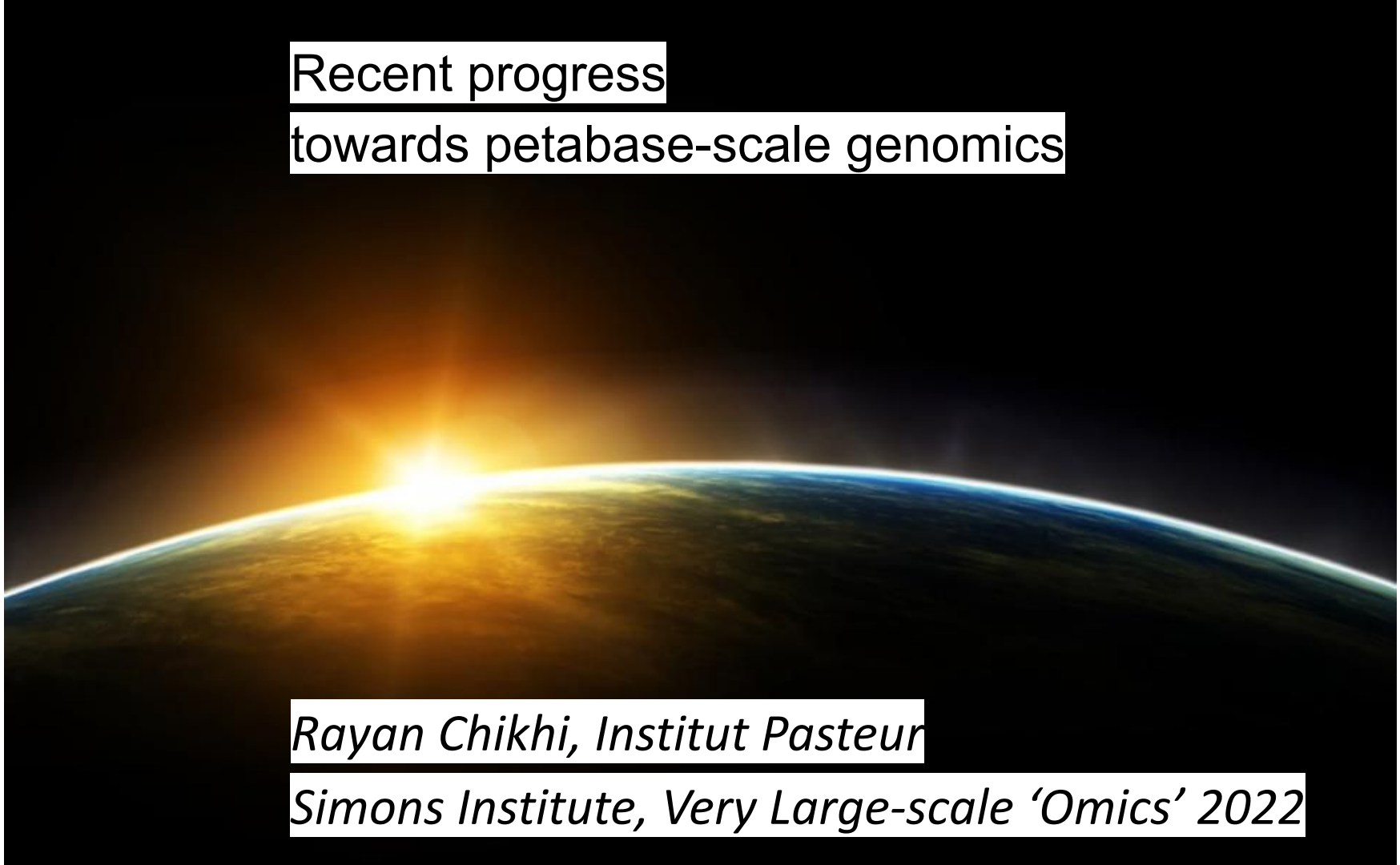
# Recent progress towards petabase-scale genomics



*Rayan Chikhi, Institut Pasteur*

*Simons Institute, Very Large-scale 'Omics' 2022*

# Hello

- PI, Bioinformatics algorithms lab @ Institut Pasteur
- CV: PhD@ENS Rennes, Postdoc@PSU, CNRS

Research:

- *de novo* assembly
- k-mer methods
- metagenomics
- large-scale bioinfo

@RayanChikhi on Twitter

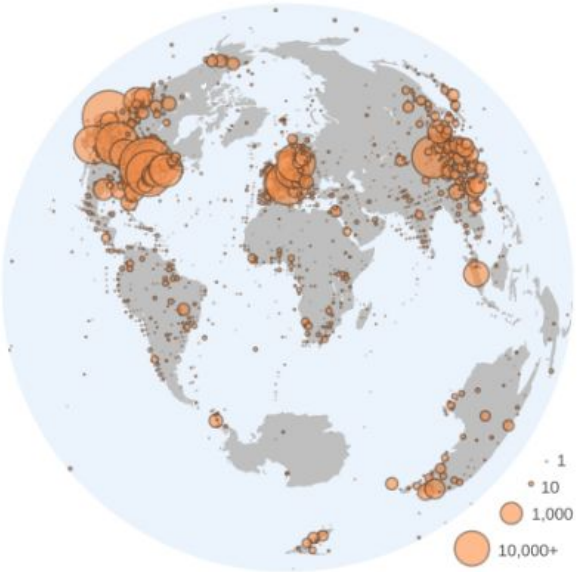
http://rayan.chikhi.name

# Petabase-scale viral discovery

*Rayan Chikhi, on behalf of the Serratus team*

**We analysed all available RNA sequencing data and discovered 10x more viruses species than previously known, including coronaviruses.**

**Nature, 2022**

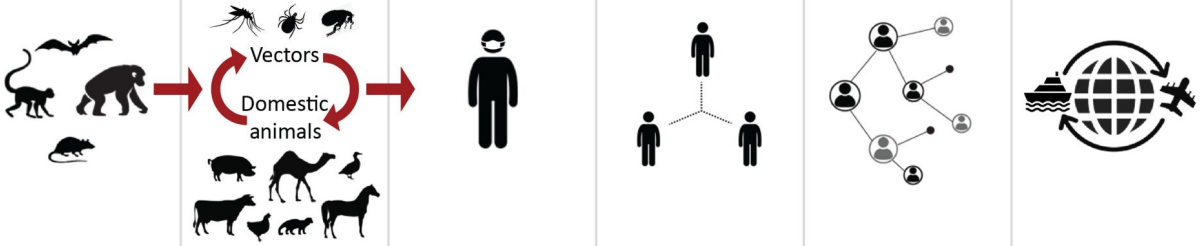



cnrs PR[AI]RIE PaRis Artificial Intelligence Research InstitutE INSTITUT PASTEUR https://serratus.io/

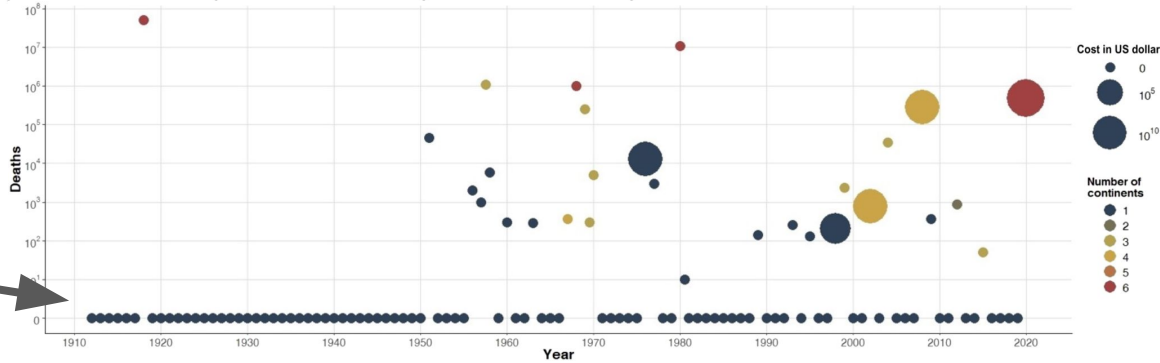
# Viral surveillance in the age of pandemics

Source: https://www.science.org/doi/10.1126/sciadv.abl4183

| | Pre-emergence | Pathogen spillover | Emergence | Localized transmission | Epidemic | Pandemic |
|---|---|---|---|---|---|---|
| Prevention and control | **Anticipation**<br>Viral surveillance | **Primary prevention**<br>Reduction of deforestation, agriculture, wildlife trade | **Early detection**<br>Syndromic surveillance | **Containment**<br>Contact tracing<br>Isolation | **Control & mitigation**<br>Widespread testing<br>quarantine<br>scale-up health care<br>school and business closure | **All prior measures brought to a larger scale** |
| Stages of pathogen spread | **Pathogen evolution in wild animals** | **Direct spillover** | | | **Human amplification** | **Global spread** |




We're here




Would be nice to contain viruses there

# SARS-CoV-2 circulate(s|d) among animals

PETS & ANIMALS

## Tiger at zoo in Knoxville tests positive for SARS-CoV-2, two others possibly infected

A veterinary team from the University of Tennessee College of Veterinary Medicine is taking care of the three tigers.

CNNWire By Joe Wenzel

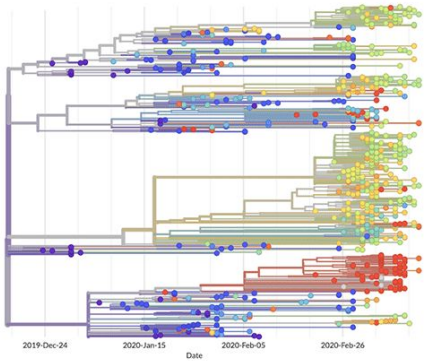
Saturday, October 31, 2020

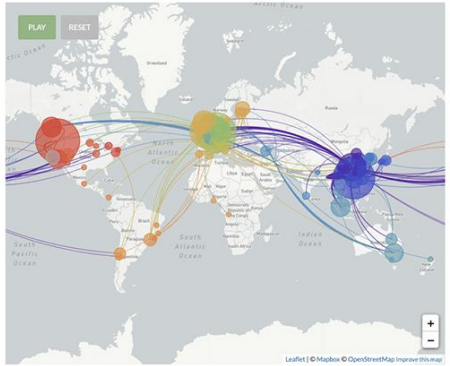
## Ontario dog believed to be first in Canada to test positive for COVID-19

Officials said that the risk of infection and illness in most domestic animals is low

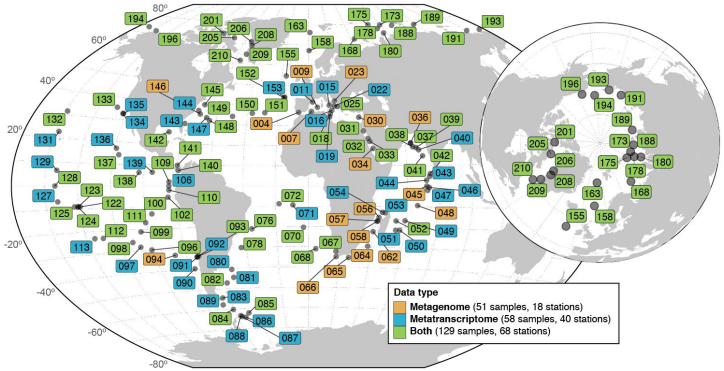
KATYA SLEPIAN / Oct. 26, 2020 1:45 p.m. / CANADA & WORLD / NEWS


# Enter sequencing efforts




Nextstrain




Tara Oceans, Salazar et al. (2019)

https://www.nature.com/articles/d41586-021-01069-w

SRA | SRA ▾ | Search
Advanced | Help

## 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 paired end sequencing (ERR3407135)

Metadata | Analysis (alpha) | Reads | Download

Filter: Find | Filtered Download | What does it do?

What can the filter be applied to?

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

< 1 1 346553 >

View: ☑ biological reads ☐ technical reads

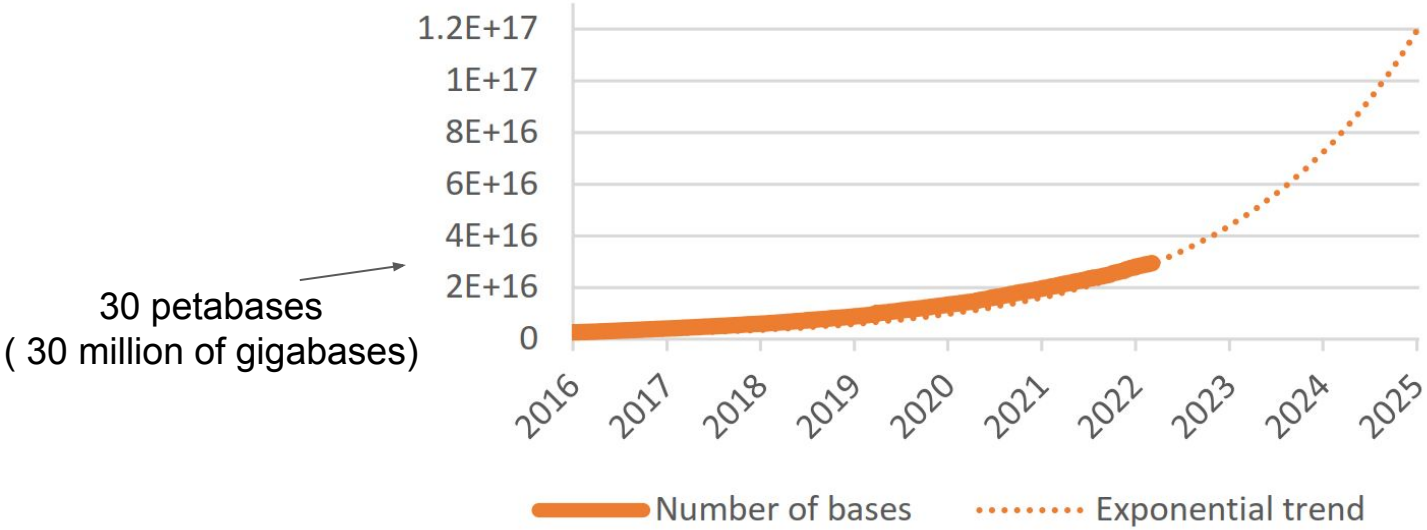
1. ERR3407135.1 ERS3549882
   name: NB551234:144:HL523AFXY:1:11101:5421:
   member: default
2. ERR3407135.2 ERS3549882
   name: NB551234:144:HL523AFXY:1:11101:2248
   member: default
3. ERR3407135.3 ERS3549882
   name: NB551234:144:HL523AFXY:1:11101:2566:
   member: default
4. ERR3407135.4 ERS3549882
   name: NB551234:144:HL523AFXY:1:11101:2119
   member: default
5. ERR3407135.5 ERS3549882
   name: NB551234:144:HL523AFXY:1:11101:2350
   member: default

### Reads (separated)

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

**>gnl|SRA|ERR3407135.1.2** NB551234:144:HL523AFXY:1:11101:5421:1076 *R (Biological)*
ATCAACAACAGCGGGAATACCACCTCTTCCAGCCGTTGTTTCCAACCAATACGCGTTAAT
TCACCGAAACCGCGACAGCGCAATGGAACGCATCATTGCGCAGGTGTTGCAGAATACGGA
AAACCGCATCCGAAACGAGATGCGCGTTAAT

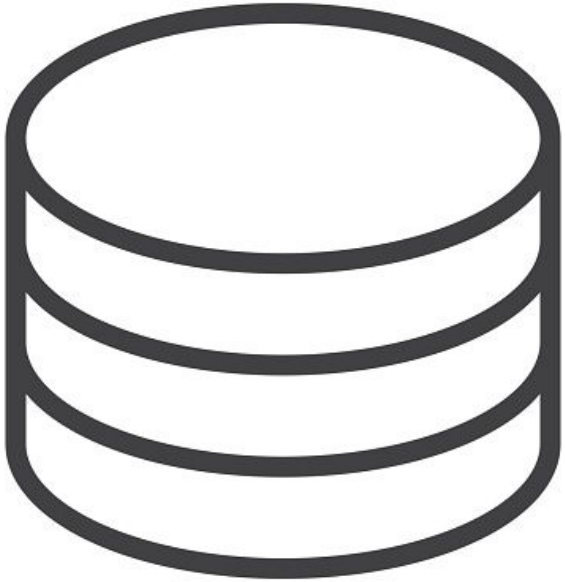
# Growth of the Sequence Read Archive



NCBI SRA database : 30 PB

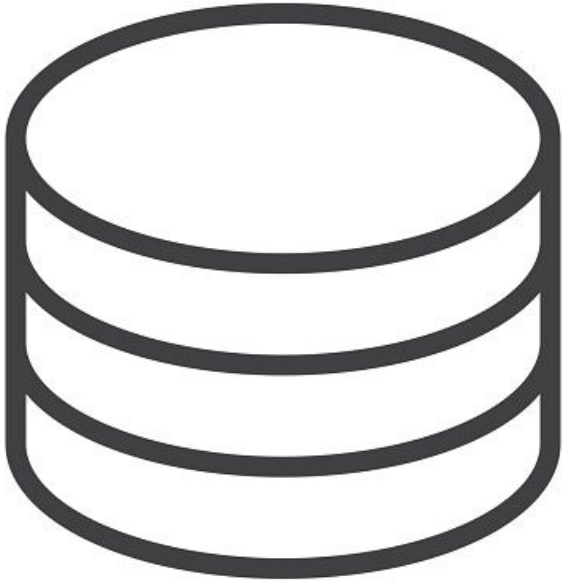

<u>Data crypt</u>

All the raw reads sleep there, undisturbed



All RNA-seqs (2008-2020)
5 million samples, 10.2 Petabases

Downloading all RNA-seq samples:

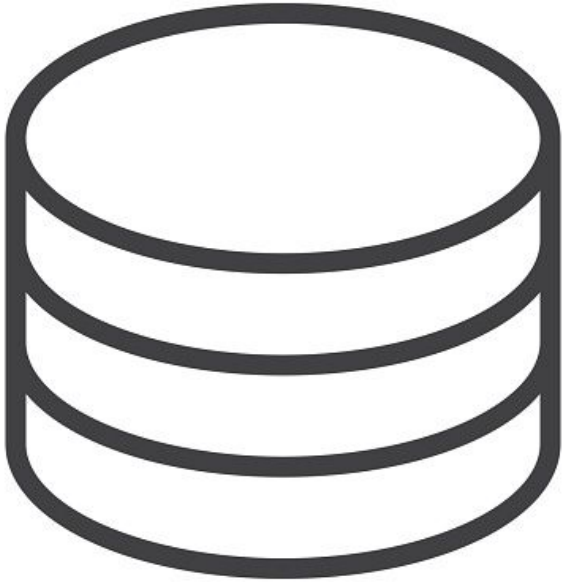


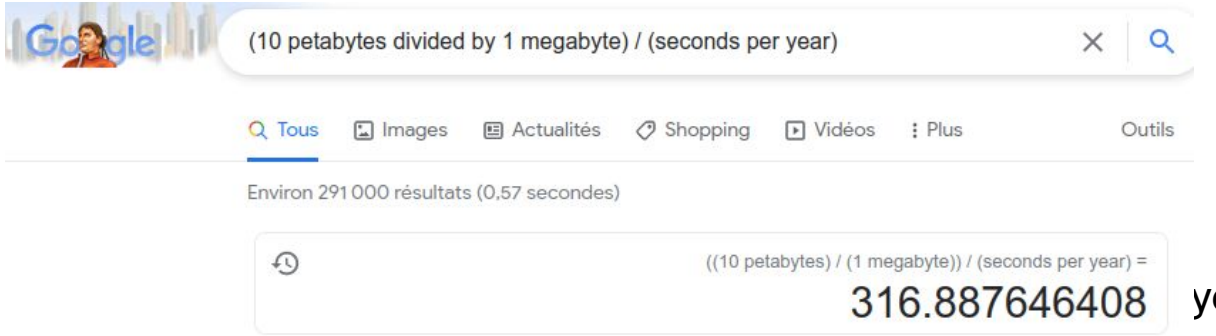
<u>Guesstimate</u>:

- How many years would it take to download 10 petabases (i.e. 10,000,000,000 MB) at 1 MB/sec?

Hint: ~30,000,000 seconds in a year

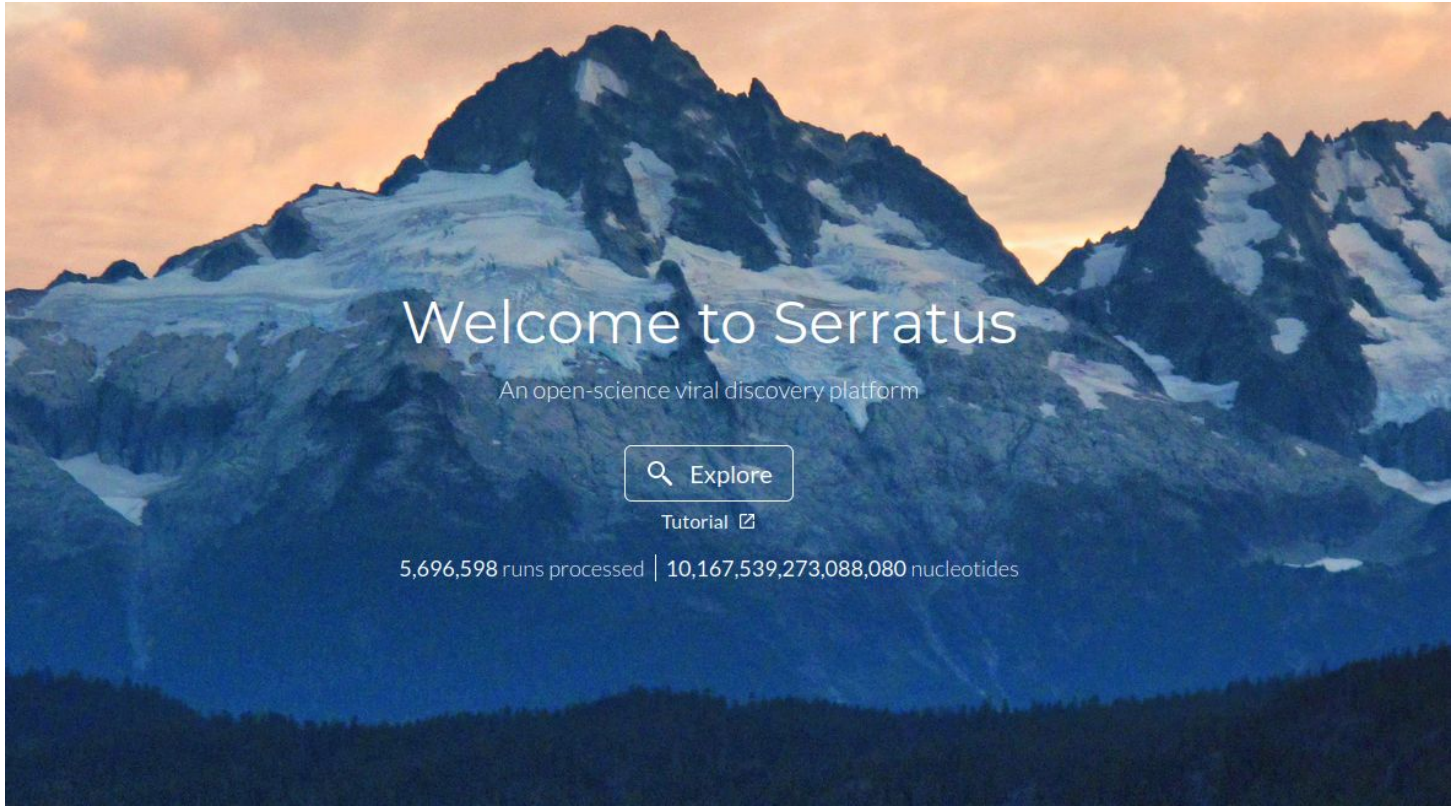
Downloading all RNA-seq samples:






years at 1 MB/s

# Serratus: a cloud analysis of all RNA-seqs



# Serratus: two analyses

1) Nucleotide alignments

    *all RNAseqs vs all RNA viral genomes*

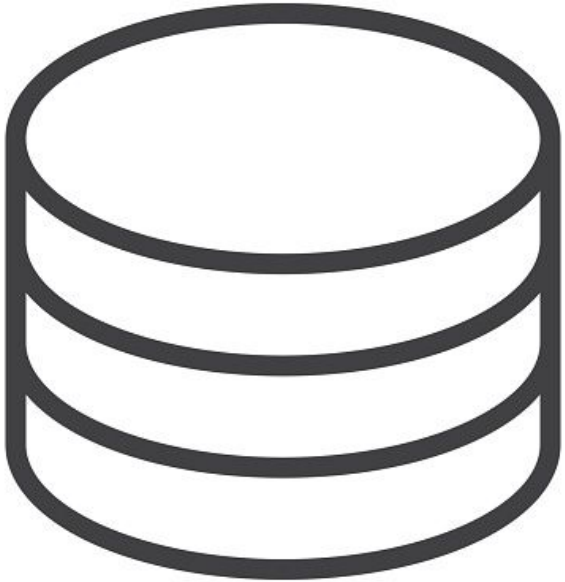
    **> Discovered new coronaviruses**

2) Protein (translated) alignments

    *all RNAseqs vs a universal RNA virus gene*

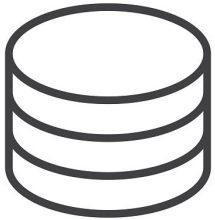
    **> Discovered 130,000 new RNA virus species**

Analysis 1:



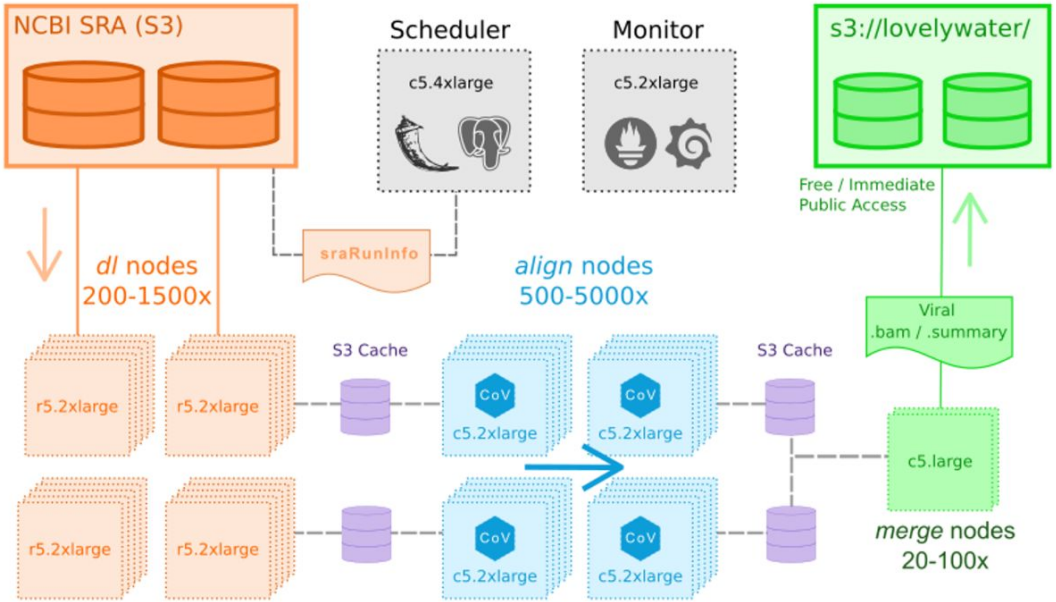
All RNA-seqs

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



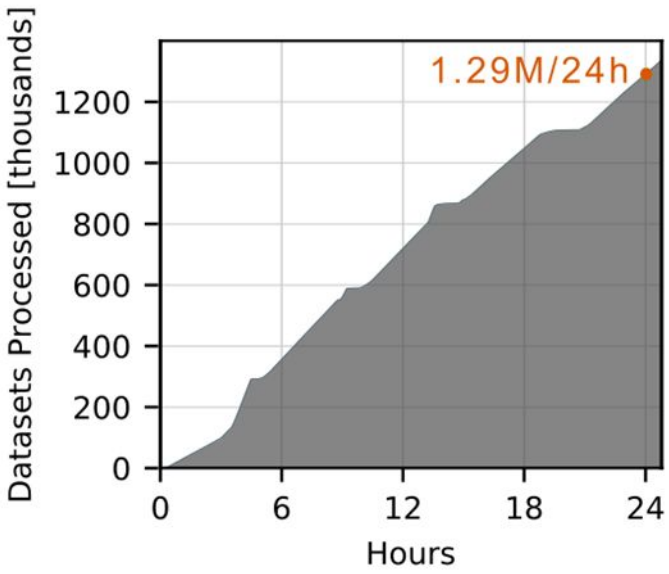
**55,715 CoV+ samples**

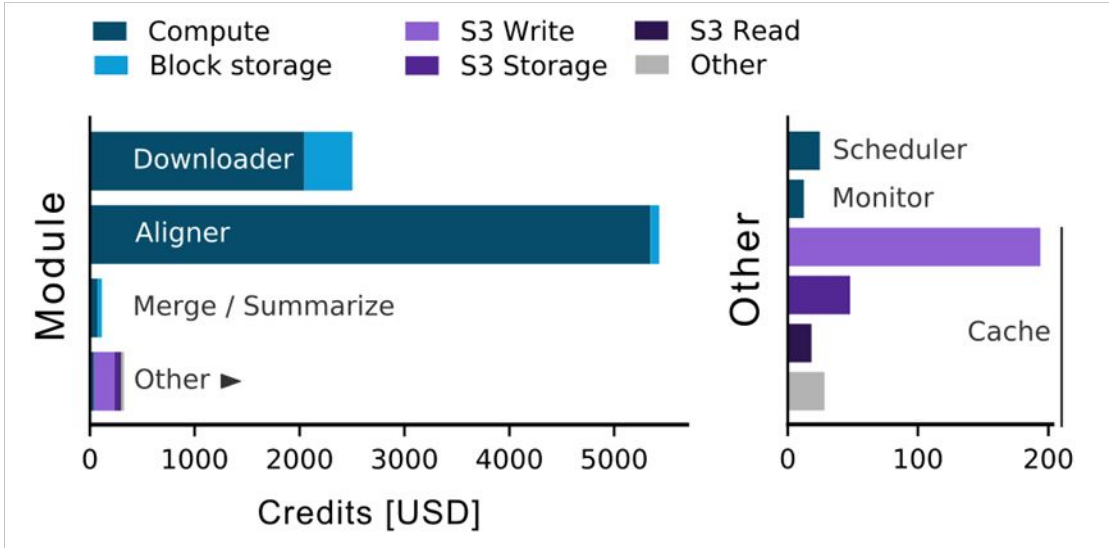
# Serratus architecture




- Aggressively cost-optimized
- Native access to SRA on S3
- Dynamic scaling up to ~22,250s vCPU
- Open Source: GPLv3

# Serratus performance & costs

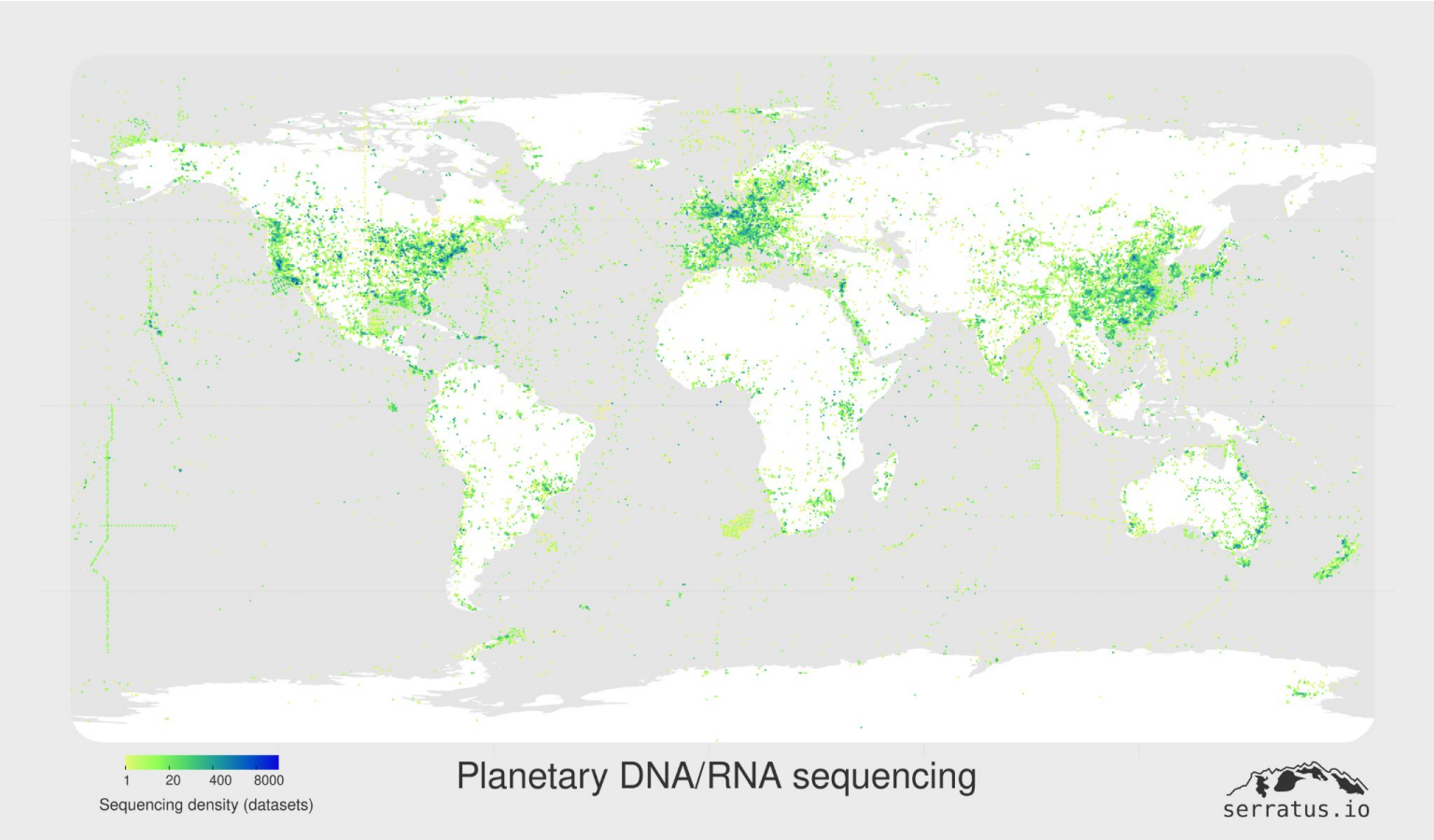


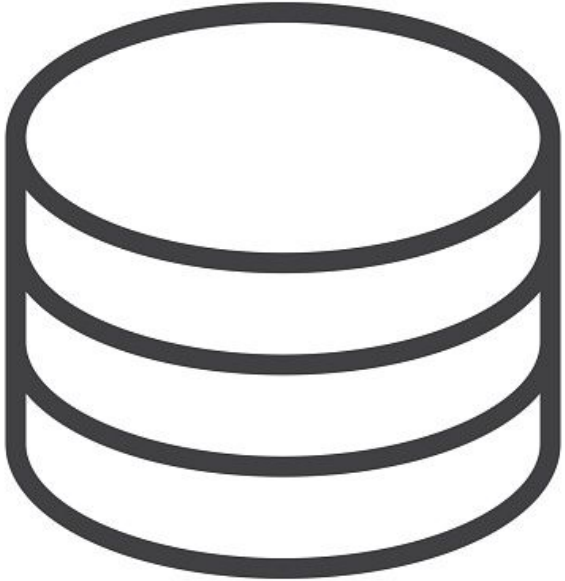

***1 million NGS libraries / day***

***$0.005 / library***

# Geography of SRA samples

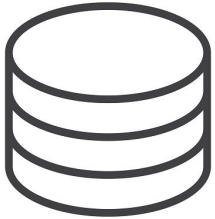


Analysis 2:



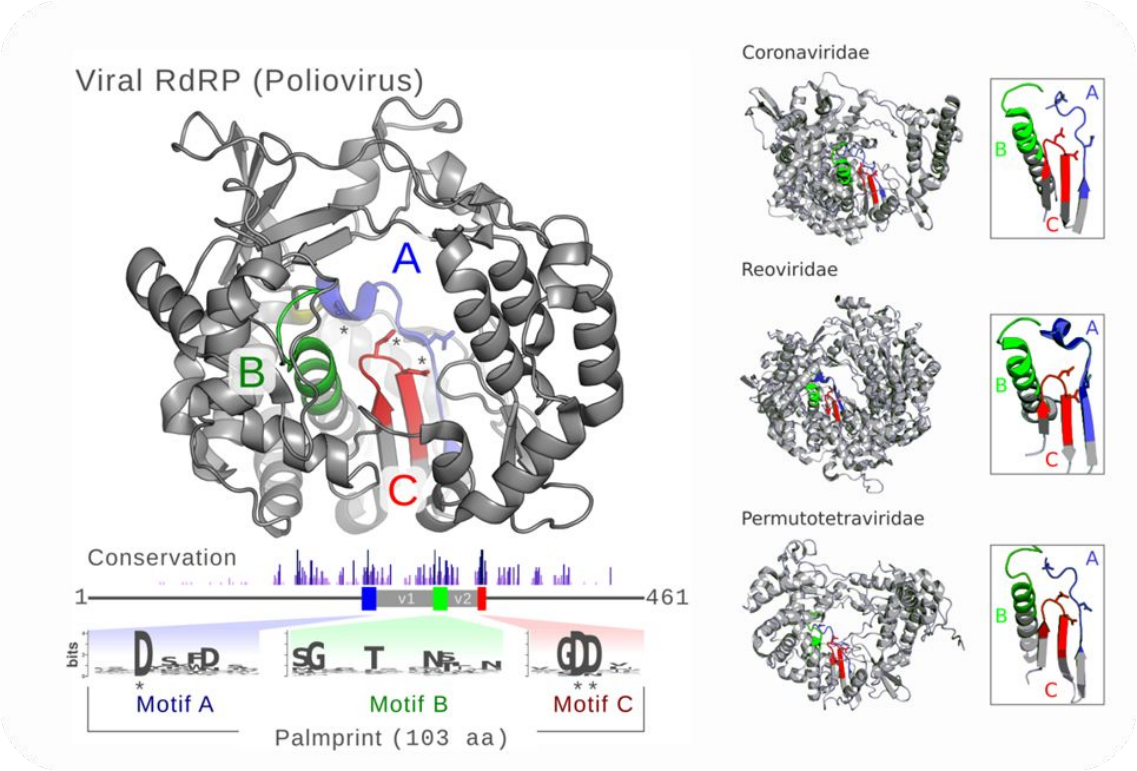
All RNA-seqs

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



**aligned reads (.bam files)**

# Analysis 2, search database: 15,060 known RNA viruses RdRP gene



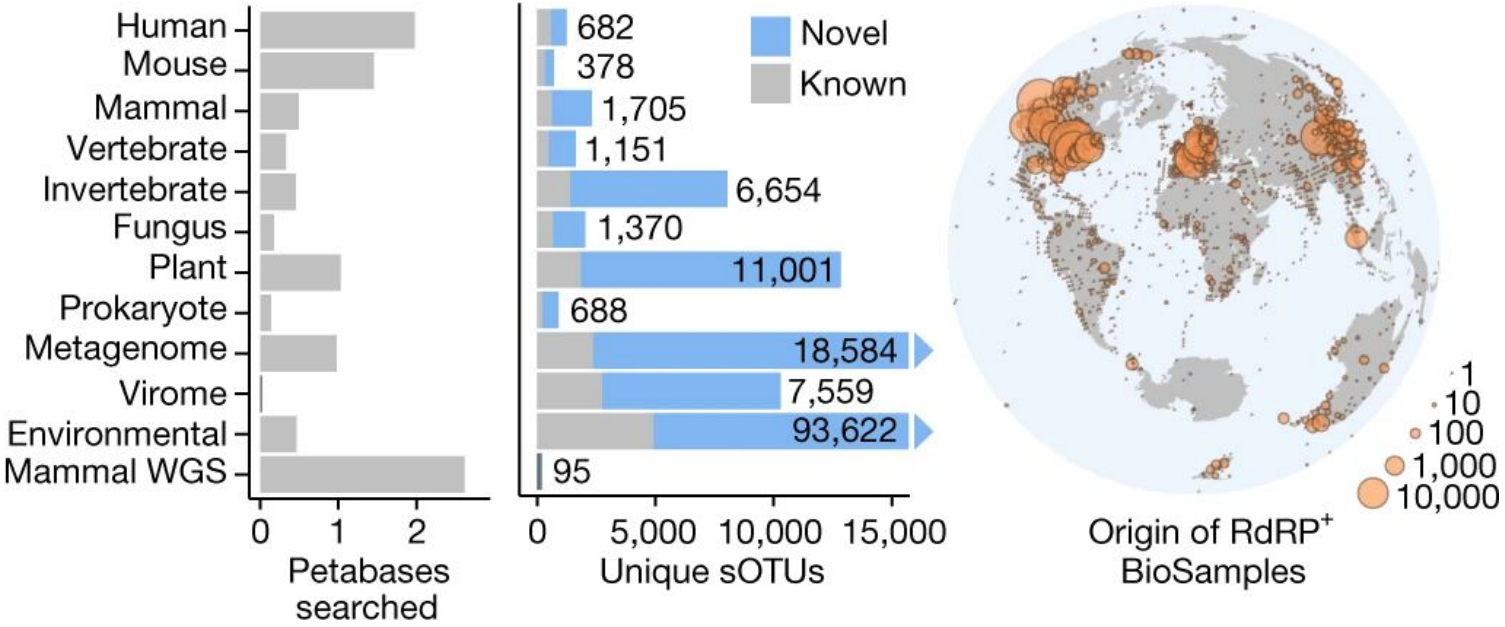

(Babaian & Edgar, 2021. bioRxiv)

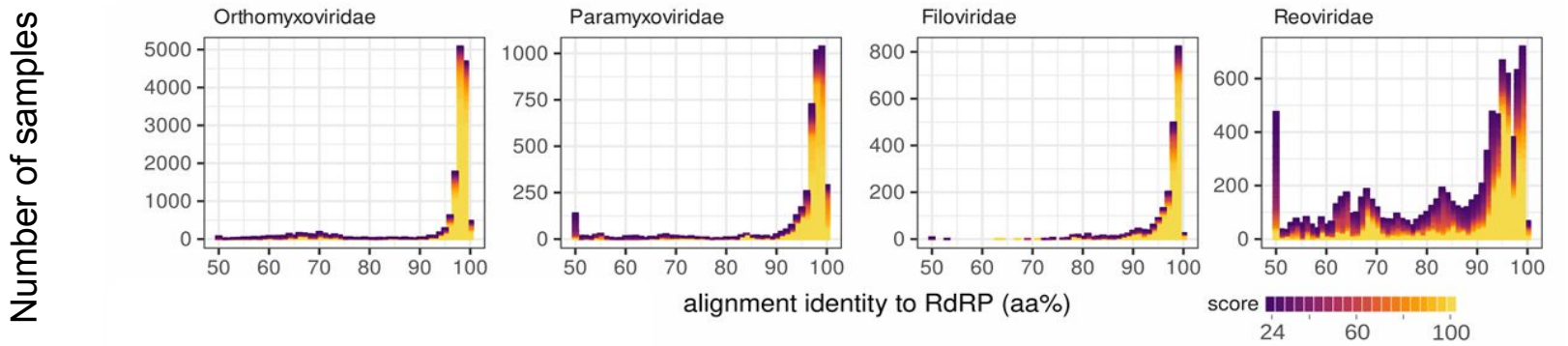
- RNA Virus “Palmprint”
- Species threshold: 90% amino-acid id

# Assembly of all viral RdRPs (Analysis 2)

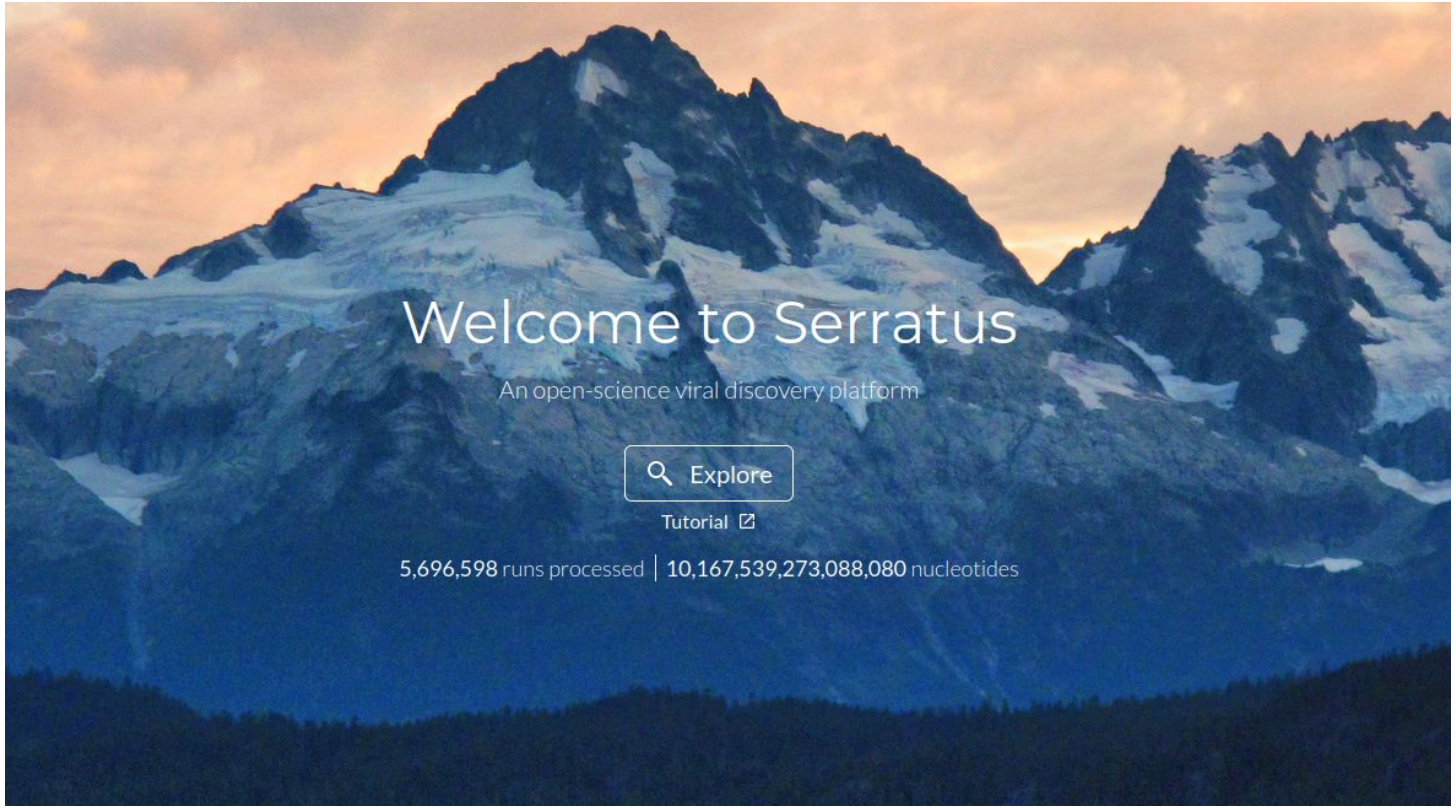
“Micro-assembly” of all RdRp-matching reads within each sample

- SPAdes assembler & GNU parallel

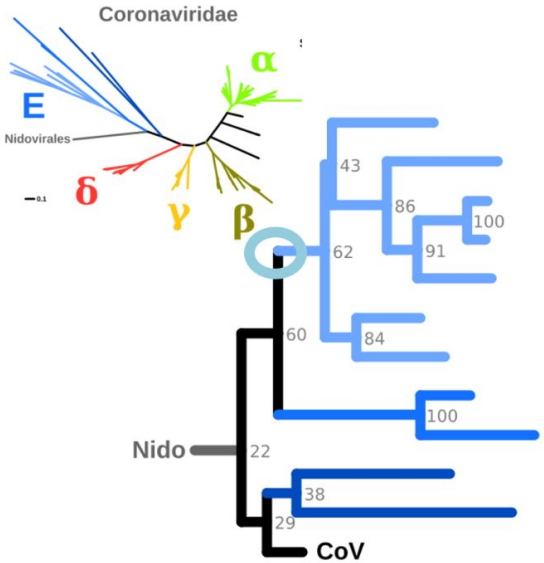

Human
Mouse
Mammal
Vertebrate
Invertebrate
Fungus
Plant
Prokaryote
Metagenome
Virome
Environmental
Mammal WGS
0
1
2
Petabases searched
682
378
1,705
1,151
6,654
1,370
11,001
688
18,584
7,559
93,622
95
Novel
Known
0
5,000
10,000
15,000
Unique sOTUs
1
10
100
1,000
10,000
Origin of RdRP+ BioSamples


Number of samples
Orthomyxoviridae
Paramyxoviridae
Filoviridae
Reoviridae
alignment identity to RdRP (aa%)
score
24
60
100

# "petabase scale" on Google, `www.serratus.io`



# Discovering new Coronaviruses

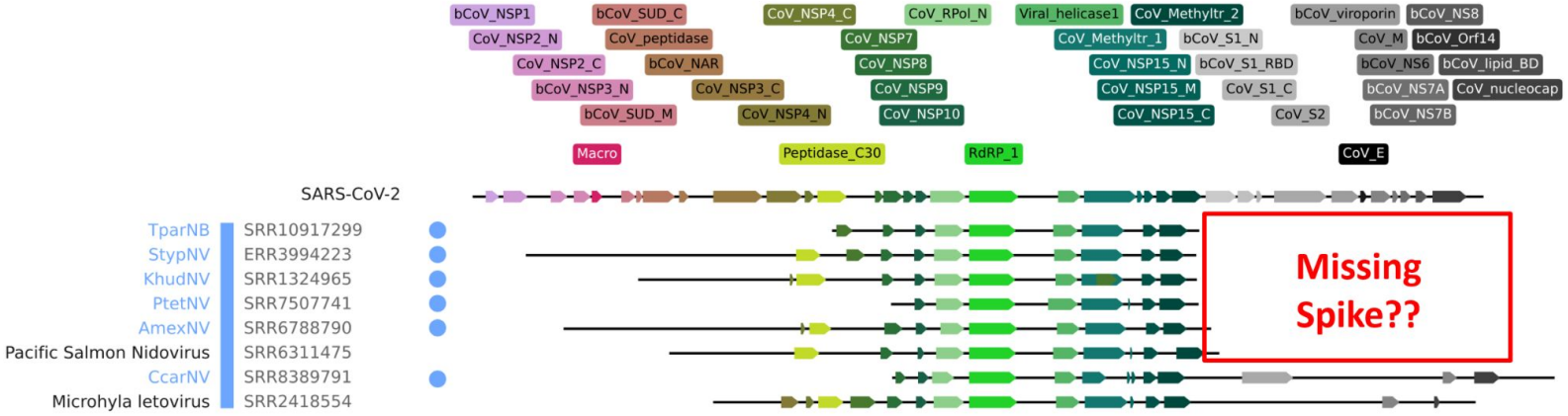



| | |
|---|---|
| AmexNV | SRR6788790 |
| PtetNV | SRR7507741 |
| HkudNV | SRR1324965 |
| StypNV | ERR3994223 |
| TparNV | SRR10917299 |
| Pacific Salmon Nidovirus | |
| AcaNV | SRR5997671 |
| SilNV | SRR12184956 |
| MalbNV | SRR10402291 |
| HtraNV | SRR8389791 |
| Microhyla Letovirus | |

Slides credit: A. Babaian

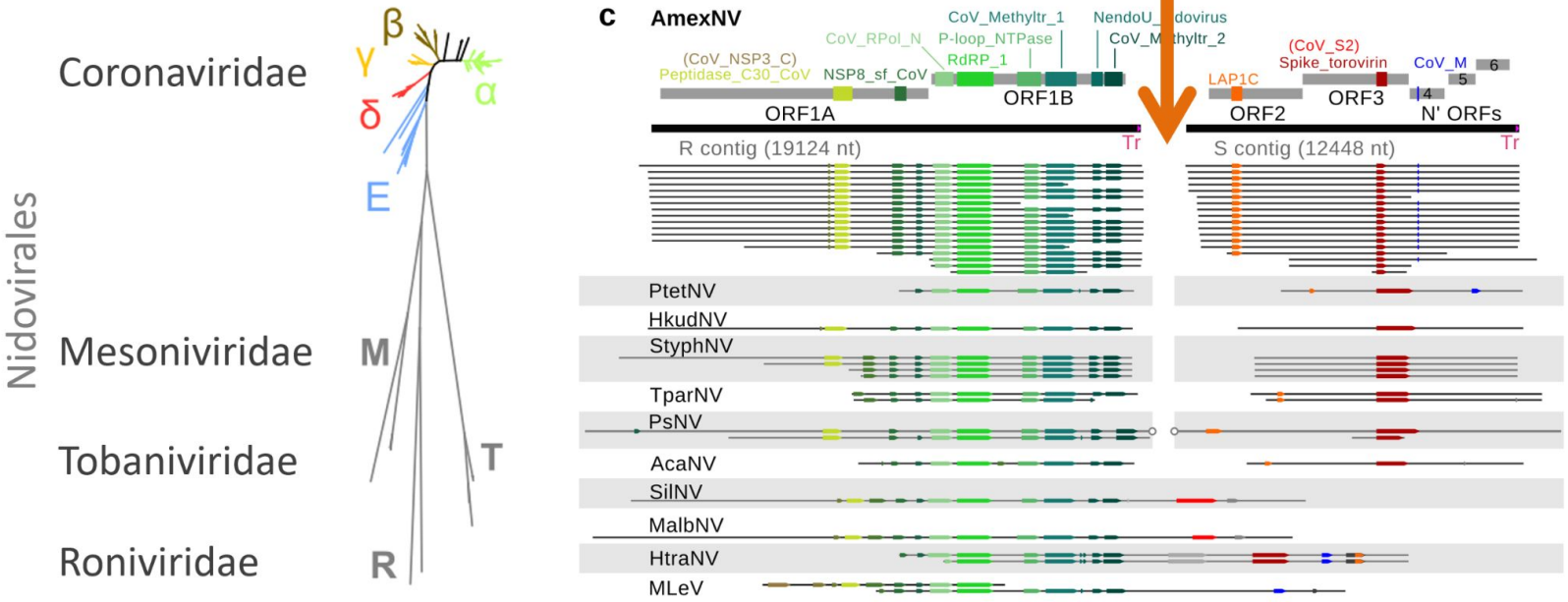
serratus.io

# Discovering new Coronaviruses




serratus.io

# *Segmented Coronaviruses?*




**Re-writing the textbook definition of a Coronavirus**

serratus.io

# Metagenome & metavirome assembly

Usually:

Reconstruct *all* the genomes

Computationnally intensive

Analysis 1:

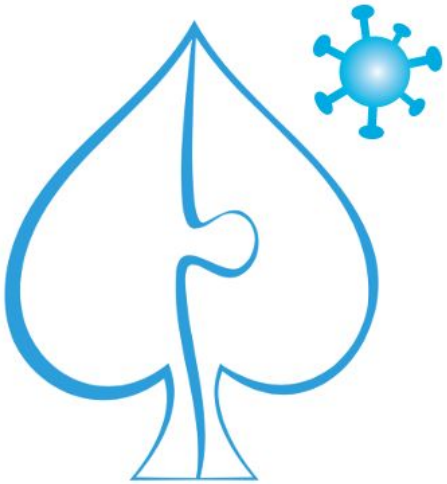
**Reconstruct only CoV genome(s)**

Computationnally intensive

Analysis 2:

**Reconstruct only RdRP genes**

Computationnally easy(/ier) :)



SPAdes assembler

rnaSPAdes

coronaSPAdes

# AWS Batch framework for large-scale assembly

**Peak:**
**~28,000 vCPUs**

| Name | Instance ID | Instance Type | Availability Zone | Instance State |
|---|---|---|---|---|
| Compute | i-004fc86f836336d17 | c5.9xlarge | us-east-2a | running |
| Compute | i-01af64dd577f162b5 | c5.9xlarge | us-east-2a | running |
| | | | us-east-2a | running |
| Compute | i-0879ad68f76a4a54e | c5.9xlarge | us-east-2a | running |
| Compute | i-094ddc9b931fde962 | c5.9xlarge | us-east-2a | running |
| Compute | i-0c8f6d93593531c32 | c5.9xlarge | us-east-2a | running |
| Compute | i-0e08ab6c5a3d0ce3f | c5.9xlarge | us-east-2a | running |
| Compute | i-0ea10648adeeabf68 | c5.9xlarge | us-east-2a | running |

(screenshot: P. Barbera)

AWS Batch > Dashboard

Last updated: 07:11:08 PM. Auto-refreshes every 60 seco

## Dashboard

### Jobs overview

| RUNNABLE | RUNNING | SUCCEEDED | FAILED |
|---|---|---|---|
| 450 | 173 | 48 | 817 |

### Job queue overview

| Job queue | SUBMITTED | PENDING | RUNNABLE | STARTING | RUNNING | SUCCEEDED | FAILED |
|---|---|---|---|---|---|---|---|
| RayanUnitigsBatchProcessingJobQueue | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| RayanSerratusDlBatchProcessingJobQueue | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| RayanSerratusAssemblyBatchJobQueue | 0 | 0 | 450 | 7 | 173 | 48 | 817 |

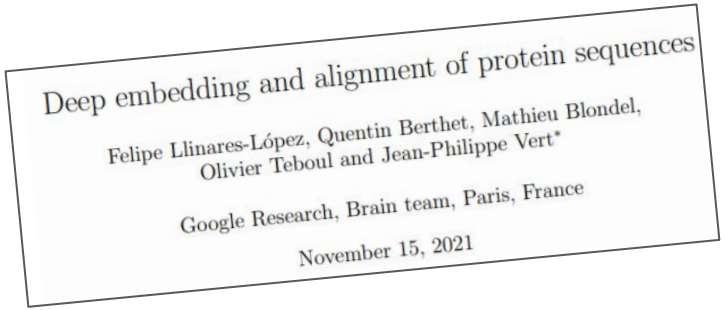
# But, for all-RdRp assembly (Analysis 2)..

With a single “bigger” instance (c6a.48xlarge, 192 cores)

# $10^5$ viral species known, $10^8$ left to discover

## What's next?

- DNA viruses
- Lower homology detection with known RdRPs
  - Replacing Bowtie 2 / Diamond by …?

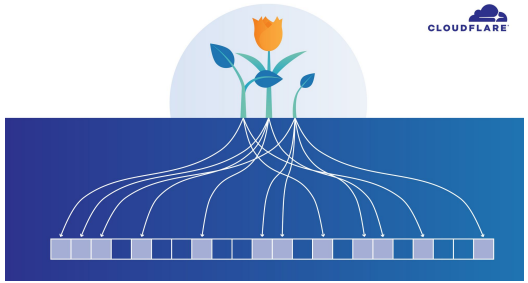

Deep embedding and alignment of protein sequences

Felipe Llinares-López, Quentin Berthet, Mathieu Blondel, Olivier Teboul and Jean-Philippe Vert*

Google Research, Brain team, Paris, France

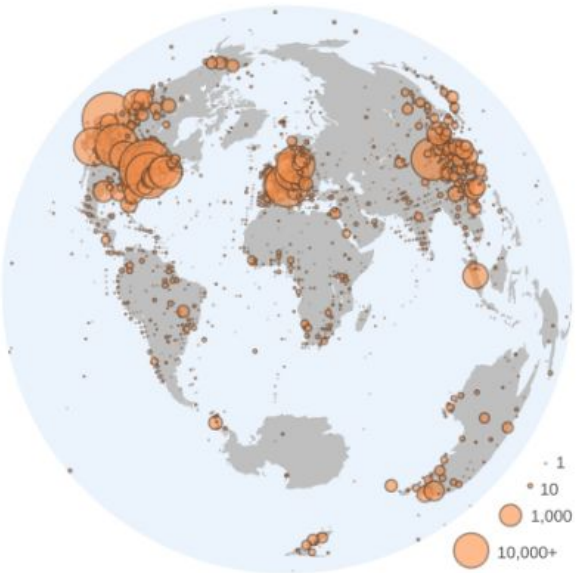
November 15, 2021

- A global **index of the SRA**
  - nearly feasible with k-mers already
  - would only support exact search



# Summary:

- **132,260 novel RNA virus species**
- **1 new group of CoV-like segmented virus**
- **hyper-compressed** (300-500 nt) Zetaviruses
  53 novel deltaviruses (cancer),
  252 huge phages, ..




All our data is accessible:
https://github.com/ababaian/serratus/wiki/Access-Data-Release
7 TB of alignments and assemblies

# More details:

https://www.nature.com/articles/s41586-021-04332-2

https://github.com/ababaian/serratus/

Chat with us on Slack:

https://join.slack.com/t/hackseq-rna/shared_invite/zt-ewlzh9qf-SiNkxvvTJflcutFN0h5jIQ

nature



Article | Published: 26 January 2022

## Petabase-scale sequence alignment catalyses viral discovery

Robert C. Edgar, Jeff Taylor, Victor Lin, Tomer Altman, Pierre Barbera, Dmitry Meleshko, Dan Lohr, Gherman Novakovsky, Benjamin Buchfink, Basem Al-Shayeb, Jillian F. Banfield, Marcos de la Peña, Anton Korobeynikov, Rayan Chikhi & Artem Babaian

*Nature* **602**, 142–147 (2022) | Cite this article



### Abstract

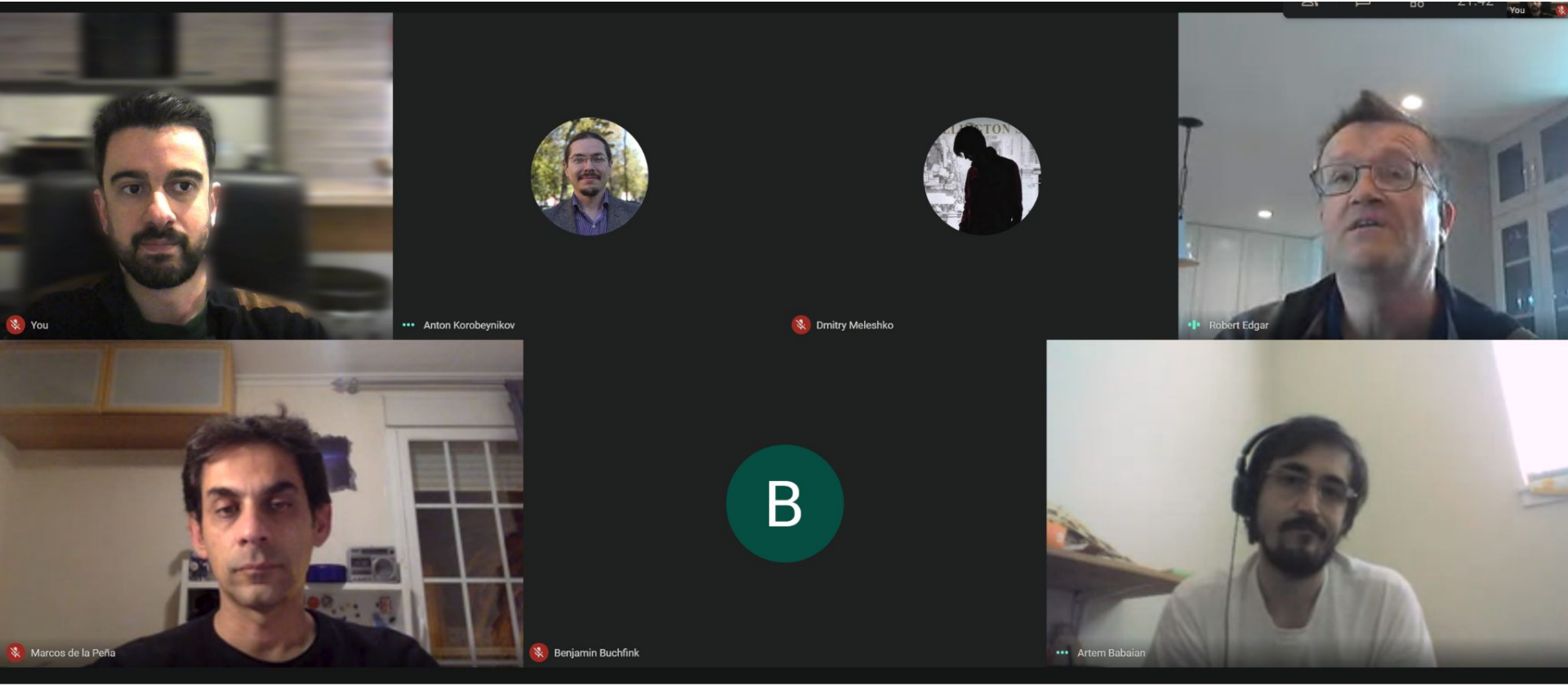
Public databases contain a planetary collection of nucleic acid sequences, but their systematic exploration has been inhibited by a lack of efficient methods for searching this corpus, which (at the time of writing) exceeds 20 petabases and is growing exponentially[1]. Here we developed a cloud computing infrastructure, Serratus, to enable ultra-high-throughput sequence alignment at the petabase scale. We searched 5.7 million biologically diverse samples (10.2 petabases) for the hallmark gene RNA-dependent RNA polymerase and identified well over $10^5$ novel RNA viruses, thereby expanding the number of known species by roughly an order of magnitude. We characterized novel viruses related to coronaviruses, hepatitis delta virus and huge phages, respectively, and analysed their environmental reservoirs. To catalyse the ongoing revolution of viral discovery, we established a free and comprehensive database of these data and tools. Expanding the known sequence diversity of viruses can reveal the evolutionary origins of emerging pathogens and improve pathogen surveillance for the anticipation and mitigation of future pandemics.

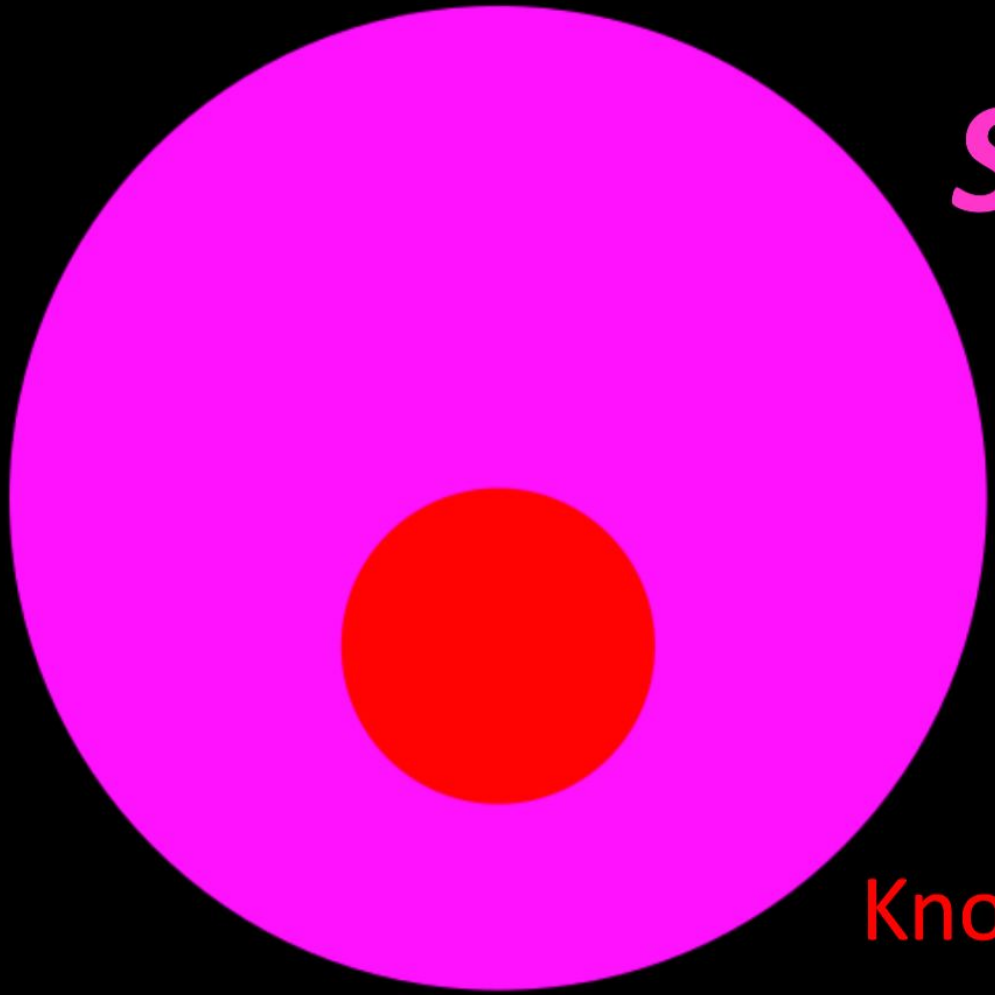

## Digital Collaboration

- Anton Korobeynikov (St. Petersburg)
- Artem Babaian (Vancouver)
- Basem Al-Shayeb (Berkeley)
- Benjamin Buchfink (Tubingen)
- Dan Lohr (Boulder)
- Dmitry Meleshko (Ithaca)
- Gherman Novakovsky (Vancouver)
- Jeff Taylor (Vancouver)
- Jillian F. Banfield (Berkeley)
- Marcos de la Pena (Valencia)
- Pierre Barbera (Heidelberg)
- Rayan Chikhi (Paris)
- Robert C. Edgar (Sonoma)
- Tomer Altman (San Francisco)
- Victor Lin (Gainsville)

# All equal contributions

# We never met IRL





*Serratus*

Known RNA Virome

# Earth's Virome

# Summary

- Lots of genomics data
- Many great analyses could be made
- Cloud helps at the largest scale

# Credits

Some of the people who initiate these “small-group but large-scale” analyses:

C. Titus Brown, Ben Langmead, Artem Babaian, Rob Finn, Adam Phillippy, Andre Kahles, Zamin Iqbal, Carl Kingsford, Rob Patro, Christina Boucher, Pierre Peterlongo, Olivier Jaillon, Dominique Lavenier, Antoine Limasset, Camille Marchet, Daniel Gautheret, Thérèse Commes, and many others I forget to mention

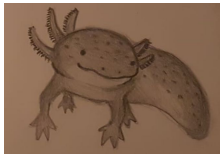
Additional credits:

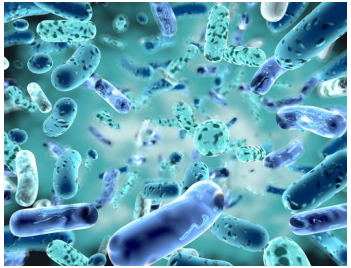
k-mer people

Slides advise: Michel Attafeu, Sophie Shaw, Na, Cami, Karin, M, Malfoy

# Sequence Bioinformatics

@ Institut Pasteur

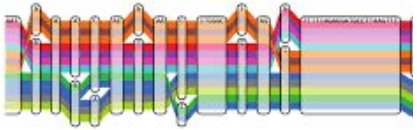




Genomes & metagenomes assembly

Algorithms and data structures on k-mers

Sequence search in very large datasets



Pangenomics

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


Vielen Dank für ihre
Aufmerksamkeit!

# Bonus slides

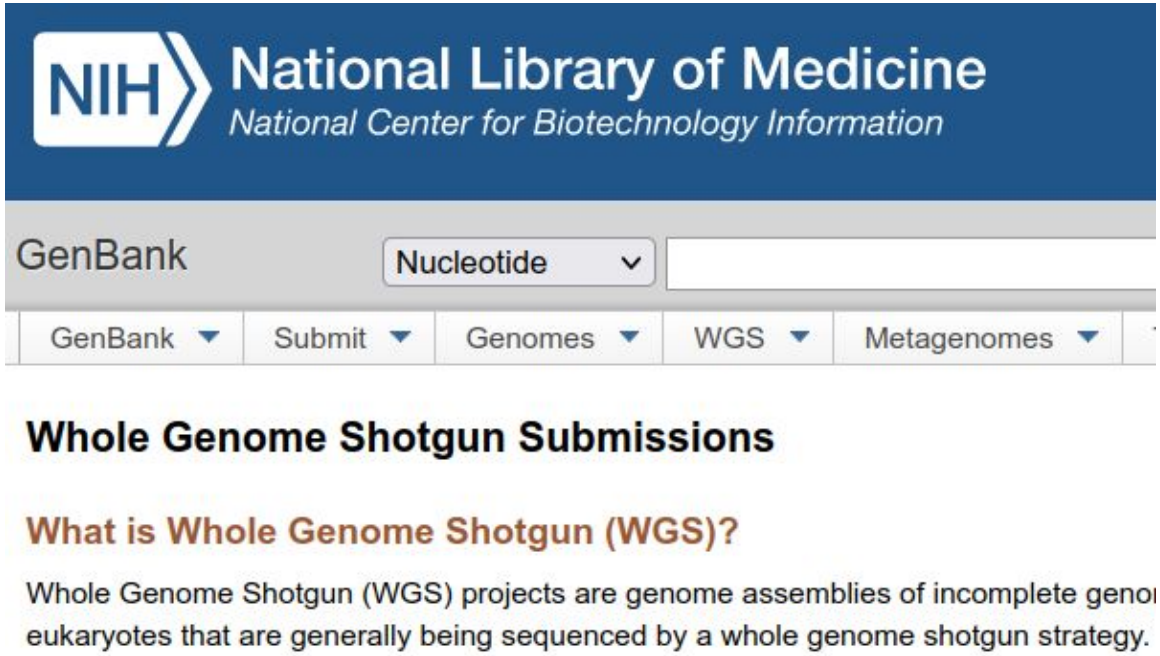
Part 1.5:
*“Spill the beans! Where is this* magical bigger data *you speak of?”*

# GenBank


**Type:** assemblies
**Size:** 1.2 TB (April 2022)
**Diversity:** high

Particularity: all sequences are *annotated*

# NCBI WGS




**Type:** assemblies
**Size:** 16 TB (April 2022)
**Diversity:** high

Difference with GenBank: sequences are not necessarily annotated

# NCBI SRA

SRA | SRA ▾ | Search
Advanced | Help

## 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
3. NextSeq 500 paire
   1 ILLUMINA (Illumina
   Accession: ERX34307

### NextSeq 500 paired end sequencing (ERR3407135)

Metadata | Analysis (alpha) | Reads | Download

Filter: Find | Filtered Download | What does it do?
What can the filter be applied to?

< 1 1 346553 >

View: ☑ biological reads ☐ technical reads

1. ERR3407135.1 ERS3549882
   name: NB551234:144:HL523AFXY:1:11101:5421:
   member: default
2. ERR3407135.2 ERS3549882
   name: NB551234:144:HL523AFXY:1:11101:2248
   member: default
3. ERR3407135.3 ERS3549882
   name: NB551234:144:HL523AFXY:1:11101:2566
   member: default
4. ERR3407135.4 ERS3549882
   name: NB551234:144:HL523AFXY:1:11101:2119
   member: default
5. ERR3407135.5 ERS3549882
   name: NB551234:144:HL523AFXY:1:11101:2350
   member: default

#### Reads (separated)

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

**>gnl|SRA|ERR3407135.1.2** NB551234:144:HL523AFXY:1:11101:5421:1076 *R (Biological)*
ATCAACAACAGCGGGAATACCACCTCTTCCAGCCGTTGTTTCCAACCAATACGCGTTAAT
TCACCGAAACCGCGACAGCGCAATGGAACGCATCATTGCGCAGGTGTTGCAGAATACGGA
AAACCGCATCCGAAACGAGATGCGCGTTAAT

# What is STAT good for?

- Say you have a model organism
  - Search for all sequencing data containing that organism
  - Find host-associations
  - Find co-occurrences with other species

- Say you have a set of samples
  - Determine set of species in them
  - Find other similar samples

- etc..

# NCBI STAT

A taxonomic index of all sequencing data

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

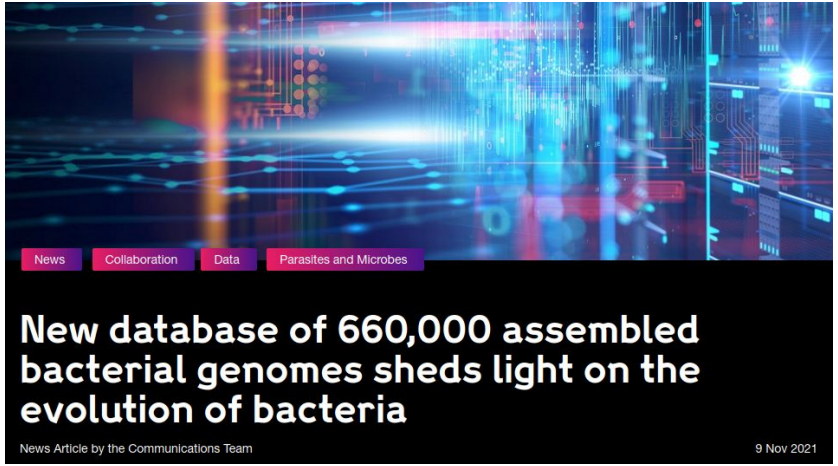
*Genome Biology* **22**, Article number: 270 (2021) | Cite this article

**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%**
    - Acidobacteria: **< 0.01%**
  - Eukaryota: **1.94%**

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

# Blackwell, .., Iqbal's 661k bacterial genomes collection




**Type:** assemblies
**Size:** 2.5 TB
**Diversity:** medium
**dBG?** yes

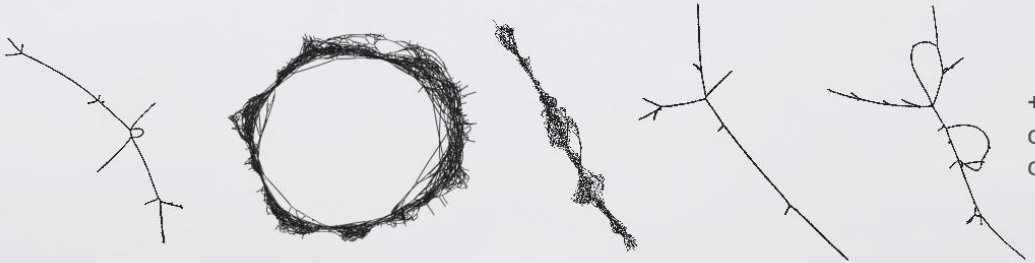
# Results: Pangenome graph of 661,405 bacterial genomes

Data from Blackwell et al, 2021:

```
2.9T 661k_assemblies.fa
1.6T 661k_assemblies.fa.lz4
```

```
rust-mdbg -k 10 -l 12 --density 0.001 --minabund 1 661k_assemblies.fa.lz4
```

Largest 5 connected components:



+ 725,820 connected components

| | | | | | |
|---|---|---|---|---|---|
| Taxons in component | 18 | 22 | 4 | 22 | 10 |
| Dominant species | *Mycobacterium tuberculosis* | *Salmonella enterica* | *Burkholderia gladioli* | *Pseudomonas protegens* | *Cupriavidus alkaliphilus* |

# Many others (often metagenomic)

**Rayan Chikhi**
@RayanChikhi

In this thread we are releasing a concatenated FASTA file of all assemblies produced by Serratus: 59,256 SRA accessions, 5.9 terabases total.

**Uros** @uki156 · Mar 22
Replying to @RayanChikhi
When you said "in this thread we are releasing", I was hoping you were actually going to tweet out the entire thing

Resource | Open Access | Published: 20 July 2020

## A unified catalog of 204,938 reference genomes from the human gut microbiome

Alexandre Almeida, Stephen Nayfach, Miguel Boland, Francesco Strozzi, Martin Beracochea, Zhou Jason Shi, Katherine S. Pollard, Ekaterina Sakharova, Donovan H. Parks, Philip Hugenholtz, Nicola Segata, Nikos C. Kyrpides & Robert D. Finn

MGNify: a database of assemblies of metagenome studies from ENA searchable by metadata

EMBL-EBI | MGnify

**MGnify**

Anton Korobeynikov
23:42 Hier
Lots of stuff in MGnify: https://ebi-metagenomics.github.io/blog/

Submit, analyse, discover and compare microbiome data

Search MGnify

Example searches: Tara oceans, MGYS00000410, Human Gut

Overview | Submit data | Text search | Sequence search | Browse data

### Search by

**Text search**
Name, biome, or keyword

**Sequence search**
Sequence search

### Or by data type

**Analysis types**

| | |
|---|---|
| 356039 | amplicon |
| 28873 | assemblies |
| 2039 | metabarcoding |
| 33827 | metagenomes |
| 2205 | metatranscriptomics |

**Public data**

| | |
|---|---|
| 8696 | studies |
| 661121 | samples |
| 444172 | analyses |
| 9421 | genomes in 4 MAG catalogues |