



Experiences Sharing Genomics Data

David Haussler

PROBLEM: Genome data held in silos, unshared, not standardized

- No one institute has enough on its own to make progress
- Every clinician should be able to compare their genomes to others



We need a network to share



UNIVERSITY OF CALIFORNIA
SANTA CRUZ

Genomics
Institute



Global Alliance
for Genomics & Health
Collaborate. Innovate. Accelerate.



200+ Genomic Data Initiatives Globally

Clinical/Genomic Medicine



40

Initiatives

Research



70

Initiatives

National



21

From 15 Countries

Cohorts



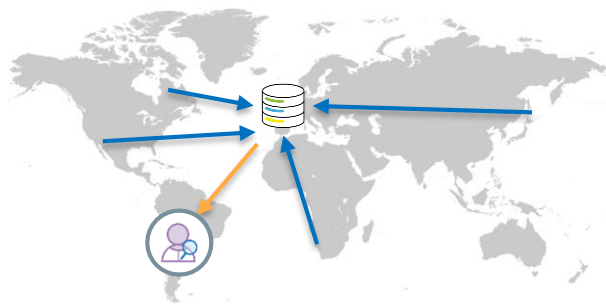
64

Globally

Data Federation Model for Global Genome Sharing

Data Commons

Basic research consented for data sharing



Aggregate data globally

Download, analyze locally

Cloud Commons

Large scale research datasets

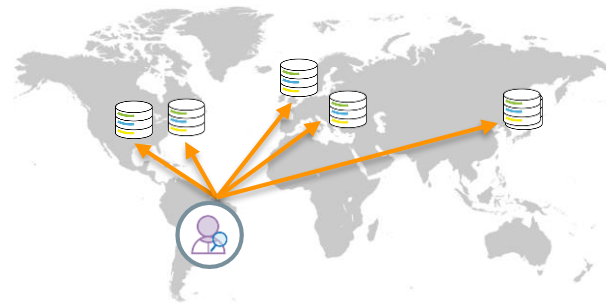


Aggregate data globally

Analyze centrally in secure cloud

Data Federation

New approach for genomics initiatives



Host data locally

Visit data remotely and collate results



User



Data transmission

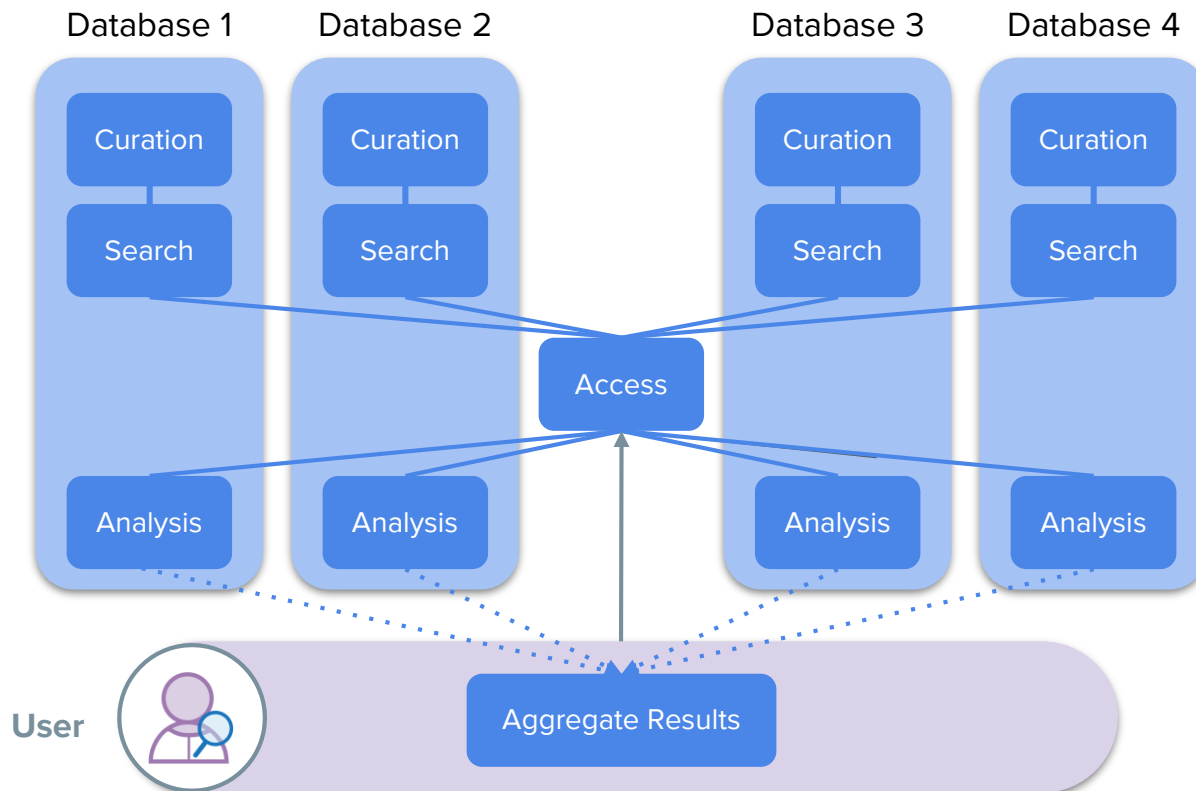


Analysis traveling to the data

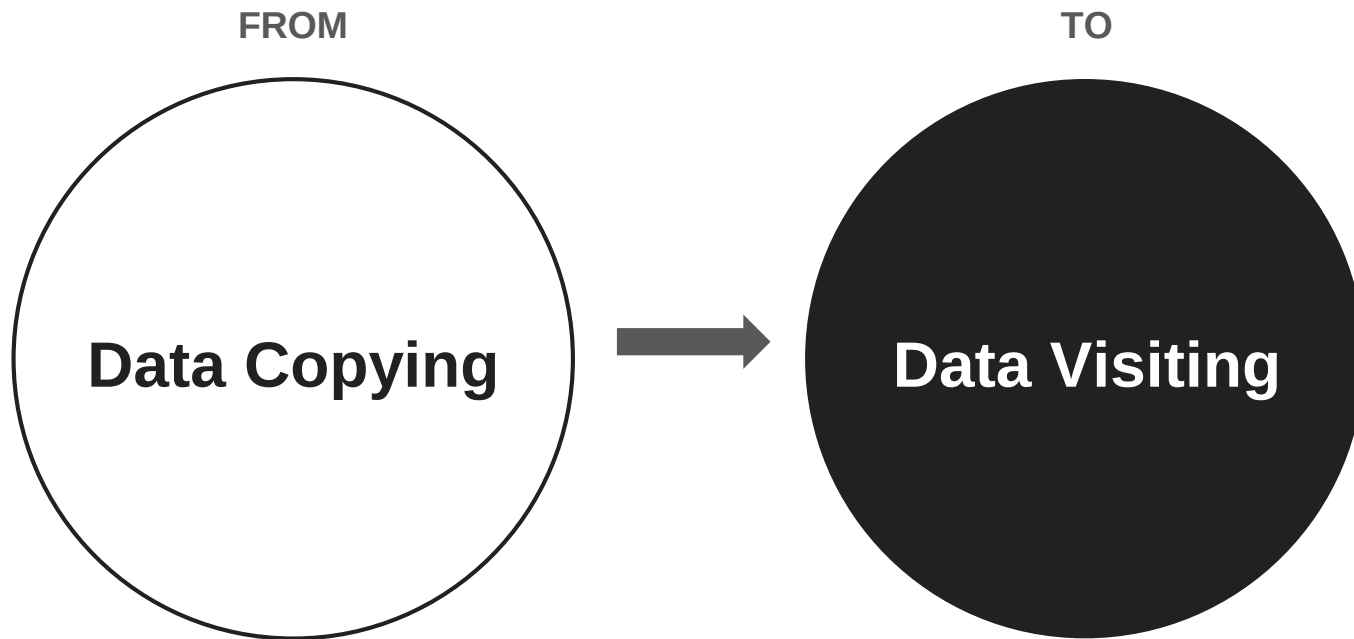
Federated analysis of cooperating genome databases



Global Alliance
for Genomics & Health



A New Paradigm for Genome Sharing



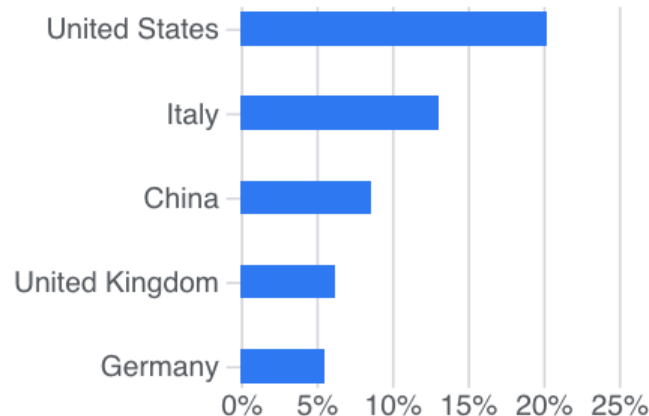
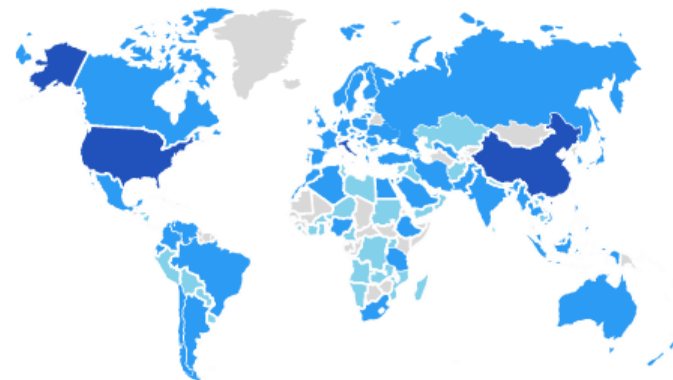
Case Study: BRCA Exchange validated more than 65,000 genetic variants of the BRCA1 and 2 genes

Solution: BRCA Exchange

BRCA Exchange aggregates "big data" on BRCA variants from research and health care databases around the globe. The Evidence-based Network for the Interpretation of Germline Mutant Alleles (ENIGMA)*



contributes expert classifications for each variant. Patients, clinicians, and others can review these details to inform their decisions on cancer prevention, screening, and intervention.



UNIVERSITY OF CALIFORNIA
SANTA CRUZ Genomics
Institute

AGATC:VARIANT:GAGTTTC:CLASSIFICATION:UT
GGAACAGACAGTAAATATTTCTGAGATGCTGAGG
GAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGG
ATGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAG
TGCAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAG
© BREAST CANCER ACTION GROUP (BCAC) 2014

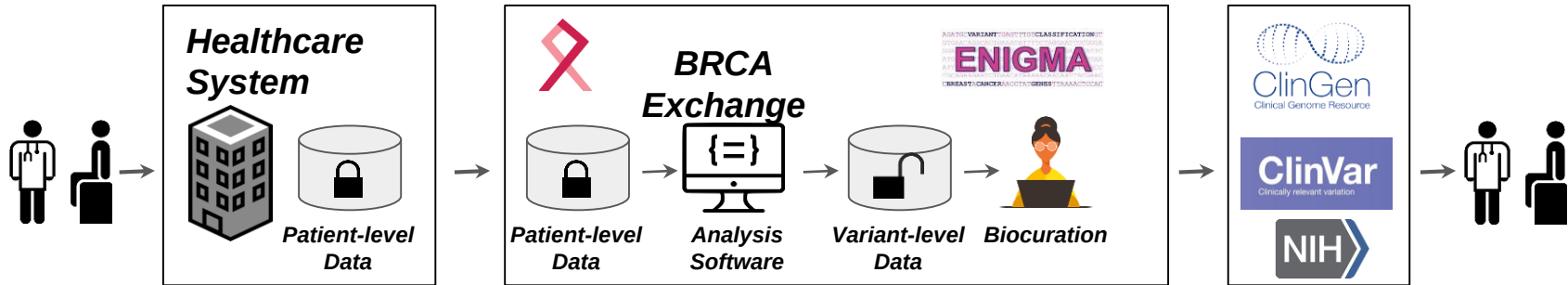
ENIGMA

How to share BRCA information from patients

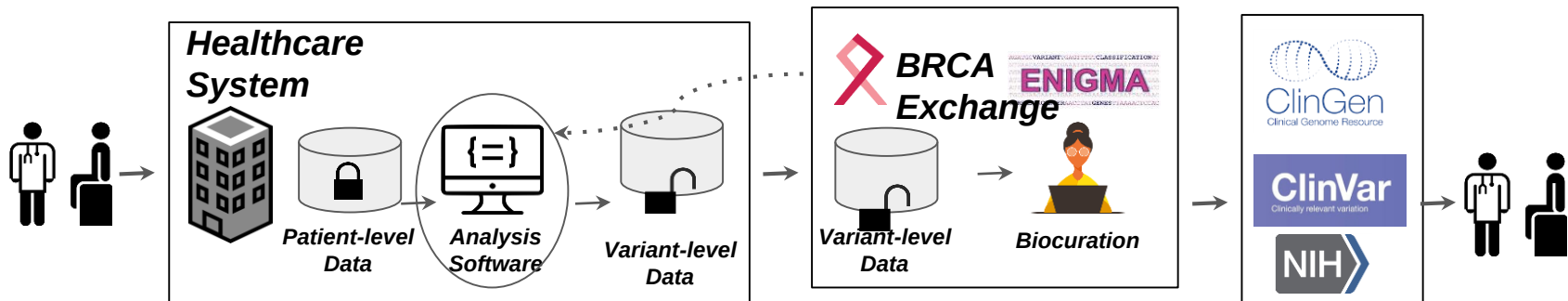
- Variant interpretation relies on data from patients and their families. It requires **sharing treatments and outcomes**.
- Patient information is private and cannot be shared directly.
- But, the information needed for variant interpretation is not the individual private, patient-level information. It is variant-level summary data.
- We can share the variant-level summary data.

Privacy-preserving data sharing through federated analysis

Traditional Data Sharing



Federated Analysis



Example: Federated analysis collaboration with BioBank Japan

- BioBank Japan has a large cohort of cancer patients and controls, which they cannot share directly.
- We shared a Docker container with them to analyze their patient cohort
- The container generated variant-level data, which we are now using together with the ENIGMA Consortium to interpret BRCA variants!



2nd Case Study: California Institute for Regenerative Medicine Stem Cell Genomics Hub



Home Data Projects Tools Standards Help About ▾

Welcome to the CESC Stem Cell Hub!



Browse Files

Filter, view, and download files

[More >](#)



View Datasets

Explore the datasets and assays available

[More >](#)



Read Articles

Read publications about CESC research

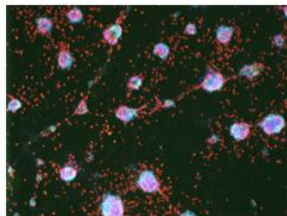
[More >](#)



Use Tools

Take advantage of CESC analysis tools

[More >](#)



Welcome to the Stem Cell Hub web site!

The CIRM Center of Excellence in Stem Cell Genomics (CESCG) was established with the goal to apply genomics and bioinformatics approaches to stem cell research to accelerate fundamental understanding of human biology and disease mechanisms, enhance cell and tissue production and advance personalized cellular therapies. The Stem Cell Hub was built with funding from CIRM to host high-quality genomic data collected under the CESC Collaborative Research Programs (CRP).

The Stem Cell Hub contains many terabytes of data that cover a large variety of sequencing assays, including a vast amount of single-cell data. It houses primary data files such as DNA reads in fastq format, as well as many types of files derived from mapping and other analysis of the primary data, and PDF and other document files describing protocols. Sequenced samples in the Stem Cell Hub include naturally developing human tissues as well as cells engineered from healthy skin and blood cells to become replacements for diseased cells and tissues, including data from ethically and scientifically careful human clinical trials.

We hope that the Stem Cell Hub will be useful to a wide range of scientists from clinicians to cell biologists to bioinformaticians doing custom analysis and combining data from multiple projects. Please see the help link for additional information on how to use the site.

If you have any questions, or if you'd like to contribute data to the Stem Cell Hub, please contact our support team at cirm-wrangler-group@ucsc.edu

84 TB of data in 180,170 files from 18 CIRM research labs

6.7 TB of that data is **currently publicly available**; the rest is pre-publication

The data are **Machine-Learning Ready***

<https://cirm.ucsc.edu>

*https://www.acd.od.nih.gov/documents/presentations/12132019AI_FinalReport.pdf

Cell Browser

The Cell Browser is a software **tool using a 2D viewer** to represent single-cell RNA expression

Integrates CIRM data with global single-cell data, including HCA

Shows expression data for individual cells

Allows for a visual comparison of large datasets consisting of many cells

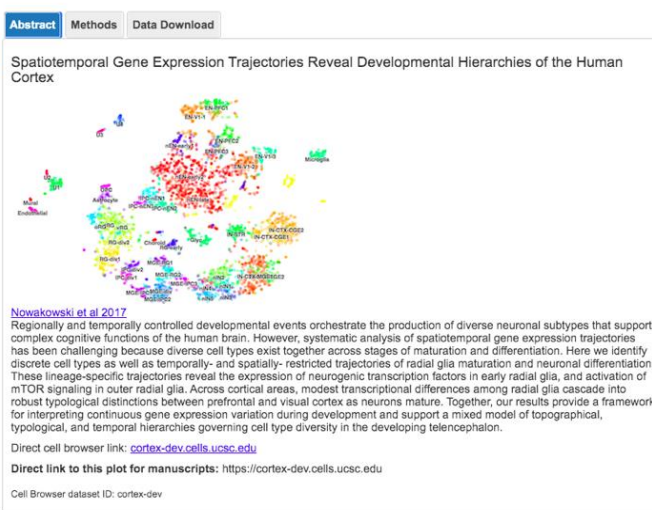
Includes overlays of metadata, marker gene levels, cell clustering and more

Useful for comparing single-cell layout/batch correction methods

Coordinated with Human Cell Atlas (HCA)

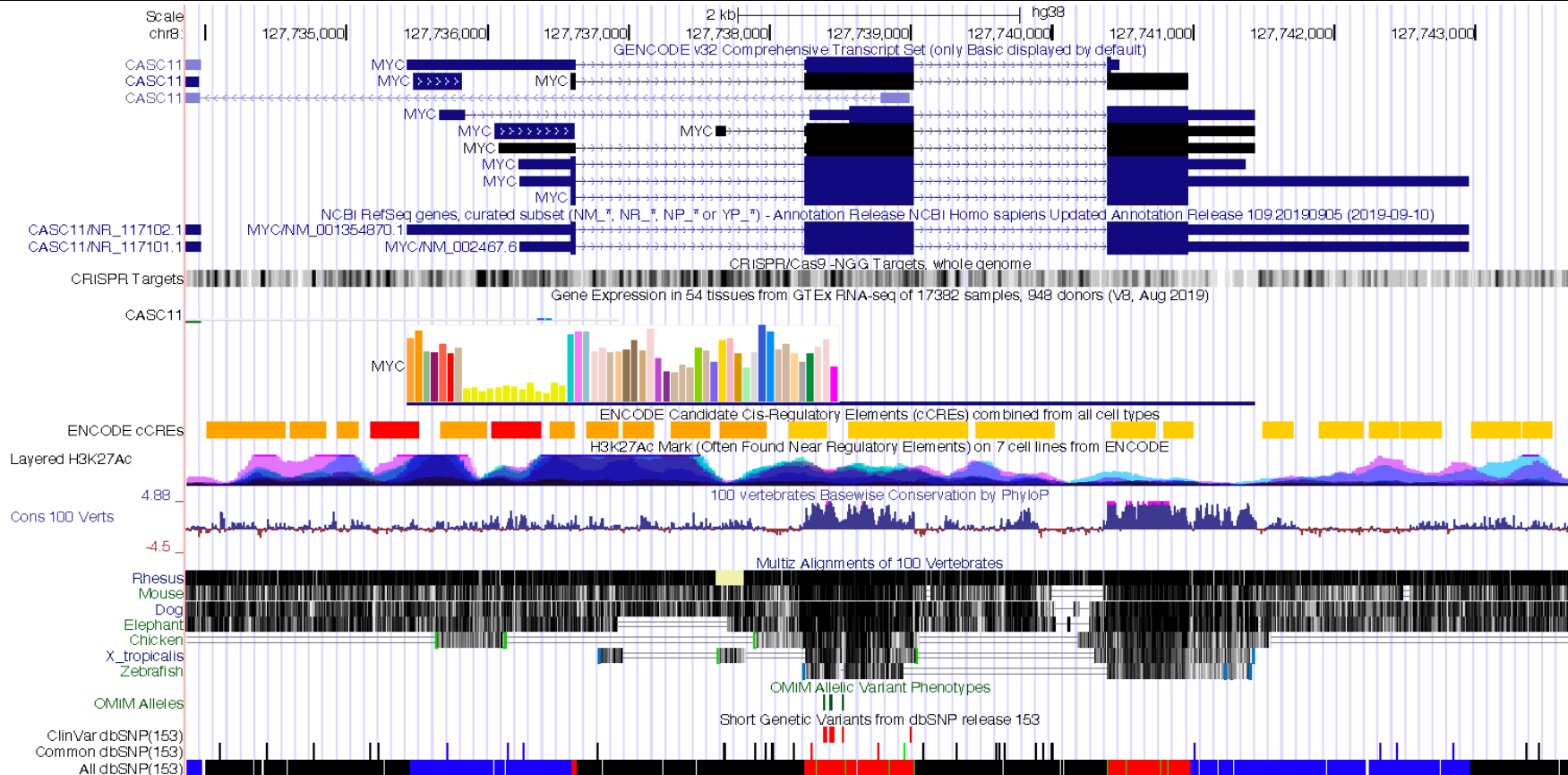
Choose Cell Browser Dataset

Overview	
Cortex development	smartseq2 4.3k Open
Glioblastoma Infiltrating vs. Tumor Core	smartseq2 GEO 3.6k Open
Oligodendrocytes in MS	10x GEO 19k Open
HCA Datasets via Xena	9 datasets Open
Adult Pancreas	smartseq2 4.6k Open
Autism	10x 105k Open
Head and Neck Cancer	smartseq2 3.6k Open
Melanoma	Drop-Seq 10x 6 datasets Open
Mouse Organogenesis	mouse 2.1M Open
Multiple Sclerosis	10x 49k Open
Organoid Report Card	10x 2 datasets Open
Tabula Muris	smartseq2 45k Open
Treehouse Cancer Compendium	bulk 6 datasets Open
Mouse cortex and hippocampus	3.9k Open
Human Kidney Cell Atlas	10x 7 datasets Open
Brain and organoid	190k Open
Tabula Muris Senis	2 collections Open



<https://cells.ucsc.edu/>

Genome Browser



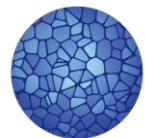
<https://genome.ucsc.edu/>

Gen-2 CIRM Data will be in the Data Biosphere

Scalable and interoperable computing resource for the genomics scientific community

Cloud-based infrastructure

- Highly elastic; shared analysis and computing environment



HUMAN
CELL
ATLAS

Data access and security

- Genomic and single cell datasets, phenotypes and metadata
- Securely housed, large datasets generated by NHGRI, NHLBI, NCI, CZI funded programs and HCA community, as well as other initiatives / agencies



AnVIL



LungMAP

Collaborative computing environment for datasets and analysis workflows

- Storage, scalable analytics, data visualization
- Security, training & outreach, with new models of data access
 - ...for both users with limited computational expertise and sophisticated data scientist users

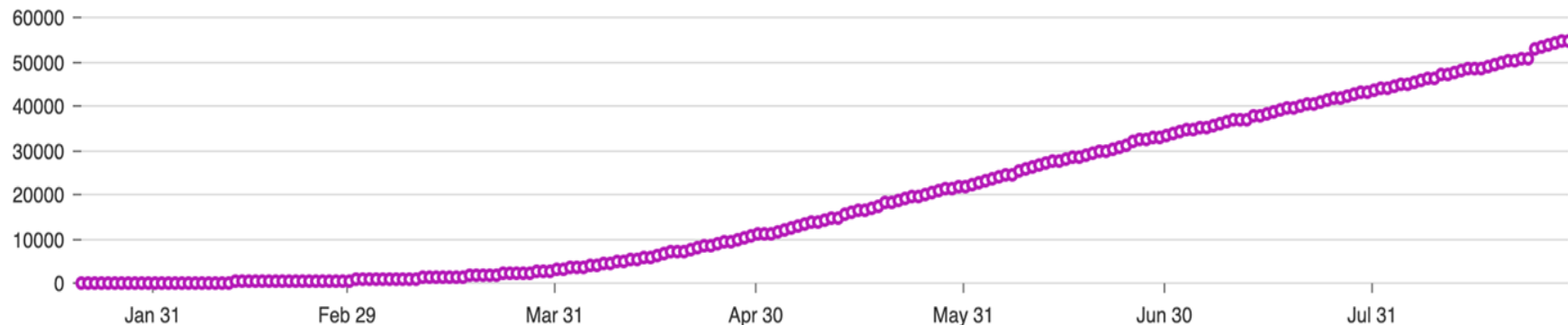


DATA
BIOSPHERE

Case Study: SARS-CoV-2 Research is generating data at an astonishing pace

Jan 21, 2020 - Aug 30, 2020

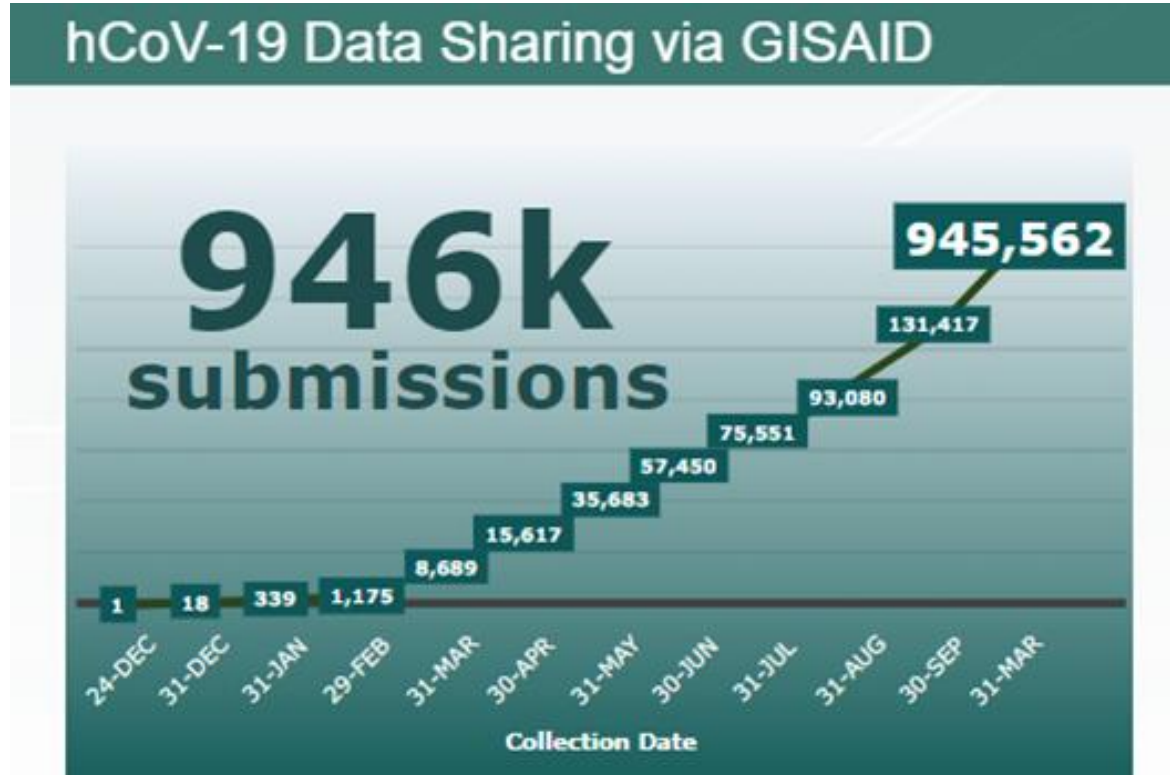
4121 New in the Past 7 Days | 54917 Cumulative Papers



In April 2020, SARS CoV-2 papers had a doubling time of ~14.5 days.

(The virus doubling time then was ~7 days)

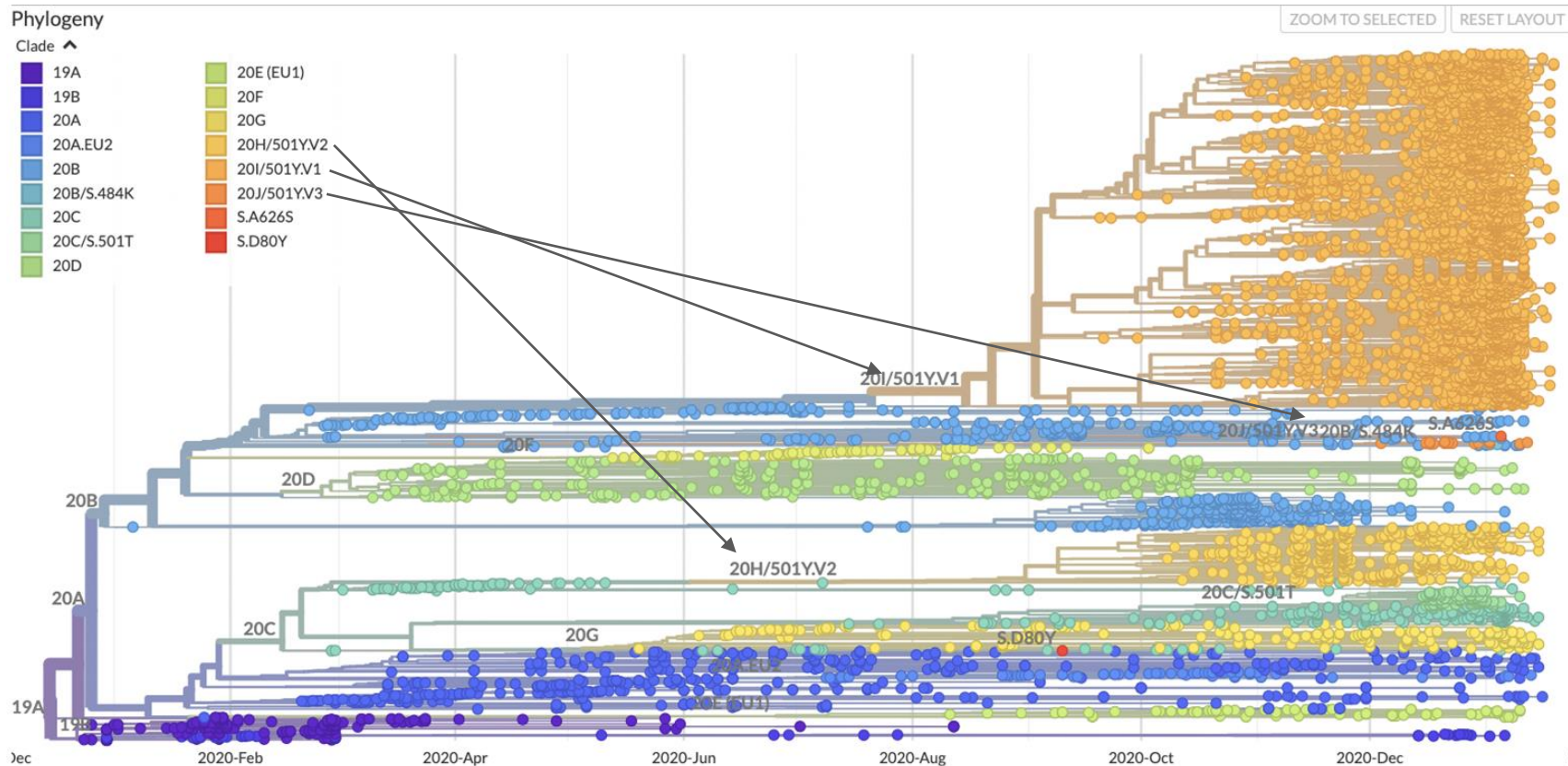
Genomic Data has also grown at an exponential rate



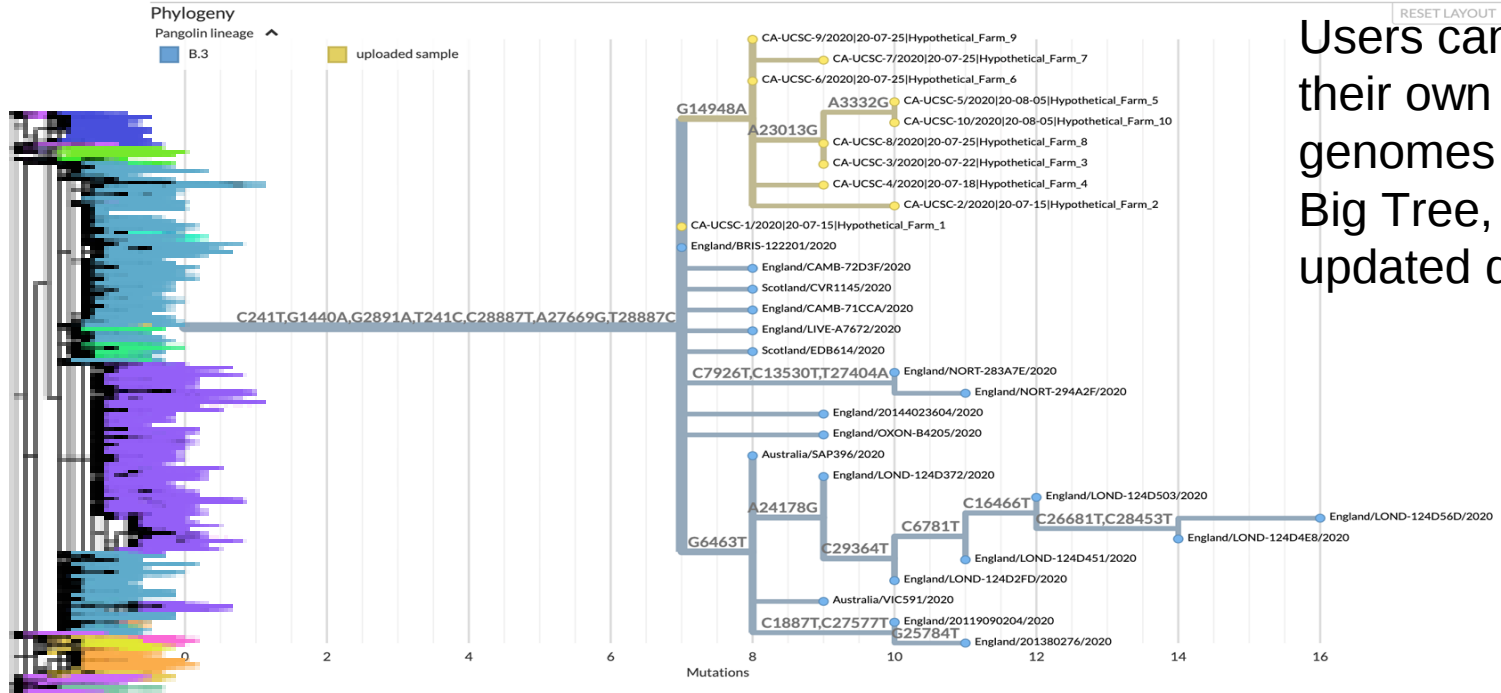
Now more than 1.2M genomes sequenced

New variants spread faster

Many show independent mutations to same alleles



We are the only lab that was able to maintain an accurate phylogenetic tree of >1M genomes



Users can add their own virus genomes to the Big Tree, which updated daily

>1M GISAID+public SARS-CoV-2 genome sequences, >1000 users/week

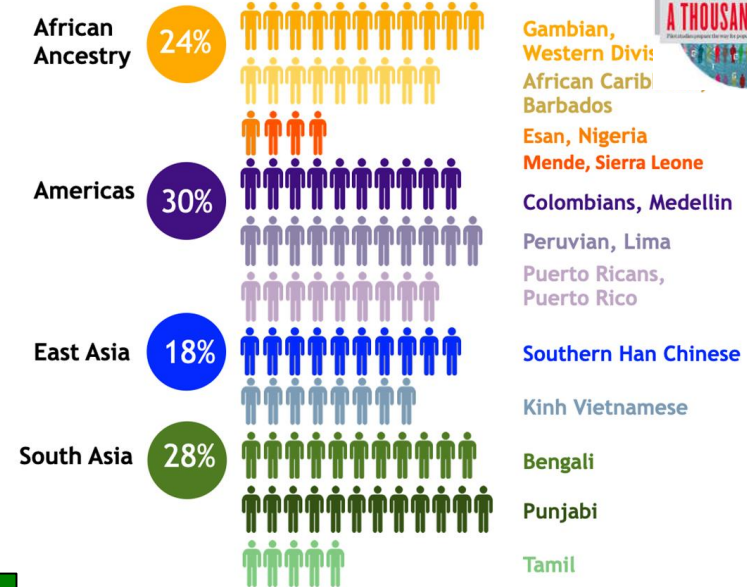
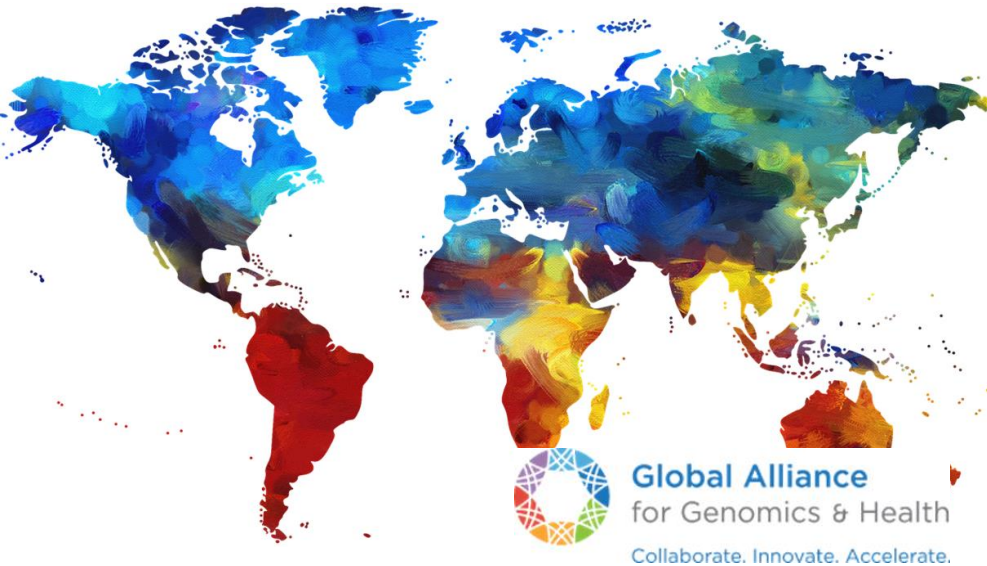
<https://genome.ucsc.edu/cgi-bin/hgPhyloPlace>

Data Sharing Issues for Pandemic Genome Tracking

- We were asked to “cease and desist” based on decade-old data sharing agreements (we refused)
- But the root issue is real: hardworking, underfunded scientists around the world contributing data deserve better recognition and reward

Latest Challenge: we must expand the one reference human genome to a Human Pangenome Reference

Global Genomic Partnerships



New international partnerships are needed to reach a more complete "human pangenome reference"

- ✓ Represents genetic and geo. diversity
- ✓ Availability of low passage cell lines
- ✗ Strong Regional Scientific Partnerships

Take Home

The most important issue in data sharing is respect.

Respect for those who make the science possible by contributing data.

But that respect has to be earned with an equal amount of generosity.