

Future Directions of AnVIL Workshop

29OCT2021



- Introduction to AnVIL -
Anthony Philippakis and Michael Schatz



Introductions



Michael Schatz
Johns Hopkins University
Computer Science and Biology



Anthony Philippakis
Broad Institute
Chief Data Officer & Institute Scientist

AnVIL Working Groups + Committees

Technical Working Group

Chairs: Michael Schatz (JHU)
Brian O'Connor (Broad)

Data Access Working Group

Chairs: Stacey Donnelly (Broad)
Carolyn Hutter (NHGRI)

Outreach Working Group

Chairs: Jeffrey Leek (JHU)
Frederick Tan (JHU)

Data Processing Working Group

Chairs: Eric Banks (Broad),
Ira Hall (WashU/Yale)

Portal Working Group

Chairs: Michael Schatz (JHU)
Benedict Paten (UCSC)

Phenotype Working Group

Chairs: David Crosslin (eMERGE - UW)
Robert Carroll (VUMC)

Data Ingestion Committee

Members: Michael Schatz (JHU)
Anthony Philippakis (Broad) et al

AHA/AnVIL Working Group

Members: Michael Schatz (JHU)
Anthony Philippakis (Broad) et al

With extensive participation from all sites

What is the AnVIL?

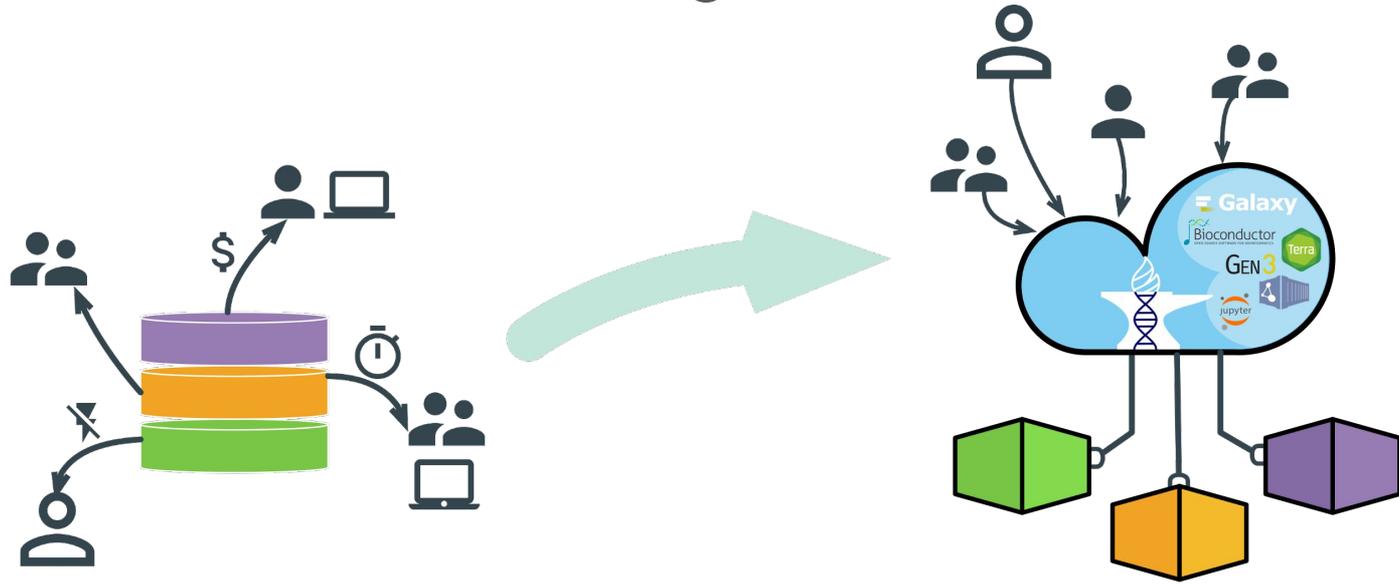
Scalable and interoperable computing resource for the genomics scientific community

- Cloud-based infrastructure
 - Highly elastic; shared analysis and computing environment
- Data access and security
 - Genomic datasets, phenotypes and metadata
 - Large datasets generated by NHGRI programs, as well as other initiatives / agencies
 - dbGaP Authenticated sharing of primary and derived datasets
- 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

The screenshot shows the AnVIL website homepage. At the top, there is a navigation bar with the AnVIL logo and the text "NHGRI Analysis Visualization and Informatics LabSpace". A search bar and utility icons are on the right. Below the navigation is a main heading "Migrate Your Genomic Research to the Cloud" with a sub-heading "Secure, cost-effective genomic analysis at scale." and buttons for "Get Started" and "Learn More". A featured article titled "A complete reference genome improves analysis of human genetic variation" is displayed. Below this is a grid of tool integrations: Terra, Gen3, Dockstore, NCPI, Bioconductor, Galaxy, Jupyter, and Seqr. A section titled "Access diverse, open and controlled access, cloud-hosted datasets" features statistics: "250+ Cohorts", "3+ Petabytes", and "300+ thousand Participants". Below this is a grid of dataset cards for CMG, CCDG, GTEx, 1000G, eMERGE, PAGE, HPRC, T2T, and Convergent Neuro, each with "Learn More" and "Datasets" links. At the bottom are buttons for "Consortia Roadmap", "Explore Datasets", and "Contribute Data".

<https://anvilproject.org>

AnVIL: Inverting the model of genomic data sharing



Traditional: Bring data to the researcher

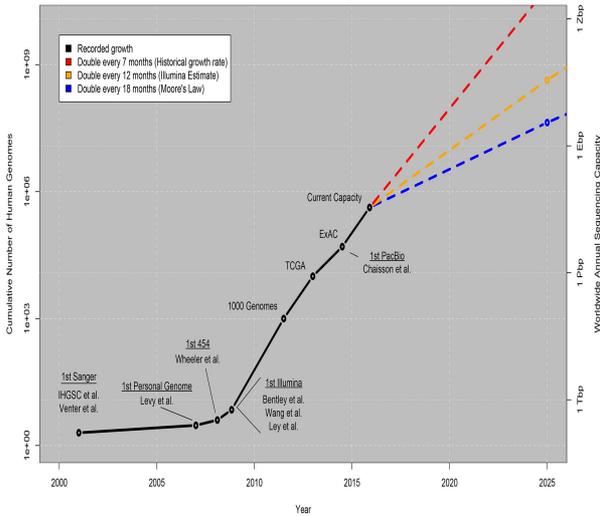
- Copying/moving data is costly
- Harder to enforce security
- Redundant infrastructure
- Siloed compute

Goal: Bring researcher to the data

- Reduced redundancy and costs
- Active threat detection and auditing
- Greater accessibility
- Elastic, shared, compute

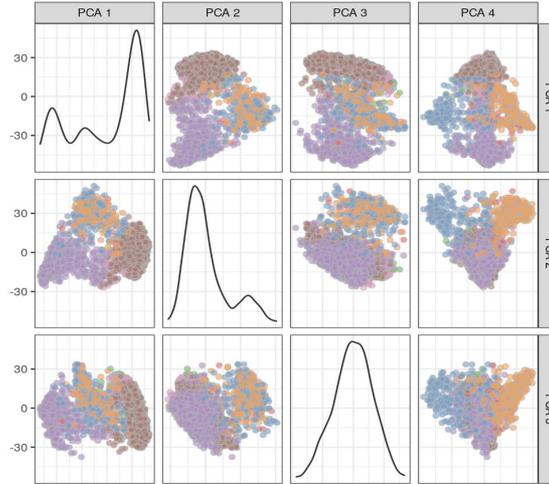
Why AnVIL?

Data



Scale,
Integration,
Sharing & Reuse

Computation



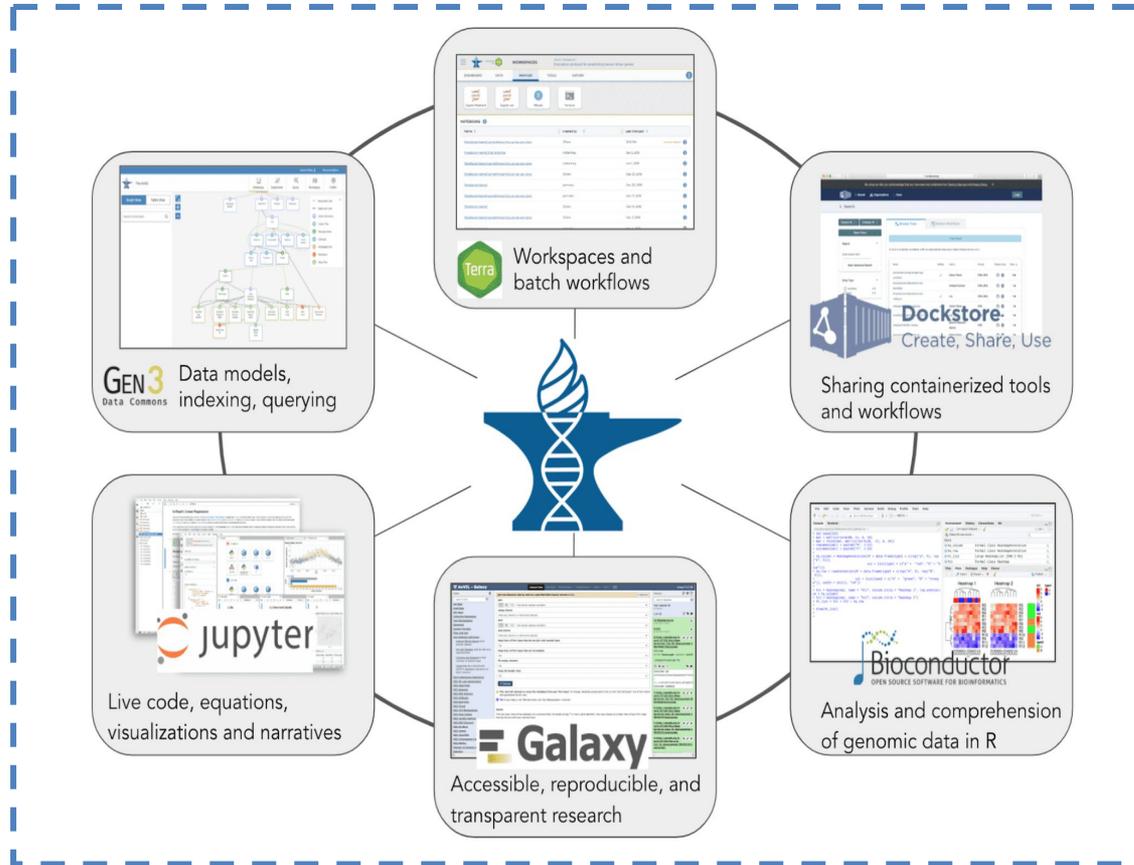
Simplicity,
Reproducibility,
Security

Users



Democratization,
Collaboration,
Discoveries

Building a Secure Federated Data Ecosystem



FedRAMP

FedRAMP certified
1 ATO



Implemented on Google Cloud Platform

AnVIL by the numbers

Data

	Gen3	Total
Consortia		9
Cohorts		254
Subjects	22,071	291,301
Samples	69,787	314,038
Size		3.87 PB

Tools & Workflows

<u>Dockstore:</u>	WDL: 840 workflows Galaxy: 28 workflows
<u>Terra:</u>	272 public workspaces 48 featured workspaces
<u>Bioconductor:</u>	2,041 software packages 977 annotation resources 406 data collections
<u>Galaxy:</u>	7,829 tools available

Users

Visits

anvilproject.org	6731 in Q3
anvil.terra.bio	440 / month

Terra usage

Users	>15,000
Public Workspaces	272
Cloned Workspaces	625

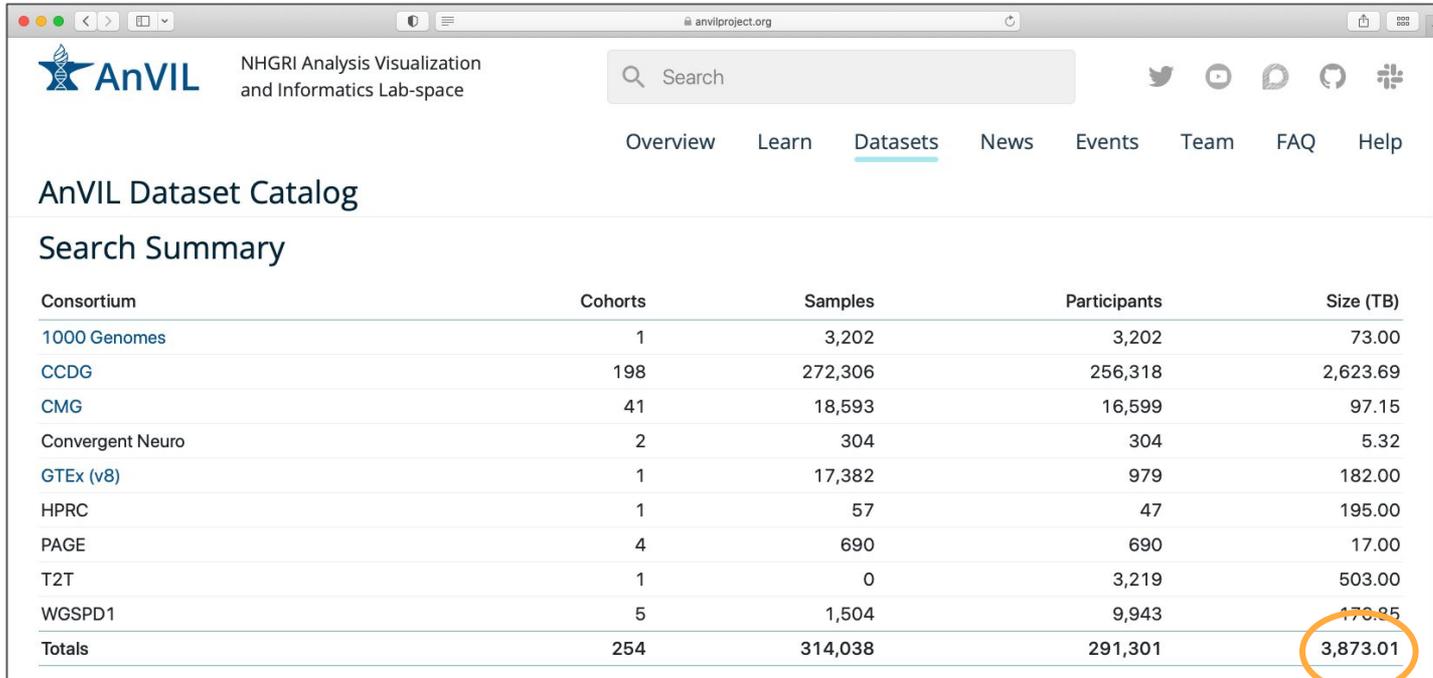
Terra Launches

Workflows	3082
Jupyter	1305
RStudio	1562
Galaxy	470

Communications

Twitter	648 followers
Slack	234 users

AnVIL Dataset Catalog



The screenshot shows the AnVIL Dataset Catalog website. The main content area displays a search summary table with the following data:

Consortium	Cohorts	Samples	Participants	Size (TB)
1000 Genomes	1	3,202	3,202	73.00
CCDG	198	272,306	256,318	2,623.69
CMG	41	18,593	16,599	97.15
Convergent Neuro	2	304	304	5.32
GTEx (v8)	1	17,382	979	182.00
HPRC	1	57	47	195.00
PAGE	4	690	690	17.00
T2T	1	0	3,219	503.00
WGSPD1	5	1,504	9,943	170.25
Totals	254	314,038	291,301	3,873.01

The total size of 3,873.01 TB is circled in orange in the original image. To the right of the table is a sidebar titled "Disease" with a list of diseases and their corresponding cohort counts. The sidebar shows "No selected terms." and a list of diseases with checkboxes and counts.

- >3.8Pb of data, >291,000 participants available
- Population-scale cohorts powers disease association studies
- Cross-project synthetic cohorts make existing data more valuable
- Connect multiple datatypes together to make new discoveries

<https://anvilproject.org/data>

NIH Cloud Platform Interoperability (NCPI) Dataset Catalog

NIH Cloud Platform Interoperability Effort

Search

e.g. disease, study name, dbGaP Id

Platform	Studies	Focus / Disease	Studies	Data Type	Studies	Study Design	Studies	Consent Code	Studies
<input type="checkbox"/> AnVIL	39	<input type="checkbox"/> Alzheimer Disease	2	<input type="checkbox"/> Allele-Specific Expression	1	<input type="checkbox"/> Case Set	35	<input type="checkbox"/> ALZ	1
<input type="checkbox"/> BDC	104	<input type="checkbox"/> Anemia, Sickle Cell	9	<input type="checkbox"/> AMPLICON	1	<input type="checkbox"/> Case-Control	27	<input type="checkbox"/> ALZ_NPU	1
<input type="checkbox"/> CRDC	27	<input type="checkbox"/> Arterial Pressure	2	<input type="checkbox"/> Bisulfite-Seq	4	<input type="checkbox"/> Clinical Trial	5	<input type="checkbox"/> ARR	1
<input type="checkbox"/> KFDRRC	17	<input type="checkbox"/> Asthma	17	<input type="checkbox"/> CHIP-Seq	3	<input type="checkbox"/> Control Set	1	<input type="checkbox"/> DS-AF-IRB-RD	2
		+ 56 more		+ 20 more		+ 6 more		+ 107 more	

No selected terms.

Download TSV [Download TSV](#) Copy URL [Copy URL](#)

Search Summary

Platform	Studies	Participants
AnVIL	39	178,609
BDC	104	429,666
CRDC	27	97,038
KFDRRC	17	14,984
Totals *	176	689,301



NIH Cloud Platform Interoperability Effort

Search

e.g. disease, study name, dbGaP Id

Platform	Studies	Focus / Disease	Studies	Data Type	Studies
<input type="checkbox"/> AnVIL	39	<input type="checkbox"/> Alzheimer Disease	2	<input type="checkbox"/> Allele-Specific Expression	1
<input type="checkbox"/> BDC	104	<input type="checkbox"/> Anemia, Sickle Cell	9	<input type="checkbox"/> AMPLICON	1
<input type="checkbox"/> CRDC	27	<input type="checkbox"/> Arterial Pressure	2	<input type="checkbox"/> Bisulfite-Seq	4
<input type="checkbox"/> KFDRRC	17	<input type="checkbox"/> Asthma	17	<input type="checkbox"/> CHIP-Seq	3
		+ 56 more		+ 20 more	

No selected terms.

Download TSV [Download TSV](#) Copy URL [Copy URL](#)

Search Summary

Platform	Studies
AnVIL	39
BDC	104
CRDC	27
KFDRRC	17
Totals *	176

Focus / Disease

Current selection: 4 Platforms 176 Studies 689,301 Participants

No selected terms.

Focus / Disease	Studies
<input type="checkbox"/> Alzheimer Disease	2
<input type="checkbox"/> Anemia, Sickle Cell	9
<input type="checkbox"/> Arterial Pressure	2
<input type="checkbox"/> Asthma	17
<input type="checkbox"/> Atherosclerosis	1
<input type="checkbox"/> Ataxia	23
<input type="checkbox"/> Blood Pressure	1
<input type="checkbox"/> Breast Neoplasms	1
<input type="checkbox"/> Cardiovascular Diseases	25
<input type="checkbox"/> Child Development Disorders, Pervasive	4
<input type="checkbox"/> Clot Lip	3
<input type="checkbox"/> Congenital Microtia	1
<input type="checkbox"/> Coronary Artery Disease	2
<input type="checkbox"/> Coronary Disease	1
<input type="checkbox"/> COVID-19	1
<input type="checkbox"/> Cranial Nerve Diseases	1
<input type="checkbox"/> Diabetes Mellitus, Type 1	1
<input type="checkbox"/> Disorders of Sex Development	1
<input type="checkbox"/> Encephalomalacia	1
<input type="checkbox"/> Epilepsy	1

11Pb / 689k participants and growing!
Cross-platform accessibility through several key technologies (RAS, DRS, FHIR)

Outreach & User Engagement

Upcoming

03 NOV 2021

Genome Informatics 2021 - POSTER SESSION

Modeling the computing requirements and costs for genomics analysis in the cloud

The 2021 Genome Informatics Meeting will cover topics including Microbial and Metagenomics; Sequencing Algorithms, Variant Discovery and Genome Assembly; Evolution, Complex Traits and Phylogenetics; Functional Genomics; Single Cell Genomics; and Epigenetics and Genome Structure.

19 JAN 2022

ASHG 2021 - INTERACTIVE WORKSHOP

Structural variant discovery from long-read sequencing data on the cloud with Galaxy in Terra

In this workshop, we will guide you through an end-to-end SV identification journey using Galaxy, a platform designed to facilitate access to computational methods for researchers without a programming background.

26 JAN 2022

ASHG 2021 - INTERACTIVE WORKSHOP

Reproducible Analysis of Human Pangenome Data using the AnVIL

This workshop will explore and demonstrate open access data from the Human Pangenome Research Consortium (HPRC), an NHGRI funded effort to create a more diverse and comprehensive reference human pangenome.

Announcing the AnVIL Cloud Credits Program (AC2) Awardees

Posted: June 03, 2021

NHGRI's Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL) cloud genomics platform is pleased to announce the awardees of the pilot phase of the AnVIL Cloud Credits (AC2) Program.

Awardees

Seventeen proposals were received from 14 different institutions and of these, the AC2 Review Committee (AC2RC) has awarded 6 proposals with cloud credits. Those awardees include:

- Alex Greiner | The University of Iowa | Graduate Student, "Burden analysis of inherited cardiac arrhythmia genes in epilepsy"
- Melissa Suzanne Cline | UC Santa Cruz Genomics Institute | Principal Investigator, "Leveraging AnVIL and Terra for secure collaboration on genetic variant interpretation"
- Andrew Davidson | University of California | Graduate Student, "Comprehensive characterization of transposable element expression across human tissues"
- Anahita Khojandi | University of Tennessee-Knoxville | Associate Professor, "Deep Learning for Accurate Tissue-Specific Prediction of Gene Expression in Large Deeply-Phenotyped Population"
- Anshul Kundaje | Stanford University | Principal Investigator, "Deciphering cis-regulatory syntax of a transcription factor binding atlas with interpretable deep learning models"
- Tychele N. Turner | Washington University in St. Louis | Principal Investigator, "A k-mer based approach to assess copy number in PacBio HiFi data"

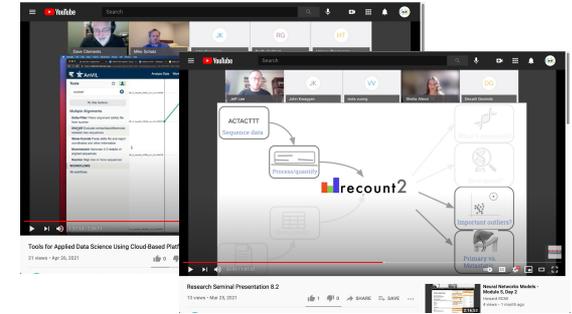
GSP/CCDG/CMG MAGIC Jamboree



PRIMED Consortium



Howard University VADSTI



Genomic Data Science Community Network



Portal: <http://anvilproject.org/>

Mailing List: help@lists.anvilproject.org

AnVIL Community Slack: <http://bit.ly/anvil-community>



New Results

 [Follow this preprint](#)

Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL)

 Michael C. Schatz, Anthony A. Philippakis, Enis Afgan, Eric Banks, Vincent J. Carey, Robert J. Carroll, Alessandro Culotti, Kyle Ellrott, Jeremy Goecks, Robert L. Grossman,  Ira M. Hall,  Kasper D. Hansen, Jonathan Lawson, Jeffrey T. Leek, Anne O'Donnell Luria, Stephen Mosher, Martin Morgan, Anton Nekrutenko, Brian D. O'Connor, Kevin Osborn, Benedict Paten, Candace Patterson, Frederick J. Tan, Casey Overby Taylor, Jennifer Vessio,  Levi Waldron, Ting Wang, Kristin Wuichet, AnVIL Team

AnVIL-powered COVID19 Analysis

Science

RESEARCH ARTICLES

Cite as: J. E. Lemieux *et al.*, *Science*
10.1126/science.abe3261 (2020).

Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events

Jacob E. Lemieux^{1,2*}†, Katherine J. Siddle^{1,3*}, Bennett M. Shaw^{1,2}, Christine Loretz¹, Stephen F. Schaffner^{1,3,4}, Adrienne Gordon Adams¹, Timelia Fink², Christopher H. Tomkins-Tinch^{1,3}, Lydia A. Krasnikova^{1,3}, Katherine C. DeRuff¹, Melissa Bauer^{1,6}, Kim A. Lagerborg^{1,6}, Erica Normandin^{1,7}, Sinéad B. Chapman¹, Steven K. Reilly^{1,3}, Melis N. Anahtar⁸, Aaron E. Carter¹, Cameron Myhrvold^{1,3}, Molly E. Kemball^{1,7}, Sushma Chaluvadi¹, Caroline Cusick¹, Katelyn Flowers¹, Anna Neum Cerrato¹, Maha Farhat^{1,9}, Damien Slater², Jason B. Harris^{2,11}, John A. Branda², David Hooper², Jessie M. Gaeta^{12,13}, Tra James O'Connell^{12,14,15}, Andreas Gnirke¹, Tami D. Lieberman^{1,16}, Anthony Philippakis¹, Meagan Burns², Catherine M. B. Luban^{1,17,18}, Edward T. Ryan^{2,4,15}, Sarah E. Turbett^{2,8,13}, Regina C. LaRoque^{2,15}, William P. Hanage¹⁹, Glen R. Gallagher²⁰, Madoff^{20,21}‡, Sandra Smole²¹‡, Virginia M. Pierce^{21,22}‡, Eric Rosenberg^{2,8*}‡, Parris C. Sabeti^{1,3,4,15,23}‡‡, Daniel J. Park²⁴‡, E. MacInnis^{1,4,13}‡†

¹Broad Institute of Harvard and MIT, 415 Main Street, Cambridge, MA 02142, USA. ²Division of Infectious Diseases, Massachusetts General Hospital. ³Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA. ⁴Department of Immunology and Infectious Chan School of Public Health, Harvard University, Boston, MA, USA. ⁵Massachusetts Department of Public Health, Boston, MA, USA. ⁶Harvard Program Biomedical Sciences, Harvard Medical School, Boston, MA 02115, USA. ⁷Department of Systems Biology, Harvard Medical School, Boston, MA, USA. ⁸Pathology, Massachusetts General Hospital, Boston, MA, USA. ⁹Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. ¹⁰Pathology, Massachusetts General Hospital, Boston, MA, USA. ¹¹Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. ¹²Institute for Medical Engineering and Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. ¹³Program in Molecular Medicine, Massachusetts Medical School, Worcester, MA 01605, USA. ¹⁴Massachusetts Consortium on Pathogen Readiness, Boston, MA, 02115, USA. ¹⁵Cent Disease Dynamics, Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA 02115, USA. ¹⁶University of Massachusetts Infectious Diseases and Immunology, Worcester, MA 01655. ¹⁷Pediatric Infectious Disease Unit, Massachusetts General Hospital for Children, Boston, MA, USA. ¹⁸Department of Pathology, Harvard Medical School, Boston, MA, USA. ¹⁹Howard Hughes Medical Institute, 4000 Jones Bridge Rd, Chevy Chase, MD, USA. ²⁰Department of Pathology, Harvard Medical School, Boston, MA, USA. ²¹Howard Hughes Medical Institute, 4000 Jones Bridge Rd, Chevy Chase, MD, USA. ²²Department of Pathology, Harvard Medical School, Boston, MA, USA. ²³Howard Hughes Medical Institute, 4000 Jones Bridge Rd, Chevy Chase, MD, USA. ²⁴Department of Pathology, Harvard Medical School, Boston, MA, USA.

*These authors contributed equally to this work.

†Corresponding author. Email: lemieux@broadinstitute.org (J.E.L.); parris@broadinstitute.org (P.C.S.); bronwyn@broadinstitute.org

‡These authors contributed equally to this work.

Analysis of 772 complete SARS-CoV-2 genomes from early in the Boston area epidemic reveals introductions of the virus, a small number of which led to most cases. The data revealed two superspreading events. One, in a skilled nursing facility, led to rapid transmission and significant impact in this vulnerable population but little broader spread, while other introductions into the facility had a large effect. The second, at an international business conference, produced sustained community spread and was exported, resulting in extensive regional, national, and international spread. The two events differed significantly in the genetic variation they generated, suggesting varying transmission dynamics in superspreading events. Our results show how genomic epidemiology can help understand the link between

COVID-19_Broad_Viral_NGS - x +

anvil.terra.bio/#workspaces/pathogen-genomic-surveillance/COVID-19_Broad_Viral_NGS

WORKSPACES pathogen-genomic-surveillance/COVID-19_Broad_Viral_NGS (read only)

DASHBOARD DATA NOTEBOOKS WORKFLOWS JOB HISTORY

ABOUT THE WORKSPACE

Massachusetts has been severely impacted by the COVID-19 pandemic, with 115,850 cases and 8,690 deaths as of August 22, 2020. Seventy percent of the state's 6.9 M population lives in the city of Boston and its surrounding communities. To understand the introduction and spread of SARS-CoV-2 in this region, the Broad Institute is sequencing viral genomes from COVID-19 cases from the Boston area for genomic epidemiological analyses.

This dataset provides the first high resolution view of the introductions and spread of SARS-CoV-2 in the greater Boston area based on viral genomic data. All genomes were obtained from nasopharyngeal swabs from individuals with confirmed SARS-CoV-2 infection from March 3rd and May 9th, 2020. These cases represent a non-random sample from a single tertiary care center whose clinical catchment area primarily involves eastern Massachusetts.

To view the related blog post on COVID-19 efforts by the Viral Genomics group at Broad, please see [here](#).

To view the Terra blog post related to this COVID-19 viral NGS workspace, please see [here](#).

Epidemiological analysis results are available for review as a pre-print on [Virological](#) and can be found [here](#).

Laboratory protocols used by the Broad Viral Genomics group can be found [here](#).

WORKSPACE INFORMATION

ORGANIZATION	LAST UPDATED
MIT	8/27/2020

OWNERS

OWNER	ACCESS LEVEL
dpaik@broadinstitute.org	Reader
cloretz@broadinstitute.org	Reader
schaluva@broadinstitute.org	Reader

TAGS

No tags yet

Google Bucket

Name: fc-061a81bb-6bbb-4306-8f07-
Location: multi-region: US
Open in browser [G](#)

The Data

In this workspace we've provided tools and data, so that labs can go from raw reads (uBAM), through to producing a phylogenetic tree with their private and publicly available data.

The data in this workspace includes:

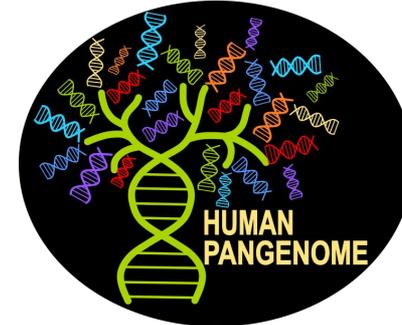
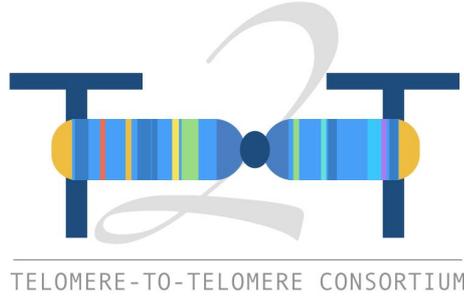
- High-quality viral genomes sequenced from nasopharyngeal swabs from individuals with confirmed SARS-CoV-2 infection from MGH and the MA DPH
- Over 5,000 viral genomes from [GISAID](#) that can be used to build NextStrain phylogenetic trees with your data

From .BAM to NextStrain Tree and GenBank Data Submission

```
graph TD
    A[Your .BAM files] --> B[Perform Assembly]
    B --> C[Your .fasta files]
    C --> D[Perform GenBank Data Submission]
    E[Broad/MGH .fasta] --> D
    F[GenBank .fasta] --> D
```

https://anvil.terra.bio/#workspaces/pathogen-genomic-surveillance/COVID-19_Broad_Viral_NGS

T2T & HPRC on AnVIL



ABOUT THE WORKSPACE

Telomere-to-Telomere (T2T) Consortium's AnVIL_T2T Workspace

The [Telomere-to-Telomere \(T2T\) consortium](#) is an open, community-based effort to de novo assemble the first complete reference human genome from the CHM13 hydatidiform mole. Using a combination of PacBio HiFi sequencing and Oxford Nanopore ultra long reads, the recently released CHM13v1 reference genome is nearly perfect, with an estimated sequence accuracy exceeding QV70 and only 5 rRNA arrays left unresolved. The genome includes more than 100 Mbp of novel sequence compared to GRCh38, corrects many structural errors in the GRCh38 reference genome, and unlocks the most complex regions of the genome to clinical and functional study for the first time.

Currently Available Data

Here we use the T2T-CHM13 reference genome to investigate how it improves variant calling for individual samples, trios, and population-scale analysis. This includes 17 samples from diverse ethnicities sequenced with long reads that we analyze for SNVs, indels and structural variants using PEPPER-Margin-DeepVariant and Sniffles, along with all 3,202 short-read samples from the recently extended 1000 Genomes Project collection that we analyze using the CATK HaplotyPcaller for SNVs and indels on the NHGRI AnVIL Cloud Platform. We demonstrate that the CHM13 reference improves read mapping and variant calling across all samples in a number of major ways:

1. Adds over 80 million base pairs of sequence that can be effectively used for variant calling with long reads.

WORKSPACE INFORMATION

CREATION DATE 2/23/2021	LAST UPDATED 3/9/2021
SUBMISSIONS 0	ACCESS LEVEL Writer
EST. \$/MONTH \$2340.52	GOOGLE PROJECT ID anvil-datasto...

OWNERS

sizarate96@gmail.com
candace@broadinstitute.org
anvil-admins@firecloud.org

TAGS

Add a tag
No tags yet

Google Bucket

Name: fc-47de7dae-e8e6-429c-b760-...
Location: multi-region: US
Open in browser

ABOUT THE WORKSPACE

Human Pangenome Reference Consortium's AnVIL_HPRC Workspace

TOWARDS A COMPLETE REFERENCE OF HUMAN GENOME DIVERSITY

This workspace holds sequencing and assembly data submitted to the [Human Pangenome Reference Consortium](#). Data is stored in this workspace to allow immediate use by researchers participating in the Human Pangenome Project. Data in this workspace is constantly being added and updated and the workspace is under active development as our production pipeline continues.

WORKSPACE INFORMATION

CREATION DATE 3/12/2020	LAST UPDATED 3/9/2021
SUBMISSIONS 0	ACCESS LEVEL Reader
GOOGLE PROJECT ID anvil-datasto...	

OWNERS

schaluva@broadinstitute.org
miten@soe.ucsc.edu
juklucas@ucsc.edu
bhannafi@ucsc.edu
arula@broadinstitute.org
esheets@ucsc.edu

TAGS

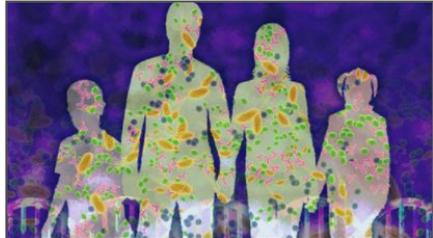
No tags yet

Google Bucket

Name: fc-4310e737-a388-4a10-8c9e-...
Location: Loading...
Open in browser

https://anvil.terra.bio/#workspaces/anvil-datastorage/AnVIL_T2T
https://anvil.terra.bio/#workspaces/anvil-datastorage/AnVIL_HPRC

Clinical Engagements



RESEARCH FUNDING
Centers for Common Disease Genomics >



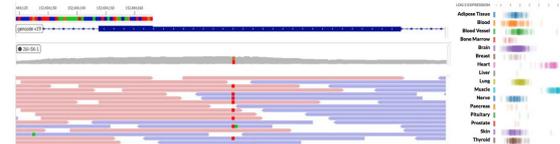
RESEARCH FUNDING
Centers for Mendelian Genomics >

NEB chr2:152404194 missense
G>T HGVS.C c.20216C>A
HGVS.P p.Trh6739Asn

3KG_WGS 0.0 26i-SK-1 27i-AK-1 28i-TK-1 29i-KA-1
3KG_WGS_POPMAX 0.0 G/T G/T G/G G/T
EXAC_V0 0.0
EXAC_V0_POPMAX 0.0

Mendelian Inheritance:

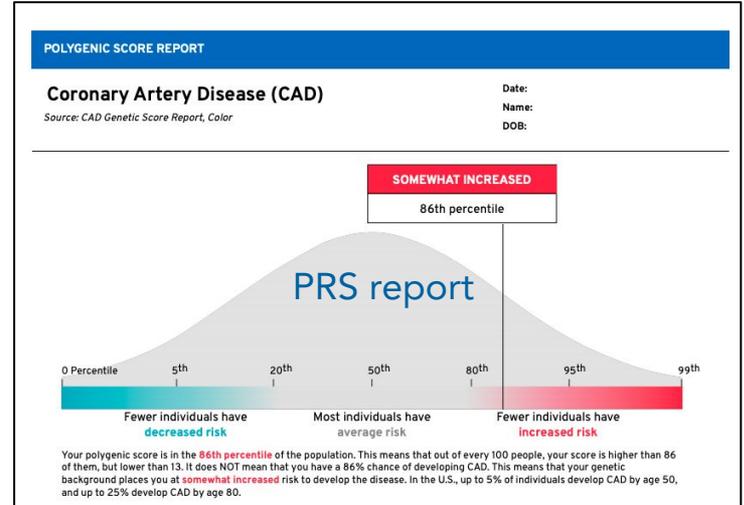
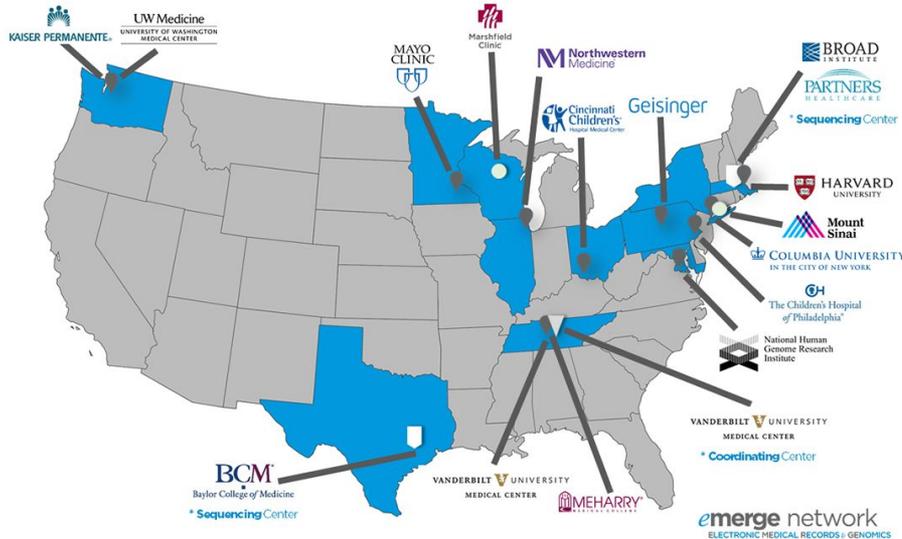
- Recessive
 - Homozygous Recessive
 - X-Linked Recessive
 - Compound Heterozygous
- Dominant
- De Novo Dominant



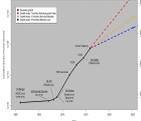
Individual **Sex** **Affected Status** **PhenoTIPS** **Candidate Genes, Variants** **MatchMakerExchange**

Individual	Sex	Affected Status	PhenoTIPS	Candidate Genes, Variants	MatchMakerExchange
29i-KA-1	Female	Unaffected	Edit View	-	
28i-TK-1	Male	Unaffected	Edit View	-	
27i-AK-1	Male	Affected	Edit View	Gene: NEB Variant: chr2:152581399 TG -> T	Submit to MME
26i-SK-1	Female	Affected	Edit View	Gene: NEB Variant: chr2:152581399 TG -> T	Submit to MME

seqr



Data

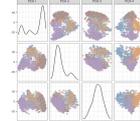


Data as a product to
lower barriers to
discovery

Support many (all!)
NHGRI consortia

Pioneer new models
of data use oversight
and data governance

Computation



Increased number
of tools & analysis
types

Interactive, batch,
and visual analytics

Expand capabilities
for predictive
biology & medicine

Users

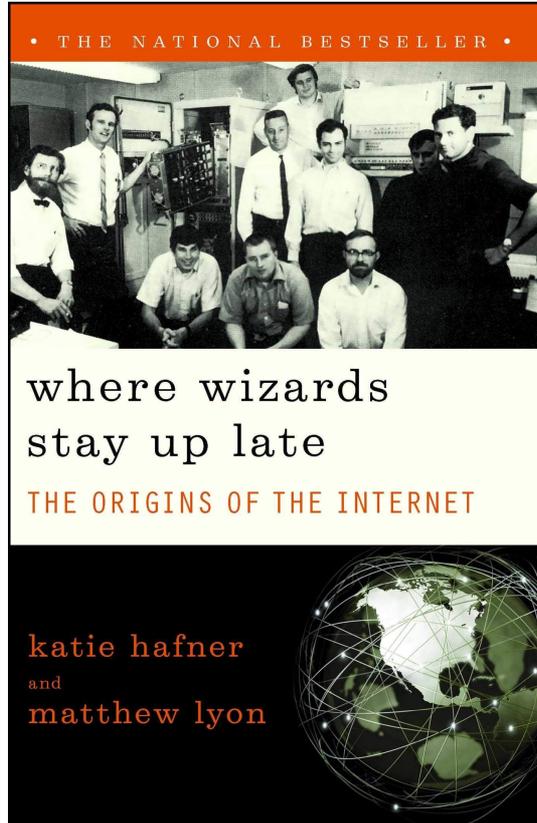


Increase number
of users &
consortiums

Empower users to do
more on their own

Serve as the platform
for cutting edge
biomedical research

Closing thought on interoperability



We should not underestimate the importance of this effort...

- If we are successful, we will catalyze the creation of an open and federated data ecosystem.
 - Others have done it before (SWIFT, the internet, the web).
- If we fail, we will degenerate into a collection of monolithic data silos
 - Others have done this before too (medical records in US hospitals)...

Today's deeper dives

Data submission and consortia engagement

Analysis tools

Concurrent Breakouts --
Session 1

Infrastructure

Outreach and training

Concurrent Breakouts --
Session 2