

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 petabytes 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 18S species) has been determined to be pApApApCpUpUpCpApApCpApApCpGpGpApUpCpUpGpGpUpUpCpUpCpGpCpApUpCpGpApUpGpApApGpApApCpGpCpApGpCpGpApApUpGpCpGpApUpApCpGpUpApApUpGpUpGpApApUpGpApApUpCpUpUpCpCpGpUpGpGpUpApUpUpCpCpApGpGpGpGpCpApUpGpCpCpUpGpUpUpGpApGpGpGpCpGpUpCpApUpUpU.

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

Information technologies scale exponentially

Sydney Brenner and Nathan Myhrvold, ~2005

		Base pairs
1995	Bacterium	2×10^6
2000/3	Mammal	3×10^9
2013	2500 humans	7.5×10^{12}
2021	~1M genomes	3×10^{15}

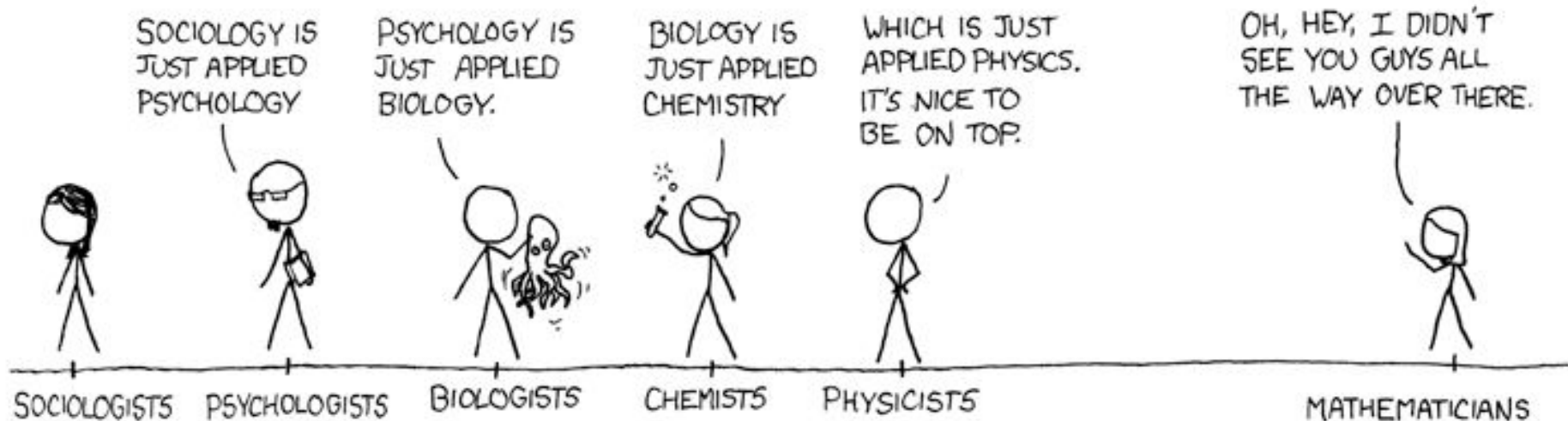
Cost drop from \$1/bp to \$10⁻⁷/bp

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

DNA sequence, genomes and computation together
Informatics is to biology what mathematics is to physics ?

*“Informatics is to biology,
what mathematics is to physics”*

Richard Durbin, RECOMB 2023 keynote



“purity” →

← “usefulness”

Big data in biology: GenBank

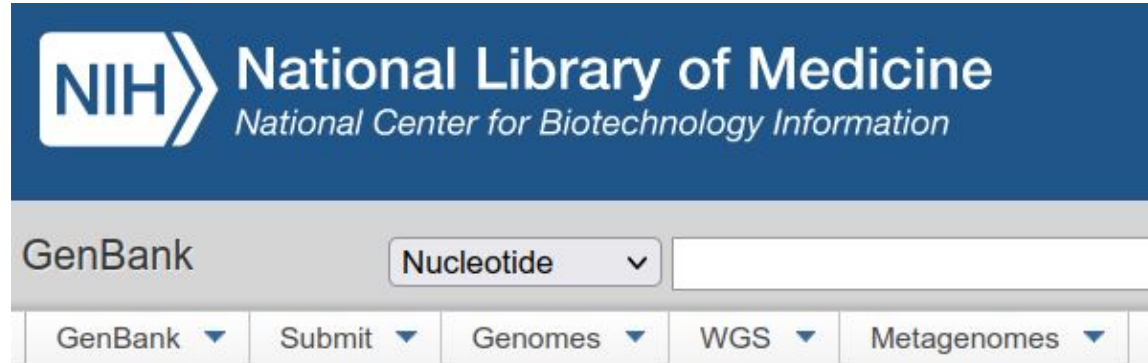


Type: assemblies of >500,000 species

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

Particularity: all sequences are *annotated*

NCBI WGS



The image shows the top section of the NCBI GenBank submission page. It features the NIH logo and the text 'National Library of Medicine' and 'National Center for Biotechnology Information'. Below this is a search bar with the label 'GenBank' and a dropdown menu set to 'Nucleotide'. At the bottom, there is a navigation bar with buttons for 'GenBank', 'Submit', 'Genomes', 'WGS', and 'Metagenomes'.

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](#))

Difference with GenBank: sequences are not necessarily annotated

NCBI SRA

Size: 30 PB

SRA

SRA

Advanced

Search

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: Find Filtered Download [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)
ACCTGAGCGCGCAGCTCCAGTAAATCAAACGCGGCGCGGAATTTGGGATGTTCCATCAGT
TTCCAGGCGCGTTTGCCCTGACGTGCGGACATGCGTAACTGAAGCTGCCAAATATCAGCG
GTAAGCGTGGTAAGCGCTTTCGGGATCGCCA

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)
ATCAACAACAGCGGGAATACCACTCTTCCAGCCGTTGTTTCCAAACCAATACGCGTTAAT
TCACCGAAACCGGACAGCGCAATGGAACGCATCATTCGCGCAGGTGTTCGAGAATACGGA
AAACCGCATCCGAAACGAGATGCGCGTTAAT

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
```

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



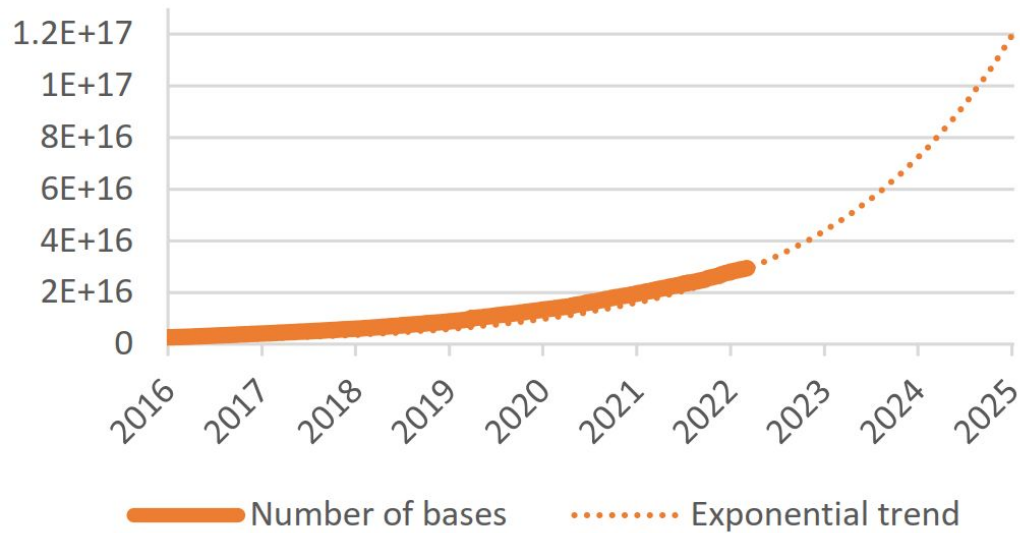
NCBI SRA database: 30 PB

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

Dreams,
Fantasy Dreams,
and 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



“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)
- ...

The background of the slide features a complex, abstract design. It includes a glowing DNA double helix structure in shades of blue, green, and yellow. Interspersed with the DNA are various binary code elements, such as '0' and '1' digits, and lines of light that suggest data flow or digital pathways. The overall aesthetic is futuristic and high-tech, typical of a presentation on bioinformatics or data science.

“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****

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

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



The human genome is *finally* complete



- Introduced nearly **200 Mb** of new sequence (vs the GRCh38)
- Finally resolved with combination of high-coverage long accurate reads (PacBio HiFi) + ultra-long data (ONT)



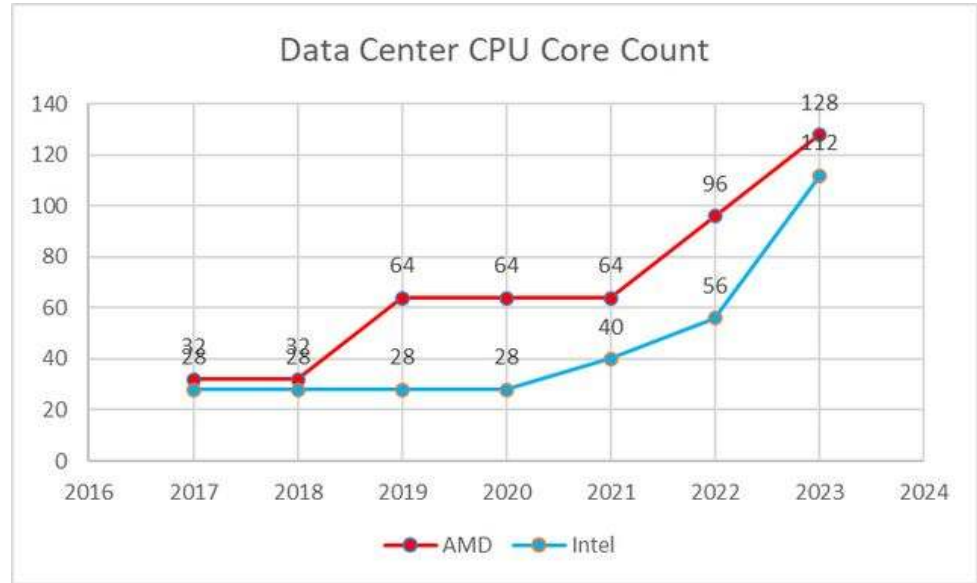
Science 100% COMPLETE
FILLING THE GAPS
Closing in on a complete human genome

Nanopore Community Meeting 2022 | @NanoporeConf | #NanoporeConf

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



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

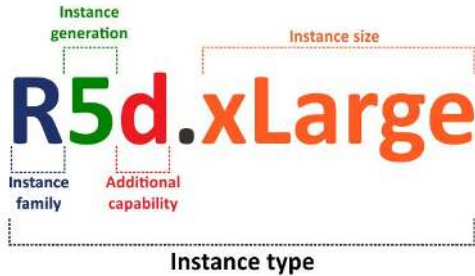


OVHcloud, Roubaix, France

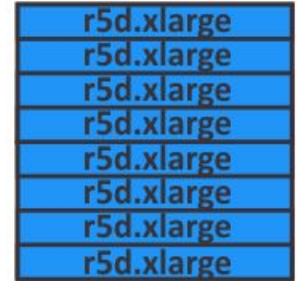
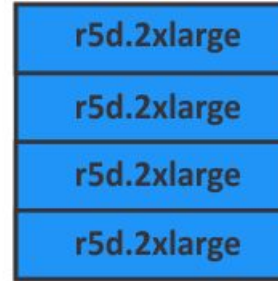


Your workshop instance: `t3a.large` : 2 CPU cores, 8 GB memory

AWS EC2 instance naming



AWS EC2 instance sizes



💖 `c6a.48xlarge` 💖 : 192 cores, 384 GB mem

What is *nearly* a cloud

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

- (This dog)



“Storing information in the cloud”?

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



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



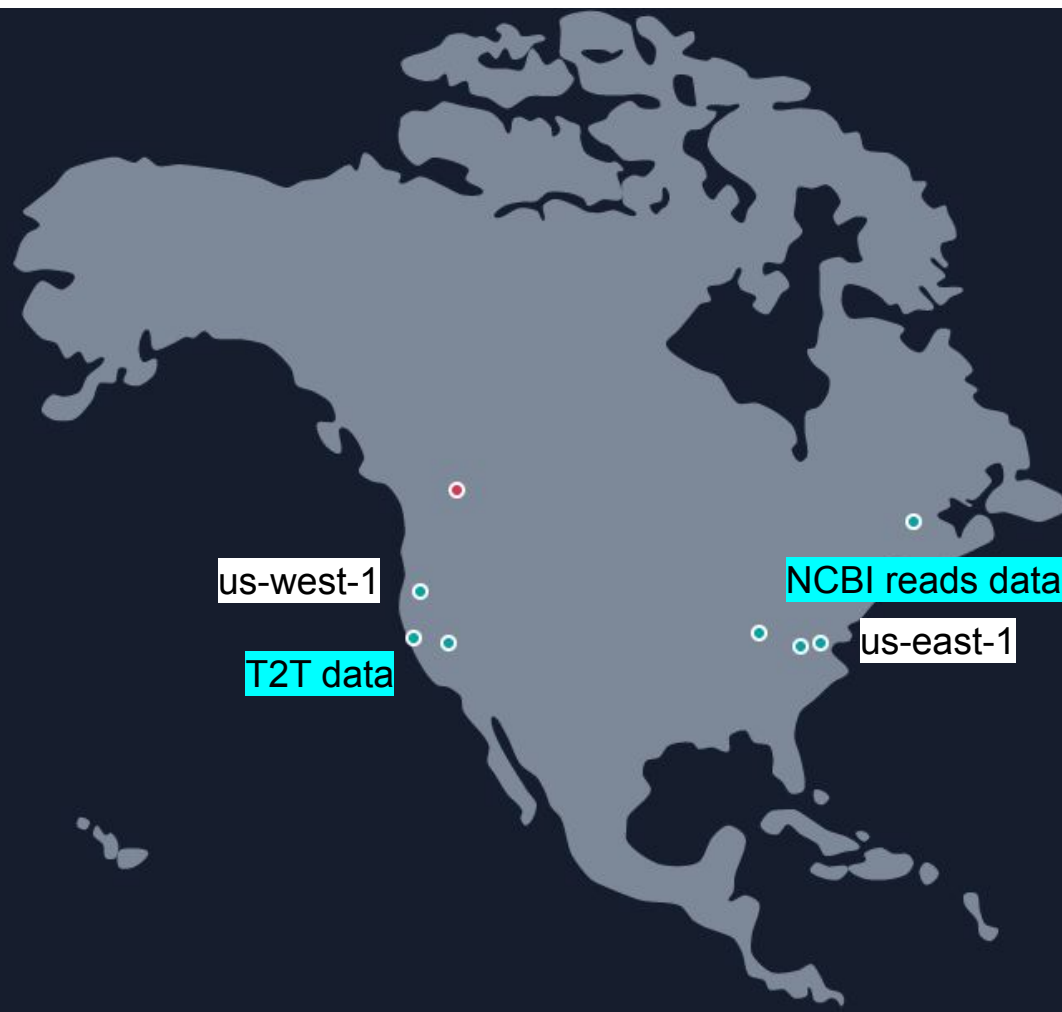
chicano joker @datLucario

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



us-west-1

T2T data

NCBI reads data

us-east-1



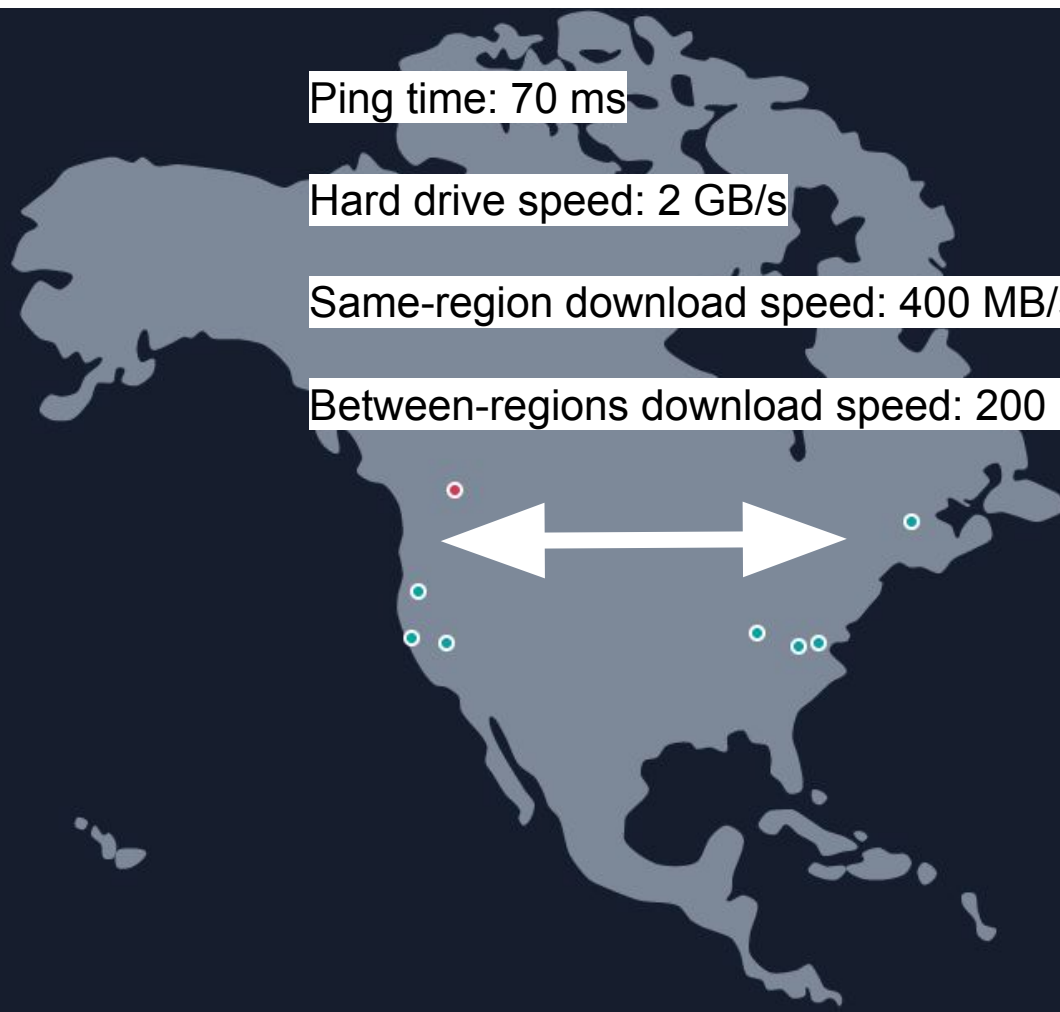
Makes a difference?!

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

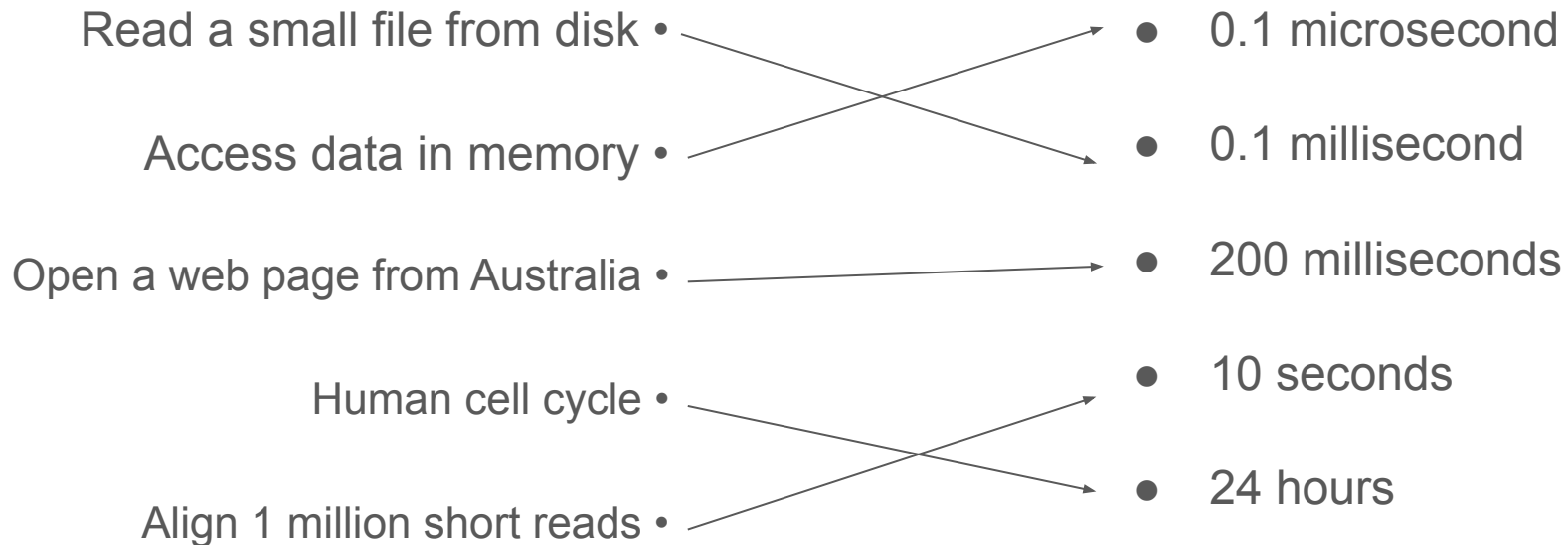
5) Align 1 million short reads



- 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

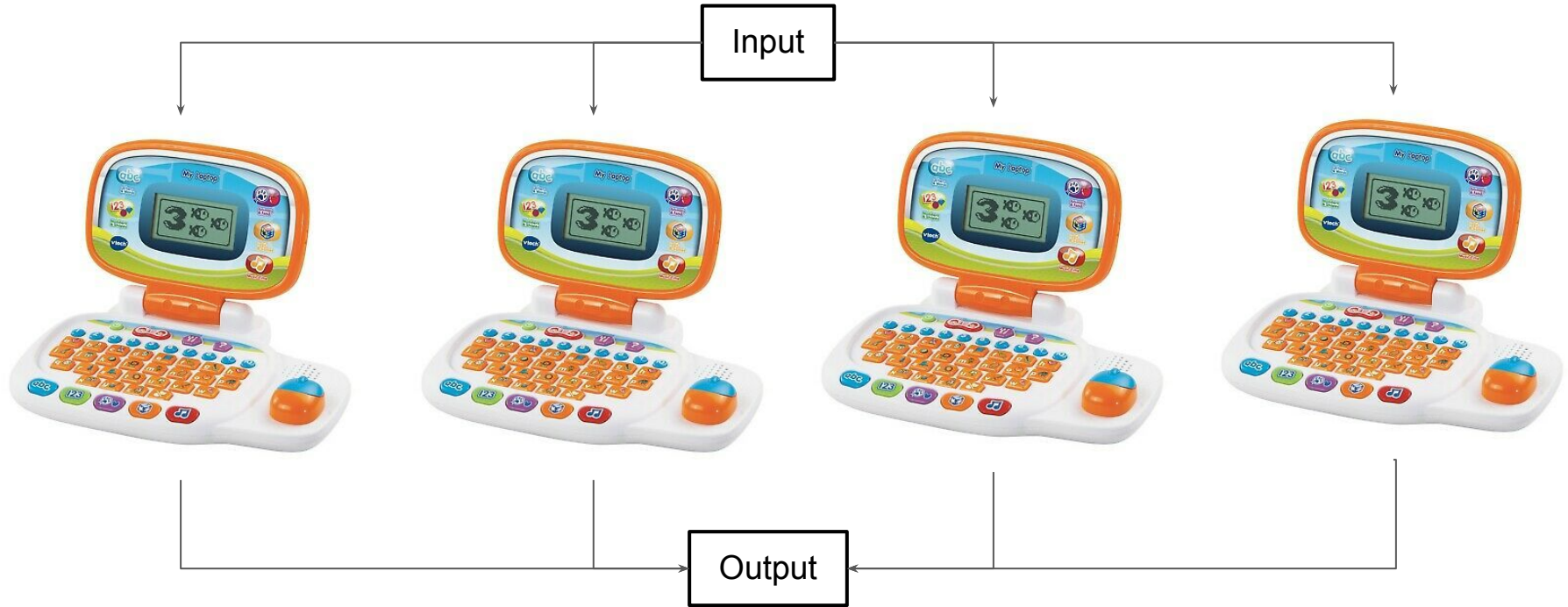


-	-	10^0	1
deci	d	10^{-1}	0,1
centi	c	10^{-2}	0,01
mili	m	10^{-3}	0,001
micro	μ	10^{-6}	0,000 001
nano	n	10^{-9}	0,000 000 001
pico	p	10^{-12}	0,000 000 000 001

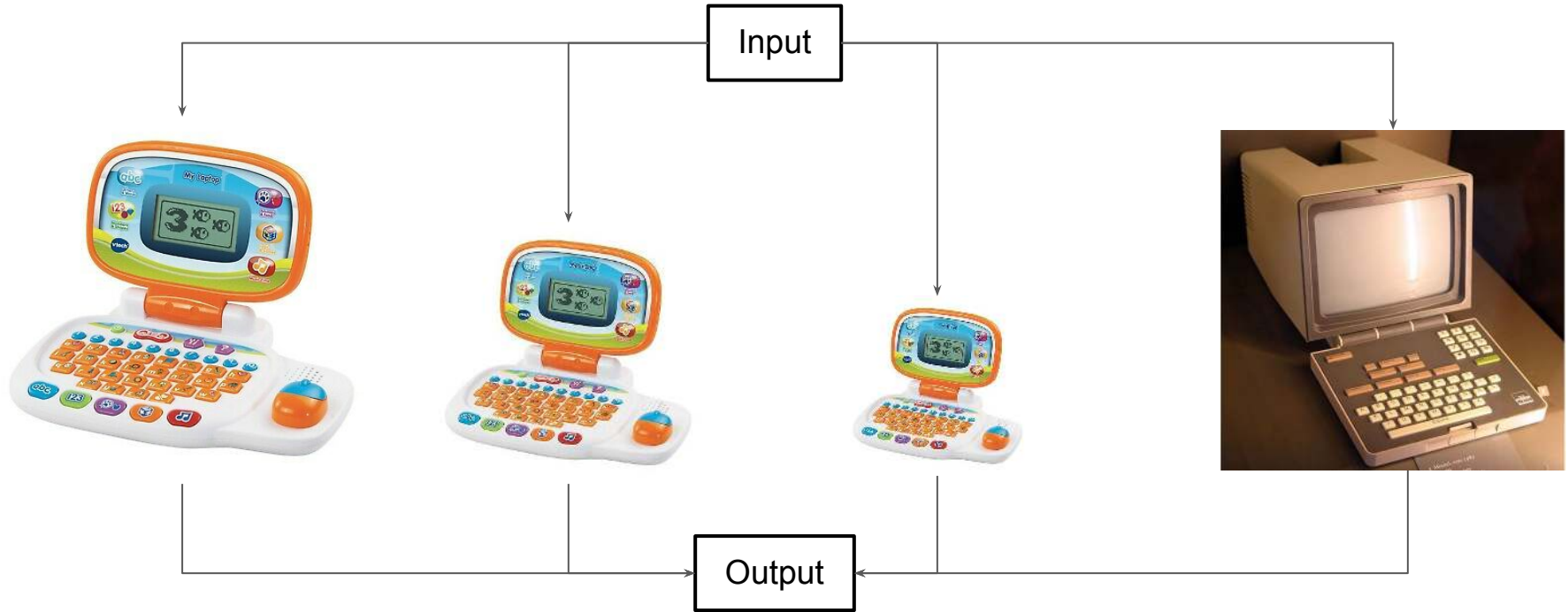
Are 200 CPUs 200x faster than 1 CPU?



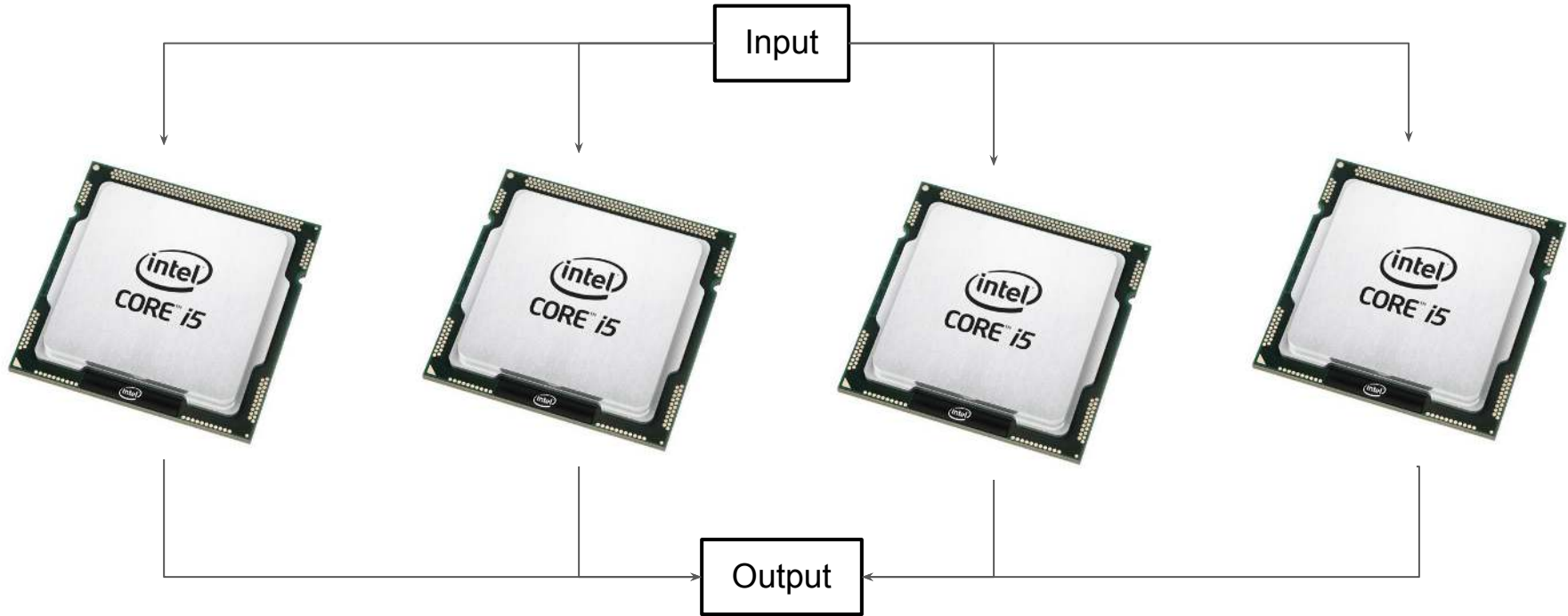
Parallelism: use many “computers” to execute one task



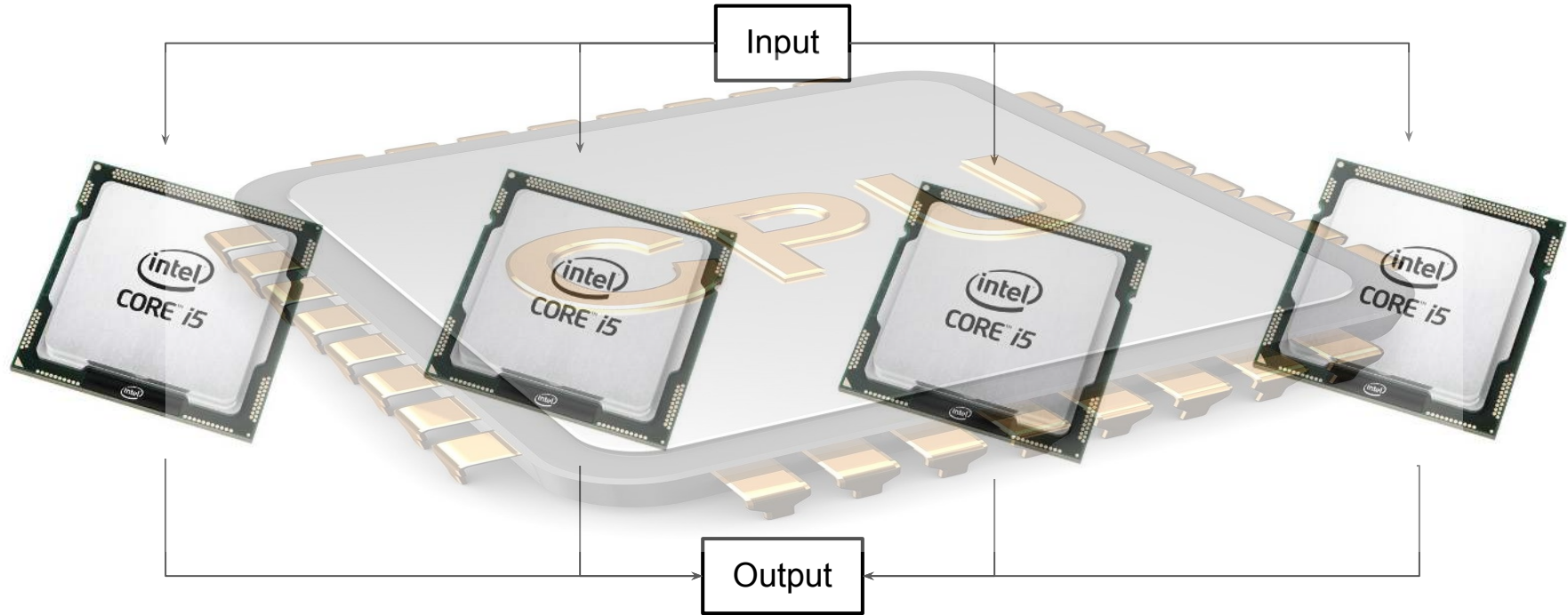
Parallelism: they don't need to be identical computers



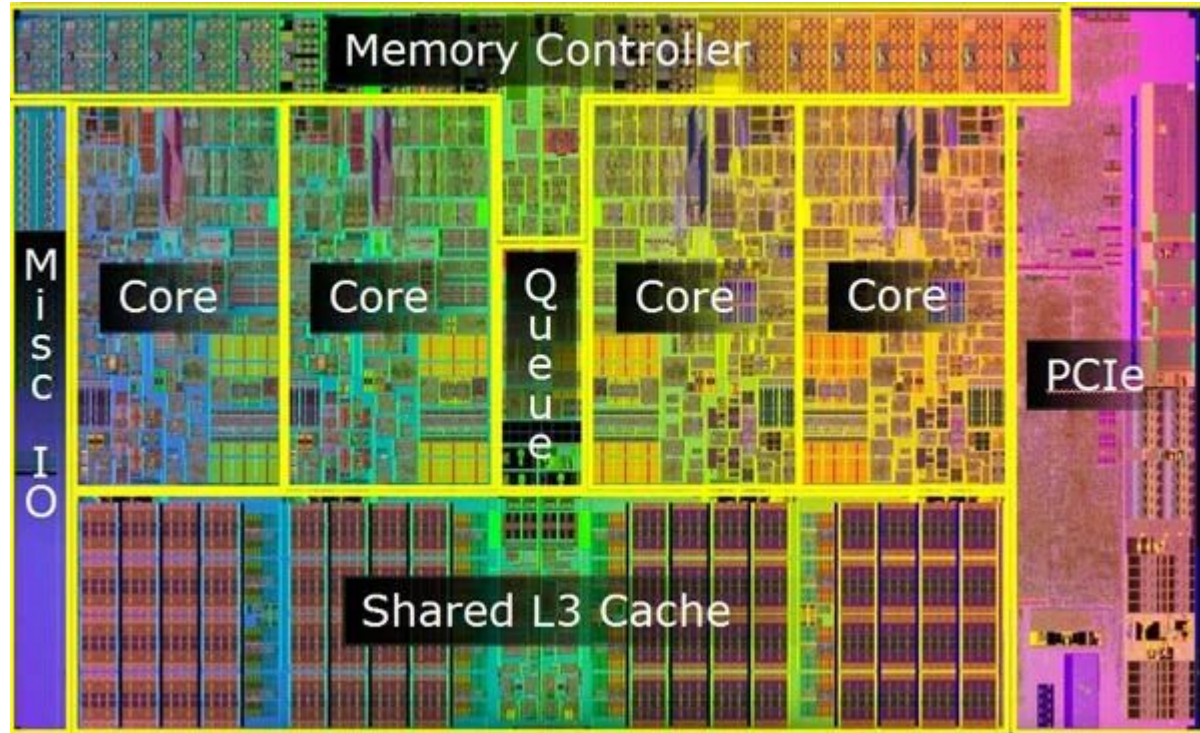
Parallelism: they don't even need to be “computers”



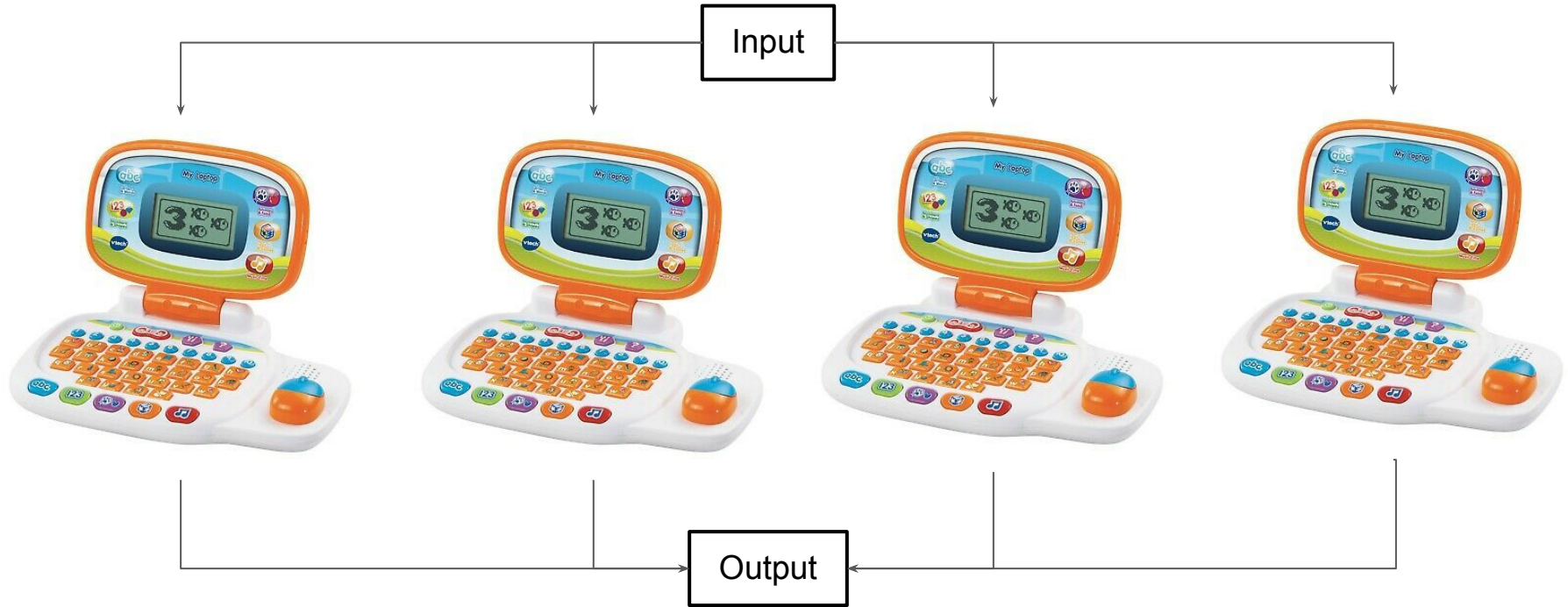
Parallelism: they don't even need to be “computers”



Parallelism: CPU = many little computers in parallel



CPU (simplified)

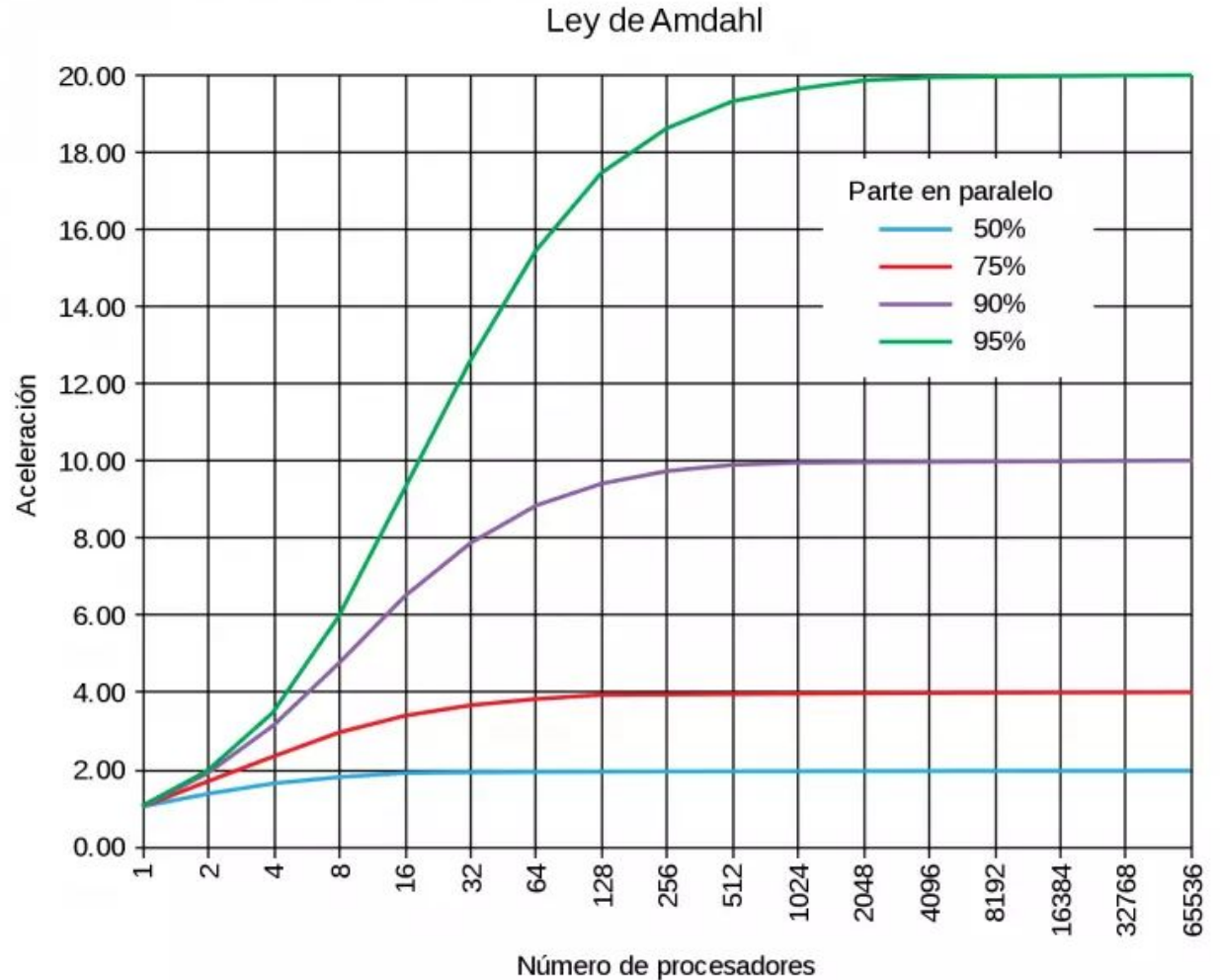
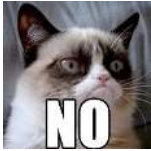


The limits of computing

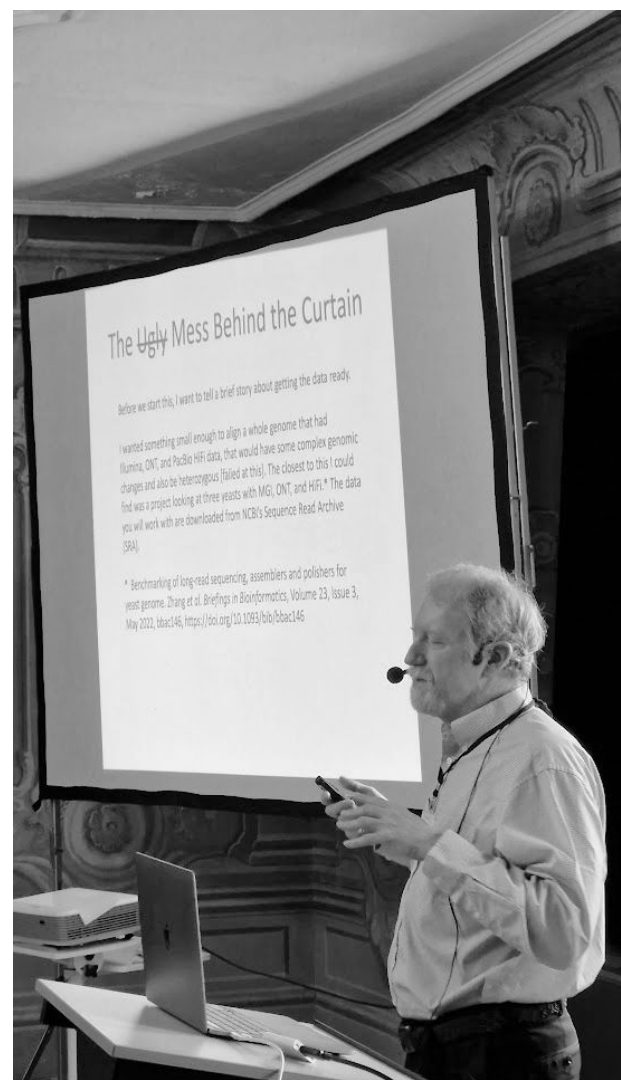
So, can we speed up indefinitely by stacking computers (or CPUs)?



Amdahl's law:



*“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

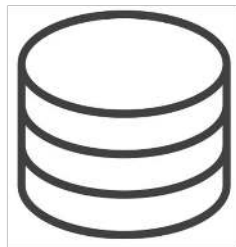
Part 3: map quickly!

<https://github.com/ekimb/mapquik>

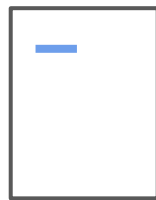
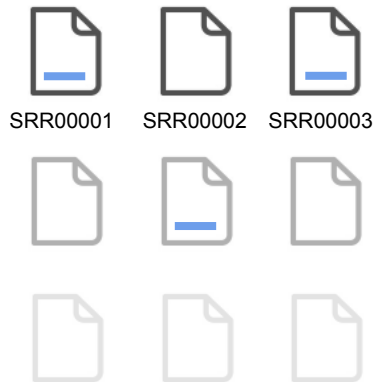


Live: Demo of mapping human 10x coverage HiFi reads using mapquik in <20 seconds, including FASTA conversion using seqkit and chatgpt

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  project)

ACTGATGGTG?
GTGAATGG?
AAAAAAAAAA?

Web pages

Google:



Search query

Any questions so far? Coffee break?



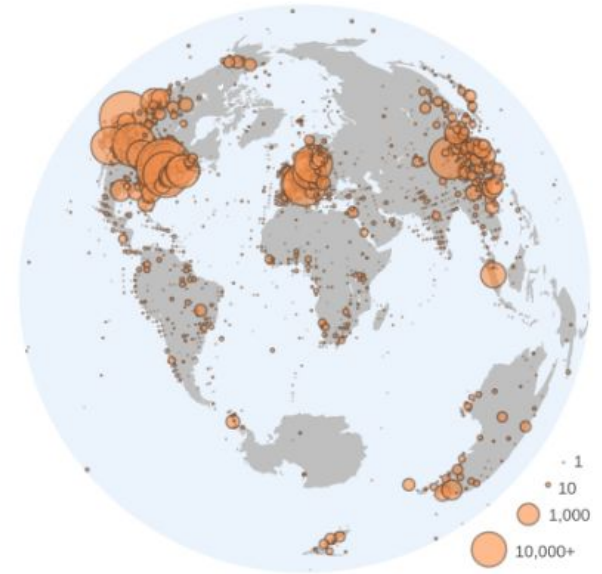
Hey don't leave, there is a part 2!!!



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.



NCBI SRA database : 30 PB

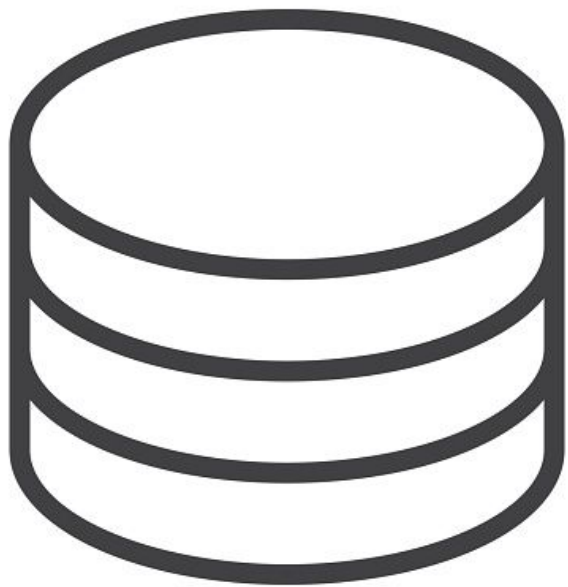


NCBI SRA database : 30 PB



Data crypt

Reads just sleep,
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 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:



[Tous](#) [Images](#) [Actualités](#) [Shopping](#) [Vidéos](#) [Plus](#) [Outils](#)

Environ 291 000 résultats (0,57 secondes)

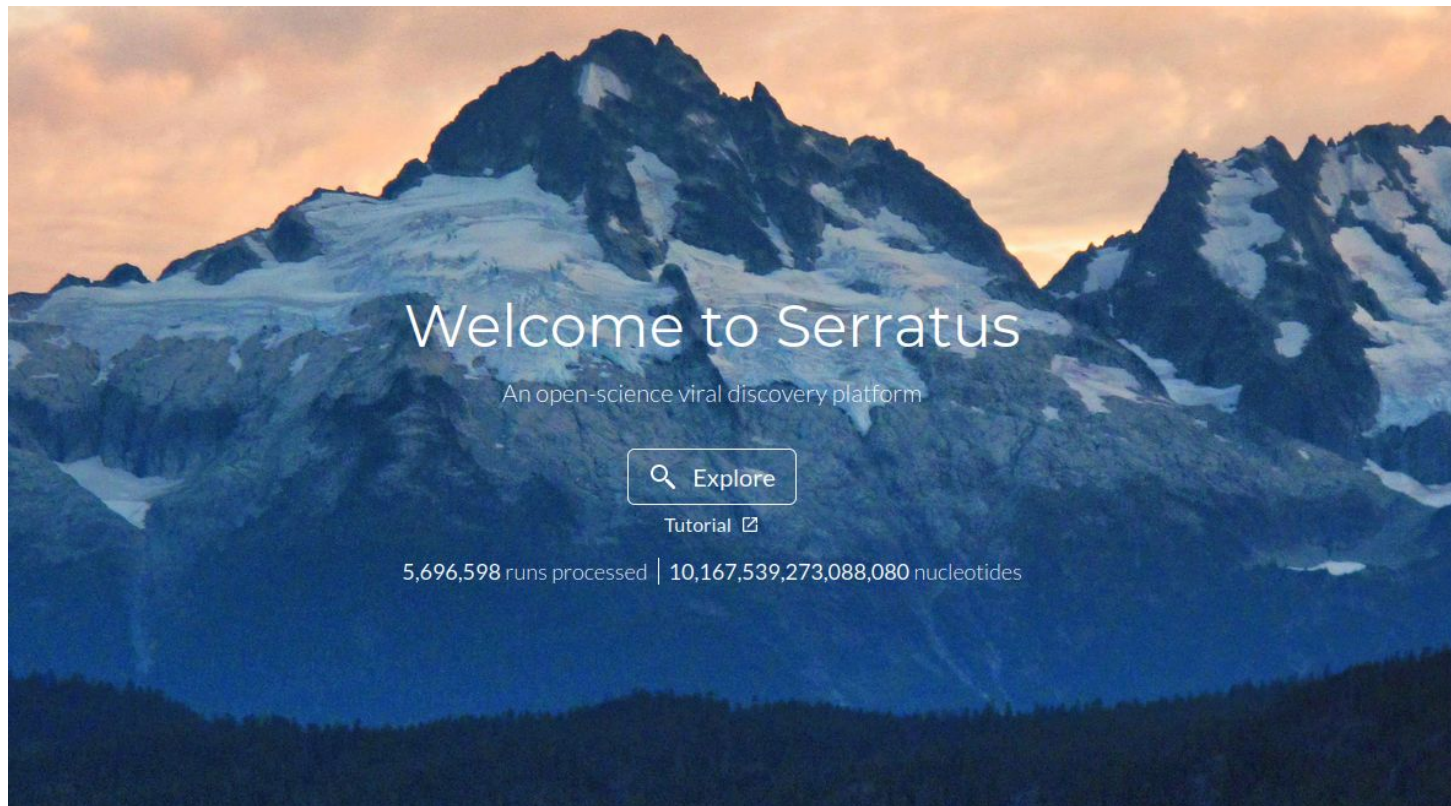


$((10 \text{ petabytes}) / (1 \text{ megabyte})) / (\text{seconds per year}) =$

316.887646408

years at 1 MB/s

Serratus: 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

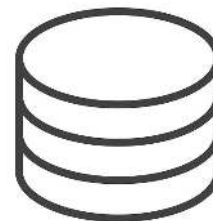
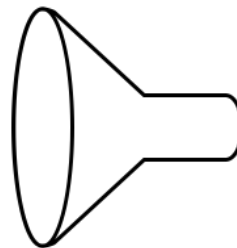
> Discovered 130,000 new RNA virus species

Analysis 1:



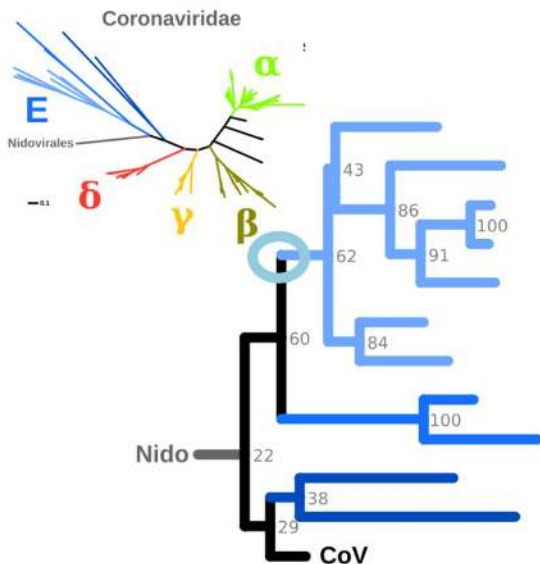
All RNA-seqs

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



**55,715 CoV+
samples**

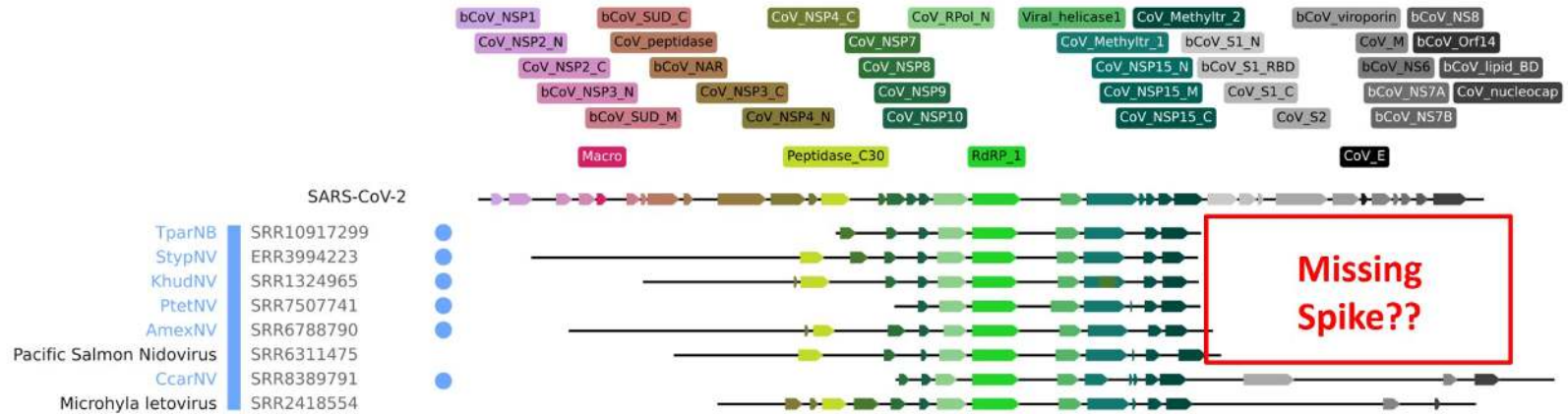
Discovering new Coronaviruses



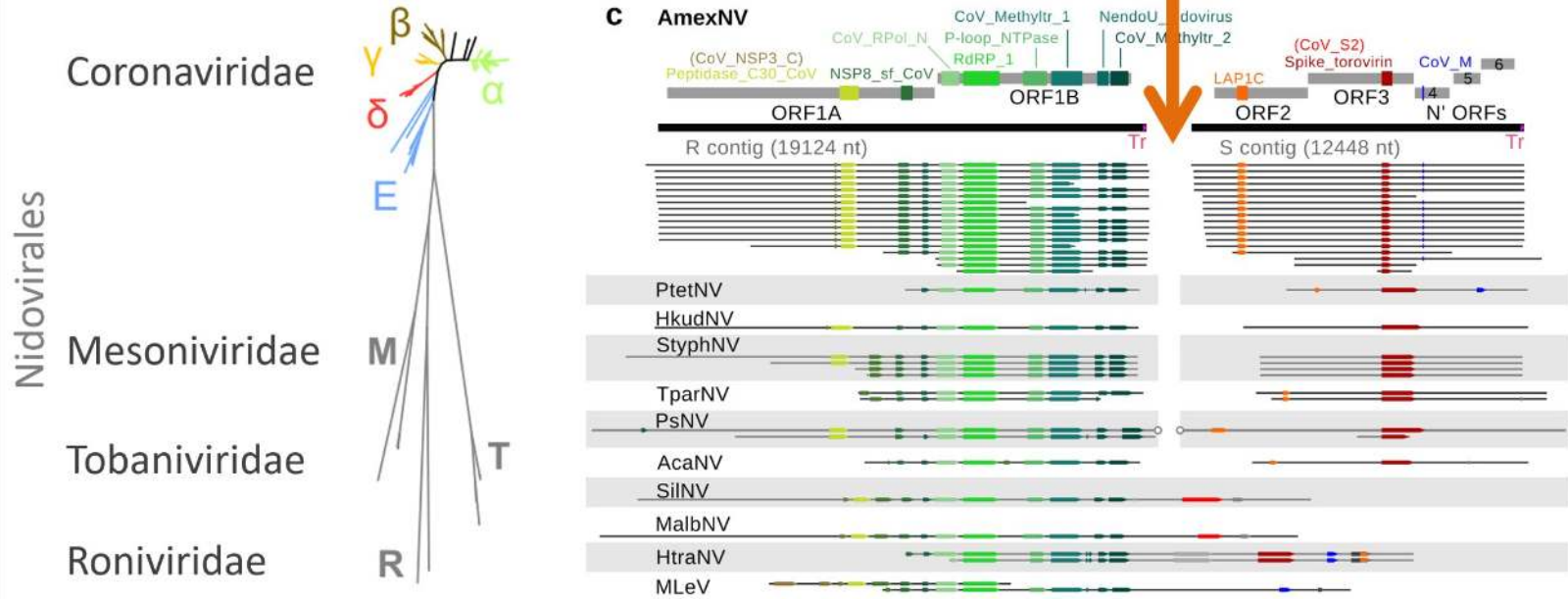
AmexNV SRR6788790
PtetNV SRR7507741
HkudNV SRR1324965
StypNV ERR3994223
TparNV SRR10917299
Pacific Salmon Nidovirus
AcaNV SRR5997671
SilNV SRR12184956
MalbNV SRR10402291
HtraNV SRR8389791
Microhyla Letovirus



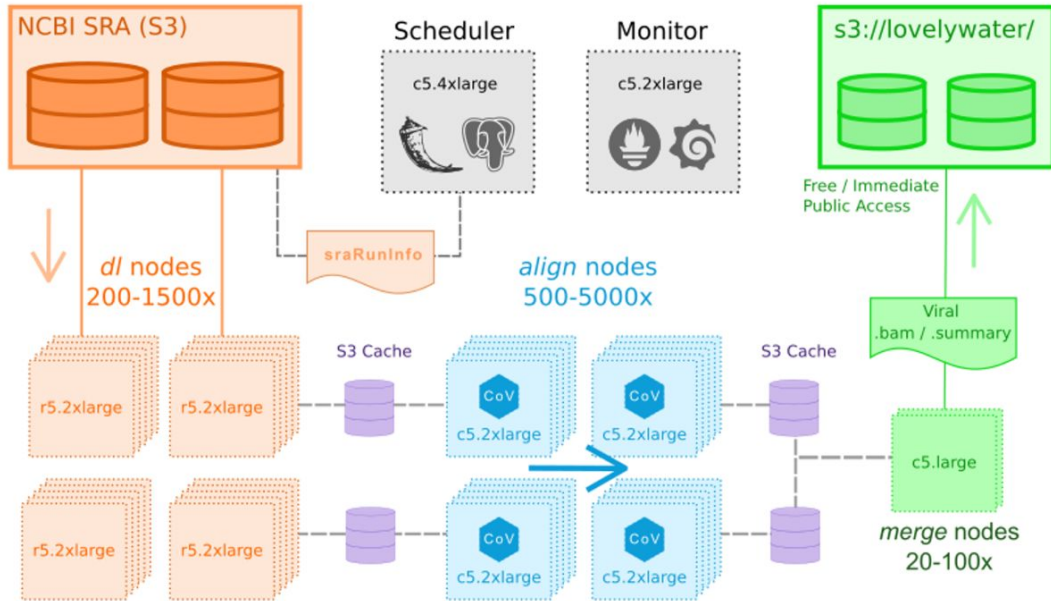
Discovering new Coronaviruses



Segmented Coronaviruses?

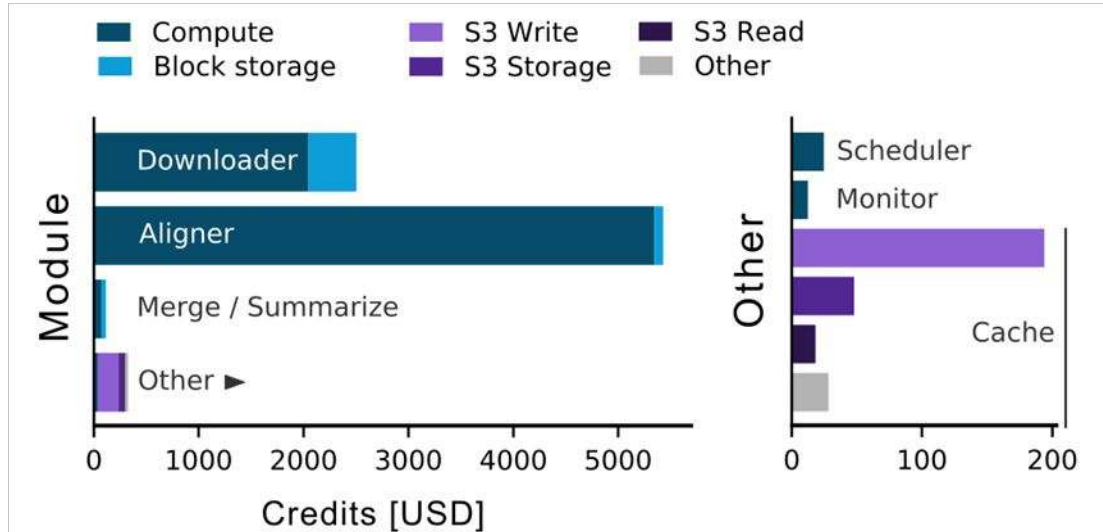
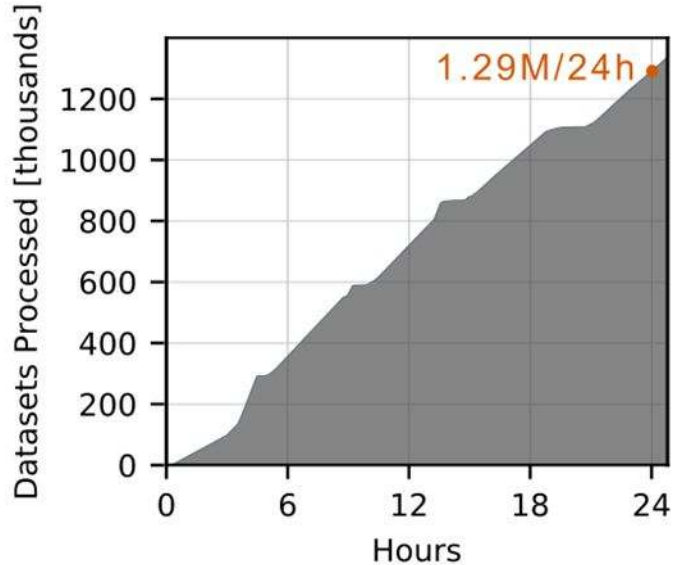


Serratus architecture



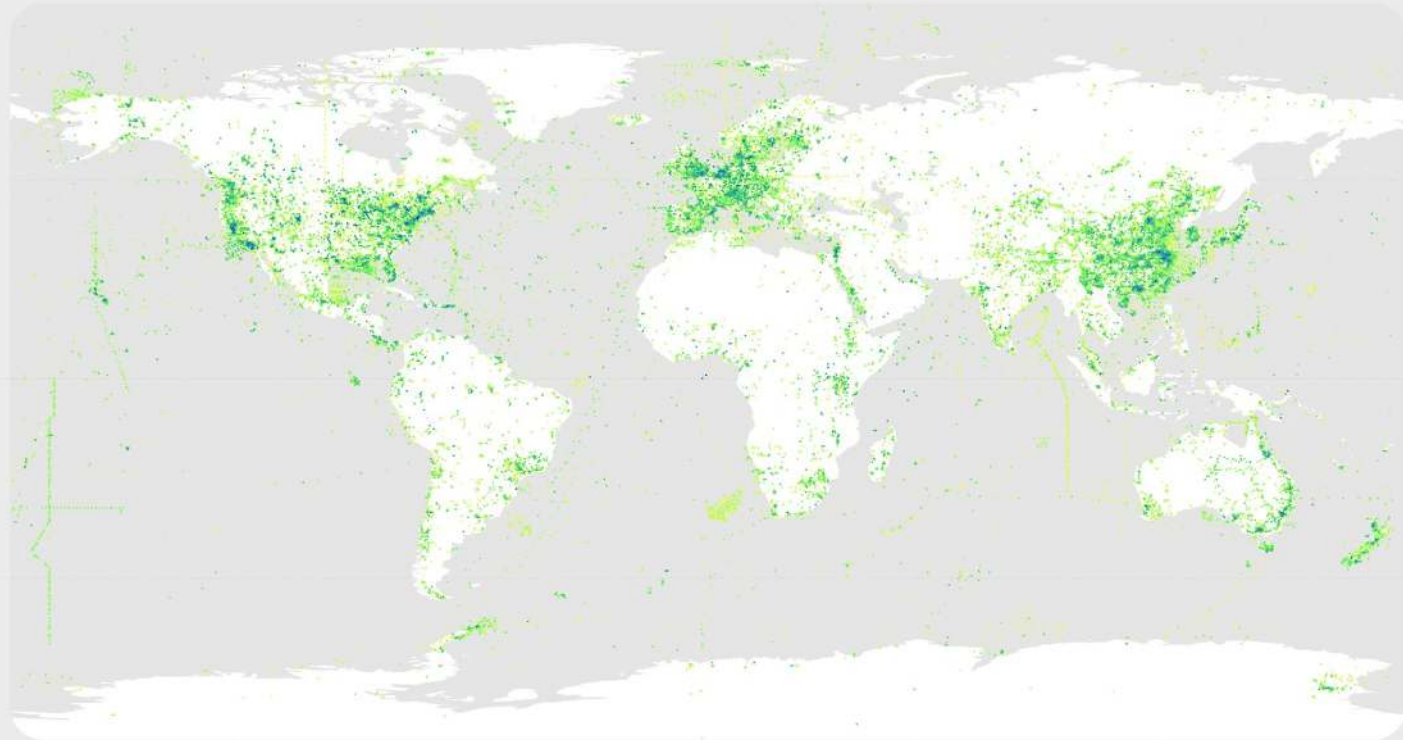
- Aggressively cost-optimized
- Native access to SRA on S3
- Dynamic scaling up to ~22,250s vCPU
- Open Source: GPLv3

Serratus performance & costs



1 million NGS libraries / day
\$0.005 / library

Geography of SRA samples



1 20 400 8000
Sequencing density (datasets)

Planetary DNA/RNA sequencing

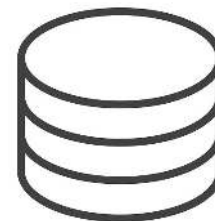
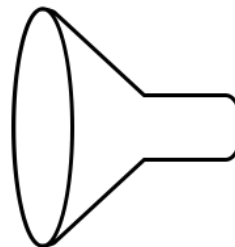


Analysis 2:



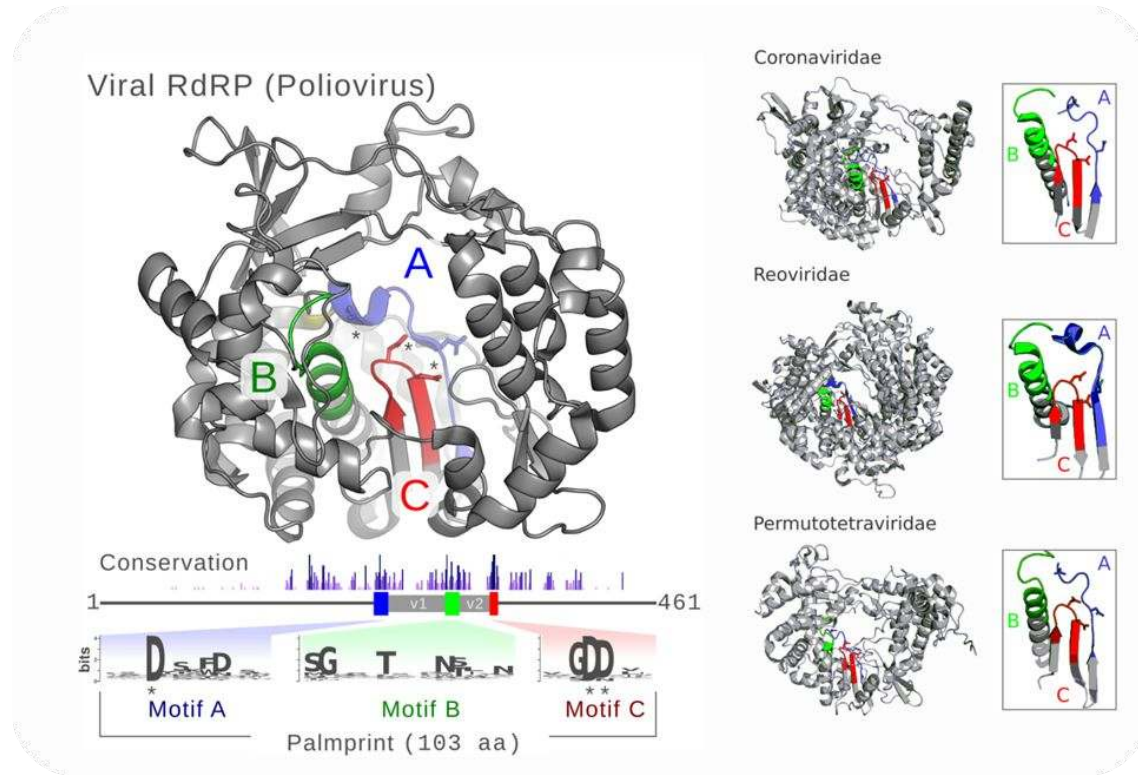
All RNA-seqs

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



**aligned reads
(.bam files)**

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



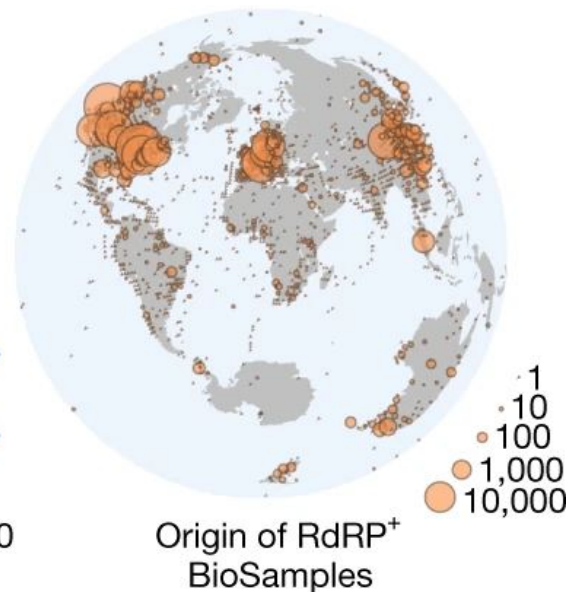
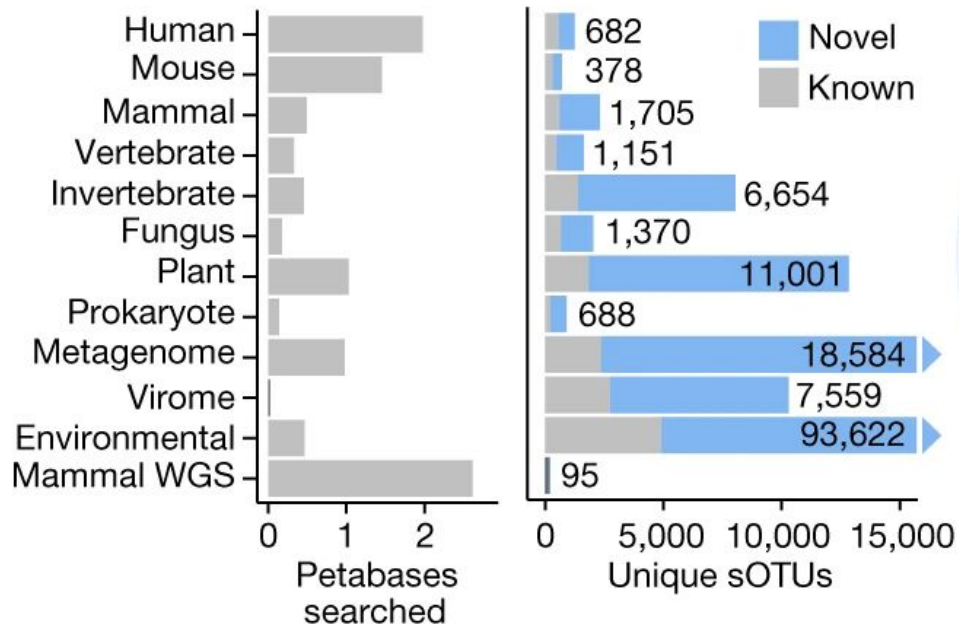
(Babaian & Edgar, 2021. bioRxiv)

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

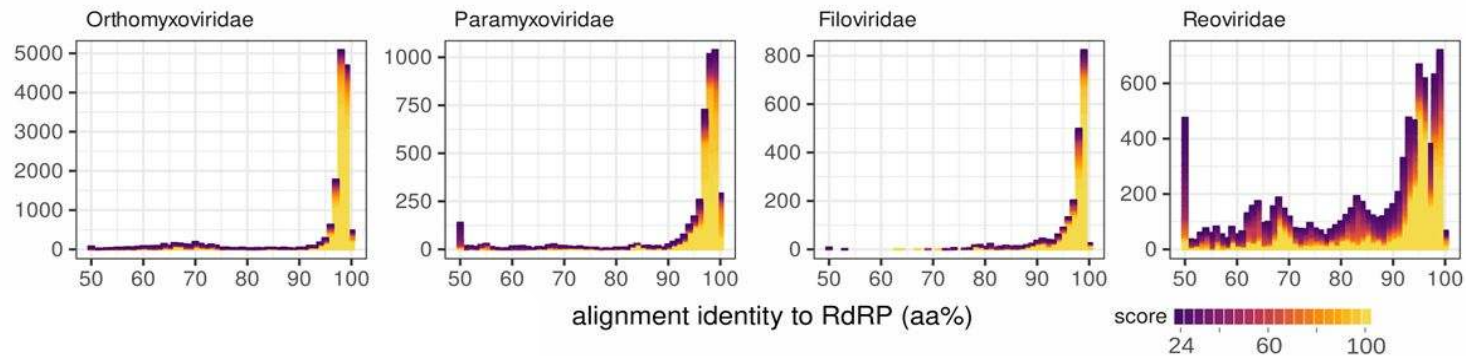
Assembly of all viral RdRPs (Analysis 2)

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

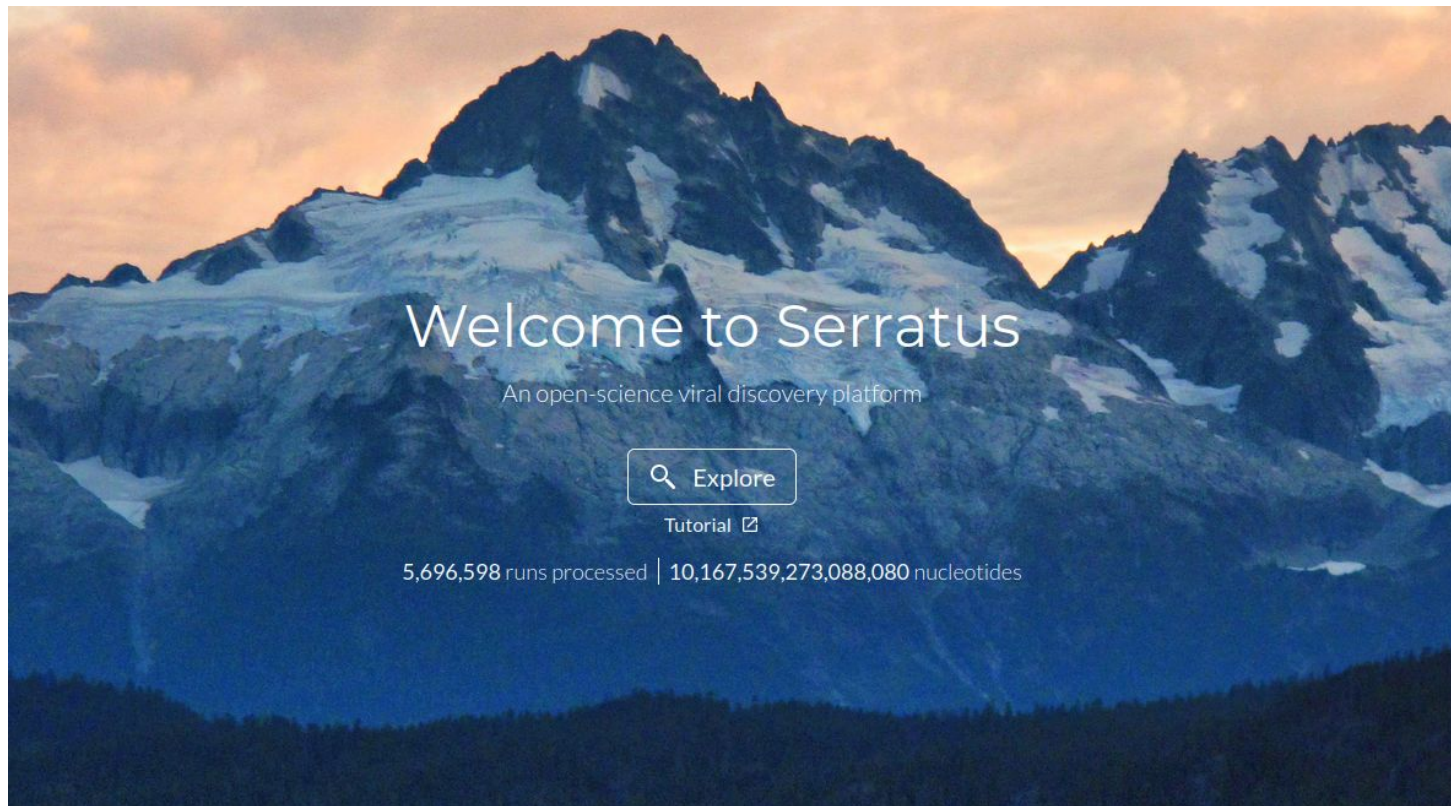
- SPAdes assembler & GNU parallel
- Single large AWS instance (`c6a.48xlarge`, 192 cores)



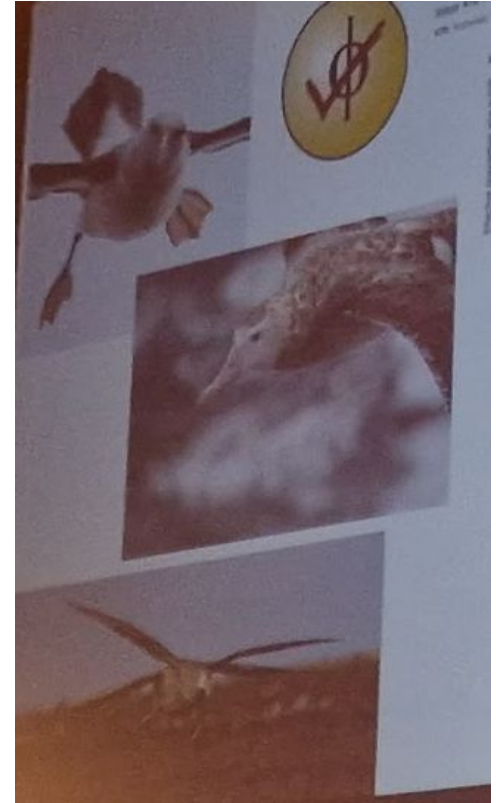
Number of samples



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

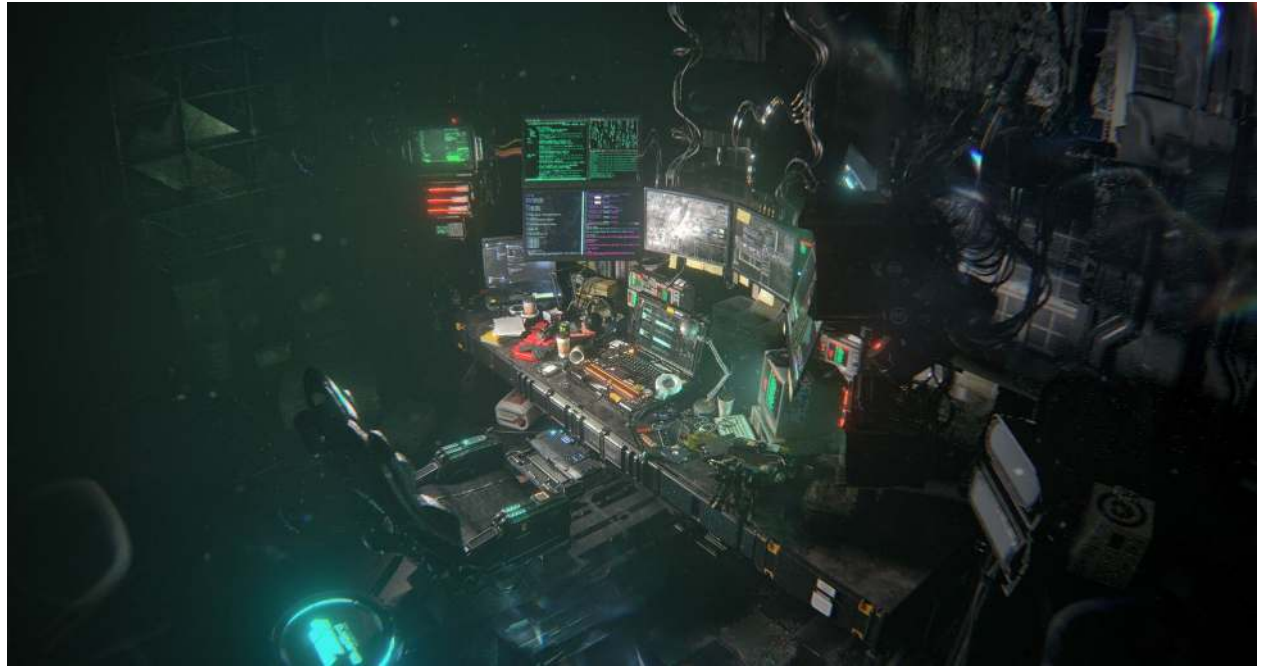


How was all of this large-scale assembly done?



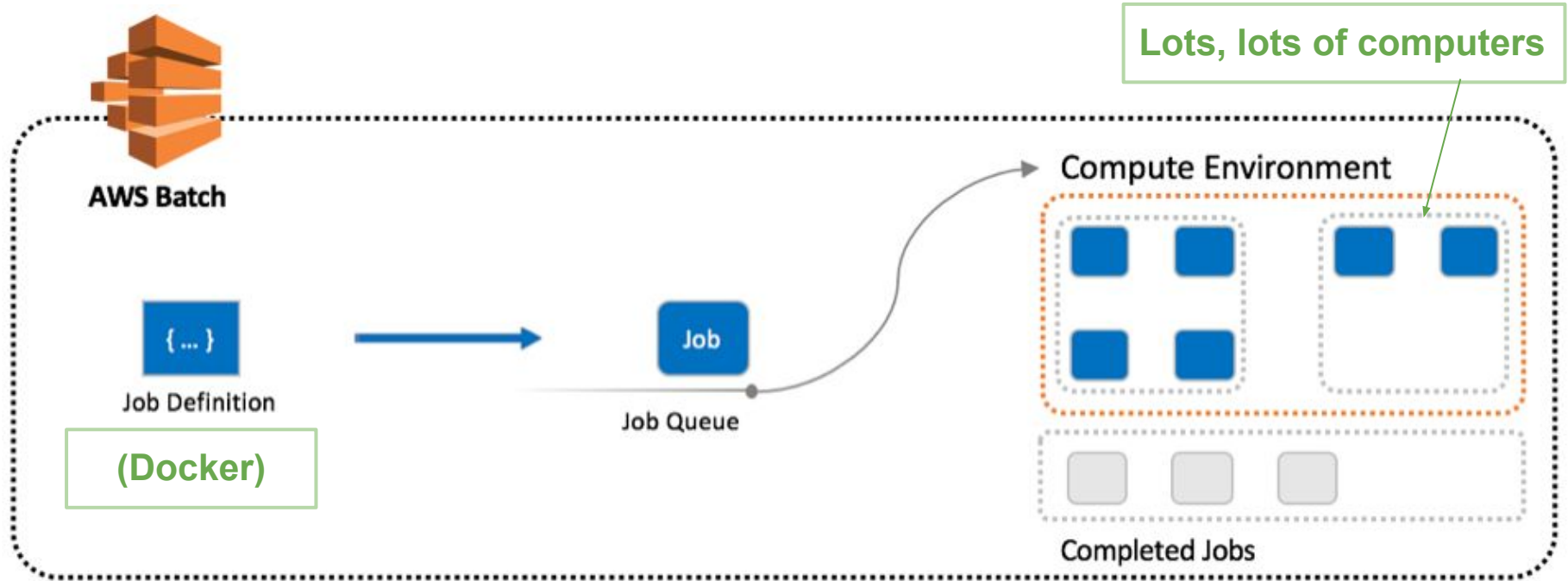
How was all of this large-scale assembly done?

cloud scripting



* (artist rendition)

AWS Batch framework for large-scale assembly



Peak:
~28,000 vCPUs

☐	Name ▾	Instance ID ▾	Instance Type ▾	Availability Zone ▾	Instance State ▲
(screenshot: P. Barbera)					
☐	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

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	 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?

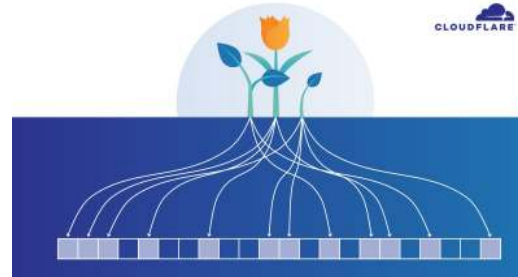
- DNA viruses
- Lower homology detection with known RdRPs
 - Replacing Bowtie 2 / Diamond by ...?
- A global **index of the SRA**
 - nearly feasible with k-mers already
 - would only support exact search
 - with ML, could do low(er) homologies

Deep embedding and alignment of protein sequences

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

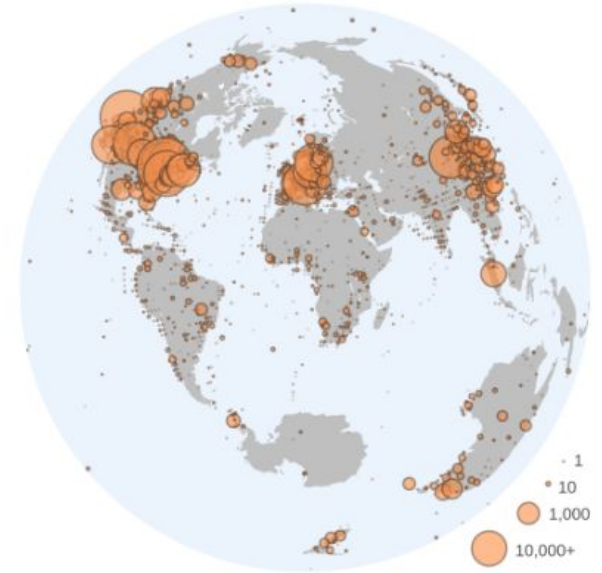
Google Research, Brain team, Paris, France

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

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 exponentially¹. 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 10⁵ 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.

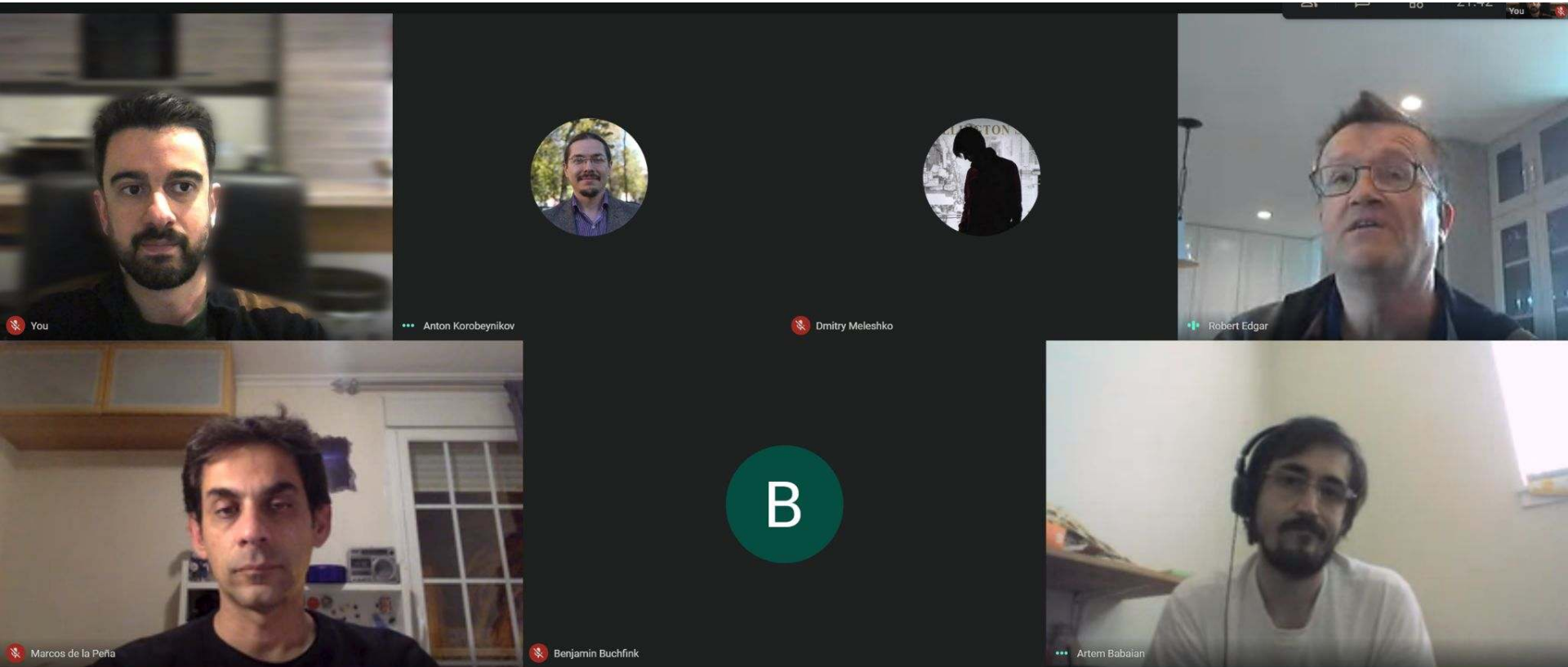


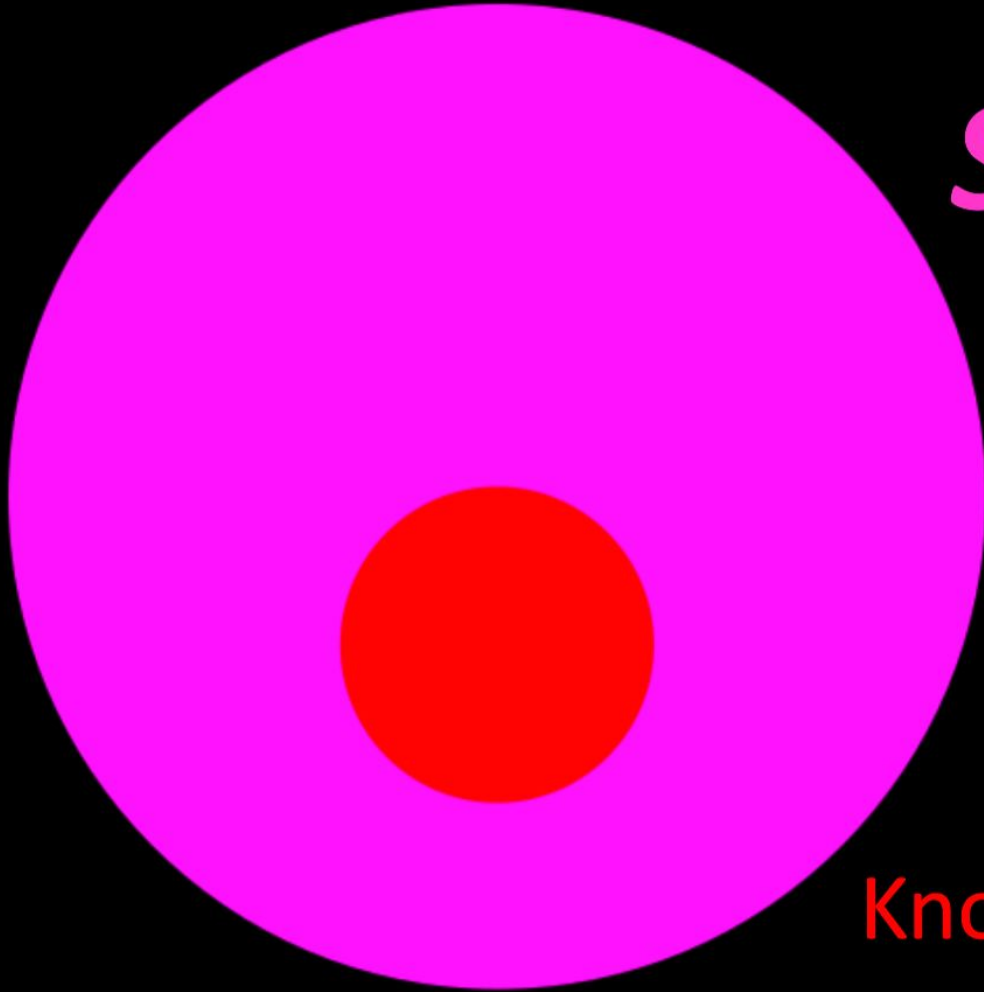
Digital Collaboration

- 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 (Gainesville)

All equal contributions

We never met IRL





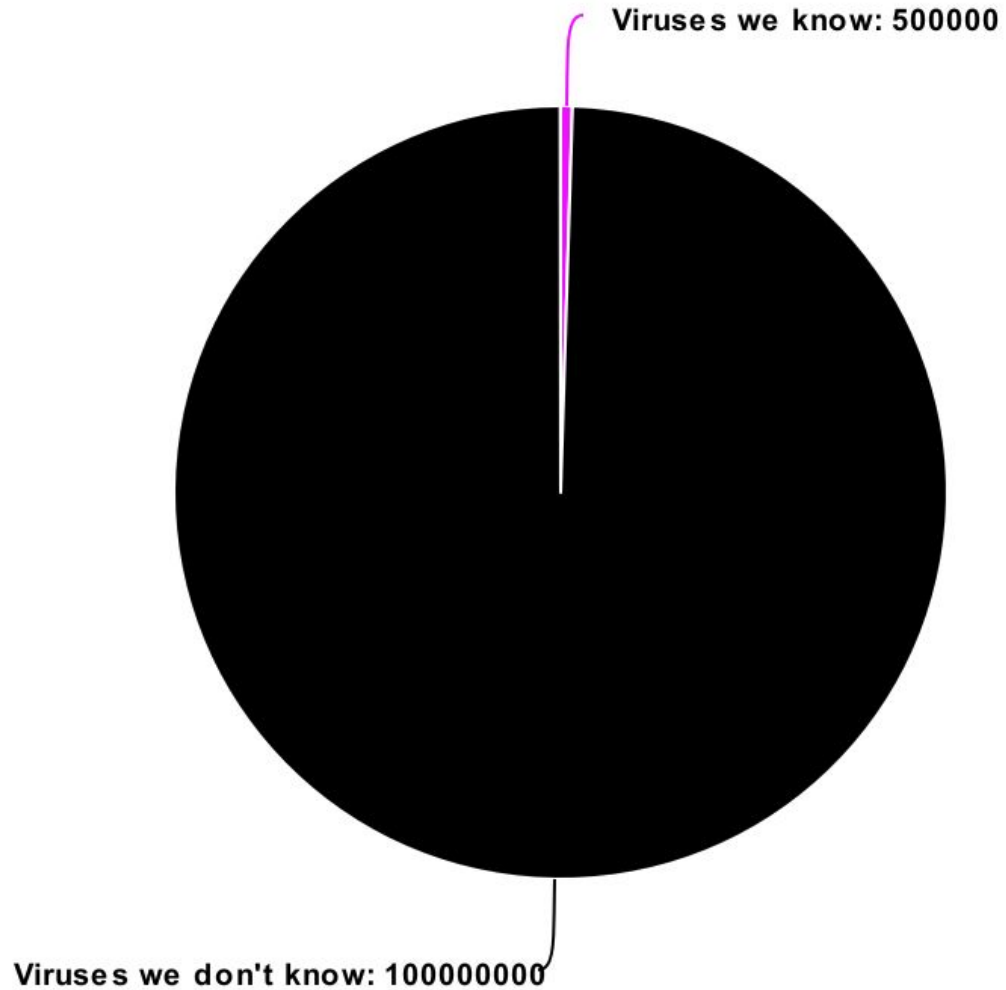
Serratus

Known RNA Virome

Earth's Virome

We are here





Pie chart for
Josie

Outro

bigger data

big data



WE'RE-GONNA-NEED



A BIGGER INSTANCE TYPE

Sequence Bioinformatics

@ Institut Pasteur



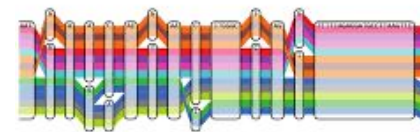
Genomes &
metagenomes
assembly



Algorithms and
data structures
on k-mers



Sequence
search in very
large datasets



Pangenomics

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...

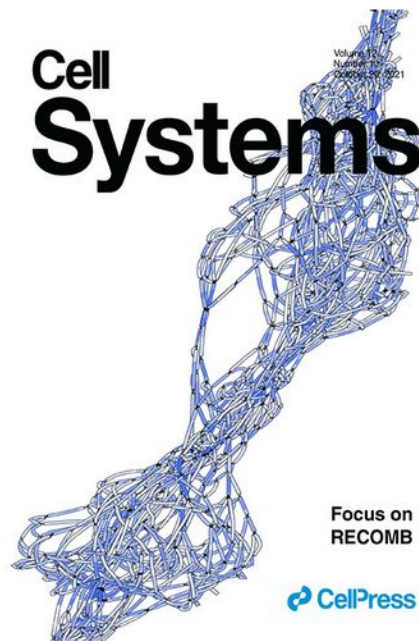
We previously introduced **minimizer-space de Bruijn graphs (mdBGs)**, where

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

Classical alphabet: $\Sigma_{\text{DNA}} = \{A, C, G, T, N\}$

A k -mer with $k = 3$: AGT

Minimizer alphabet: Set of ℓ -mer minimizers as letters



A detour...

Fixed set of
universe minimizers



AATGACATGATCATGA

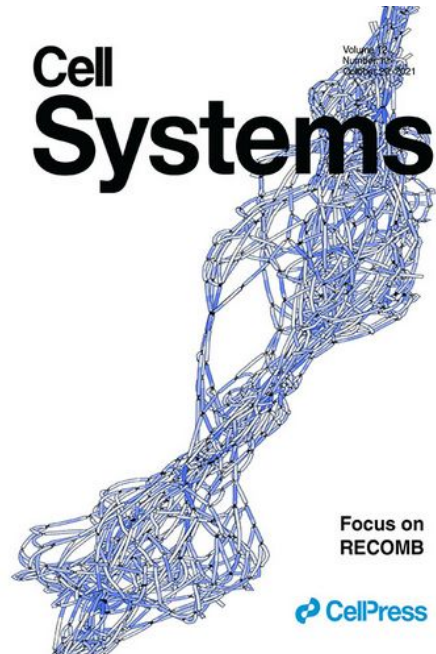
GA

TC

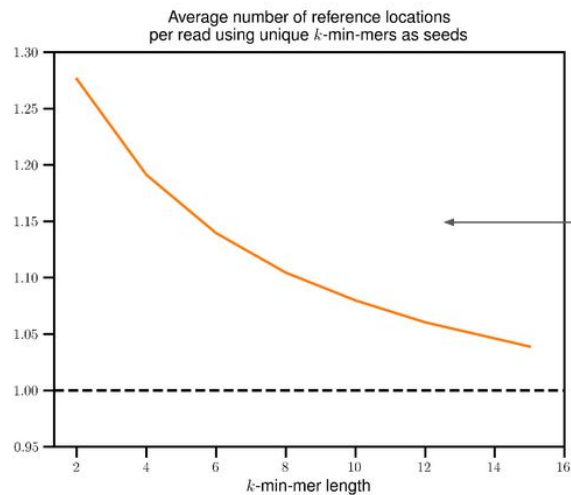
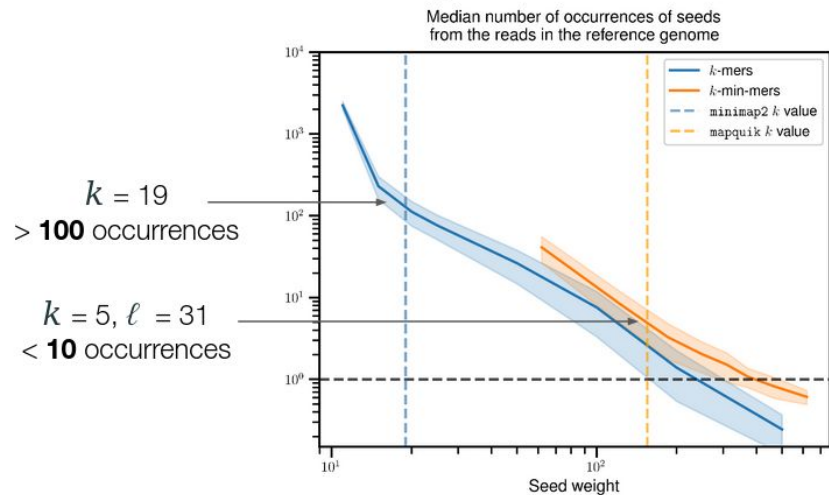
$\Sigma^\ell = \{ \text{all minimizers of length } \ell \} = \{m_1, m_2, m_3, \dots\}$

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?



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