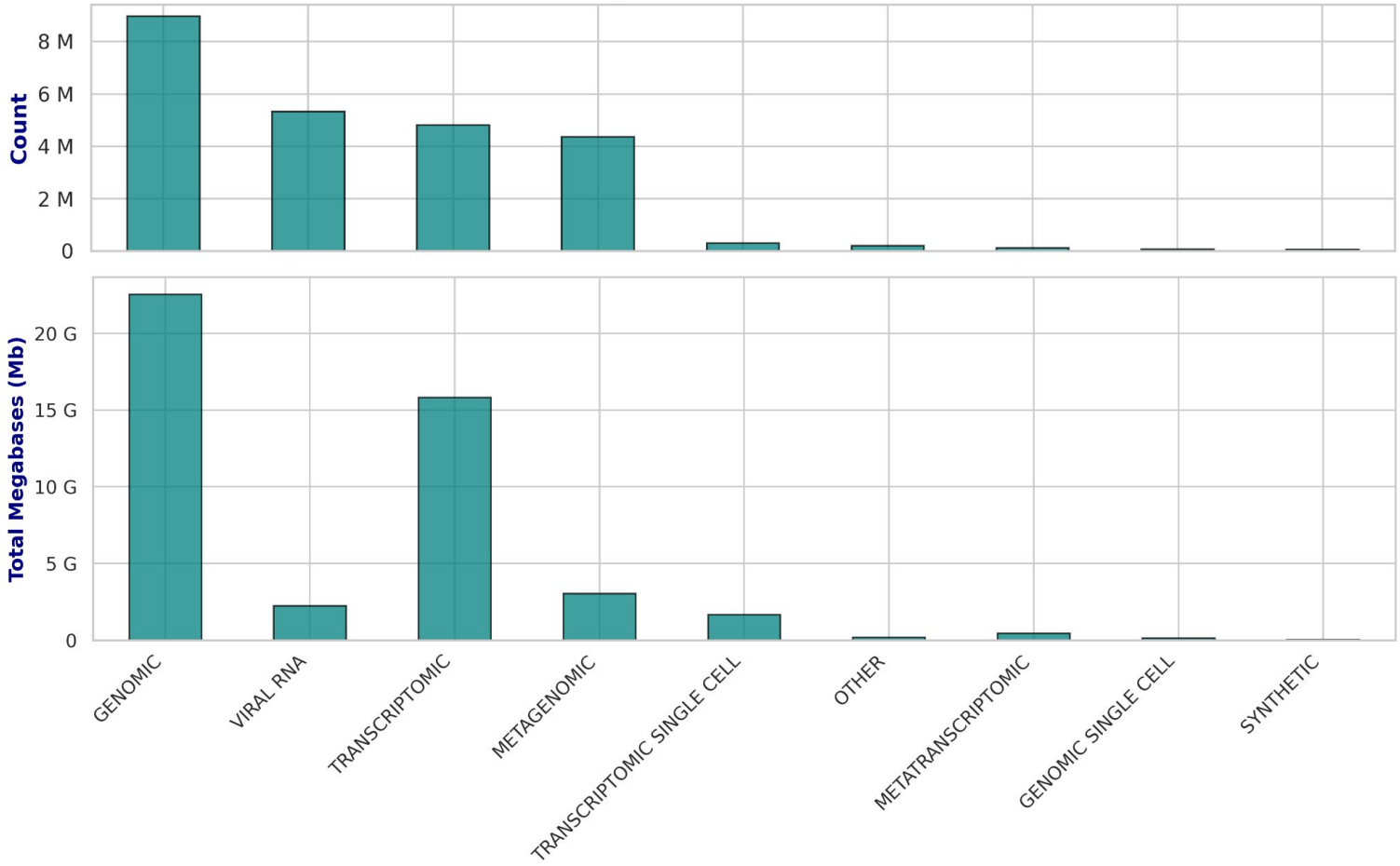
https://trace.ncbi.nlm.nih.gov/Traces/index.html?view=mirroring

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



SRA structured metadata

SRA accessions types (2023)



SRA taxonomy analysis

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%**

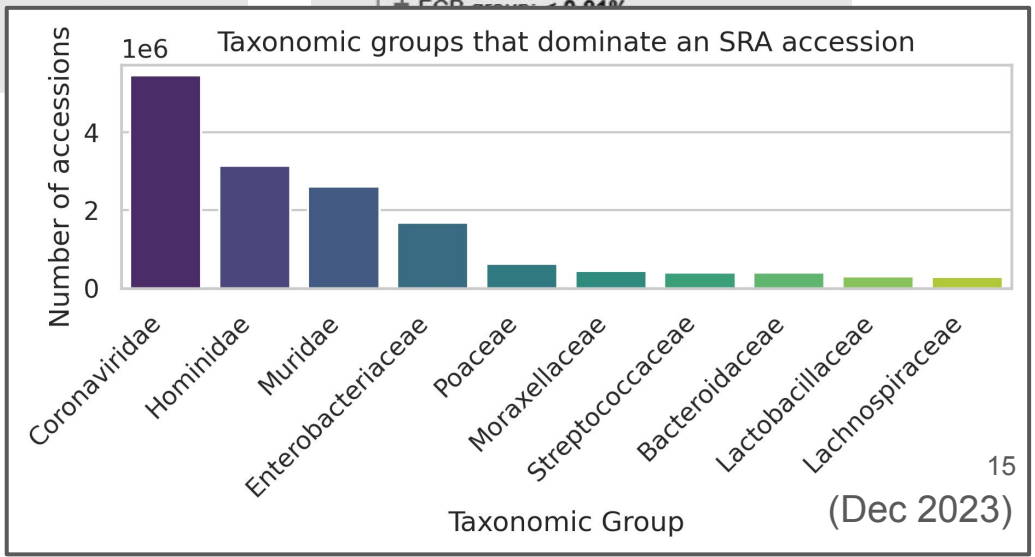
Method | Open Access | Published: 20 September 2021

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

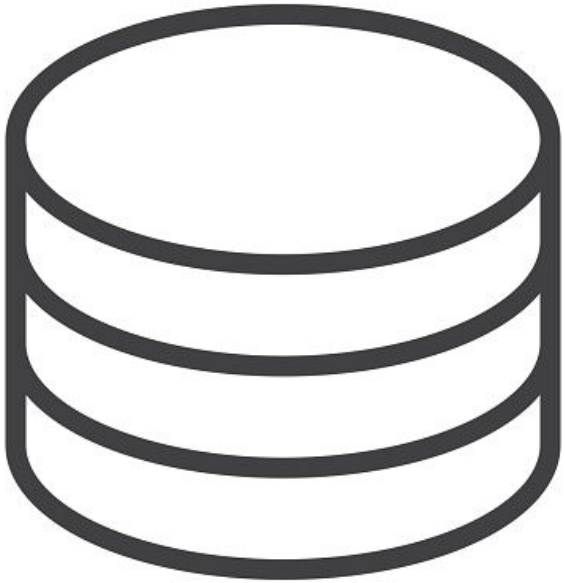
Kenneth S. Katz, Oleg Shutov, Richard Lapoint, Michael Kimelman, J. Rodney Brister & Christopher O'Sullivan

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

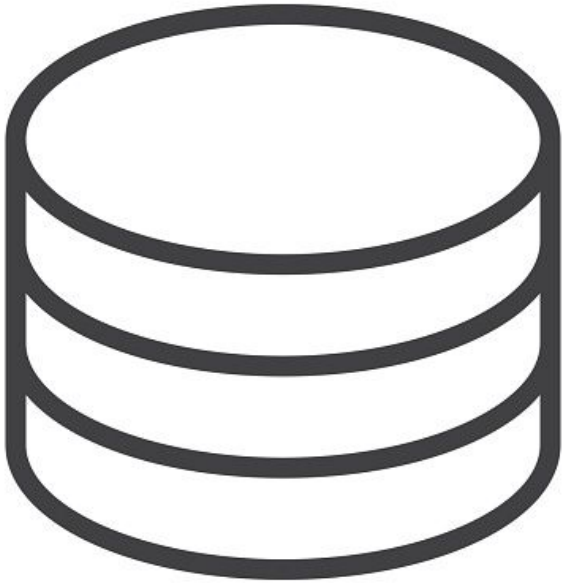
"we have processed more than 27.9 Peta base pairs from runs"



How to analyze the entire SRA? (before Logan)



- How much time to download 20 petabytes at 200 MB/sec?



- How much time to download 20 petabytes at 200 MB/sec?

  ~ 3 years

How to analyze the entire SRA?
(before Logan)
You can’t

Serratus infrastructure

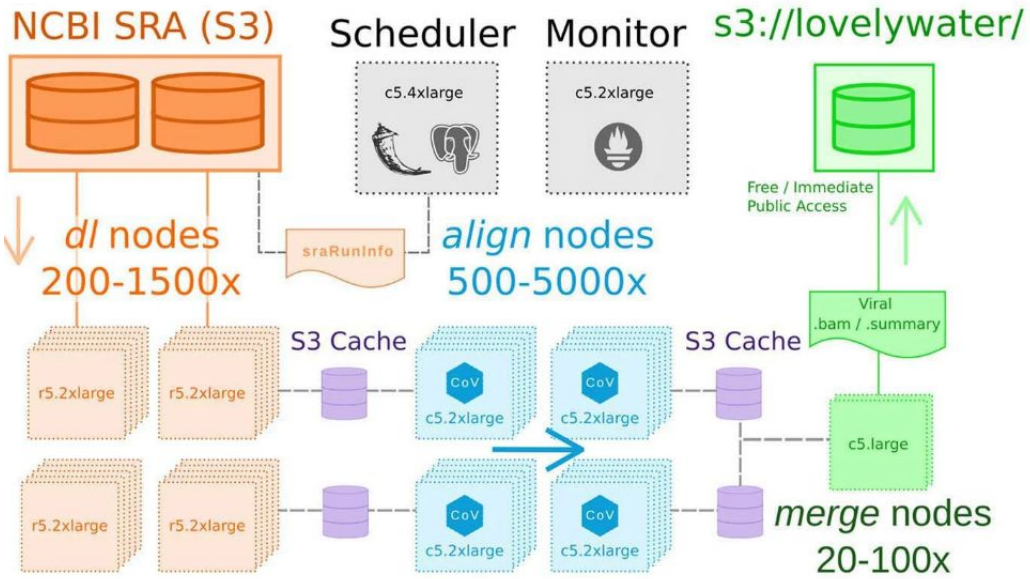


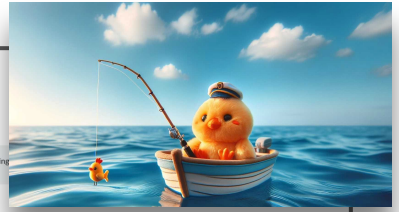
Fig: A. Babaian

How to analyze the entire SRA?
(before Logan)
OK you can but with cloud-scale efforts

How to search sequences in (parts of) the SRA

Branchwater Metagenome Query

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



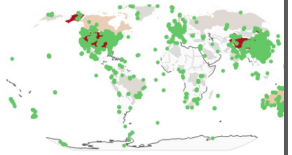
Your query returned 11100 accession IDs. The returned metadata can be pre-filtered prior to .CSV download and plotting

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,-109.139	
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 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.

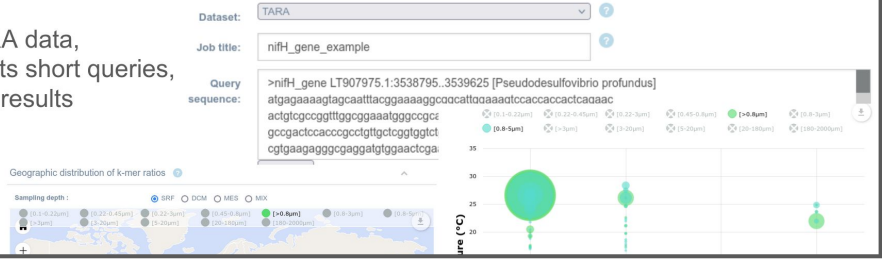
Seven databases currently available are as follows:

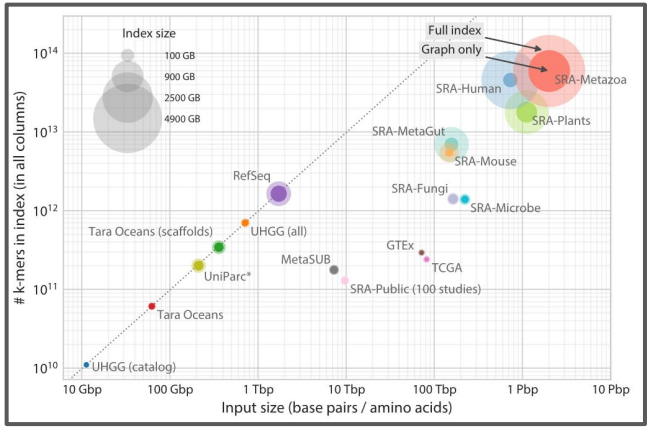
1. **Metagenomic:** All metagenomic and metatranscriptomic runs released in public SRA before the end of 2021
2. **WGS:** All assemblies for the Whole Genome Shotgun sequencing projects available as of Feb 14, 2022
3. **RefSeq:** All assemblies available in the Reference Sequence collection as of April 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

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





New Results Follow this preprint Previous

Indexing All Life's Known Biological Sequences

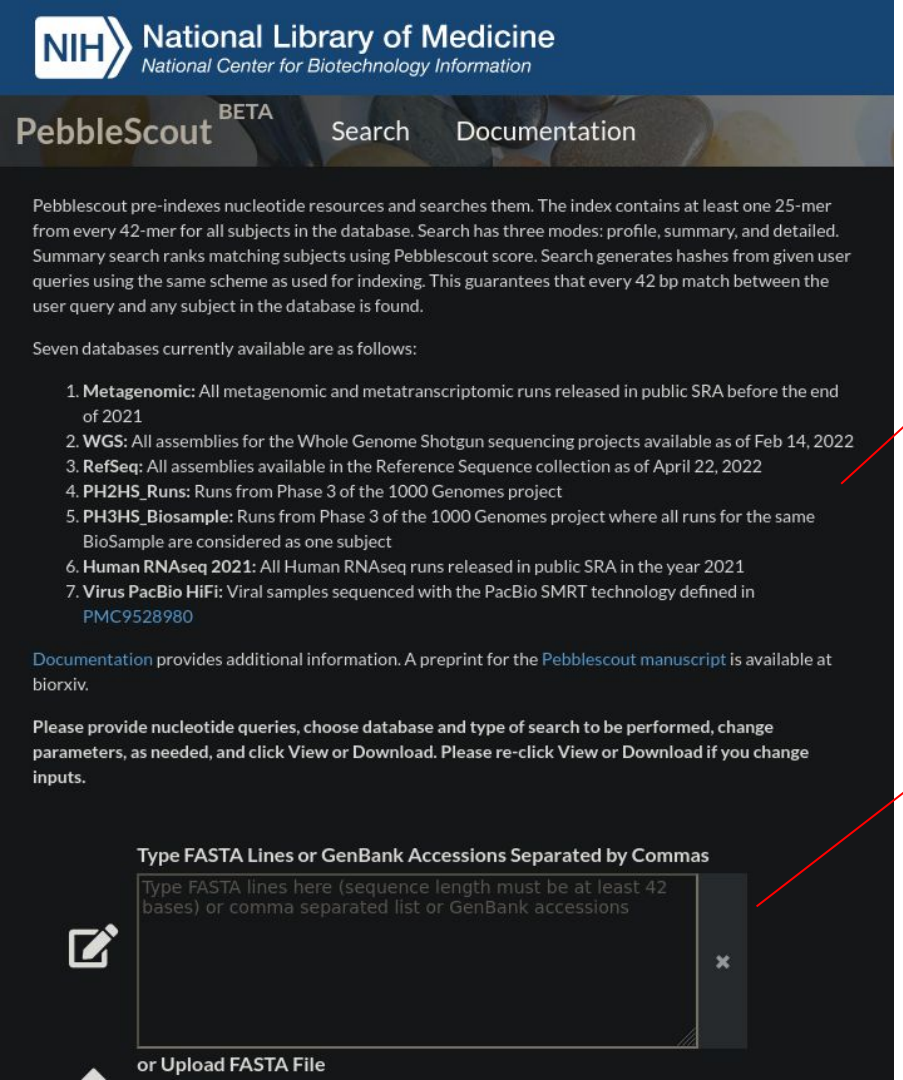
Posted May 14, 2024.

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?].

Download PDF
Print/Save Options
Supplementary Material
Revision Summary



- All metagenomes, all assemblies (WGS), all human RNAseq, RefSeq

- Search for any sequence > 42 nt, using k-mers (minimizers)

Pebblescout usage example

Collaborator needs to search SRA for all samples containing Wolbachia to find new hosts

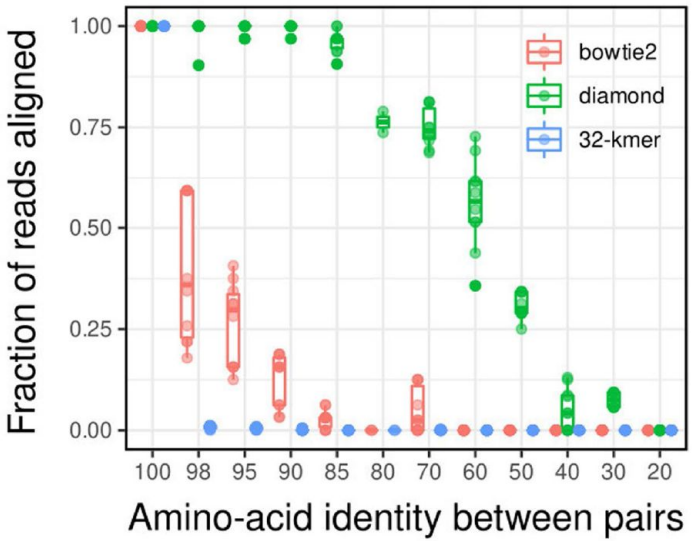
PebbleScout BETA We did exactly this in our paper!

- (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)

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)



Credit: RC Edgar

Logan

Please do not tweet this part

An announcement will be made later

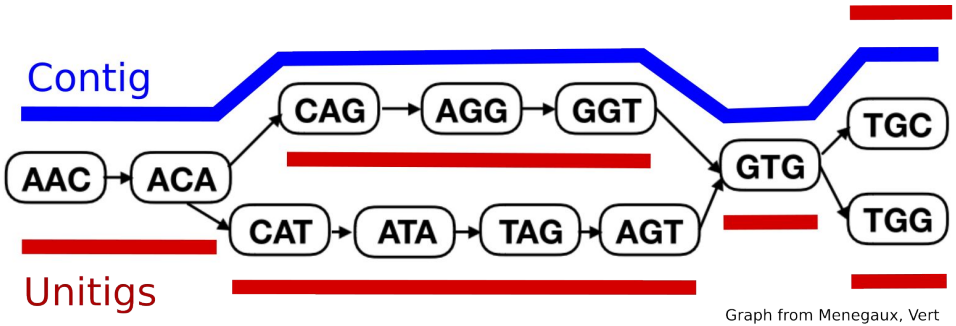
Logan: Outline

- **Assembled the entire SRA** into unitigs (cuttlefish2) and contigs (minia3)
- 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 and 0.4 PB of contigs
- It’s done, finally
- k=31

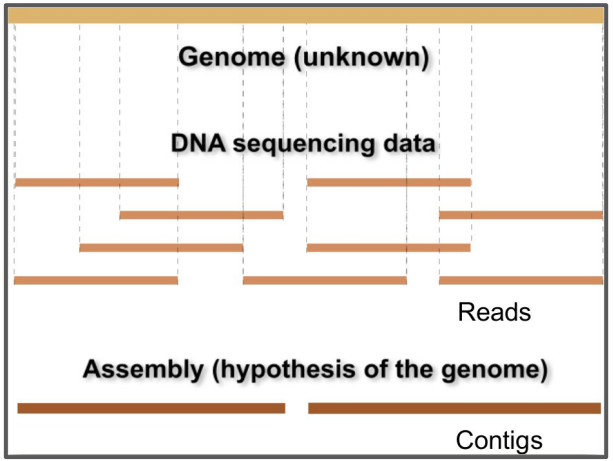
Unitigs? Contigs?

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, ..)

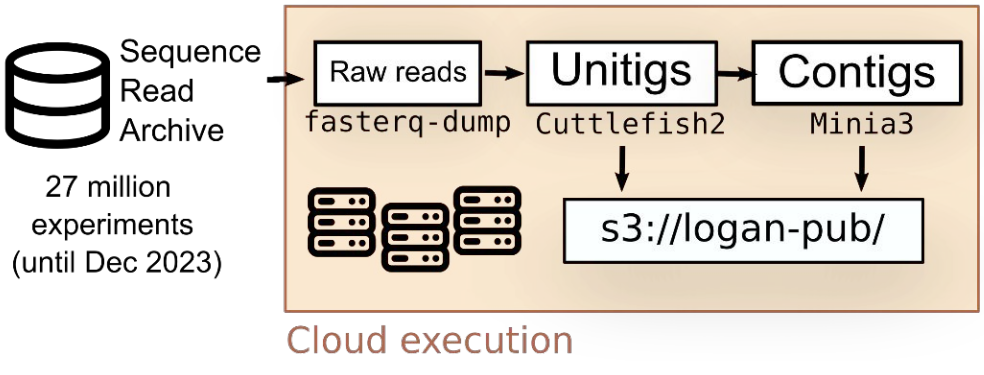
Contigs are consensuses

More on this tomorrow:
K-mer session 11:30, Cohen room. Will be an introduction from scratch

Logan: project steps

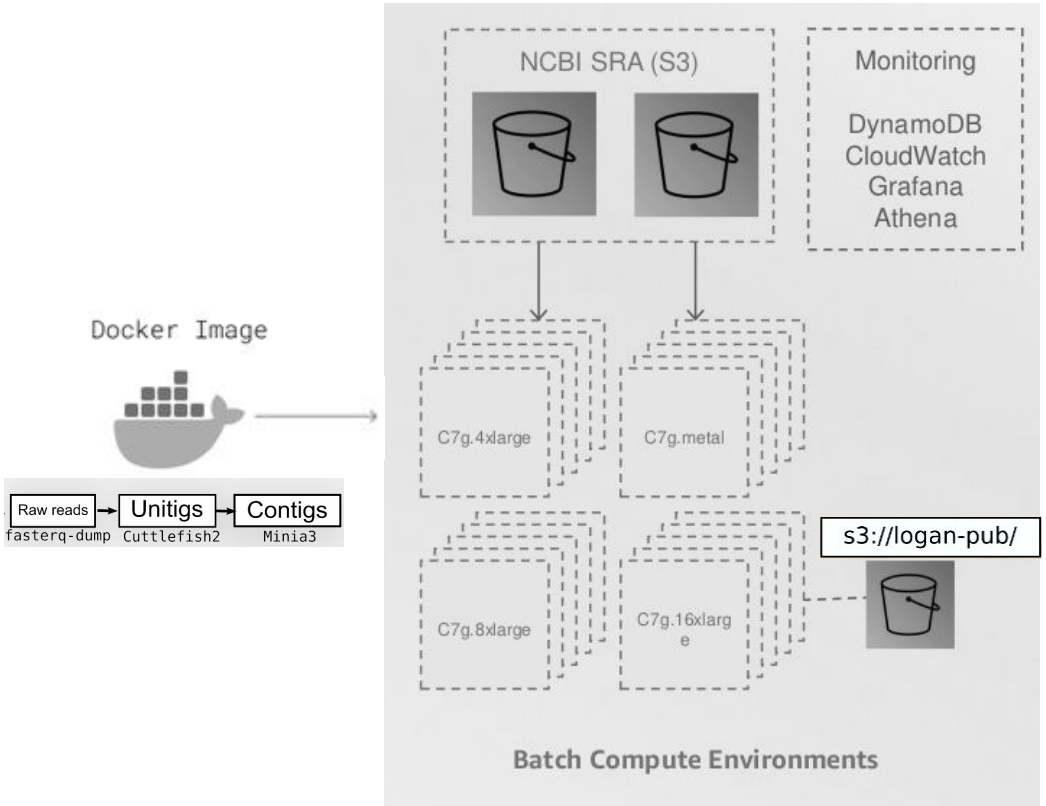
- **Step 1 (2024):** Download all of SRA, assemble each sample, host results publicly [done]

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



- **Step 2 (2025):** Index assemblies, create a search engine (“searching YouTube”) [in progress]

Logan: infrastructure



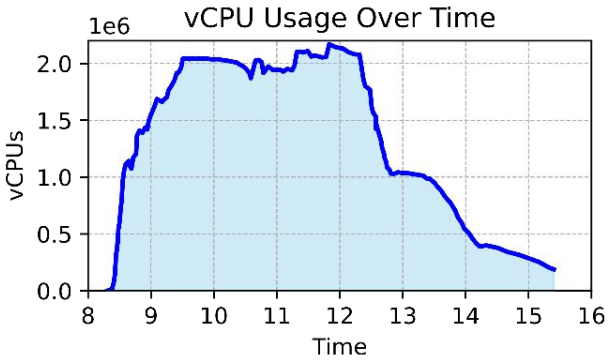
AWS services used:

- 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

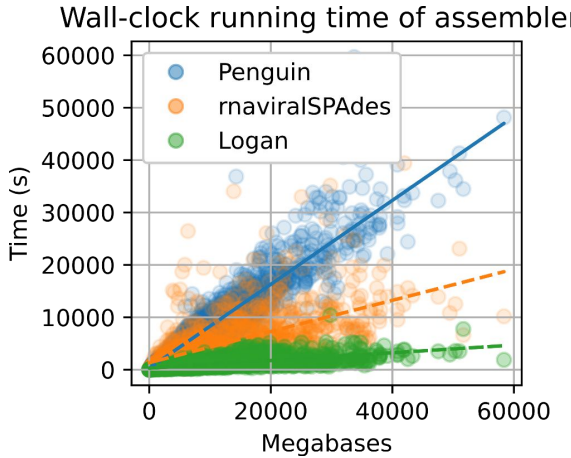
Input data	19.6 petabases
Runtime*	7 hours
Peak Number of Instances	73,100
Peak Number of vCPUs	2.18 million
Peak Total EBS storage	52 petabytes

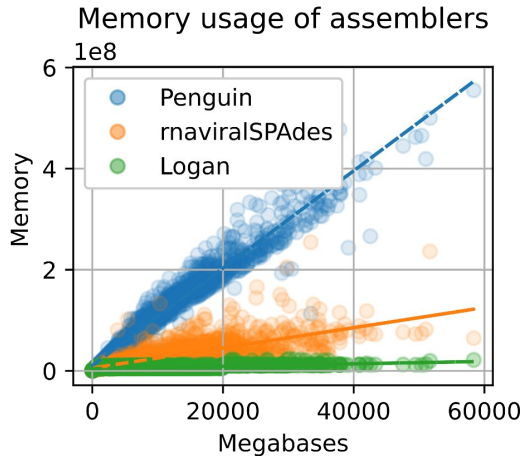
Many failures:

- 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 algorithmic research





Algorithmic components used in Logan

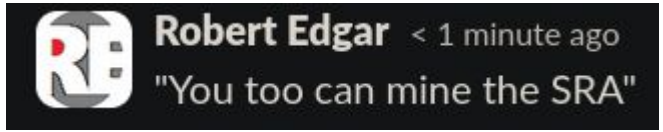
- Minimizer-based kmer counting (KMC inside cuttlefish2)
- Parallel unitigs construction (cuttlefish2)
- Minimum perfect hashing (BBHash inside cuttlefish2, Minia)
- SPAdes assembly graph simplifications (Minia)
- FASTA-aligned zstd block format (f2sz)

Accessing Logan

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

From anywhere, no account needed

Want to dive in Logan data ?



- We do whole-SRA alignments regularly: include your sequence(s) in the next batch

- All Logan unitigs & contigs are public, but if you need assistance: contact me

Many planned analyses

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

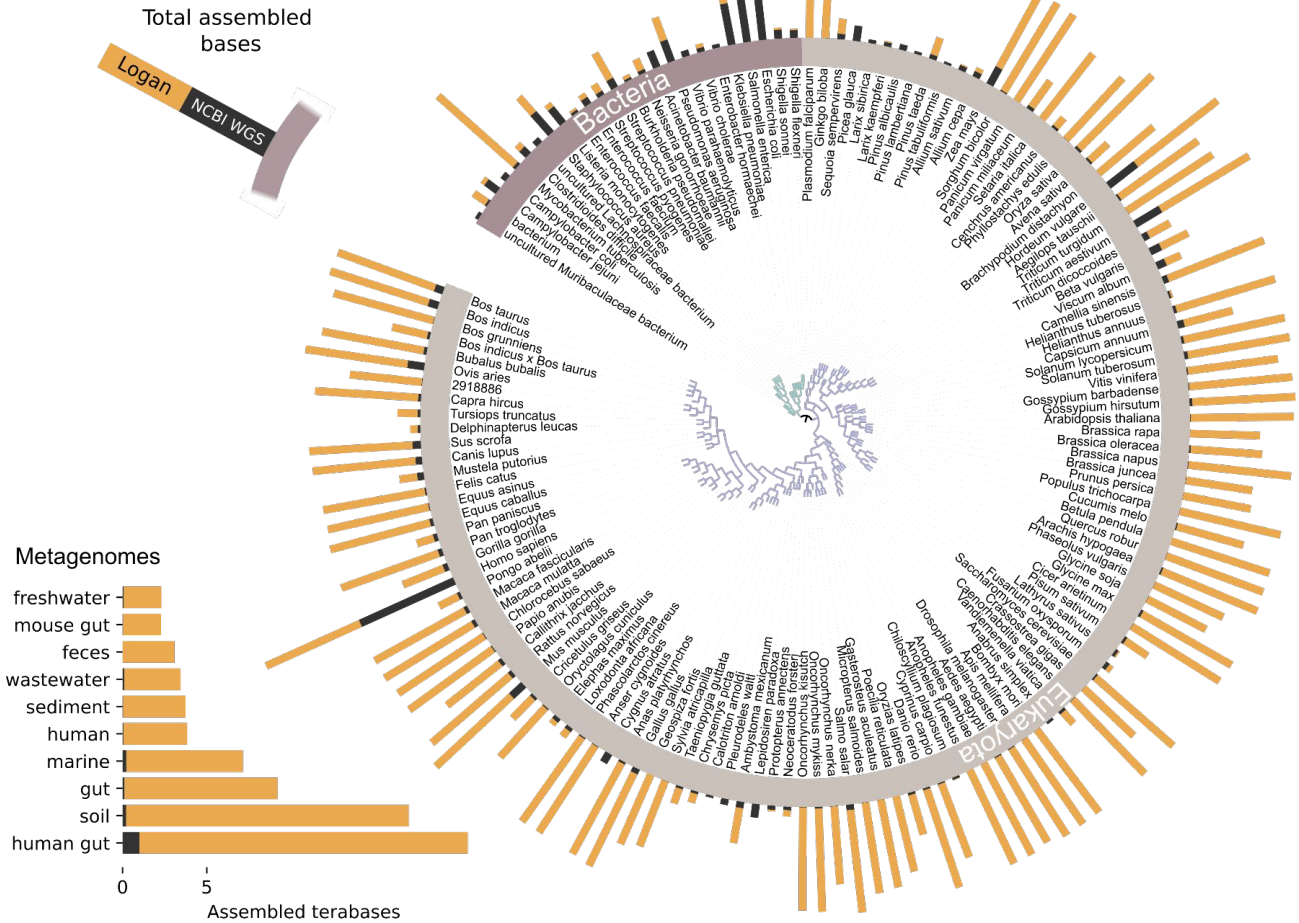
Call for collaborations

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

What’s in Logan

Expansion of assembled contigs across the Tree of Life



Public sequence datasets

50 Pb

SRA (not assembled)

6 Pb

Logan (2024)

24 Tb

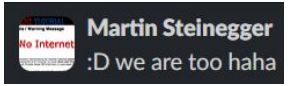
NCBI WGS (2023)

2.5 Tb

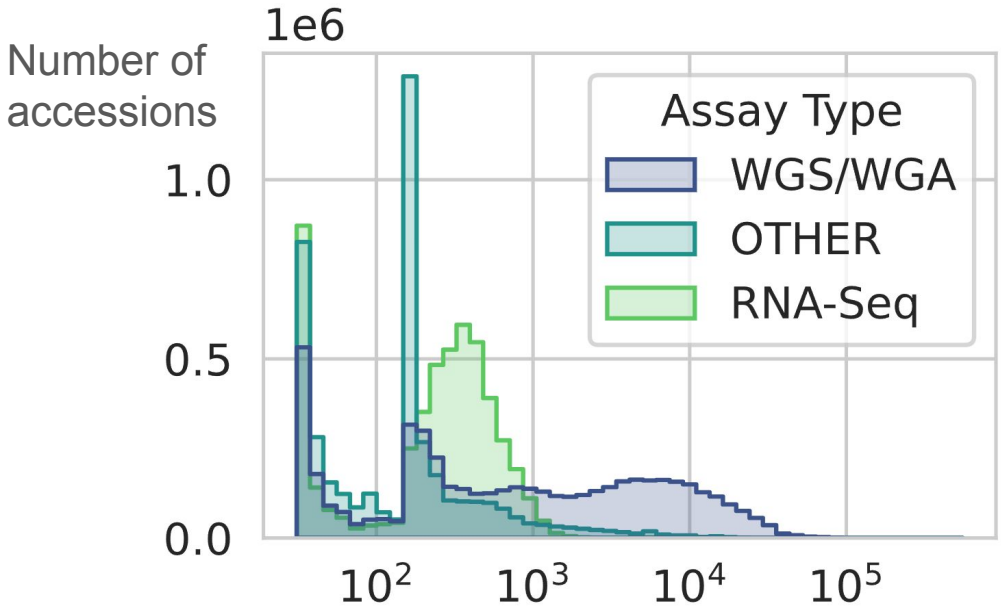
NCBI GenBank (2023)

283 GB

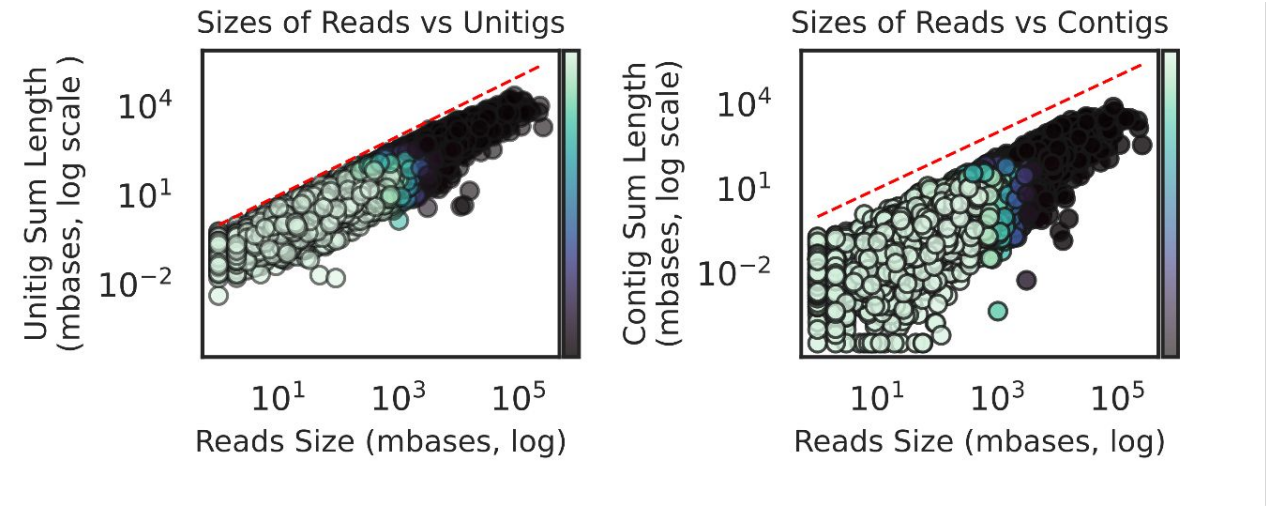
NCBI BLAST nt



Draft-level assembly contiguity



10-100x smaller unitigs/contigs vs reads



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$)

How can Logan be useful?

A “fun” experiment..

Pick an organism: Chicken 

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 (which one?)
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?

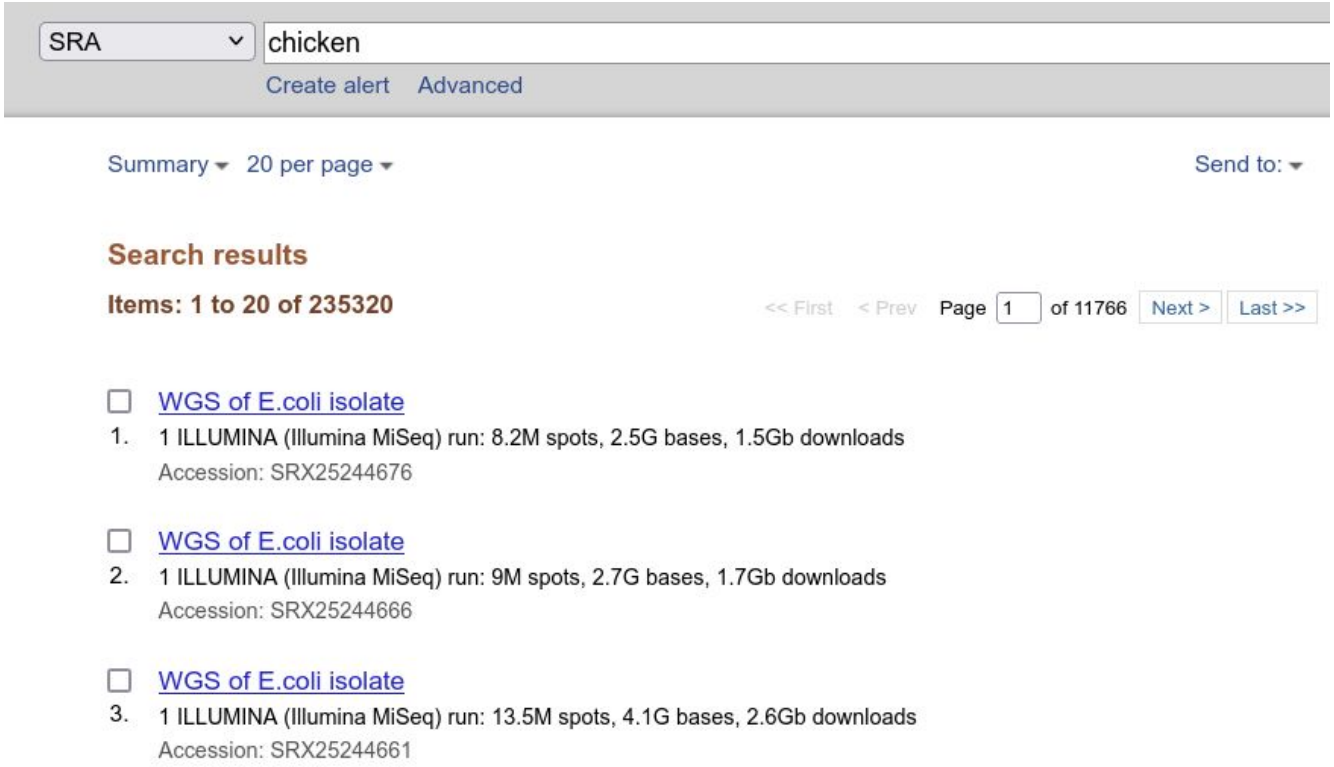
~~0) BLAST~~ Not enough individuals in nt

~~1) NCBI Pebblescout~~ Only has metagenomes

2) SRA metadata query

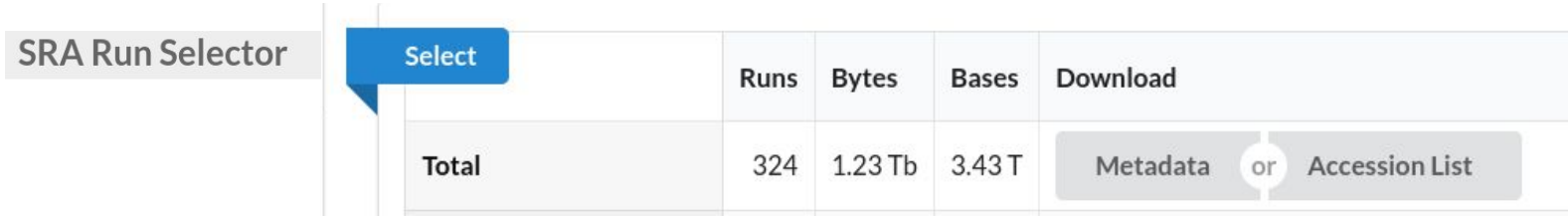
3) SRA taxonomy query

SRA metadata query 1: fail

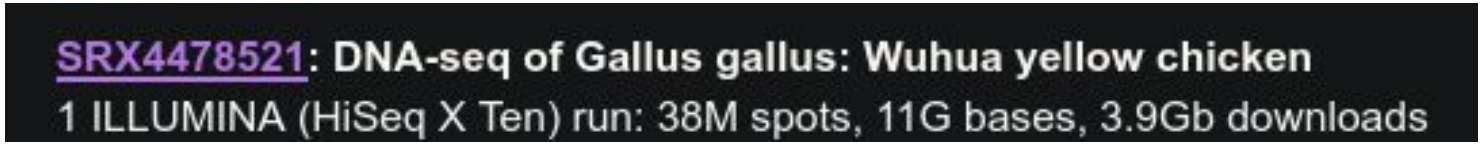


SRA metadata query 2: better

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



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



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
  - 4.3 terabases of contigs
  - Raw data is **374 terabases** of reads 😱 (= 1000GP twice)

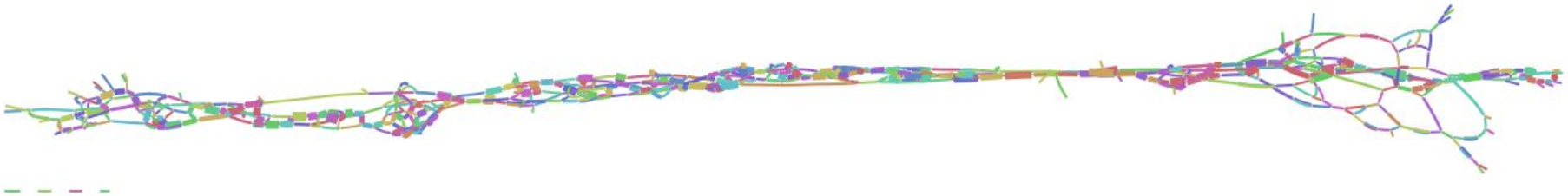
Logan analysis

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)

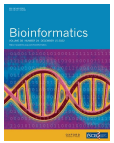


GWAS directly from sequences (skips SNP detection):

TGGGGGTCATCGCCGTGGACCGCTACATCG..

TGGGGGTCATCGCCGTGGACCGCTACATC**A**..

$p<10^{-7}$



JOURNAL ARTICLE

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

Téo Lemane, Rayan Chikhi, Pierre Peterlongo

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

Conclusion

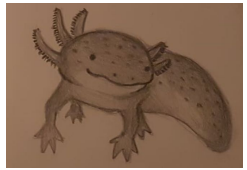
- **SRA-scale analyses now 100x more tractable**
- **Logan: all of Life’s genomic data at your fingertips**

Technically:

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

Sequence Bioinformatics

INSTITUT PASTEUR



Lab members:

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

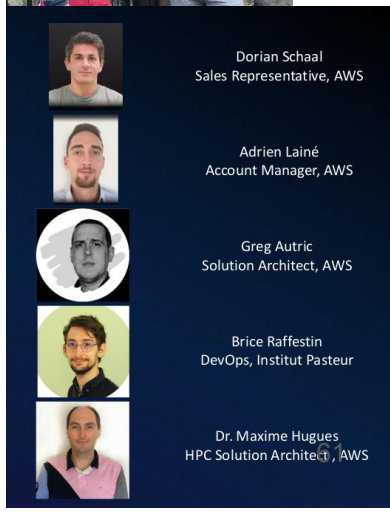
Alumni:

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

AWS support (Dorian Schaal, Adrien Lainé)



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

Thank you for your attention!