

*We're gonna need a bigger instance*



*Rayan Chikhi, Institut Pasteur*

# Your instructor is..

- PI in bioinformatics algorithms
- Before that, permanent researcher @ CNRS
- Before that, PhD+postdoc in bioinformatics @ ENS Rennes, Penn State

## My research:

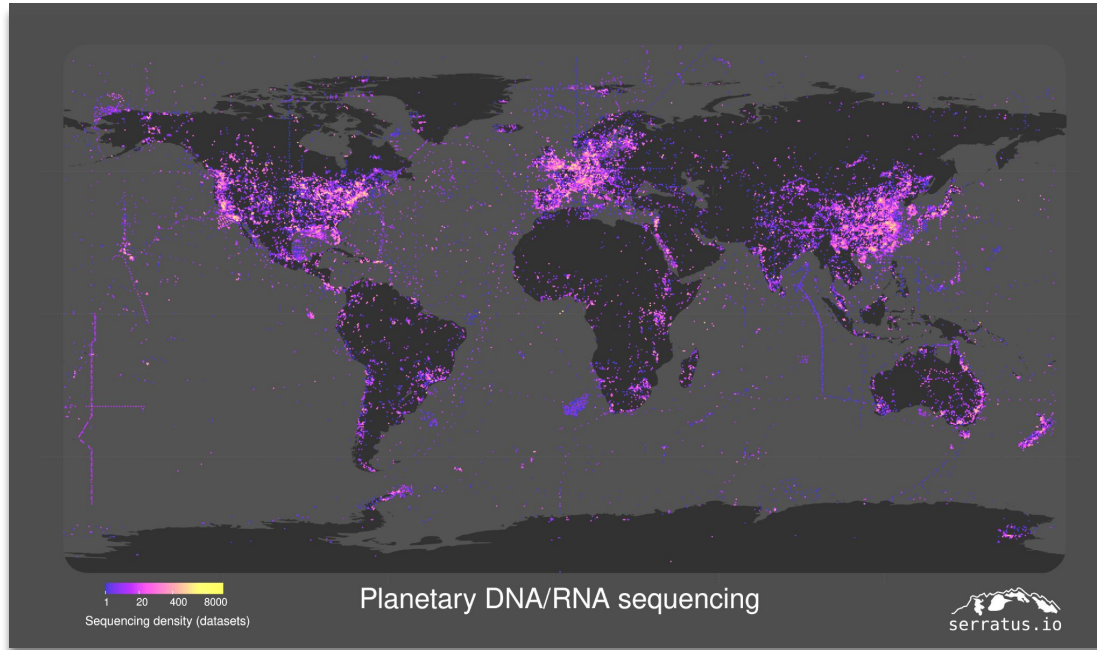
- *de novo* assembly
- k-mers
- metagenomics
- viruses



@RayanChikhi on Twitter 

<http://rayan.chikhi.name>

# “Planet” answer during speed networking

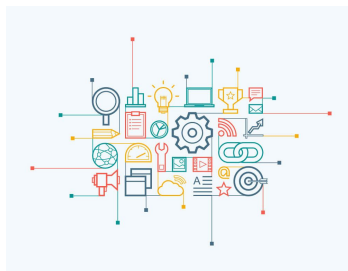


# Sequence Bioinformatics

@ Institut Pasteur



Genomes &  
metagenomes  
assembly



Algorithms and  
data structures  
on k-mers



Sequence  
search in very  
large datasets



Pangenomics



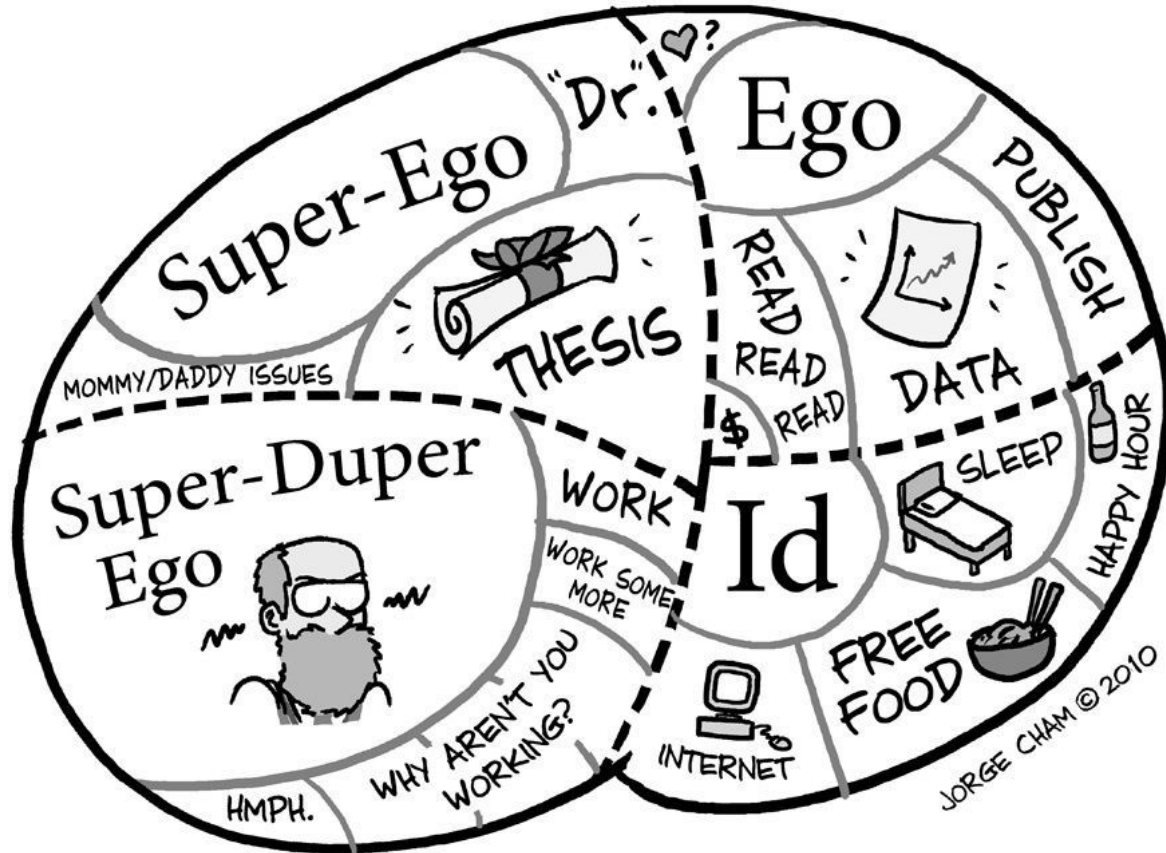
# Today's talk

- *Part 1: For the last week, the Workshop on Genomics has given you access and made you use an infinitely powerful resource during the labs. Perhaps you didn't even notice it. Here I will reveal what it is and how you could harness its power as part of your own research.*
- **Part 2: The Story of Serratus: Petabase-scale alignment for viral discovery**

**Part 1:** For the last week, the Workshop on Genomics has given you access & asked you use an infinitely valuable resource and perhaps you did not even notice it. Here I will reveal what it is and how to harness its power.

I know what you're thinking

(because I've been there)

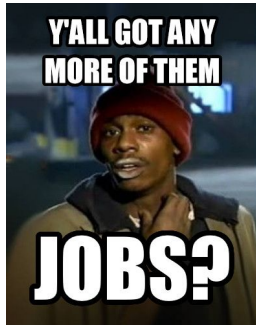


1st year PhD: *“Is my project any good?”*

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

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

Postdoc:



> *No time to learn new things*

WHAT IF I TOLD YOU

This past week you have been using

- infinite\* computation
- &
- infinitely\* fast access to data

\* but, limited by Guy







Download from  
**Dreamstime.com**  
This watermark can be removed for previewing purposes only.

id 83717812  
© Curnypah | Dreamstime.com

And with it, one could  
perform wonderful,  
ground-breaking  
genomics analyses

# Part 1: (Really) Large-scale Genomics

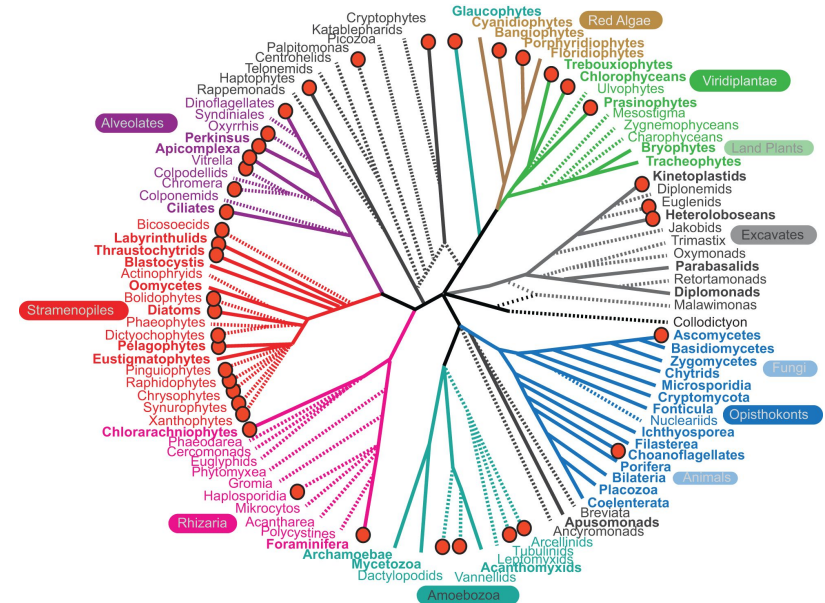
# Part 1.1: Some examples



# Sonya's talk

- MMETSP (Marine Microbial Eukaryote Transcriptome Sequencing Project)
- 650 transcriptomes

*“[..] Transcriptome assembly was carried out using NCGR's BPA1.0 (Batch Parallel Assembly v. 1.0) and BPA2.0 pipelines, as the methods were refined during the 2 year effort [..]”*



# Tara Oceans' analyses

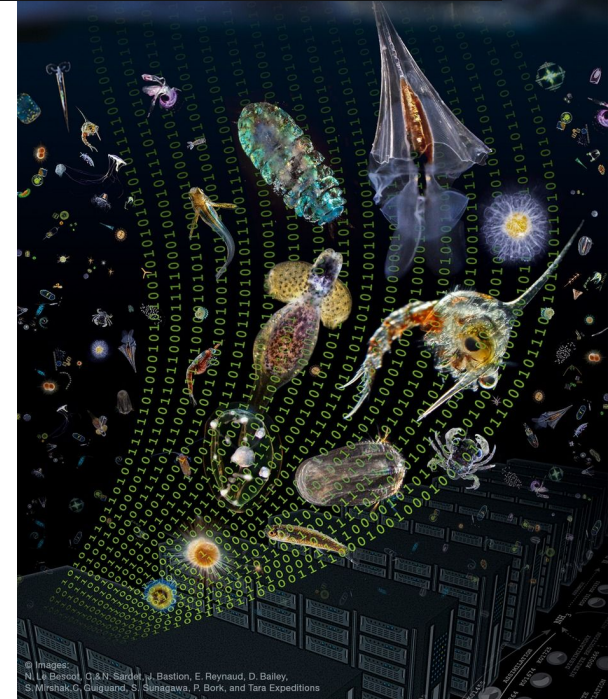


the *Tara Oceans* project (Fig 1), which has arguably been one of the wettest wet laboratory experiments ever.

- > 35,000 samples, sampled from 2009 to 2013

2015:

*“We analyzed 7.2 terabases of metagenomic data from 243 Tara Oceans samples”*



Science

Current Issue First release papers Archive About Submit ma

HOME > SCIENCE > VOL. 348, NO. 6237 > STRUCTURE AND FUNCTION OF THE GLOBAL OCEAN MICROBIOME

SPECIAL ISSUE RESEARCH ARTICLE

## Structure and function of the global ocean microbiome

SHINICHI SUNAGAWA, LUIS PEDRO COELHO, SAMUEL CHAFFRON, JENS ROAT KULTIMA, KARINE LABADIE, GUILLEM SALAZAR, BARDYA DJAHANSCHIRI, GEORG ZELLER, DANIEL R. MENDE, PEER BORK

+42 authors [Authors Info & Affiliations](#)

SCIENCE • 22 May 2015 • Vol 348, Issue 6237 • DOI: 10.1126/science.1261359

© Images: N. Le Besou, G.N. Sardet, Bastion, E. Reynaud, D. Bailey, S. Mirshak, G. Gauguand, S. Sunagawa, P. Bork, and Tara Expeditions



# Tara Oceans' analyses

2018:

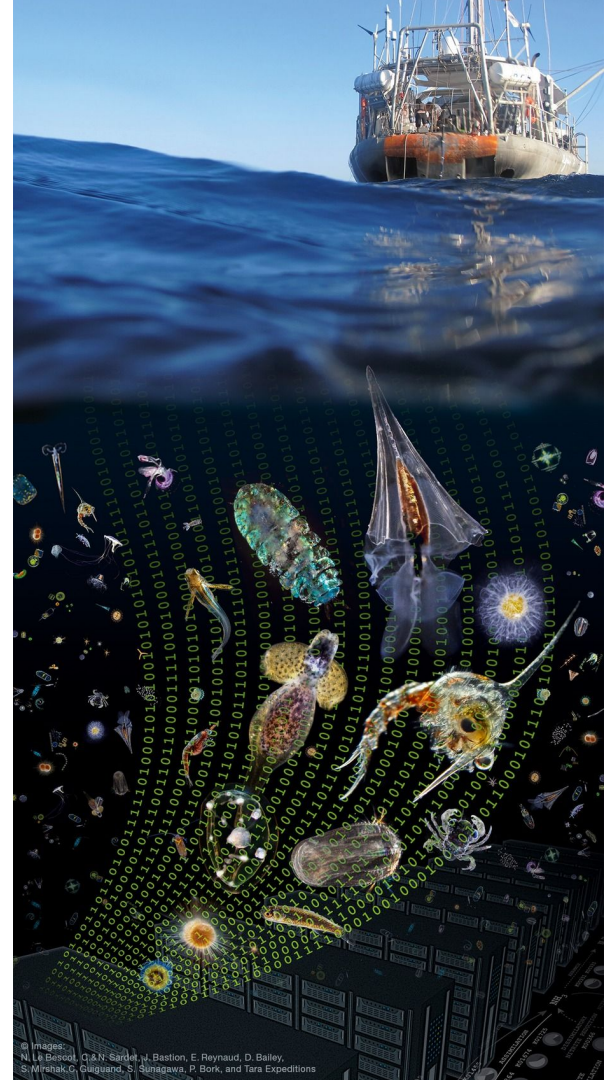
*“441 size-fractionated plankton communities [..], resulting in 16.5 terabases of raw data”*

[Nat Commun](#). 2018; 9: 373.

Published online 2018 Jan 25. doi: [10.1038/s41467-017-02342-1](https://doi.org/10.1038/s41467-017-02342-1)

A global ocean atlas of eukaryotic genes

[Quentin Carradec](#),<sup>#1,2,3</sup> [Eric Pelletier](#),<sup>#1,2,3</sup> [Corinne Da Silva](#),<sup>1</sup> [Adriana Alberti](#),<sup>1</sup> [Yoann Seeleuthner](#),<sup>1,2,3</sup>



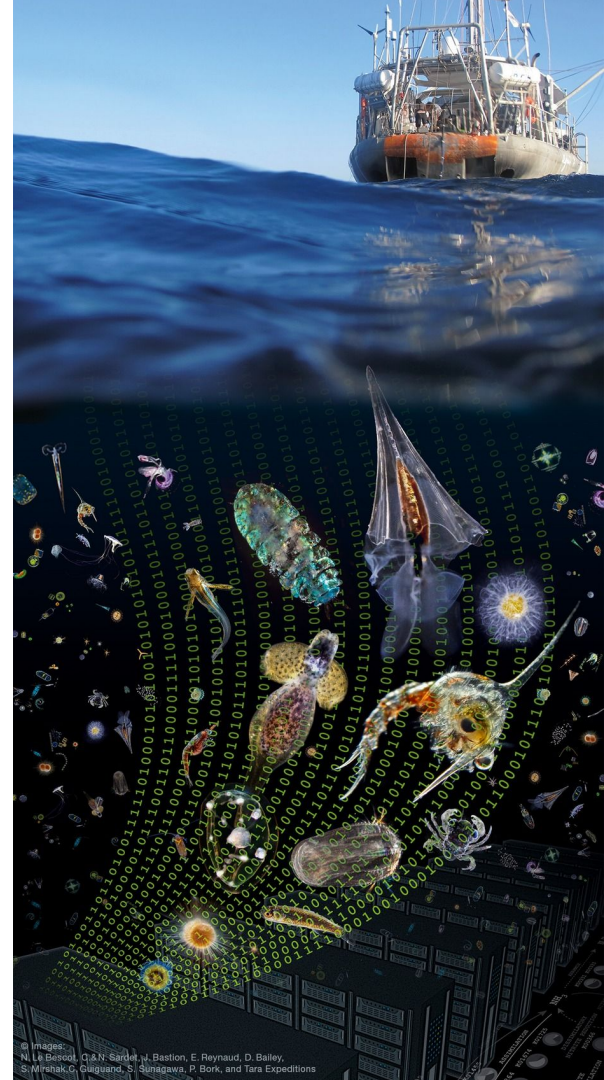
# Tara Oceans' analyses

2022:

*“28 terabases [..] from 771 metatranscriptomes [..]“*

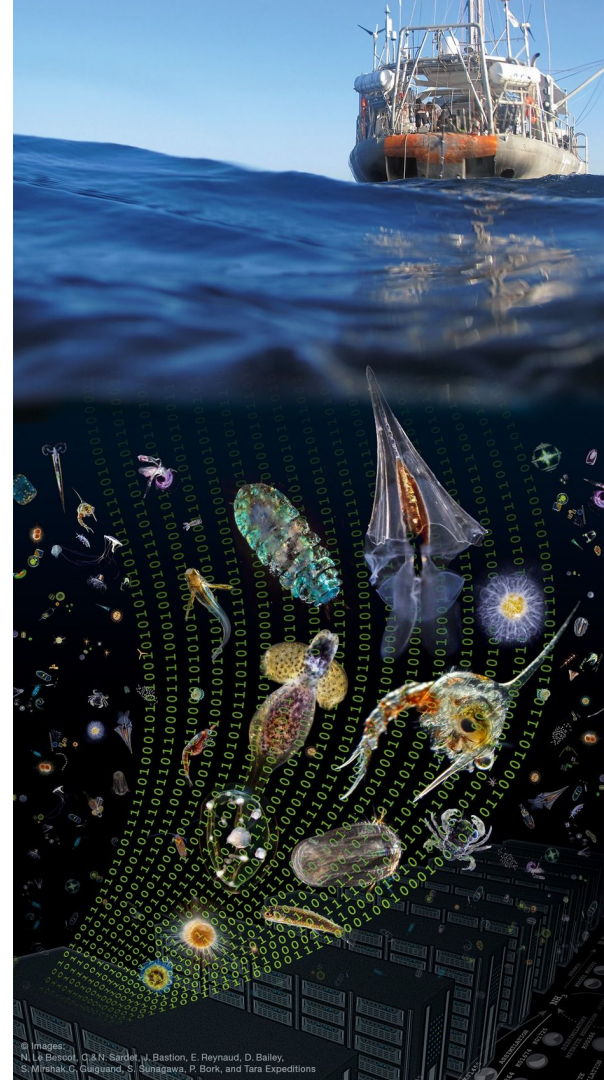


The image shows a screenshot of a Science journal article page. The Science logo is at the top left. Navigation links include 'Current Issue', 'First release papers', 'Archive', and 'About'. A 'Submit manuscript' button is on the right. The article title is 'Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome'. The authors listed are Ahmed A. Zayed, James M. Wainaina, Guillermo Dominguez-Huerta, Eric Pelletier, Jiarong Guo, Mohamed Mohssen, Funing Tian, Akbar Adjie Pratama, Benjamin Bolduc, and Matthew B. Sullivan, with '+23 authors' and a link to 'Authors Info & Affiliations'. The publication date is 7 Apr 2022, and the DOI is 10.1126/science.abm5847.



# Tara Oceans' analyses

- Fantastic science
- Bottleneck is sequencing data analysis?
- They need a bigger instance





# It's not just the oceans



- Main challenge is speed
  - ~2 weeks per assembly on ~320 cores

- 100 tomato genomes, 238,490 structural variants
- *“the most comprehensive panSV genome for a major crop”*
- KLUH story

"WE'VE TAKEN PROCESSES THAT USED TO TAKE HUNDREDS, OR IN SOME CASES EVEN THOUSANDS, OF YEARS, AND PERFORMED THEM VERY RAPIDLY."

—Michael Schatz

Bloomberg Distinguished Associate Professor of computer science and biology

# Humans, too



**bioRxiv**  
THE PREPRINT SERVER FOR BIOLOGY

bioRxiv posts many COVID19-related papers. A reminder: they have not been formally peer-reviewed and should not guide health-related behavior or be reported in the press as conclusive.

Posted March 01, 2022.

New Results

[Follow this preprint](#)

## The sequences of 150,119 genomes in the UK biobank

Bjarni V. Halldorsson, Hannes P. Eggertsson, Kristjan H.S. Moore, Hannes Hauswedell, Ogmundur Eiriksson, Magnus O. Ulfarsson, Gunnar Palsson, Martelinn T. Hardarson, Asmundur Oddsson, Brynjar O. Jenson, Snaedis Kristmundsdottir, Brynja D. Sigurpalsdottir, Olafur A. Stefansson, Doruk Beyter, Guillaume Holley, Vinicius Tragante, Arnaldur Gylfason, Pall I. Olason, Florian Zink, Margret Asgeirsdottir, Sverrir T. Sverrisson, Brynjar Sigurdsson, Sigurjon A. Gudjonsson, Gunnar T. Sigurdsson, Gisli H. Halldorsson, Gardar Sveinbjornsson, Kristjan Norland, Unnur Styrkarsdottir, Droplaug N. Magnusdottir, Steinunn Snorraddottir, Karl Kristinsson, Emilia Sobech, Helgi Jonsson, Arni J. Geirsson, Isleifur Olafsson, Palmi Jonsson, Ole Birger Pedersen, Christian Erikstrup, Søren Brunak, Sisse Rye Ostrowski, DBDS Genetic Consortium, Gudmar Thorleifsson, Frosti Jonsson, Pall Melsted, Ingileif Jonsdottir, Thorunn Rafnar, Hilma Holm, Hreinn Stefansson, Jona Saemundsdottir, Daniel F. Gudbjartsson, Olafur T. Magnusson, Gisli Masson, Unnur Thorsteinsdottir, Agnar Helgason, Hakon Jonsson, Patrick Sulem, Karl Stefansson

doi: <https://doi.org/10.1101/2021.11.16.468246>

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

- 585 million SNPs (7.0% of all possible human SNPs)
- 58 million indels, 900k SVs, microsatellites
- associations for rare variants with large effects



# Guy doesn't have a GPU

- Nanopore basecalling takes 2 weeks

These computation problems aren't limited to big projects. They apply even if your lab is "small".



Any other examples from the audience?



Part 1.2:  
Can you really  
analyze  
*everything?*



# 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]  $10^1 = 10$

# To Petabytes and beyond

<https://academic.oup.com/nar/article/48/10/5217/5825624>

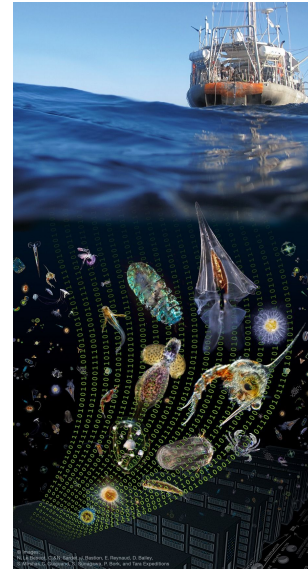
HiSeq 2500  
rapid run:  
300 gigabases



Your laptop: 1 terabyte



Tara Oceans DNA and  
RNA: 60 terabytes



<https://www.ncbi.nlm.nih.gov/pubmed/25228796.pdf>

30  
petabytes

All  
public  
DNA  
sequencing  
data

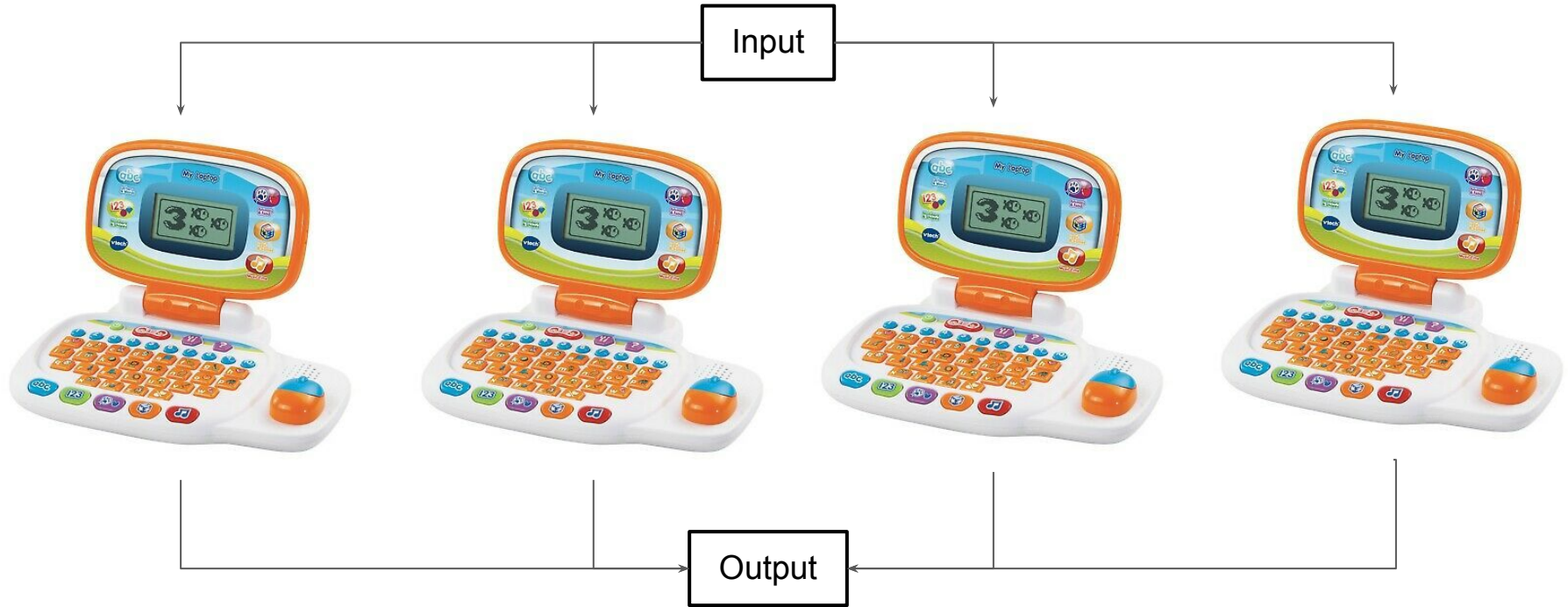


# Parallelism

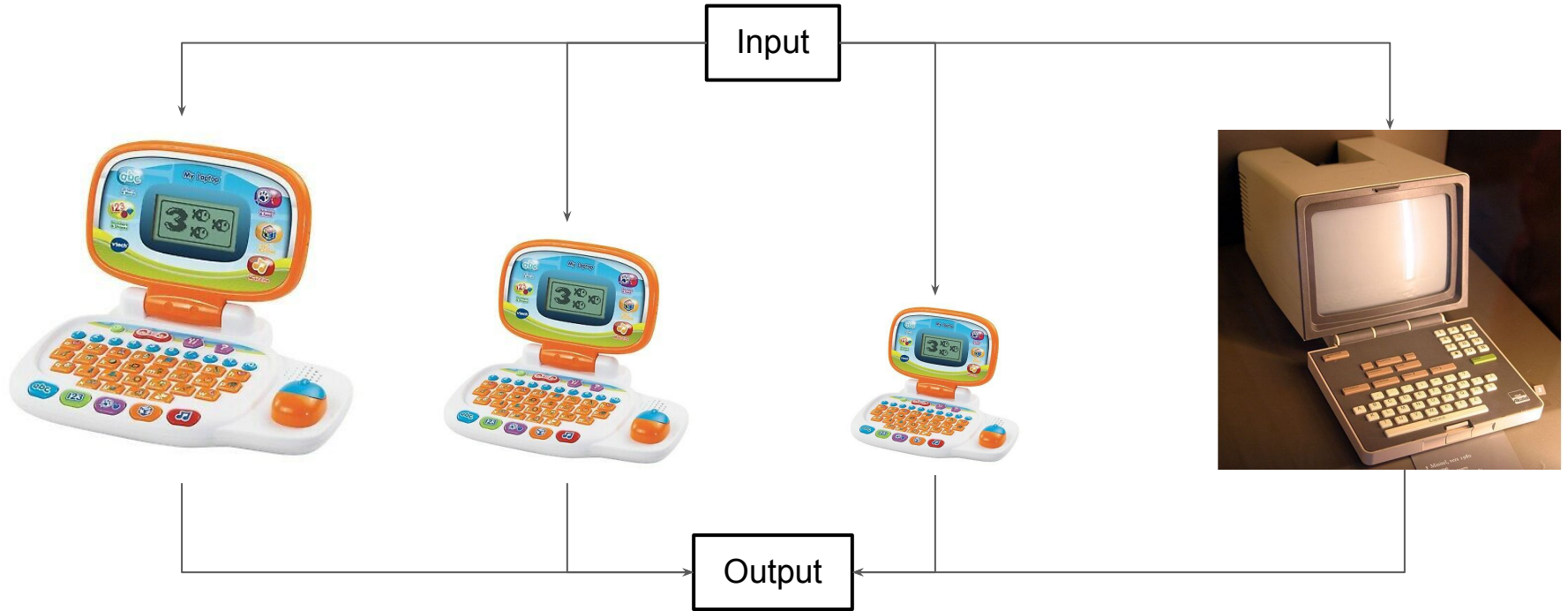
Rationale: one computer is never enough



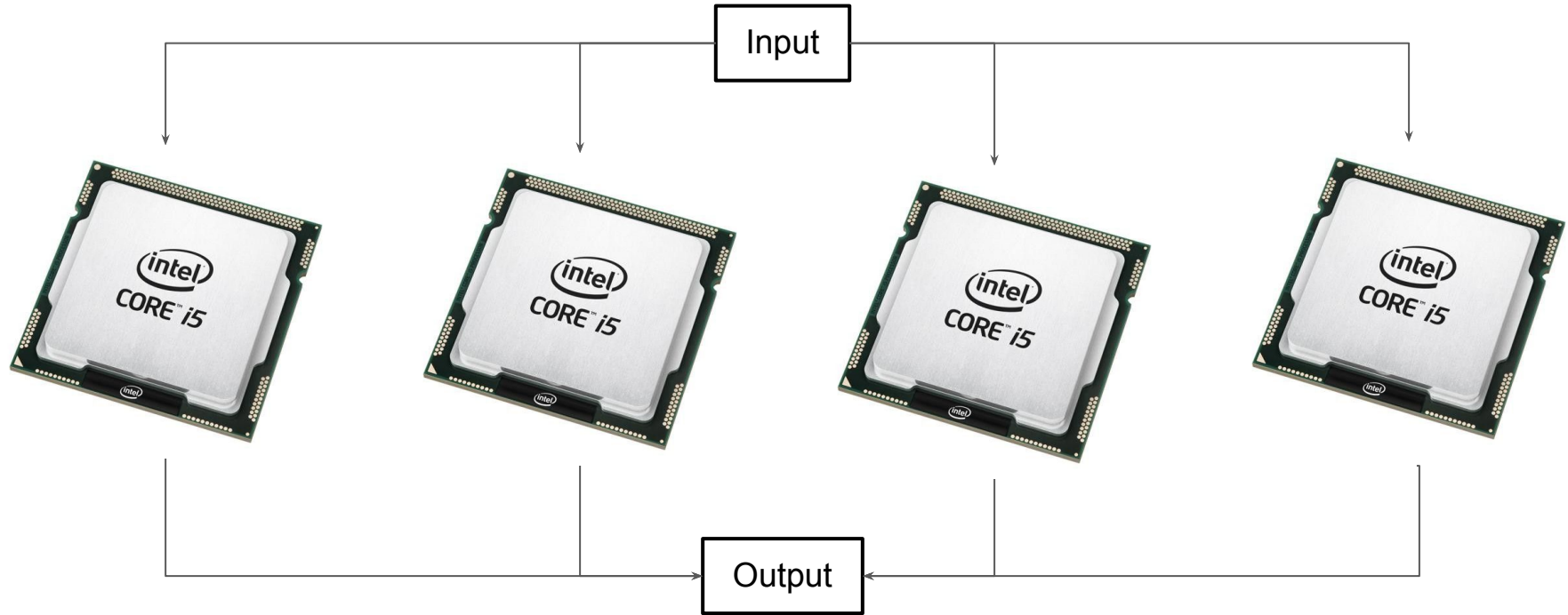
Parallelism: use many “computers” to execute one task



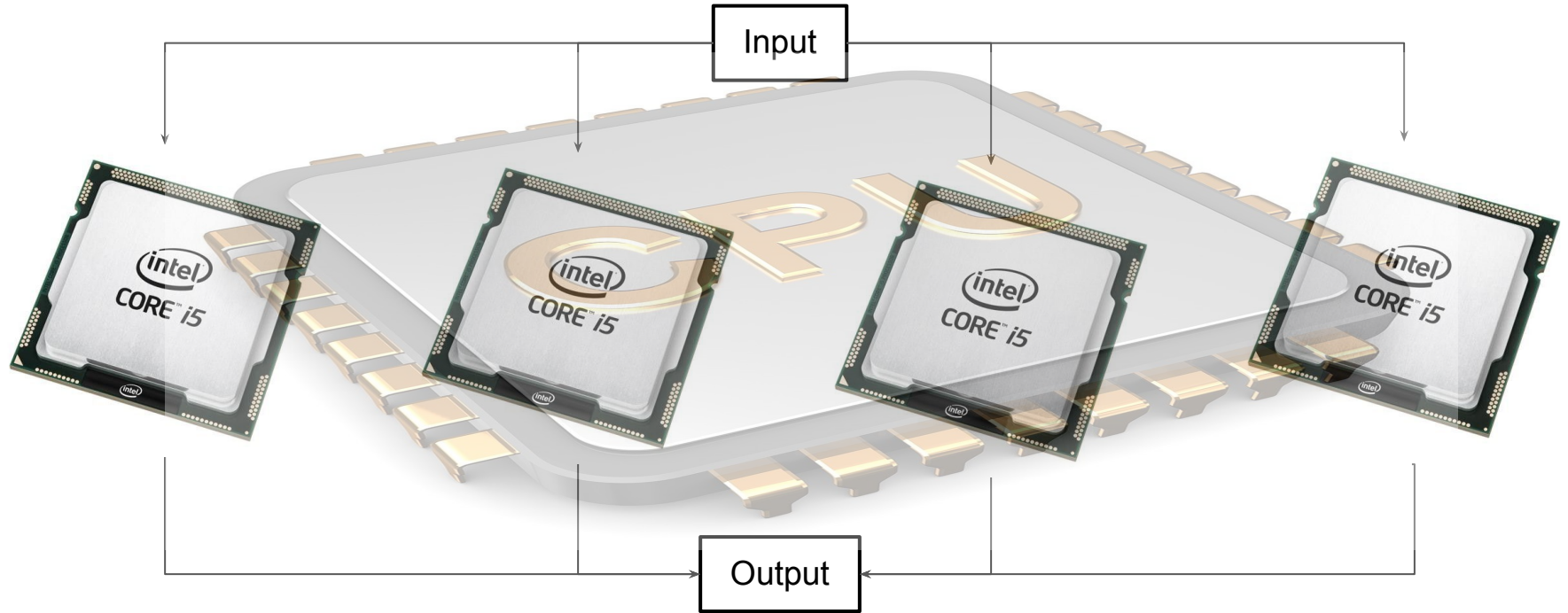
# Parallelism: they don't need to be identical computers



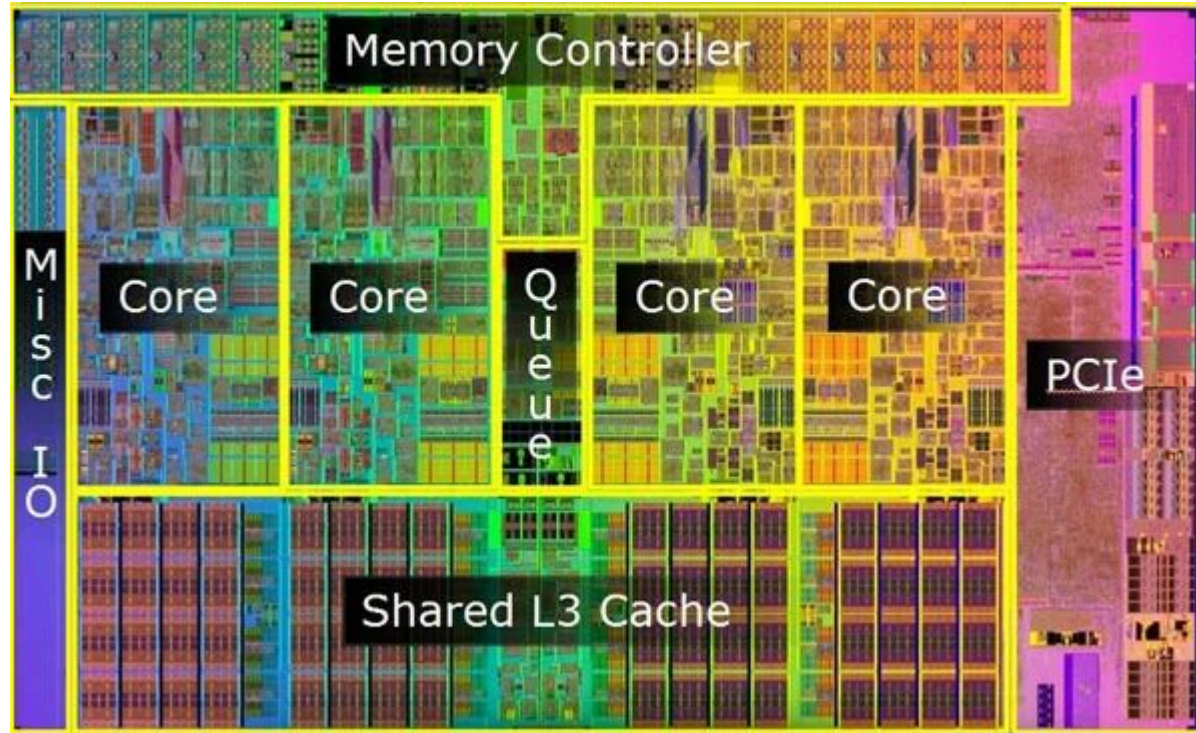
Parallelism: they don't even need to be “computers”



Parallelism: they don't even need to be “computers”

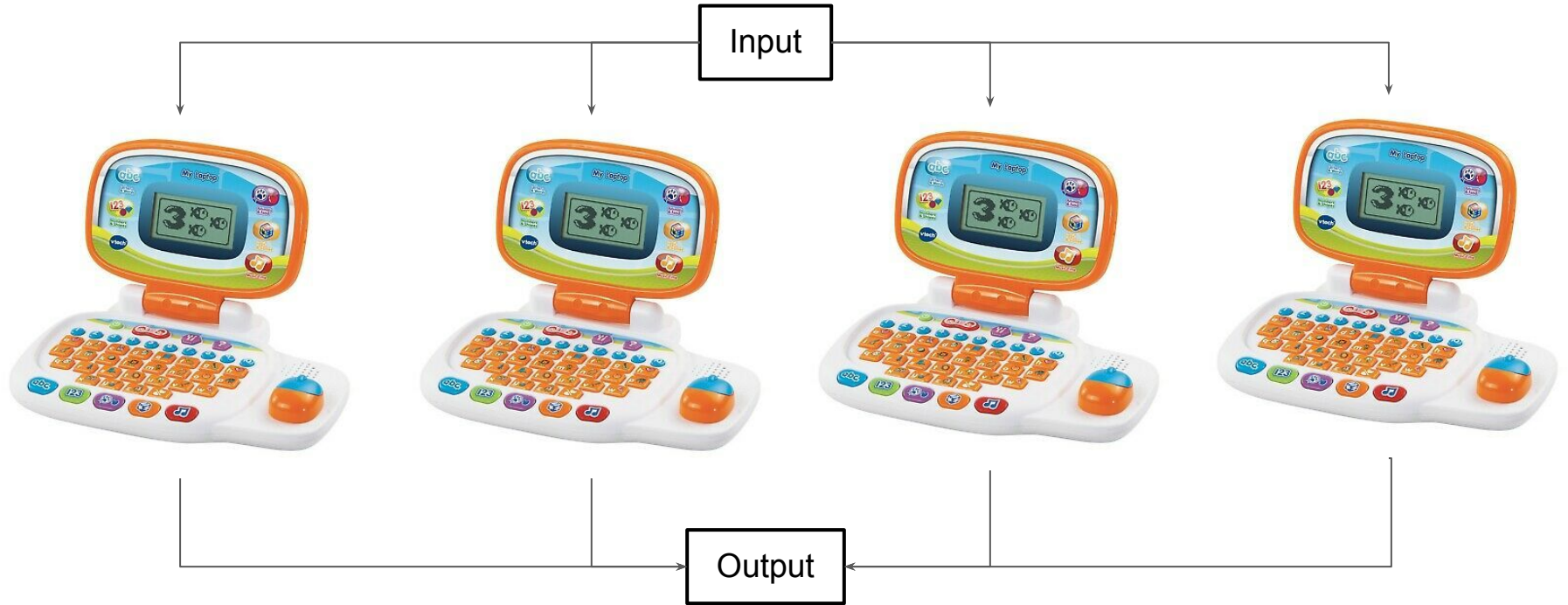


Parallelism: CPU = many little computers in parallel





# CPU (simplified)



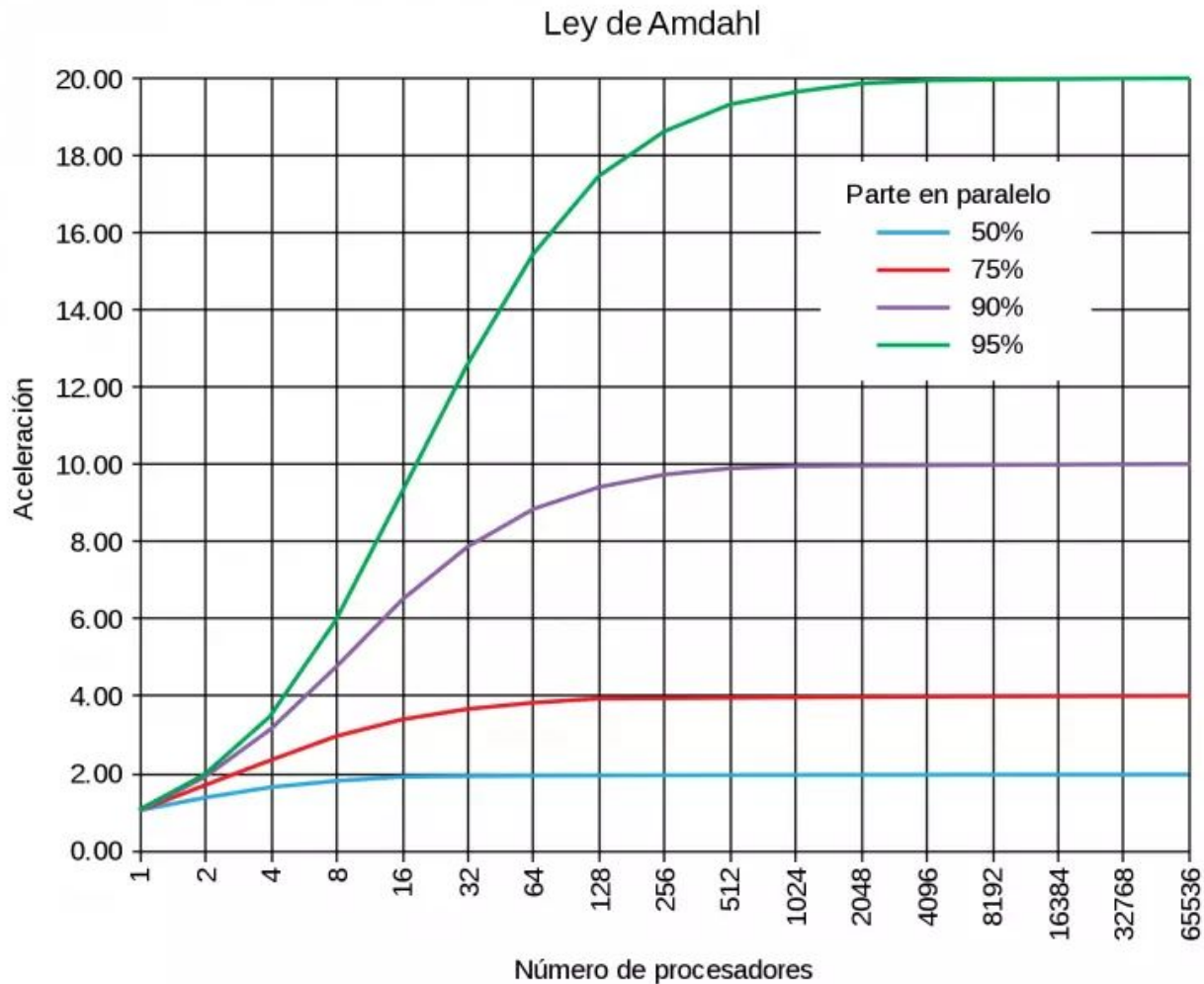


# The limits of computing

So, can we speed up indefinitely by stacking computers (or CPUs)?



# Amdahl's law



# Connect the dots from left to right

Read a small file from disk •

Access data in memory •

Open a web page from Australia •

Align 1 million reads •



• 100 nanoseconds

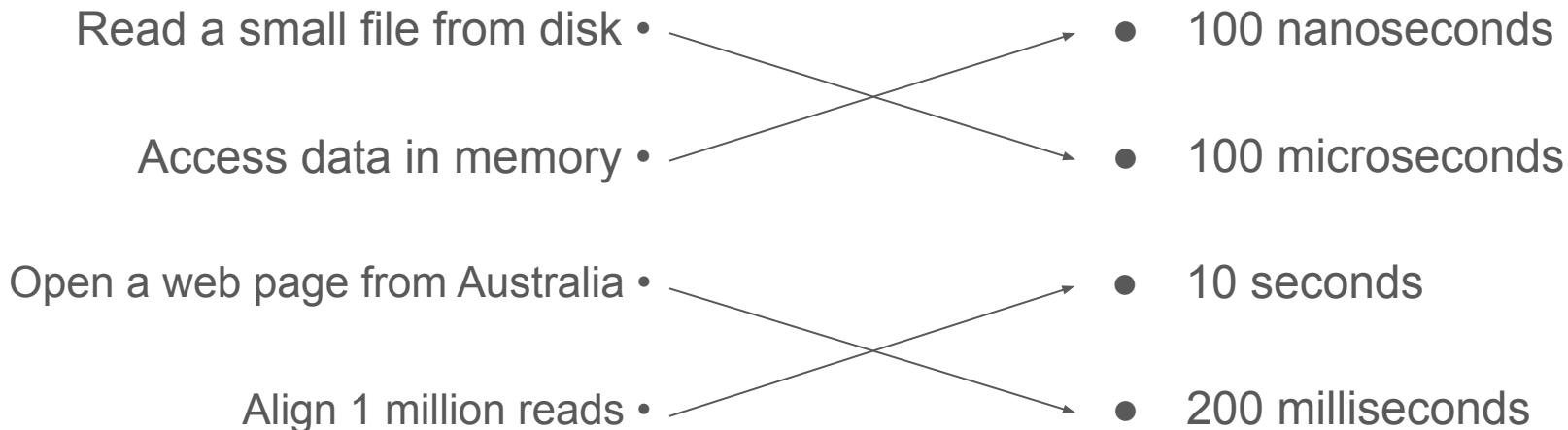
• 100 microseconds

• 10 seconds

• 200 milliseconds

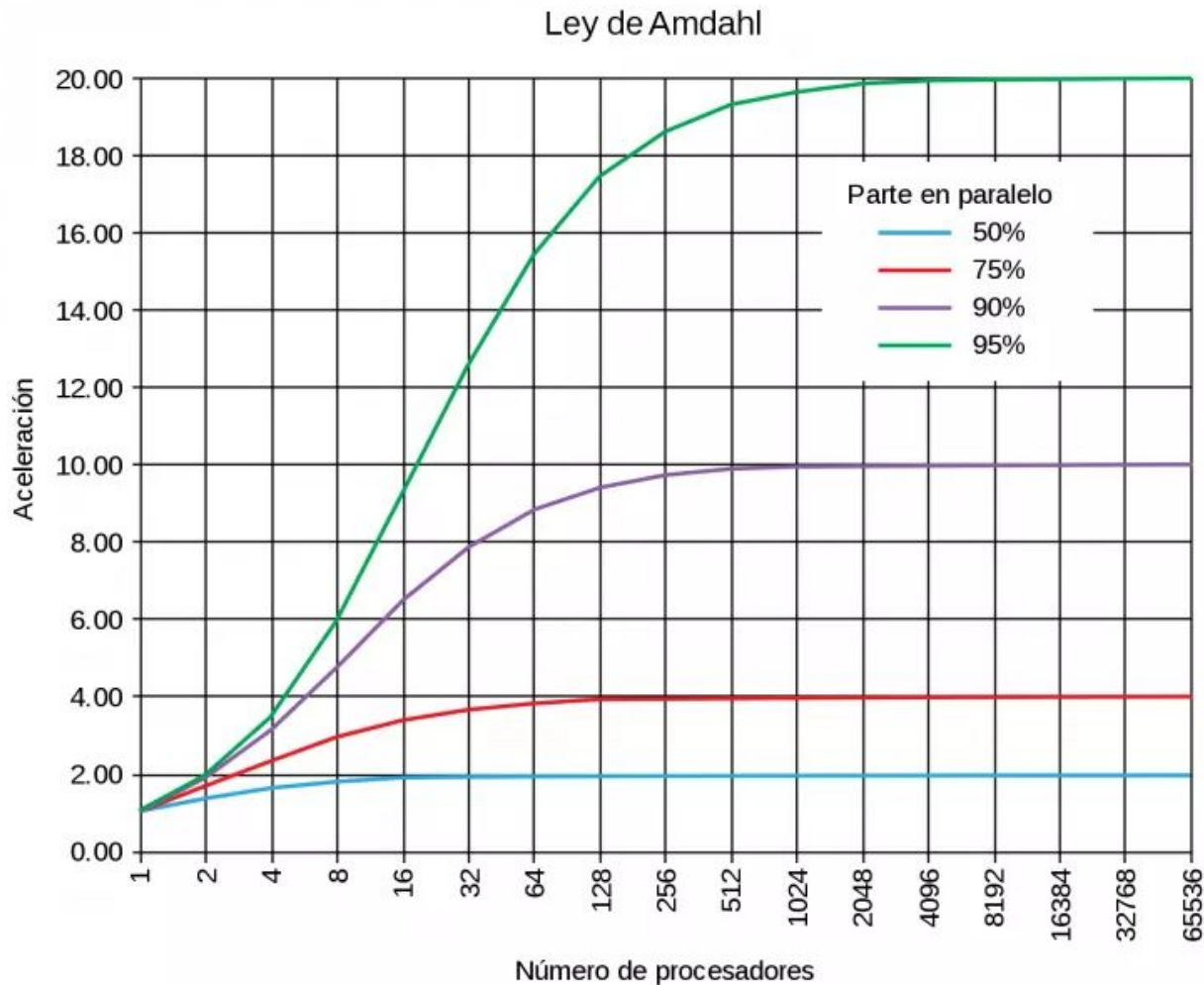
-	-	$10^0$	1
deci	d	$10^{-1}$	0,1
centi	c	$10^{-2}$	0,01
mili	m	$10^{-3}$	0,001
micro	$\mu$	$10^{-6}$	0,000 001
nano	n	$10^{-9}$	0,000 000 001
pico	p	$10^{-12}$	0,000 000 000 001

# Connect the dots from left to right



-	-	$10^0$	1
deci	d	$10^{-1}$	0,1
centi	c	$10^{-2}$	0,01
mili	m	$10^{-3}$	0,001
micro	μ	$10^{-6}$	0,000 001
nano	n	$10^{-9}$	0,000 000 001
pico	p	$10^{-12}$	0,000 000 000 001

Amdahl's law  
=  
small  
bottlenecks  
add up to big  
bottlenecks



# Programming languages: an aside





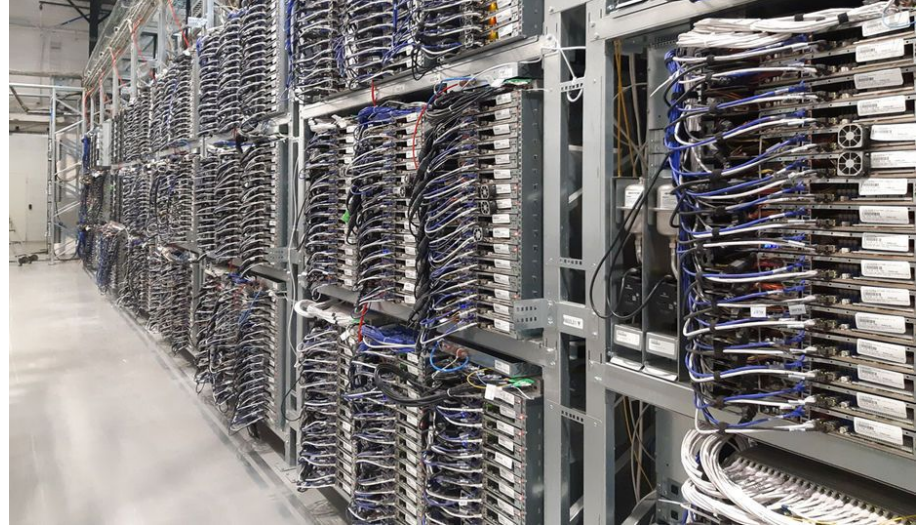
What is the biggest “computer” one can get today?

## Part 1.3: The cloud



# What is the cloud

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



OVHcloud, Roubaix, France



# What is *not* the cloud



Which of these terms *really* apply to the cloud?

# What is *not* the cloud



“Desktop”, “Mobile”, “Laptop”  
*are just ways to access the cloud*

“Server”  
*is what the cloud is made of*

“Network”  
*is what connects servers*

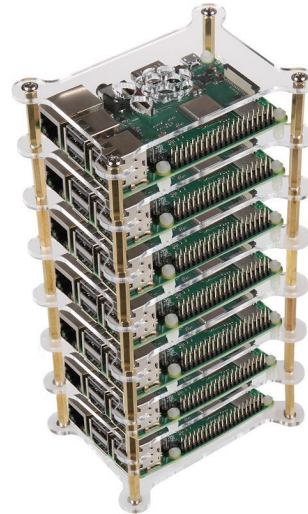
“Database”  
*is a possible application*

“Other”  
;)

# What is *nearly* the cloud

- Your university cluster
- Your 2-week access to the Workshop on Genomics 2022's resources
- 7 Raspberry Pi's stacked together

- (This dog)





# Some terms

**EC2:** *Amazon's cloud*

**Instance:** *Amazon's jargon for "a computer that is running and that you can connect to"*

**AMI:** *Amazon's jargon for something I honestly never remember except vaguely:  
a 'snapshot' of an operating system with pre-installed software on it*

**S3:** *Amazon's big cloud hard drive*

# “Storing information in the cloud”?

It just means the data is somewhere on a computer on Internet

“Cloud”, for us bioinformaticians, is really about doing some long task



chicano joker @datLucario

Apr 24

when information is “stored in the cloud” that means a samoyed, somewhere, knows it. the trick is knowing which samoyed has your data

Apr 24, 2022 · 11:27 AM UTC

45 2,357 106 9,632



chicano joker @datLucario







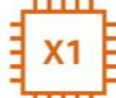




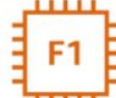

Apr 24

this samoyed, for example, does not know anything. it has not had a single thought its entire life



6 231 8 1,418

# Instance types

General Purpose	Compute Optimised	Memory Optimised	Accelerated Computing	Storage Optimised
 ARM based core and custom silicon	 Compute - CPU intensive apps and DBs	 RAM - Memory intensive apps and DB's	 Processing optimised- Machine Learning	 High Disk Throughput - Big data clusters
 Tiny - Web servers and small DBs		 Xtreme RAM - For SAP/Spark	 Graphics Intensive - Video and streaming	 IOPS - NoSQL DBs
 Main - App servers and general purpose		 High Compute and High Memory - Gaming	 Field Programmable - Hardware acceleration	 Dense Storage - Data Warehousing

Full list:  
<https://instances.vantage.sh/>

API Name	Memory	vCPUs	Instance Storage	Network Performance	Linux On Demand cost
<input type="text" value="Search"/>	<input type="text" value="Search"/>	<input type="text" value="Search"/>	<input type="text" value="Search"/>	<input type="text" value="Search"/>	<input type="text" value="Search"/>
m6a.24xlarge	384.0 GiB	96 vCPUs	EBS only	37.5 Gigabit	\$4.147200 hourly
m5dn.xlarge	16.0 GiB	4 vCPUs	150 GB NVMe SSD	Up to 25 Gigabit	\$0.272000 hourly
c6a.8xlarge	64.0 GiB	32 vCPUs	EBS only	12.5 Gigabit	\$1.224000 hourly
g5.16xlarge	256.0 GiB	64 vCPUs	1900 GB NVMe SSD	25 Gigabit	\$4.096000 hourly

## Costs (2022)

\$0.15/hour for “your laptop” (8 GB ram, 4 CPUs) on the cloud

\$2.5/hour for a beefy cluster node (128 GB ram, 64 CPUs)

~\$200 for 1 week analysis, no shutdown, 32 cores, 64 GB RAM

# What it looks like

Instances (1/5) [Info](#)

Launch instances

Search

	Name	Instance ID	Instance state	Instance type	Status check	Alarm
<input type="checkbox"/>	serratus-ryan	i-08ad942d5d8931995	⊖ Stopped	t3a.medium	–	No alarm
<input type="checkbox"/>	serr-api	i-06cba368af1300836	⊕ Running	t2.micro	⊕ 2/2 checks passed	No alarm
<input type="checkbox"/>	serratus-sum...	i-0749cb5cf2172867a	⊕ Running	t2.micro	⊕ 2/2 checks passed	No alarm
<input checked="" type="checkbox"/>	serratus-sum...	i-08e804b7d41c63ec3	⊕ Running	t2.micro	⊕ 2/2 checks passed	No alarm
<input type="checkbox"/>	artem-work	i-0145b70da07b26393	⊖ Stopped	c5n.2xlarge	–	No alarm

- Launch instances
- Launch instance from template
- Migrate a server
- Connect
- Stop instance
- Start instance
- Reboot instance
- Hibernate instance



## What if you're not swimming in Amazon credits?

Alternatives:

- GPU
- FPGA (Dragen, etc)
- Your local cluster (see next slide)
- Your national cluster (see next slide)
- Quantum Computing?
- DNA computing?
- (any other idea?)



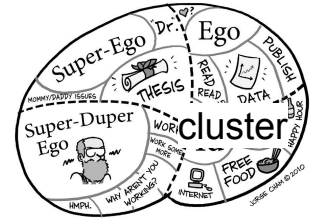


# University/Country computers

(Use them!)

Typically:

- 1) “Pre”-get an account, even if you have nothing to compute, just to get familiar
- 2) Experiment with `sbatch/srun`
- 3) Sometimes need to fill a project application for large jobs (short/worth it)

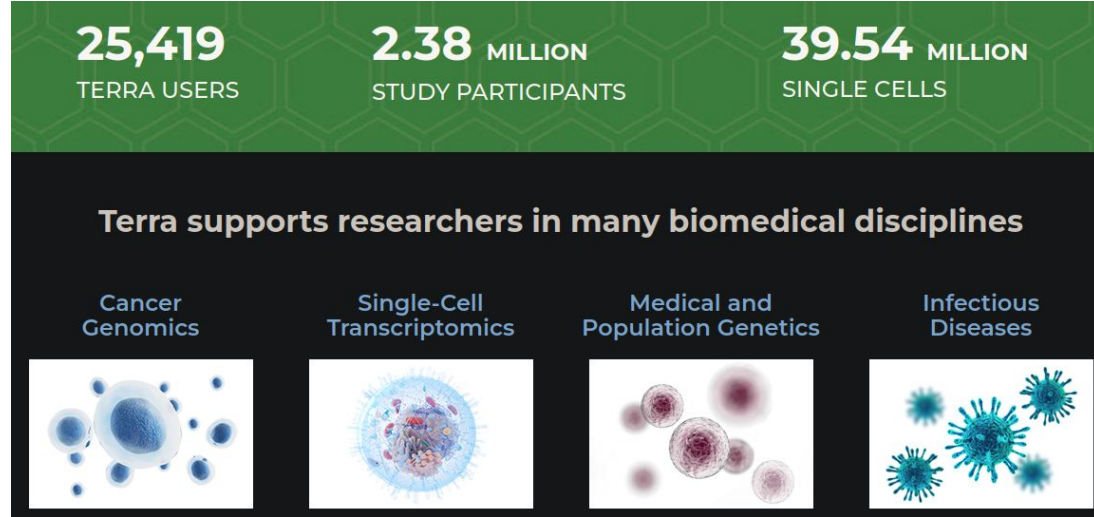


**compute** | **calcul**  
canada | canada



Accessible workflows

Terra.bio

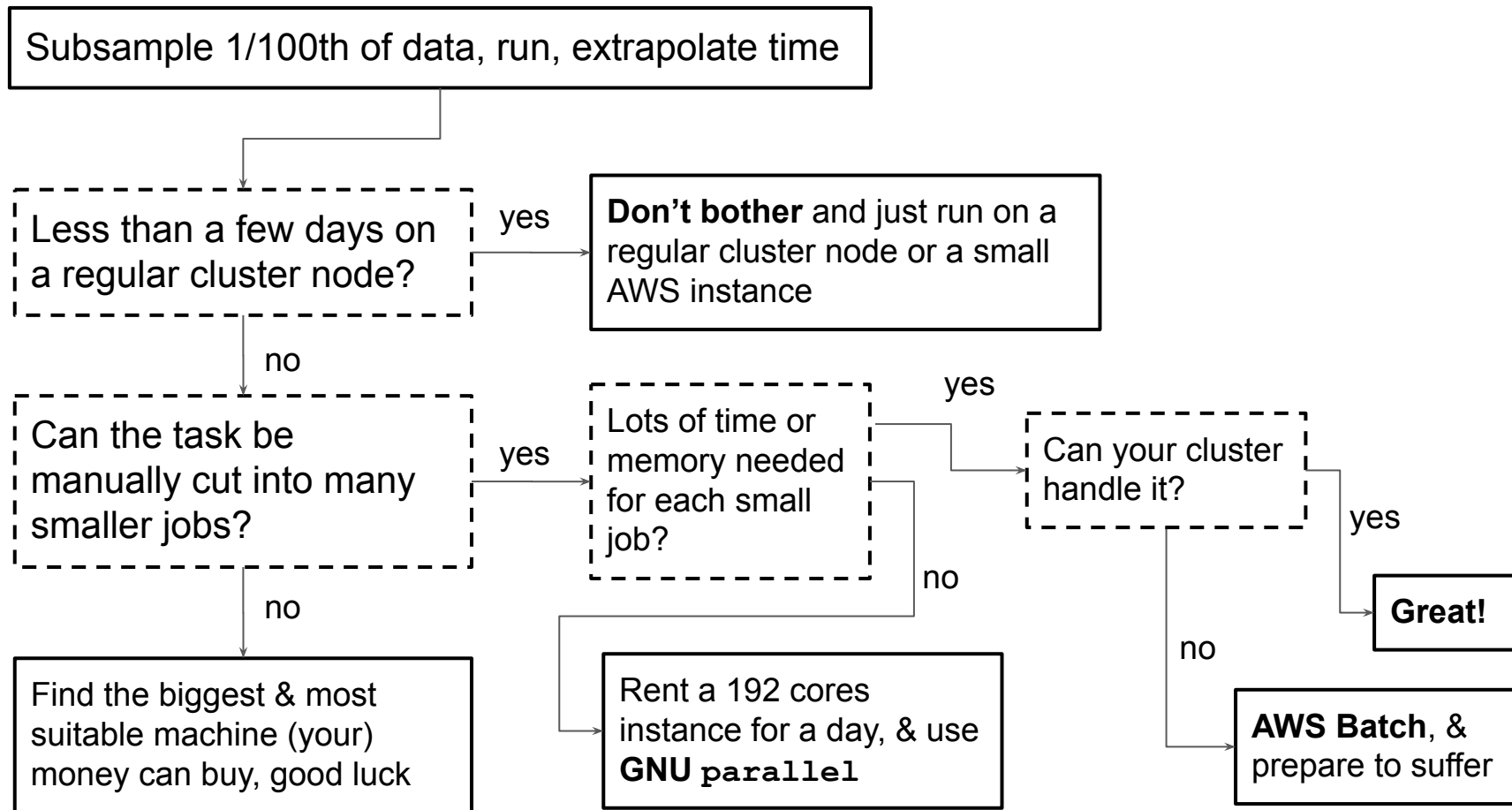


With some more effort:

nextflow

snakemake

# ✨ Rayan's "big compute" cheat sheet ✨



Part 1.4:  
Some large-scale  
genomics analyses



Part 1.4.1:  
Some large-scale  
genomics analyses



# The “nr” database of BLAST

*“The nucleotide collection consists of **GenBank+EMBL+DDBJ+PDB+RefSeq** sequences, but excludes EST, STS, GSS, WGS, TSA. [...] The database is non-redundant.”*

125 GB compressed, <ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz>

## The “refseq\_genomes” database:

*“This database contains NCBI Refseq genomes across all taxonomy groups.”*

1.5 TB [ref: STAT]



Part 1.4.2:  
Some large-scale  
genomics  
analyses:

Ultra-rapid  
Nanopore  
sequencing



# From sequencing to diagnostic in < 6 hours

[nature](#) > [nature biotechnology](#) > [articles](#) > [article](#)

Article | [Open Access](#) | [Published: 28 March 2022](#)

## **Accelerated identification of disease-causing variants with ultra-rapid nanopore genome sequencing**

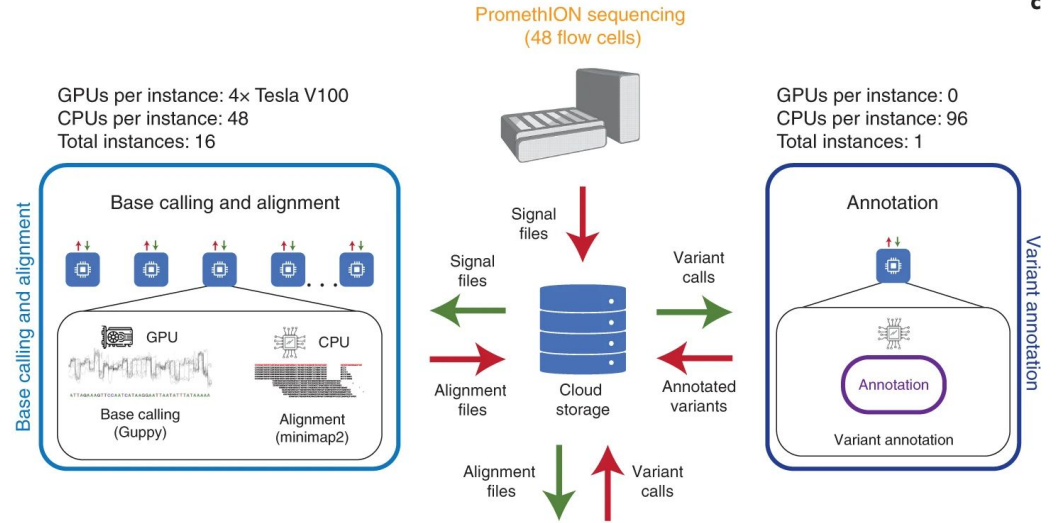
[Sneha D. Goenka](#), [John E. Gorzynski](#), [Kishwar Shafin](#), [Dianna G. Fisk](#), [Trevor Pesout](#), [Tanner D. Jensen](#), [Jean Monlong](#), [Pi-Chuan Chang](#), [Gunjan Baid](#), [Jonathan A. Bernstein](#), [Jeffrey W. Christle](#), [Karen P. Dalton](#), [Daniel R. Garalde](#), [Megan E. Grove](#), [Joseph Guillory](#), [Alexey Kolesnikov](#), [Maria Nattestad](#), [Maura R. Z. Ruzhnikov](#), [Mehrzaad Samadi](#), [Ankit Sethia](#), [Elizabeth Spiteri](#), [Christopher J. Wright](#), [Katherine Xiong](#), [Tong Zhu](#), [Miten Jain](#), [Fritz J. Sedlazeck](#), [Andrew Carroll](#), [Benedict Paten](#) & [Euan A. Ashley](#) ✉

— Show fewer authors

[Nature Biotechnology](#) (2022) | [Cite this article](#)

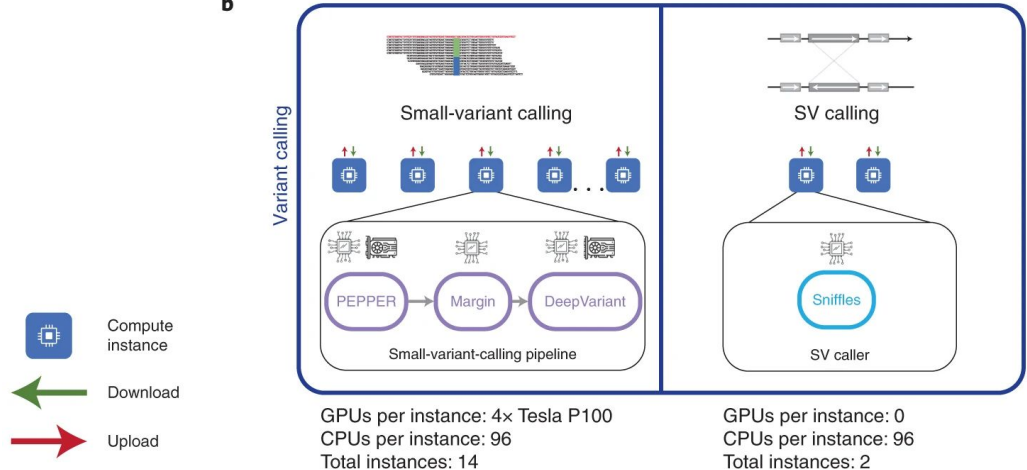
# How? cloud.

a



c

b



Part 1.4.3:  
Some large-scale  
genomics  
analyses:

160,000 *E. coli*'s



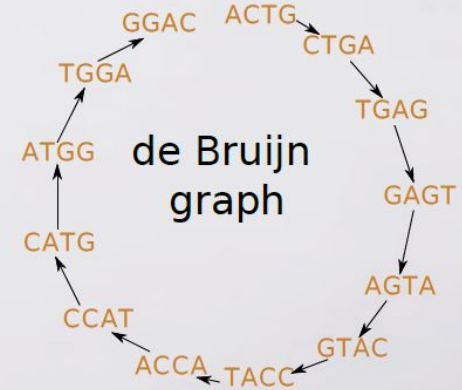
# Recall the de Bruijn graph

Reference genome  
ACTGAGTACCATGGAC  
ACTGAGTAC  
Reads  
CTGAGTACCAT  
GAGTACCATGGAC

*k*-mers

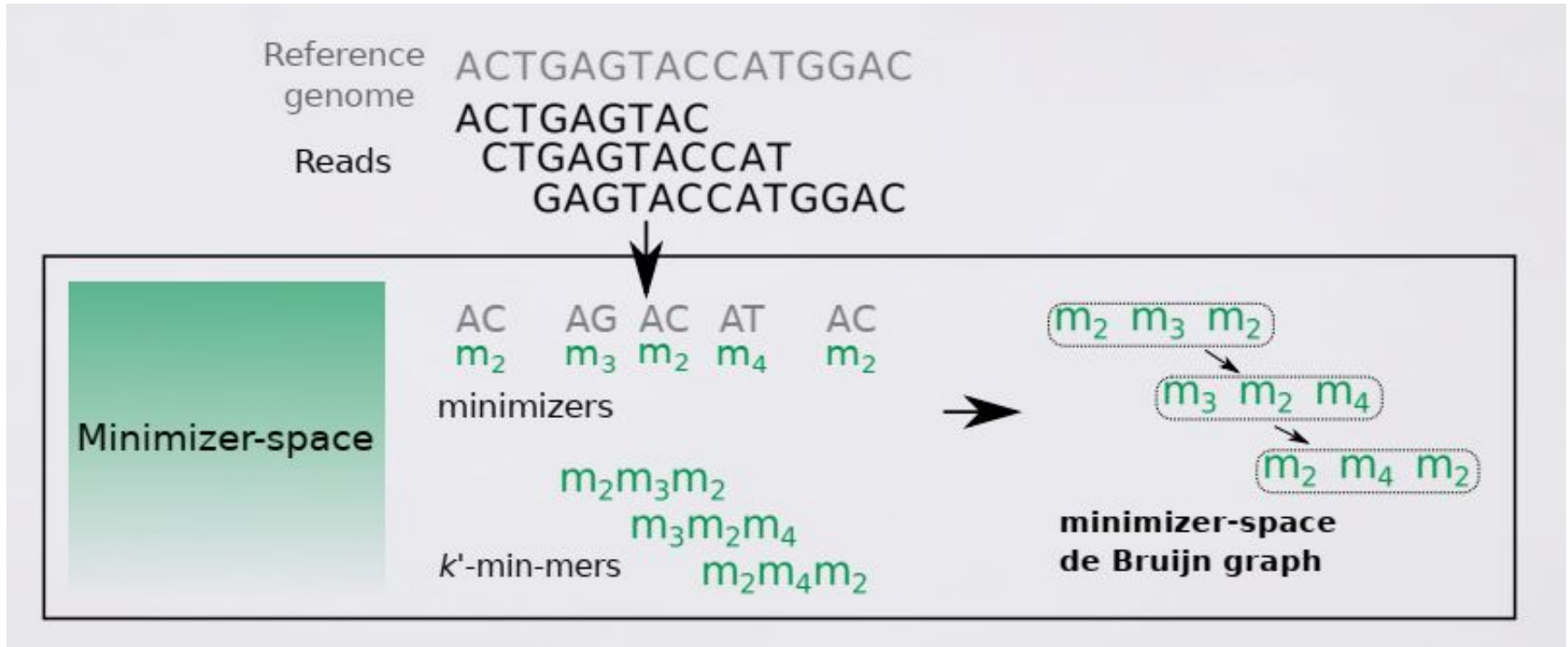
ACTG TACC GGAC  
CTGA ACCA  
TGAG CCAT  
GAGT CATG  
AGTA ATGG  
GTAC TGGA

Base-space



# Now glance at an improved de Bruijn graph

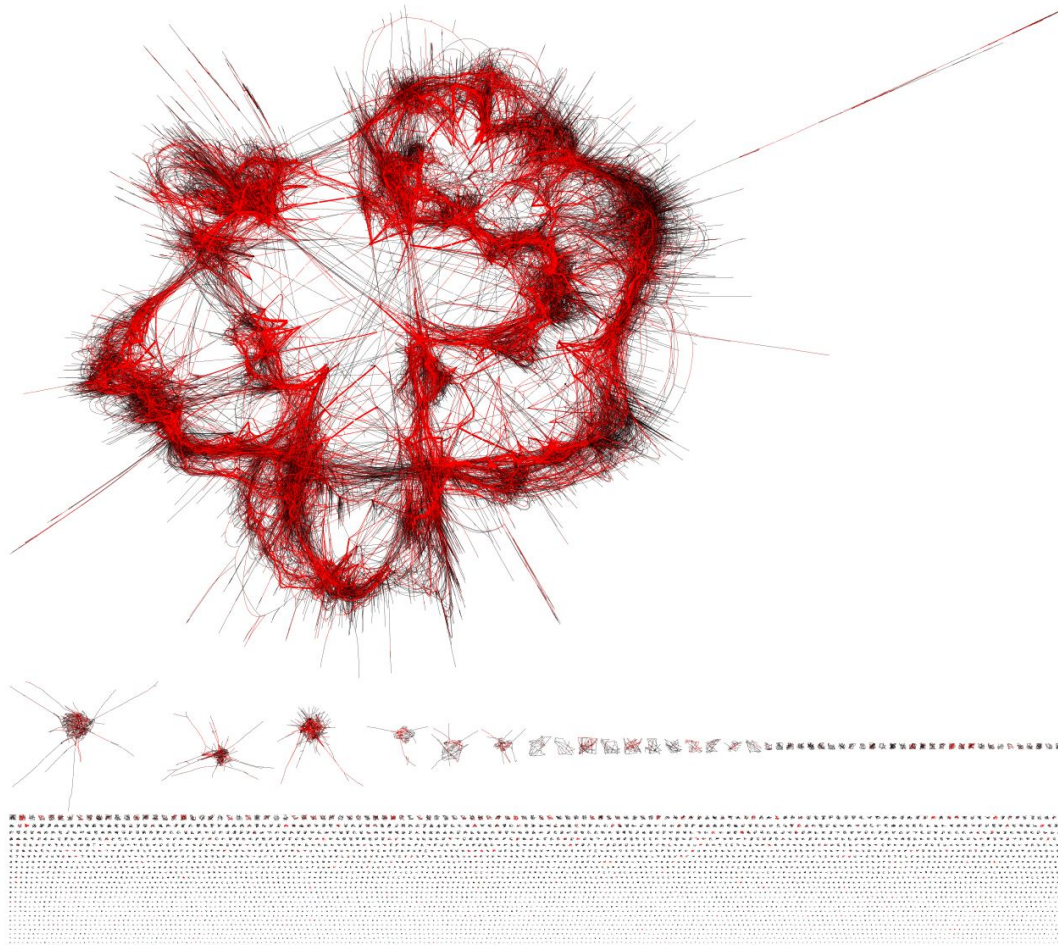
(disclaimer: not technically an improvement, more like a powerful variant)





167,000 E. coli's  
graph

~500k nodes



# Exploring 167,000 E. coli

**Graph drawing**

Scope:

Node(s):

Match:  Exact  Partial

Distance:

Style:  Single  Double

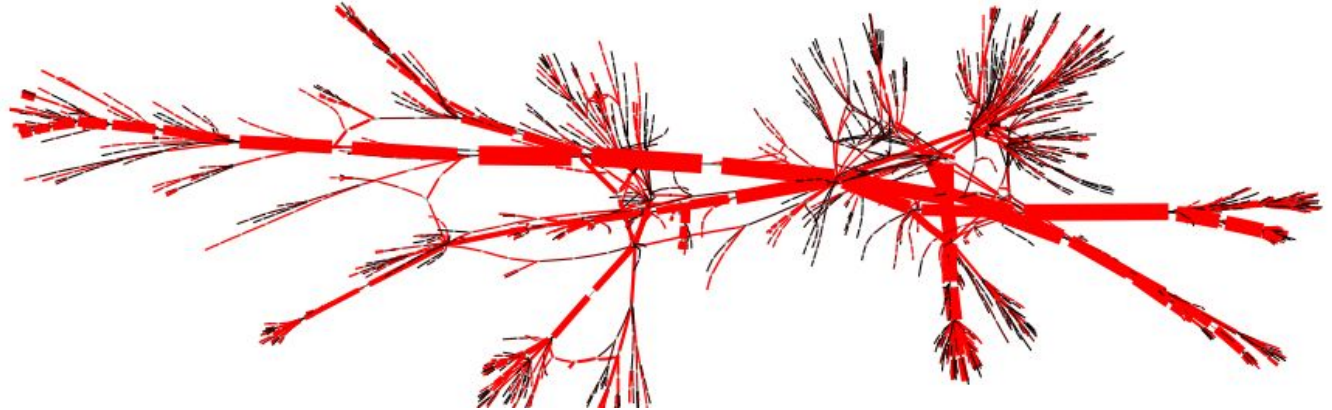
**Graph display**

Zoom:

Node width:

Colour by depth

**Node labels**

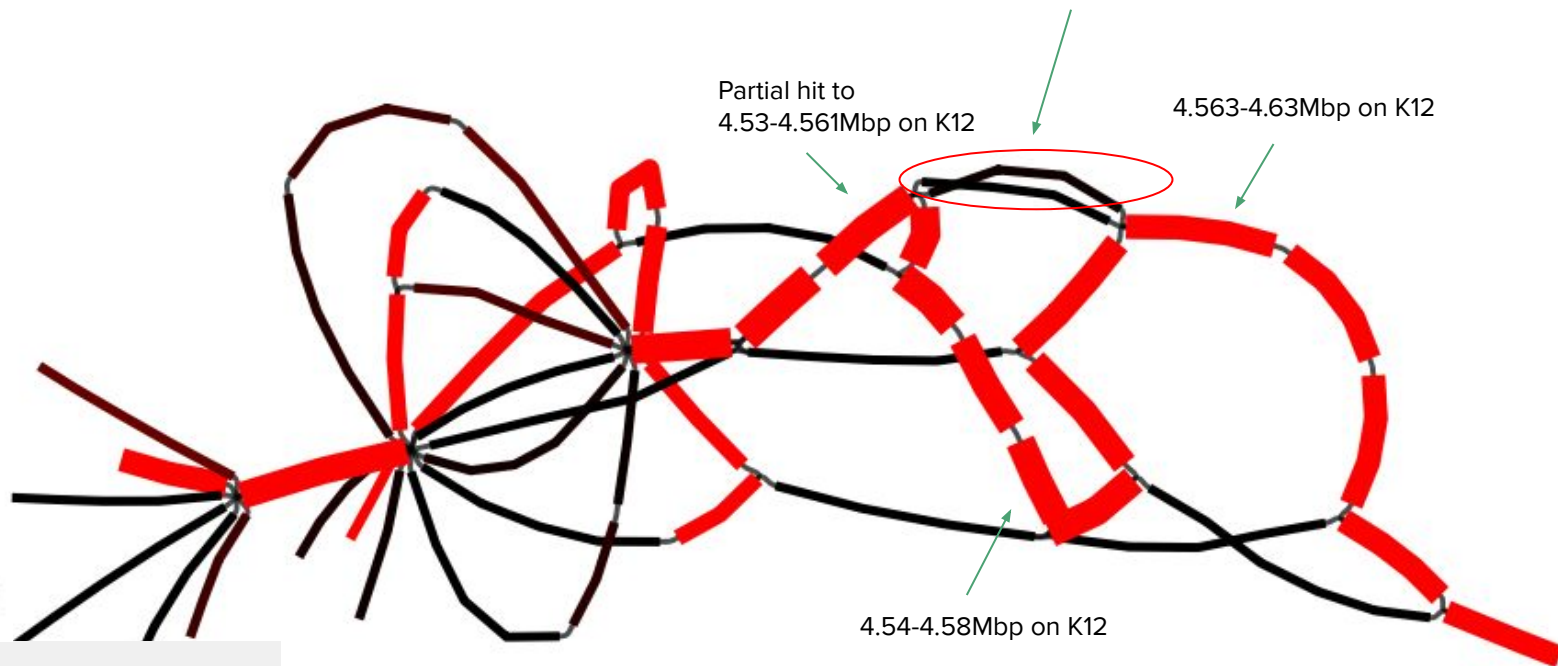


# Graph comparative genomics

**Split hit on K12 -> exchange?**


4.53-4.57Mbp (start -> 34kbp)


4.58-4.62Mbp (44kbp -> end)



 Bandage

BLAST

 Create/view BLAST search

 Query: none

Summary:

species-scale  
graphical bacterial  
pangenomics



This all has been foretold..



## Notes from the Datapocalypse

C. Titus Brown  
School of Veterinary Medicine;  
Genome Center & Data Science Initiative  
UC Davis



March 23, 2017



@ctitusbrown

#jgi2017

A large blue USS Enterprise NCC-1701-C is shown from a low-angle perspective, appearing to fly towards the viewer. The ship's saucer section is prominent, with the registration number 'NCC-1701-C' visible on its underside. The nacelles and engines are illuminated with a bright blue glow. The ship is set against a black background of space, with the curved horizon of Earth and its blue atmosphere visible in the lower-left corner. The text 'bigger data' is overlaid in white on the saucer section, and 'big data' is overlaid in white on the Earth's horizon.

bigger data

big data



Is it coffee break time?



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



# GenBank



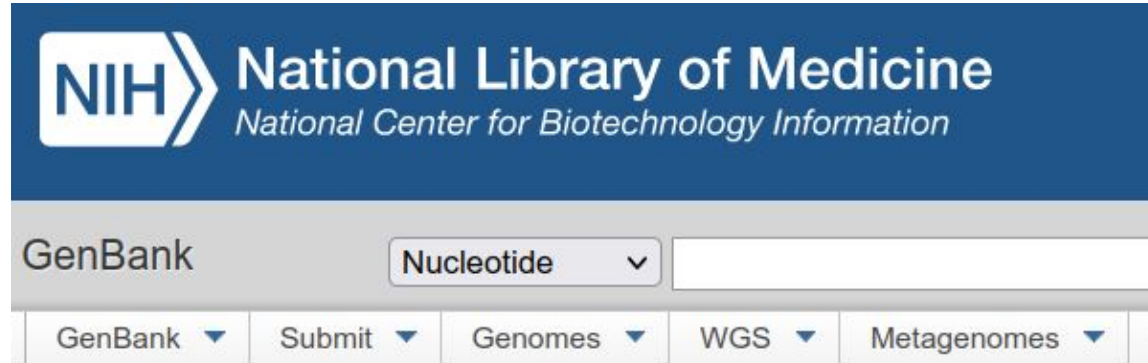
**Type:** assemblies

**Size:** 1.2 TB ([April 2022](#))

**Diversity:** high

Particularity: all sequences are *annotated*

# NCBI WGS



## Whole Genome Shotgun Submissions

### What is Whole Genome Shotgun (WGS)?

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

**Type:** assemblies


**Size:** 16 TB ([April 2022](#))

**Diversity:** high

Difference with GenBank: sequences are not necessarily annotated

# NCBI SRA

SRA   [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](#)

[NextSeq 500 paire](#)

1. 1 ILLUMINA (Illumina)  
Accession: ERX34307

[NextSeq 500 paire](#)

2. 1 ILLUMINA (Illumina)  
Accession: ERX34307

[NextSeq 500 paire](#)

3. 1 ILLUMINA (Illumina)  
Accession: ERX34307

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

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

< 1 1 346553 >

View:  biological reads  technical reads

#### Reads (separated)

1. [ERR3407135.1 ERS3549882](#)  
name: NB551234.144:HL523AFXY.1:11101:5421:  
member: default

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

```
ACCTGAGCGCGCAGCTCCAGTAAATCAAACGCGCGCGGAAATTTGGGATGTTCCATCAGT  
TTCCAGGCGCGTTTGCCCTGACGTCGCGACATGCGTAACTGAAGCTGCCAAATATCACGG  
GTAAGCGTGTGTAAGCGCTTTCGGATCGCCA
```

2. [ERR3407135.2 ERS3549882](#)  
name: NB551234.144:HL523AFXY.1:11101:2248:  
member: default

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

```
ATCAACAACAGCGGGAATACCACCTCTTCCAGCCGTTGTTTCCAAACAAATACGCGTTAAT  
TCACCGAAACCGGACAGCGCAATGGAACGCATCATTTGCCGAGGTGTTGCAGAAATACGGA  
AAACCGCATCCGAAACGAGATGCGCGTTAAT
```

3. [ERR3407135.3 ERS3549882](#)  
name: NB551234.144:HL523AFXY.1:11101:2566:  
member: default

4. [ERR3407135.4 ERS3549882](#)  
name: NB551234.144:HL523AFXY.1:11101:2199:  
member: default


5. [ERR3407135.5 ERS3549882](#)  
name: NB551234.144:HL523AFXY.1:11101:2350:  
member: default

# 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](#)

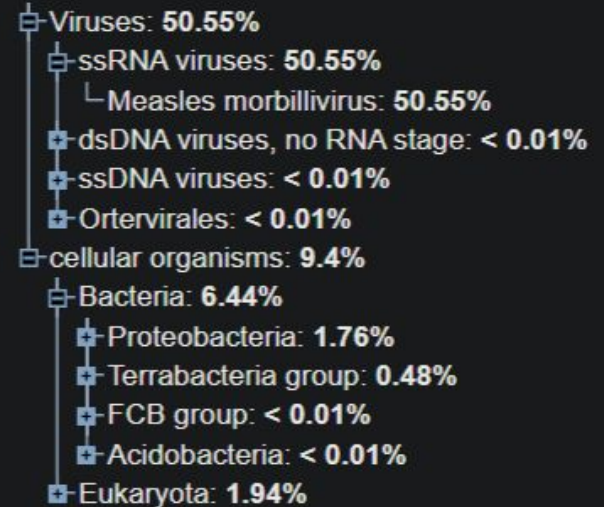
[Genome Biology](#) **22**, Article number: 270 (2021) | [Cite this article](#)

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

## Taxonomy Analysis

Unidentified reads: 40.04%

Identified reads: 59.96%

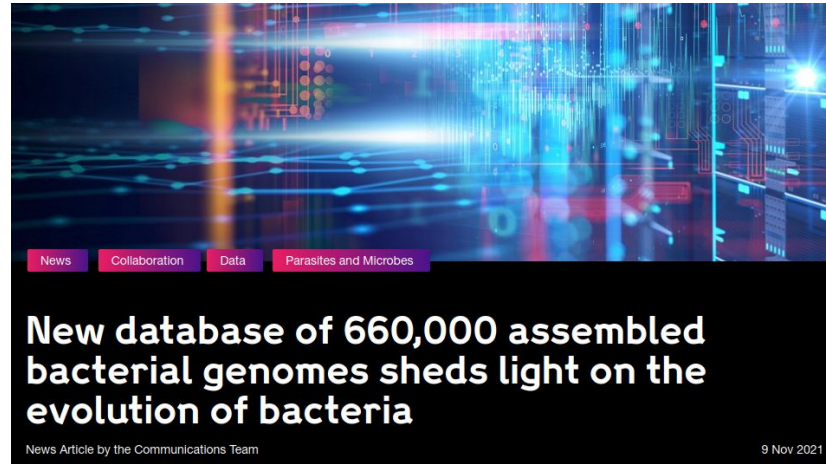




# What is STAT good for?

- 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..

# 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:



Taxons in component

18

22

4

22

10

Dominant species

*Mycobacterium  
tuberculosis*

*Salmonella  
enterica*

*Burkholderia  
gladioli*

*Pseudomonas  
protegens*

*Cupriavidus  
alkaliphilus*

# Many others (often metagenomic)



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

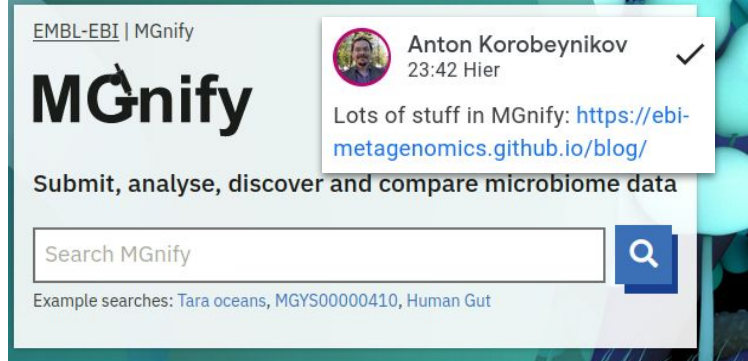
When you said "in this thread we are releasing", I was hoping you were actually going to tweet out the entire thing

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**

Submit, analyse, discover and compare microbiome data

Search MGNify

Example searches: [Tara oceans](#), [MGYS00000410](#), [Human Gut](#)

Anton Korobeynikov  
23:42 Hier ✓

Lots of stuff in MGNify: <https://ebi-metagenomics.github.io/blog/>

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

#### xxx 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

# Summary of Part 1

- Lots of genomics data
- Many great analyses could be made
- Cloud helps at the largest scale
- What the field needs: biologists who think big and know computing. You?
  - What large-scale project would you do?



Download from  
**Dreamstime.com**  
This watermark-free image is for previewing purposes only.

id 83717812  
© Curnypah | Dreamstime.com

Recall: with infinite computation, one could perform wonderful, ground-breaking genomics



(Although in practice it never works the first time)

nor the second time and when it works the third time you're not sure why



# 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

Slide help: Michel Attafeu, Sophie Shaw, Cami, Karin, M, Malfoy

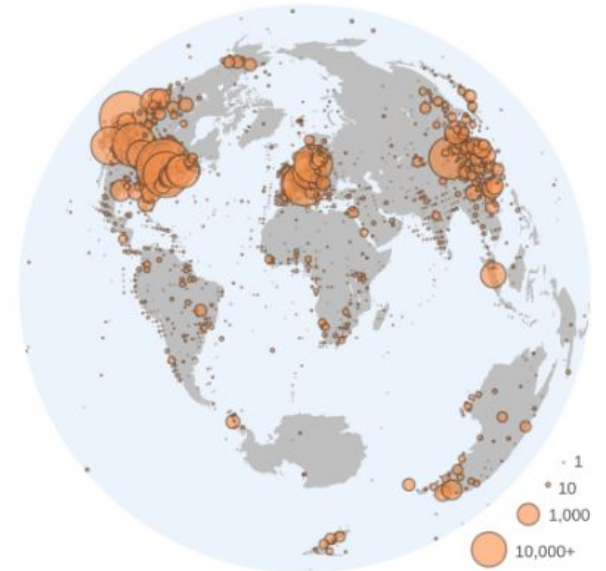
Any questions?



# Part 2: Petabase-scale viral discovery

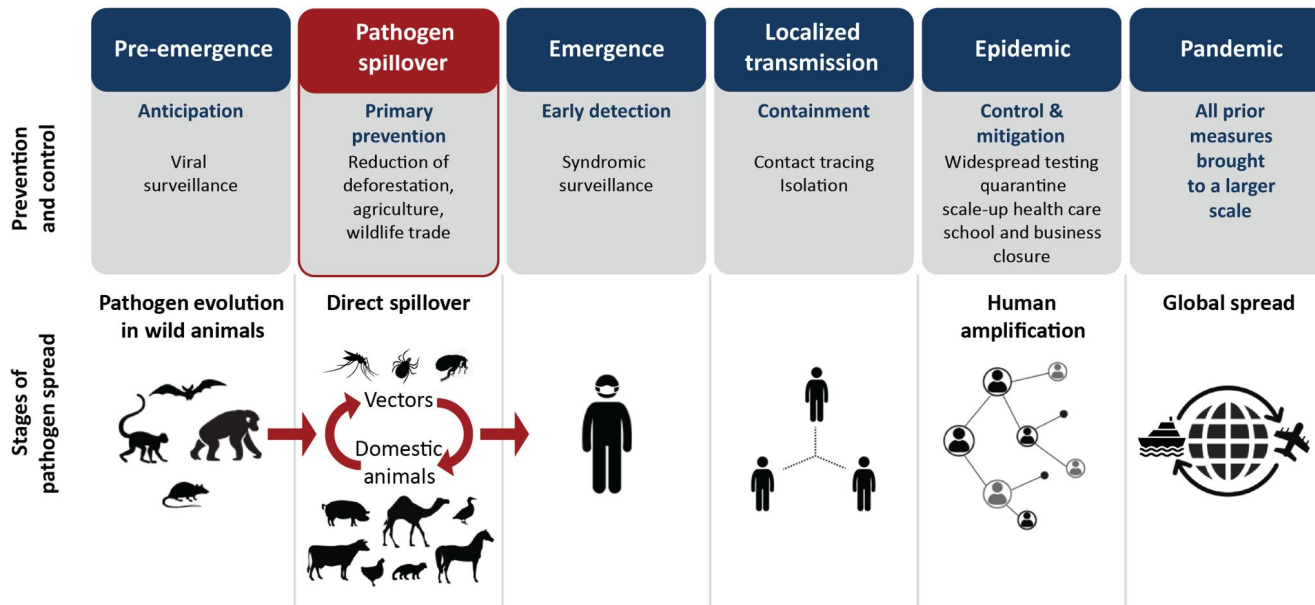
*Rayan Chikhi, on behalf of the Serratus team*

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

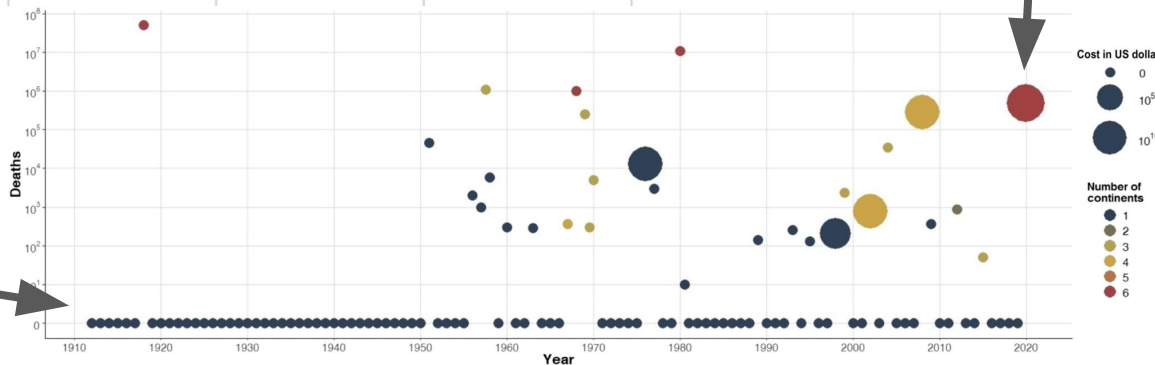


# Viral surveillance in the age of pandemics

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



We're here (quite literally in Cesky)



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](#)

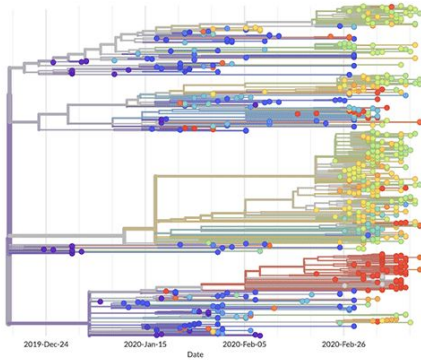


Denmark to cull mink herd over coronavirus mutation fears – here's what the science says

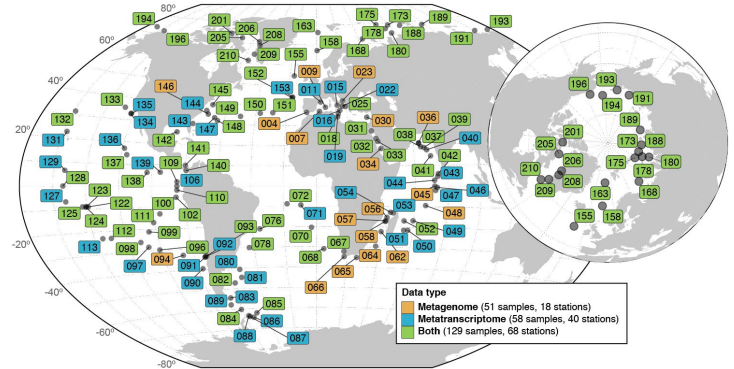
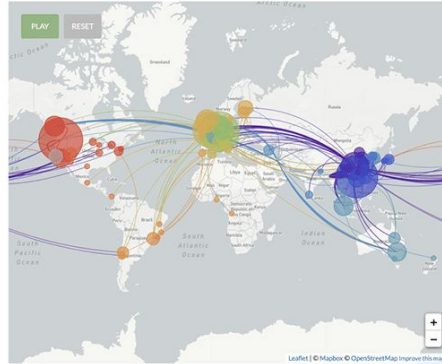
November 9, 2020 @ 4:56m EST



# Enter sequencing efforts



Nextstrain



Tara Oceans, Salazar et al. (2019)





## 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

[NextSeq 500 paire](#)

1. 1 ILLUMINA (illumina)  
Accession: ERX34307

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

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

[NextSeq 500 paire](#)

2. 1 ILLUMINA (illumina)  
Accession: ERX34307

[NextSeq 500 paire](#)

3. 1 ILLUMINA (illumina)  
Accession: ERX34307

< 1 1 346553 >

View:  biological reads  technical reads

## Reads (separated)

1. [ERR3407135.1](#) [ERS3549882](#)

name: NB551234:144:HL523AFXY:1:11101:5421:  
member: default

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

ACCTGAGCGCGCAGCTCCAGTAAATCAAACGCGGCGGGAATTTGGGATGTTCCATCAGT  
TTCAGGCGCGTTTGCCCTGACGTCGCGACATGCGTAACTGAAGCTGCCAAATATCACGG  
GTAAGCGTGGTAAGGCGTTTCGGGATCGCCA

2. [ERR3407135.2](#) [ERS3549882](#)

name: NB551234:144:HL523AFXY:1:11101:2248:  
member: default

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

ATCAACAACAGCGGGAATACCACCTCTTCCAGCCGTTGTTCCAAACCAATACGCGTTAAT  
TCACCGAAACCGCGACAGCGCAATGGAAACGCATCATTTGCCAGGTTTGCAGAATACGGA  
AAACCGCATCCGAAACGAGATCGCGGTTAAT

3. [ERR3407135.3](#) [ERS3549882](#)

name: NB551234:144:HL523AFXY:1:11101:2566:  
member: default

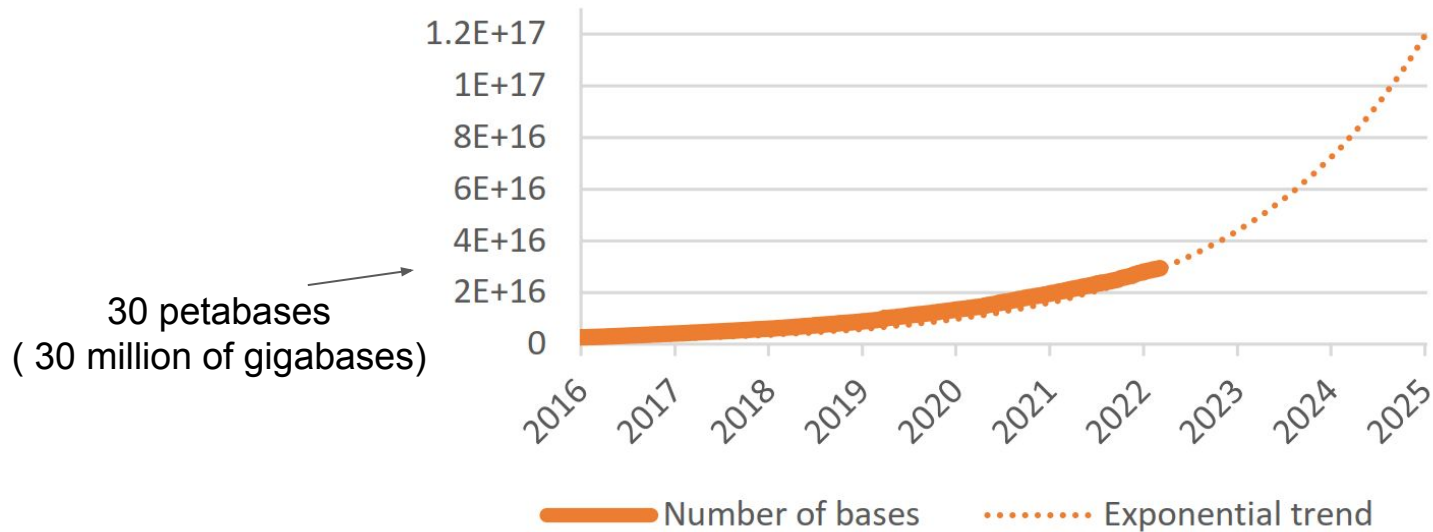
4. [ERR3407135.4](#) [ERS3549882](#)

name: NB551234:144:HL523AFXY:1:11101:21195:  
member: default

5. [ERR3407135.5](#) [ERS3549882](#)

name: NB551234:144:HL523AFXY:1:11101:23504:  
member: default

# Growth of the Sequence Read Archive



YouTube: 100-1000 PB



NCBI SRA database: 30 PB



Institut Pasteur: 8 PB



Your laptop: 0.001 PB



NCBI SRA database : 30 PB



NCBI SRA database : 30 PB



Data crypt

All the raw reads sleep  
there, undisturbed



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



Downloading all  
RNA-seq samples:



Guesstimate:

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

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

Downloading all  
RNA-seq samples:



(10 petabytes divided by 1 megabyte) / (seconds per year)



[Tous](#)

[Images](#)

[Actualités](#)

[Shopping](#)

[Vidéos](#)

[Plus](#)

Outils

Environ 291 000 résultats (0,57 secondes)



$((10 \text{ petabytes}) / (1 \text{ megabyte})) / (\text{seconds per year}) =$

316.887646408

years at 1 MB/s

# Serratus: two analyses

## 1) Nucleotide alignment

all RNAseqs vs all RNA viral genomes

## 2) Protein (translated) alignment

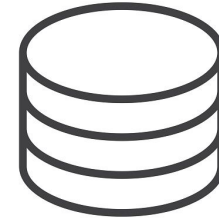
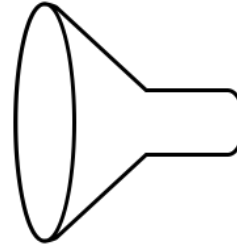
all RNAseqs vs a universal RNA virus gene

# Analysis 1:



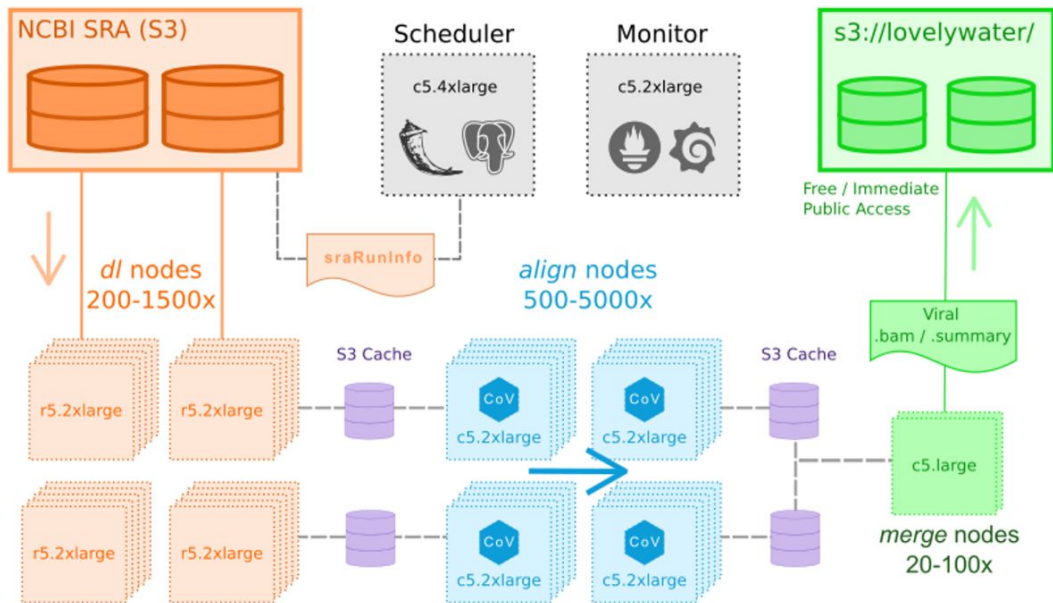
All RNA-seqs

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



**55,715 CoV+  
samples**

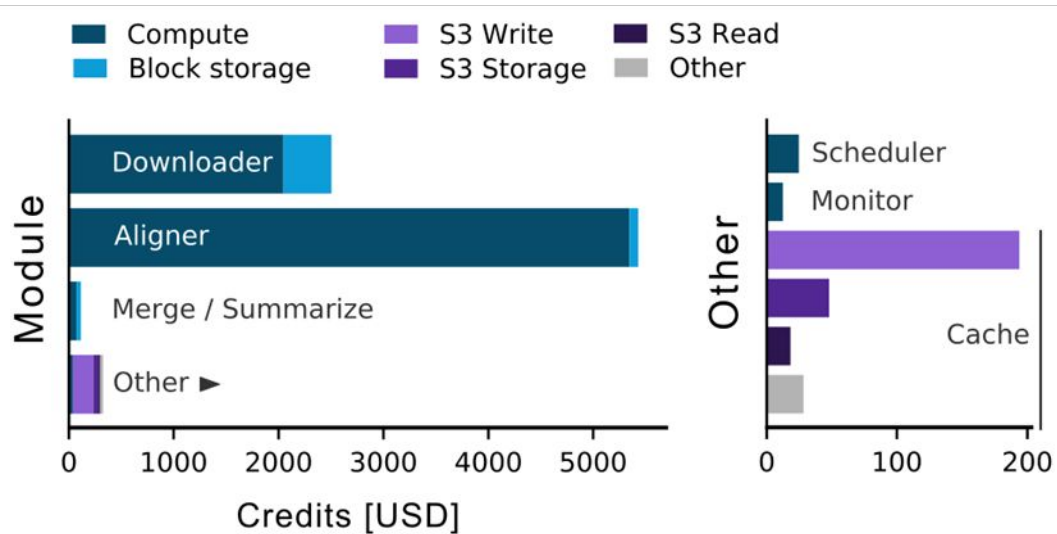
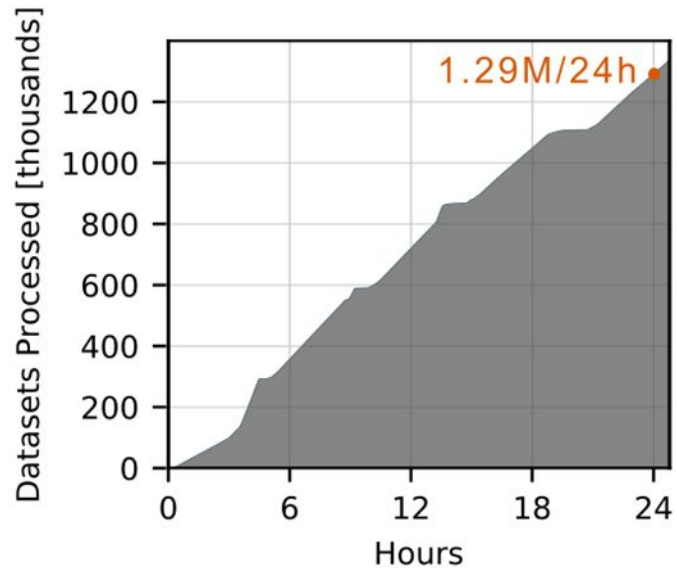
# Serratus architecture



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

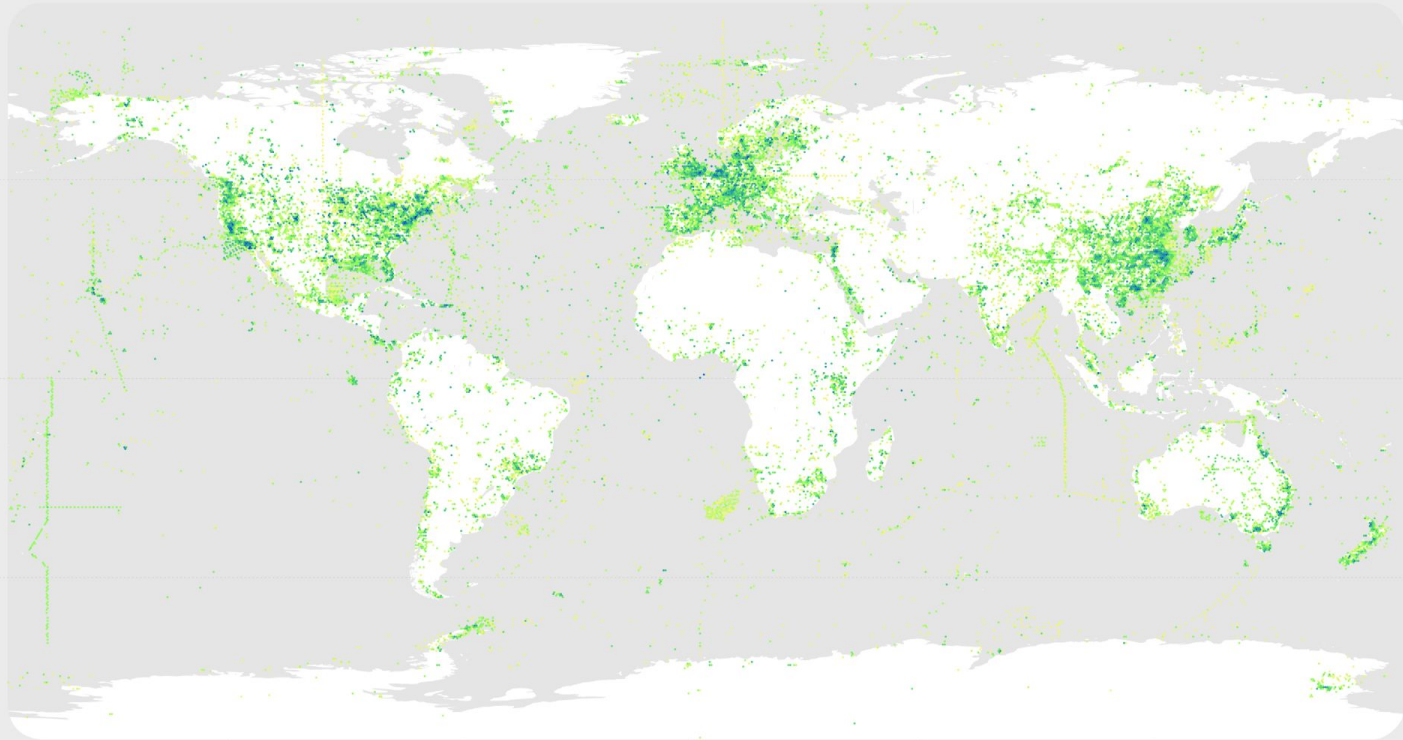
SRA also available  
@ Google Cloud,  
<https://datascience.nih.gov/strides>

# Serratus performance & costs





# Geography of SRA samples



1 20 400 8000  
Sequencing density (datasets)

Planetary DNA/RNA sequencing

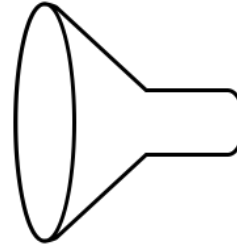


## Analysis 2:



All RNA-seqs

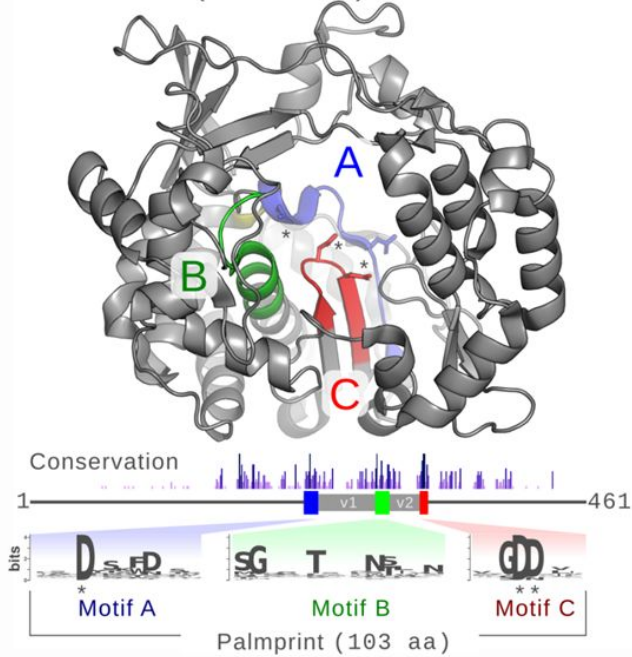
**Serratus download &  
sensitive align  
(DIAMOND2)  
to all known versions of  
RNA virus universal gene**



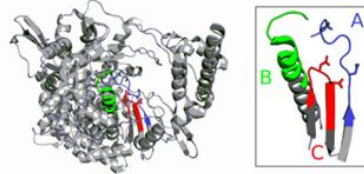
**aligned reads  
(.bam files)**

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

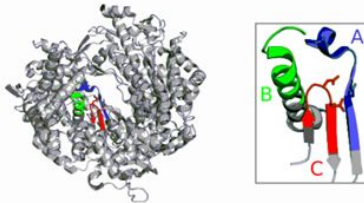
Viral RdRP (Poliovirus)



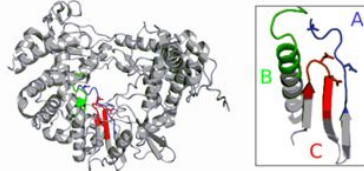
Coronaviridae



Reoviridae



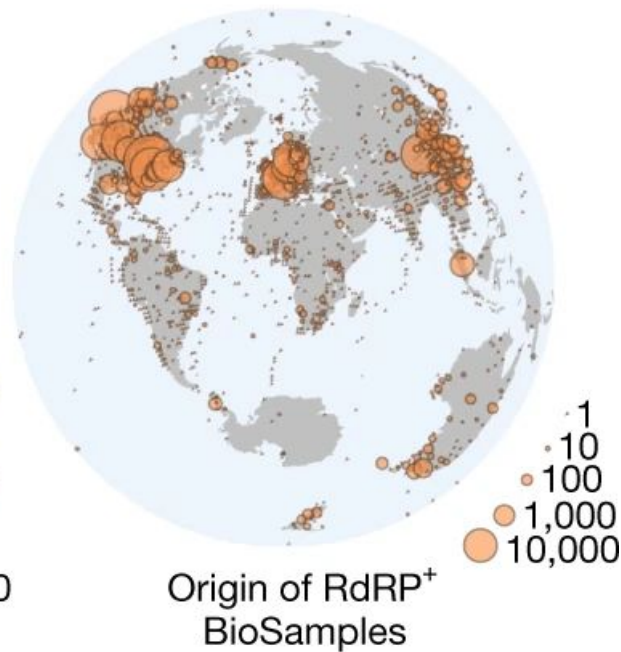
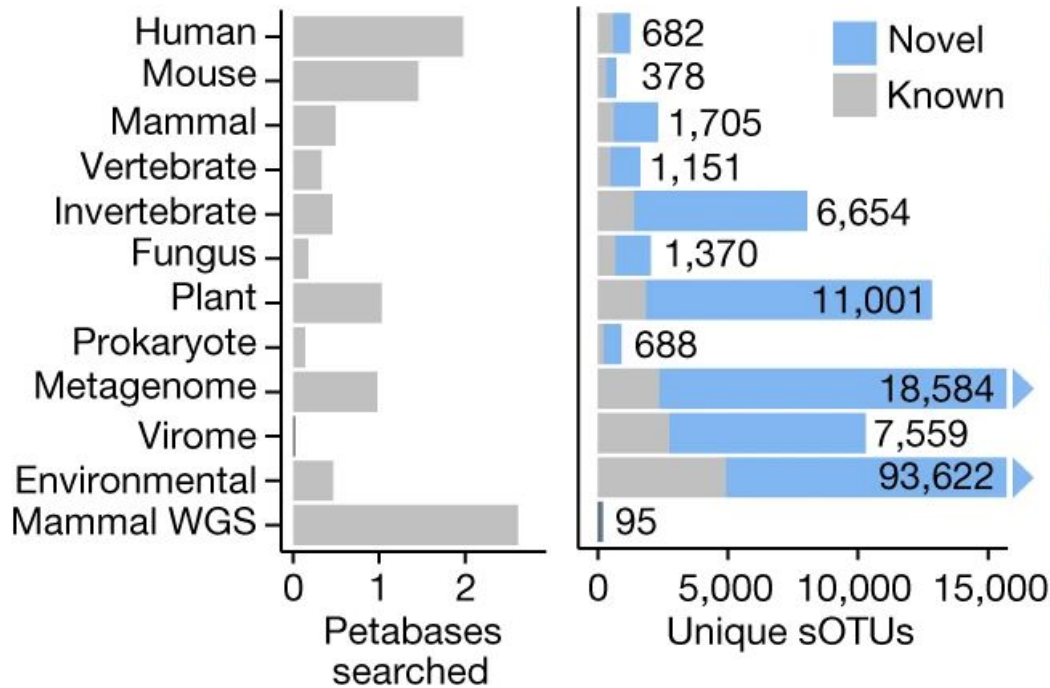
Permutotetraviridae



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

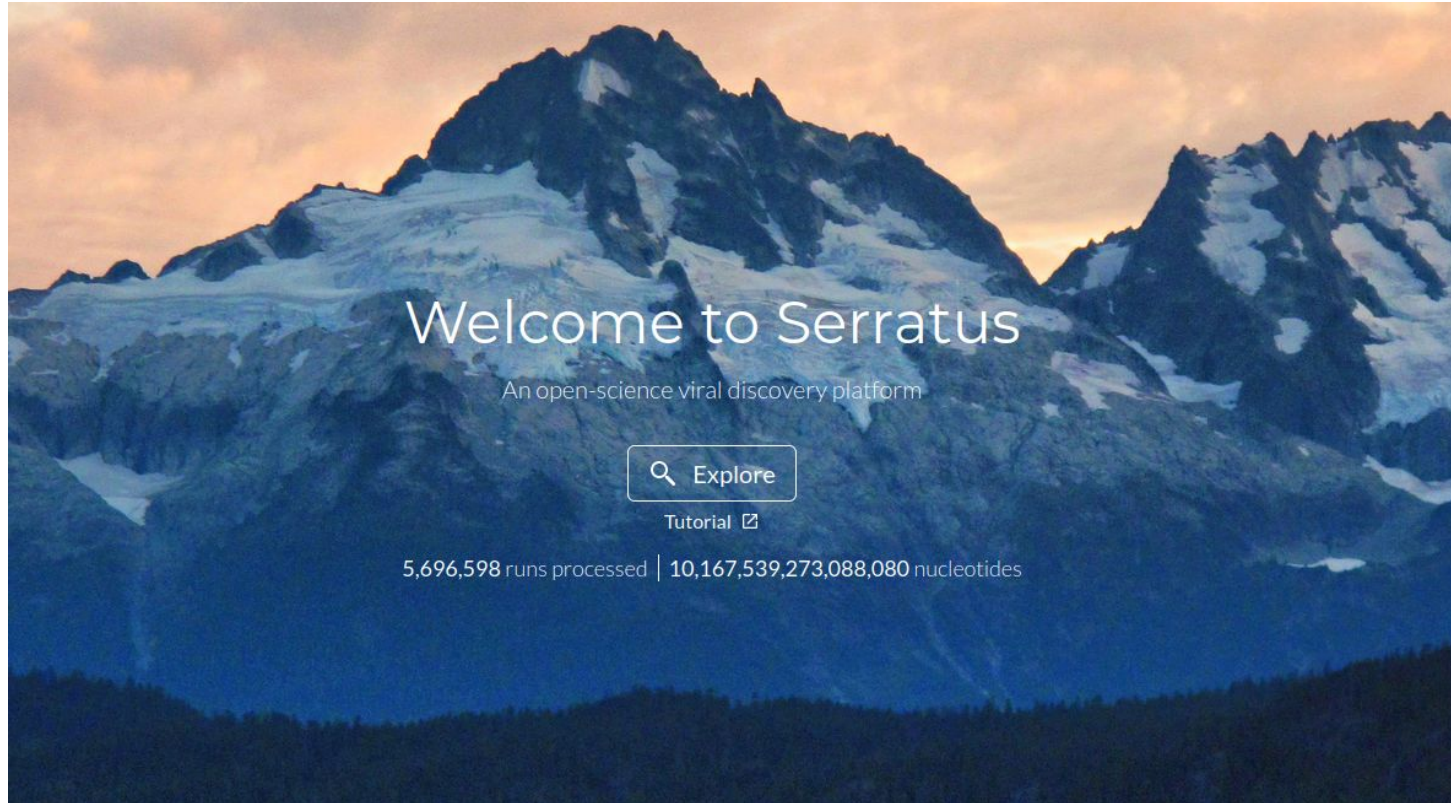
## Analysis 2, assembly

Then we “micro-assembled” all RdRp-matching reads within each sample



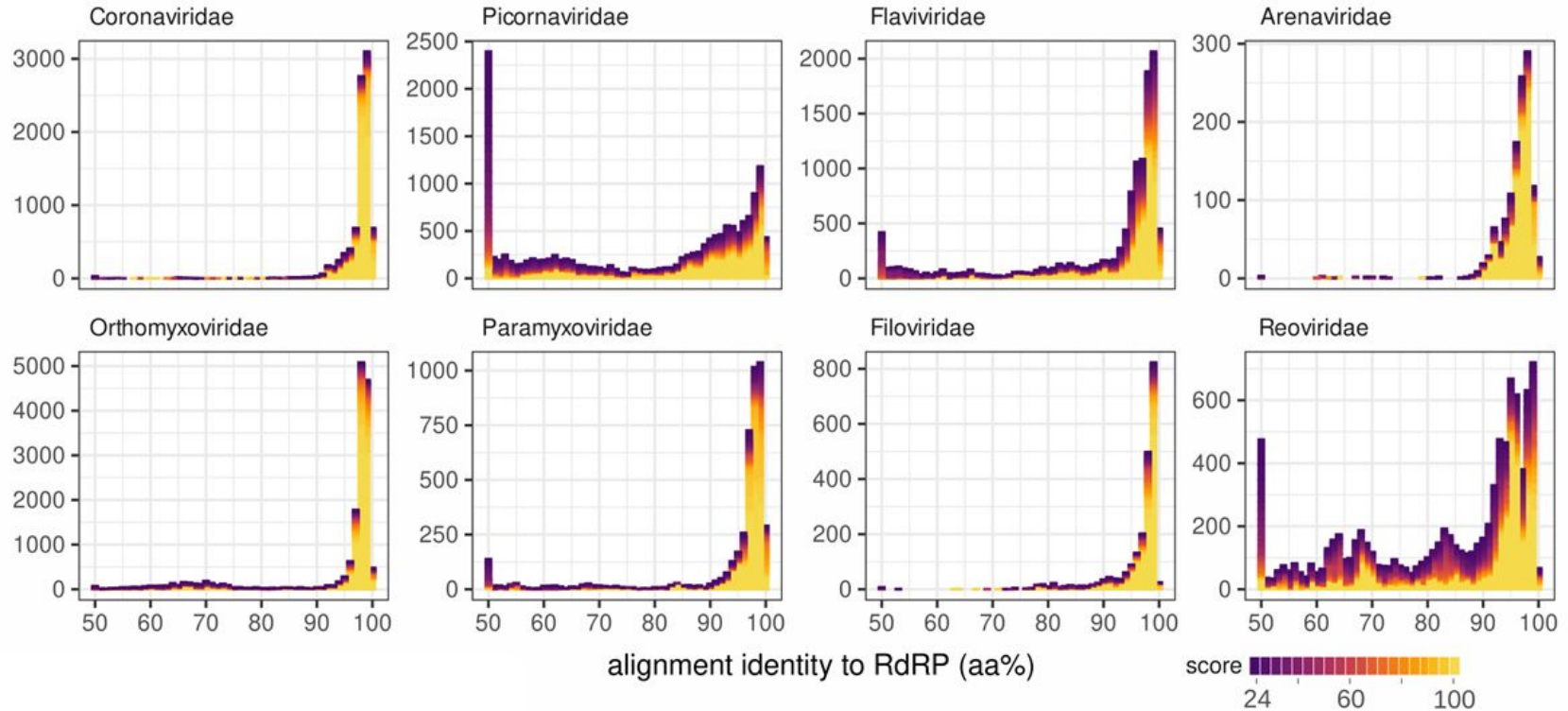
*Serratus* can process in excess of  
**1 million NGS libraries / day**  
 for a cost of  
**\$0.005 / library**

Type "petabase scale" on Google, or `www.serratus.io`

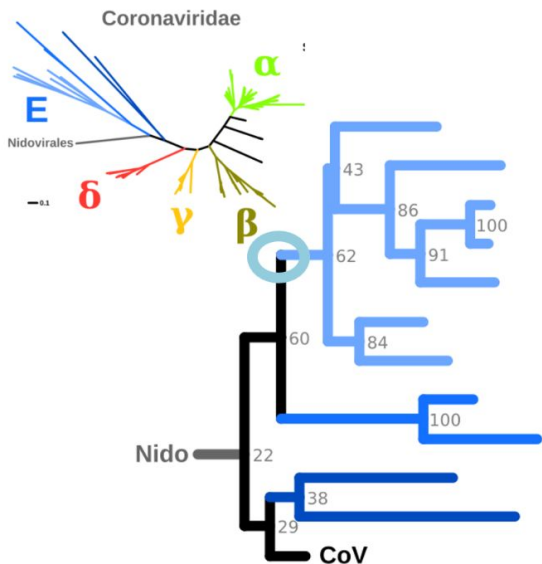




# Discovering viral species by families, by homology to known ones



# Discovering new Coronaviruses



**AmexNV** SRR6788790

**PtetNV** SRR7507741

**HkudNV** SRR1324965

**StypNV** ERR3994223

**TparNV** SRR10917299

Pacific Salmon Nidovirus

**AcaNV** SRR5997671

**SiINV** SRR12184956

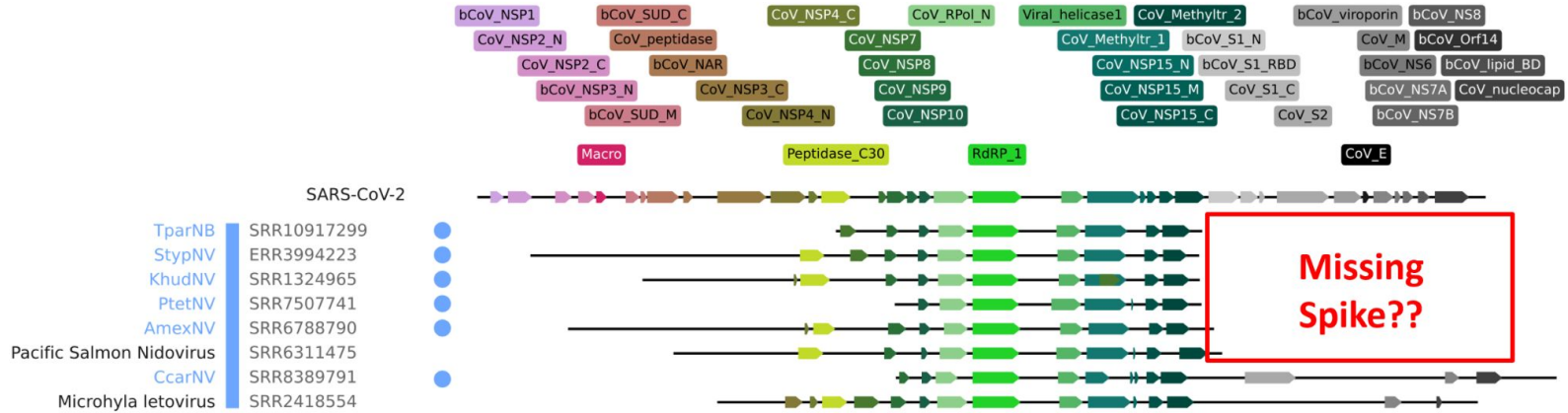
**MalbNV** SRR10402291

**HtraNV** SRR8389791

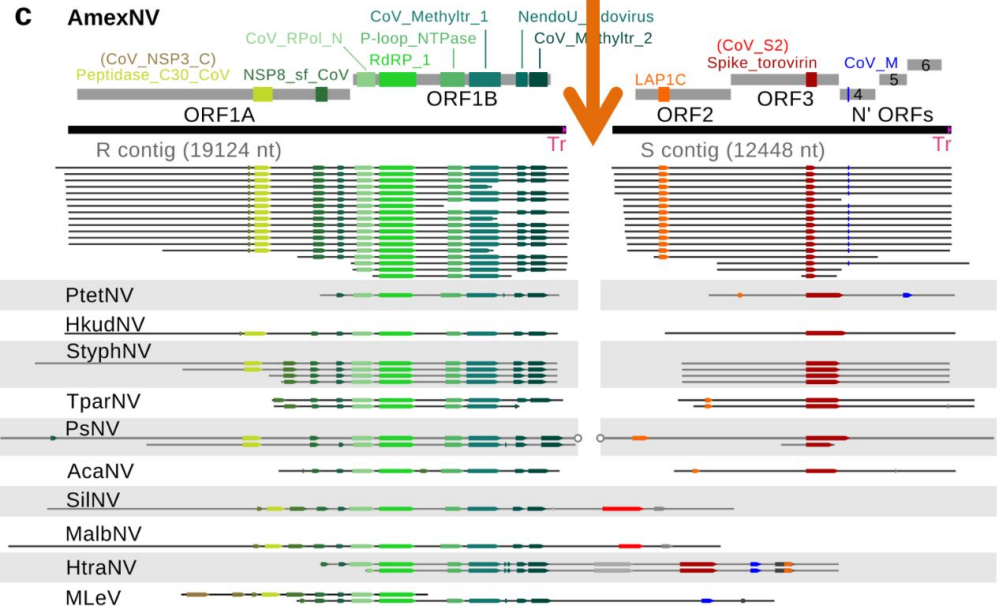
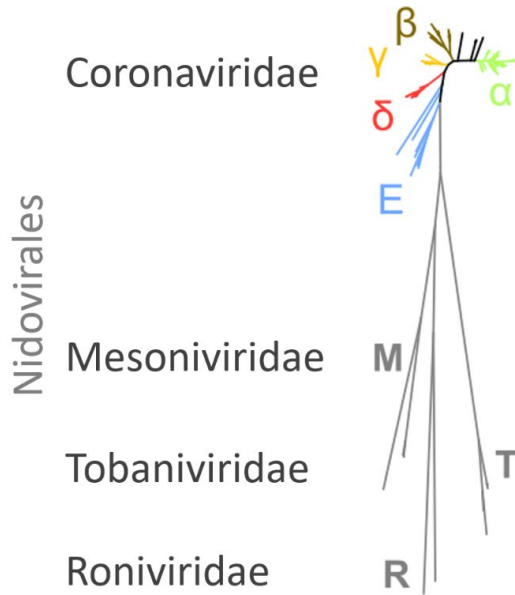
Microhyala Letovirus



# Discovering new Coronaviruses



# Segmented Coronaviruses?



Re-writing the textbook definition of a Coronavirus



# Metagenome / metavirome assembly

Usually:

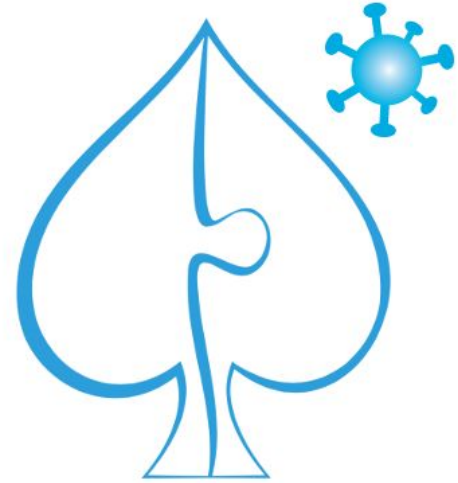
Reconstruct *all* the genomes in  
a sample

Analysis 1:

**Reconstruct CoV genome(s) in a  
sample**

Analysis 2:

**Reconstruct RdRP genes(s) in a  
sample**



SPAdes assembler

rnaSPAdes

coronaSPAdes

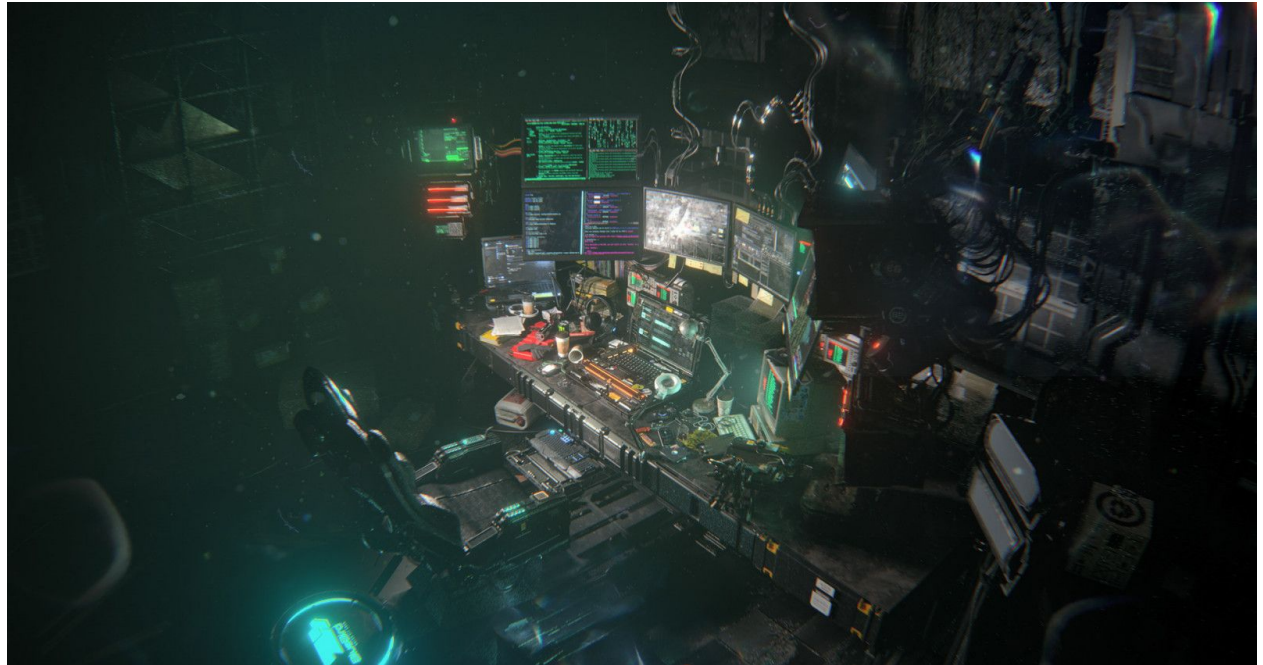
# How was all of this large-scale assembly done?





# How was all of this large-scale assembly done?

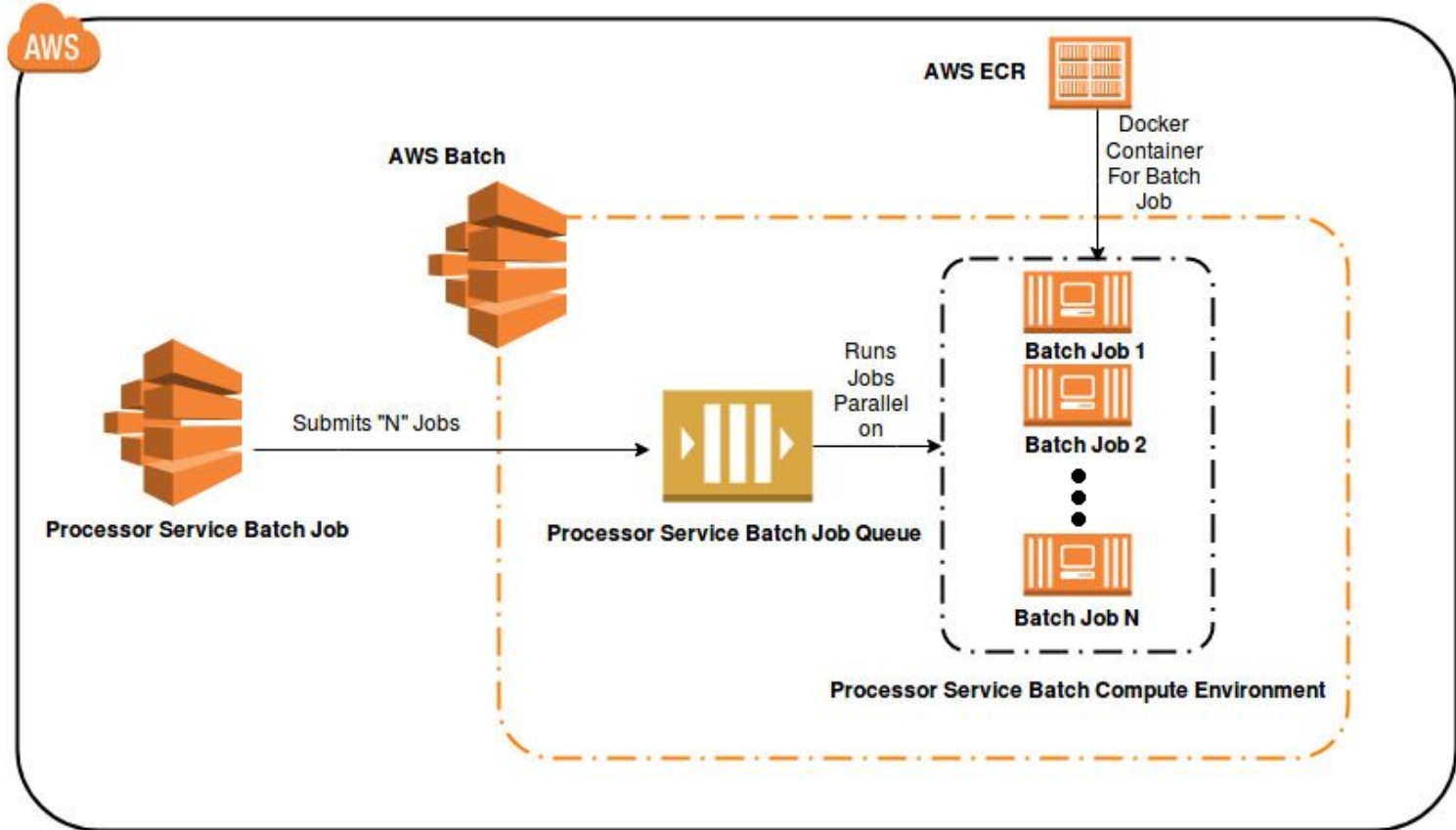
cloud scripting



\* (artist's rendition)



# AWS Batch framework for large-scale assembly



Peak:  
~28,000 vCPUs

<input type="checkbox"/>	Name	Instance ID	Instance Type	Availability Zone	Instance State
<input type="checkbox"/>	Compute	i-004fc86f836336d17	c5.9xlarge	us-east-2a	<span style="color: green;">●</span> running
<input type="checkbox"/>	Compute	i-01af64dd577f162b5	c5.9xlarge	us-east-2a	<span style="color: green;">●</span> running
<input type="checkbox"/>	Compute	i-064fe18ba8316f79f	c5.9xlarge	us-east-2a	<span style="color: green;">●</span> running
<input type="checkbox"/>	Compute	i-0879ad68f76a4a54e	c5.9xlarge	us-east-2a	<span style="color: green;">●</span> running
<input type="checkbox"/>	Compute	i-094ddc9b931fde962	c5.9xlarge	us-east-2a	<span style="color: green;">●</span> running
<input type="checkbox"/>	Compute	i-0c8f6d93593531c32	c5.9xlarge	us-east-2a	<span style="color: green;">●</span> running
<input type="checkbox"/>	Compute	i-0e08ab6c5a3d0ce3f	c5.9xlarge	us-east-2a	<span style="color: green;">●</span> running
<input type="checkbox"/>	Compute	i-0ea10648adeeabf68	c5.9xlarge	us-east-2a	<span style="color: green;">●</span> running

(screenshot: P. Barbera)

AWS Batch > Dashboard

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

## Dashboard

### Jobs overview

RUNNABLE

450

RUNNING

173

SUCCEEDED

48

FAILED

817

### Job queue overview

Job queue	SUBMITTED	PENDING	RUNNABLE	STARTING	RUNNING	SUCCEEDED	FAILED
RayanUnitigsBatchProcessingJobQueue	0	0	0	0	0	<span style="color: green;">●</span> 0	<span style="color: red;">✘</span> 0
RayanSerratusDLBatchProcessingJobQueue	0	0	0	0	0	<span style="color: green;">●</span> 0	<span style="color: red;">✘</span> 0
RayanSerratusAssemblyBatchJobQueue	0	0	450	7	173	<span style="color: green;">●</span> 48	<span style="color: red;">✘</span> 817

## But for Analysis 2..

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

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

## What's next?

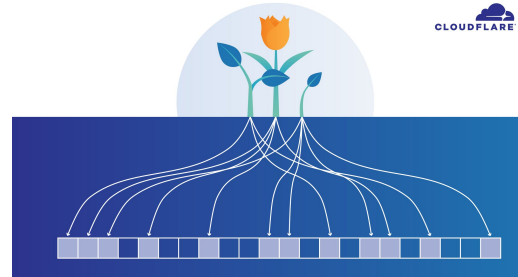
- DNA viruses
- Lower homology detection with known RdRPs
  - Replacing Bowtie 2 / Diamond by ...?
- A global index of the SRA
  - nearly feasible with k-mers already
  - would only support exact search
  - with ML, could do low(er) homologies

Deep embedding and alignment of protein sequences

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

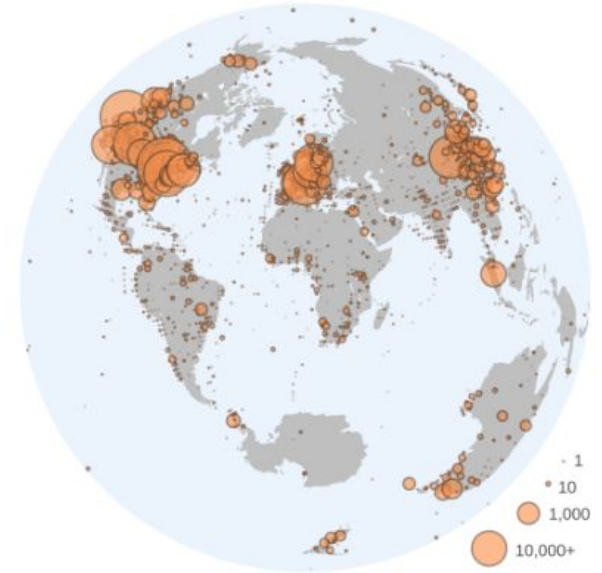
Google Research, Brain team, Paris, France

November 15, 2021



# 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](https://join.slack.com/t/hackseq-rna/shared_invite/zt-ewlzh9qf-SiNkxvVTJflcutFN0h5jIQ)

## 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](#)

**32k** Accesses | **1024** Altmetric | [Metrics](#)

### 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<sup>1</sup>. 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<sup>5</sup> 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.



hackseq

- ▲ May 22 - 24 2020
- ▲ Vancouver, BC
- ▲ 72-hour collab. RNA-hackathon
- ▲ Travel Grants

HACKSEQ RNA  
VANCOUVER - EMICS HACKATHON

Official Hackseq Vancouver

## Digital Collaboration

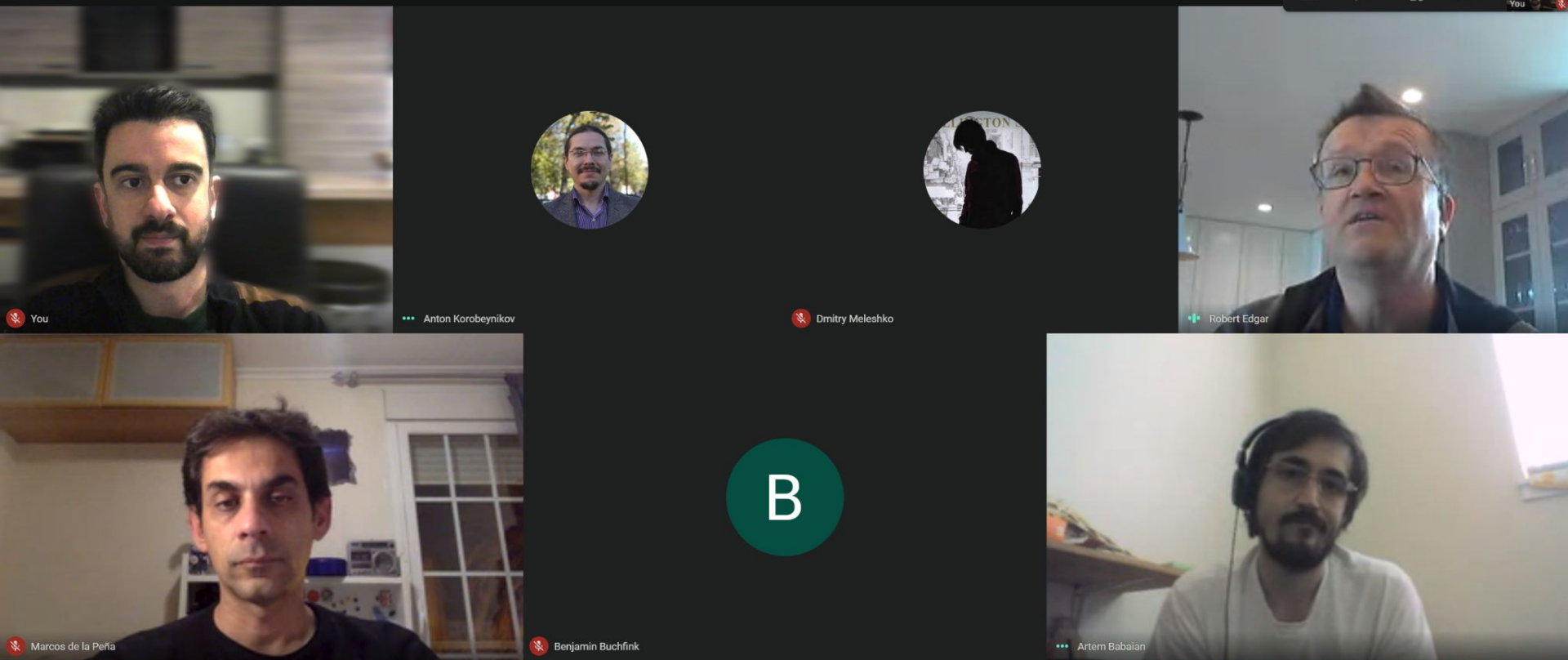
- Anton Korobeynikov (St. Petersburg)
- Artem Babaian (Vancouver)
- Basem Al-Shayeb (Berkeley)
- Benjamin Buchfink (Tubingen)
- Dan Lohr (Boulder)
- Dmitry Meleshko (Ithaca)
- Gherman Novakovsky (Vancouver)
- Jeff Taylor (Vancouver)
- Jillian F. Banfield (Berkeley)
- Marcos de la Pena (Valencia)
- Pierre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

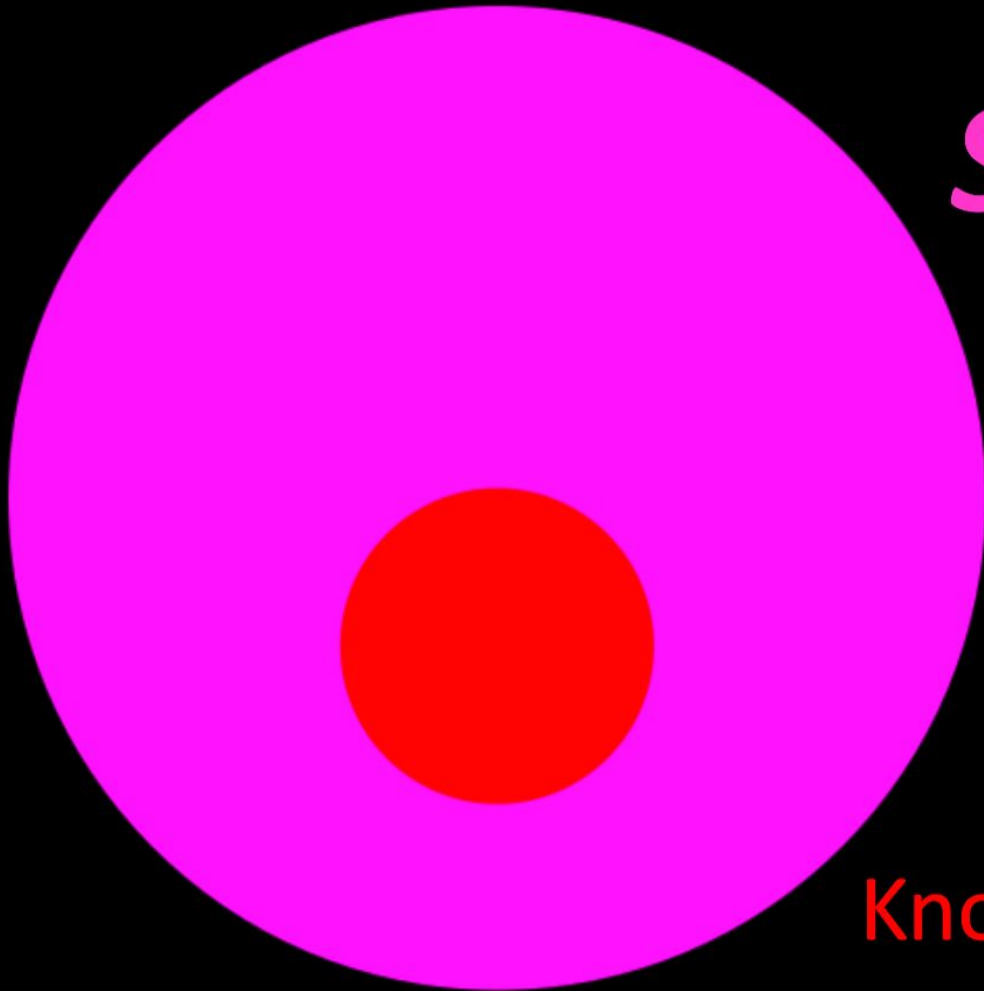
# All equal contributions





# We never met IRL





*Serratus*

Known RNA Virome

# Earth's Virome

*We are here*



**WE'RE-GONNA-NEED**



**A BIGGER INSTANCE TYPE**

Vielen Dank für ihre  
Aufmerksamkeit!

