SNIA CSTI | CLOUD STORAGE TECHNOLOGIES

Moving Genomics to the Cloud: Compute and Storage Considerations

Live Webcast

September 9, 2021

10:00 am PT / 1:00 pm ET

Today's Presenters

Alex McDonald
Independent Consultant
Chair, SNIA Cloud Technologies Initiative

Michael McManus
Director, Precision Medicine & Principal Engineer
Intel

Torben Kling Petersen
Principal Engineer
HPC Storage BU
HPE

Christopher Davidson
Application & Performance Engineering, Manager
HPE

SNIA-at-a-Glance

180
industry leading organizations

2,500
active contributing members

50,000
IT end users & storage pros worldwide

Learn more: **snia.org/technical** @SNIA

SNIA. CSTI | CLOUD STORAGE TECHNOLOGIES

SNIA. CSTI | CLOUD STORAGE TECHNOLOGIES

What We Do

Educate vendors and users on cloud storage, data services and orchestration

Support & promote business models and architectures: OpenStack, Software Defined Storage, Kubernetes, Object Storage

Understand Hyperscaler requirements
Incorporate them into standards and programs

Collaborate with other industry associations

SNIA Legal Notice

The material contained in this presentation is copyrighted by the SNIA unless otherwise noted.

Member companies and individual members may use this material in presentations and literature under the following conditions:

- Any slide or slides used must be reproduced in their entirety without modification
- The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.

This presentation is a project of the SNIA.

Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.

The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

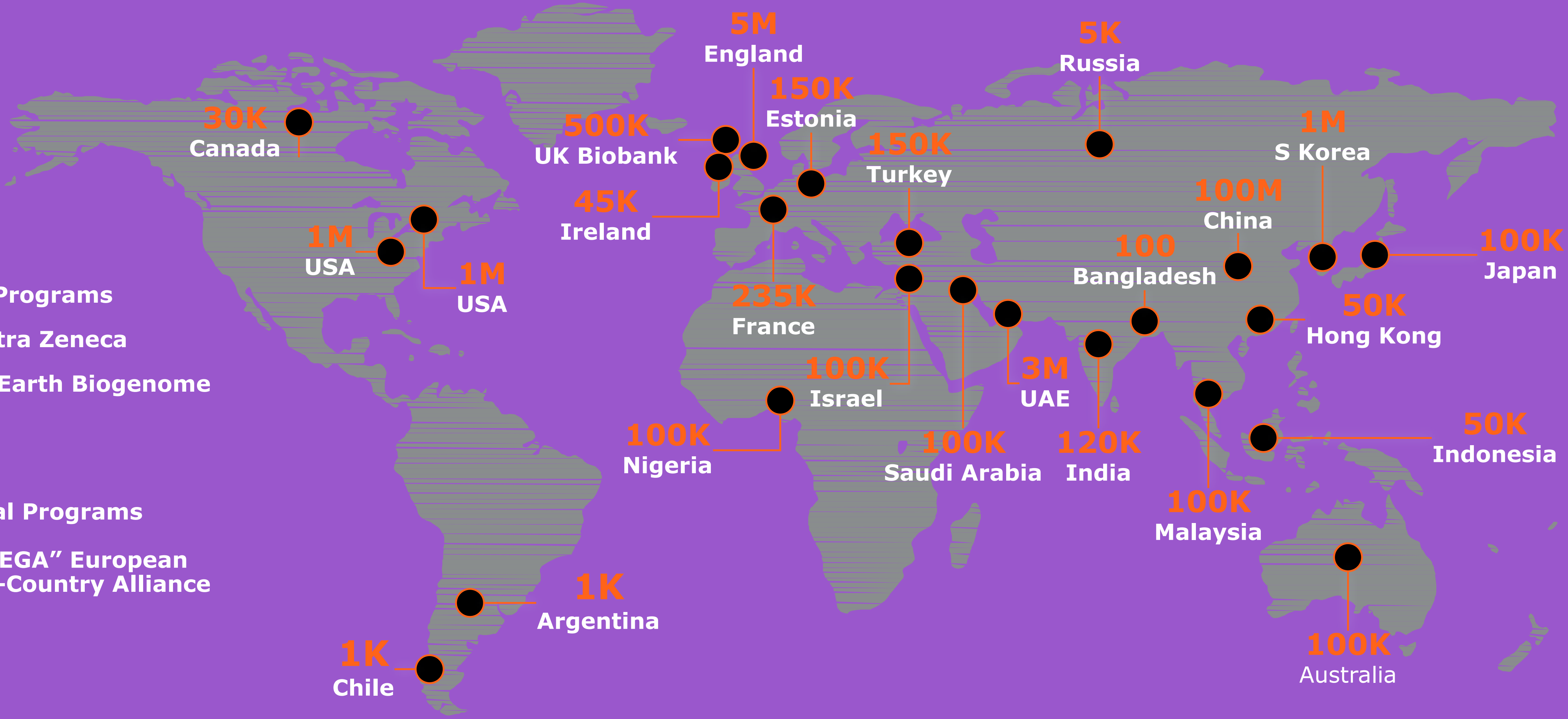
Agenda

- Trends in Genomics & the Need for Data Storage & Management
- Data Management & Storage Considerations
- Public Science in Practice

Trends in Genomics

Michael McManus

Large Scale Sequencing is already underway!



Global Programs

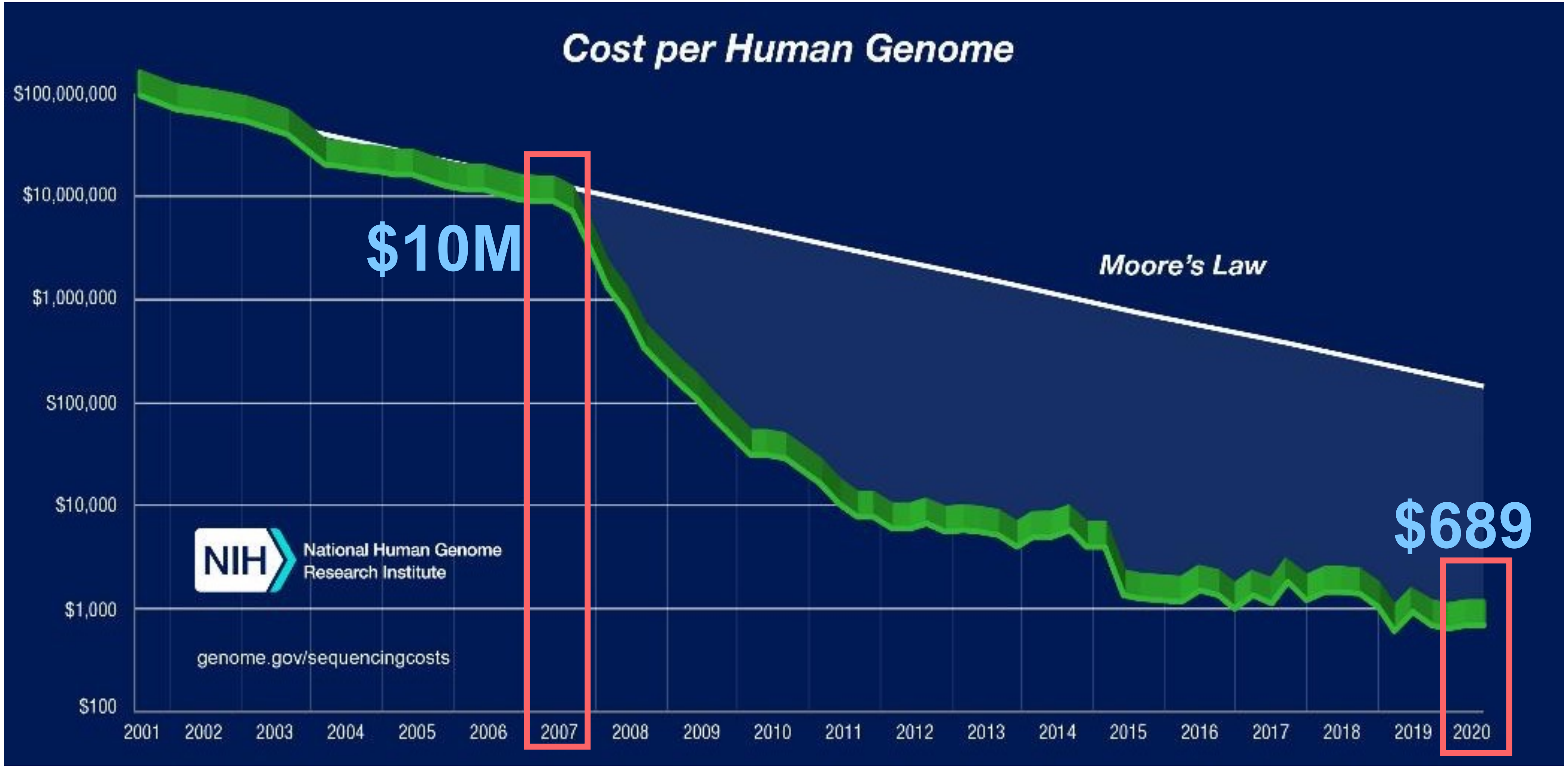
- **2M** Astra Zeneca
- **1.5M** Earth Biogenome

Regional Programs

- **1M** "MEGA" European 20-Country Alliance

- Source: Frost & Sullivan, "Global Precision Market Growth Opportunities, Forecast to 2025," January 2017 and Intel's own market research 2017-2021

Cost per Genome is Dropping



- Source: "Sequencing Data Cost", from NHGRI, https://www/genome.gov/sequencingcostsdata/

Sequencers commonly used for Genomics


Illumina MiniSeq®


Illumina iSeq® 100


Illumina MiSeq®


Illumina NextSeq® 550


Illumina NextSeq® 1000/2000


Illumina NovaSeq 6000®


DNBSEQ-G50


DNBSEQ-G400


DNBSEQ-T7


ThermoFisher Ion GeneStudio S5


ThermoFisher Ion Torrent Genexus

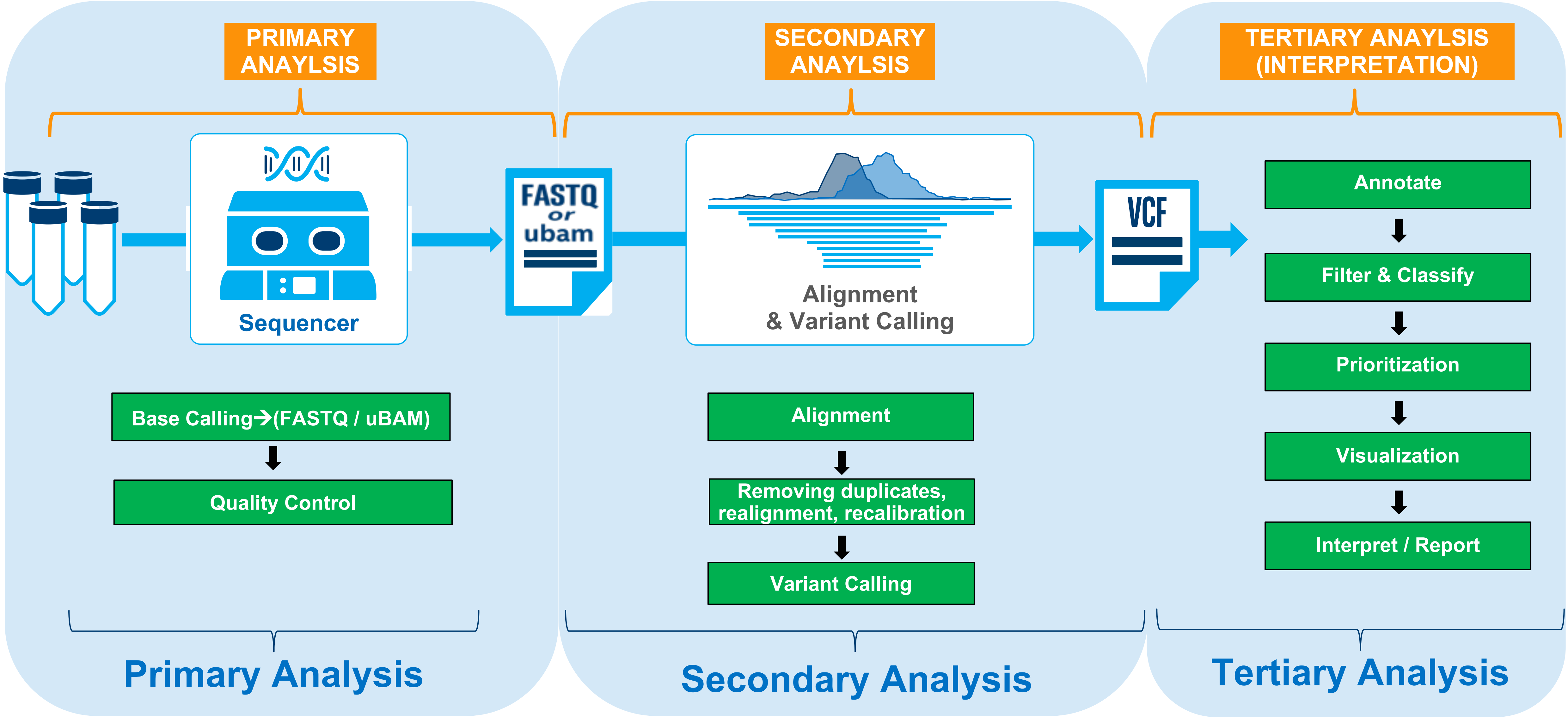

PacBio Sequel IIe System


Oxford Nanopore MinION


Oxford Nanopore GridION

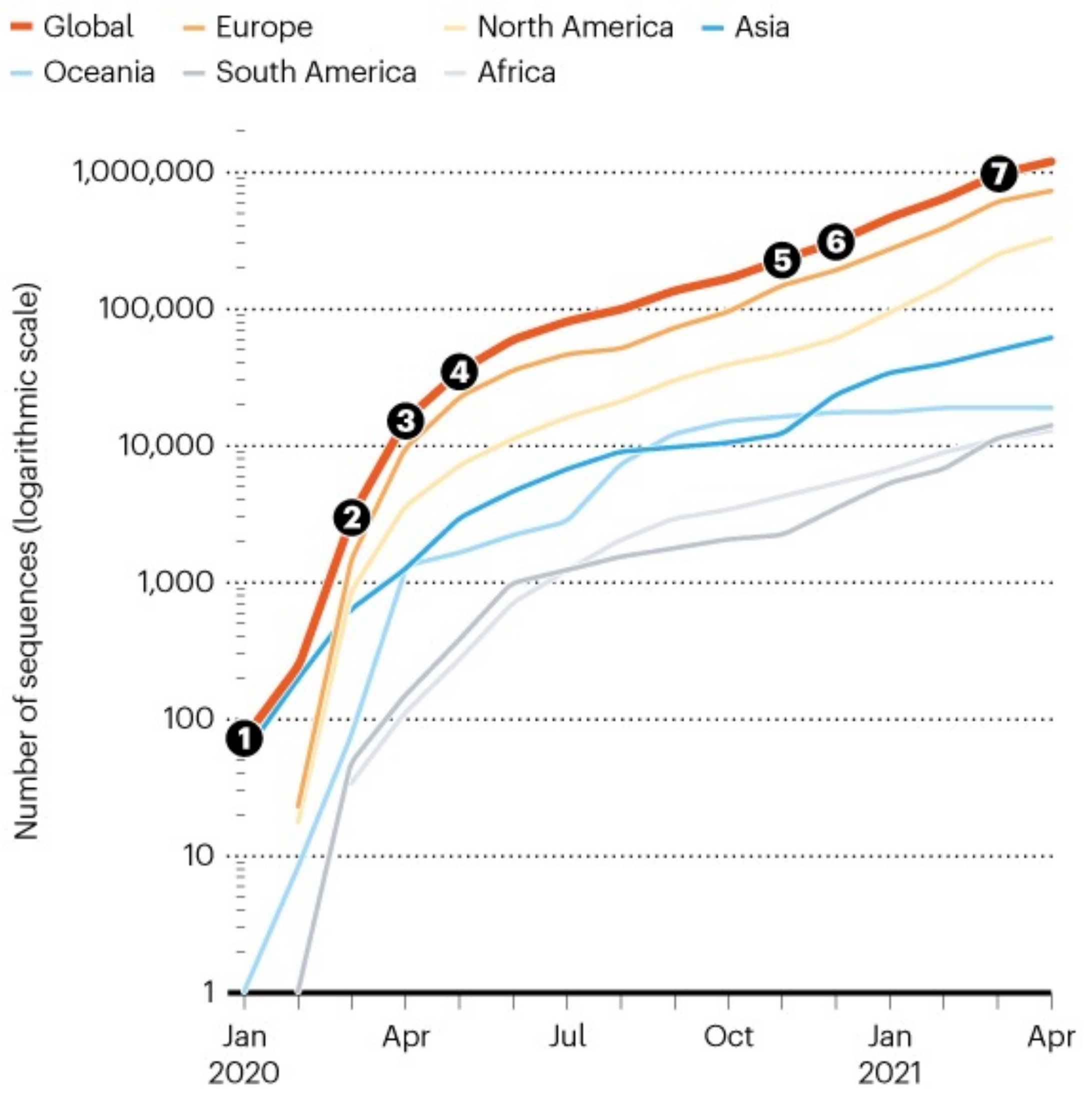

Oxford Nanopore PromethION

Overview of the Human Genomics Workflow



SNIA. CSTI | CLOUD STORAGE TECHNOLOGIES

Collaboration in the Time of COVID

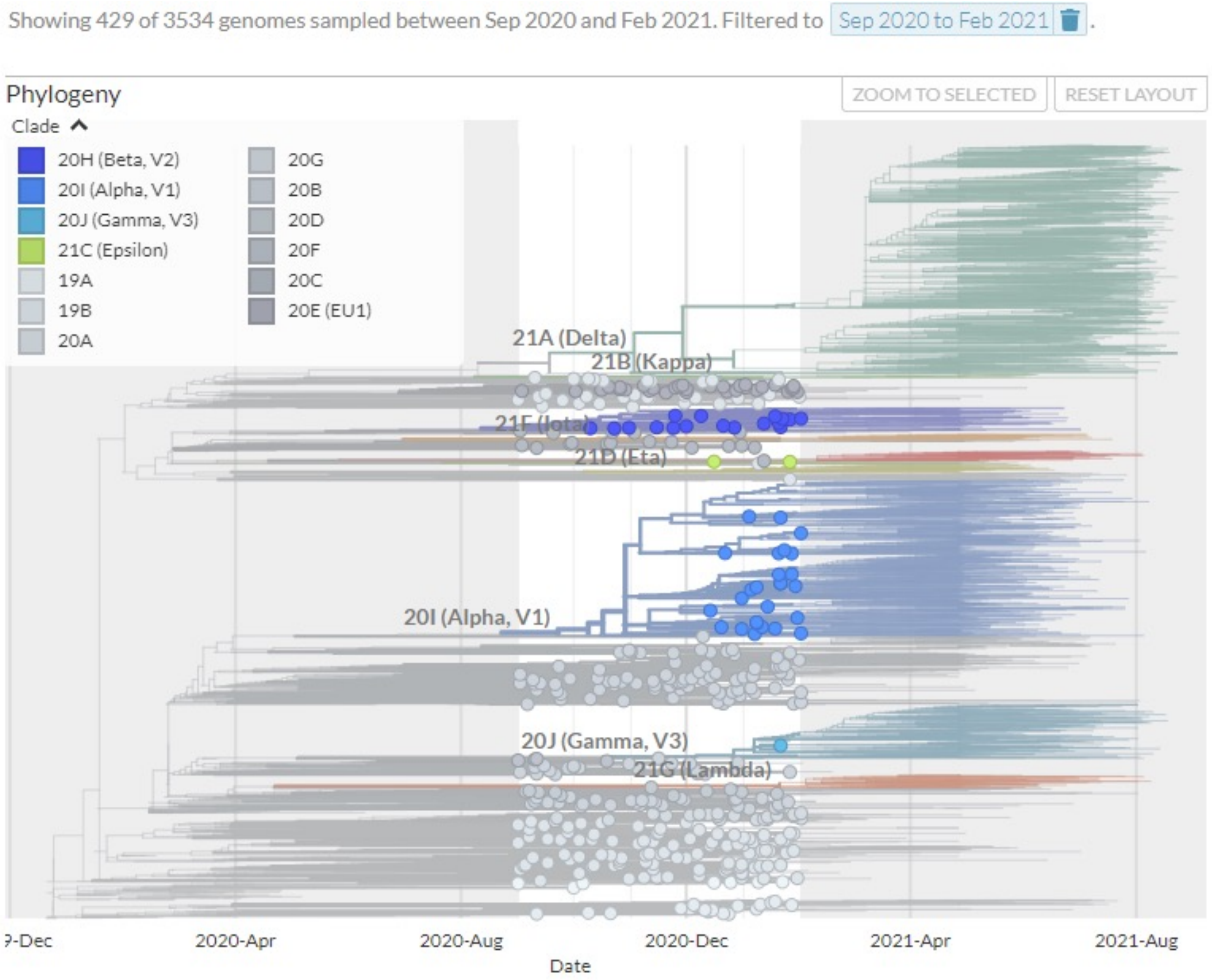


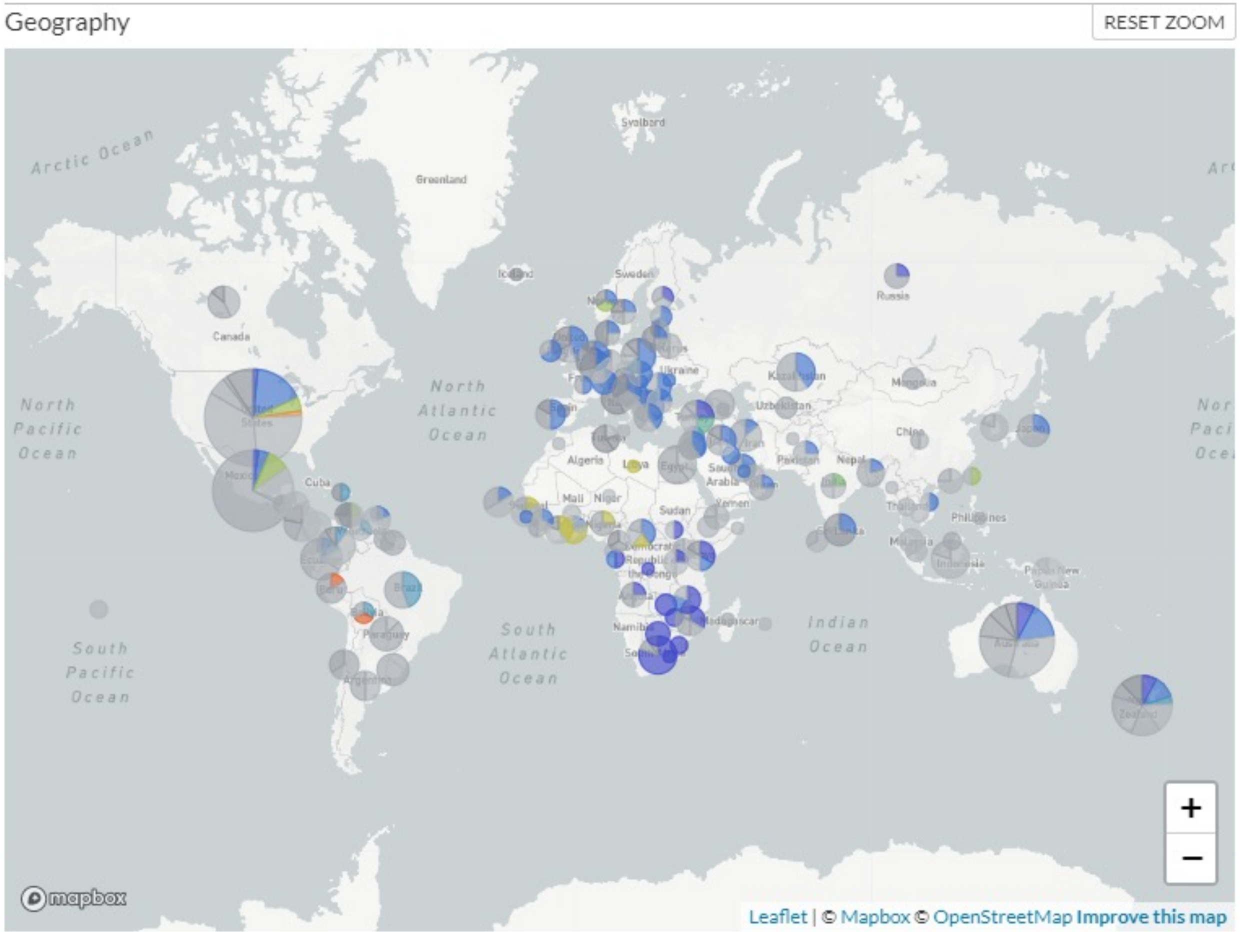
More than one million SARS-CoV-2 genome sequences have been shared on the GISAID data-sharing platform since January 2020, and are helping researchers to track the spread of viral variants. Most are from the United States and Europe, but contributions come from every region of the world.

1. **January:** First SARS-CoV-2 genome, from China.
2. **March:** First African sequence, from Nigeria.
3. **April:** Victoria, Australia, has 1,300 cases; 80% are sequenced, identifying clusters from cruise ships and hospitality venues.
4. **May:** UK sequences 6% of cases, more than any other country.
5. **November:** South African surge prompts intensified surveillance. Researchers find a widespread new variant - B.1.351.
6. **December:** 40% of genomes sequenced in Manaus, Brazil, are of the P.1 variant, with mutations linked to increased transmissibility and immune evasion.
7. **March:** US sequencing rate doubles, owing to a government mandate for surveillance and funding from the Centers for Disease Control and Prevention.

https://media.nature.com/lw800/magazine-assets/d41586-021-01069-w/d41586-021-01069-w_19094110.png

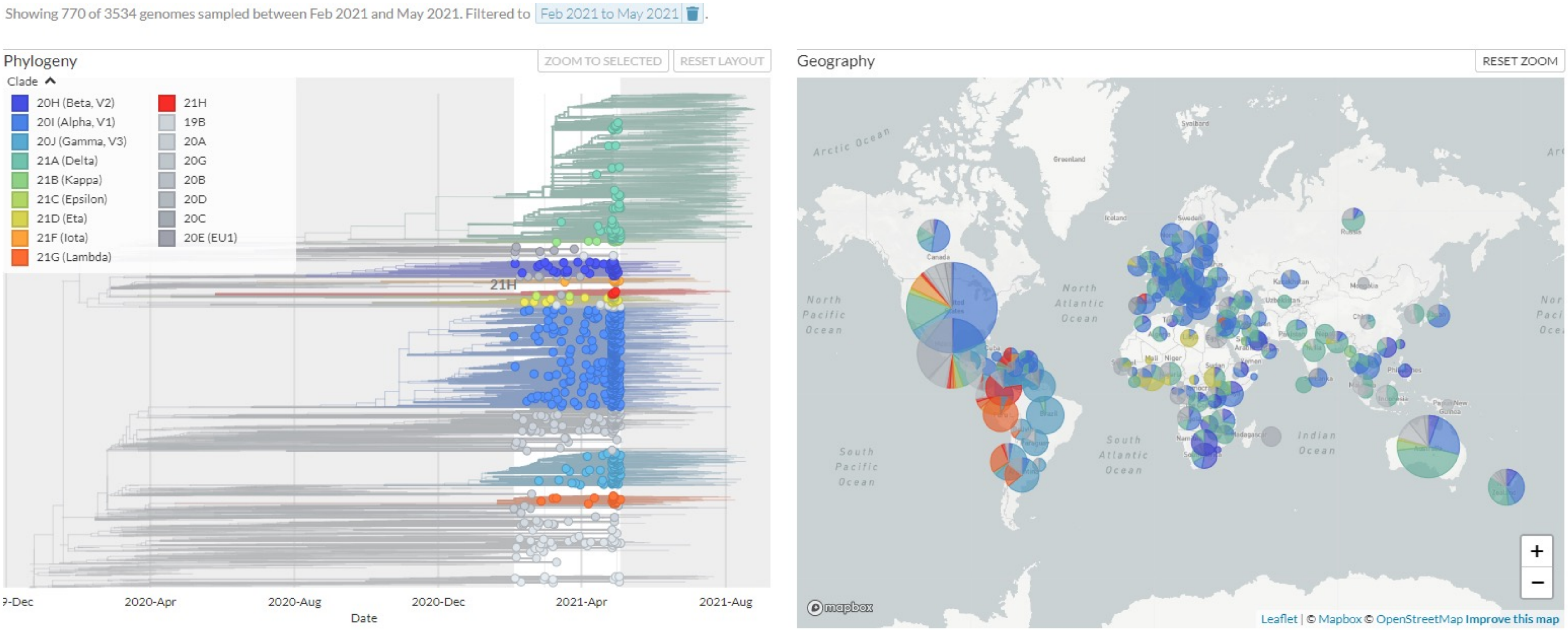
Genomic epidemiology of novel coronavirus - Global subsampling





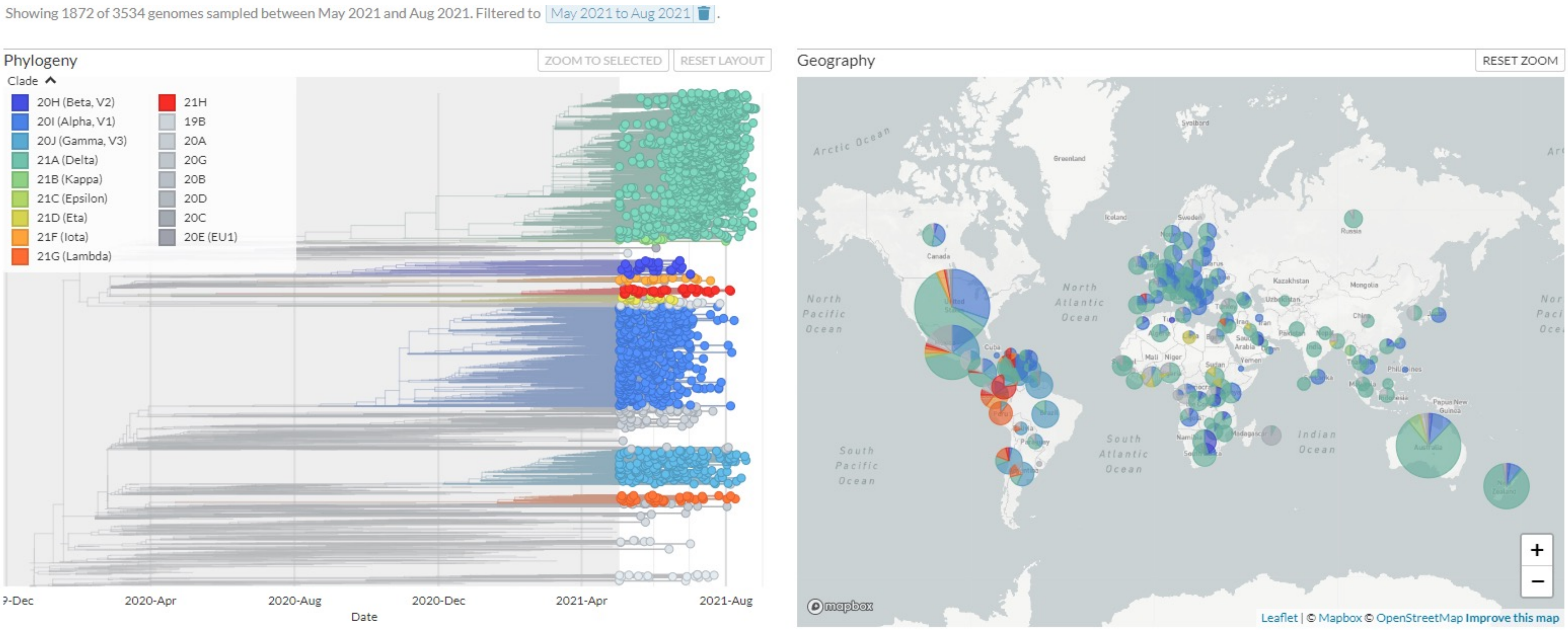
SNIA. CSTI | CLOUD STORAGE TECHNOLOGIES

Genomic epidemiology of novel coronavirus - Global subsampling



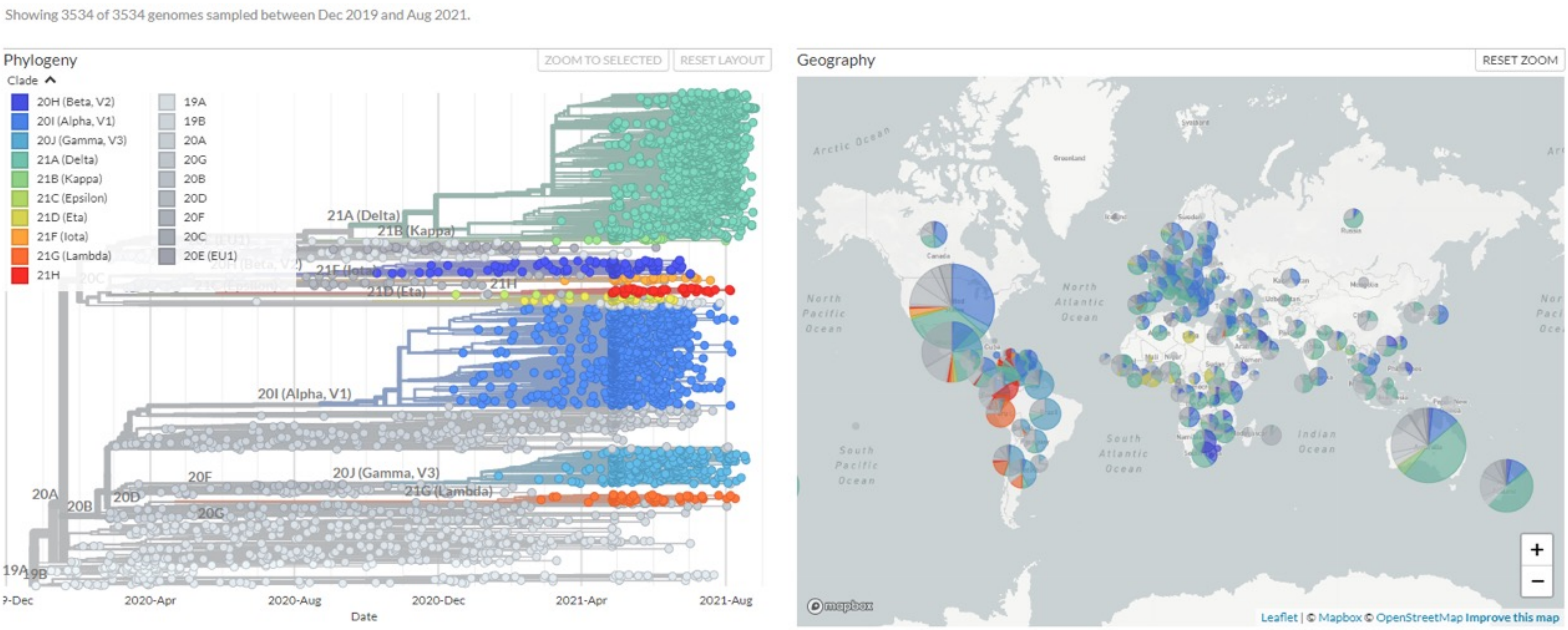
SNIA. CSTI | CLOUD STORAGE TECHNOLOGIES

Genomic epidemiology of novel coronavirus - Global subsampling

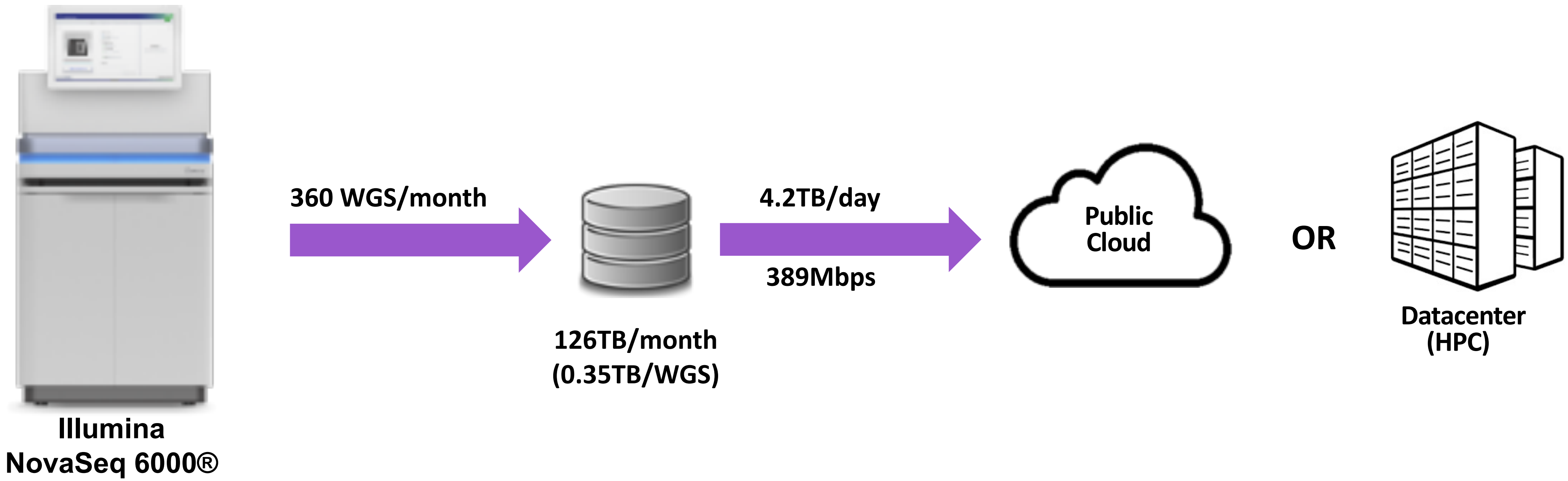


SNIA. CSTI | CLOUD STORAGE TECHNOLOGIES

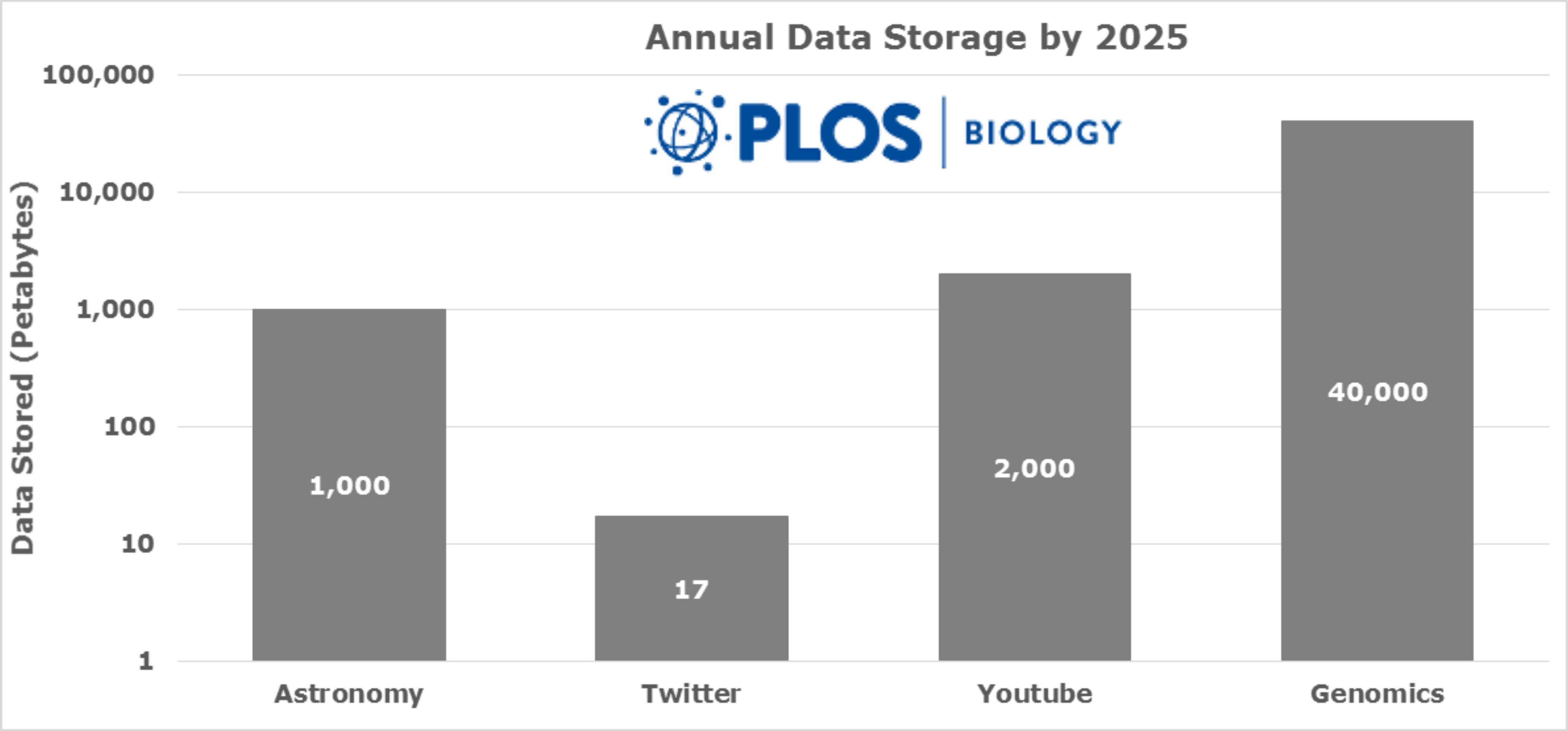
Genomic epidemiology of novel coronavirus - Global subsampling



Data Volume Example

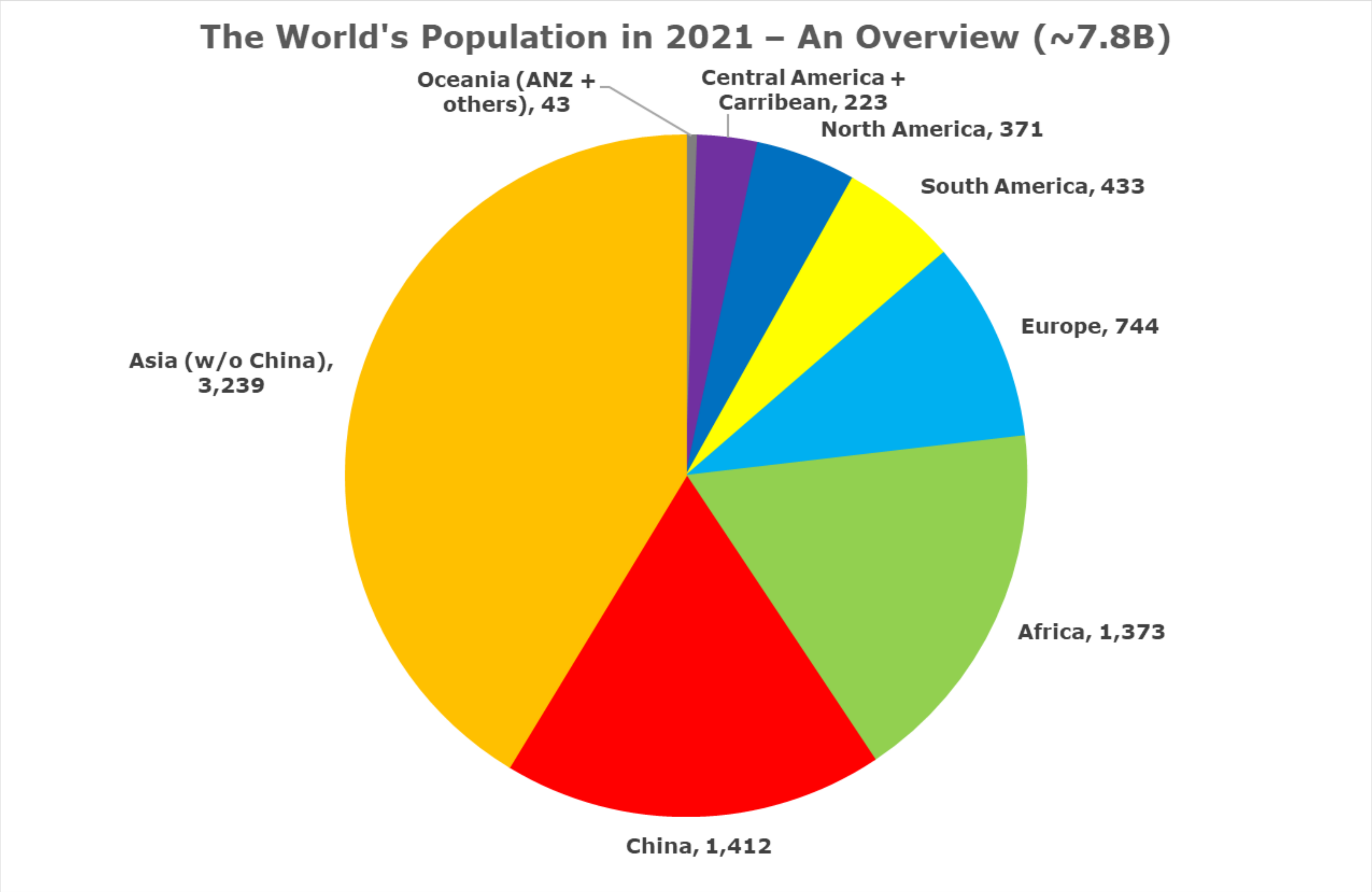


Four Data Storage Domains of Big Data in 2025



Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, et al. (2015) Big Data: Astronomical or Genomical?. PLoS Biol 13(7): e1002195. doi:10.1371/journal.pbio.1002195 http://journals.plos.org/plosbiology/article?id=info:doi/10.1371/journal.pbio.1002195

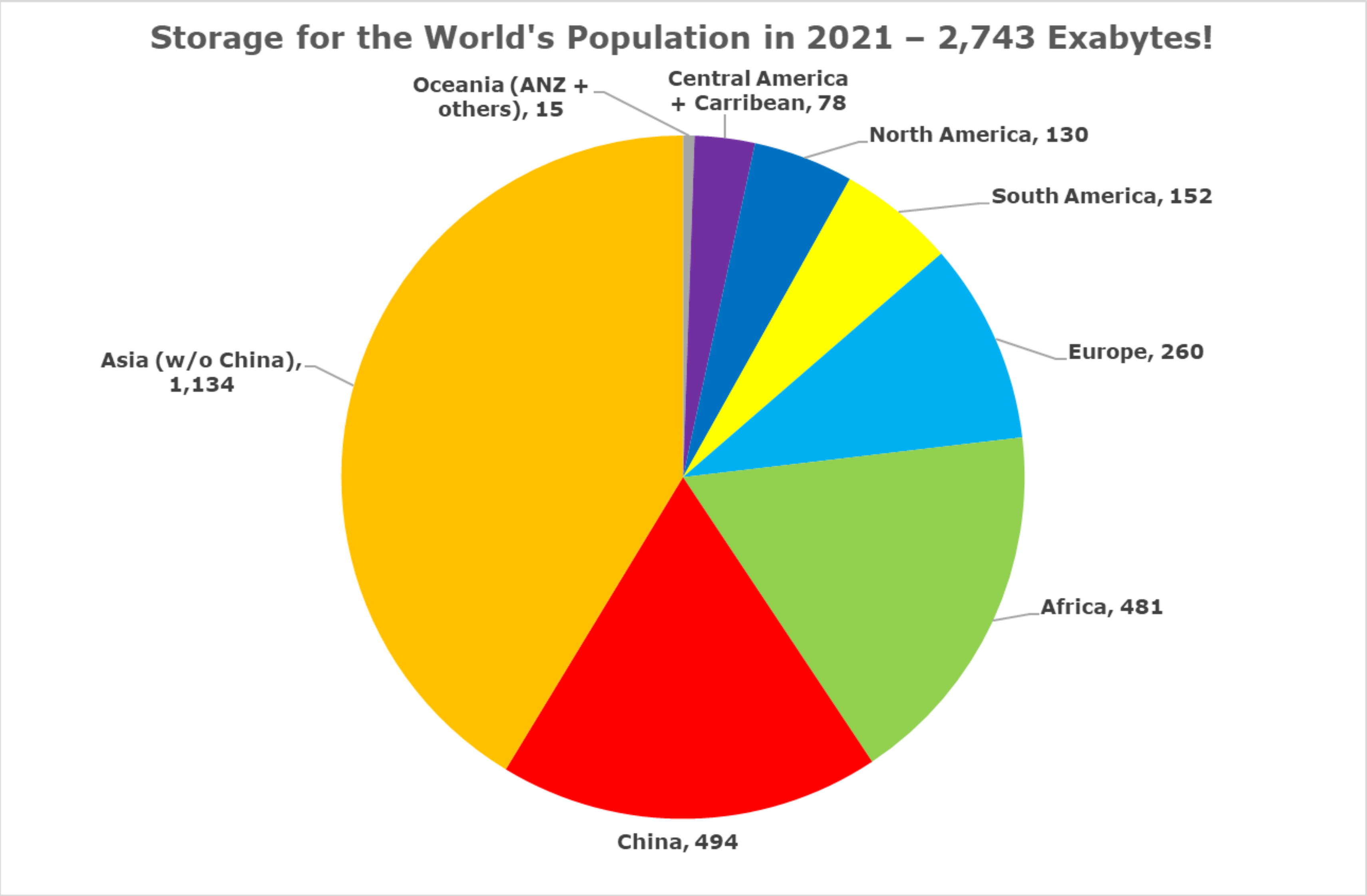
The World's Population – An Overview (~7.8B)



Source: 2021 World Population Data Sheet https://interactives.prb.org/2021-wpds/

Genomics Storage Associated with Populations

2,743 Exabytes for all ~7.8B people on Earth

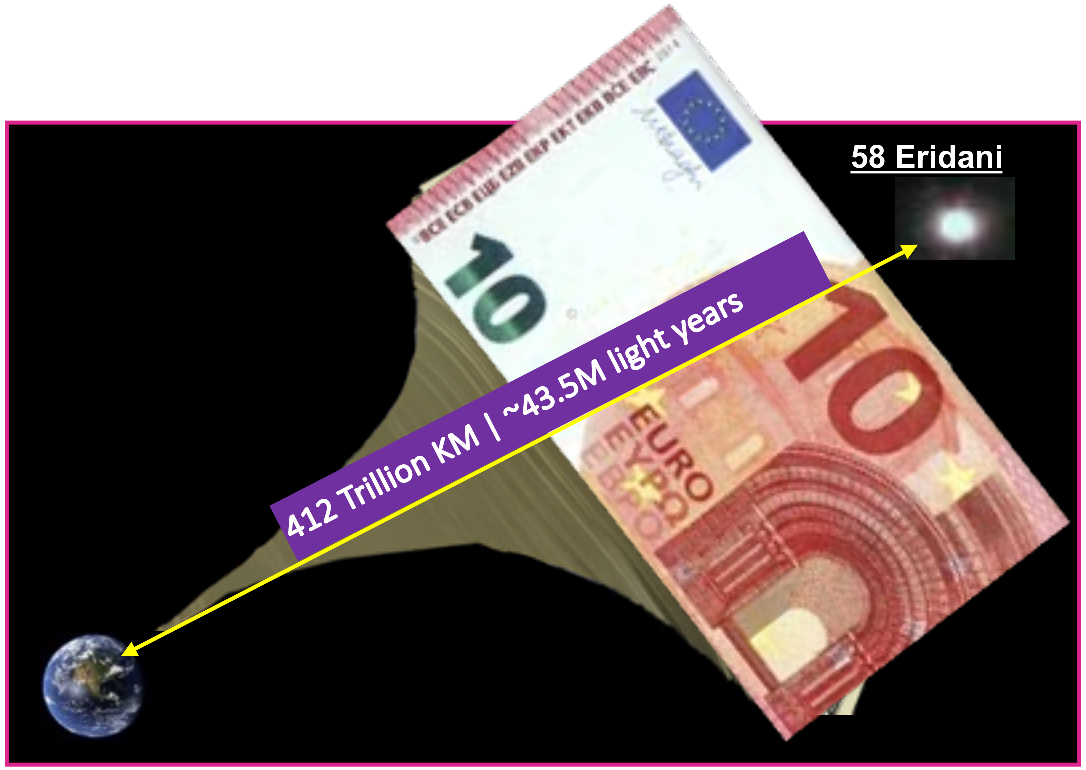


Applying Intel's Genomics Sizing Method to data from the 2021 World Population Data Sheet *Source:* https://interactives.prb.org/2021-wpds/

SNIA. CSTI | CLOUD STORAGE TECHNOLOGIES

Sequencing Everyone on Earth – Storage Fun Fact

- Assume 1 byte of storage = thickness of a 10€ note (0.15mm)
- Imagine a stack of 10€ notes, 1 note for each byte of storage need to sequence the entire Earth's population
- The required 2,743 Exabytes of storage could be represented as a stack of 10€ notes that would *extend all the way from the Earth to the star 58 Eridani, ~43.5 light-years away*.*



* Ignoring the laws of Physics and General Relativity ☺

†List of start systems within 40-45 light-years, https://en.Wikipedia.org/wiki/List_of_star_systems_within_40-45_light-years

The Storage Challenge

Torben Kling Petersen

Convergence of High Performance Storage

- Era of convergence of traditional simulation and AI requires NEW HPC storage

- Mainly **WRITING**
- **LARGE** files
- In mainly **SEQUENTIAL** order.
- Capacity measured in **PETABYTES**

Examples of traditional HPC storage:

- Cray ClusterStor L300
- DDN EXAScaler
- IBM ESS 3000

CHALLENGE IN NEW ERA:
An ORDER OF MAGNITUDE less performance
for small, random I/O compared to traditional AI storage

Modeling & Simulation

Machine Learning

Converged workloads running on one machine
in mission- or business-critical workflows

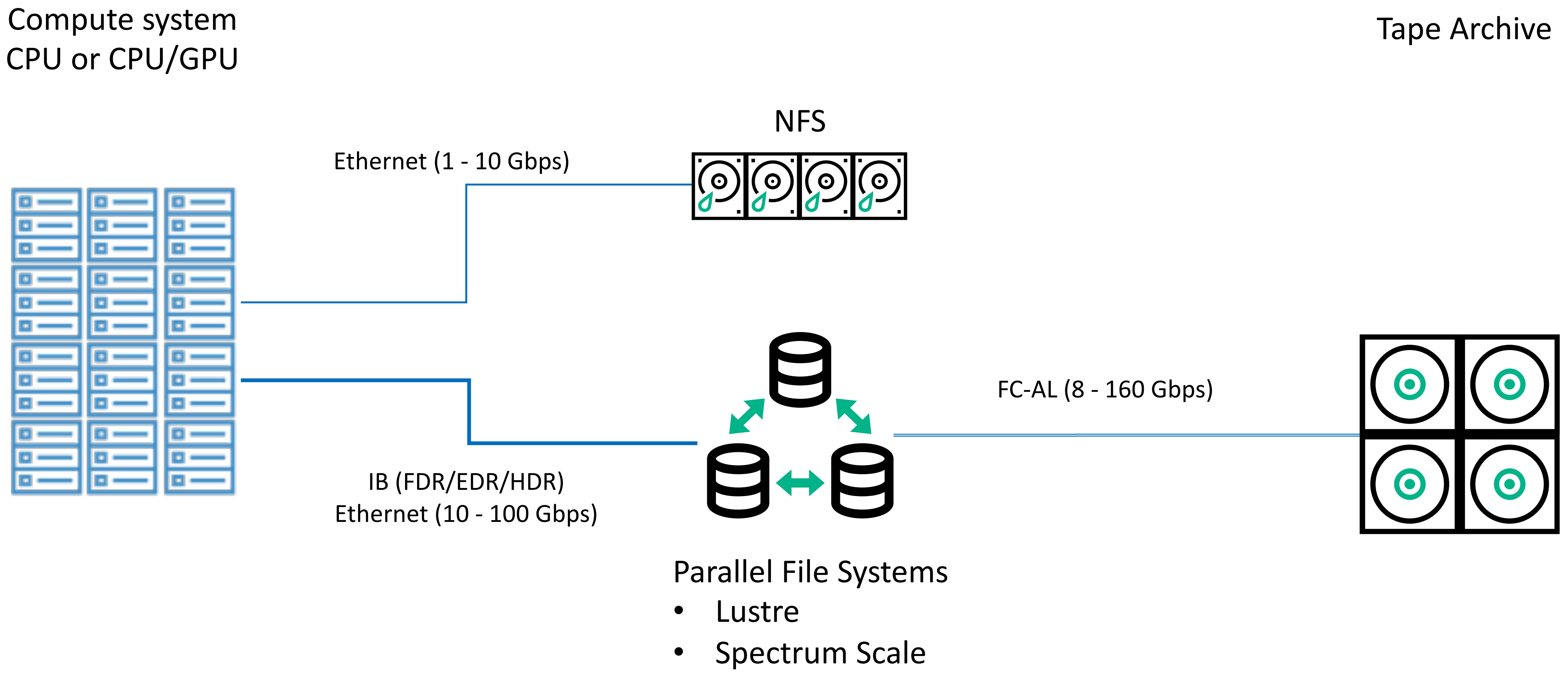
- Mainly **READING**
- Files of **ALL SIZES**
- In mainly **RANDOM** order.
- Capacity measured in **TERABYTES**

Examples of traditional AI storage:

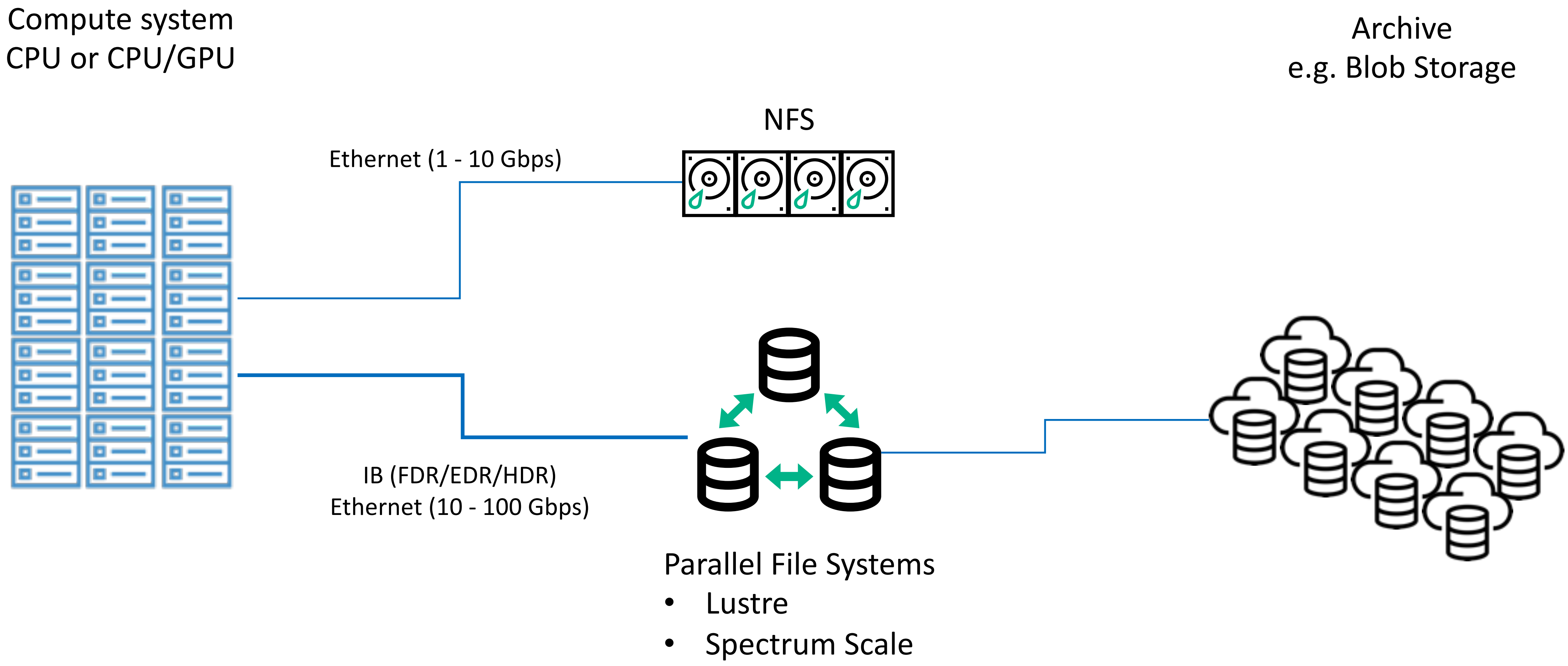
- NetApp AFF
- Dell EMC Isilon F-Series
- Pure Storage FlashBlade

CHALLENGE IN NEW ERA:
An ORDER OF MAGNITUDE more expensive
per terabyte compared to traditional HPC storage

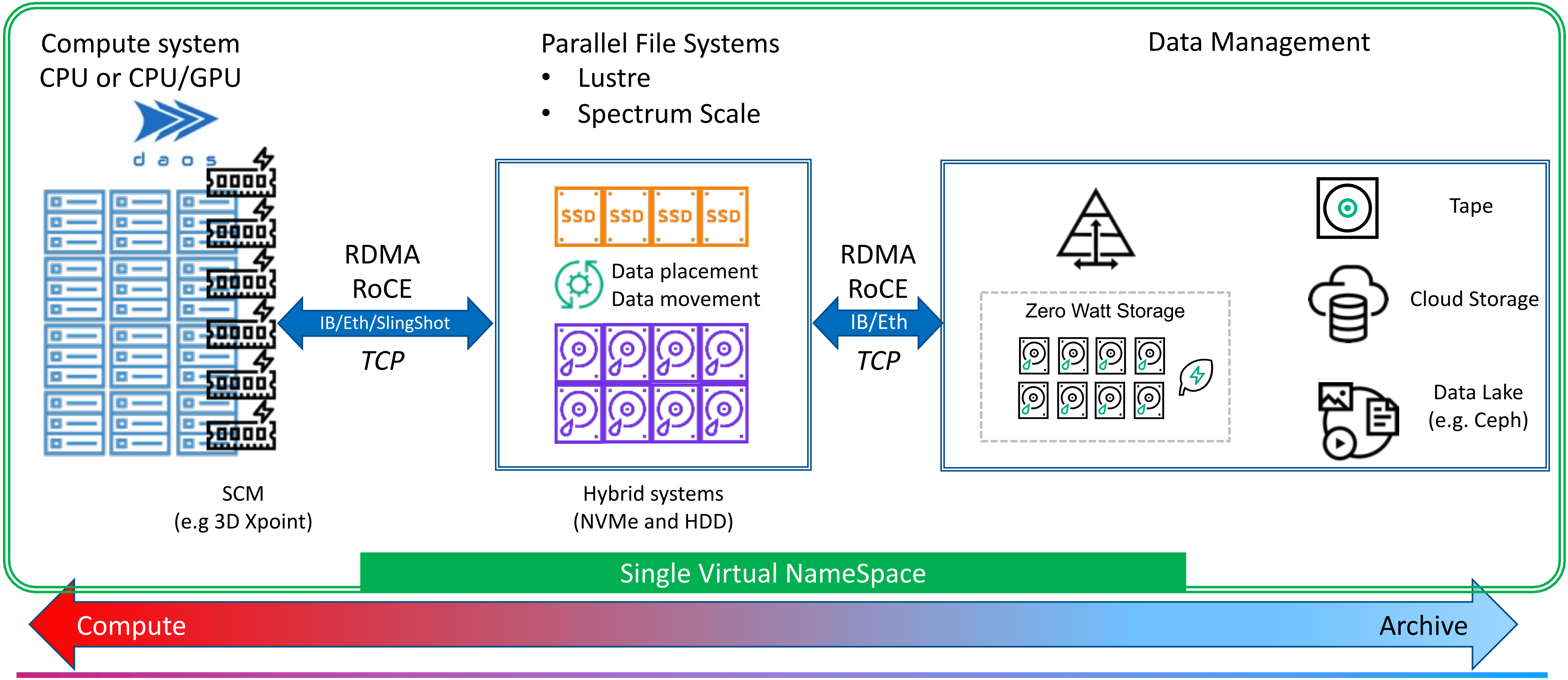
Traditional HPC design – On Prem



Traditional HPC design – Cloud Based



The "NEW" world – On prem or cloud based



SNIA. CSTI | CLOUD STORAGE TECHNOLOGIES

Data Services Requirements

- Data movement NVMe <-> HDDs
  - Policy based data migration based on capacity and age
  - Manual data migration
  - File purging policies
  - WLM directives
- Rapid search facility
  - External to file system -> low impact
  - Query function for advanced searching
  - HSM aware

Data Management

- Data movement - Primary FS to:
  - hot archive
  - object store
  - tape
  - cloud
- Policy based data migration based on
  - Age
  - Size
  - Type
  - Project
  - Classification
  - Usage history etc
- Manage multiple front ends
- Horizontal data movement
- Maintain full namespace mirror
- HSM **and** Incremental Backups
- Tiers gated by:
  - Cost per PB
  - Capacity growth
  - Retention requirements
  - Access performance

Moving it all "to the cloud"

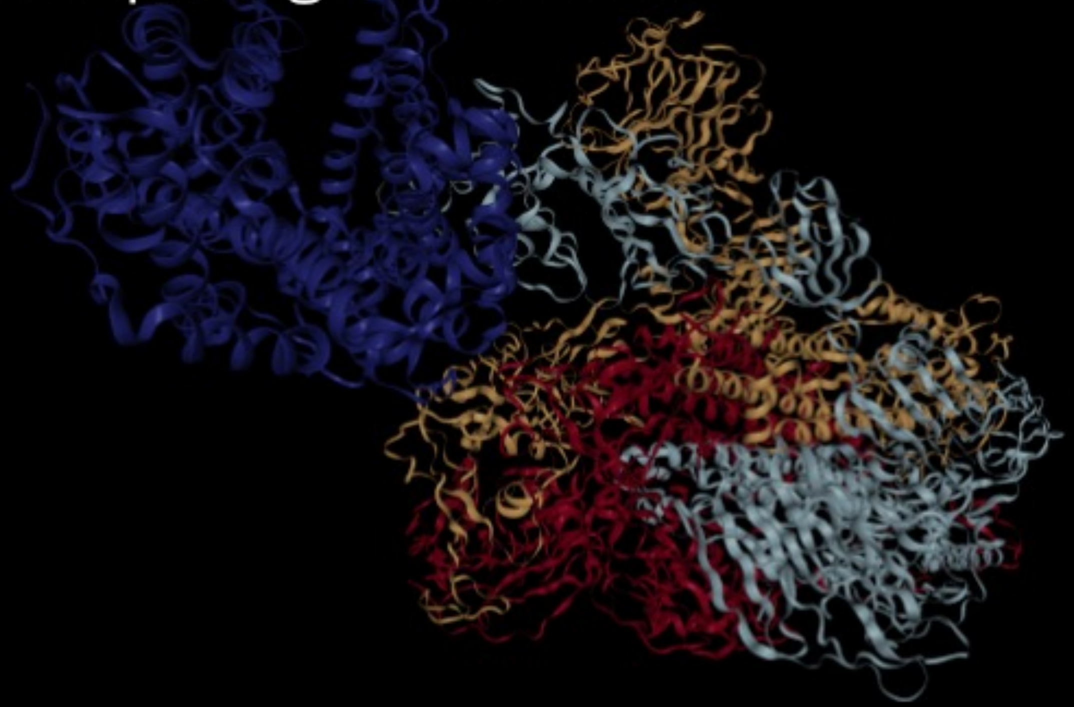
- Benefits:
  - CAPEX vs OPEX
  - Pay as you grow
  - Shared resources
  - Simple reconfiguration
- Challenges:
  - Moving data to/from the cloud
  - Multi-tenancy / Security
  - Legal constraints
  - Performance
    - Containers kill throughput and IOPS
  - Fair cost:
    - By capacity used ?
    - By performance ?
  - Simple reconfiguration
    - Significant data movement

So What’s the Next Step?

Christopher Davidson

Public Science In Practice

The COVID-19 High Performance Computing Consortium



Bringing together the Federal government, industry, and academic leaders to provide access to the world's most powerful high-performance computing resources in support of COVID-19 research.

100
—
Projects

600
—
Petaflops

SNIA. CSTI | CLOUD STORAGE TECHNOLOGIES

COVID-19 HPC Consortium - Facts & Timeline

THEN…

- 1 of 13 original members & 1 of 5 original members from Industry
- One of the largest private-public collaborations ever with members from Government, Industry, & Academia

NOW…

- 43 consortium members & global
- Phase 2 – focus on therapeutics and patient impact
- 100 projects
- 600 Petaflops

Dec 2019
First cases of COVID-19

March 22nd
Launch of COVID-19 HPC Consortium

April 2020
Global cases reaches 1M

May 2020
Call for research proposals

May 2020
Europe
PRACE & UKRI join Consortium

Aug 2020
Australia
NCI & Pawsey join Consortium

Nov 2020
HPCwire award for best "HPC Collaboration"

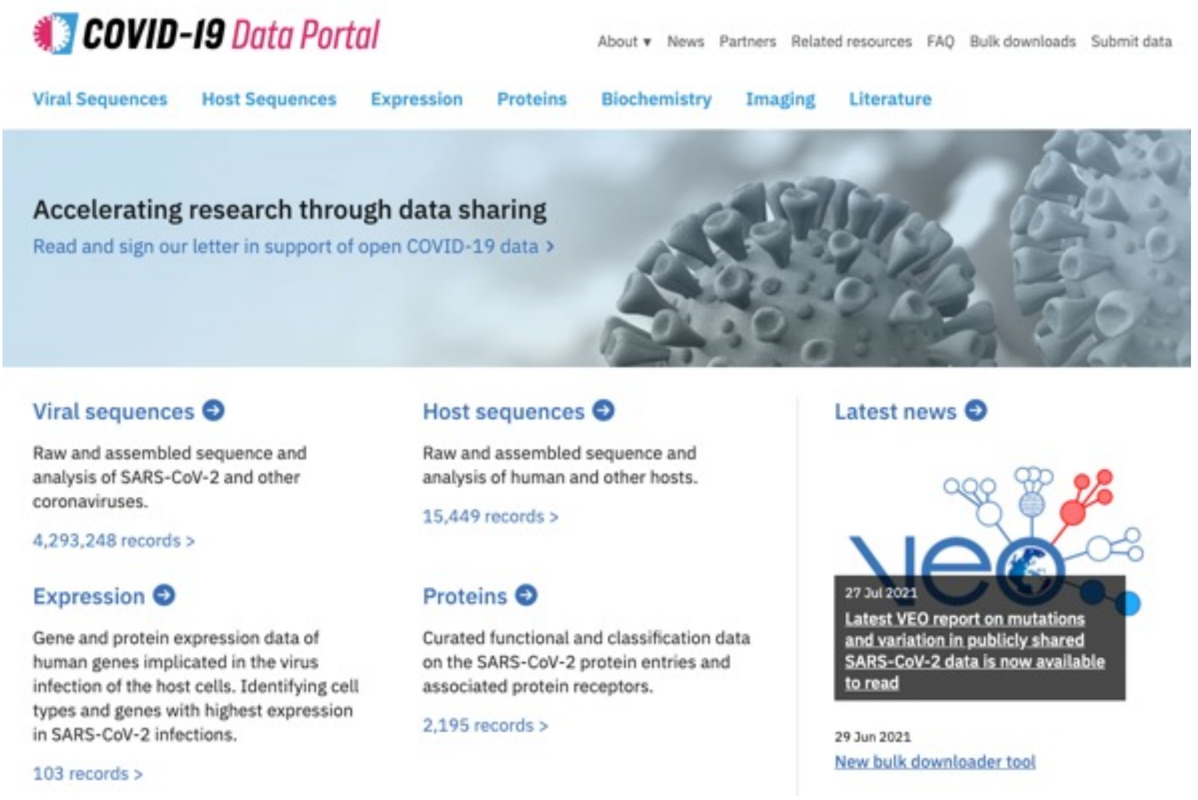
Nov 2020
Phase II
Focus on Therapeutics & Patient Impact

Jan 27, 2021
Global cases reaches >100M

What's next?

SNIA. CSTI | CLOUD STORAGE TECHNOLOGIES

Public Science In Practice



https://www.covid19dataportal.org/

https://covid.molssi.org/

SNIA. CSTI | CLOUD STORAGE TECHNOLOGIES

Summary

- Genomic data is growing at an exponential rate
- Work smarter, not harder
- Compute is a small part of the problem; data management & storage are of utmost importance
- Public & Private cloud provide a means to keep pace with the science and democratize the process
- Cloud provides a number of challenges but nothing is impossible

Thanks for Viewing this Webcast

- Please rate the webcast and provide us with feedback
- This webcast and a copy of the slides will be available at the SNIA Educational Library https://www.snia.org/educational-library
- A Q&A from this webcast will be posted to the SNIA Cloud blog: www.sniacloud.com/
- Follow us on Twitter @SNIACloud

Thank you

Questions?

Brief Genomics File Format Overview

From: Michael J. McManus

SNIA. CSTI | CLOUD STORAGE TECHNOLOGIES

A Quick Overview of File Formats (FASTQ)

- **FASTQ** – A text-based format for storing the DNA bases (the A's C's, G's, and T's) and the corresponding quality scores for each DNA base. An ASCII character is used to represent a base and another ASCII character for the quality score.

- Source: "FASTQ Format." Wikipedia: The Free Encyclopedia. Wikimedia Foundation, Inc. 29 May 2016. 7 July 2016. <https://en.wikipedia.org/wiki/FASTQ_format)

    - Average file size for a 30X whole genome - ~180GB

```
SEQUENCE ID:      @SEQ_ID
DNA BASES:        GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
SEPARATOR:        +
QUALITY SCORE:    !''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```


SNIA. CSTI | CLOUD STORAGE TECHNOLOGIES

A Quick Overview of File Formats (BAM)

- **BAM** – A BAM file is the binary version of a SAM file. A SAM file is a tab-delimited text file that contains sequence alignment data.
  Source: Broad Institute, <https://www.broadinstitute.org/igv/BAM>
  - Average BAM file size for a 50X whole genome - ~200-300GB

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```


SNIA. CSTI | CLOUD STORAGE TECHNOLOGIES

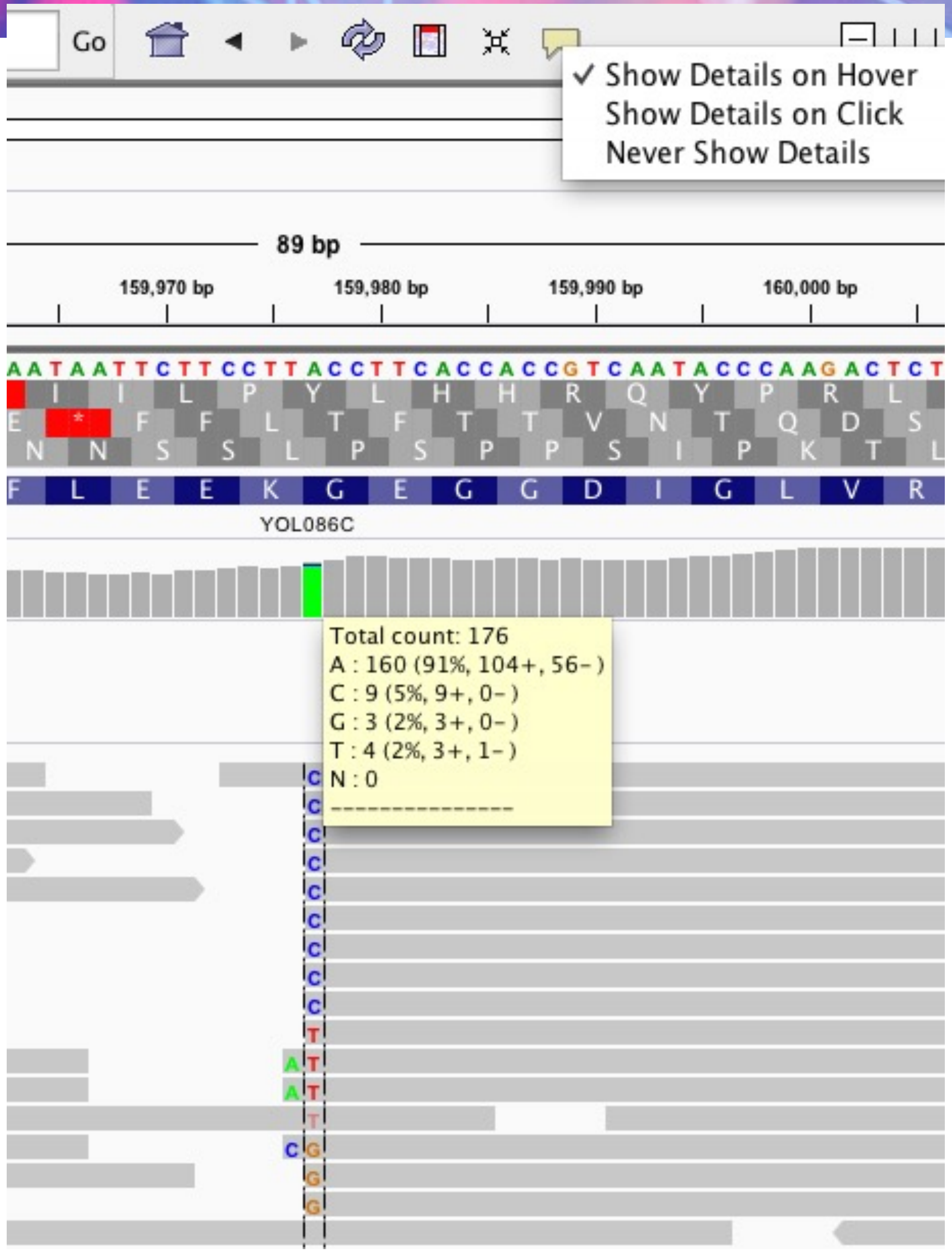
A Quick Overview of File Formats (VCF)

- **VCF** – The Variant Call Format specifies the format of a text file used in bioinformatics for storing gene sequence variations.
  - Source: "Variant Call Format," Wikipedia: The Free Encyclopedia. Wikimedia Foundation, Inc. 29 June 2016. 7 July 2016. <https://en.wikipedia.org/wiki/Variant_Call_Format>

```
##fileformat=VCFv4.0
##fileDate=20110705
##reference=1000GenomesPilot-NCBI37
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF  ALT     QUAL FILTER INFO                              FORMAT      Sample1        Sample2        Sample3
2      4370    rs6057    G    A       29   .      NS=2;DP=13;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:52,51 1|0:48:8:51,51 1/1:43:5:.,.
2      7330    .         T    A       3    q10    NS=5;DP=12;AF=0.017               GT:GQ:DP:HQ 0|0:46:3:58,50 0|1:3:5:65,3   0/0:41:3
2      110696  rs6055    A    G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
2      130237  .         T    .       47   .      NS=2;DP=16;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:56,51 0/0:61:2
2      134567  microsat1 GTCT G,GTACT 50   PASS   NS=2;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
chr1    45796269        .       G       C
chr1    45797505        .       C       G
chr1    45798555        .       T       C
```

Genome Coverage

- **Coverage (Depth)** – refers to the number of times a nucleotide is read during the sequencing process. Deep sequencing indicates that the total number of reads is many times larger than the length of the sequence under study.
  Source: "Deep Sequencing." Wikipedia: The Free Encyclopedia. Wikimedia Foundation, Inc. 29 May 2016. 7 July 2016. <https://en.wikipedia.org/wiki/Deep_sequencing>
- Example: 50x coverage means the nucleotides in the sequence have been "read" 50 times
- Coverage enables the distinction between the inherent error in the sequencing instrument and a real genetic variant as compared to the reference genome.



SNIA. CSTI | CLOUD STORAGE TECHNOLOGIES