BIG data

Rayan Chikhi

Institut Pasteur

BIG data

Rayan Chikhi

BIG data

Rayan Chikhi

Living in the future of genomics

Rayan Chikhi

Institut Pasteur

Hello! Self-intro:

- PI in bioinformatics algorithms
- Workshop on Genomics fan:
  - Genome Assembly course 2013-2019
  - Co-director 2020-2023

Research:

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

@RayanChikhi on Twitter

http://rayan.chikhi.name

Big data is a natural continuation in biology

1972: single gene sequenced

2000: 1 high-quality human genome

2011: low-quality human genomes

2021: 10 petabases of reads analyzed

2022: 1 million humans VCFs

2022: 50 high-quality human genomes

2023–: ?

The pGpOpApTp summary paragraph

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

(Received for publication, November 20, 1972)

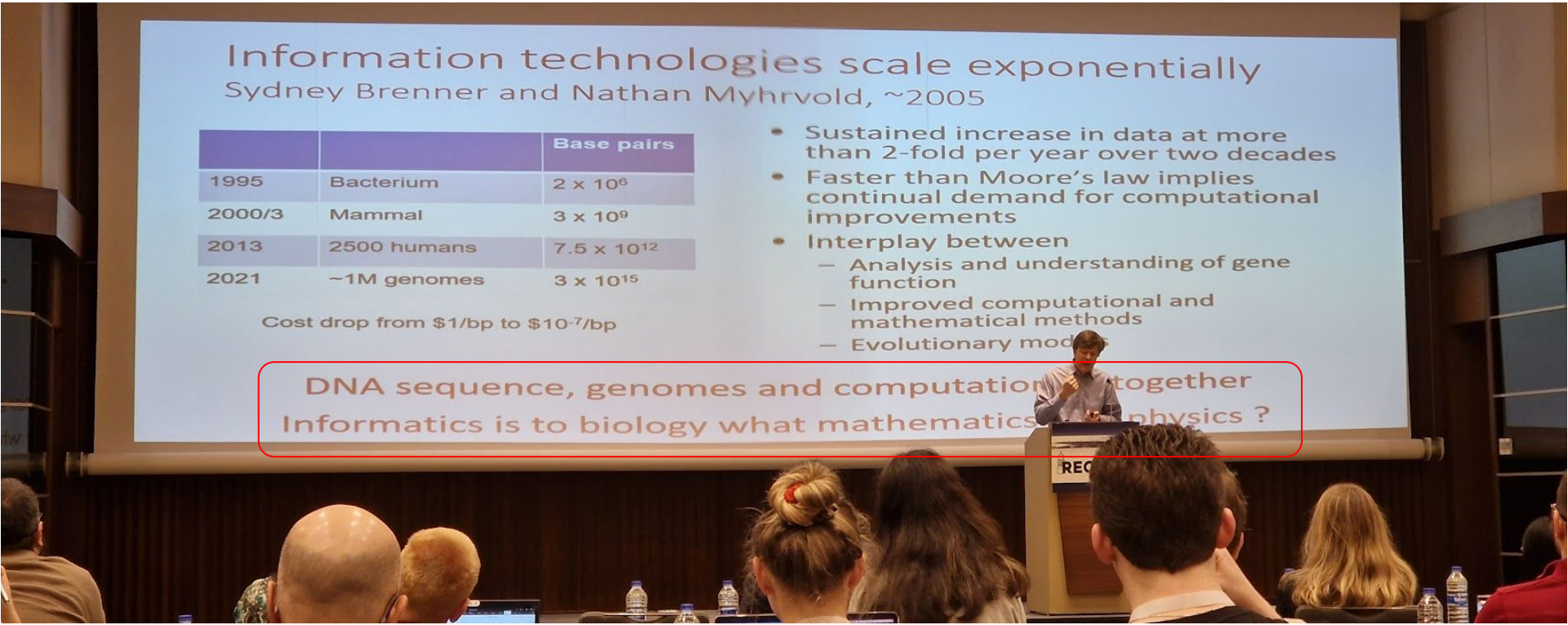
GERALD M. RUBIN*
From the Medical Research Council Laboratory of Molecular Biology, Cambridge, CB2 2QH, England

SUMMARY

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

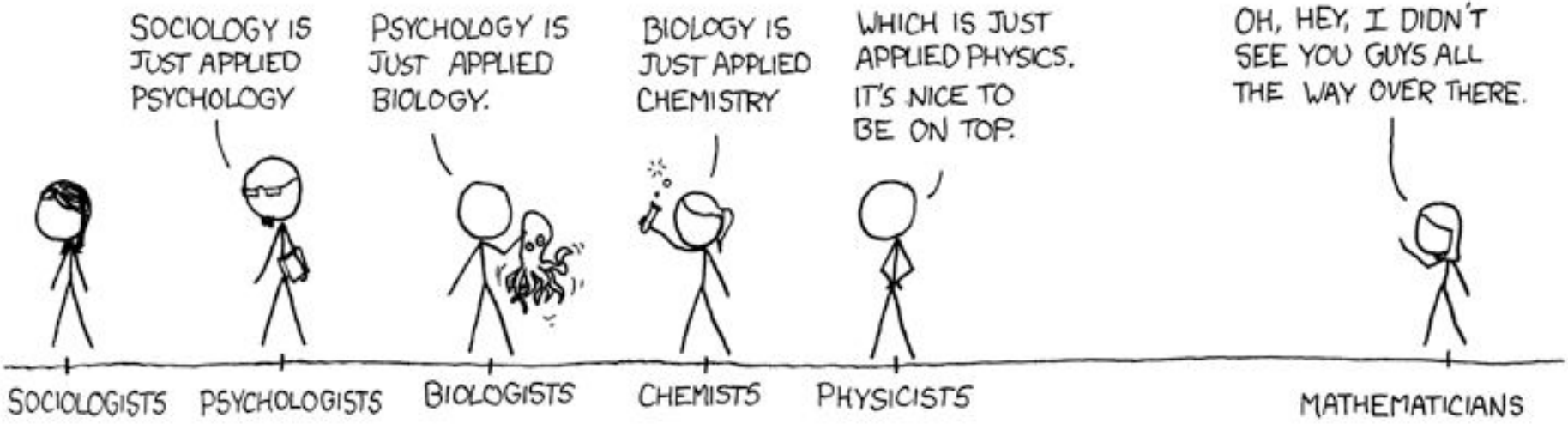
Low Phosphate Medium—Inorganic phosphate was precipitated (as $MgNH_4PO_4$) from 10% Bacto-yeast extract and 20% Bacto-peptone by the addition of 10 ml of 1 M $MgSO_4$ and 10 ml of concentrated aqueous ammonia per liter. The phosphates were allowed to precipitate at room temperature for 30 min, and the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

Credit: @SynBio1



"Informatics is to biology,
what mathematics is to physics"

Richard Durbin, RECOMB 2023 keynote



“purity” →

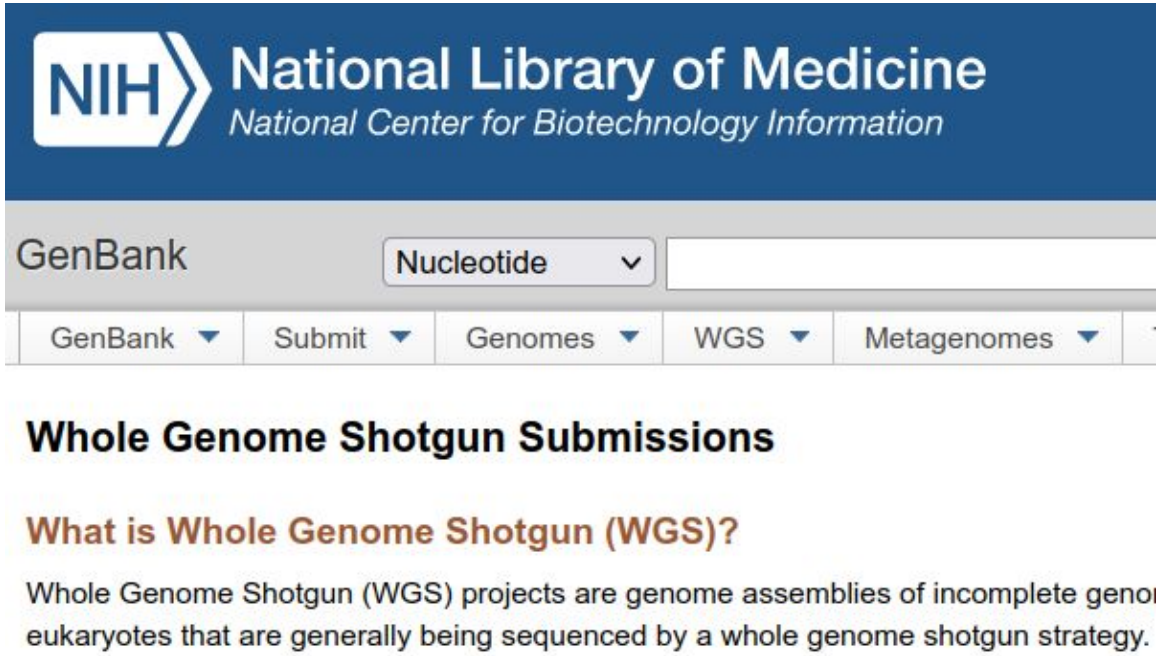
“usefulness” ←

Big data in biology: GenBank

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

Particularity: all sequences are *annotated*

NCBI WGS



Type: assemblies
Size: 16 TB (April 2022)

Difference with GenBank: sequences are not necessarily annotated

NCBI SRA

Size: 30 PB

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

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

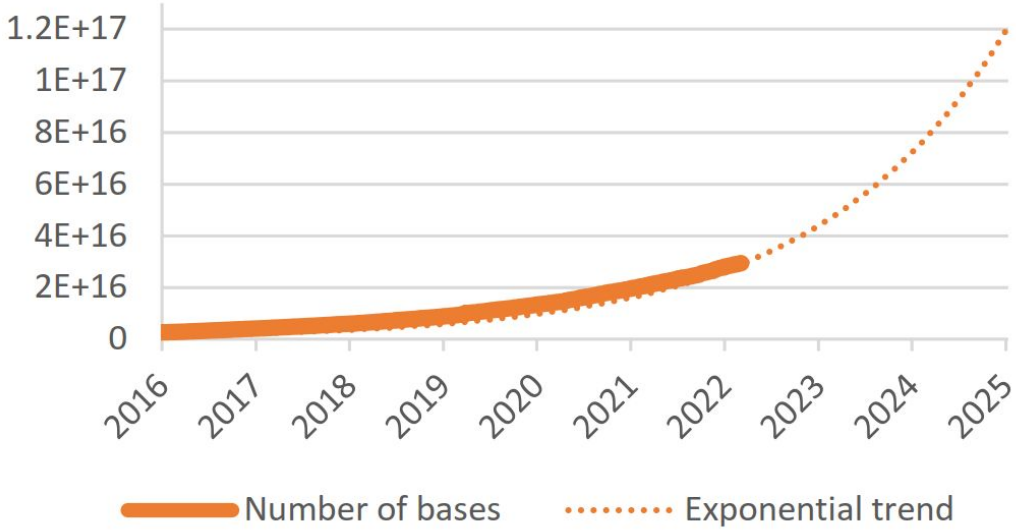
YouTube: 100-1000 PB

NCBI SRA database : 30 PB

Institut Pasteur: 8 PB

Your laptop: 0.001 PB

Growth of the Sequence Read Archive



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

… But how?

People at the leading edge of a rapidly changing field "live in the future."

- Paul Buchheit (GMail creator)

You want to know how to paint a perfect painting? It's easy. Make yourself perfect and then just paint naturally.

- Robert Pirsig (Philosopher, 1928-2017)
In: Zen and the Art of Motorcycle Maintenance

Credit: Paul Graham

“Living in the future” in biology?

- Have a lab technique only a few know
- Have data that will only be public later
- Work on “sci-fi” projects (e.g. creating a cell from scratch)
- ...

“Living in the future” in ~~biology~~ bioinformatics

- Have a ~~lab~~ computational technique only a few know
- Have data that will only be public later
- Work on “sci-fi” projects

  (e.g. analyze data so big no-one would believe it can be done)

How are some people living in the future?

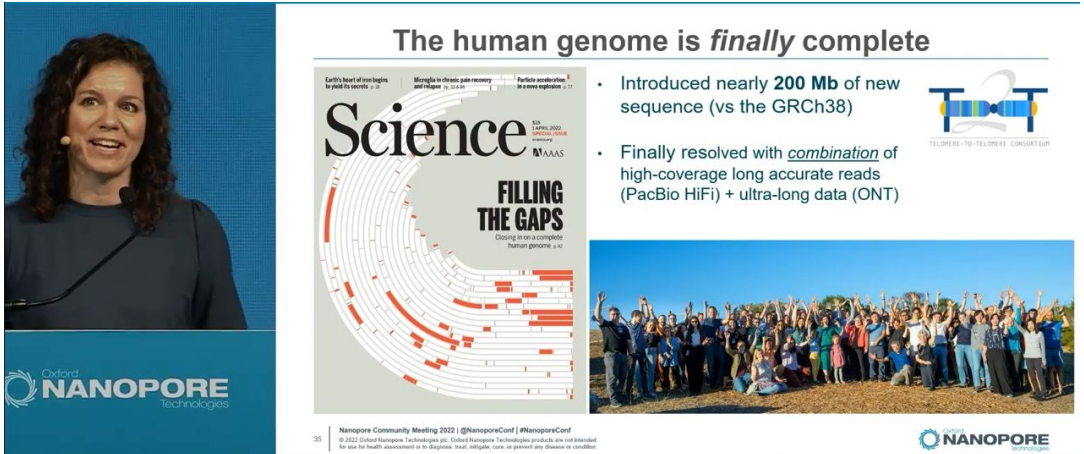
- George Church, Craig Venter
- Karen Miga & T2T team*
- Evan Eichler
- Erik
- **ALL OF YOU****

MICROBIOLOGY

Scientists Create the Smallest-Ever Moving Cell

Just two genes get tiny synthetic cells moving, offering clues to life's evolution

By Saugat Bolakhe on April 1, 2023



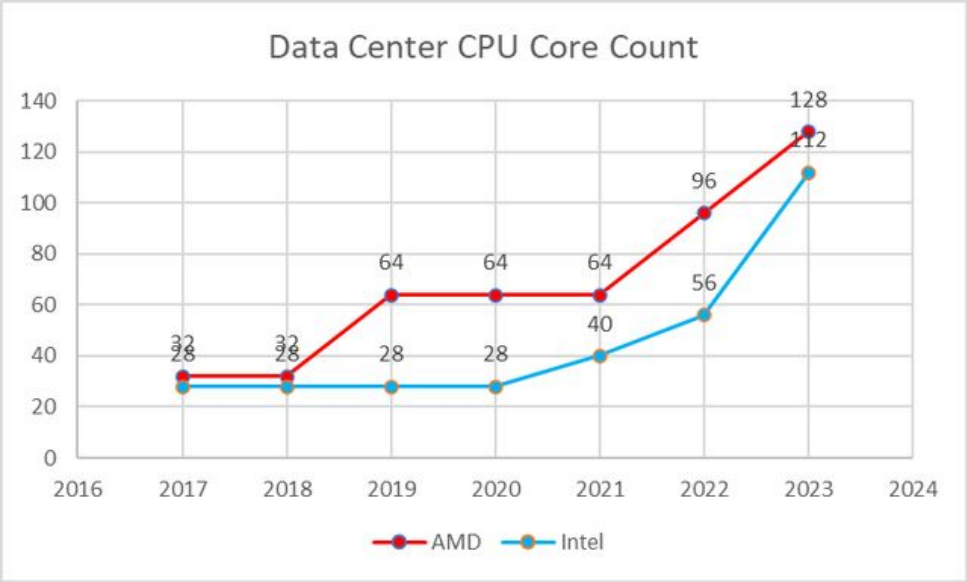
* While the rest of the world still uses GRCh38/hg19

** Generally ~months ahead of the present with your research

Future genomics, today?

Using “future” computers!

A small demo



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

Part 1: Getting reads

```
@SRR11606871.1 1 length=425
CCGGGATGTGCTTGCCTTTCGGCACCATGTA
CTGGATGCCAAAGAACACGGTGCCGTTATCC
TACCGCTCATGAAGTCACGGGGCTGATTCGT
AAGGCGCGACGGTGTGCGGTGTTCGTGTACG
AACCATCTCACCGGCGAAACTCAGGCCCTTC
TGCACCTGTCGTGGTTAATGCCGCTGGGATC
GGGGGCAACACATTGCCGAATATGCCGATCT
CGCATTCGCATGTTCCCGGCGAAAGGATCGC
GCTGATCATGGATCACCGCATTAACCAGCAT
TGATCAACCGCTGCCGTAAACC
```

Part 1: Getting reads

Nowadays: Downloading reads to your laptop / cluster

```
# human CHM13 HiFi on us-east-1, 10x coverage
s3://sra-pub-src-2/SRR11292120/m64062_190806_063919.fastq.1
```

(demo)

Live: Demo of downloading reads at 1 MB/sec

Part 1: Getting reads

Future: *Locally sourced, homegrown, data ©*

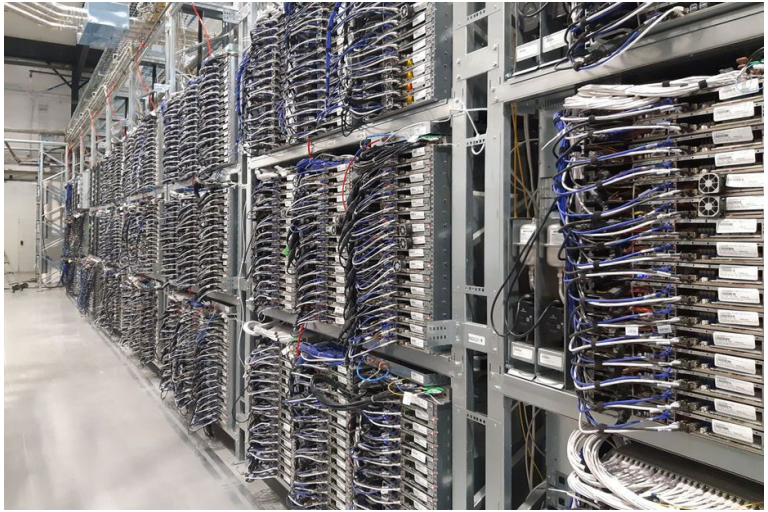
i.e. reads do not leave their host country. You go to them.

How? By renting computers in the same datacenter.

Live: Demo of downloading reads at 400 MB/sec

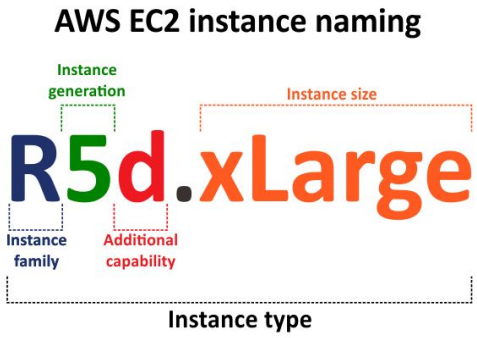
Cloud

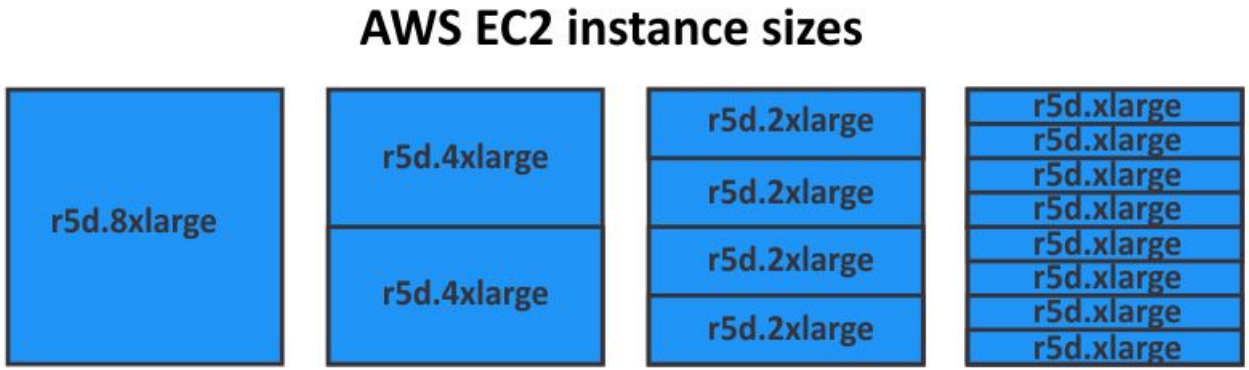
= A collection of computers owned by a single organization and accessible from the Internet



OVHcloud, Roubaix, France

Your workshop instance: `t3a.large` : 2 CPU cores, 8 GB memory





💕 c6a.48xlarge 💕 : 192 cores, 384 GB mem

What is *nearly* a cloud

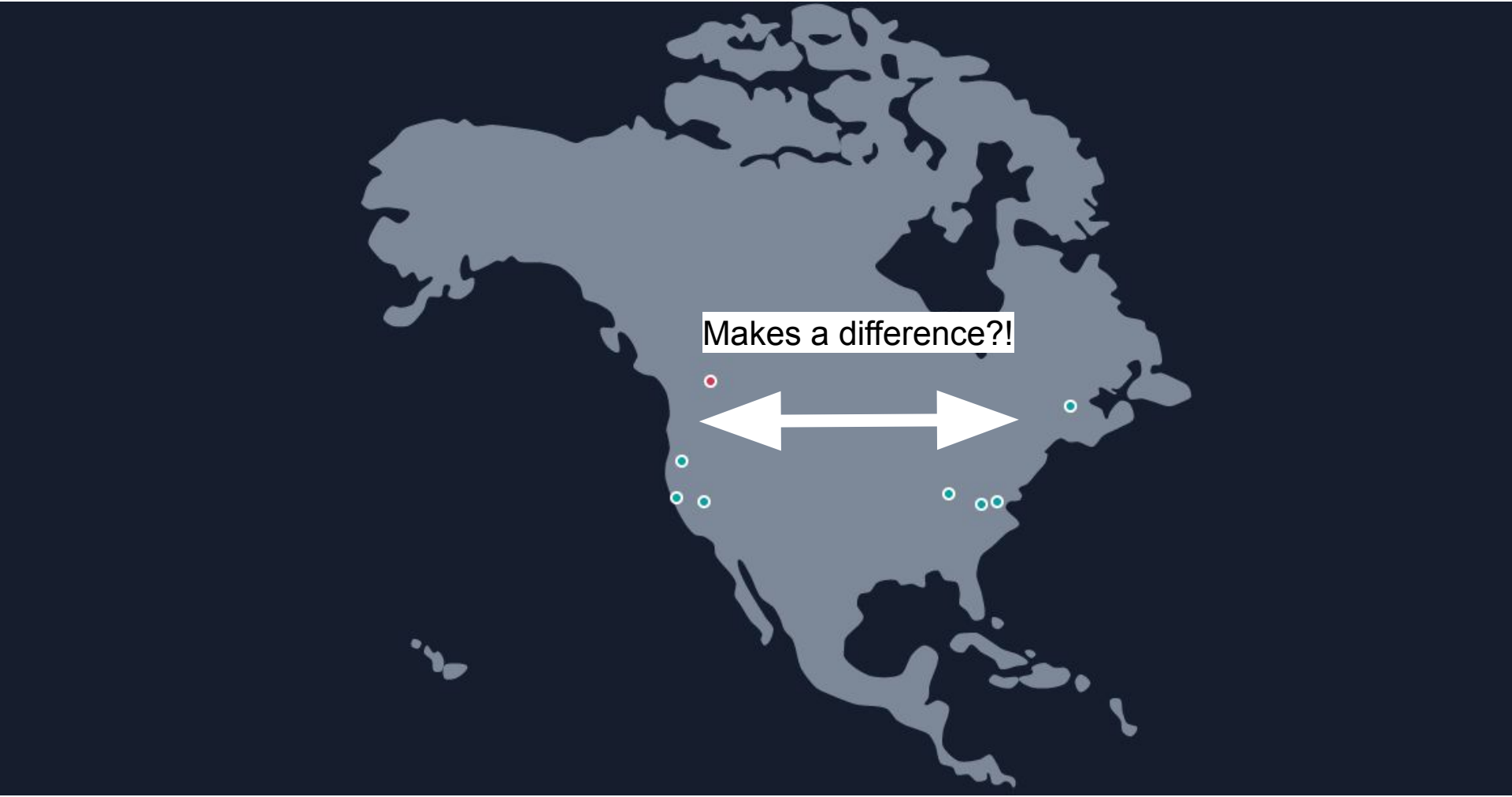
- (This dog)

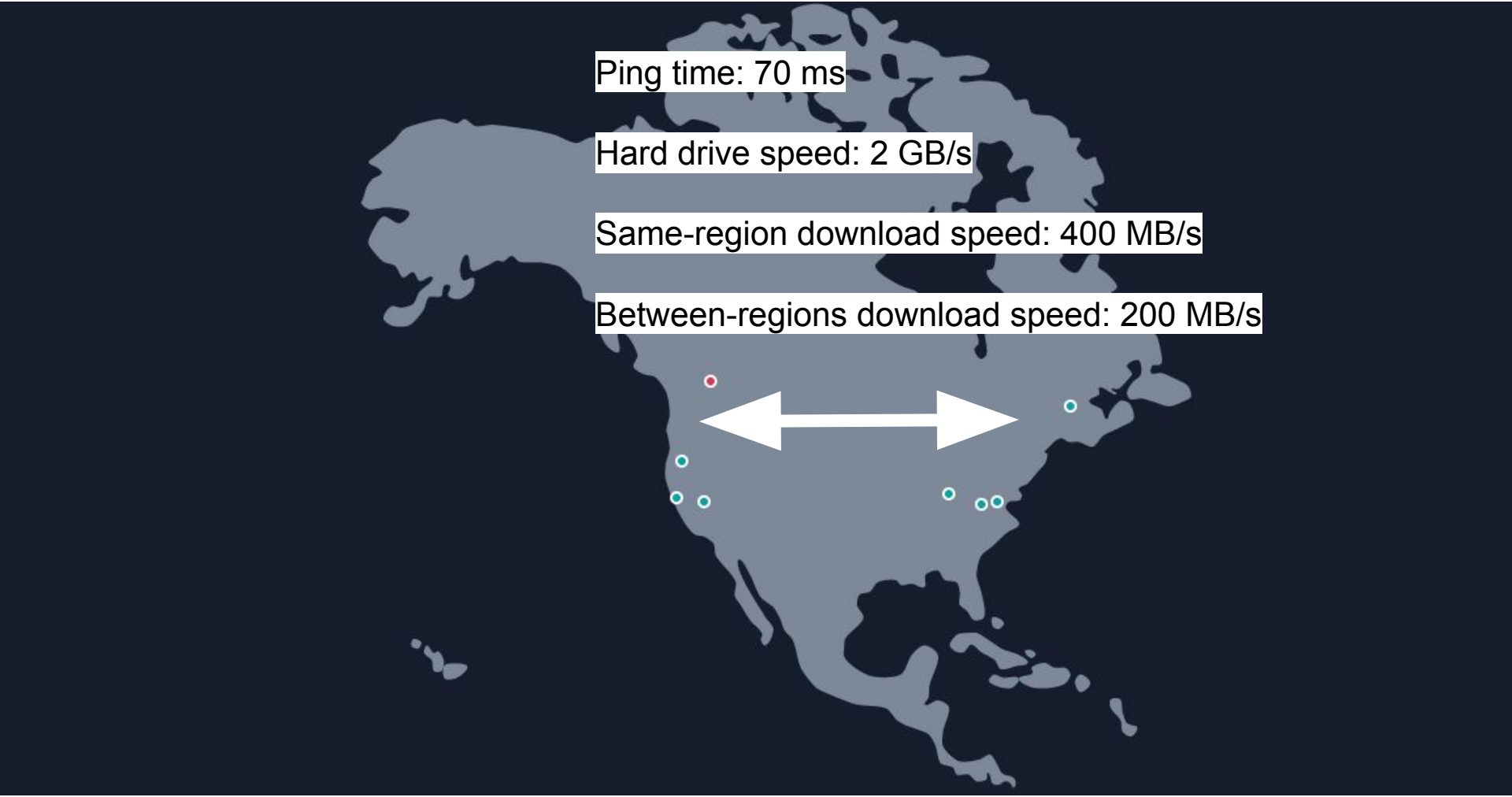
- Your university cluster
- Your 2-week access to the Workshop on Genomics 2022's resources
- 7 Raspberry Pi's stacked together

“Storing information in the cloud”?

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


us-west-1
T2T data
NCBI reads data
us-east-1


Makes a difference?!


Ping time: 70 ms
Hard drive speed: 2 GB/s
Same-region download speed: 400 MB/s
Between-regions download speed: 200 MB/s

Connect the dots from left to right

1) Read a small file from disk

2) Access data in memory

3) Open a web page from Australia

4) Human cell cycle

5) Align 1 million short reads

- 0.1 microsecond
- 0.1 millisecond
- 200 milliseconds
- 10 seconds
- 24 hours

-	-	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

Connect the dots from left to right

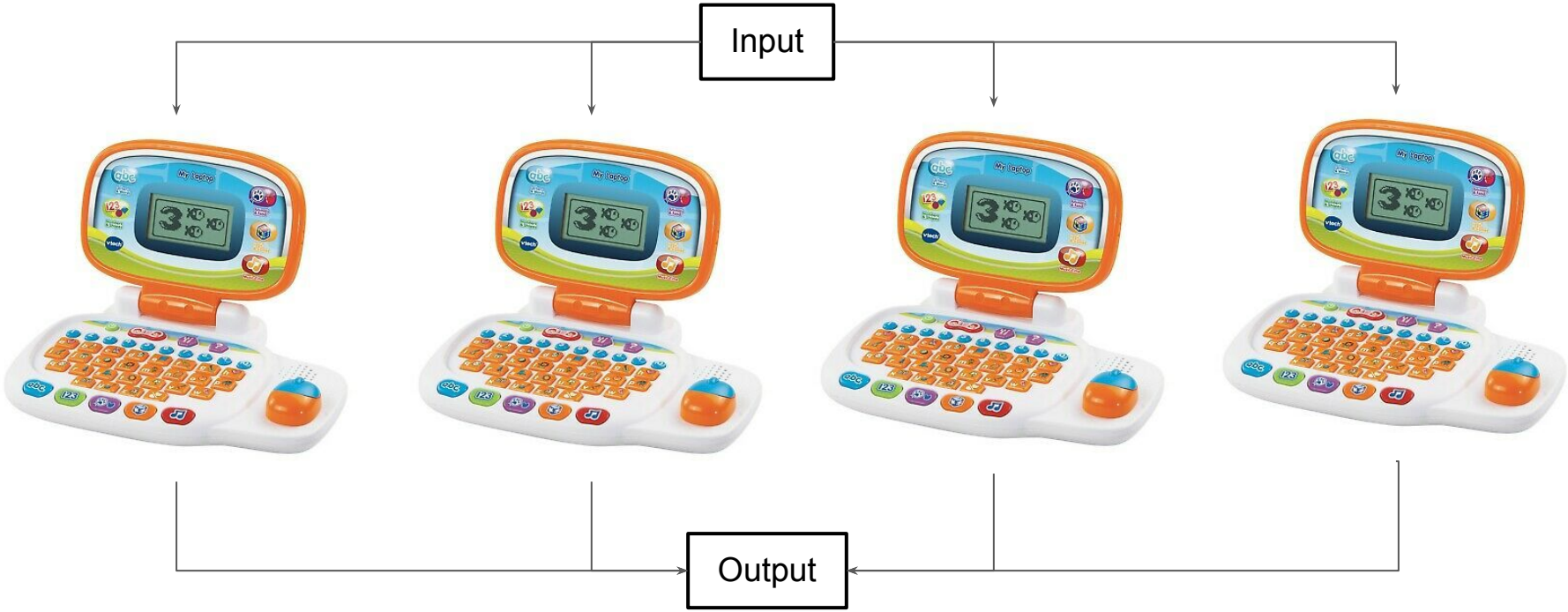
- Read a small file from disk
- Access data in memory
- Open a web page from Australia
- Human cell cycle
- Align 1 million short reads

- 0.1 microsecond
- 0.1 millisecond
- 200 milliseconds
- 10 seconds
- 24 hours

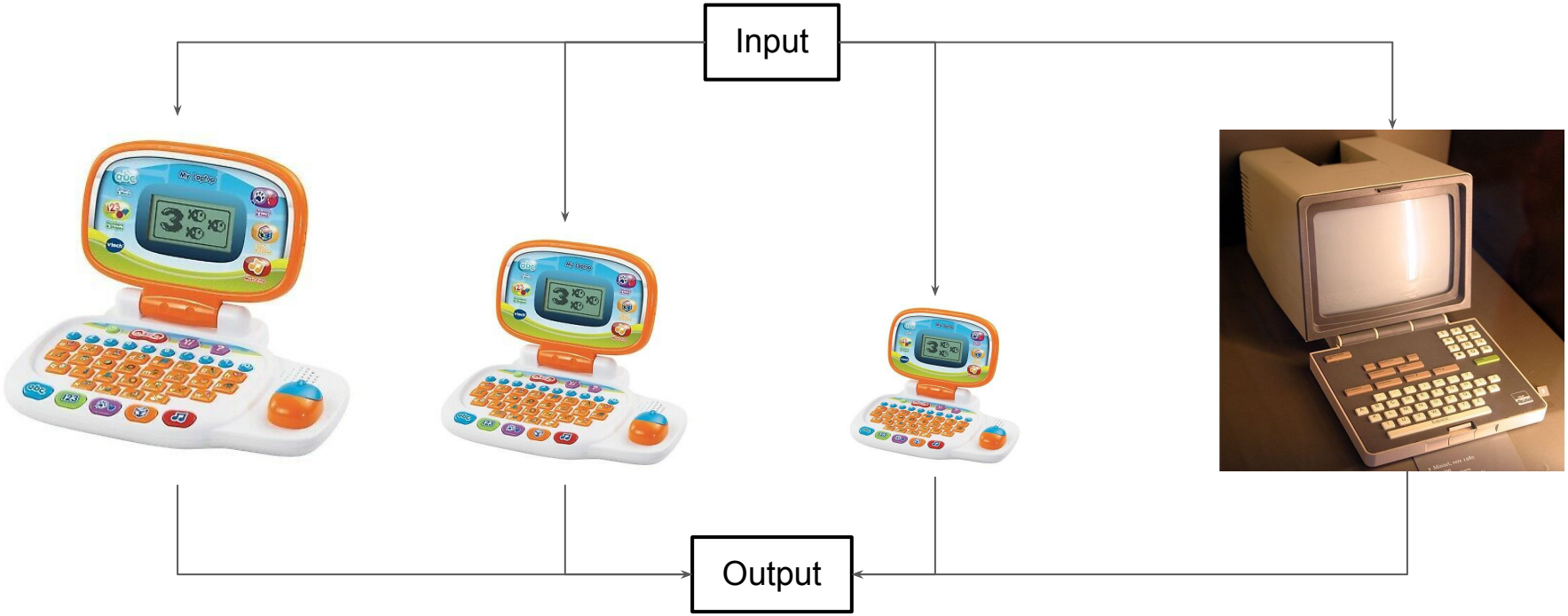
-	-	10^{0}	1
deci	d	10^{-1}	0,1
centi	c	10^{-2}	0,01
mili	m	10^{-3}	0,001
micro	µ	10^{-6}	0,000 001
nano	n	10^{-9}	0,000 000 001
pico	p	10^{-12}	0,000 000 000 001

Are 200 CPUs 200x faster than 1 CPU?

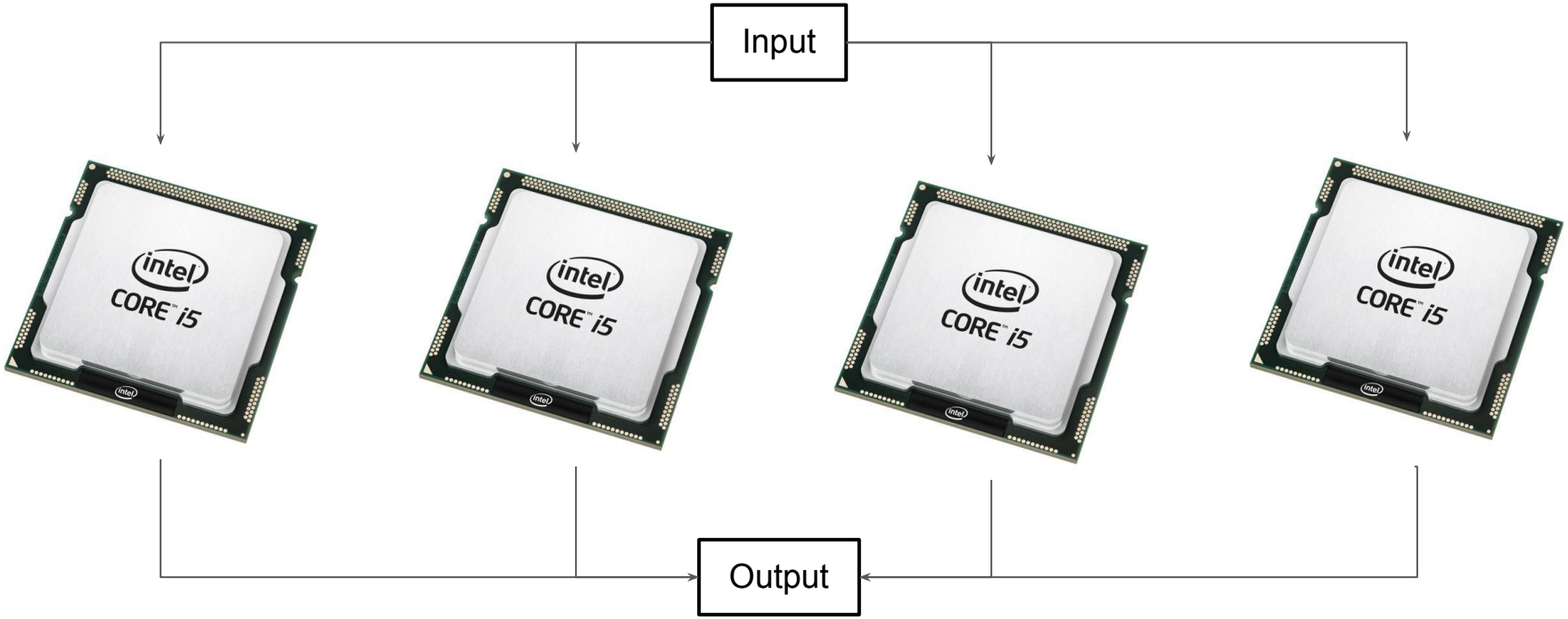
Parallelism: use many “computers” to execute one task



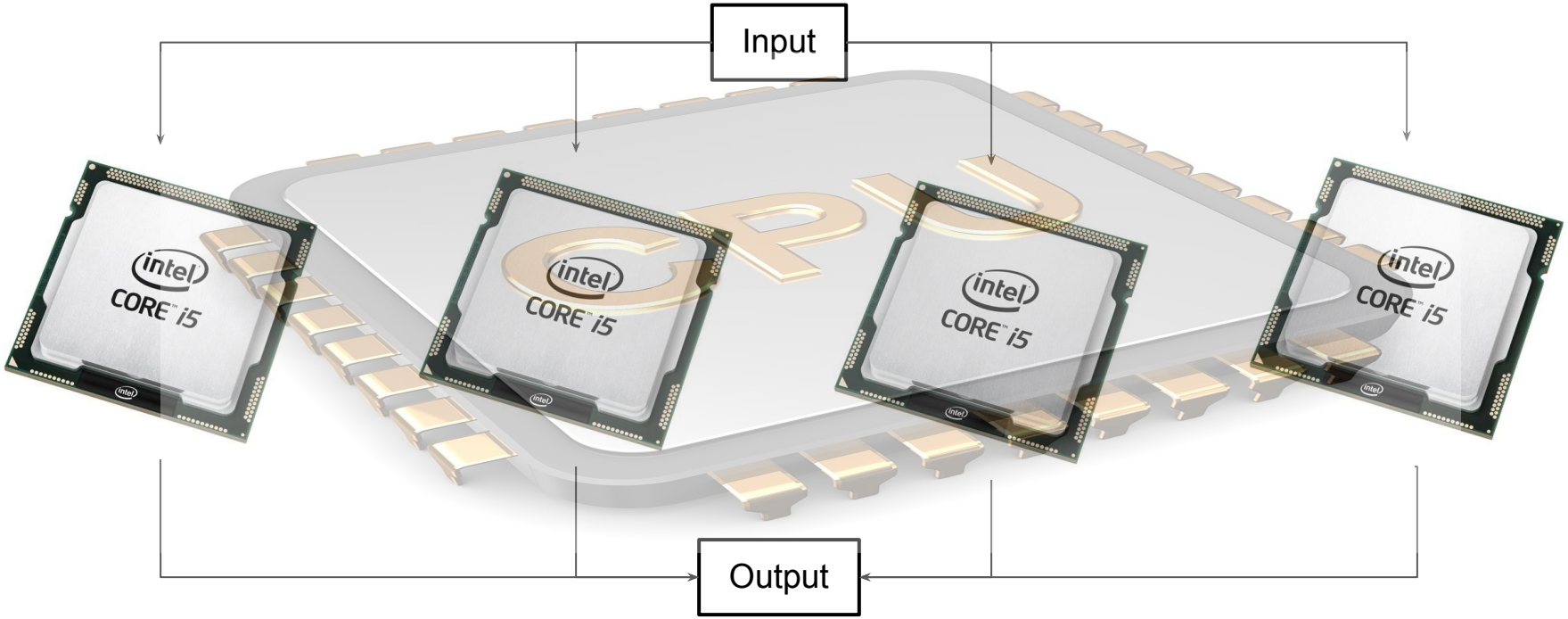
Parallelism: they don’t need to be identical computers



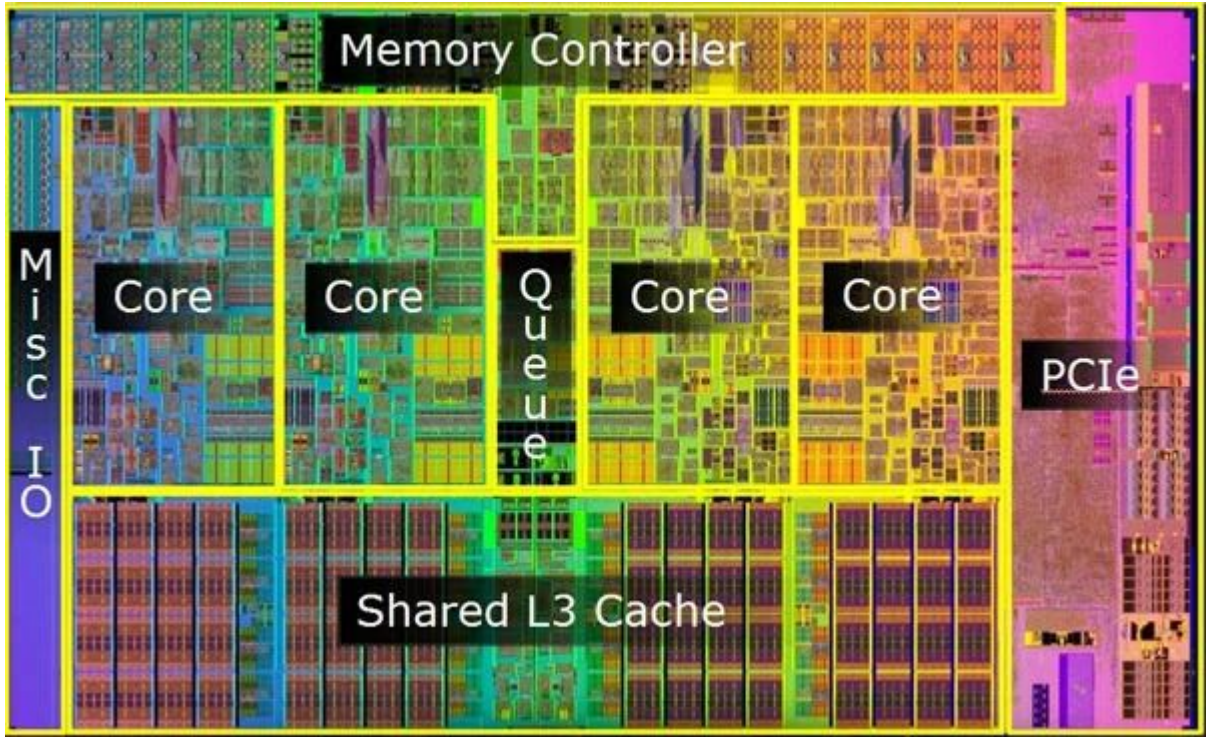
Parallelism: they don’t even need to be “computers”



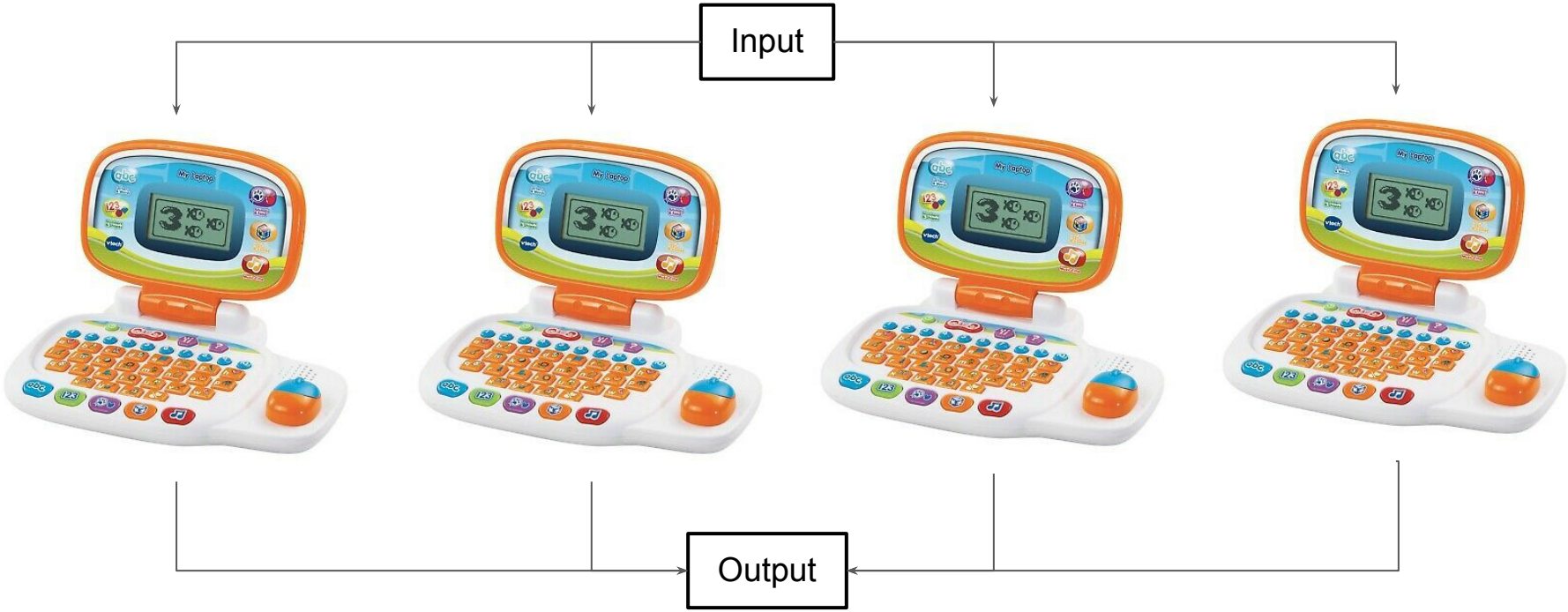
Parallelism: they don’t even need to be “computers”



Parallelism: CPU = many little computers in parallel



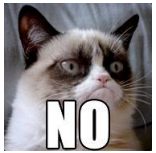
CPU (simplified)

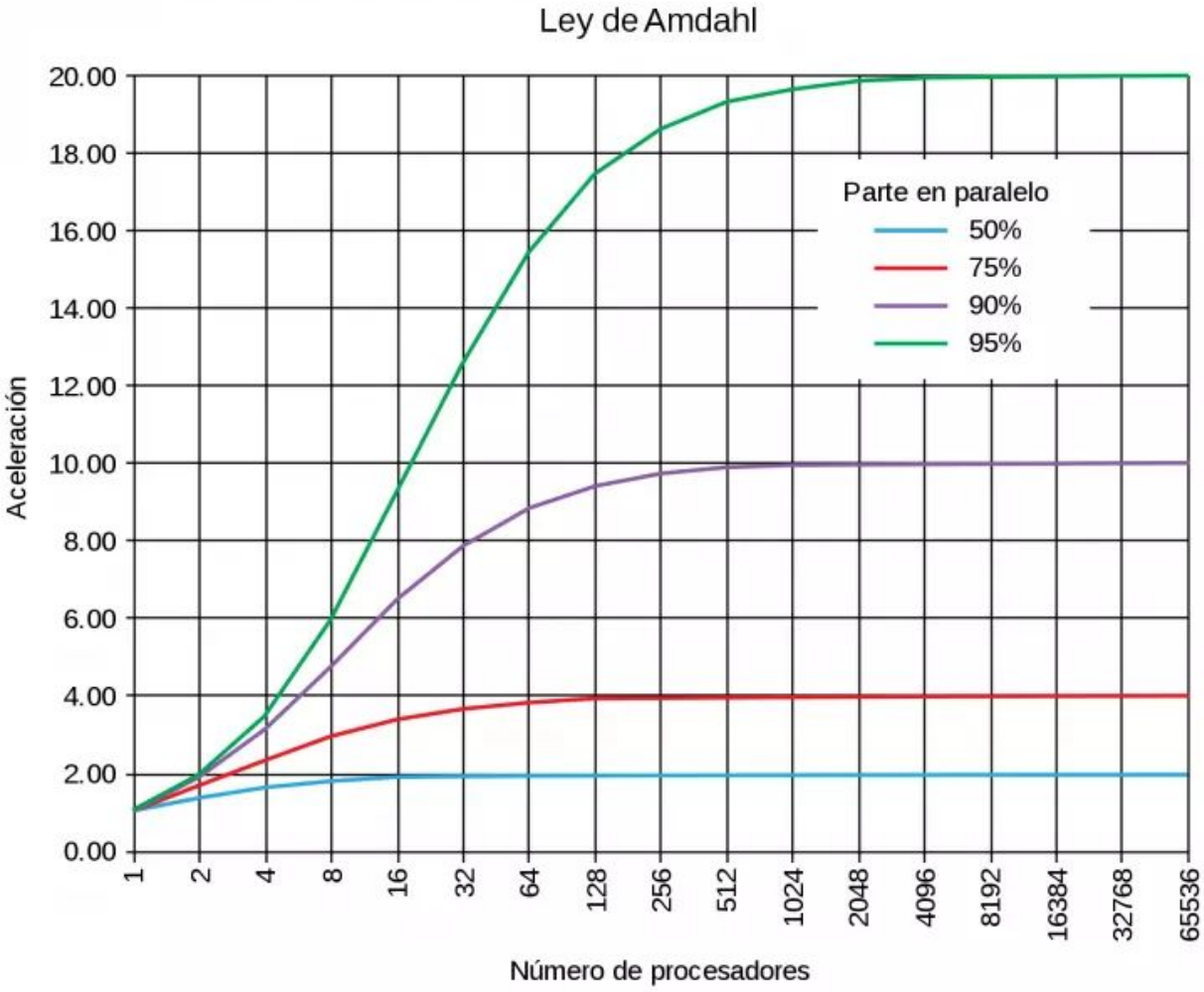


The limits of computing

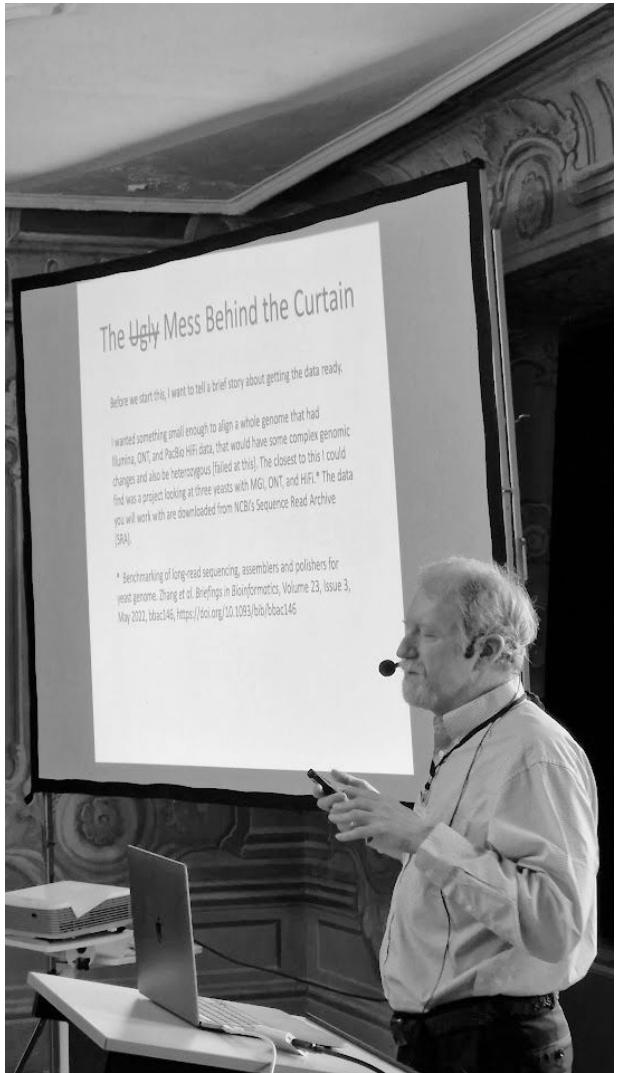
So, can we speed up indefinitely by stacking computers (or CPUs)?

Amdahl's law:


NO


Ley de Amdahl
Aceleración
20.00
18.00
16.00
14.00
12.00
10.00
8.00
6.00
4.00
2.00
0.00
1
2
4
8
16
32
64
128
256
512
1024
2048
4096
8192
16384
32768
65536
Número de procesadores
Parte en paralelo
50%
75%
90%
95%

“We cannot map a human genome in the time it takes to do a workshop“



Part 2: minimap2 on steroids

```
\time  minimap2                          \
       -t 192 -x map-hifi                \
       chm13v2.0.fa                      \
       m64062_190806_063919.fastq.1      \
       > all.sam
```

Live: Demo of mapping human 10x coverage HiFi reads using minimap2 in 2 minutes using 192 cores

T2T genome:

https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/CHM13/assemblies/analysis_set/chm13v2.0.fa.gz

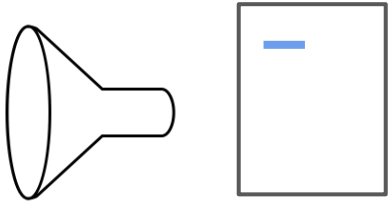
Part 3: map quickly!

https://github.com/ekimb/mapquik

Live: Demo of mapping human 10x coverage HiFi reads using mapquik in <20 seconds, including FASTA conversion using seqkit and chatgpt

My “future” project: Searching all of life’s sequence data

SRA



Index

=

Assemblies of all of ENA, searchable

(doesn’t exist - want to create it)

(5 years erc project)

ACTGATGGTG?
GTGAATGG?
AAAAAAAAAA?

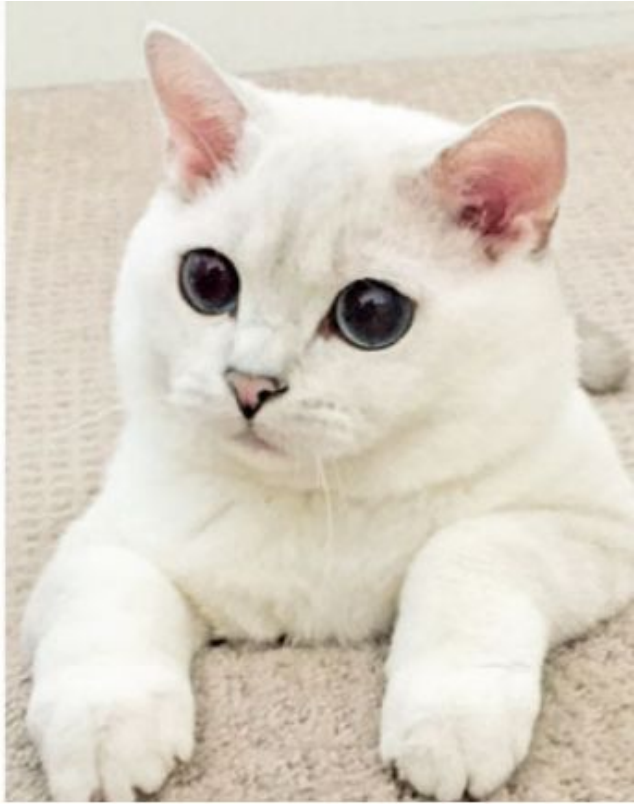
Web pages

Google:

Search query

ty Joan

Any questions so far? Coffee break?

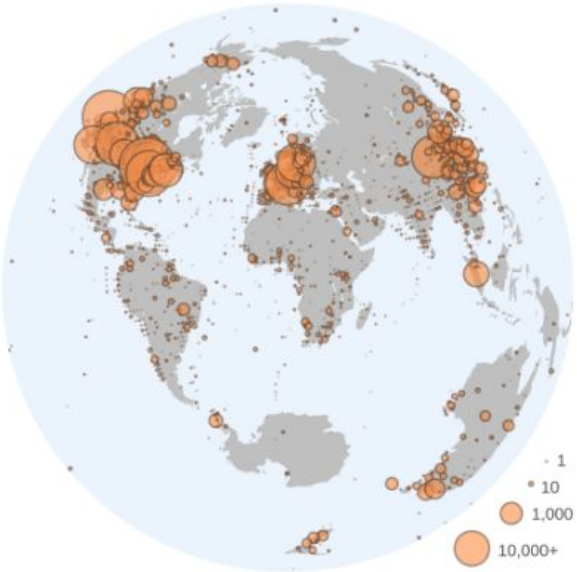


Hey don’t leave, there is a part 2!!!

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.



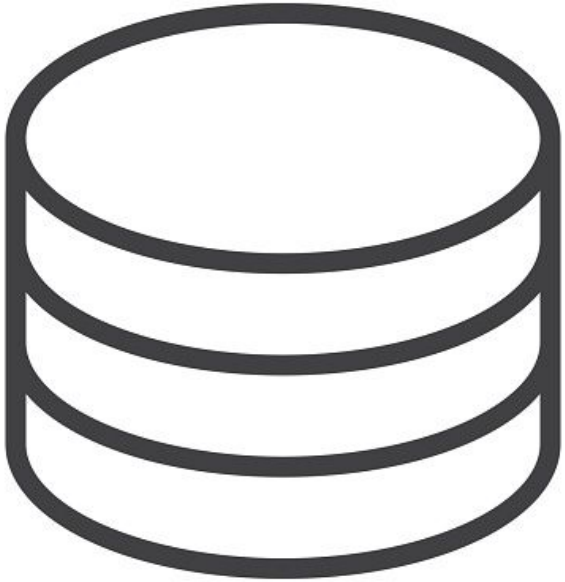
cnrs PR[AI]RIE PaRis Artificial Intelligence Research InstitutE INSTITUT PASTEUR https://serratus.io/

NCBI SRA database : 30 PB

NCBI SRA database : 30 PB

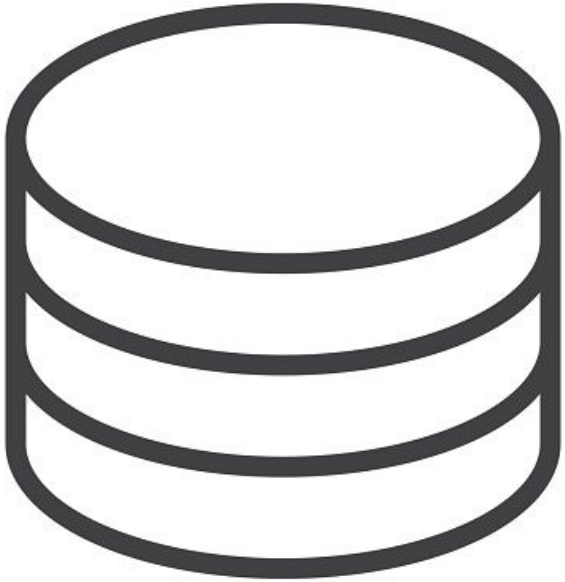
<u>Data crypt</u>

Reads just sleep, undisturbed



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

Downloading all
RNA-seq samples:

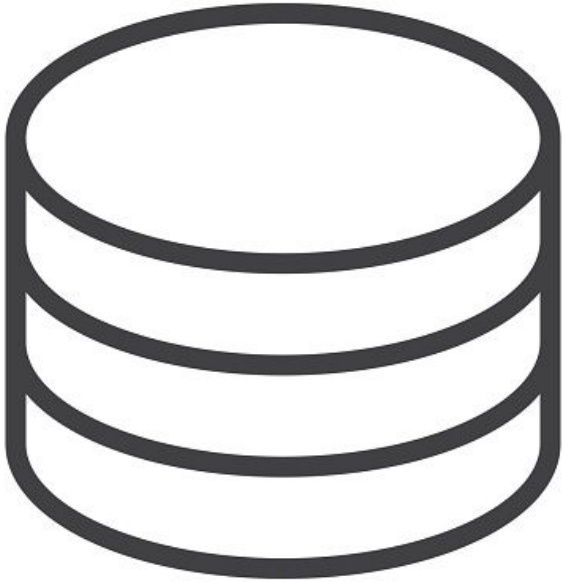


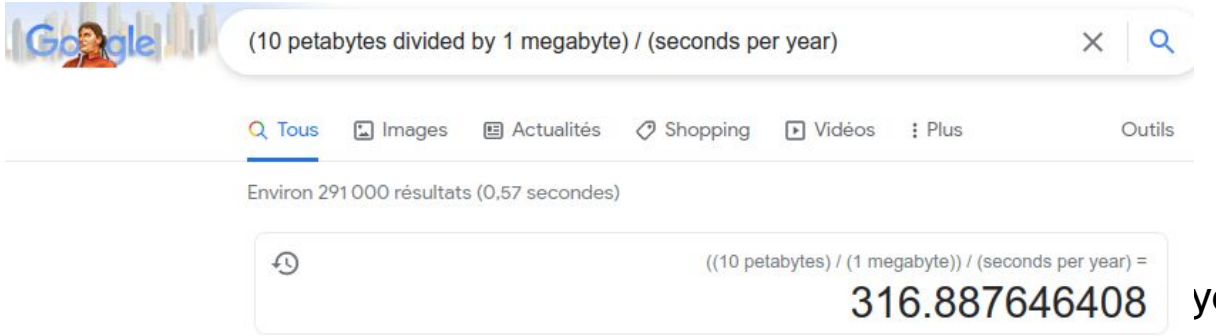
<u>Guesstimate</u>:

- 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

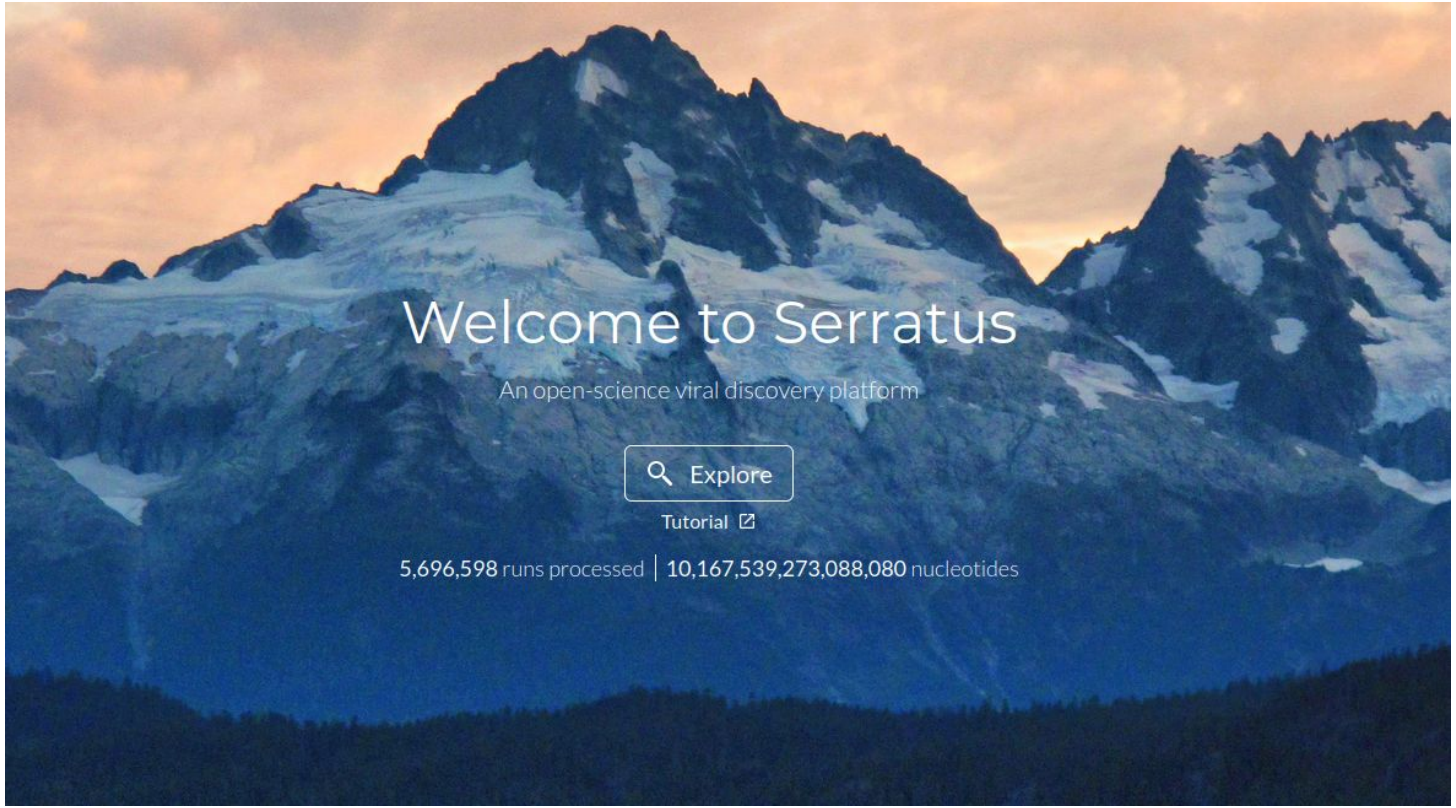
Downloading all RNA-seq samples:





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

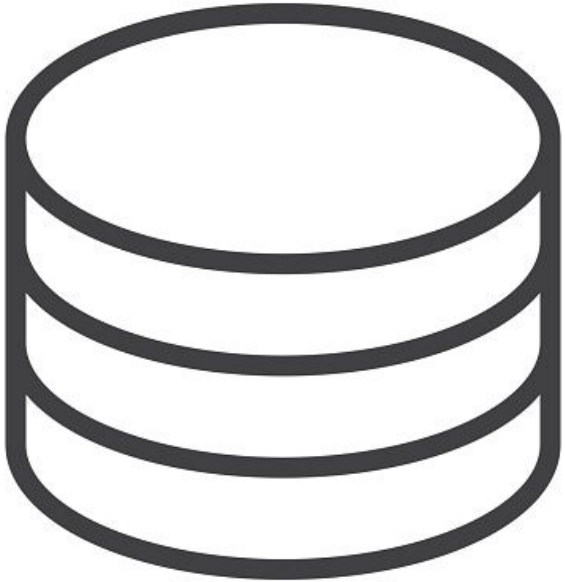
    > Discovered new coronaviruses

2) Protein (translated) alignments

    all RNAseqs vs a universal RNA virus gene

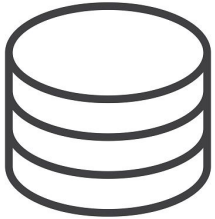
    > Discovered 130,000 new RNA virus species

Analysis 1:



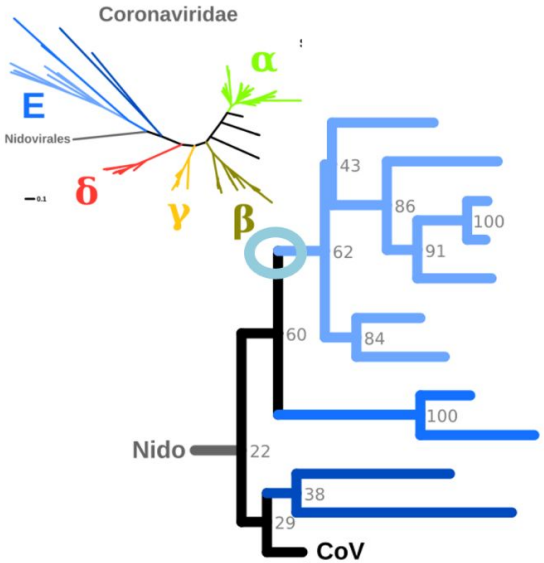
All RNA-seqs

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



55,715 CoV+ samples

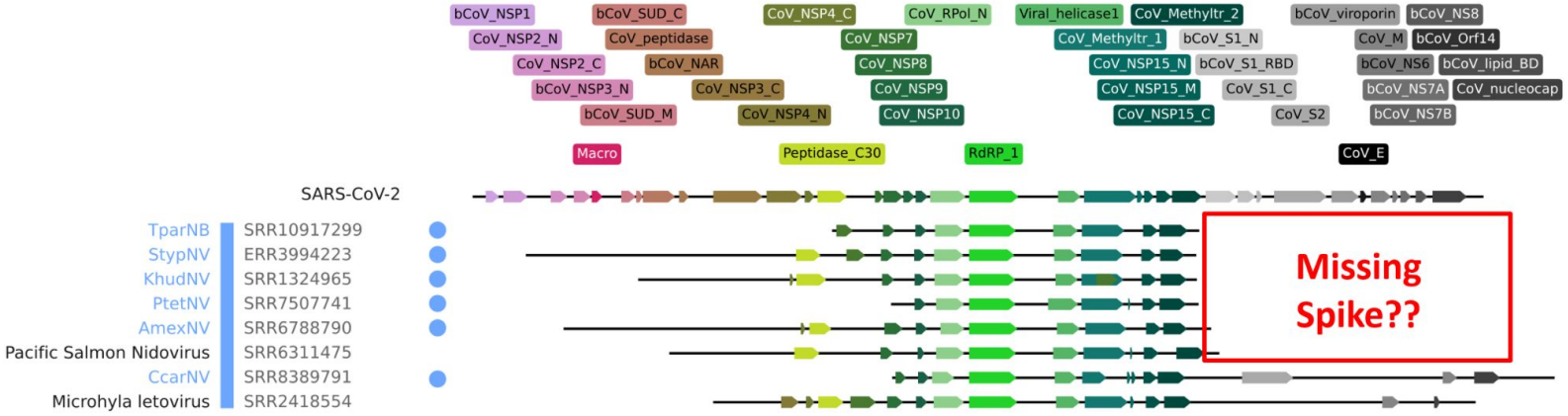
Discovering new Coronaviruses



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

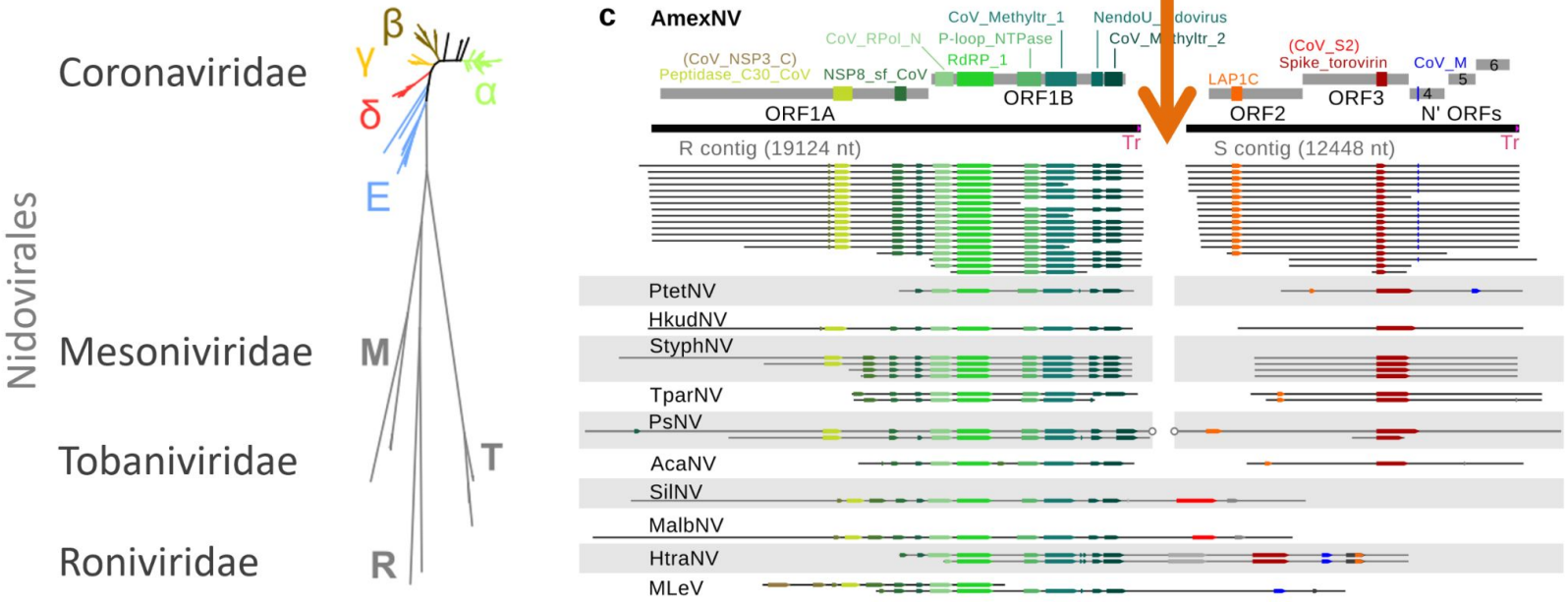
serratus.io

Discovering new Coronaviruses



serratus.io

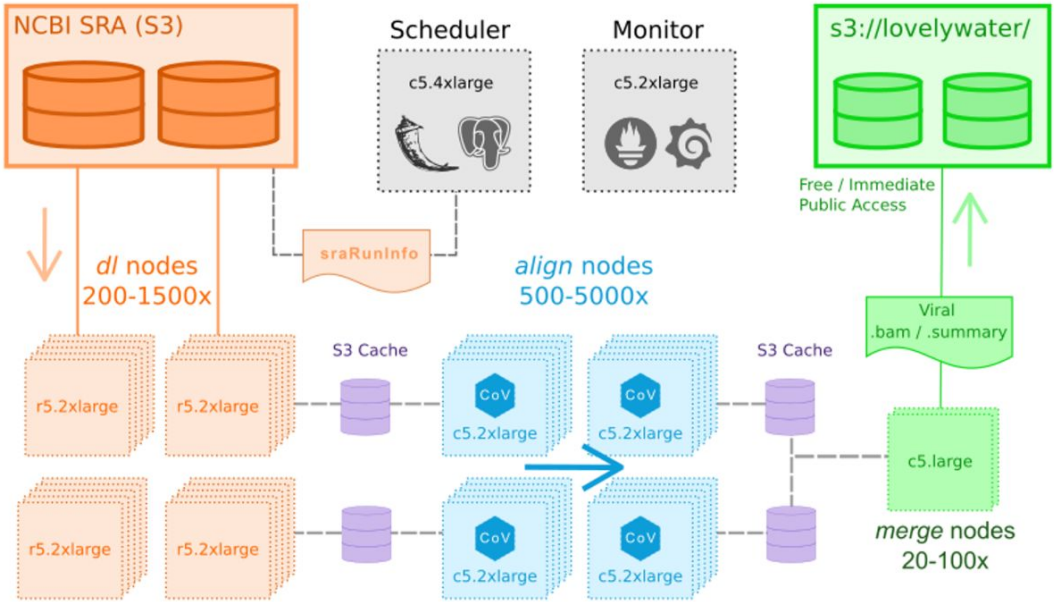
Segmented Coronaviruses?



Re-writing the textbook definition of a Coronavirus

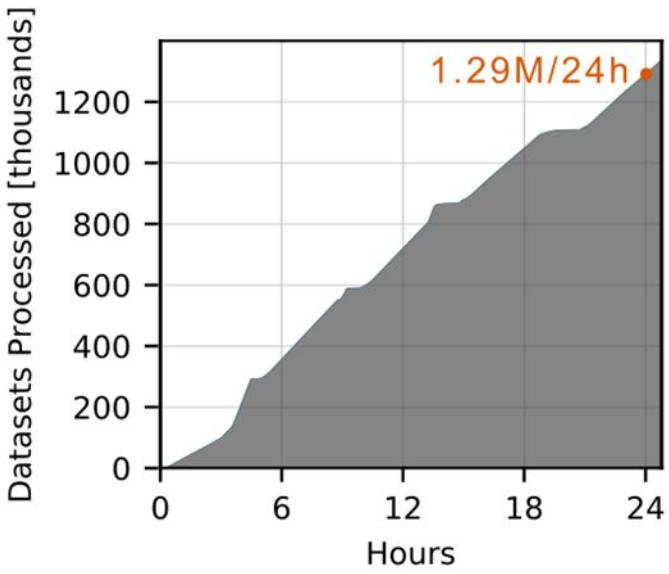
serratus.io

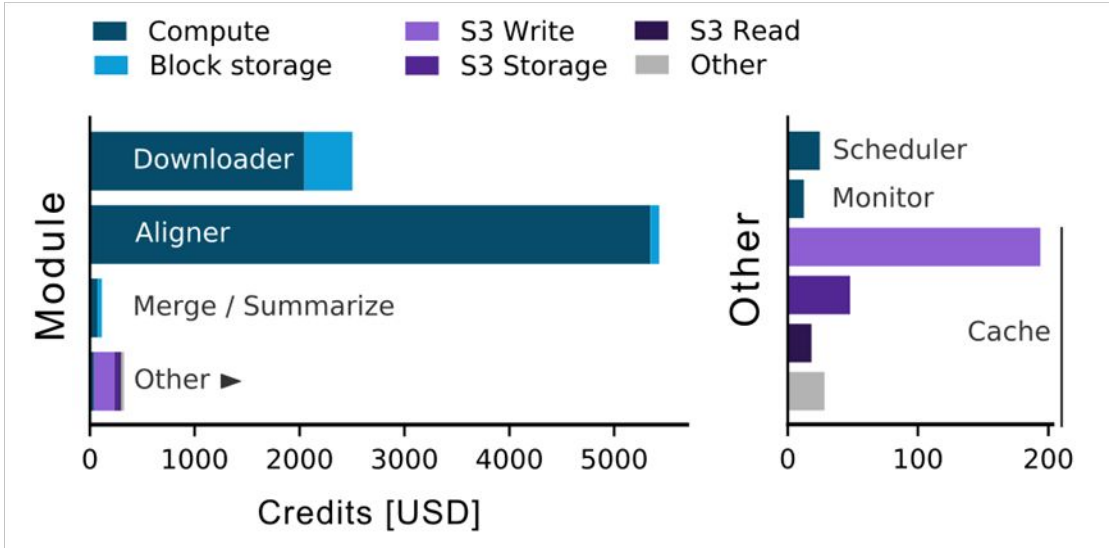
Serratus architecture



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

Serratus performance & costs

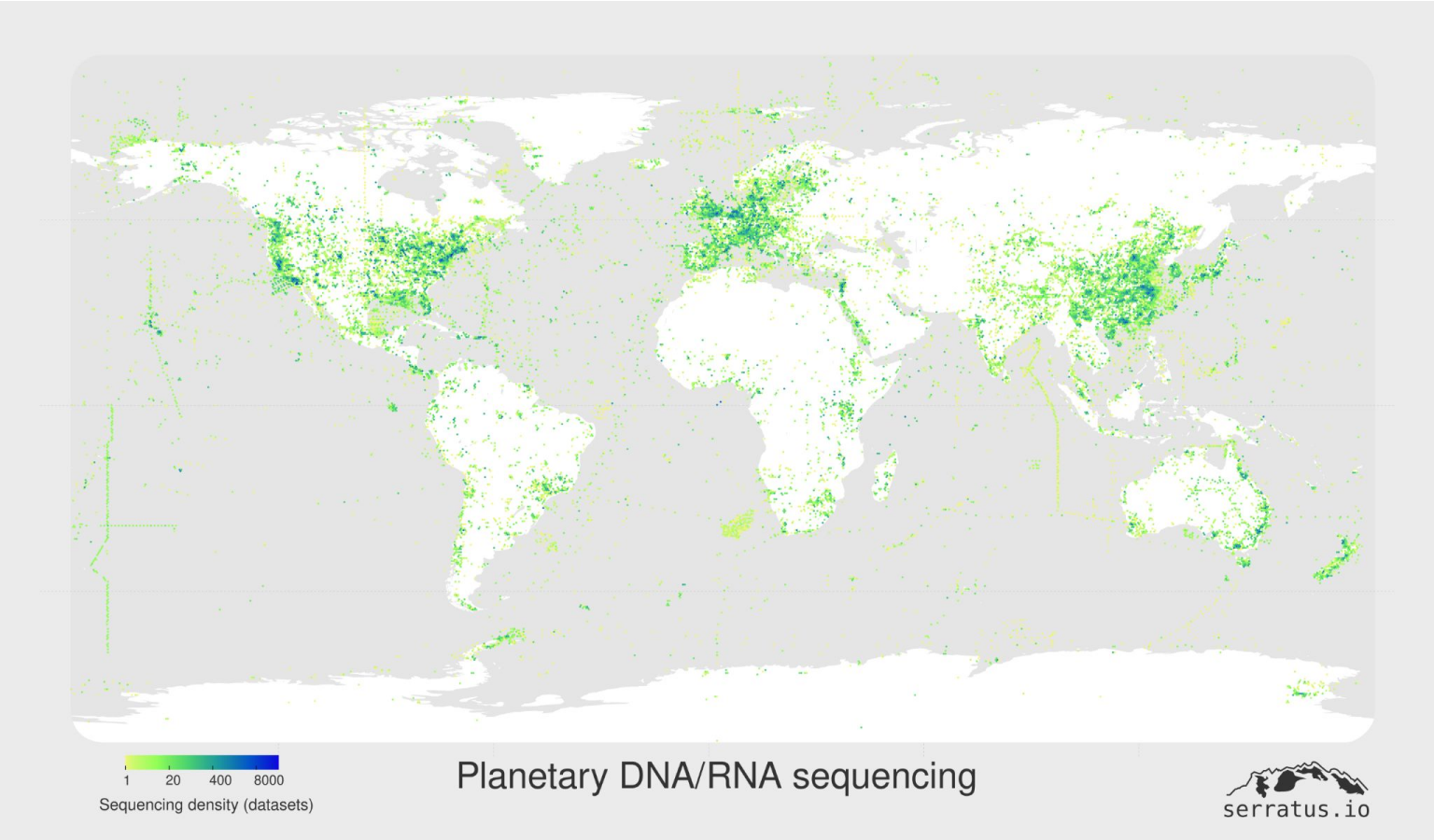




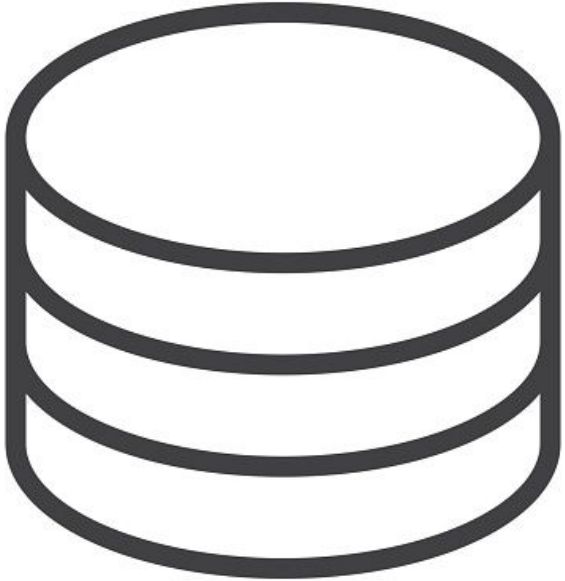
1 million NGS libraries / day

$0.005 / library

Geography of SRA samples

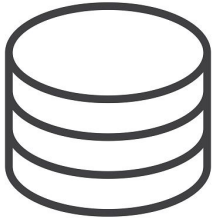


Analysis 2:



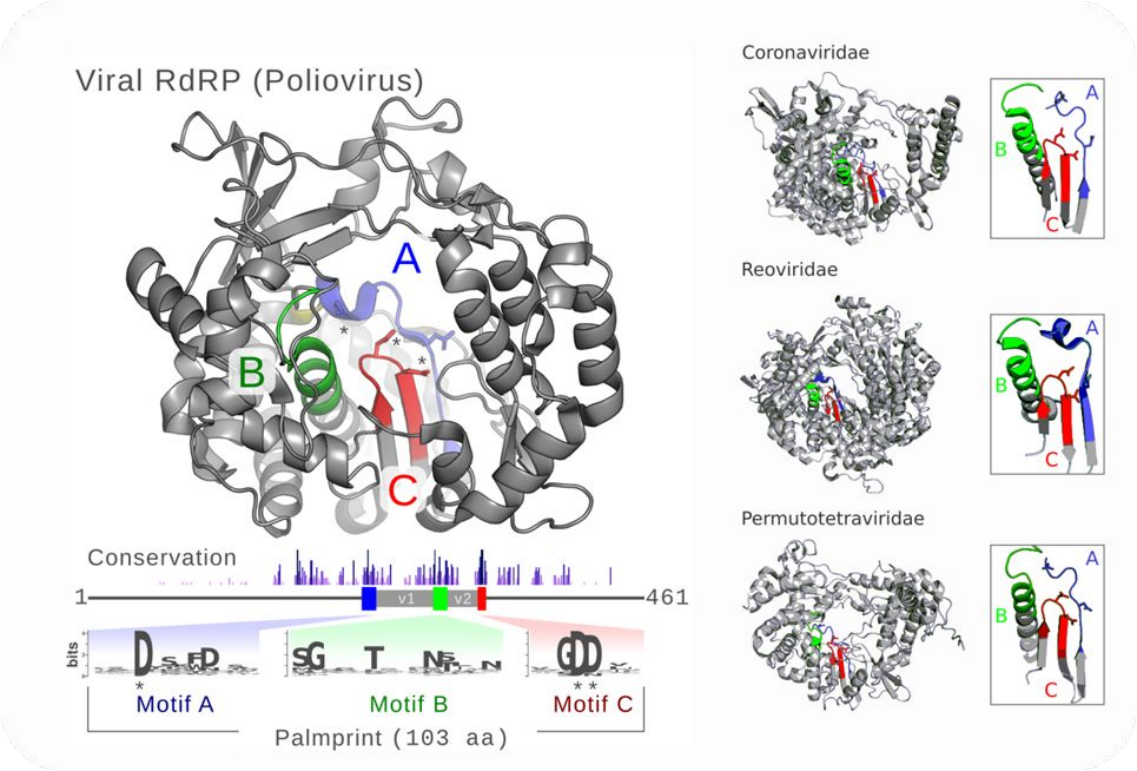
All RNA-seqs

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



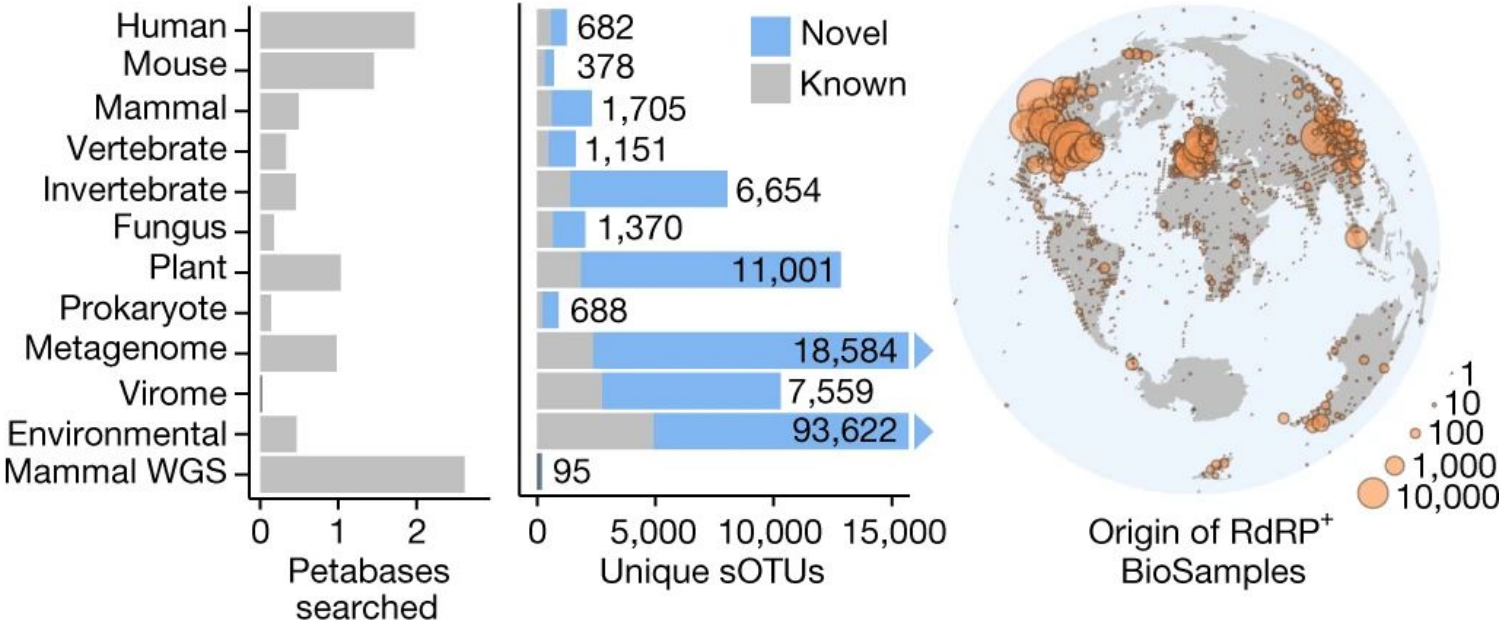
(Babaian & Edgar, 2021. bioRxiv)

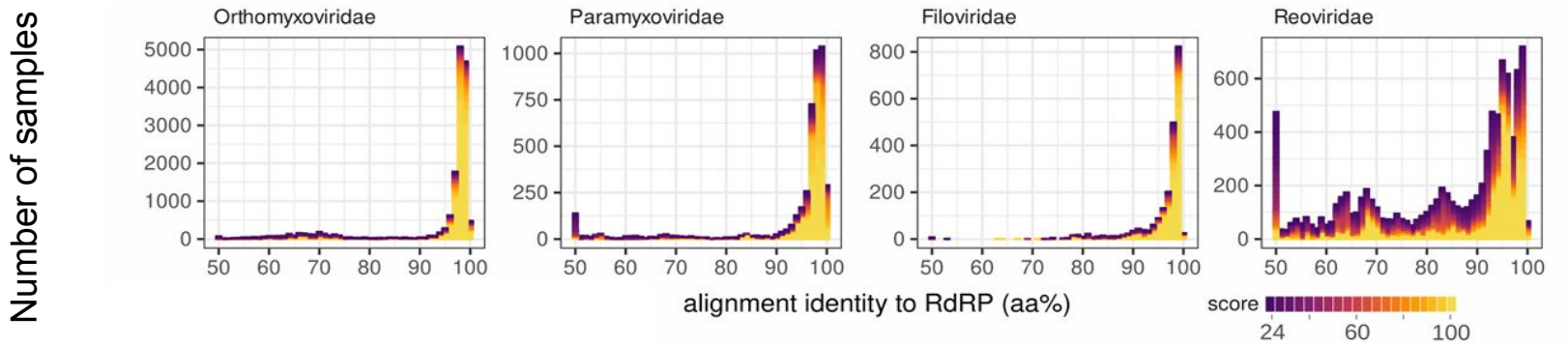
- RNA Virus “Palmprint”
- Species threshold: 90% amino-acid id

Assembly of all viral RdRPs (Analysis 2)

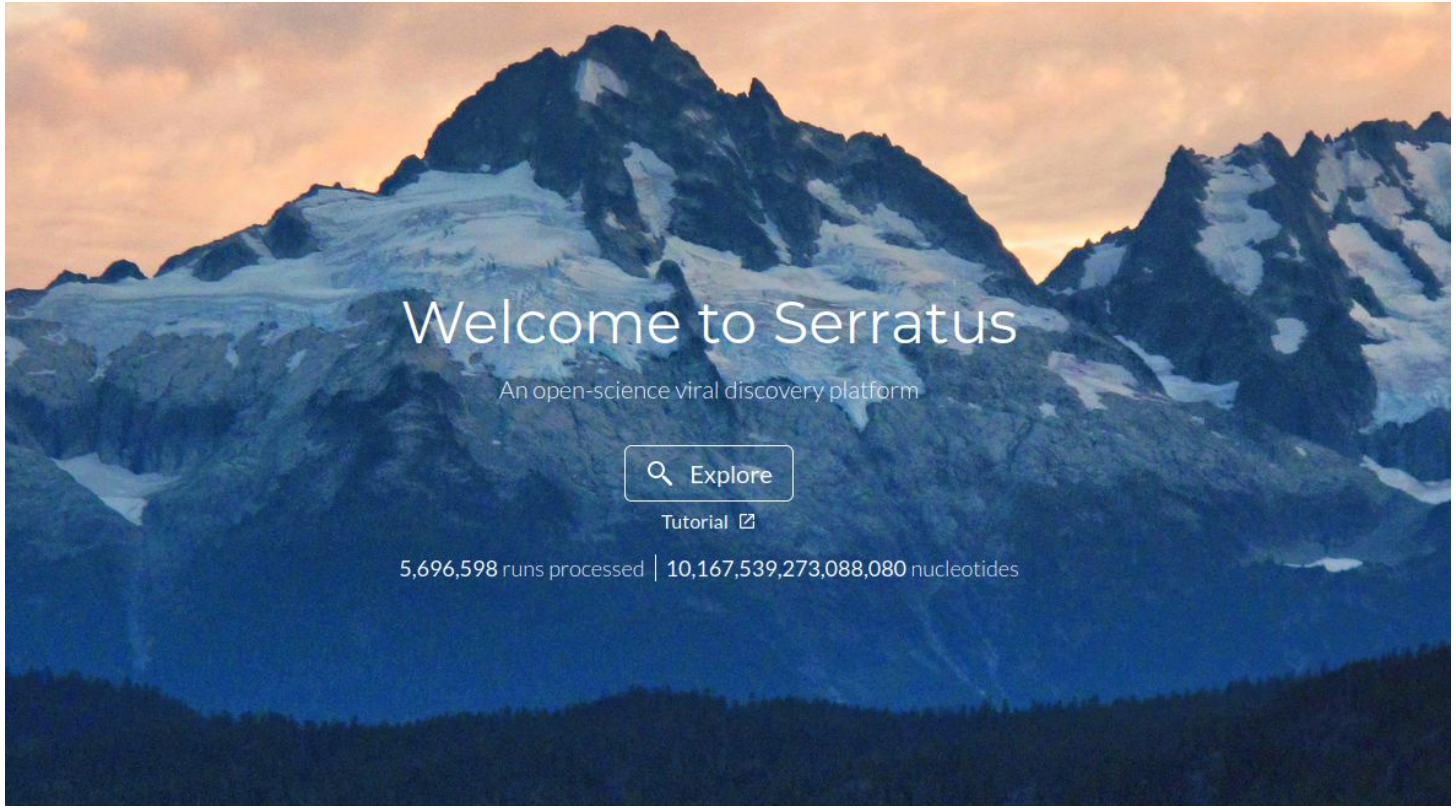
“Micro-assembly” of all RdRp-matching reads within each sample

- SPAdes assembler & GNU parallel
- Single large AWS instance (`c6a.48xlarge`, 192 cores)

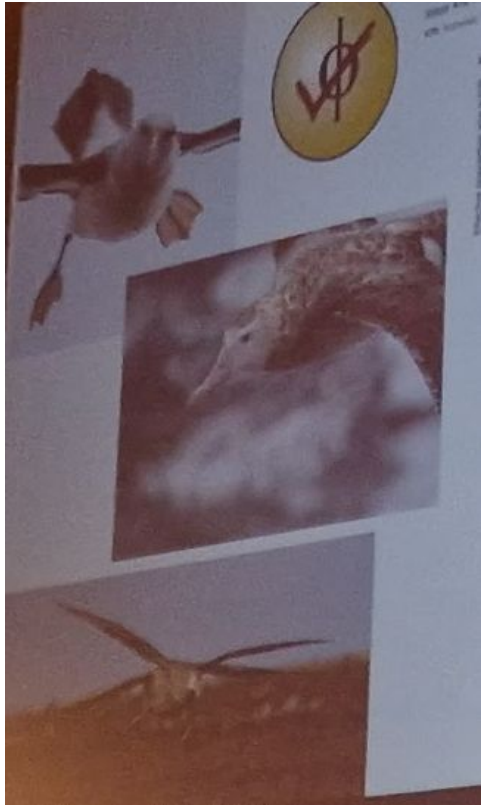

Human
Mouse
Mammal
Vertebrate
Invertebrate
Fungus
Plant
Prokaryote
Metagenome
Virome
Environmental
Mammal WGS
0
1
2
Petabases searched
682
378
1,705
1,151
6,654
1,370
11,001
688
18,584
7,559
93,622
95
Novel
Known
0
5,000
10,000
15,000
Unique sOTUs
1
10
100
1,000
10,000
Origin of RdRP+ BioSamples


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

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

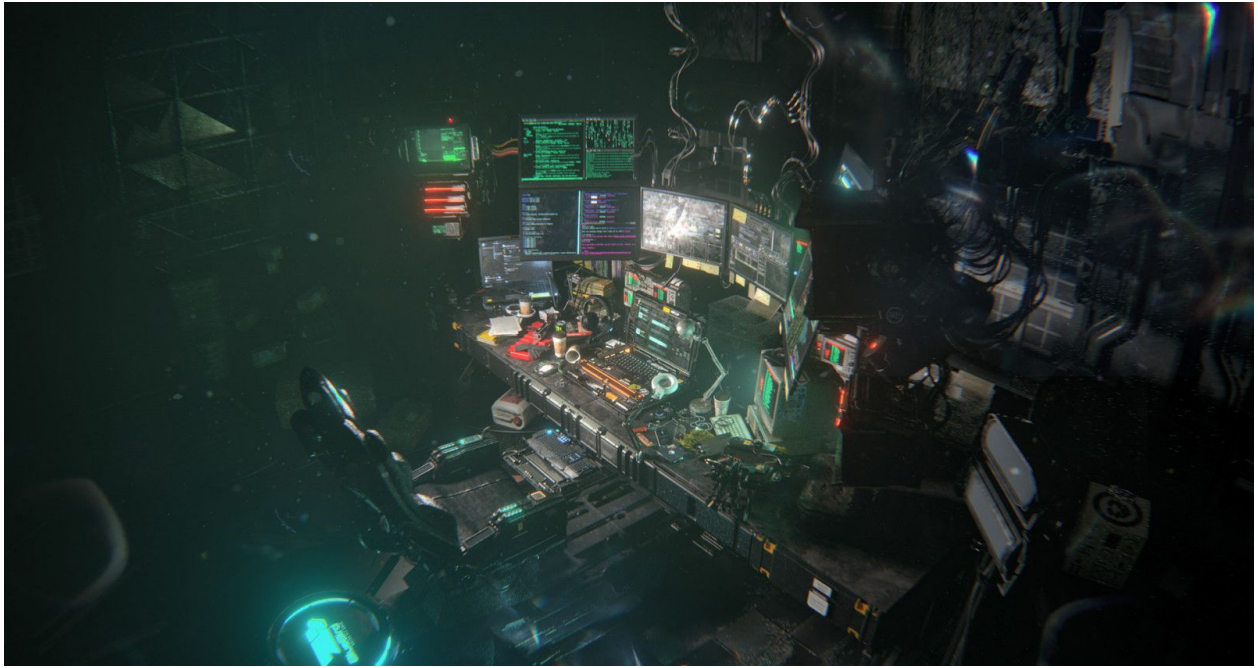


How was all of this large-scale assembly done?



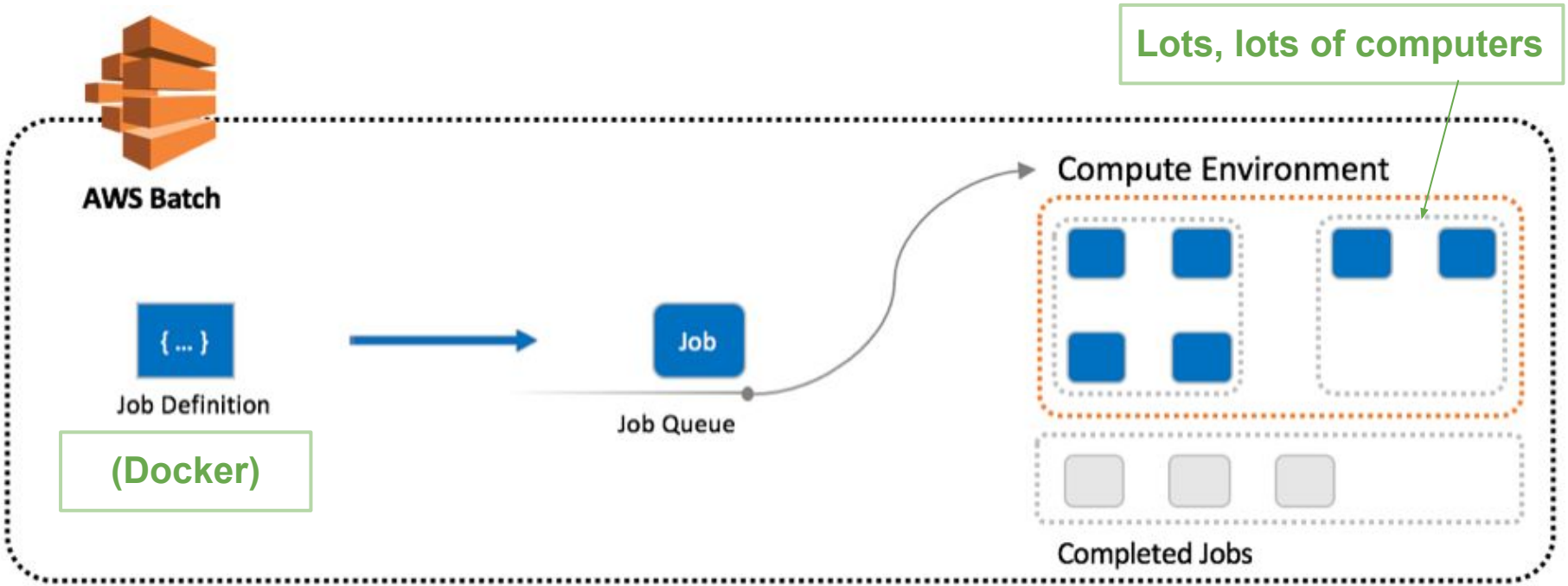
How was all of this large-scale assembly done?

cloud scripting



* (artist rendition)

AWS Batch framework for large-scale assembly



(Fig source: Amazon)

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
Compute	i-064fe18ba8316f79f	c5.9xlarge	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

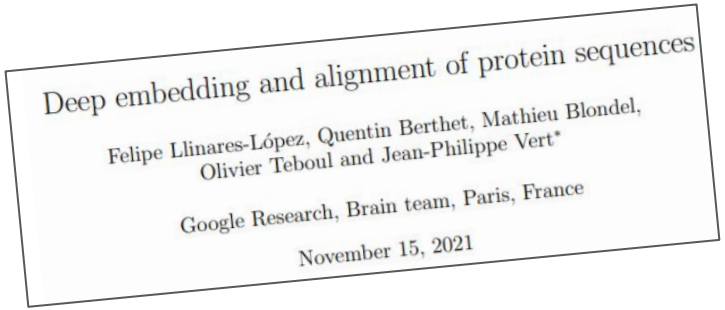
Job queue overview

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

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 …?

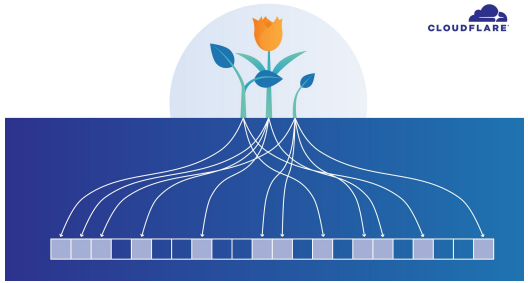

Deep embedding and alignment of protein sequences

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

Google Research, Brain team, Paris, France

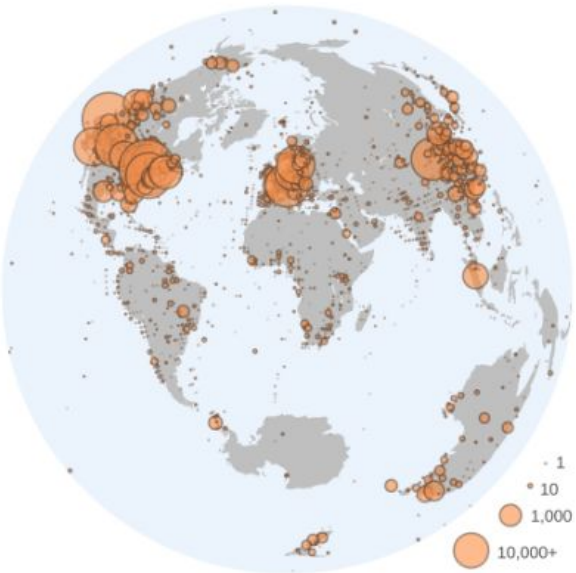
November 15, 2021

- A global **index of the SRA**
  - nearly feasible with k-mers already
  - would only support exact search
  - with ML, could do low(er) homologies



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

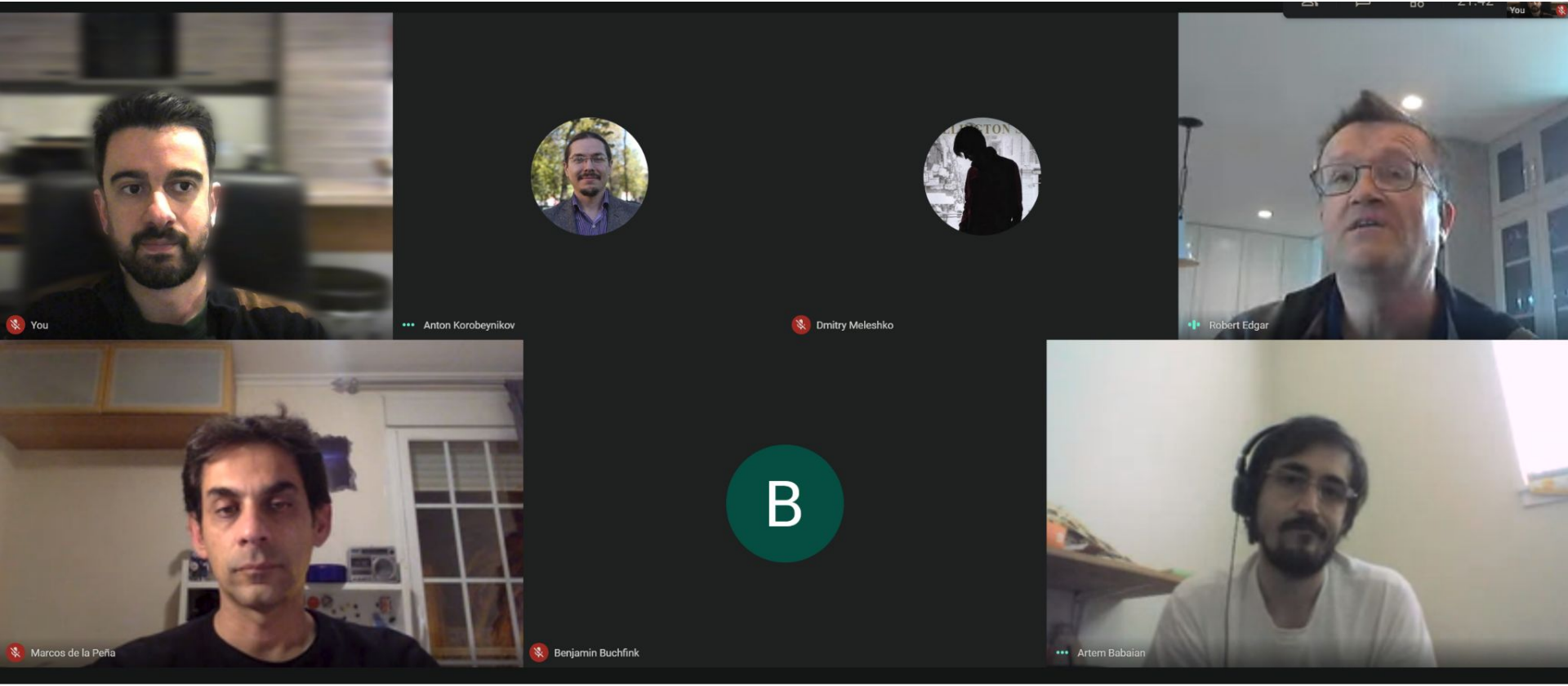
Public databases contain a planetary collection of nucleic acid sequences, but their systematic exploration has been inhibited by a lack of efficient methods for searching this corpus, which (at the time of writing) exceeds 20 petabases and is growing exponentially[1]. Here we developed a cloud computing 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.

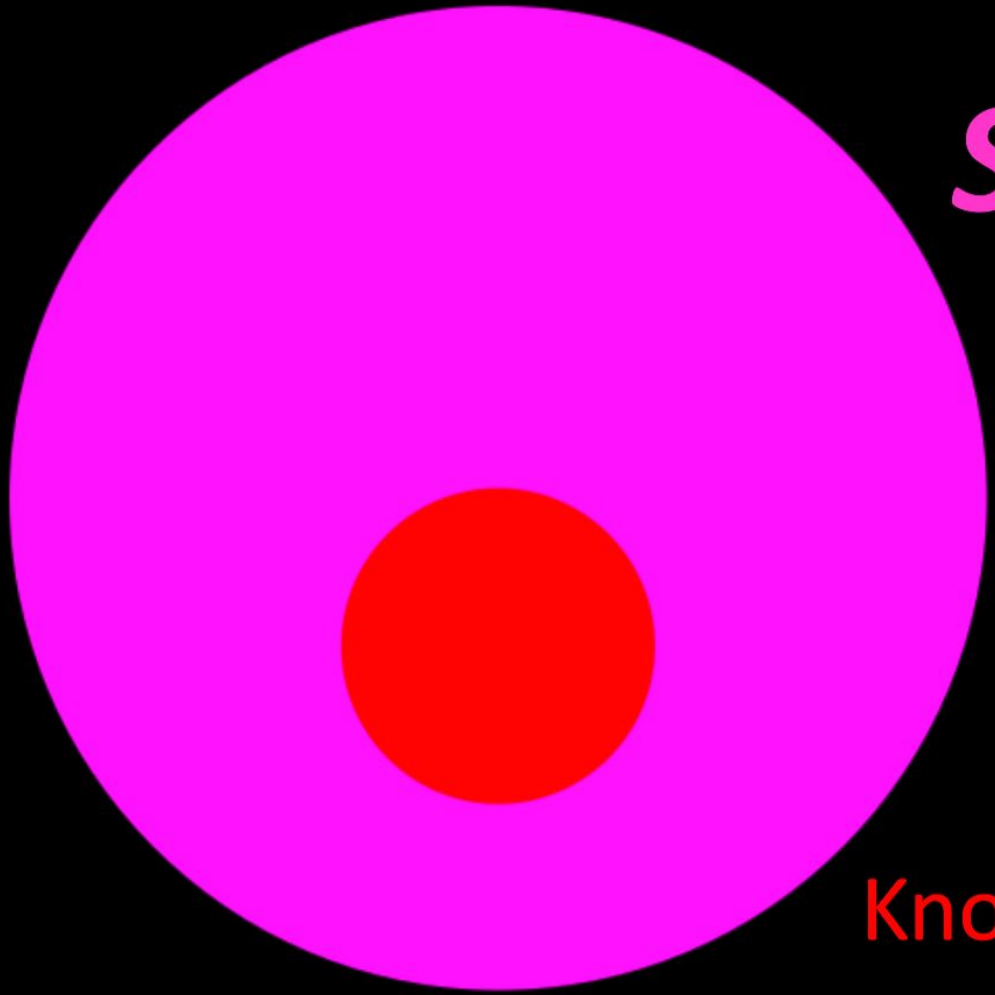
Digital Collaboration

- 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

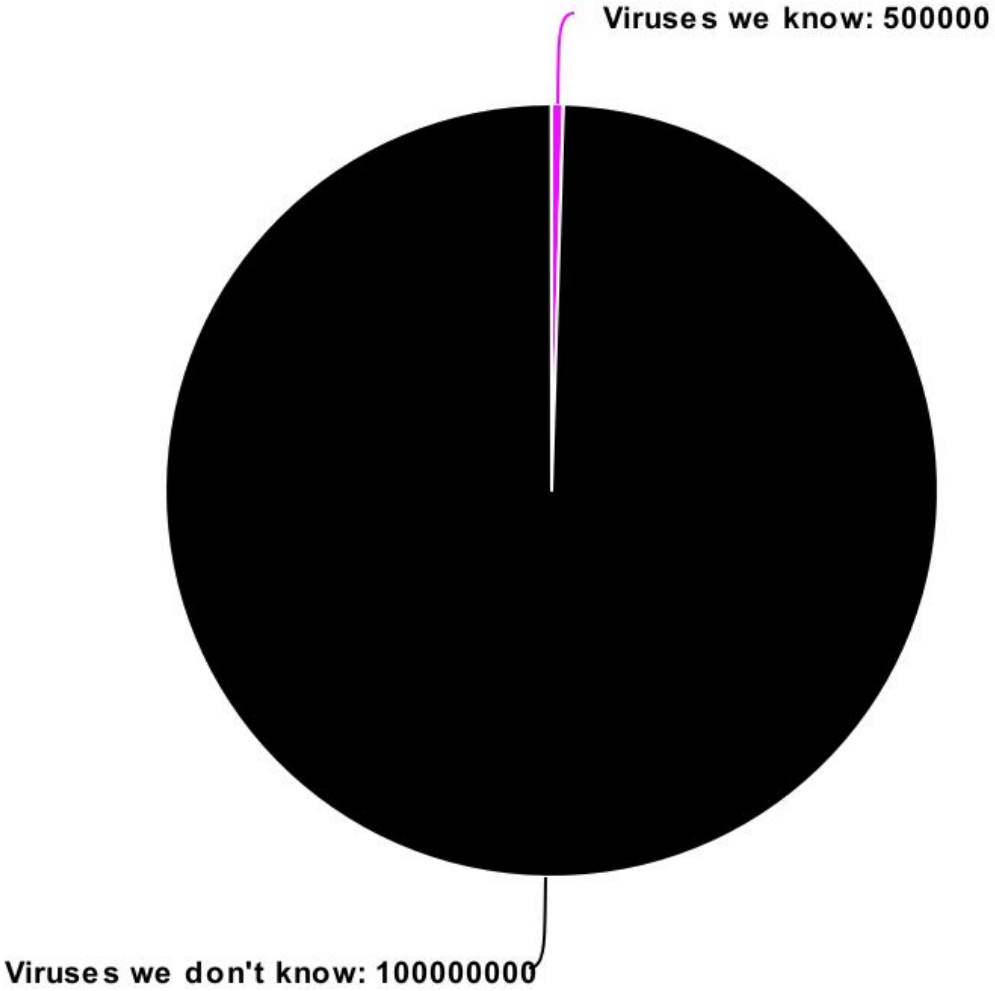




Serratus

Known RNA Virome

Earth's Virome



Pie chart for Josie

Outro

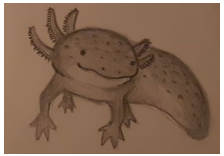

bigger data
NCC-1701-C
big data

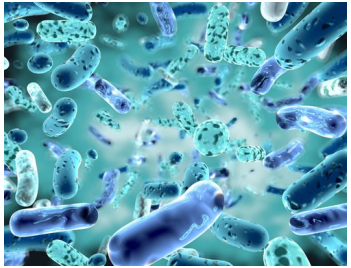
WE'RE GONNA NEED

A BIGGER INSTANCE TYPE

Sequence Bioinformatics

@ Institut Pasteur

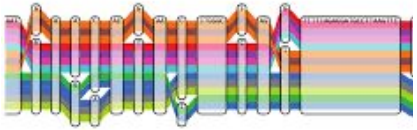




Genomes & metagenomes assembly

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

Congratulations for completing part 1! ⭐

- Francesco
- Erik
- Torda/Tammy
- Yu
- Janina
- Thomas
- Danilo
- Zoey :)
- Sam
- Daniel
- Alena
- Beatriz :D

but read more to get to part 2

Thank you for your attention!


Vielen Dank für ihre
Aufmerksamkeit!

❤️ To the amazing job done this week by Janina, Milos, Kartik, Alena, Madee, Joan, Mercè, and **Josie**!!

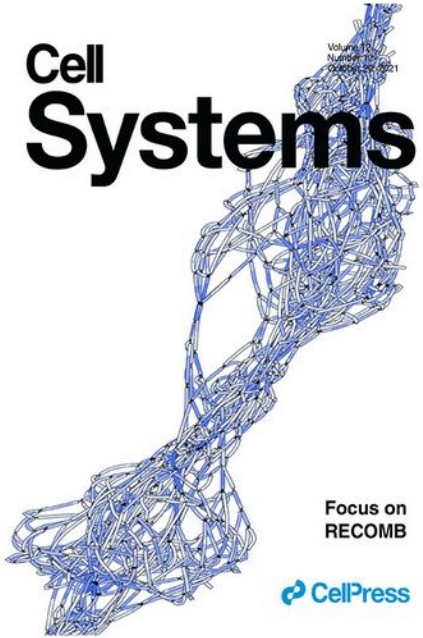
Supplementary slides
(surely won’t need them)

A detour…

We previously introduced **minimizer-space de Bruijn graphs (mdBGs)**, where

Instead of k-mers as nodes in the graph, we build **k-min-mers**

Classical alphabet: $\Sigma_{\text{DNA}} = \{\texttt{A}, \texttt{C}, \texttt{G}, \texttt{T}, \texttt{N}\}$
A k-mer with $k = 3$: `AGT`
Minimizer alphabet: Set of ℓ-mer minimizers as letters



A detour…

Fixed set of universe minimizers

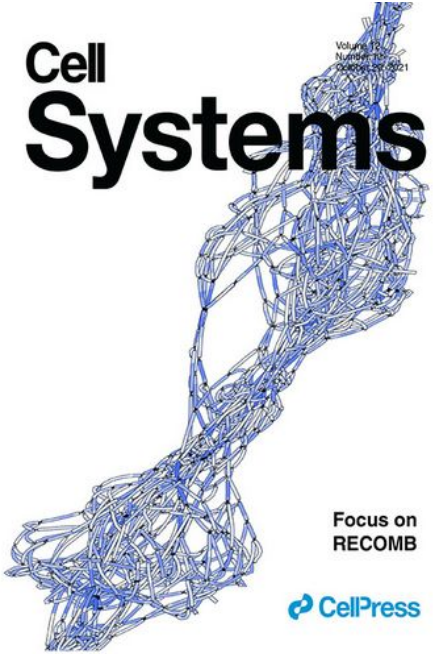
AATGACATGATCATGA

GA TC

Σ^ℓ = { all minimizers of length ℓ } = {m_1, m_2, m_3, …}

where e.g., $\ell = 2$, m_1 = AA, m_2 = AC, …

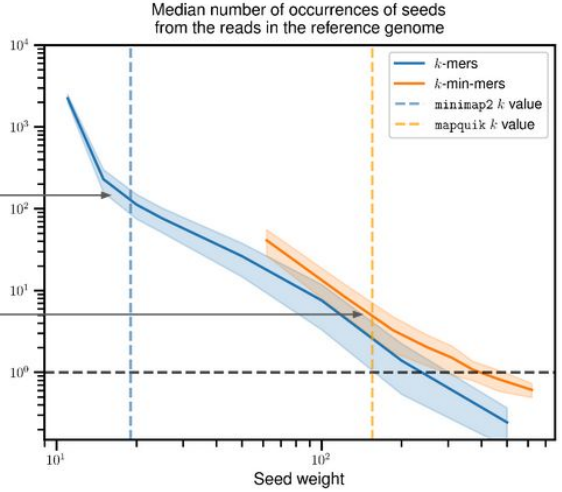
A ***k*-min-mer** (k-mer over Σ^ℓ): $m_1 m_3 m_2$

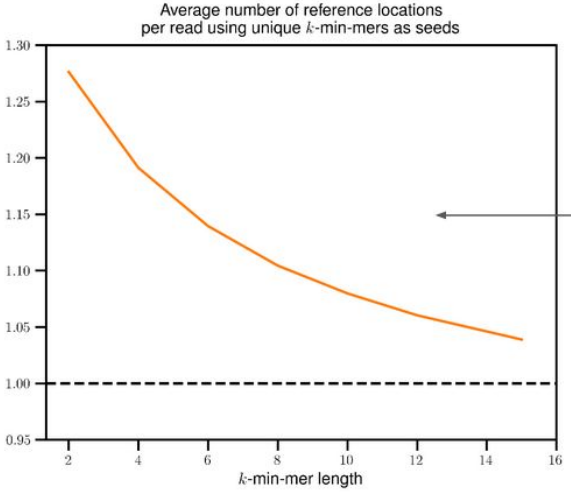


k-min-mers as alignment seeds instead of k-mers?

k = 19
> **100** occurrences

k = 5, ℓ = 31
< **10** occurrences





1.05 – 1.30 candidate locations on average

likely to find the right location by querying **all** k-min-mers + check those that occur **once**