Genomics at scale with the NHGRI AnVIL

(Analysis, Visualization, and Informatics Lab-space)

Michael Schatz

November 7, 2023

@mike_schatz

Schatzlab Overview

Human Disease Genetics

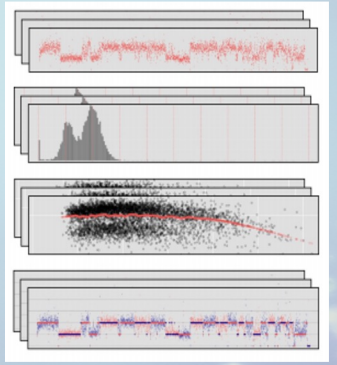
Nurk *et al.* (2022)
Aganezov *et al.* (2020)

Agricultural Genomics

Alonge *et. al.* (2022)
Naish *et al.* (2021)

Algorithmics & Systems Research

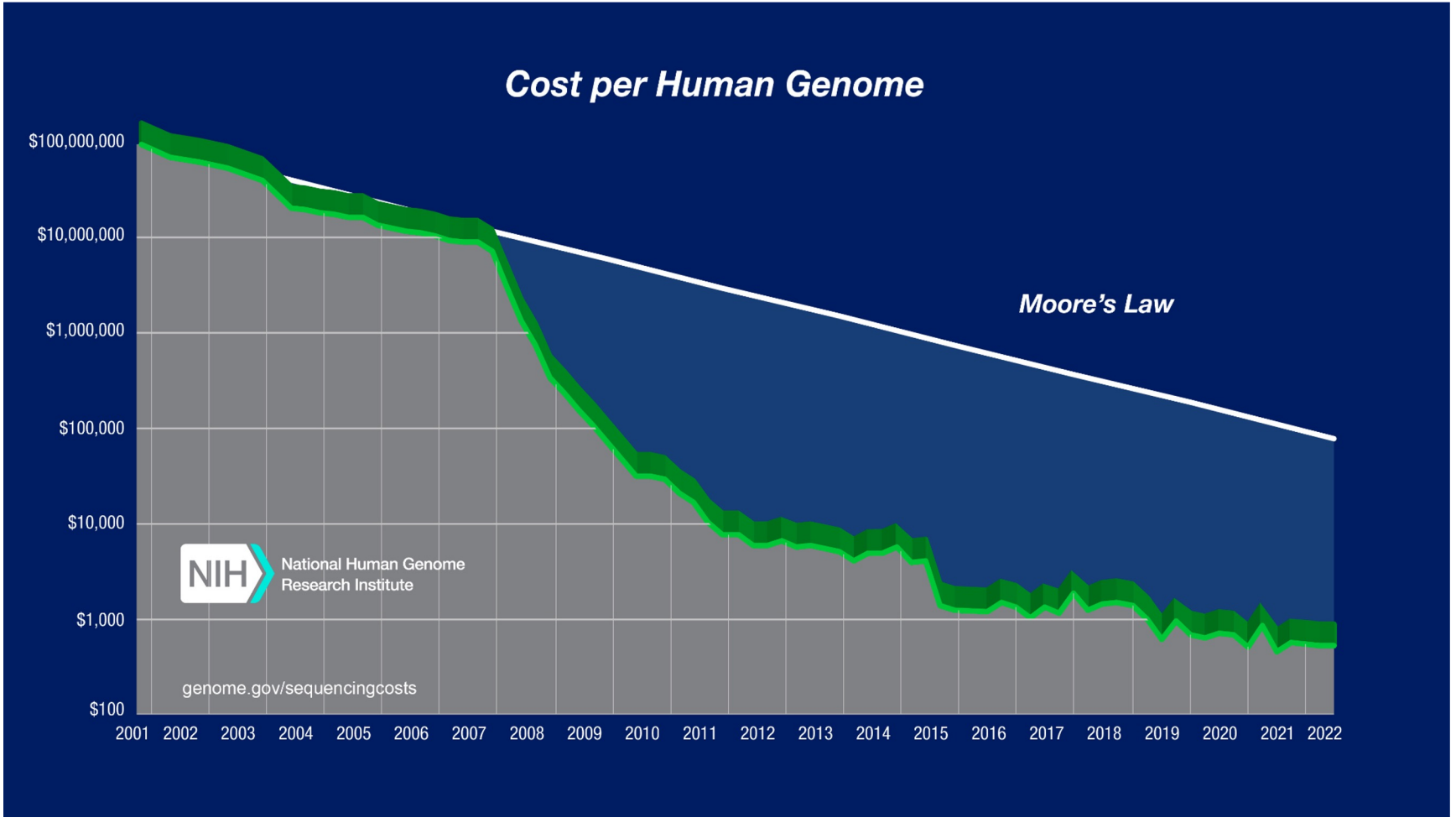
Kirsche *et al.* (2023)
Schatz, Philippakis *et al.* (2021)



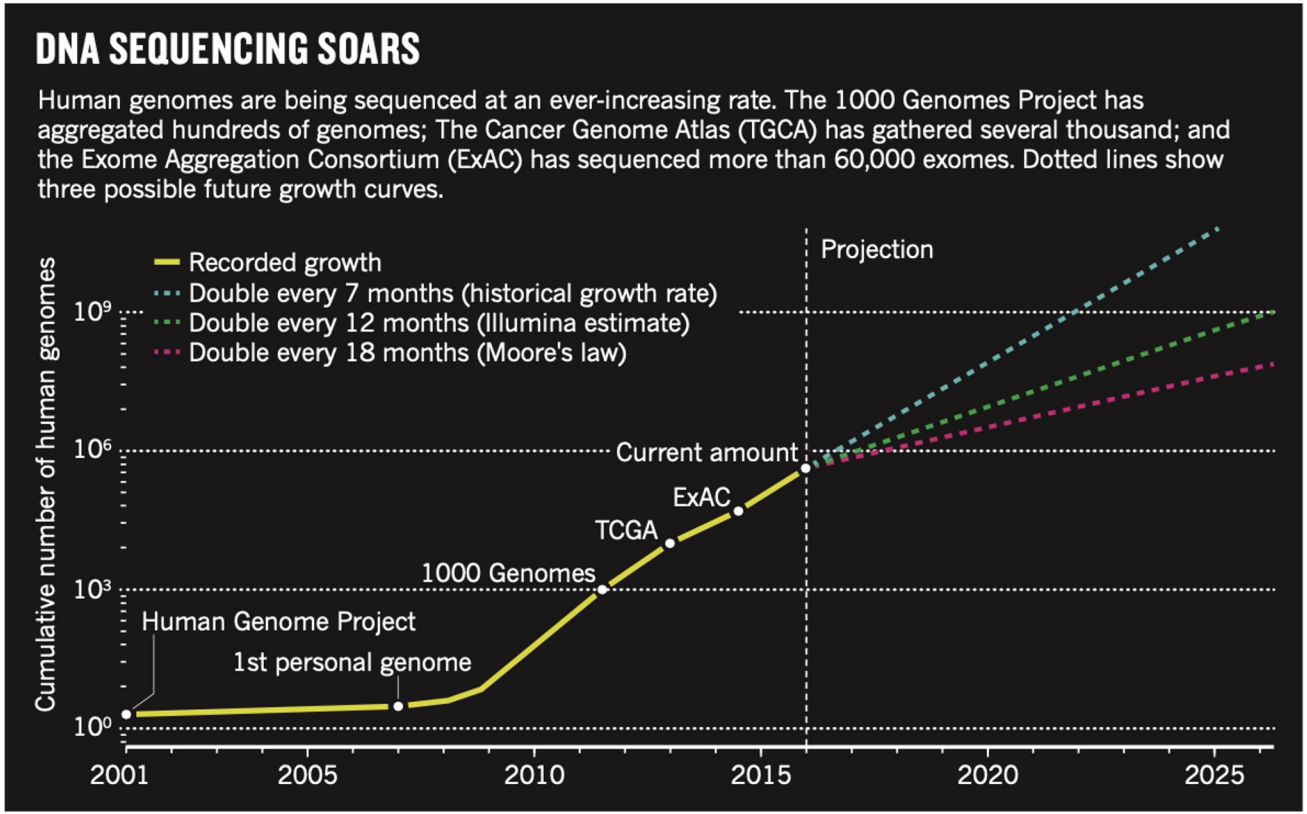
Single Cell & Single Molecule

Kirsche *et al.* (2023)
Kovaka *et al.* (2020)

Growth of Genomics



Growth of Genomics



Big Data: Astronomical or Genomical?

Stephens, Z, et al. (2015) PLOS Biology doi: 10.1371/journal.pbio.1002195

How much is a zettabyte?

Unit	Size
Byte	1
Kilobyte	1,000
Megabyte	1,000,000
Gigabyte	1,000,000,000
Terabyte	1,000,000,000,000
Petabyte	1,000,000,000,000,000
Exabyte	1,000,000,000,000,000,000
Zettabyte	1,000,000,000,000,000,000,000

How much is a zettabyte?

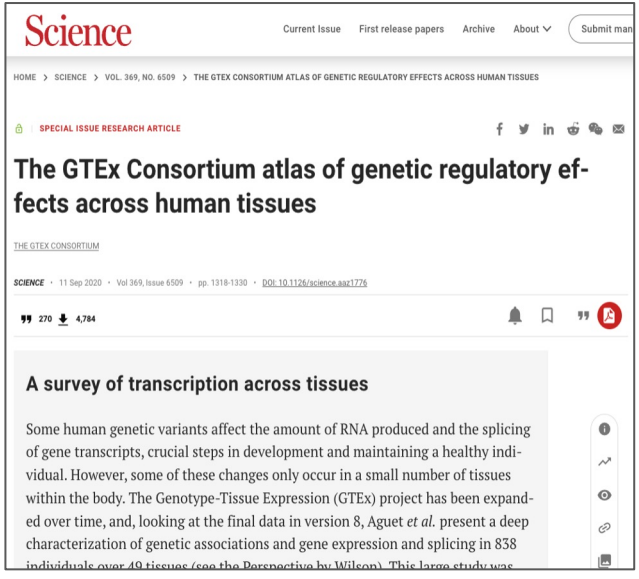
100 GB / Genome
4.7GB / DVD
~20 DVDs / Genome
X
10,000,000,000 Genomes
=
1ZB Data
200,000,000,000 DVDs

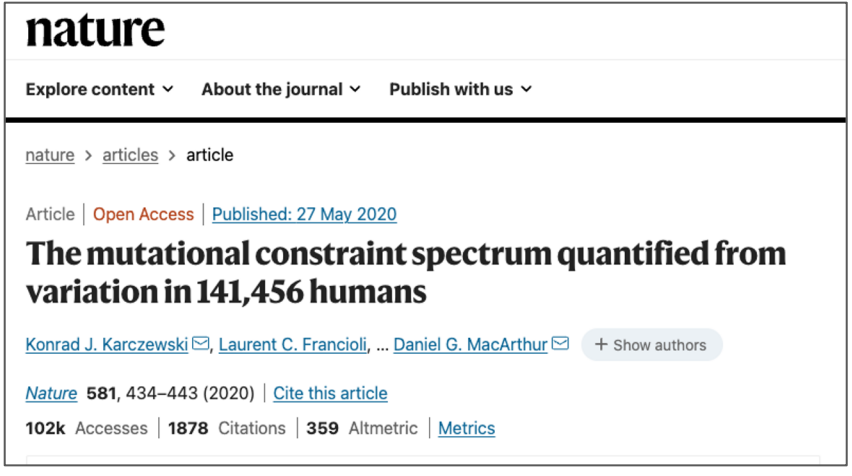
150,000 miles of DVDs
~ ½ distance to moon

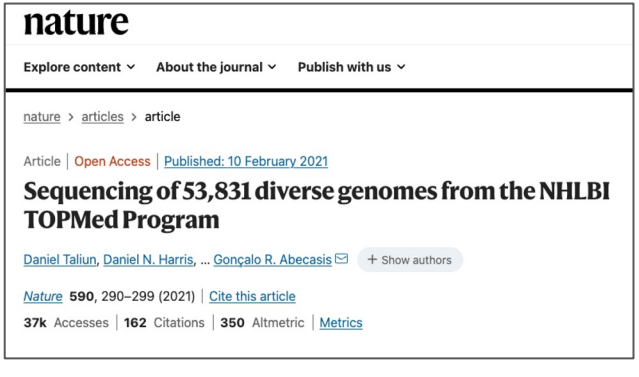
YouTube

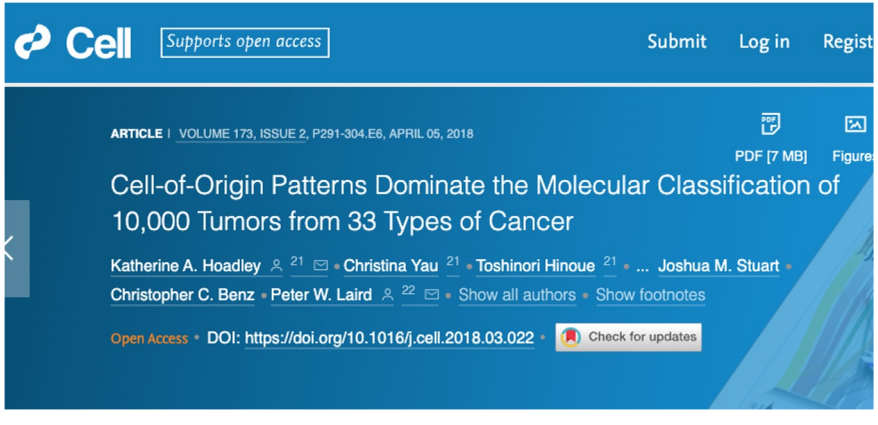
Both currently ~1Eb
And growing exponentially

Growth of Genomics

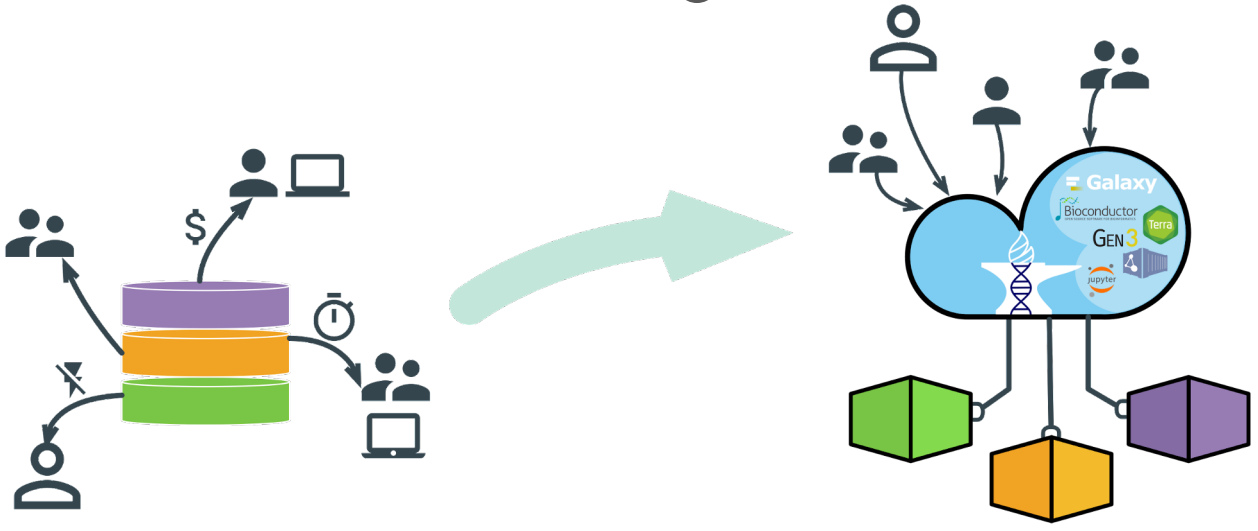








AnVIL: Inverting the model of genomic data sharing



Traditional: Bring data to the researcher

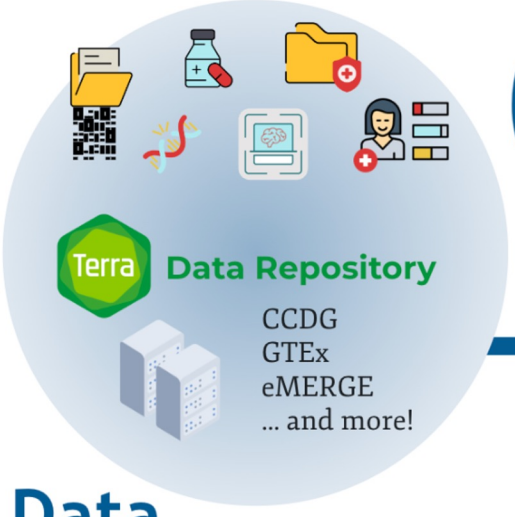
- Copying/moving data is costly
- Harder to enforce security
- Redundant infrastructure
- Siloed compute

Goal: Bring researcher to the data

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

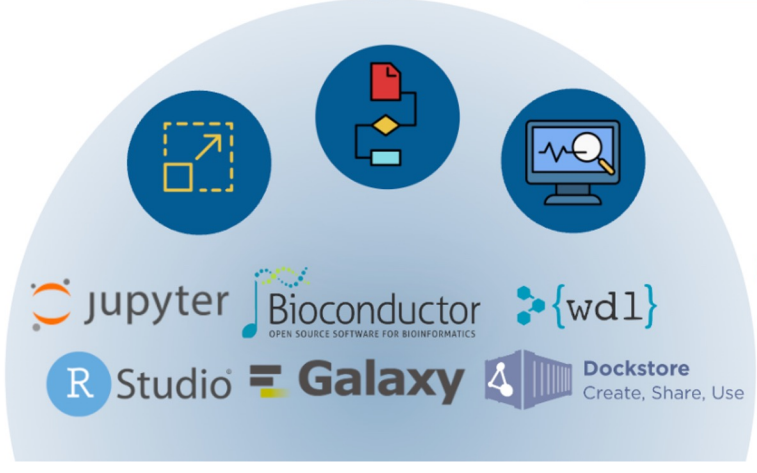
NHGRI AnVIL

Secure data storage, analysis, and sharing platform



Data

- Bring your structured data or use a flexible data schema
- Access public and managed-access data hosted in AnVIL



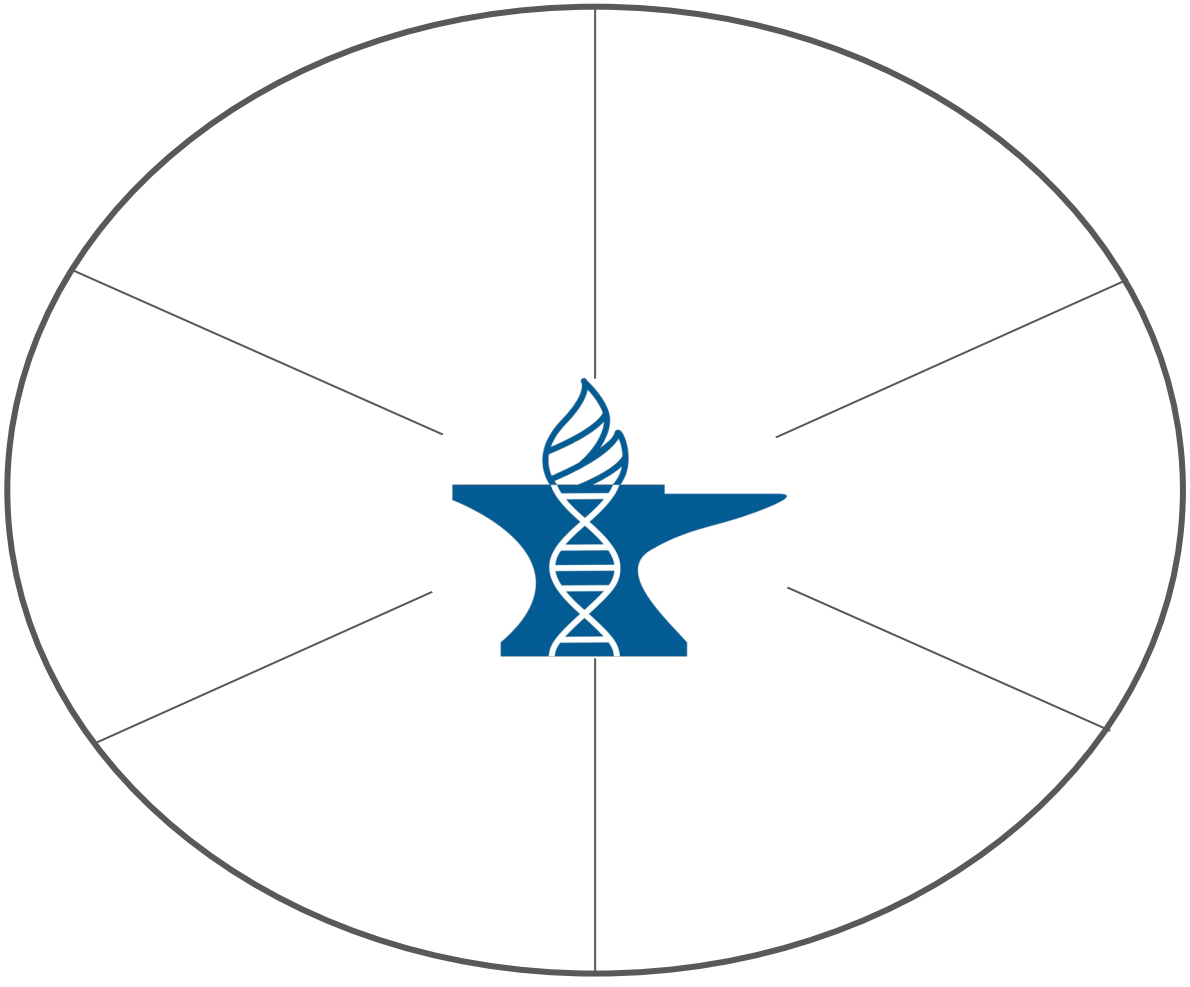
Community

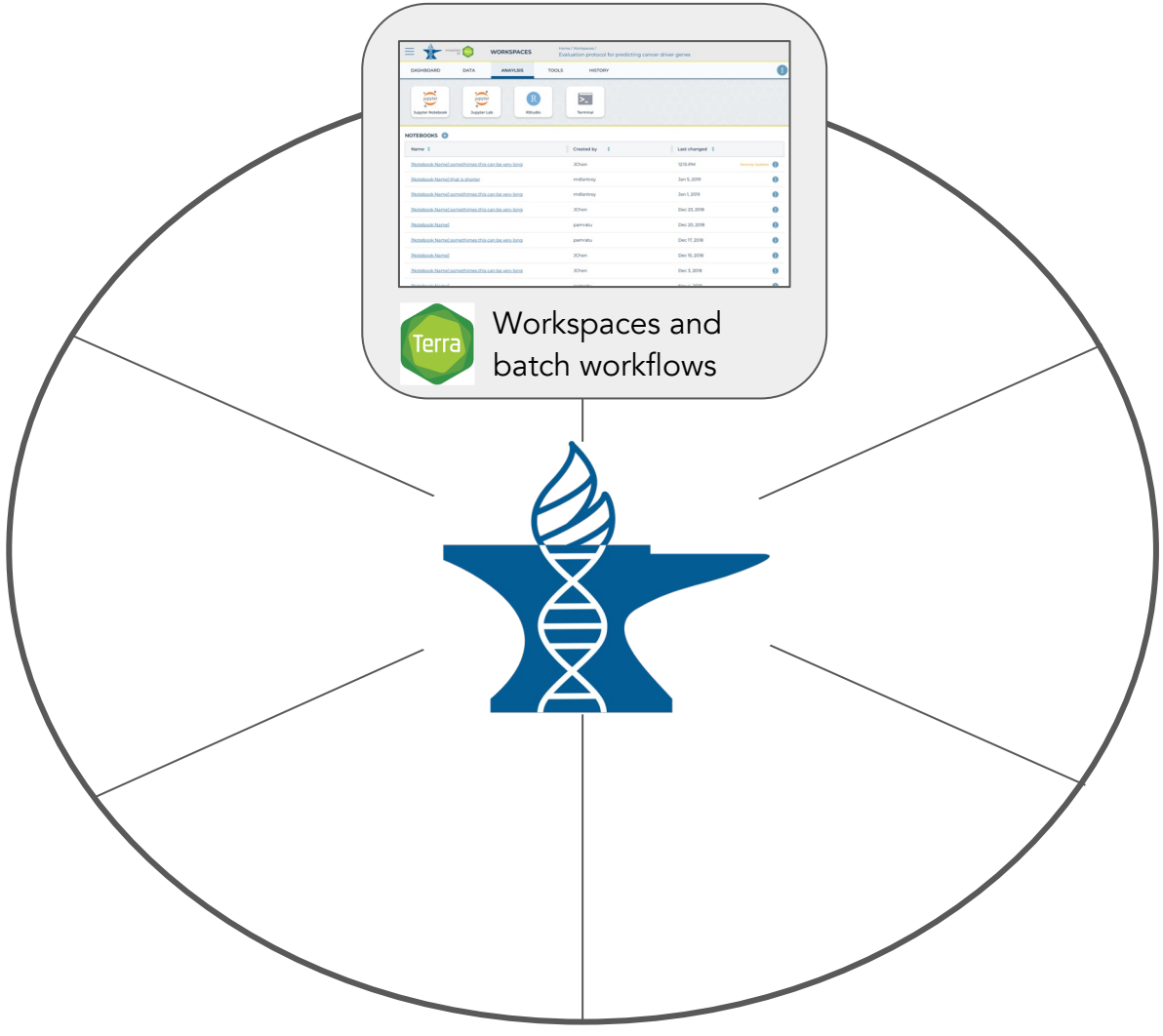
- Share your analysis and data with collaborators or with the world
- Get support at help.anvilproject.org

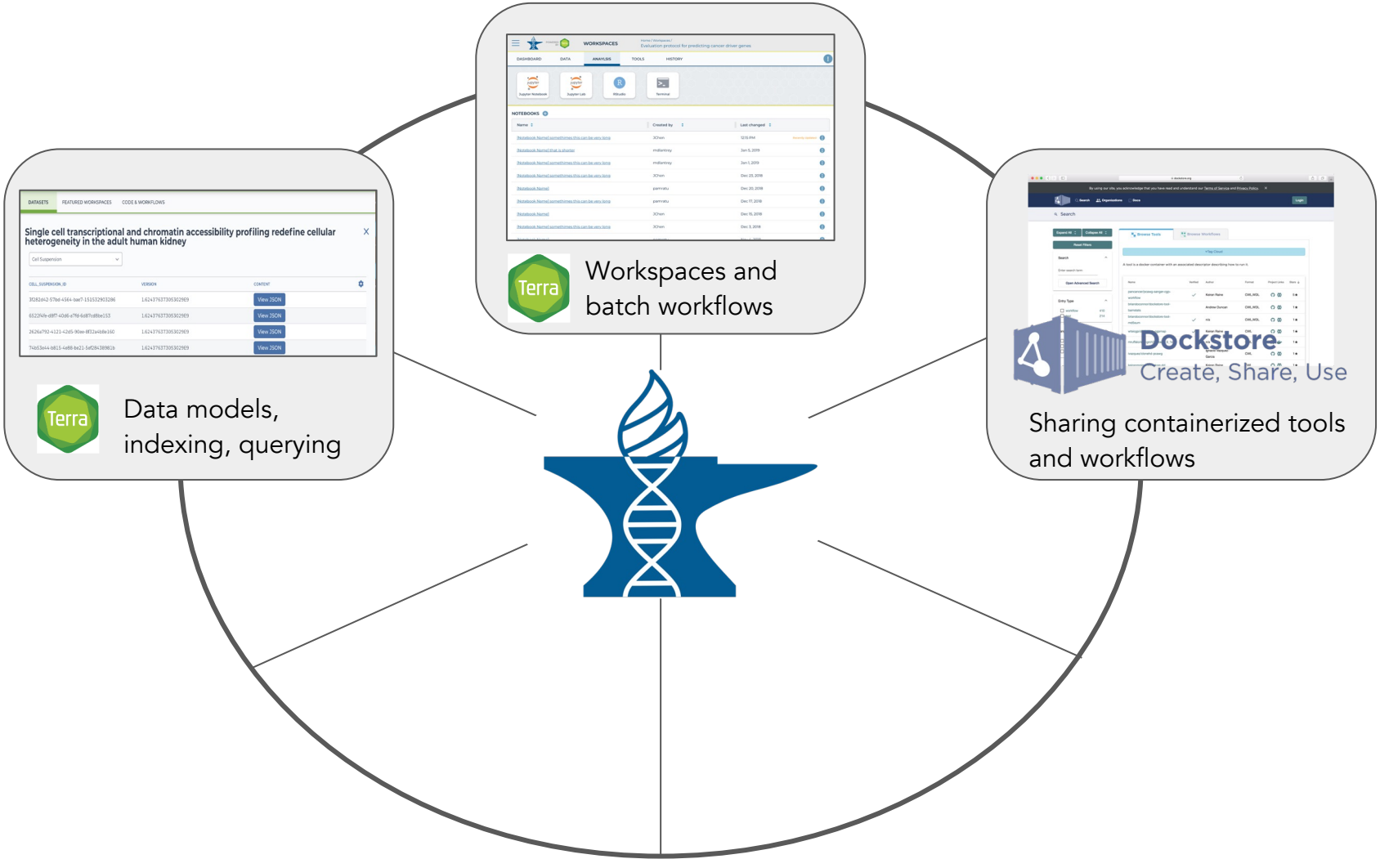
Compute

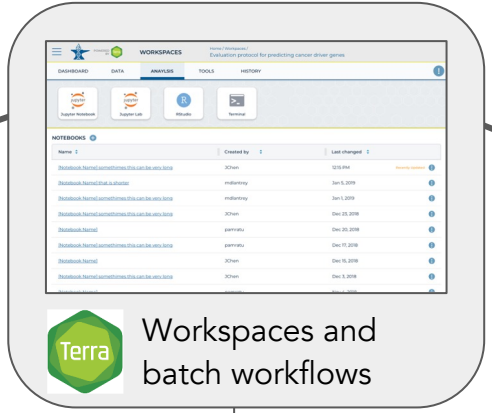
- Leverage elastic, scalable compute
- Bring your own pipelines, find shared
- Use powerful bioinformatics platforms

anvilproject.org



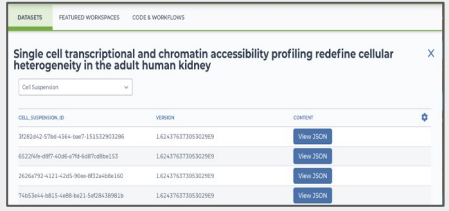

WORKSPACES
Evaluation protocol for predicting cancer driver genes
DASHBOARD
DATA
ANALYSIS
TOOLS
HISTORY
NOTEBOOKS
Name
Created by
Last changed
Terra
Workspaces and
batch workflows


WORKSPACES
DASHBOARD
DATA
ANALYSIS
TOOLS
HISTORY
NOTEBOOKS
Terra
Workspaces and batch workflows
DATASETS
FEATURED WORKSPACES
CODE & WORKFLOWS
Single cell transcriptional and chromatin accessibility profiling redefine cellular heterogeneity in the adult human kidney
Cell Suspension
View JSON
Terra
Data models, indexing, querying
Dockstore
Create, Share, Use
Sharing containerized tools and workflows



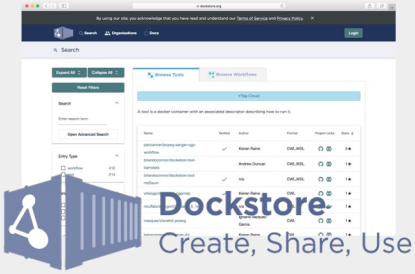
Terra

Workspaces and batch workflows

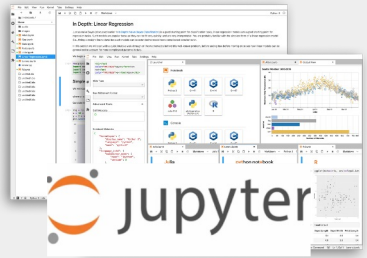


Terra

Data models, indexing, querying

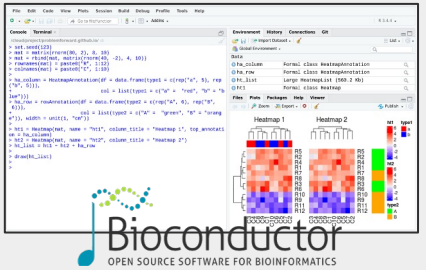


Sharing containerized tools and workflows



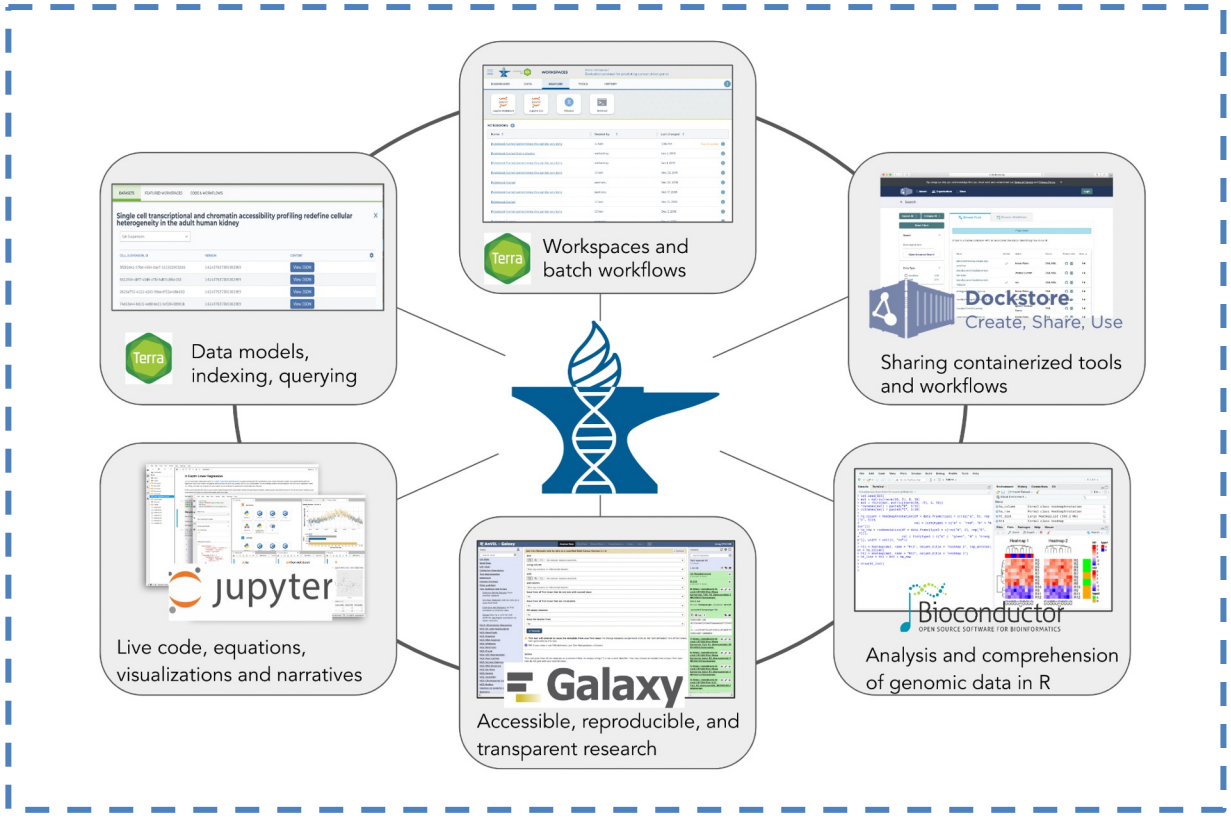
Live code, equations, visualizations and narratives

Accessible, reproducible, and transparent research



Analysis and comprehension of genomic data in R

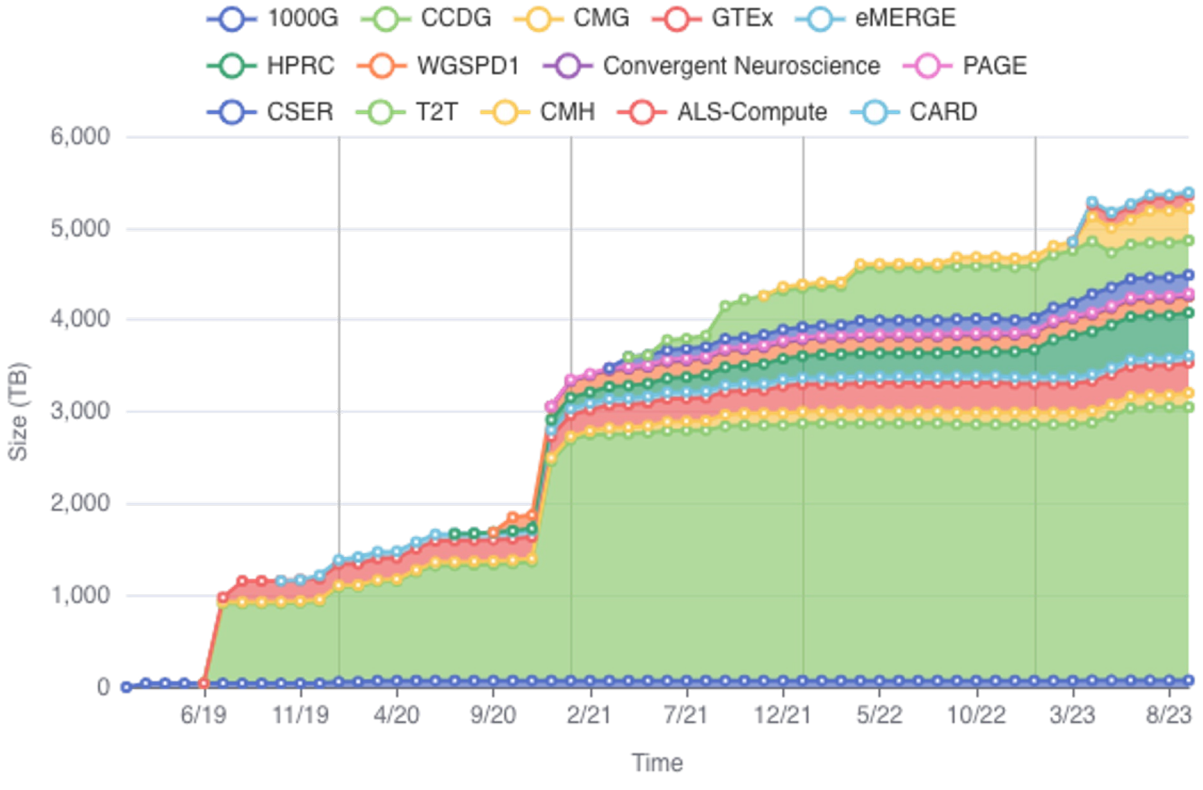
AnVIL: A Secure Federated Data & Compute Ecosystem



FedRAMP

Implemented on Google Cloud Platform

Data ingestion & harmonization



Current Consortia

- CCDG
- CMG
- GTEx
- 1000 Genomes
- eMERGE
- PAGE
- T2T & HPRC

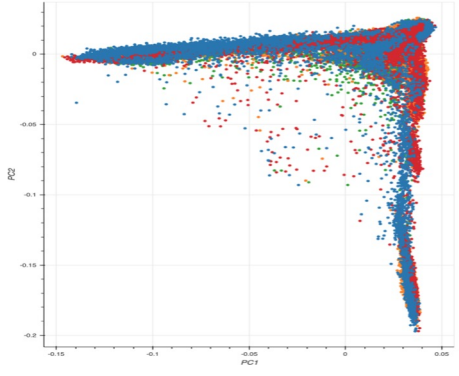
Ongoing Consortia

- GREGOR
- PRIMED
- IGVF
- Covid19hg
- CSER
- NIA, NIMH, UDN

Over 5 petabytes, 600,000 genomes and growing every day!

Powered by AnVIL

Centers for Common Disease Genomics



136,959 WGS samples

Baylor College: 35,493

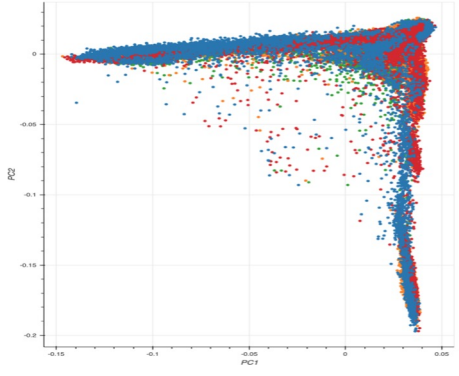
Broad Institute: 16,828

NYGC: 40,975

WashU: 43,620

Powered by AnVIL

Centers for Common Disease Genomics



136,959 WGS samples

Baylor College: 35,493

Broad Institute: 16,828

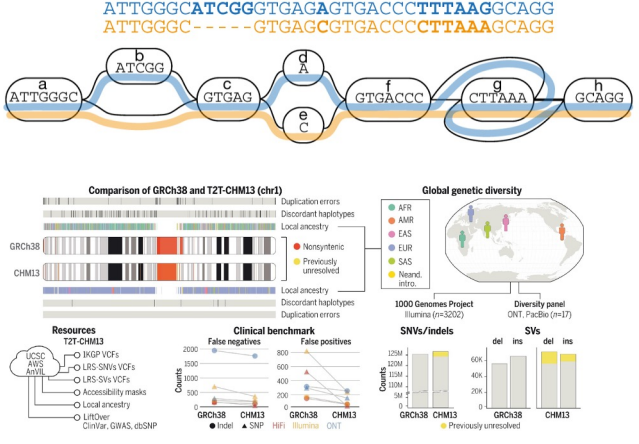
NYGC: 40,975

WashU: 43,620

T2T TELOMERE-TO-TELOMERE CONSORTIUM

TOWARDS A COMPLETE REFERENCE OF HUMAN GENOME DIVERSITY

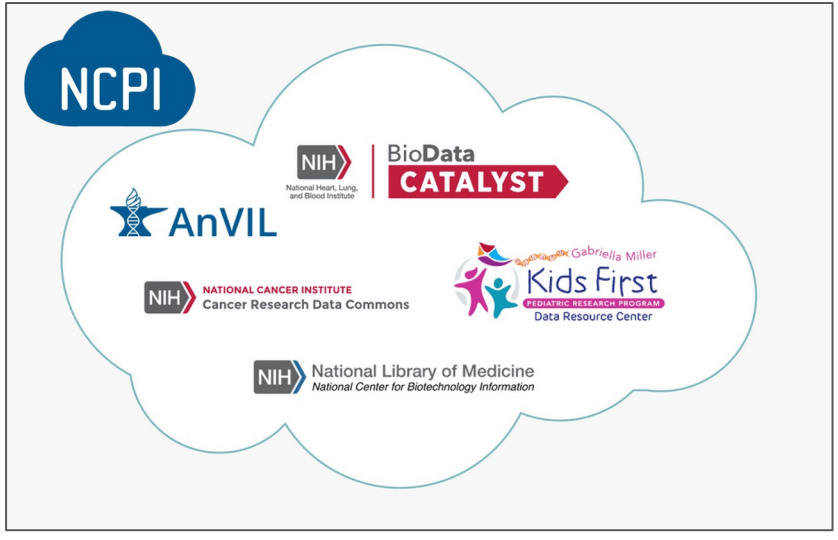
HUMAN PANGENOME

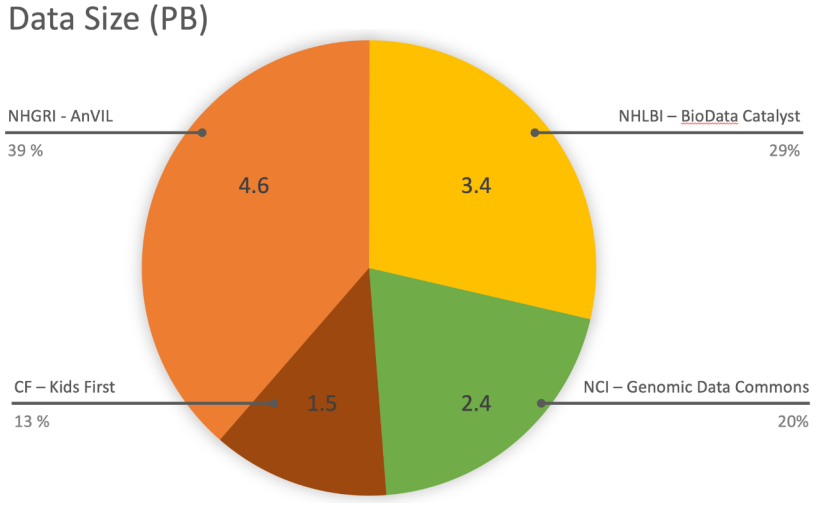


Nurk et al. (2022) Science

Liao et al. (2023) Nature

NCPI & The NCPI Dataset Catalog





Researcher Auth Service

Data Repository Service

Fast Healthcare Interoperability Resources

12Pb / 828k participants and growing!
Cross-platform accessibility through several key technologies

DUOS - Data Use Oversight System

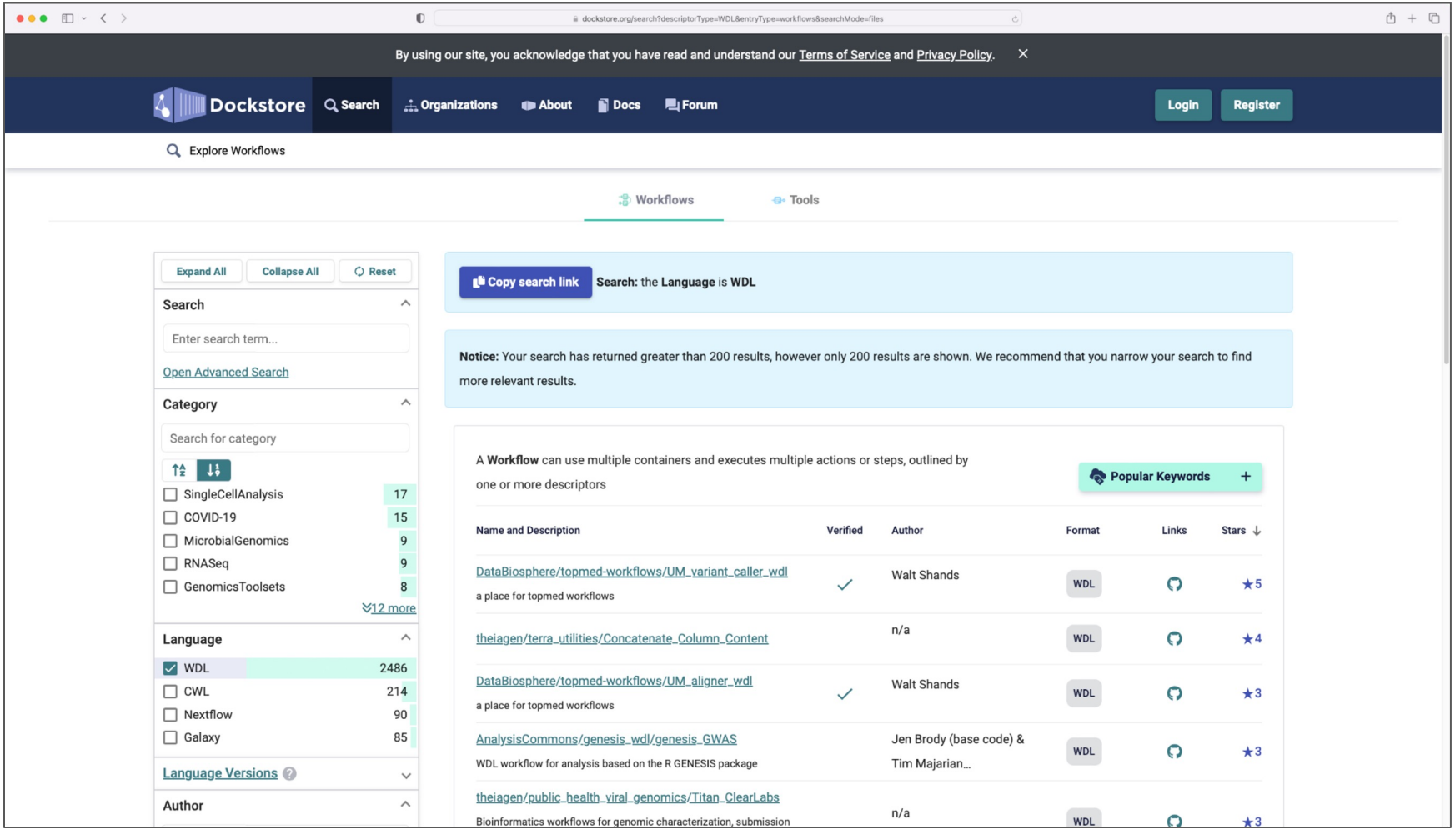
- Interfaces to transform data use limitations and data access requests to human and machine-readable codes (GA4GH Data Use Ontology (DUO))
- A matching algorithm that checks if access requests are within data use limitations
- Interfaces for Data Access Committees to adjudicate whether structuring and matching has been done appropriately

https://duos.org/

- Over 50 GTEx access requests processed by the NHGRI DAC in DUOS
- Initial testing with NHGRI, NHLBI, NIAID, and JAAMH has been very positive
- DUOS Algorithm 95% concordant with manual review (slightly more conservative)

What can you do with AnVIL?

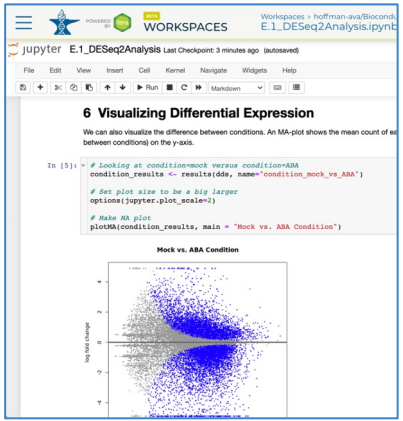
What can you do with AnVIL?



What can you do with AnVIL?

What can you do with AnVIL?

AnVIL Analysis Platforms

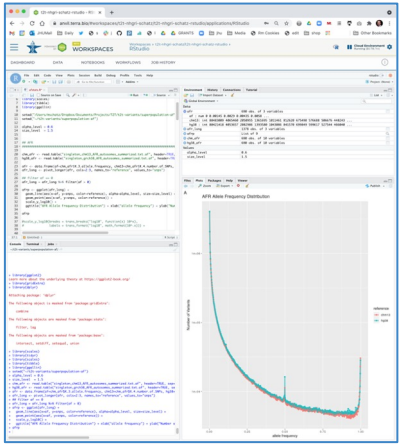


jupyter

+ Code, text and plots in one document
+ Supports coding in Python or R
- Least scalable, not a complete IDE

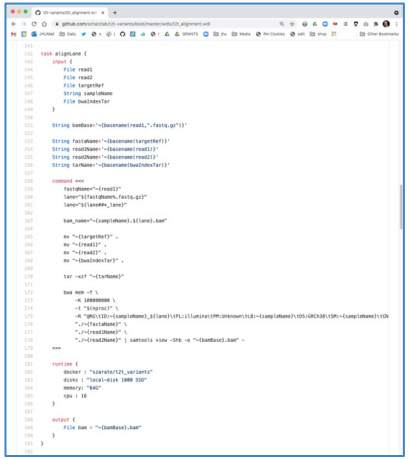
Galaxy

+ Graphical interface for thousands of tools and workflows
+ Highly accessible and reproducible
- Tools must be preconfigured to use



R Bioconductor

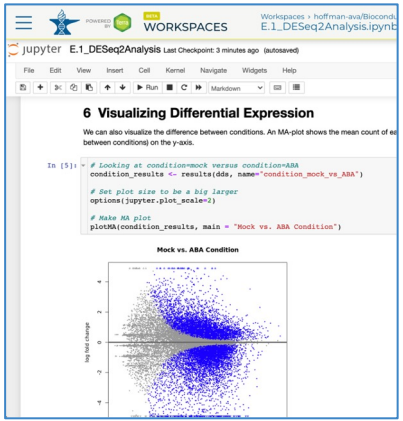
+ Feature rich IDE for programming in R
+ Rich statistics & ML and visualizations
- Limited support for other programming languages



{wdl}

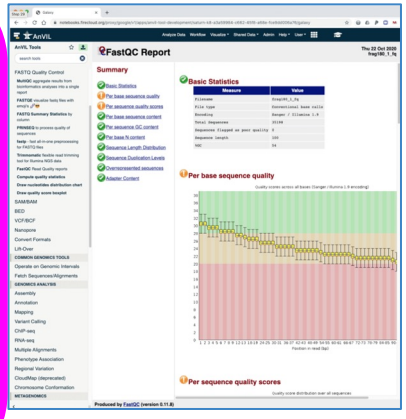
+ Extremely scalable and flexible
- Most technically demanding
- Requires careful cost management

AnVIL Analysis Platforms (coding)



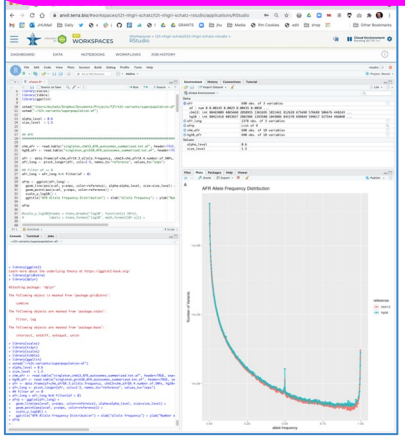
jupyter

+ Code, text and plots in one document
+ Supports coding in Python or R
- Least scalable, not a complete IDE



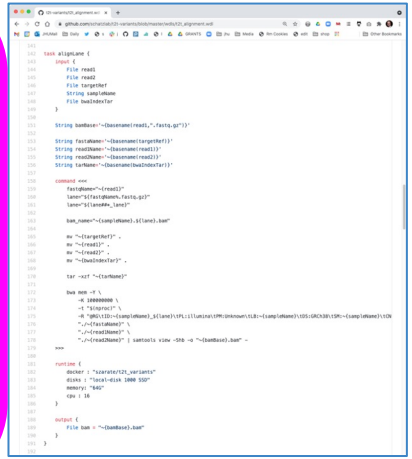
Galaxy

+ Graphical interface for thousands of tools and workflows
+ Highly accessible and reproducible
- Tools must be preconfigured to use



R Bioconductor

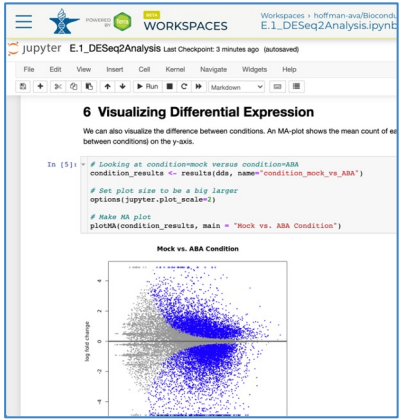
+ Feature rich IDE for programming in R
+ Rich statistics & ML and visualizations
- Limited support for other programming languages



{wdl}

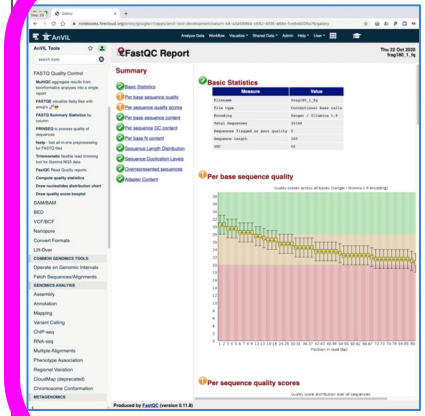
+ Extremely scalable and flexible
- Most technically demanding
- Requires careful cost management

AnVIL Analysis Platforms (graphical workflows)



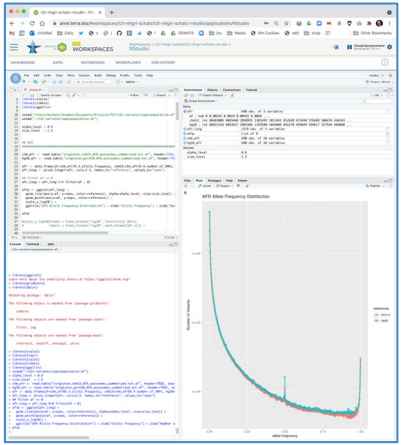
jupyter

+ Code, text and plots in one document
+ Supports coding in Python or R
- Least scalable, not a complete IDE



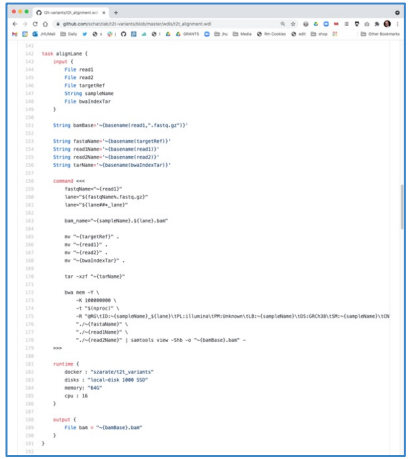
Galaxy

+ Graphical interface for thousands of tools and workflows
+ Highly accessible and reproducible
- Tools must be preconfigured to use



R Bioconductor

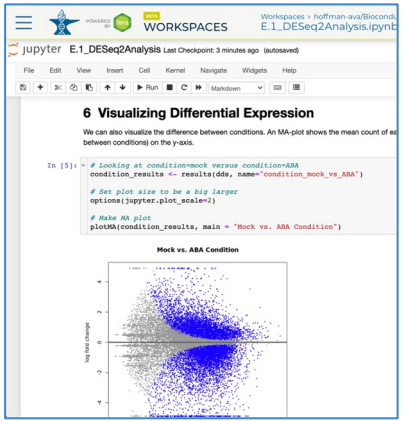
+ Feature rich IDE for programming in R
+ Rich statistics & ML and visualizations
- Limited support for other programming languages



{wdl}

+ Extremely scalable and flexible
- Most technically demanding
- Requires careful cost management

AnVIL Analysis Platforms (coding workflows)

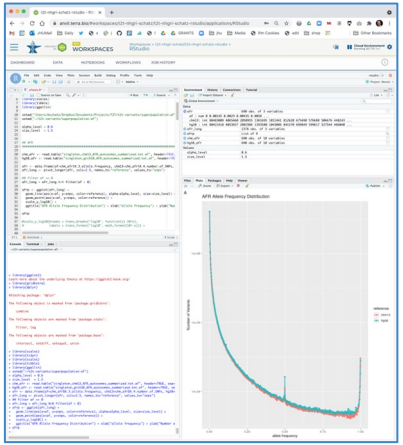


jupyter

+ Code, text and plots in one document
+ Supports coding in Python or R
- Least scalable, not a complete IDE

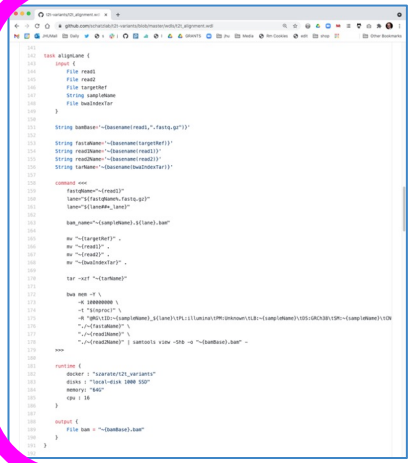
Galaxy

+ Graphical interface for thousands of tools and workflows
+ Highly accessible and reproducible
- Tools must be preconfigured to use



R Bioconductor

+ Feature rich IDE for programming in R
+ Rich statistics & ML and visualizations
- Limited support for other programming languages



{wdl}

+ Extremely scalable and flexible
- Most technically demanding
- Requires careful cost management

WDL 101: Reproducible genomics at scale

```
task samtoolsStats {
    input {
        File inputCram
        File cramIndex
        File targetRef
        String sampleName
    }

    command <<<
        samtools stats -r "~{targetRef}" \
            --reference "~{targetRef}" \
            -@ "$(nproc)" \
            "~{inputCram}" > "~{sampleName}.samtools.stats.txt"
    >>>

    Int diskGb = ceil(2.0 * size(inputCram, "G"))

    runtime {
        docker : "szarate/t2t_variants:v0.0.2"
        disks : "local-disk ${diskGb} SSD"
        memory: "12G"
        cpu : 16
        preemptible: 3
        maxRetries: 3
    }

    output {
        File stats = "~{sampleName}.samtools.stats.txt"
    }
}
```

What are the inputs?
Automatically copy from buckets to VMs

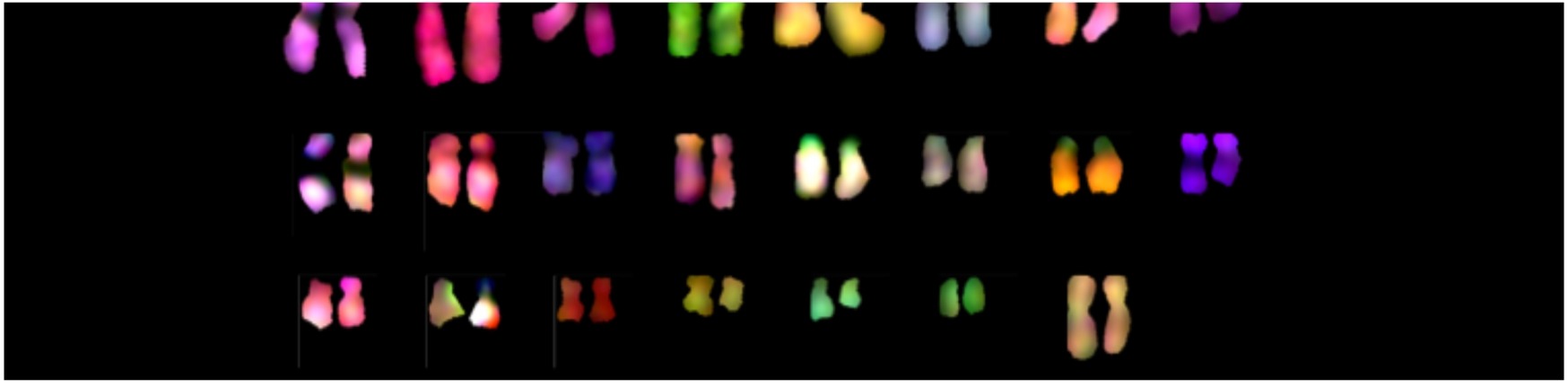
What should we run?
Essentially any command line tool!

What type of computers should we use?
Match resources (cores, RAM, disk) to needs

What are the outputs?
Automatically copy to buckets for next stage

Let's finish a human genome

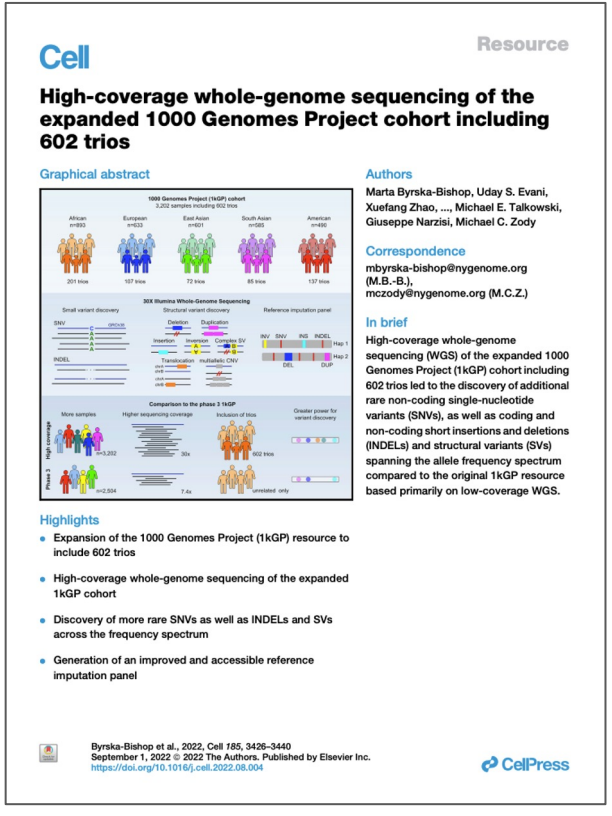

T2T Working Group Home · Technology · Data · CHM13 Cell Line · Remaining Challenges · Who We Are · Join Us



The Telomere-to-Telomere (T2T) consortium is an open, community-based effort to generate the first complete assembly of a human genome.

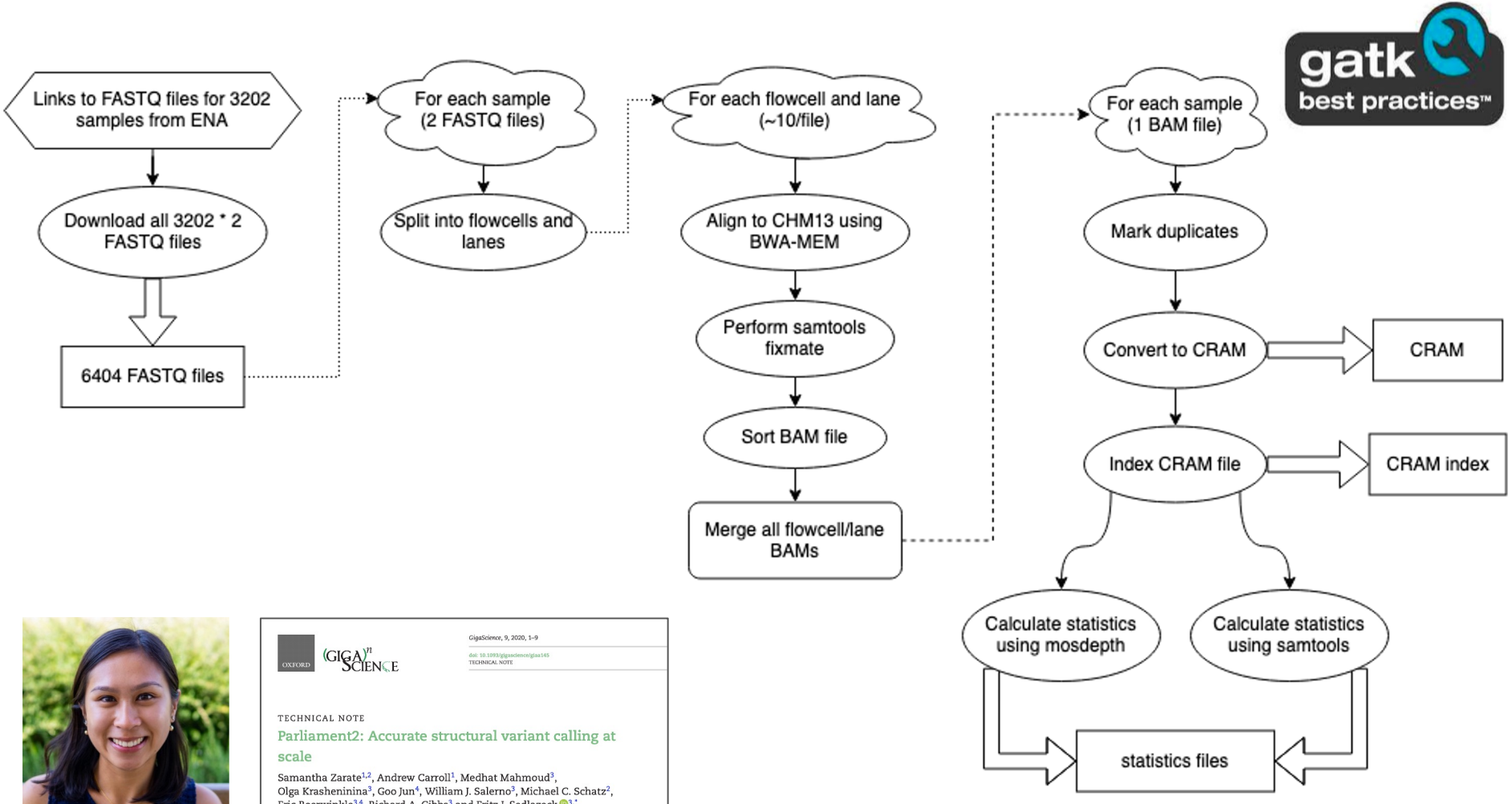
CHM13 homozygous 46,XX cell line from Urvashi Surti, Pitt; SKY karyotype from Jennifer Gerton, Stowers

T2T Powered by AnVIL!



3202 samples from 26 populations

3202 samples x 30Gb = 96Tb input data



Samantha Zarate

OXFORD (GIGA)n SCIENCE

GigaScience, 9, 2020, 1–9

doi: 10.1093/gigascience/giaa145
TECHNICAL NOTE

TECHNICAL NOTE

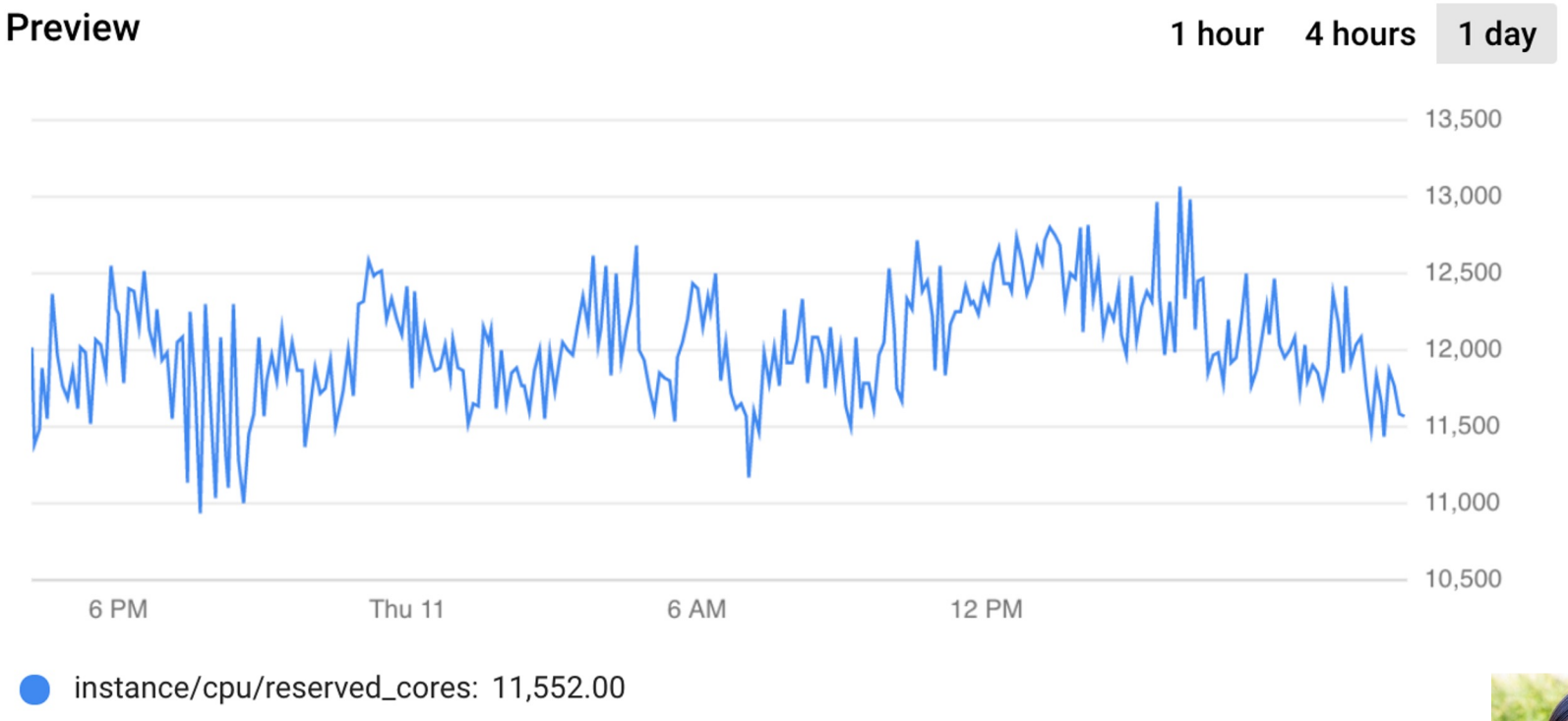
Parliament2: Accurate structural variant calling at scale

Samantha Zarate[1,2], Andrew Carroll[1], Medhat Mahmoud[3], Olga Krasheninina[3], Goo Jun[4], William J. Salerno[3], Michael C. Schatz[2], Eric Boerwinkle[3,4], Richard A. Gibbs[3] and Fritz J. Sedlazeck[3,*]

[1]DNAnexus, 1975 W El Camino Real #204, Mountain View, CA 94040, USA; [2]Department of Computer Science, 3400 N. Charles St. Johns Hopkins University, Baltimore, MD 21218, USA; [3]Human Genome Sequencing Center, One Baylor Plaza, Baylor College of Medicine, Houston, TX 77030, USA and [4]Human Genetics Center, 1200 Pressler Street, University of Texas Health Science Center at Houston, Houston, TX 77040, USA

*Correspondence address: Fritz J. Sedlazeck, Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA. E-mail: fritz.sedlazeck@bcm.edu http://orcid.org/0000-0001-6040-2691

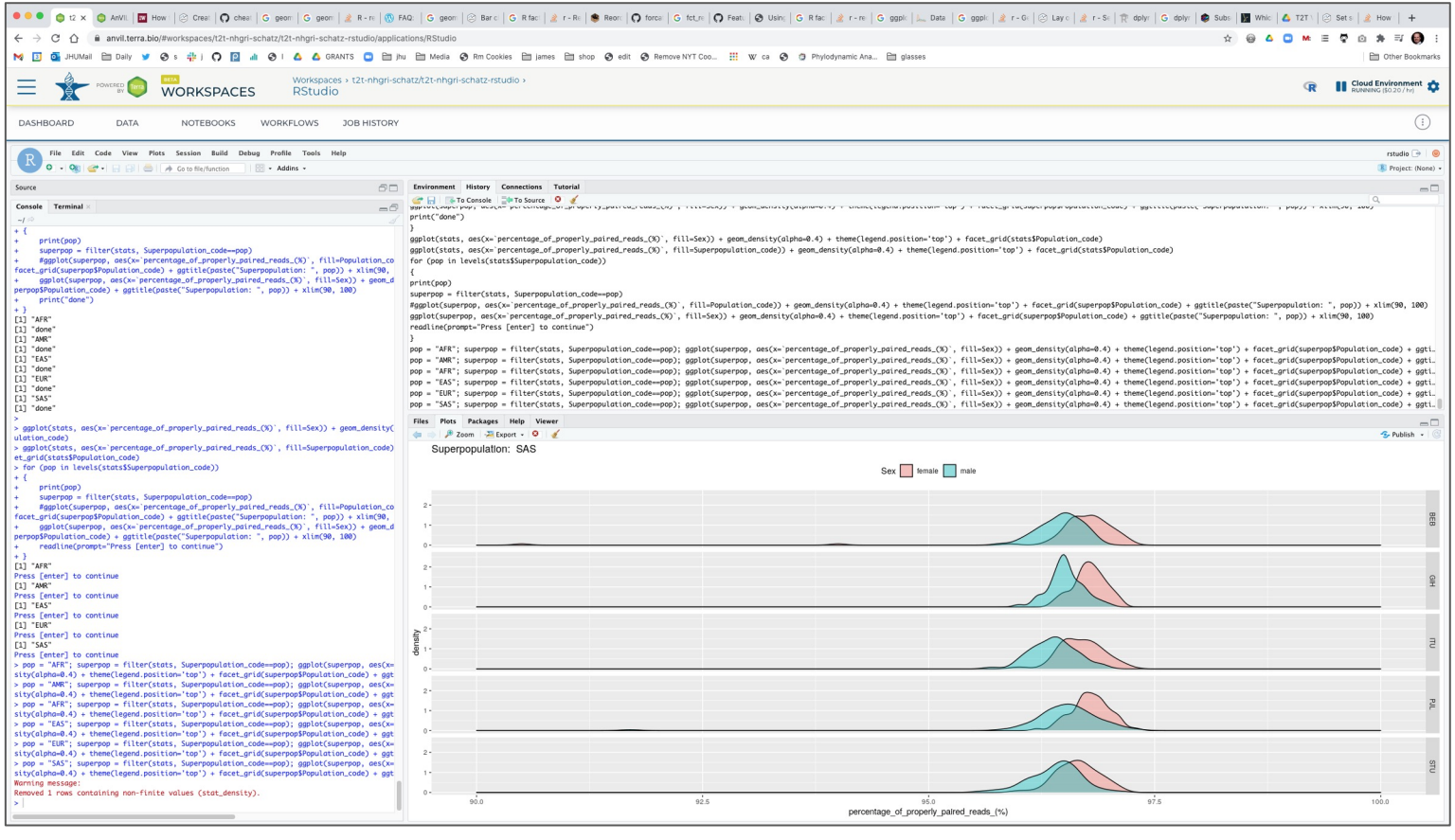
Core usage over 24 hours



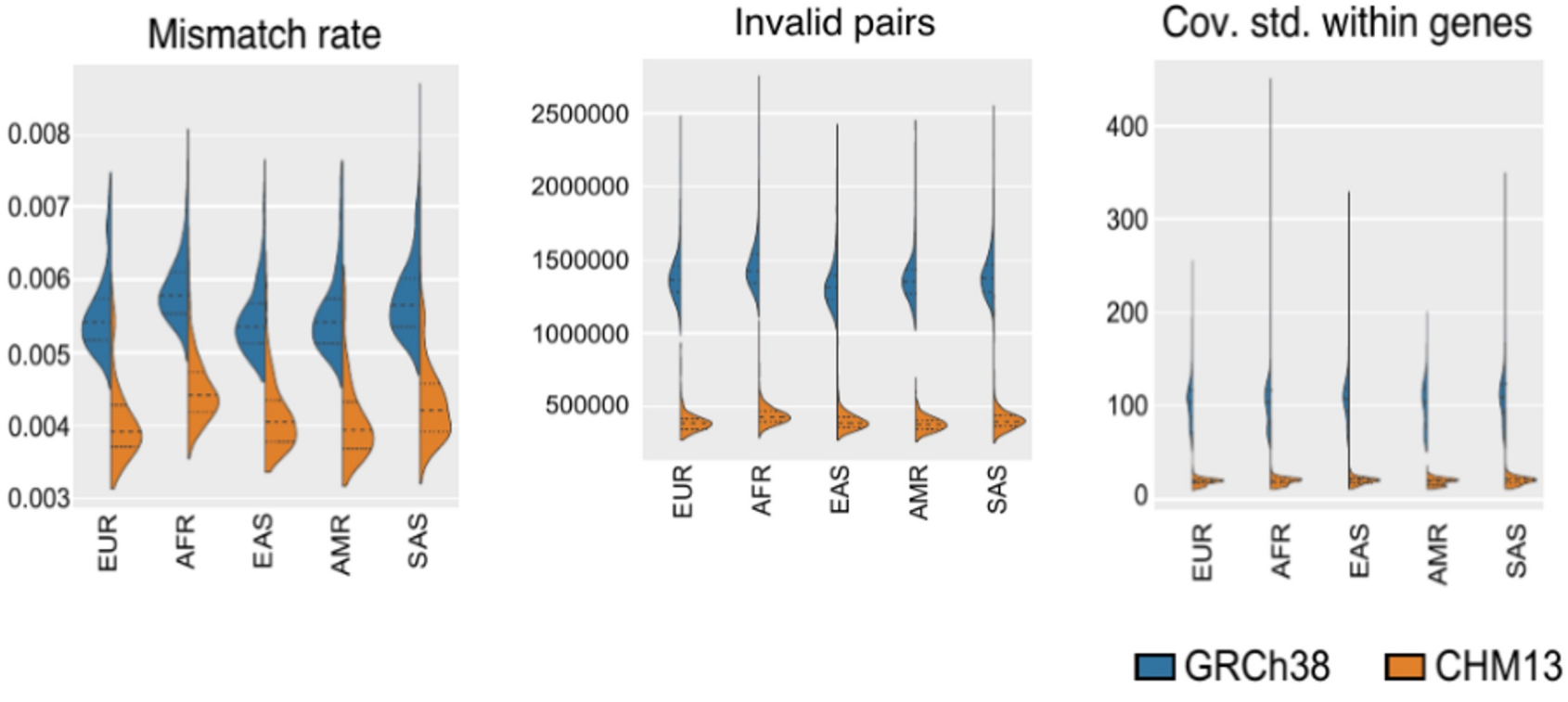
https://dockstore.org/workflows/github.com/schatzlab/t2t-variants/T2T_alignment

Samantha Zarate

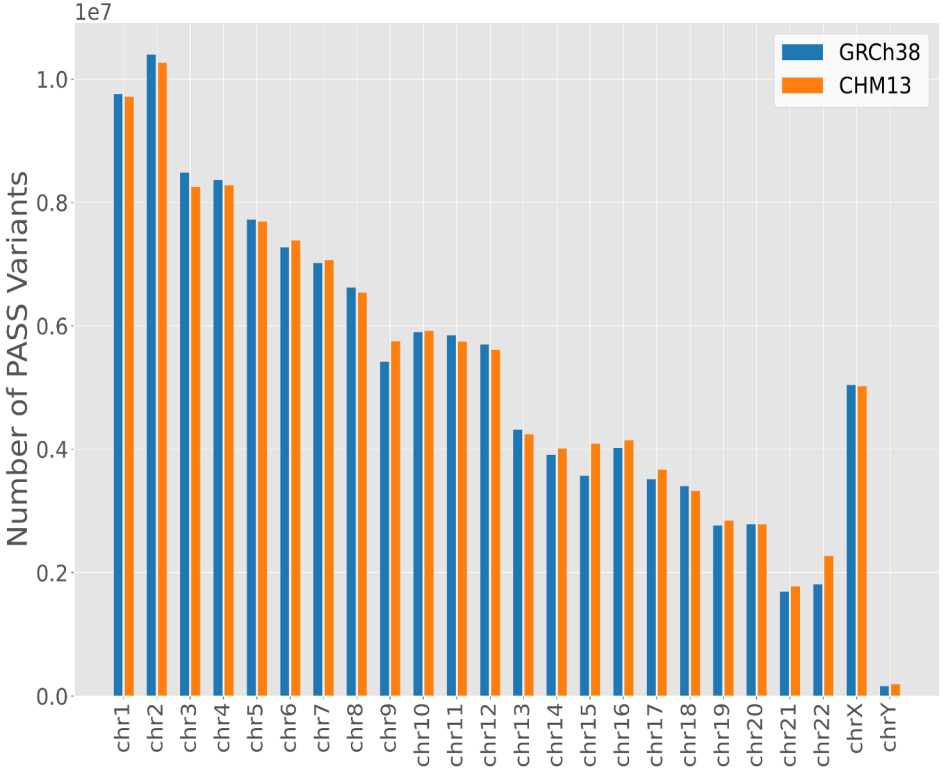
AnVIL: RStudio in the Cloud



1000G Mapping on T2T-CHM13

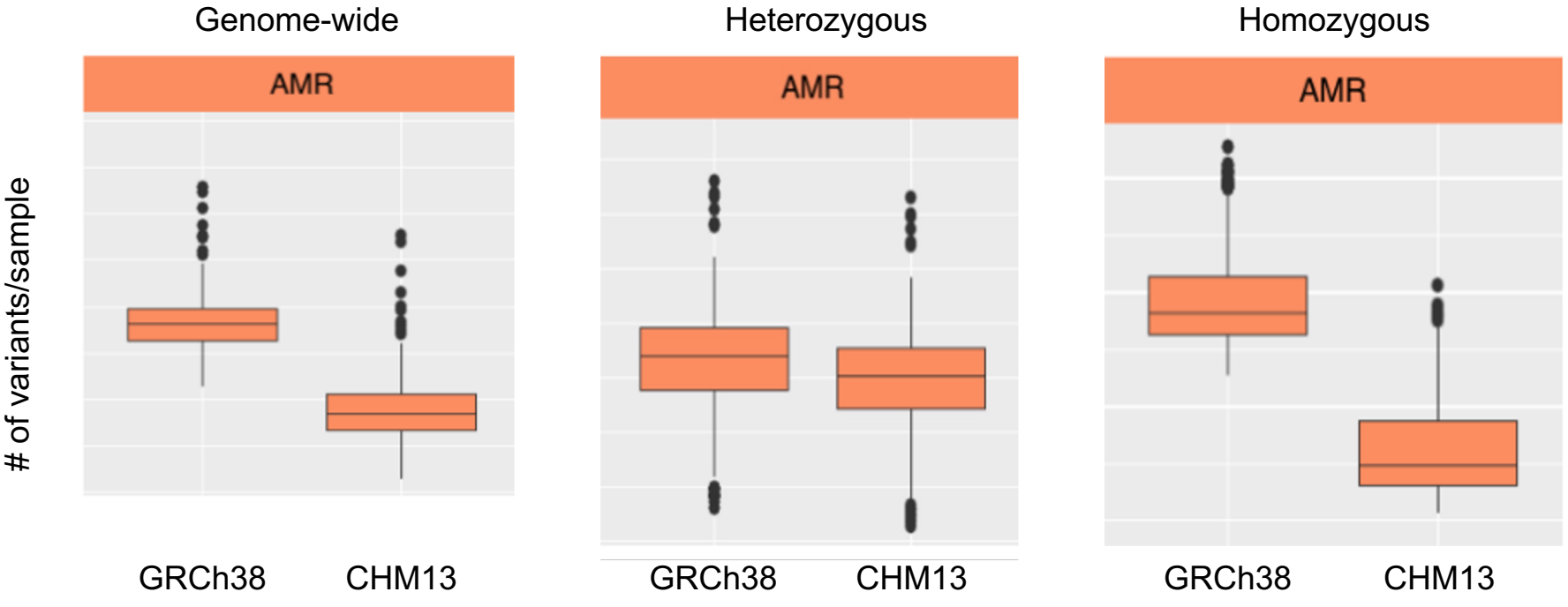


More Variants Found Using CHM13

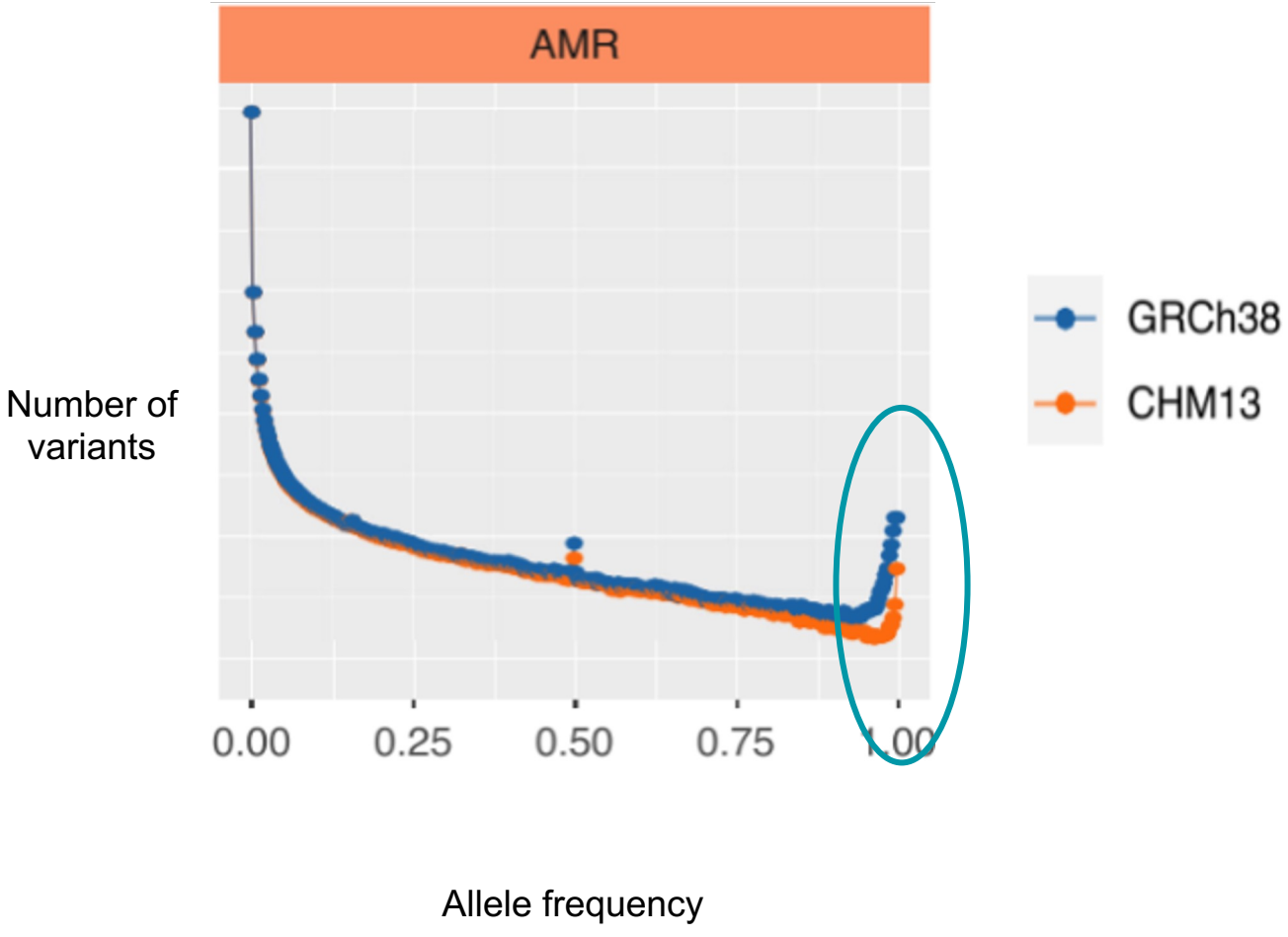


	GRCh38	CHM13
# of PASS variants	125,484,020	126,591,489

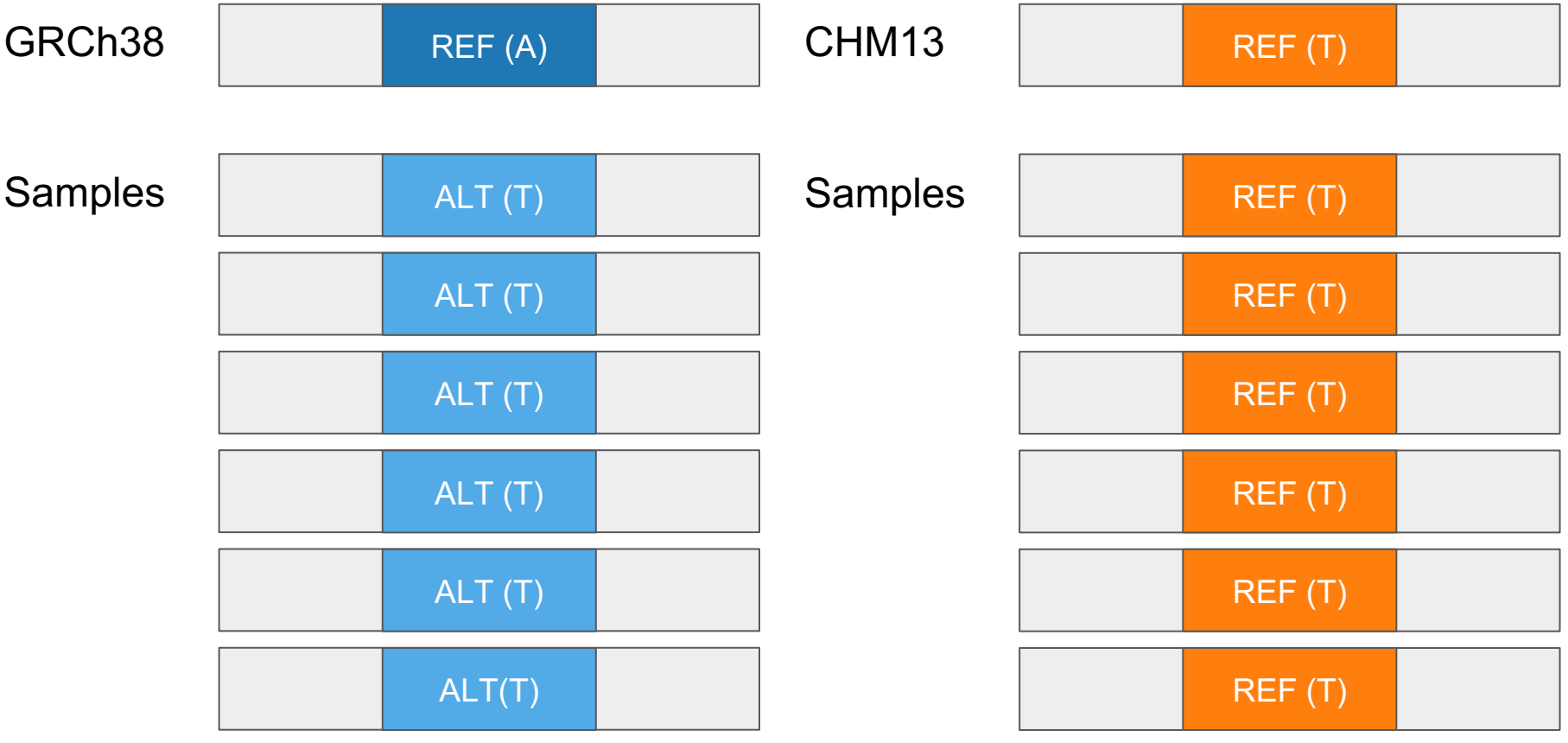
1000G Per-Sample Variant Counts on T2T-CHM13



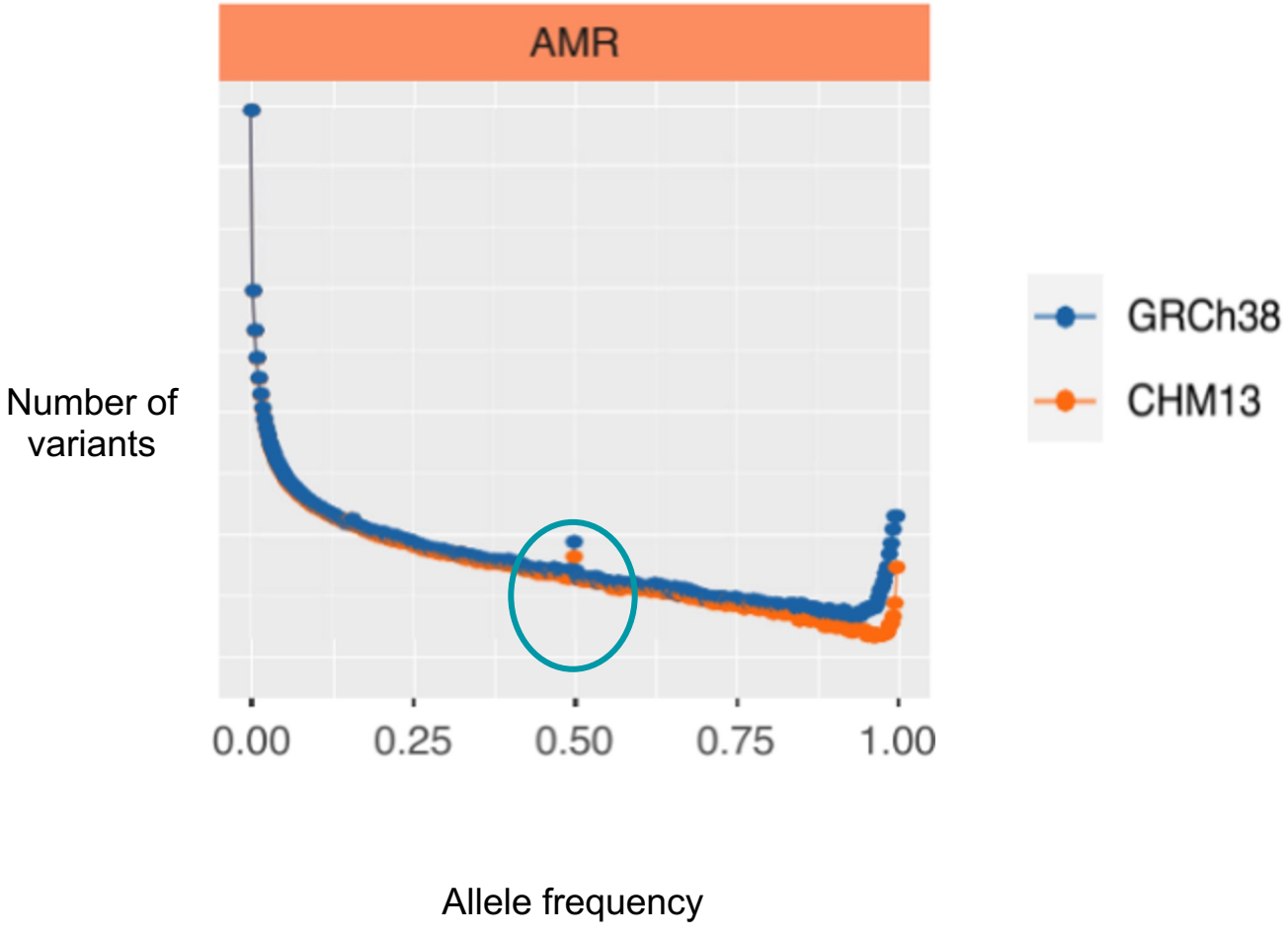
Explaining Decreased Per-Sample Count



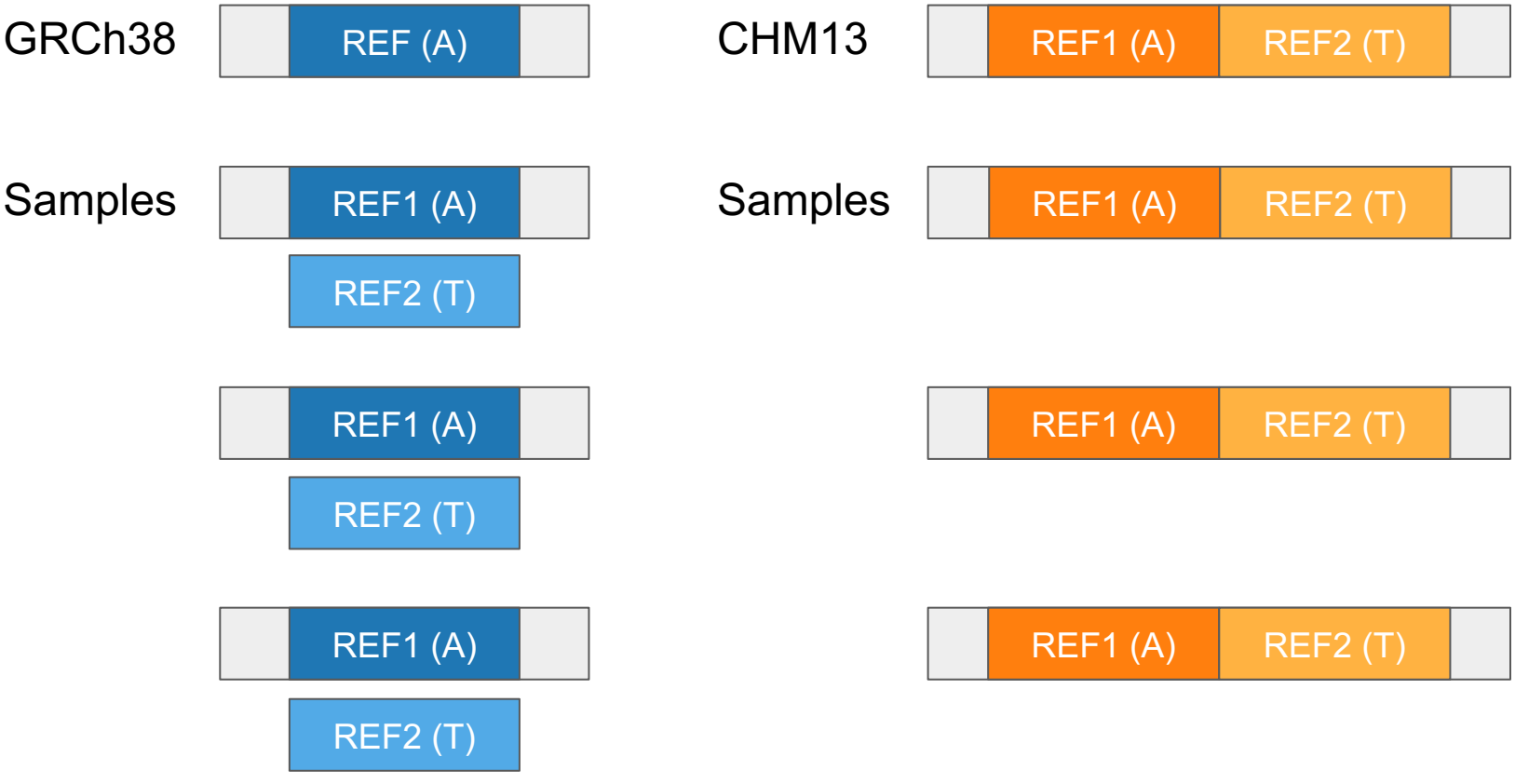
Allele Frequency = 1: Reference error / private variant



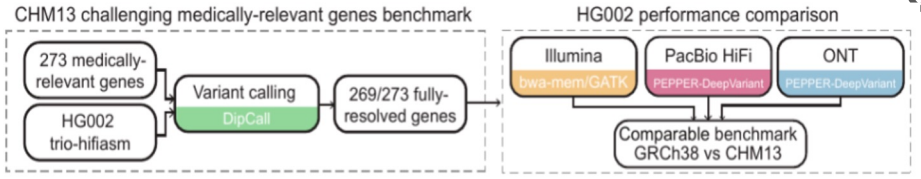
Explaining Decreased Per-Sample Count

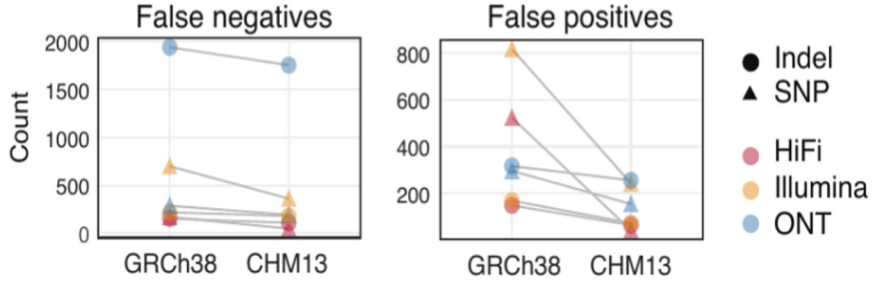


Allele Frequency ≈ 0.5: Collapsed duplication



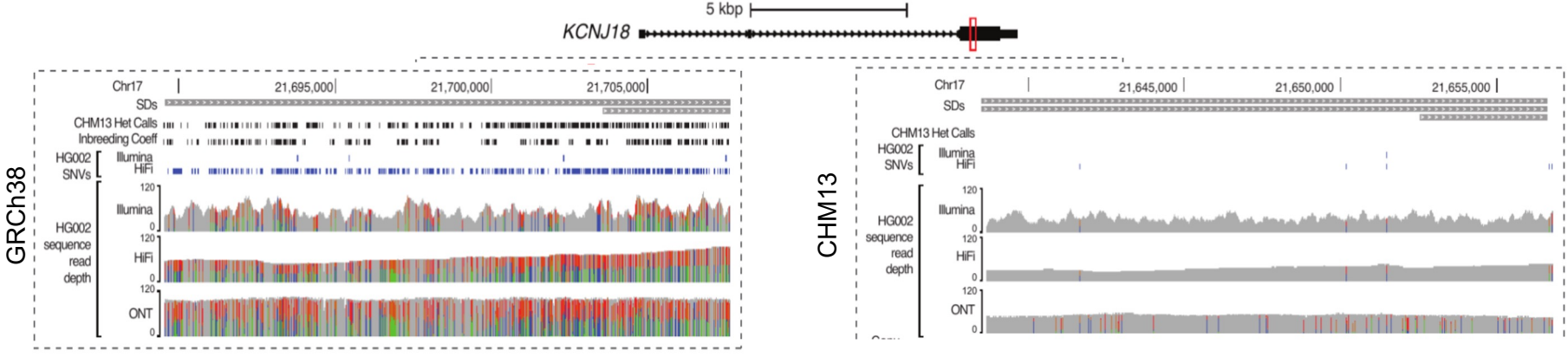
Improved Variant Calling in Clinical Genes





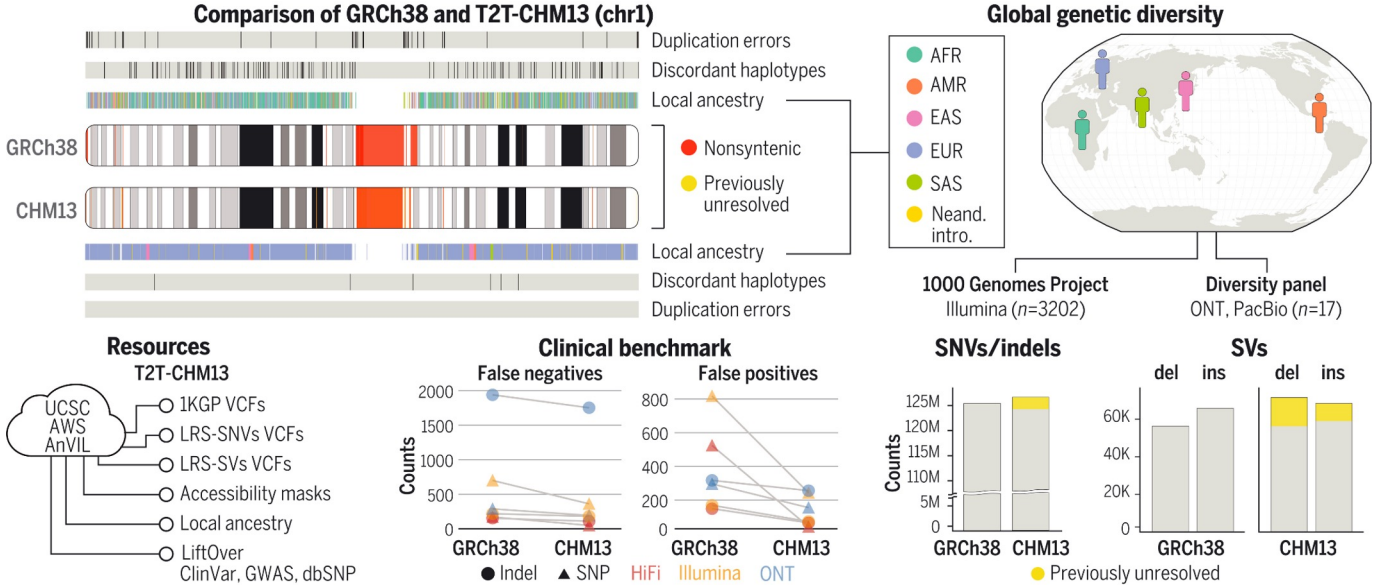
Justin Zook

Fritz Sedlazeck



- *KCNJ18*: periodic muscle paralysis
- Collapsed duplication, excessive variant calls in GRCh38

T2T-Variants Analysis Summary




Sergey Aganezov Stephanie Yan

Daniela Soto


Melanie Kirsche Samantha Zarate

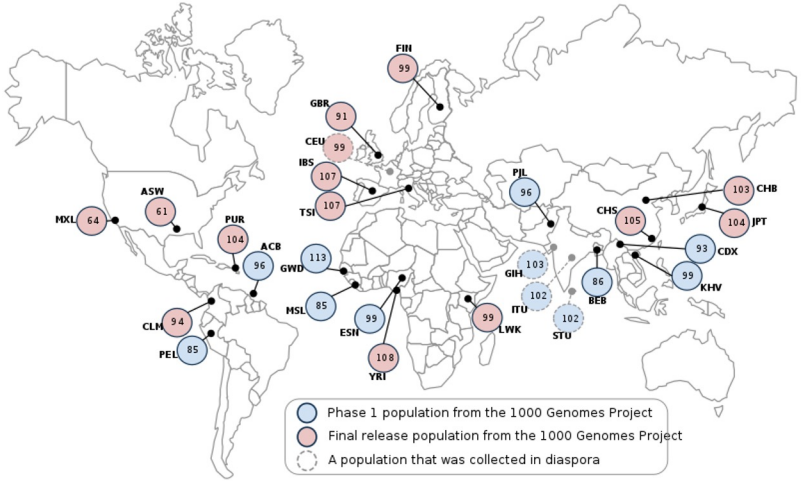
A complete reference genome improves analysis of human genetic variation

Aganezov, S*, Yan, SM*, Soto, DC*, Kirsche, M*, Zarate, S*, *et al.* (2022) *Science*. doi: 10.1126/science.abl3533

T2T-chrY: Human variation across 156 populations

1000 Genomes Project (1KGP)

3,202 samples from 26 populations



(Byrska-Bishop et al., Cell, 2022)

Simons Genome Diversity Project (SGDP)

279 open access samples from 130 populations

(Mallick et al., Nature, 2016)

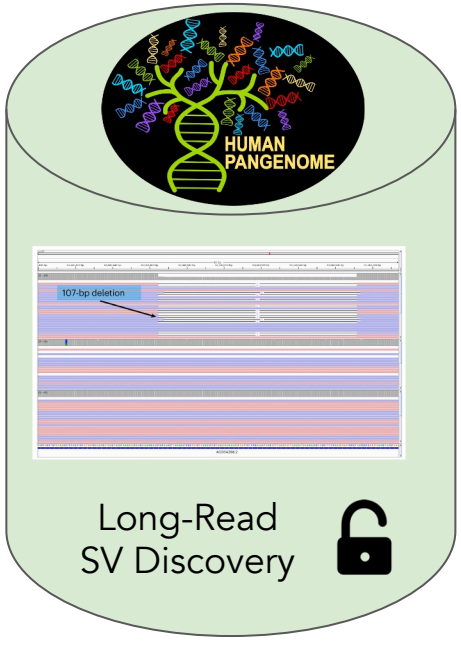
The complete sequence of a human Y chromosome

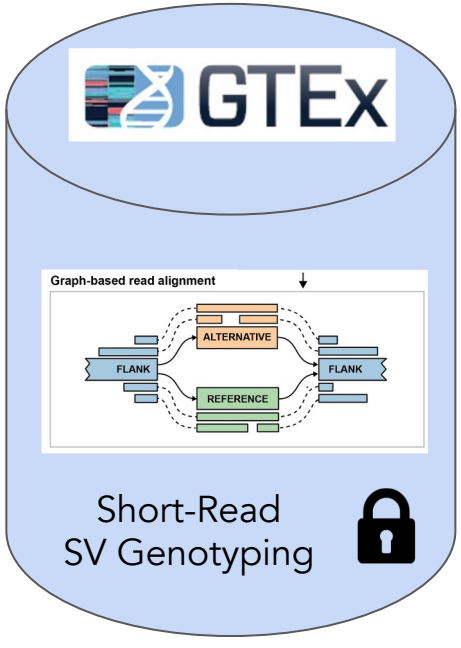
Rhie *et al.* (2023) *Nature.* https://doi.org/10.1038/s41586-023-06457-y

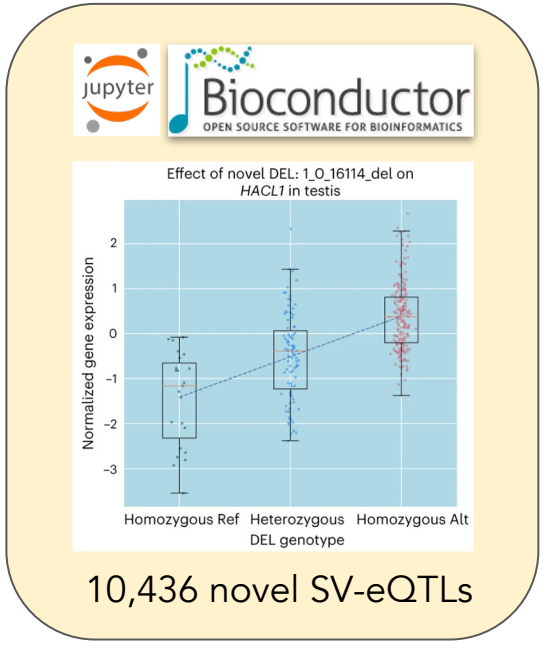
Stephen Hwang

Dylan Taylor

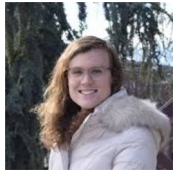
Data Integration and Reuse





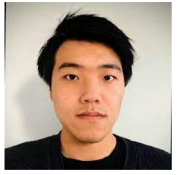


Jasmine and Iris: population-scale structural variant comparison and analysis
Kirsche *et al.* (2023) *Nature Methods*. https://doi.org/10.1038/s41592-022-01753-3



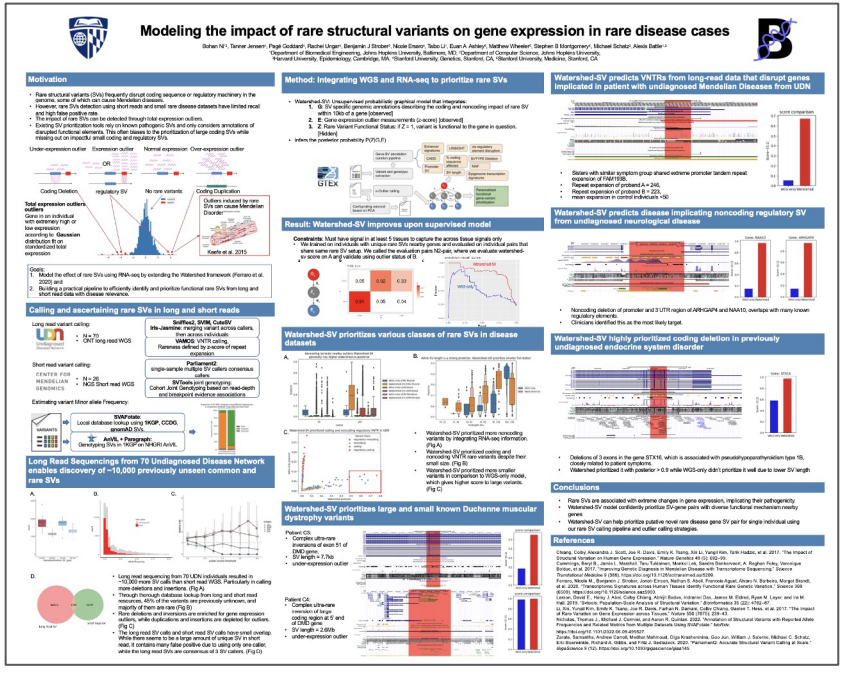
Melanie Kirsche

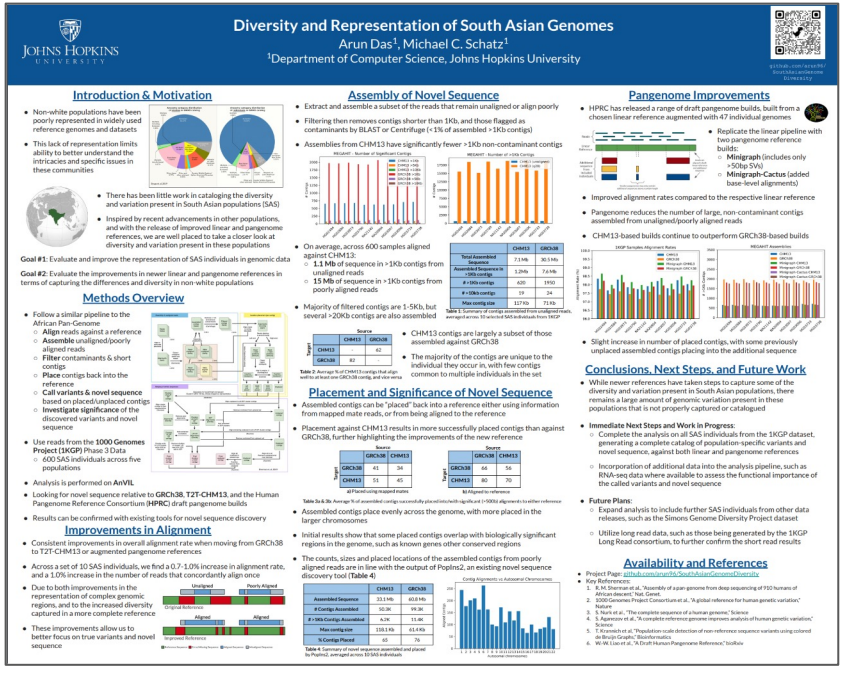
Ongoing Projects



Bohan Ni
Rare Variant Analysis

Arun Das
South Asian Pan-Genome





Diversifying Genomics

Perspective

Diversifying the genomic data science research community

The Genomic Data Science Community Network[1]

Over the past 20 years, the explosion of genomic data collection and the cloud computing revolution have made computational and data science research accessible to anyone with a web browser and an internet connection. However, students at institutions with limited resources have received relatively little exposure to curricula or professional development opportunities that lead to careers in genomic data science. To broaden participation in genomics research, the scientific community needs to support these programs in local education and research at underserved institutions (UIs). These include community colleges, historically Black colleges and universities, Hispanic-serving institutions, and tribal colleges and universities that support ethnically, racially, and socioeconomically underrepresented students in the United States. We have formed the Genomic Data Science Community Network to support students, faculty, and their networks to identify opportunities and broaden access to genomic data science. These opportunities include expanding access to infrastructure and data, providing UI faculty development opportunities, strengthening collaborations among faculty, recognizing UI teaching and research excellence, fostering student awareness, developing modular and open-source resources, expanding course-based undergraduate research experiences (CUREs), building curriculum, supporting student professional development and research, and removing financial barriers through funding programs and collaborator support.

[Supplemental material is available for this article.]

Foundations for justice in genomic data science

Despite growing opportunities in data science careers, systemic barriers have limited the participation of underrepresented groups in genomic data science research and education (Canner et al. 2017). Among bachelor's degree recipients in biological sciences, computer sciences, mathematics, and statistics from 2006–2016, 8.7% were Hispanic or Latinx, 7.8% were Black or African American, and 1.9% were multiracial and/or indigenous American (National Science Foundation 2019a). Meanwhile, these groups represent 16.3%, 12.3%, and 2.5% of the US resident population, respectively (National Science Foundation 2019a). Disparities are more pronounced in graduate education (Wiley et al. 2020). Affinity organizations in which members of underrepresented groups come together are vital to developing a sense of belonging and support system (Supplemental Table S1). However, for true representation in research, science needs inclusive spaces where researchers can communicate actively with educators and where students are supported in developing science, technology, engineering, and mathematics (STEM) identities.

The technological advancements of high-throughput sequencing in the past two decades have enabled the rapid proliferation of genomic data (Goodwin et al. 2016) but they have also led to an even greater access imbalance. Over 60 petabases of data (National Center for Biotechnology Information 2021), or about a million times the size of the original human genome project (International Human Genome Sequencing Consortium 2001), is currently available within the US National Center for Biotechnology Information (NCBI) genomic sequencing repositories. This wealth of data will help scientists determine disease risk, diagnose rare conditions, improve drug safety and efficacy (Manolio et al. 2019), survey pathogens for public health applications (Khoury et al. 2020), and even combat the effects of climate change (Hoffmann et al. 2021). Our greatest limitation is personnel to interpret these data. Yet, genomic data science currently lacks a scaffolded mechanism that supports all individuals and provides a hub of intellectual capital, curated genomic data, and the infrastructure required for authentic learning gained through research experiences. Broader, more diverse participation should be the starting point for creating a more inclusive genomic data science field (Mapes et al. 2020). Focusing on participation is not only ethical but desirable for more novel solutions to problems (Hofstra et al. 2020) and is necessary for bringing different perspectives to the table (Zook et al. 2017).

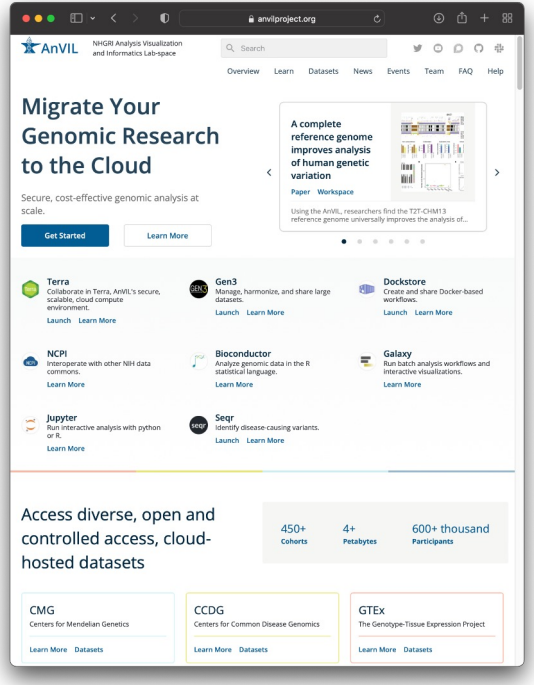
Our vision for a diverse scientific community engaged in genomic data science research is one in which researchers, educators, and students thrive in a just and fair system, not limited by their institution's scientific clout, resources, geographical location, or infrastructure (Fig. 1). Here, we focus on traditionally underserved institutions (UIs) in the United States, which include minority serving institutions (MSIs) defined by the US Department of Education: historically Black colleges and universities (HBCUs), Hispanic-serving institutions (HSIs), and tribal colleges and universities (TCUs) (Li and Carroll 2007). UIs also include community colleges (CCs) and some primarily undergraduate institutions that overlap substantially with MSIs (Nguyen et al. 2015). Collectively, UIs play a critical role in educating ethnically, racially, and socioeconomically underrepresented students despite limited access to resources (Li and Carroll 2007). In addition to the number of traditionally underrepresented students educated at UIs, these colleges and universities possess unique strengths, such as the greater sense of belonging, more positive mentoring relationships, and

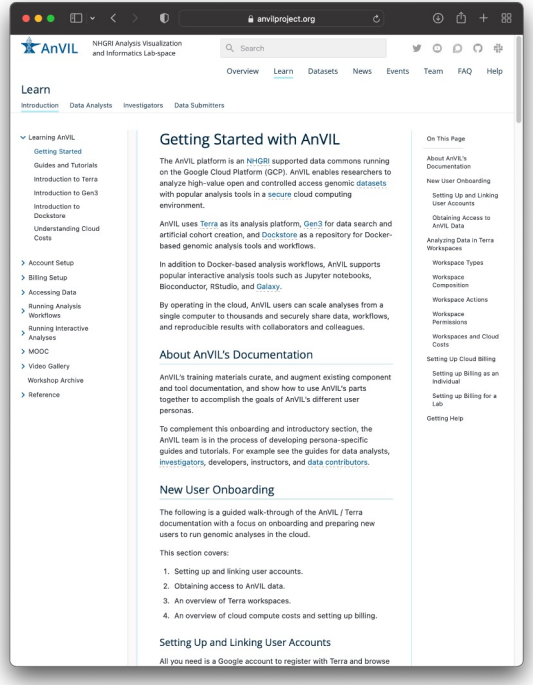
[1]A complete list of GDSCN participants appears at the end of this paper.

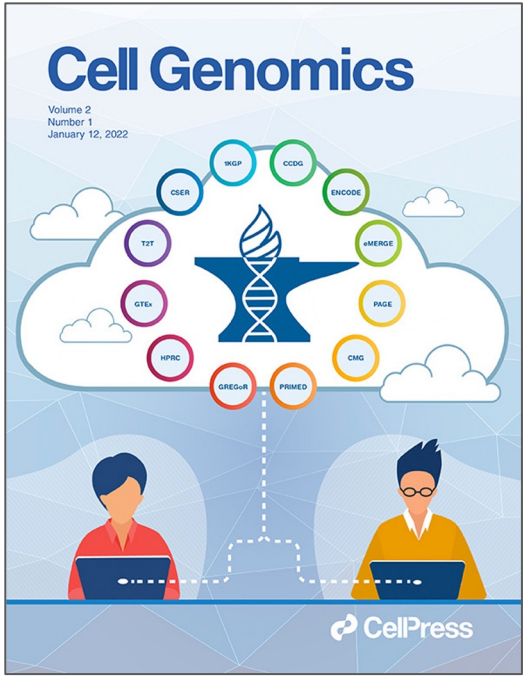
Corresponding authors: rosa.alcazar@cloviscollege.edu, ava.hoffman@jhu.edu, jslee@fortlewis.edu, sroy1@utep.edu

Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.276496.121. Freely available online through the *Genome Research* Open Access option.

Learning More!







Portal: https://anvilproject.org/
Discourse: https://help.anvilproject.org/
Office Hours & Events: https://anvilproject.org/events
Paper: Schatz, Phillippakis *et al.* (2022) *Cell Genomics.* doi: 10.1016/j.xgen.2021.100085

AnVIL Team

NIH National Human Genome Research Institute

Johns Hopkins University
Michael Schatz, Kasper Hansen, Enis Afgan, *Alex Ostrovsky, John Davis, Jenn Vessio, John Muschelli, Stephen Mosher, Natalie Kucher, Dannon Baker, Aysam Guerler, Katie Cox, Benjamin Harvey, Kai Hammers, Alex Ostrovsky, Keith Suderman, Ahmed Awan, Michelle Savage, Tyler Collins, Samantha Zarate, Bohan Ni*

Fred Hutchinson Cancer Center
Jeff Leek, Ava Hoffman, *Elizabeth Humphries*

Penn State University
Anton Nekrutenko, *John Chilton, Nate Coraor, Marten Cech, Emil Bouvier, Nicholas Stoler, Jennifer Jackson, Assunta Desanto, Delphine Lariviere*

Oregon Health & Sciences University
Jeremy Goecks, Kyle Ellrott, *Brian Walsh, Luke Sargent, Vahid Jalili, Qiange Gu*

Roswell Park Cancer Institute
Martin Morgan, *Jiefei Wang, Lori Kern, Kayla Interdonato*

Harvard Medical School
Vincent Carey, *Alexandru Mahmoud, Shweta Gopaulakrishnan, BJ Stubbs*

Broad Institute
Anthony Philippakis, *Rachel Liao, Kate Balaconis, Alex Baumann, Adrian Sharma, David Bernick, Jonathan Lawson, Kristian Cibulskis, Namrata Gupta, Rob Title, Eric Banks, Alessandro Culotti*

University of Chicago
Robert Grossman, *Radhika Reddy, Alex Van Tol, Fantix King*

University of California Santa Cruz
Benedict Paten, *Ben Vizzier, Hannes Schmidt, Dave Rogers, Nneka Denis Yuen, Charles Overbeck, Louise Cabansay, Natalie Perez, Ash O'Farrell, Walt Shands*

Vanderbilt University
Robert Carroll, *Lakhan Swamy, Katie Banasiewicz*

Washington University
Adam Coffman, Allison Reieir, Haley Abel, Jason Walker

Carnegie Institution for Science
Frederick Tan

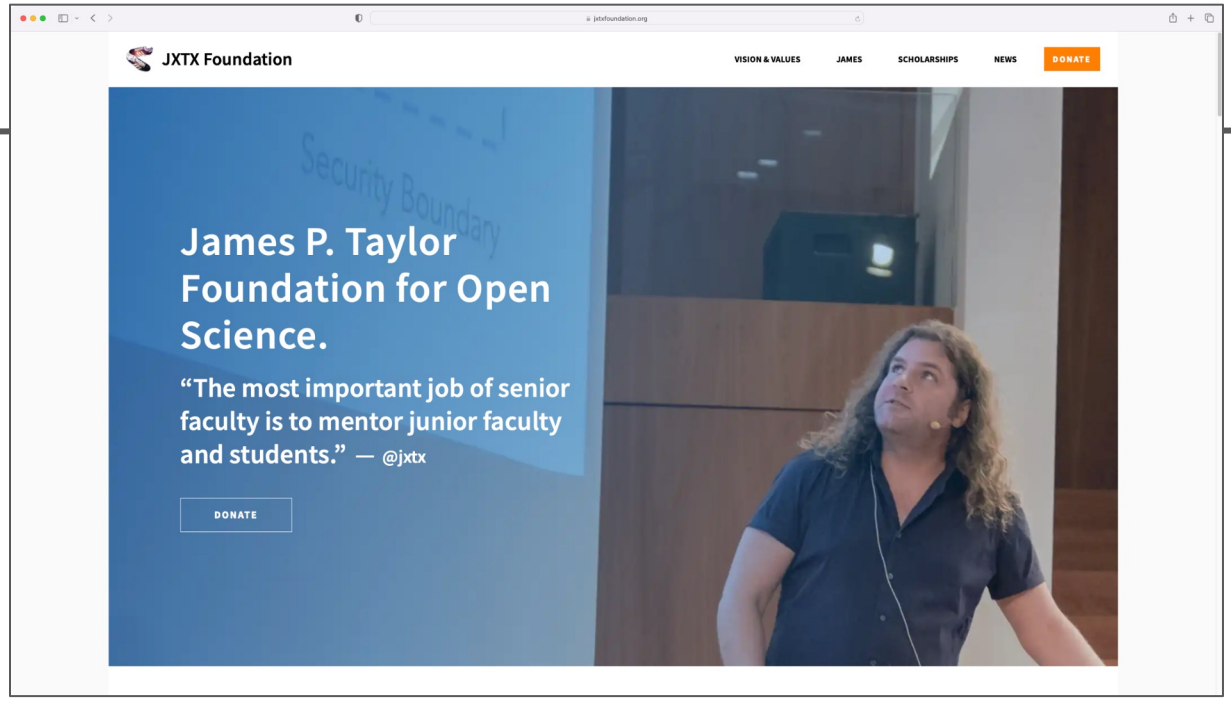
City University of New York
Levi Waldron, *Sehyun Oh, Ludwig Geistlinger*

Galaxy

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Dockstore

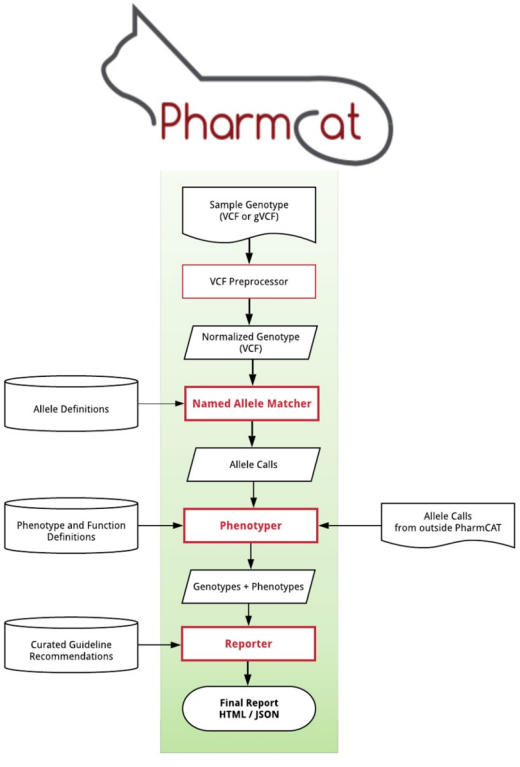
Terra

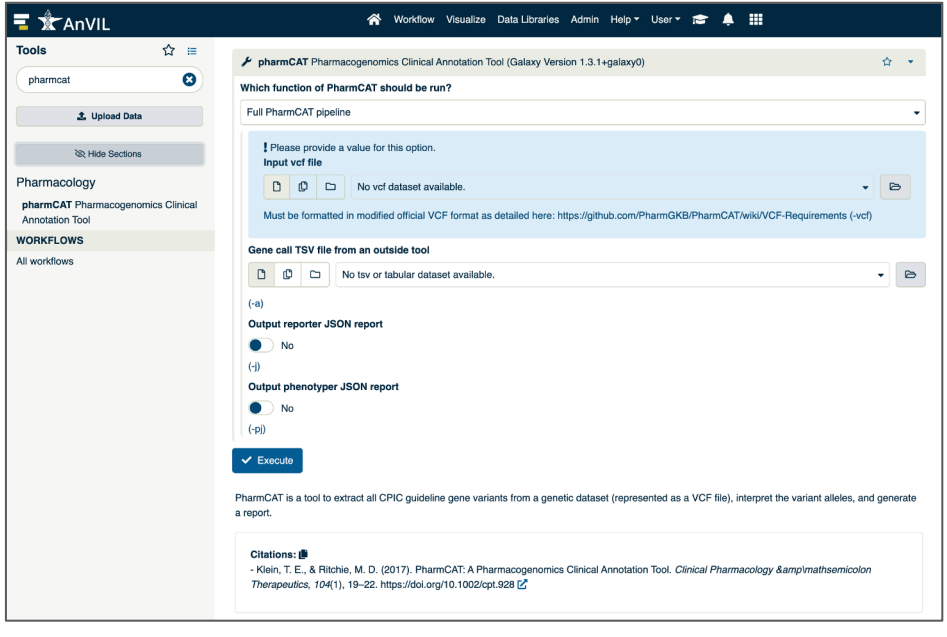


Thank you!

http://schatz-lab.org

PharmCAT: Pharmacogenomics Clinical Annotation Tool





PharmCAT: A Pharmacogenomics Clinical Annotation Tool
TE Klein, MD Ritchie. (2018) Clinical Pharmacology & Therapeutics 104(1):19-22.

Diversifying Genomics

Perspective

Diversifying the genomic data science research community

The Genomic Data Science Community Network[1]

Over the past 20 years, the explosion of genomic data collection and the cloud computing revolution have made computational and data science research accessible to anyone with a web browser and an internet connection. However, students at institutions with limited resources have received relatively little exposure to curricula or professional development opportunities that lead to careers in genomic data science. To broaden participation in genomics research, the scientific community needs to support these programs in local education and research at underserved institutions (UIs). These include community colleges, historically Black colleges and universities, Hispanic-serving institutions, and tribal colleges and universities that support ethnically, racially, and socioeconomically underrepresented students in the United States. We have formed the Genomic Data Science Community Network to support students, faculty, and their networks to identify opportunities and broaden access to genomic data science. These opportunities include expanding access to infrastructure and data, providing UI faculty development opportunities, strengthening collaborations among faculty, recognizing UI teaching and research excellence, fostering student awareness, developing modular and open-source resources, expanding course-based undergraduate research experiences (CUREs), building curriculum, supporting student professional development and research, and removing financial barriers through funding programs and collaborator support.

[Supplemental material is available for this article.]

Foundations for justice in genomic data science

Despite growing opportunities in data science careers, systemic barriers have limited the participation of underrepresented groups in genomic data science research and education (Canner et al. 2017). Among bachelor's degree recipients in biological sciences, computer sciences, mathematics, and statistics from 2006–2016, 8.7% were Hispanic or Latinx, 7.8% were Black or African American, and 1.9% were multiracial and/or indigenous American (National Science Foundation 2019a). Meanwhile, these groups represent 16.3%, 12.3%, and 2.5% of the US resident population, respectively (National Science Foundation 2019a). Disparities are more pronounced in graduate education (Wiley et al. 2020). Affinity organizations in which members of underrepresented groups come together are vital to developing a sense of belonging and support system (Supplemental Table S1). However, for true representation in research, science needs inclusive spaces where researchers can communicate actively with educators and where students are supported in developing science, technology, engineering, and mathematics (STEM) identities.

The technological advancements of high-throughput sequencing in the past two decades have enabled the rapid proliferation of genomic data (Goodwin et al. 2016) but they have also led to an even greater access imbalance. Over 60 petabases of data (National Center for Biotechnology Information 2021), or about a million times the size of the original human genome project (International Human Genome Sequencing Consortium 2001), is currently available within the US National Center for Biotechnology Information (NCBI) genomic sequencing repositories. This wealth of data will help scientists determine disease risk, diagnose rare conditions, improve drug safety and efficacy (Manolio et al. 2019), survey pathogens for public health applications (Khoury et al. 2020), and even combat the effects of climate change (Hoffmann et al. 2021). Our greatest limitation is personnel to interpret these data. Yet, genomic data science currently lacks a scaffolded mechanism that supports all individuals and provides a hub of intellectual capital, curated genomic data, and the infrastructure required for authentic learning gained through research experiences. Broader, more diverse participation should be the starting point for creating a more inclusive genomic data science field (Mapes et al. 2020). Focusing on participation is not only ethical but desirable for more novel solutions to problems (Hofstra et al. 2020) and is necessary for bringing different perspectives to the table (Zook et al. 2017).

Our vision for a diverse scientific community engaged in genomic data science research is one in which researchers, educators, and students thrive in a just and fair system, not limited by their institution's scientific clout, resources, geographical location, or infrastructure (Fig. 1). Here, we focus on traditionally underserved institutions (UIs) in the United States, which include minority serving institutions (MSIs) defined by the US Department of Education: historically Black colleges and universities (HBCUs), Hispanic-serving institutions (HSIs), and tribal colleges and universities (TCUs) (Li and Carroll 2007). UIs also include community colleges (CCs) and some primarily undergraduate institutions that overlap substantially with MSIs (Nguyen et al. 2015). Collectively, UIs play a critical role in educating ethnically, racially, and socioeconomically underrepresented students despite limited access to resources (Li and Carroll 2007). In addition to the number of traditionally underrepresented students educated at UIs, these colleges and universities possess unique strengths, such as the greater sense of belonging, more positive mentoring relationships, and

[1]A complete list of GDSCN participants appears at the end of this paper.
Corresponding authors: rosa.alcazar@cloviscollege.edu, ava.hoffman@jhu.edu, jslee@fortlewis.edu, sroy1@utep.edu
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.276496.121. Freely available online through the *Genome Research* Open Access option.

WANTED—Tools and infrastructure to broaden the path into genomics

Rosa Alcazar, Ph.D.
Department of Biology, Clovis Community College
CSHL Biological Data Science Conference
November 9-12, 2022