



Genomics at scale with the NHGRI AnVIL

(Analysis, Visualization, and Informatics Lab-space)

Michael Schatz
November 7, 2023



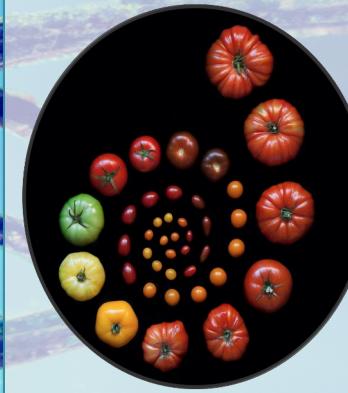
@mike_schatz

Schatzlab Overview



Human Disease Genetics

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



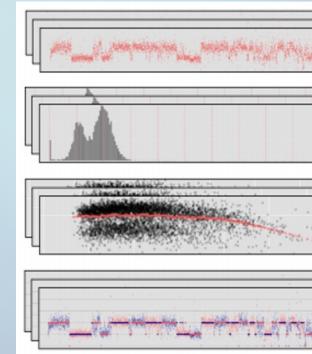
Agricultural Genomics

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



Algorithmics & Systems Research

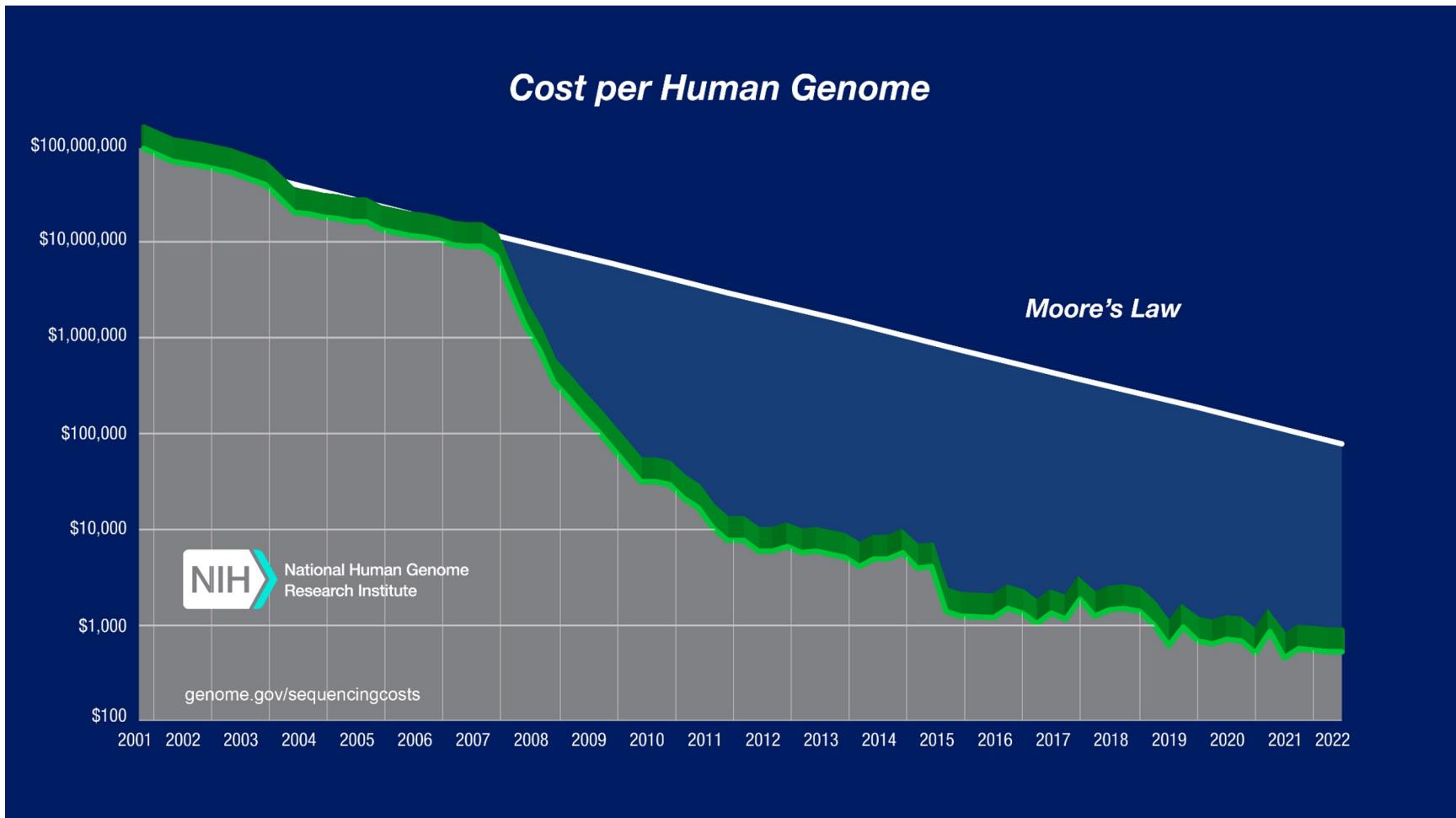
Kirsche *et al.* (2023)
Schatz, Philippakis *et al.* (2021)



Single Cell & Single Molecule

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

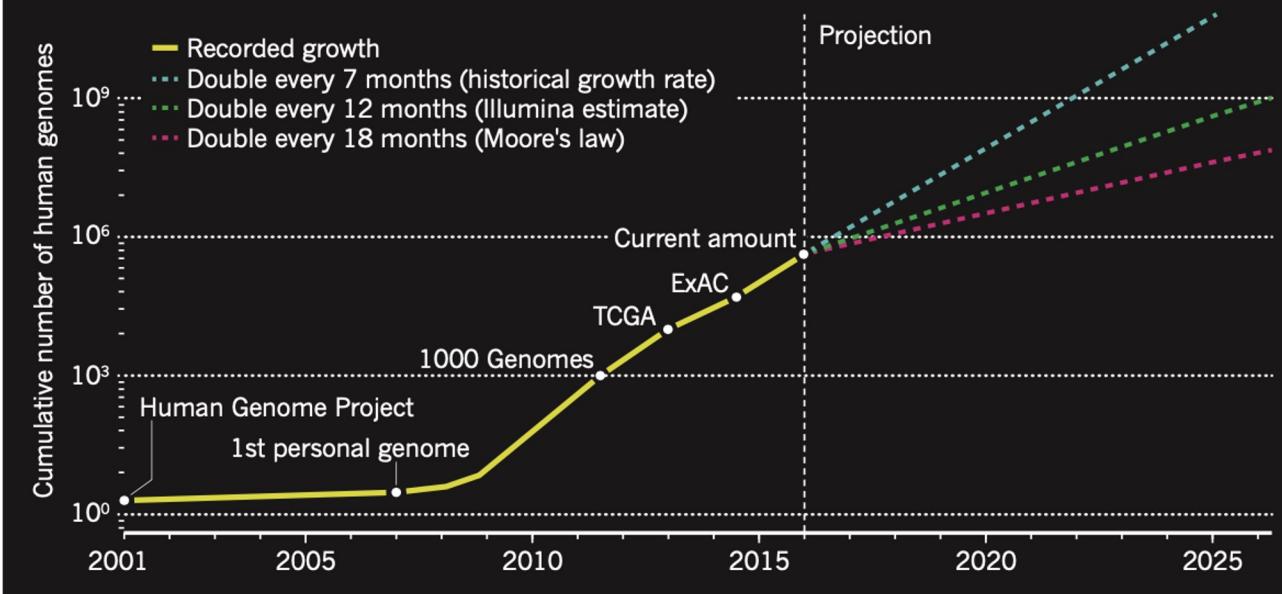
Growth of Genomics



Growth of Genomics

DNA SEQUENCING SOARS

Human genomes are being sequenced at an ever-increasing rate. The 1000 Genomes Project has aggregated hundreds of genomes; The Cancer Genome Atlas (TCGA) has gathered several thousand; and the Exome Aggregation Consortium (ExAC) has sequenced more than 60,000 exomes. Dotted lines show three possible future growth curves.



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



Both currently ~1Eb
And growing exponentially

Growth of Genomics

The screenshot shows a Science journal page. At the top, there are navigation links: Current Issue, First release papers, Archive, About, and a 'Submit man' button. Below this is a breadcrumb trail: HOME > SCIENCE > VOL. 369, NO. 6509 > THE GTEx CONSORTIUM ATLAS OF GENETIC REGULATORY EFFECTS ACROSS HUMAN TISSUES. A 'SPECIAL ISSUE RESEARCH ARTICLE' is highlighted. The main title is 'The GTEx Consortium atlas of genetic regulatory effects across human tissues'. Below the title is a subheader 'THE GTEx CONSORTIUM'. The text discusses how some human genetic variants affect RNA production and splicing. It mentions the Genotype-Tissue Expression (GTEx) project, which has expanded over time to include 838 individuals across 49 tissues. The text ends with 'This large study was...'. There are social media sharing icons (Facebook, Twitter, LinkedIn, etc.) and a sidebar with various buttons.

The screenshot shows a Nature journal page. At the top, there are navigation links: Explore content, About the journal, and Publish with us. Below this is a breadcrumb trail: nature > articles > article. The article title is 'The mutational constraint spectrum quantified from variation in 141,456 humans'. The authors listed are Konrad J. Karczewski, Laurent C. Francioli, Daniel G. MacArthur, and others. Below the authors is a link to 'Nature 581, 434–443 (2020) | Cite this article'. At the bottom of the article summary are metrics: 102k Accesses, 1878 Citations, 359 Altmetric, and Metrics.

The screenshot shows a Nature journal page. At the top, there are navigation links: Explore content, About the journal, and Publish with us. Below this is a breadcrumb trail: nature > articles > article. The article title is 'Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program'. The authors listed are Daniel Taliun, Daniel N. Harris, Gonçalo R. Abecasis, and others. Below the authors is a link to 'Nature 590, 290–299 (2021) | Cite this article'. At the bottom of the article summary are metrics: 37k Accesses, 162 Citations, 350 Altmetric, and Metrics.

The screenshot shows a Cell journal page. At the top, there is a logo for 'Cell' and a 'Supports open access' button. To the right are links for Submit, Log in, and Register. Below this is a blue header bar with the text 'ARTICLE | VOLUME 173, ISSUE 2, P291-304.E6, APRIL 05, 2018'. On the right side of the header are PDF and Figure links. The main title of the article is 'Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer'. The authors listed are Katherine A. Hoadley, Christina Yau, Toshinori Hinoue, Joshua M. Stuart, Christopher C. Benz, Peter W. Laird, and others. Below the authors is a link to 'Open Access DOI: https://doi.org/10.1016/j.cell.2018.03.022 | Check for updates'.

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

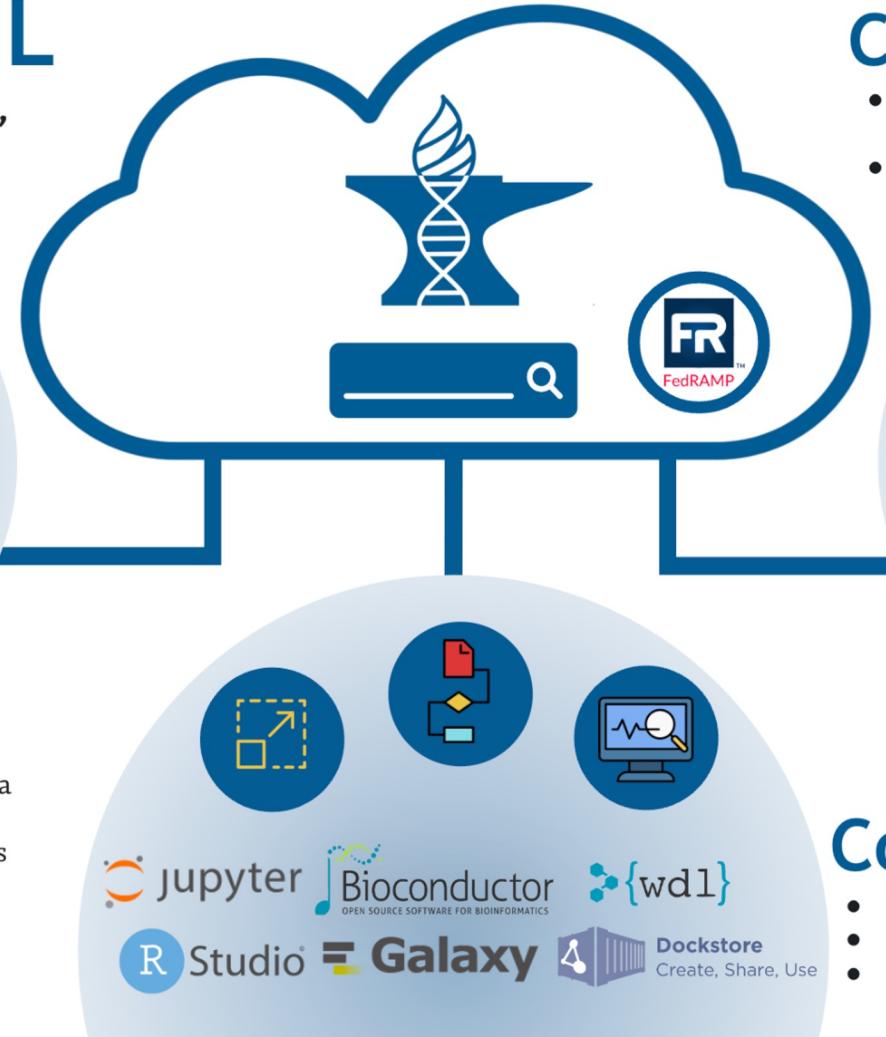


Data Repository

CCDG
GTEx
eMERGE
... and more!

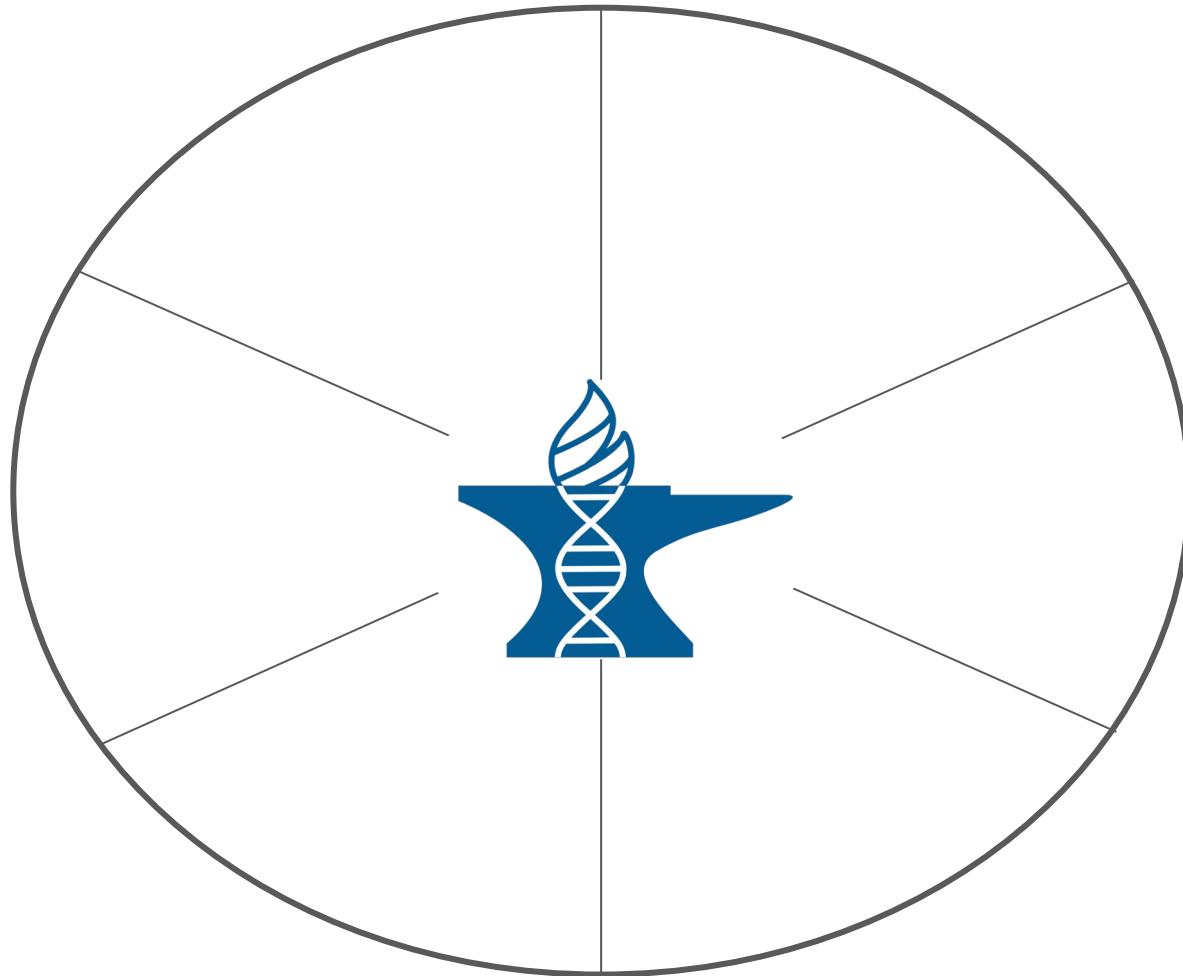
Data

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



Community

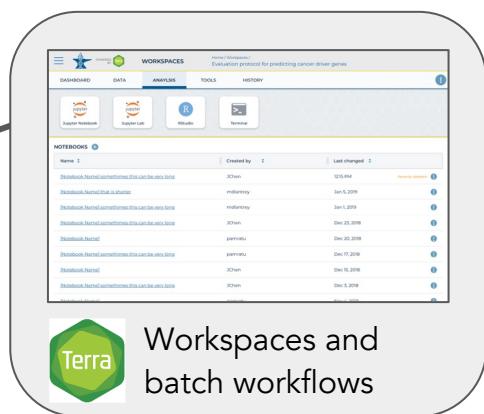
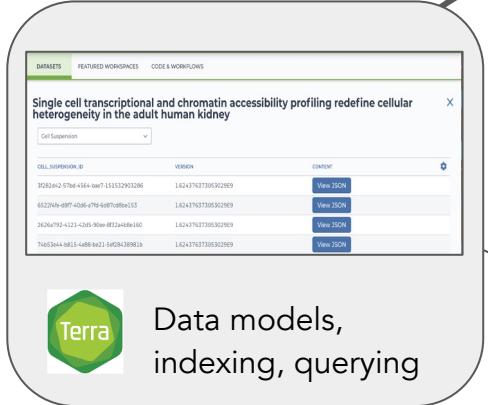
- Share your analysis and data with collaborators or with the world
- Get support at help.anvilproject.org

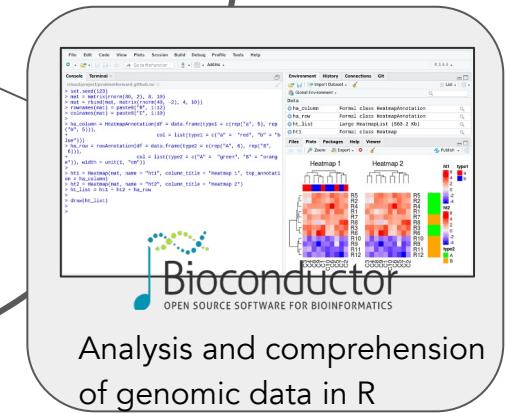
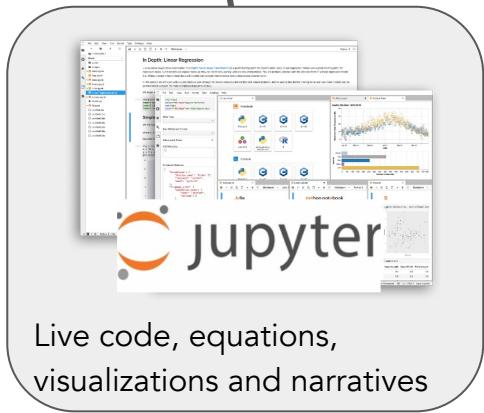
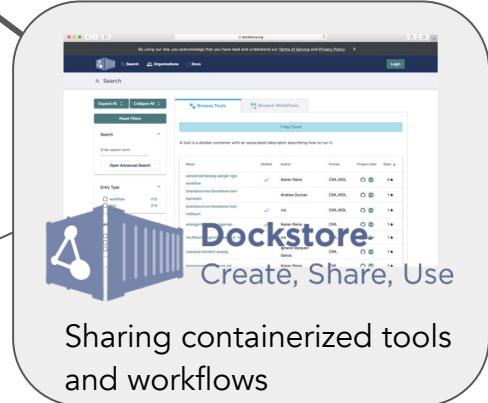
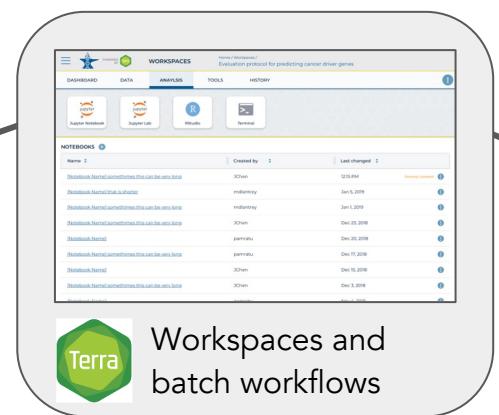
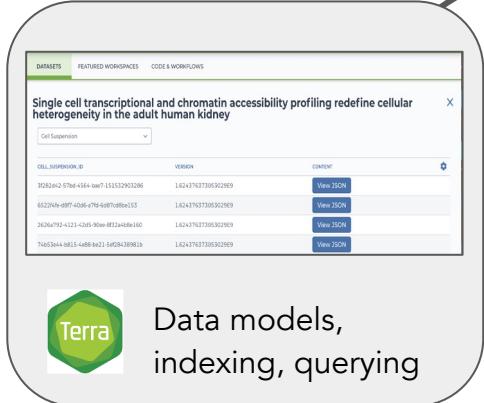




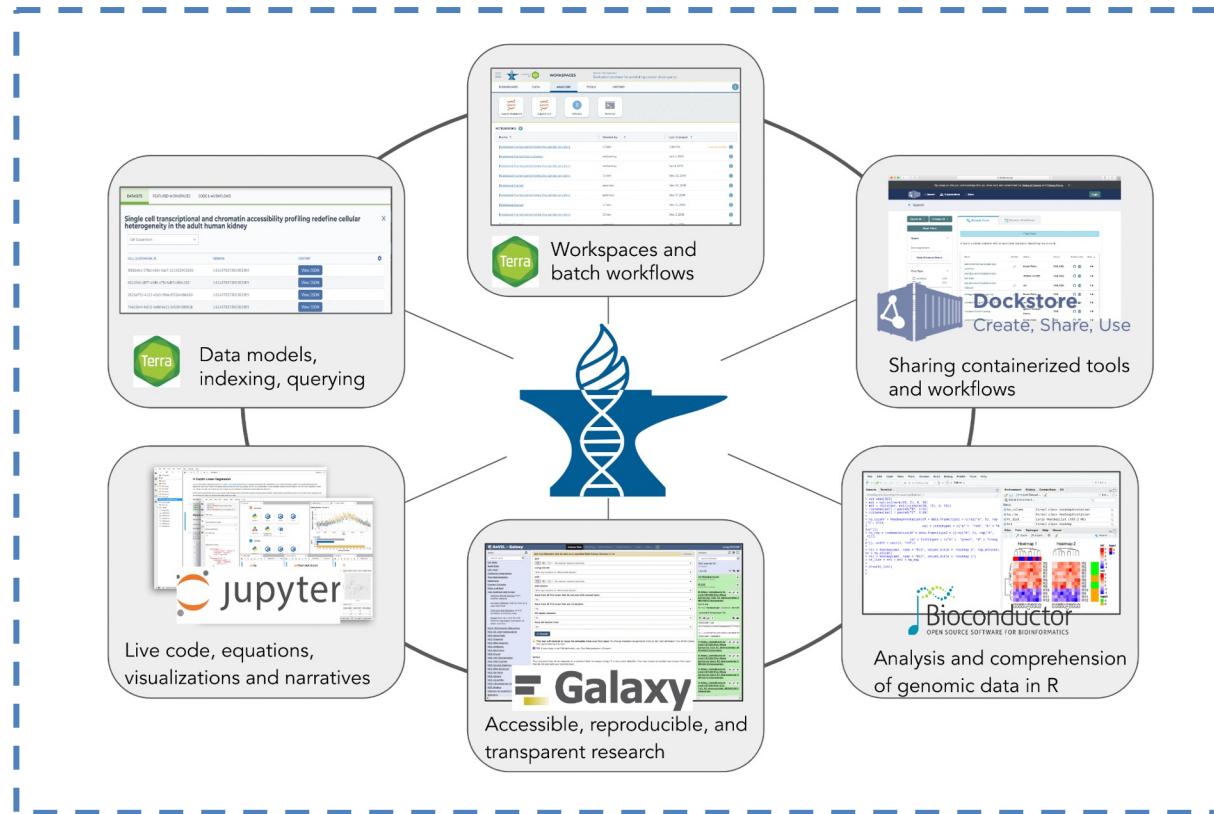
Workspaces and
batch workflows







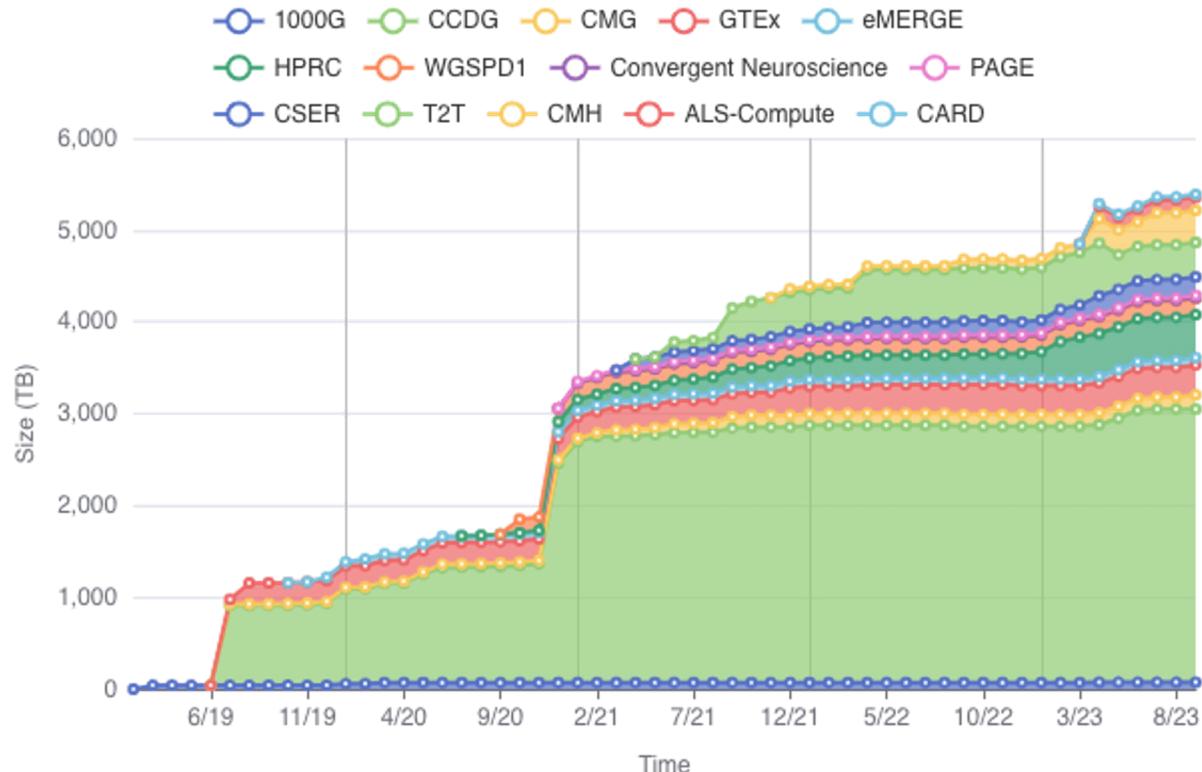
AnVIL: A Secure Federated Data & Compute Ecosystem



Implemented on Google Cloud Platform



Data ingestion & harmonization



Current Consortia

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

Ongoing Consortia

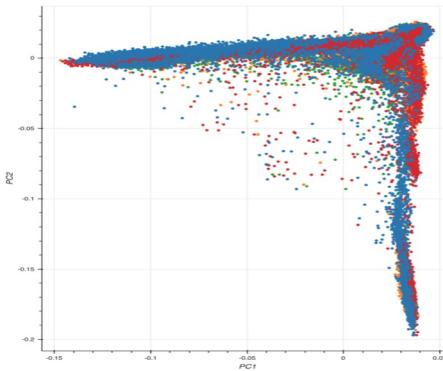
- GREGOR
- PRIMED
- IGVF
- Covid19hg
- CSER
- NIA, NIMH, UDN

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

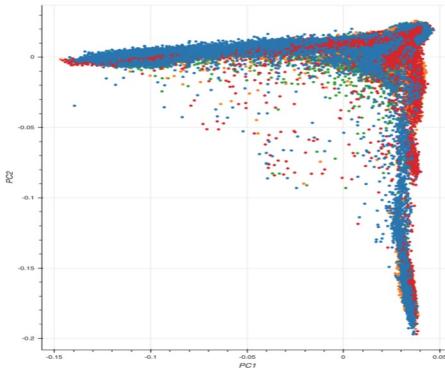
NYGC: 40,975

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

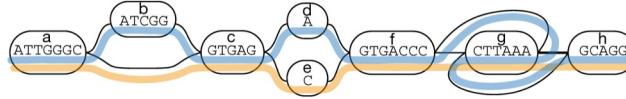
WashU: 43,620



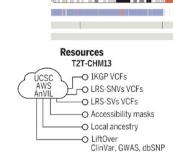
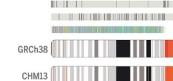
TOWARDS A
COMPLETE
REFERENCE OF
HUMAN GENOME
DIVERSITY



ATTGGGGCATCGGGTGAGAGTGACCCCTTAAGGCAGG
ATTGGGC - - - GTGAGCGTGACCCCTTAAAGCAGG



Comparison of GRCh38 and T2T-CHM13 (chr1)



Resources: T2T-CHM13

UCSC AWS AnVIL

OLBSE SMNis VCFs

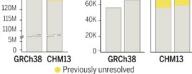
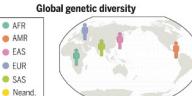
LRS-SVs VCFs

Accessibility masks

Local ancestry

LiftOver

ClinVar, GWAS, dbSNP

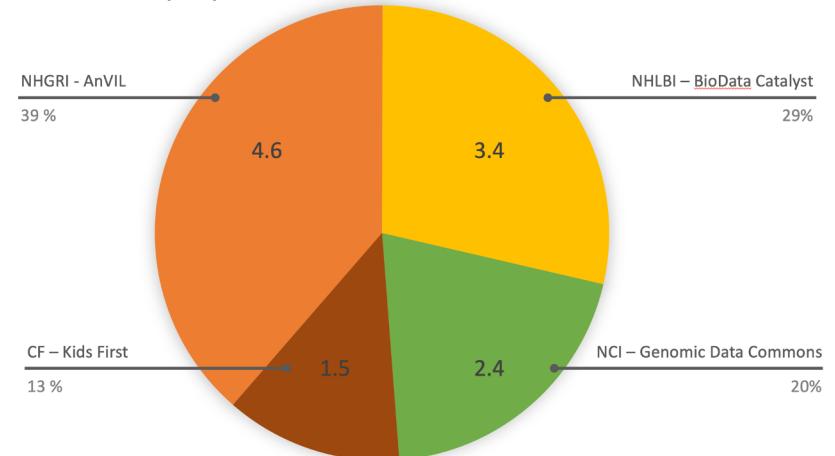


Nurk et al. (2022) Science
Liao et al. (2023) Nature

NCPI & The NCPI Dataset Catalog



Data Size (PB)



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?



By using our site, you acknowledge that you have read and understand our [Terms of Service](#) and [Privacy Policy](#). [X](#)

Dockstore [Search](#) [Organizations](#) [About](#) [Docs](#) [Forum](#) [Login](#) [Register](#)

[Explore Workflows](#)

[Workflows](#) [Tools](#)

[Copy search link](#) Search: the Language is WDL

Notice: Your search has returned greater than 200 results, however only 200 results are shown. We recommend that you narrow your search to find more relevant results.

A Workflow can use multiple containers and executes multiple actions or steps, outlined by one or more descriptors

Popular Keywords +

Name and Description	Verified	Author	Format	Links	Stars ↓
DataBiosphere/topmed-workflows/UM_variant_caller.wdl a place for topmed workflows	✓	Walt Shands	WDL	Link	★ 5
theilagen/terra_utilities/Concatenate_Column_Content	n/a		WDL	Link	★ 4
DataBiosphere/topmed-workflows/UM_aligner.wdl a place for topmed workflows	✓	Walt Shands	WDL	Link	★ 3
AnalysisCommons/genesis_wdl/genesis_GWAS WDL workflow for analysis based on the R GENESIS package		Jen Brody (base code) & Tim Majarian...	WDL	Link	★ 3
theilagen/public_health_viral_genomics/Titan_ClearLabs Bioinformatics workflows for genomic characterization, submission	n/a		WDL	Link	★ 3

Expand All Collapse All Reset

Search

Enter search term...

Open Advanced Search

Category

Search for category

SingleCellAnalysis 17

COVID-19 15

MicrobialGenomics 9

RNASeq 9

GenomicsToolsets 8

▼12 more

Language

WDL 2486

CWL 214

Nextflow 90

Galaxy 85

Language Versions ?

Author

What can you do with AnVIL?



What can you do with AnVIL?



AnVIL Analysis Platforms



jupyter E_1_DESeq2Analysis Last Checkpoint: 3 minutes ago (autosaved)

File Edit View Insert Cell Kernel Navigate Widgets Help

Markdown □

6 Visualizing Differential Expression

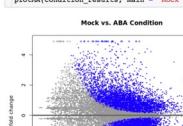
We can also visualize the difference between conditions. An MA-plot shows the mean count of expression between conditions on the y-axis.

In [5]:

```
# Looking at condition-mock versus condition-AAA
# Create MA plot -- resultsid, name="condition_mock_vs_AAA"
# Set plot size to be a big larger
options(jupyter.plot_scale=2)

# Make MA plot
pma(condition_results, main = "Mock vs. AAA Condition")
```

Mock vs. AAA Condition





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

The screenshot shows the FastQC Report interface. At the top, there's a navigation bar with links like Analysis, Home, Visualise, Shared Data, Admin, Help, User, and Logout. Below the header, the main content area has a title "FastQC Report" and a summary section. The summary includes a "Basic Statistics" table and a "Per base sequence quality" chart. The chart displays quality scores across a sequence length of 1 to 176. A legend indicates that green bars represent "Good" quality and yellow bars represent "Warning". The bottom of the page features a "Per sequence quality scores" chart showing the distribution of quality scores for each sequence.



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



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



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

AnVIL Analysis Platforms (coding)



jupyter E1_DESeq2Analysis Last Checkpoint: 3 minutes ago (autosaved)

File Edit View Insert Cell Kernel Navigate Widgets Help

Markdown

6 Visualizing Differential Expression

We can also visualize the difference between conditions. An MA-plot shows the mean count of each gene between condition on the y-axis.

In [51]:

```
# Looking at condition=mock versus condition=ABA
condition_results <- resultsdds, name="condition_mock_vs_ABA"

# Set plot size to be a big larger
options(sup器.plot.scale=2)

# Make MA plot
pma(condition_results, main = "Mock vs. ABA Condition")
```

Mock vs. ABA Condition

The MA-plot displays the relationship between the log2 fold change (y-axis) and the mean log2 count (x-axis). The y-axis ranges from -10 to 10, and the x-axis ranges from 0 to 10. A horizontal grey line at y=0 represents the zero-fold change threshold. Most data points are clustered around the origin, indicating genes with low differential expression. A few points are located in the upper right quadrant, representing genes that are highly upregulated in the ABA condition. A faint diagonal line is visible, representing the identity line where log2 fold change equals the mean log2 count.



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



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



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

AnVIL Analysis Platforms (graphical workflows)



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

The screenshot shows the FasTQC report interface. The top navigation bar includes links for Home, Help, About, and Log In. Below the header, there's a search bar and a main content area with two tabs: 'Basic Statistics' and 'Per base sequence quality'. The 'Basic Statistics' tab displays a table with various metrics like GC content, base bias, and sequencing depth. The 'Per base sequence quality' tab shows a detailed chart of quality scores across the genome. The bottom of the page features a footer with the FasTQC logo and a link to the GitHub repository.



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

The screenshot shows the DataRobot AI Platform interface. The top navigation bar includes 'File', 'Edit', 'Project', 'Data', 'Machine Learning', 'Analytics', 'Visualizations', 'Dashboards', 'Workspaces', and 'Help'. Below the navigation is a 'WORKSPACES' section with a star icon and the text 'WORKSPACES'. A 'New Project' button is visible. The main workspace area contains several project cards:

- Project 1**: Status 'Active', Model 'Random Forest', Last Run '2023-09-01 10:00 AM', Last Model '2023-09-01 10:00 AM'.
- Project 2**: Status 'Active', Model 'Random Forest', Last Run '2023-09-01 10:00 AM', Last Model '2023-09-01 10:00 AM'.
- Project 3**: Status 'Active', Model 'Random Forest', Last Run '2023-09-01 10:00 AM', Last Model '2023-09-01 10:00 AM'.
- Project 4**: Status 'Active', Model 'Random Forest', Last Run '2023-09-01 10:00 AM', Last Model '2023-09-01 10:00 AM'.
- Project 5**: Status 'Active', Model 'Random Forest', Last Run '2023-09-01 10:00 AM', Last Model '2023-09-01 10:00 AM'.

A central panel displays a distribution plot titled 'AIFR Above Frequency Distribution' with the y-axis labeled 'Number of Projects' and the x-axis labeled 'AIFR Above Frequency'. The plot shows a long tail of projects with high AIFR values.

At the bottom left, there is a code editor window with the following Python code:

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score

# Load dataset
df = pd.read_csv('https://raw.githubusercontent.com/databricks/MLlib-Samples/master/mllib-guide/mllib-guide-data/mllib-guide-data.csv')

# Split into training and testing sets
train, test = train_test_split(df, test_size=0.2)

# Train a Random Forest model
model = RandomForestClassifier(n_estimators=100)
model.fit(train[["AIFR", "AIFR_Bin", "AIFR_Threshold", "AIFR_Probability"]], train["Label"])

# Make predictions
preds = model.predict(test[["AIFR", "AIFR_Bin", "AIFR_Threshold", "AIFR_Probability"]])

# Calculate accuracy
accuracy = accuracy_score(test["Label"], preds)
print(f"Accuracy: {accuracy:.2f}")

# Print feature importance
importance = pd.DataFrame(model.feature_importances_, index=["AIFR", "AIFR_Bin", "AIFR_Threshold", "AIFR_Probability"])
print(importance)
```



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



- + 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



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

The logo consists of a blue hexagonal icon followed by the text "wdl" in a black monospace-style font, enclosed in blue curly braces.

- + 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?

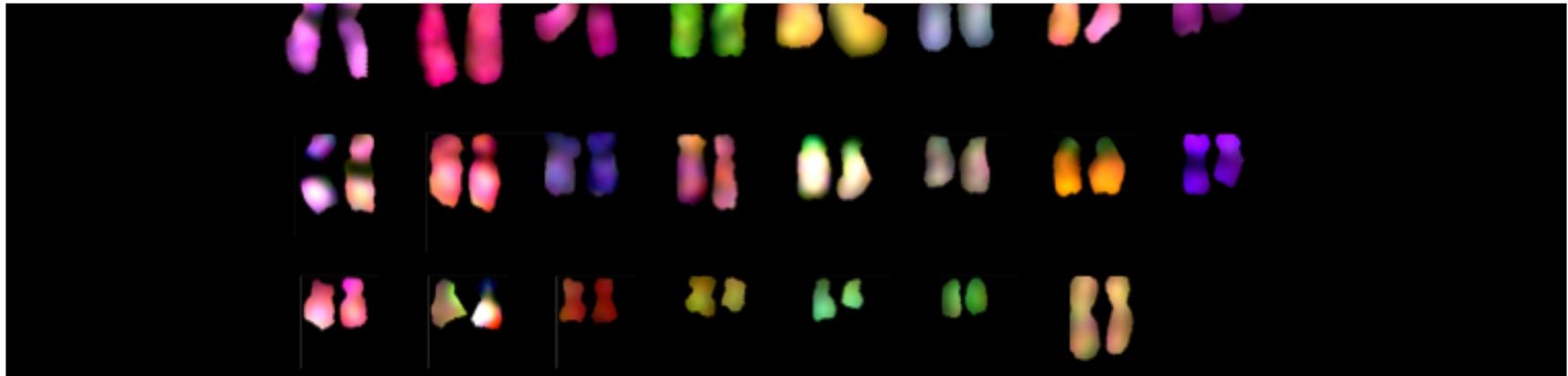
Automatically copy to buckets for next stage

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!



Cell Resource

High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios

Graphical abstract

Authors
Marta Byrska-Bishop, Uday S. Evani, Xuefang Zhao, ..., Michael E. Talkowski, Giuseppe Narzisi, Michael C. Zody

Correspondence
mbyrska-bishop@nygenome.org (M.B.-B.), mczody@nygenome.org (M.C.Z.)

In brief
High-coverage whole-genome sequencing (WGS) of the expanded 1000 Genomes Project (1kGP) cohort including 602 trios led to the discovery of additional rare non-coding single-nucleotide variants (SNVs), as well as coding and non-coding short insertions and deletions (INDELS) and structural variants (SVs) spanning the allele frequency spectrum compared to the original 1kGP resource based primarily on low-coverage WGS.

Highlights

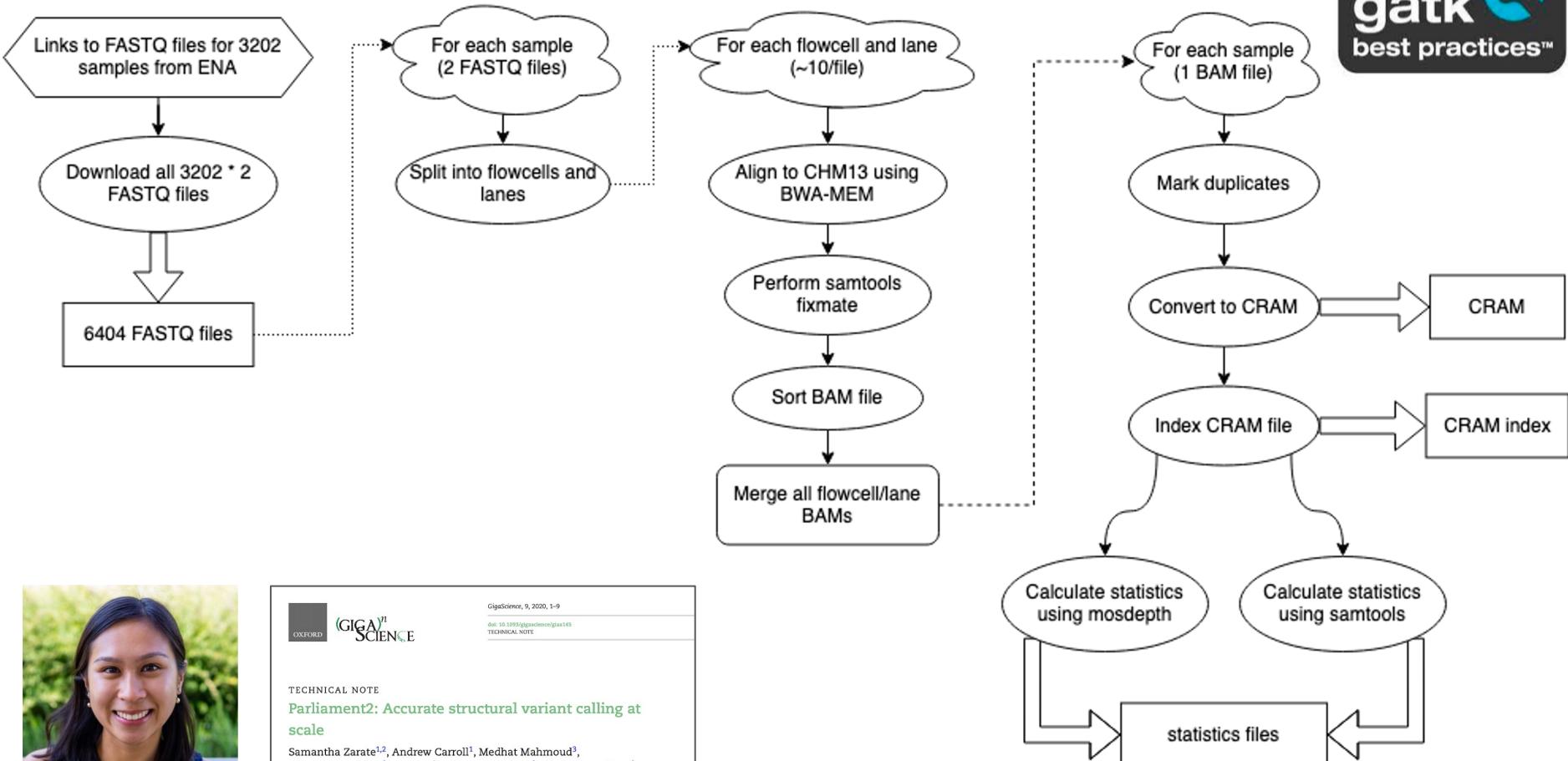
- Expansion of the 1000 Genomes Project (1kGP) resource to include 602 trios
- High-coverage whole-genome sequencing of the expanded 1kGP cohort
- Discovery of more rare SNVs as well as INDELS and SVs across the frequency spectrum
- Generation of an improved and accessible reference imputation panel

Byrska-Bishop et al., 2022, Cell 185, 3426–3440
September 1, 2022 © 2022 The Authors. Published by Elsevier Inc.
<https://doi.org/10.1016/j.cell.2022.08.004>

CellPress

3202 samples from 26 populations

3202 samples x 30Gb = 96Tb input data



Samantha Zarate



Core usage over 24 hours

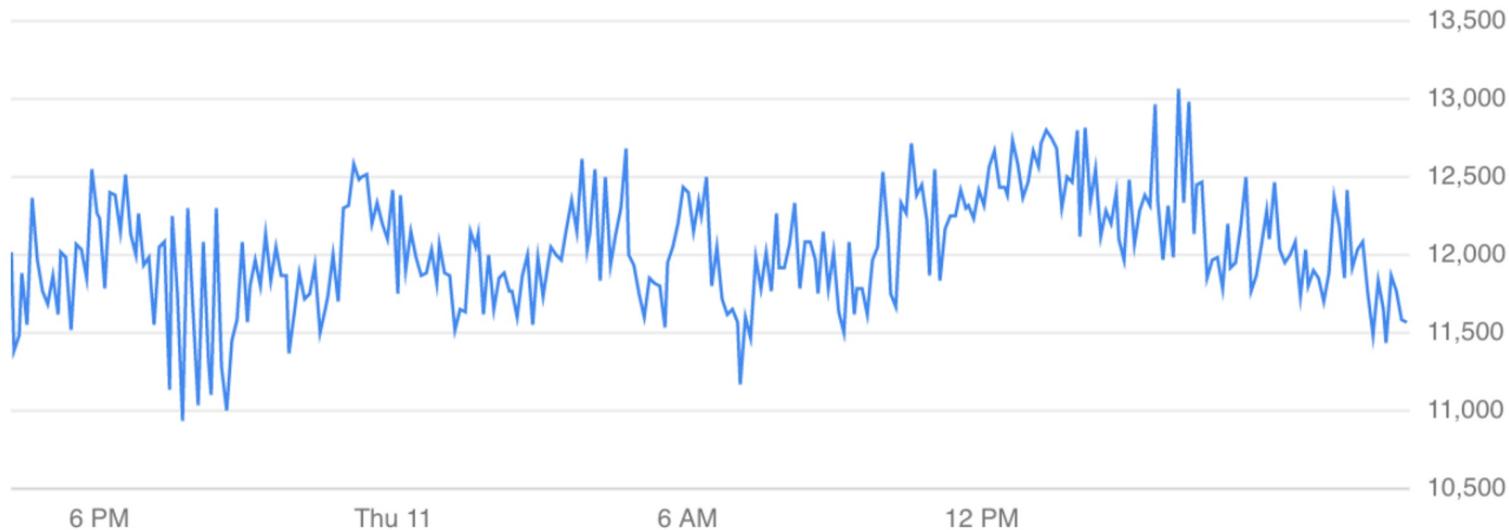


Preview

1 hour

4 hours

1 day



- instance/cpu/reserved_cores: 11,552.00



Samantha Zarate

AnVIL: RStudio in the Cloud

WORKSPACES RStudio

DASHBOARD DATA NOTEBOOKS WORKFLOWS JOB HISTORY

```
R File Edit Code View Plots Session Build Debug Profile Tools Help
```

Source

```
Console Terminal
```

```
print(pop)
superpop = filter(stats, Superpopulation_code==pop)
#ggplot(superpop, aes(x=percentage_of_properly_paired_reads_%, fill=Sex)) + geom_density(alpha=0.4) + theme(facet_grid.stats$Population_code)
#ggplot(stats, aes(x=percentage_of_properly_paired_reads_%, fill=Superpopulation_code)) + geom_density(alpha=0.4) + theme(facet_grid.stats$Population_code)
for (pop in levels(stats$Superpopulation_code)) {
  print(pop)
  superpop = filter(stats, Superpopulation_code==pop)
  #ggplot(superpop, aes(x=percentage_of_properly_paired_reads_%, fill=Sex)) + geom_density(alpha=0.4) + theme(facet_grid.stats$Population_code)
  #ggplot(superpop, aes(x=percentage_of_properly_paired_reads_%, fill=Sex)) + geom_density(alpha=0.4) + theme(facet_grid.stats$Population_code)
  #ggplot(superpop, aes(x=percentage_of_properly_paired_reads_%, fill=Sex)) + geom_density(alpha=0.4) + theme(facet_grid.stats$Population_code)
  print("done")
}
[1] "AFR"
[1] "done"
[1] "AMR"
[1] "done"
[1] "EAS"
[1] "done"
[1] "EUR"
[1] "done"
[1] "SAS"
[1] "done"
> ggplot(stats, aes(x=percentage_of_properly_paired_reads_%, fill=Sex)) + geom_density(alpha=0.4) + theme(facet_grid.stats$Population_code)
> ggplot(stats, aes(x=percentage_of_properly_paired_reads_%, fill=Superpopulation_code))
> et_grid(stats$Population_code)
> for (pop in levels(stats$Superpopulation_code)) {
+   print(pop)
+   superpop = filter(stats, Superpopulation_code==pop)
+   #ggplot(superpop, aes(x=percentage_of_properly_paired_reads_%, fill=Population_code)) + geom_density(alpha=0.4) + theme(facet_grid.stats$Population_code)
+   #ggplot(superpop, aes(x=percentage_of_properly_paired_reads_%, fill=Sex)) + geom_density(alpha=0.4) + theme(facet_grid.stats$Population_code)
+   #ggplot(superpop, aes(x=percentage_of_properly_paired_reads_%, fill=Sex)) + geom_density(alpha=0.4) + theme(facet_grid.stats$Population_code)
+   #ggplot(superpop, aes(x=percentage_of_properly_paired_reads_%, fill=Sex)) + geom_density(alpha=0.4) + theme(facet_grid.stats$Population_code)
+   print("done")
+ }
[1] "AFR"
Press [enter] to continue
[1] "AMR"
Press [enter] to continue
[1] "EAS"
Press [enter] to continue
[1] "EUR"
Press [enter] to continue
[1] "SAS"
Press [enter] to continue
> ggplot(superpop, filter(state, Superpopulation_code==pop)) + ggplot(superpop, aes(x=percentage_of_properly_paired_reads_%, fill=Sex)) + geom_density(alpha=0.4) + theme(facet_grid.superpop$Population_code) + gpt
> pop = "AMR"; superpop = filter(stats, Superpopulation_code==pop); ggplot(superpop, aes(x=percentage_of_properly_paired_reads_%, fill=Sex)) + geom_density(alpha=0.4) + theme(facet_grid.superpop$Population_code) + gpt
> pop = "AFR"; superpop = filter(stats, Superpopulation_code==pop); ggplot(superpop, aes(x=percentage_of_properly_paired_reads_%, fill=Sex)) + geom_density(alpha=0.4) + theme(facet_grid.superpop$Population_code) + gpt
> pop = "EAS"; superpop = filter(stats, Superpopulation_code==pop); ggplot(superpop, aes(x=percentage_of_properly_paired_reads_%, fill=Sex)) + geom_density(alpha=0.4) + theme(facet_grid.superpop$Population_code) + gpt
> pop = "EUR"; superpop = filter(stats, Superpopulation_code==pop); ggplot(superpop, aes(x=percentage_of_properly_paired_reads_%, fill=Sex)) + geom_density(alpha=0.4) + theme(facet_grid.superpop$Population_code) + gpt
> pop = "SAS"; superpop = filter(stats, Superpopulation_code==pop); ggplot(superpop, aes(x=percentage_of_properly_paired_reads_%, fill=Sex)) + geom_density(alpha=0.4) + theme(facet_grid.superpop$Population_code) + gpt
Warning message:
Removed 1 rows containing non-finite values (stat_density).
> |
```

File Plot Package Help Viewer

Zoom Export ⚙

Superpopulation: SAS

Sex female male

density

percentage_of_properly_paired_reads_(%)

888

He

TU

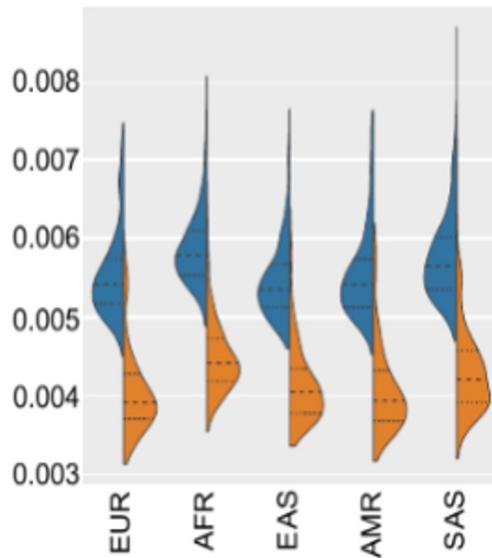
PJL

SUS

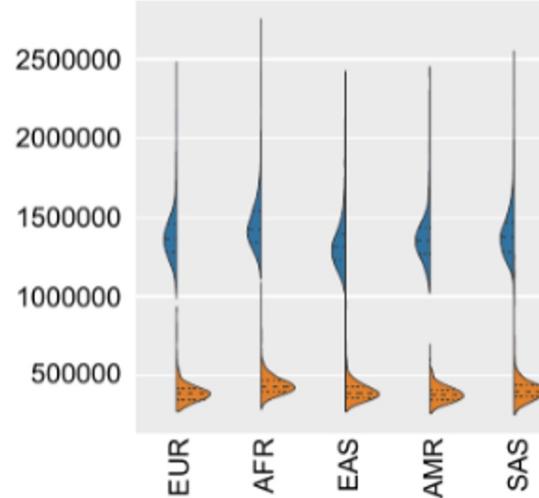
1000G Mapping on T2T-CHM13



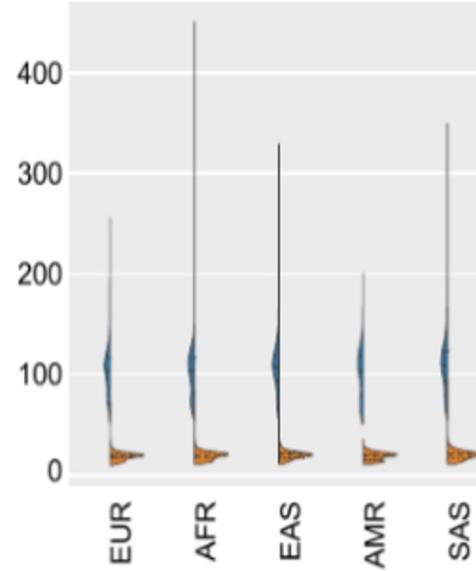
Mismatch rate



Invalid pairs

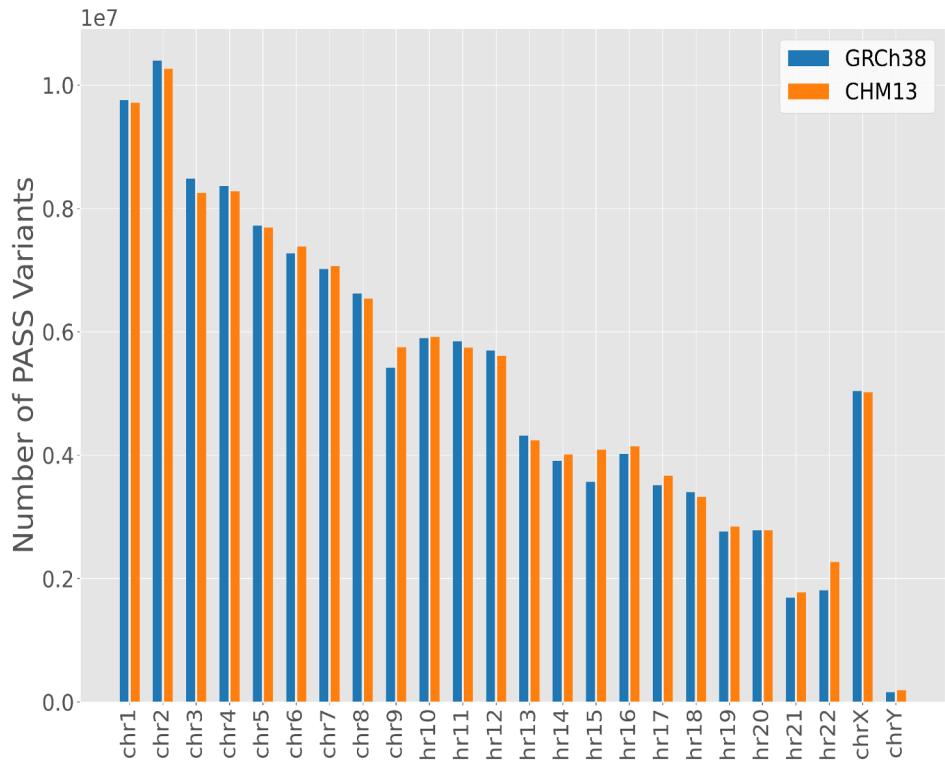


Cov. std. within genes



■ GRCh38 ■ CHM13

More Variants Found Using CHM13

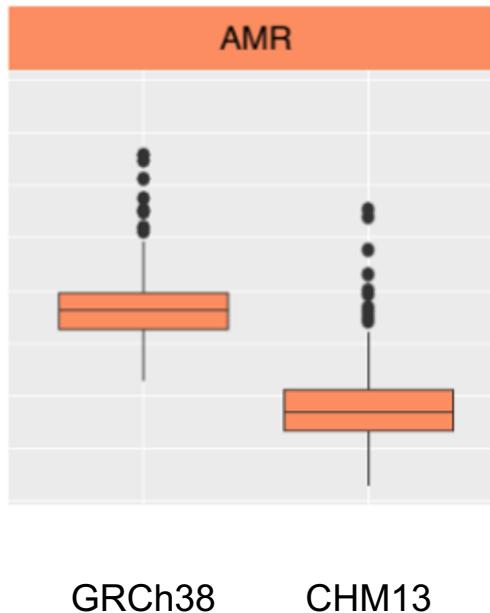


# of PASS variants	GRCh38	CHM13
125,484,020	125,484,020	126,591,489

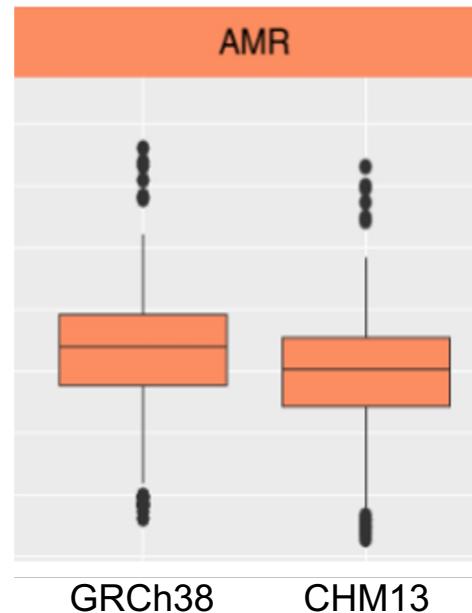
1000G Per-Sample Variant Counts on T2T-CHM13



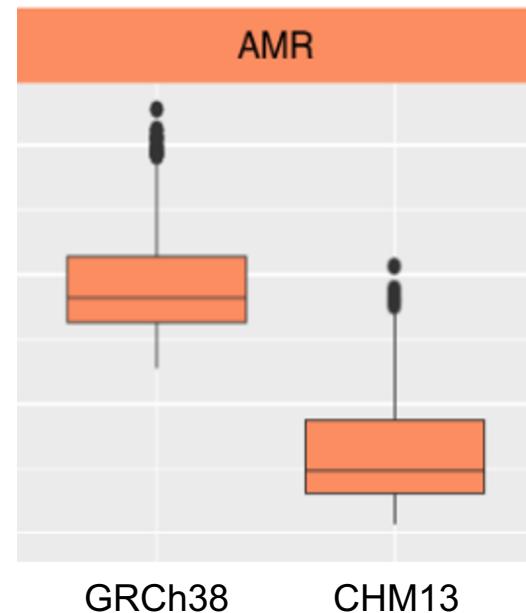
Genome-wide



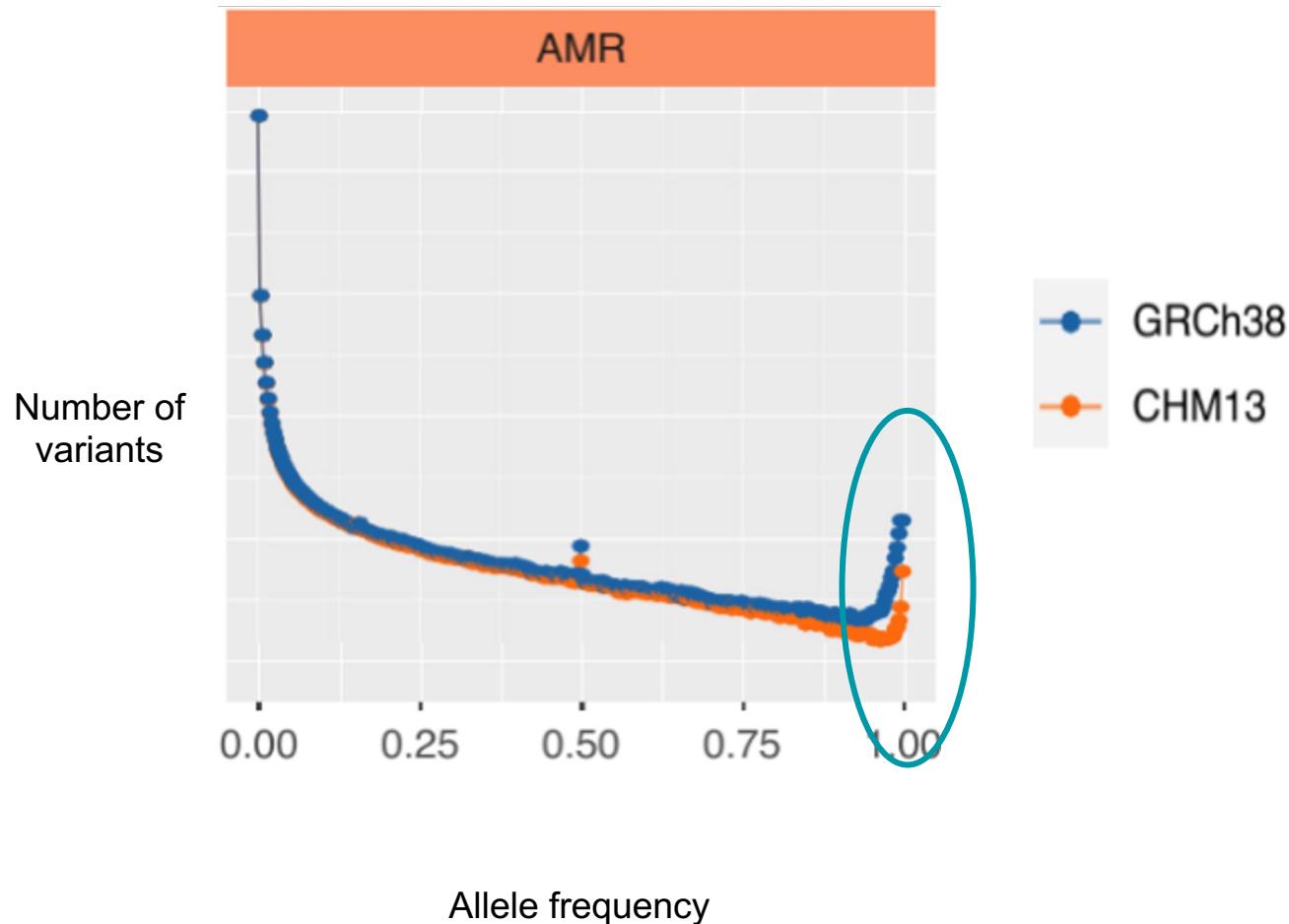
Heterozygous



Homozygous



Explaining Decreased Per-Sample Count



Allele Frequency = 1: Reference error / private variant

GRCh38



CHM13



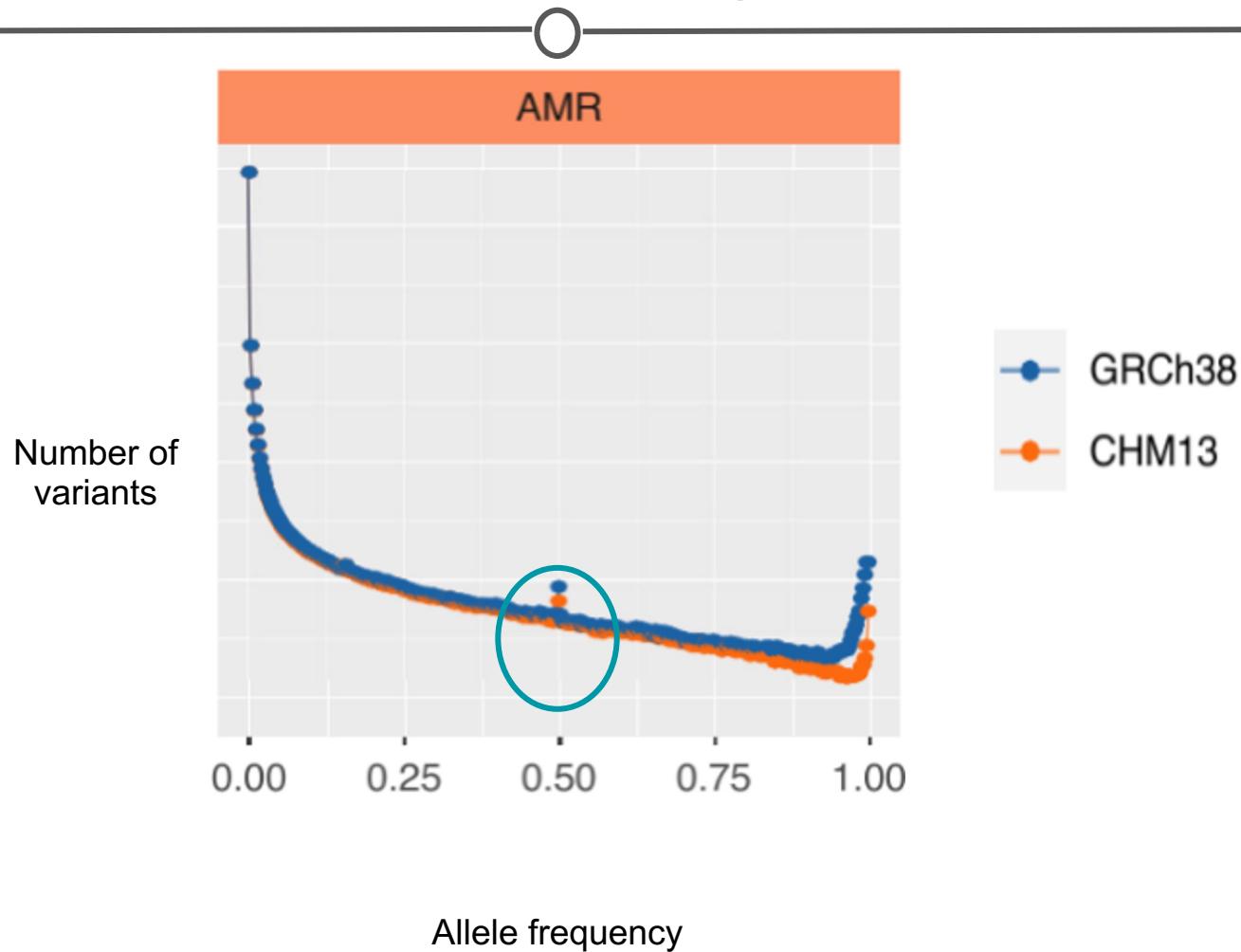
Samples



Samples



Explaining Decreased Per-Sample Count



Allele Frequency \approx 0.5: Collapsed duplication

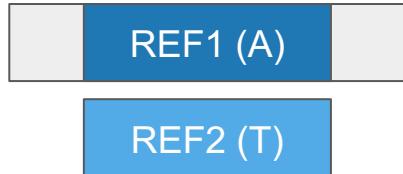
GRCh38



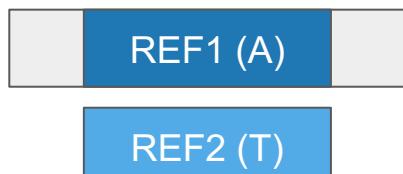
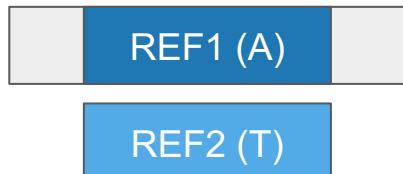
CHM13



Samples



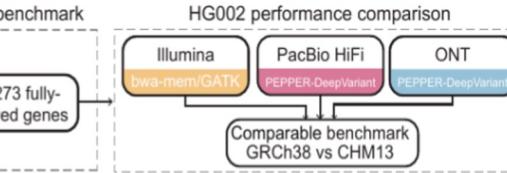
Samples



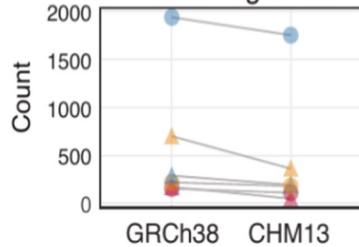
Improved Variant Calling in Clinical Genes

CHM13 challenging medically-relevant genes benchmark

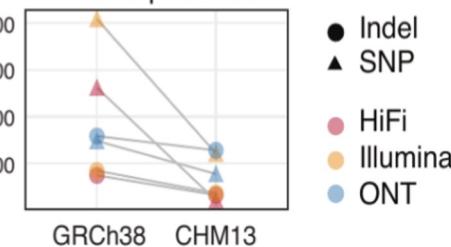
273 medically-relevant genes
HG002 trio-hifiasm



False negatives



False positives



- Indel
- ▲ SNP
- HiFi
- Illumina
- ONT

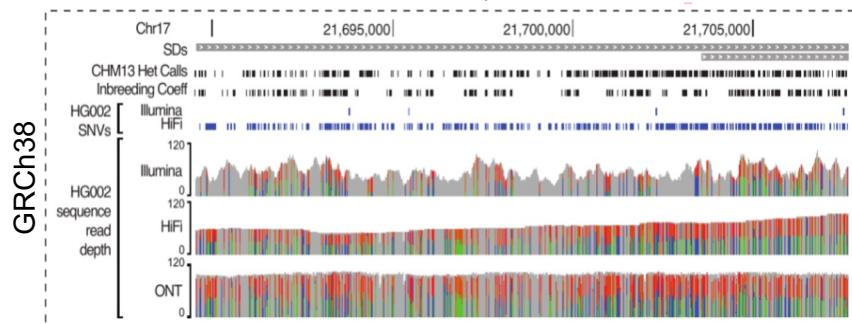


Justin Zook

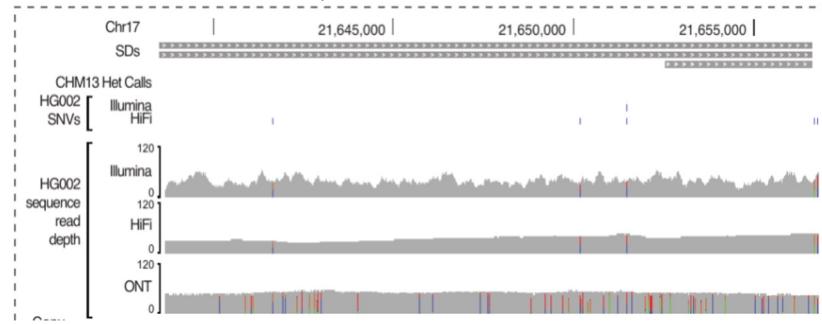


Fritz Sedlazeck

GRCh38

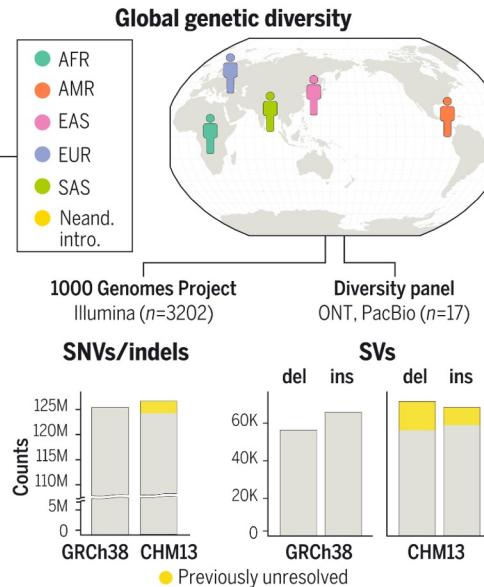
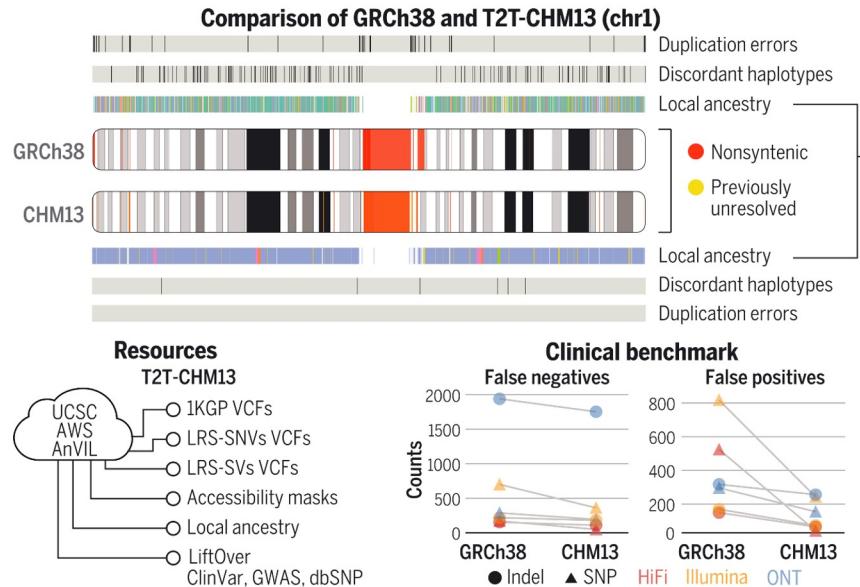


CHM13



- *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

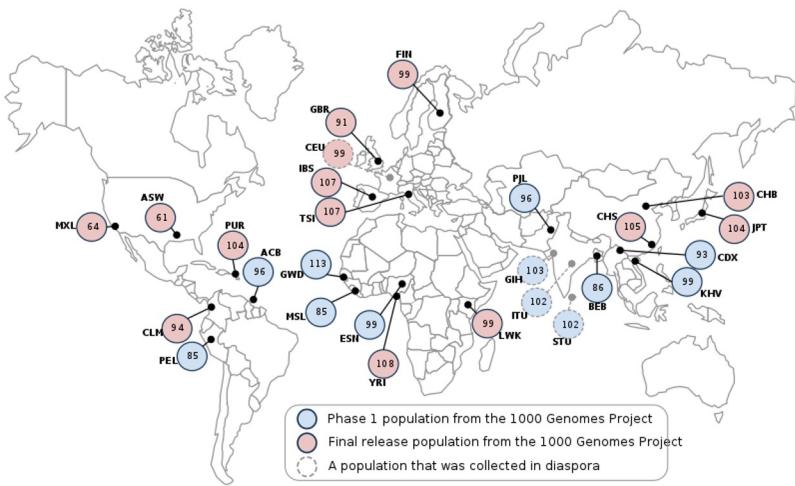
Aganezov, S*, Yan, SM*, Soto, DC*, Kirsche, M*, Zarate, S*, et al. (2022) Science. doi: 10.1126/science.abl3533

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>

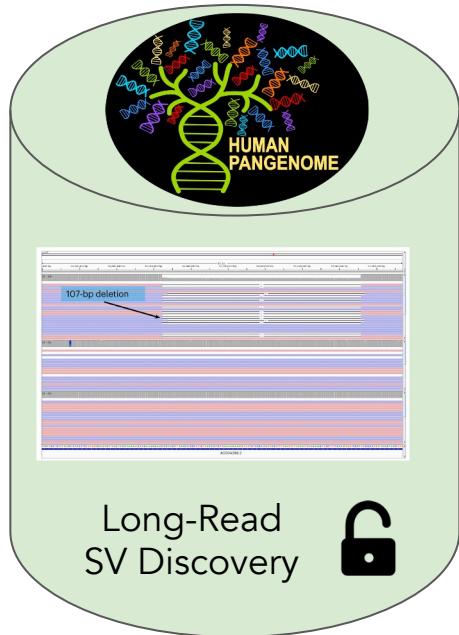


Stephen
Hwang

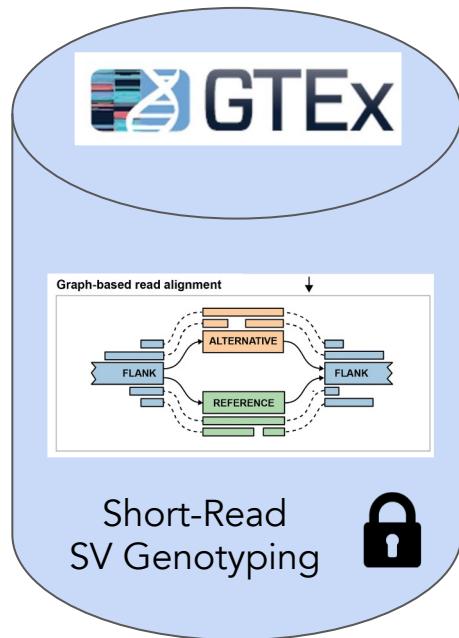


Dylan Taylor

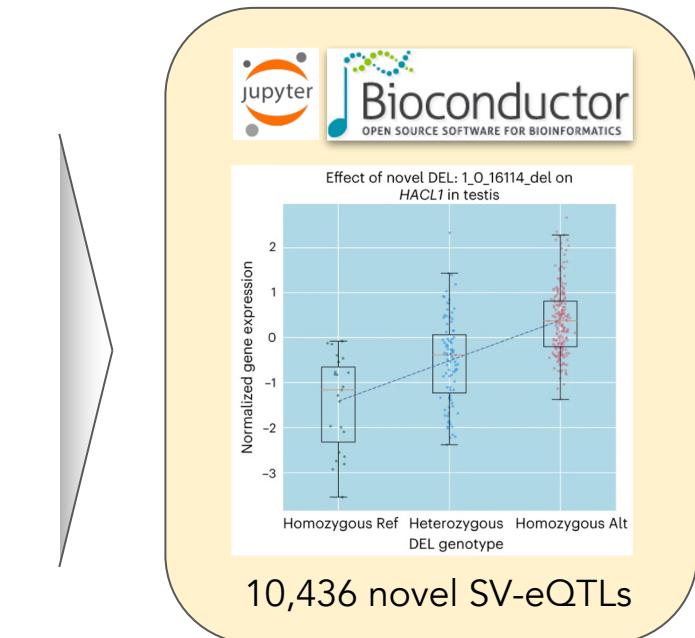
Data Integration and Reuse



Long-Read
SV Discovery



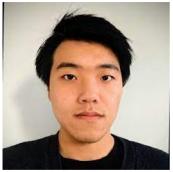
Short-Read
SV Genotyping



Melanie Kirsche

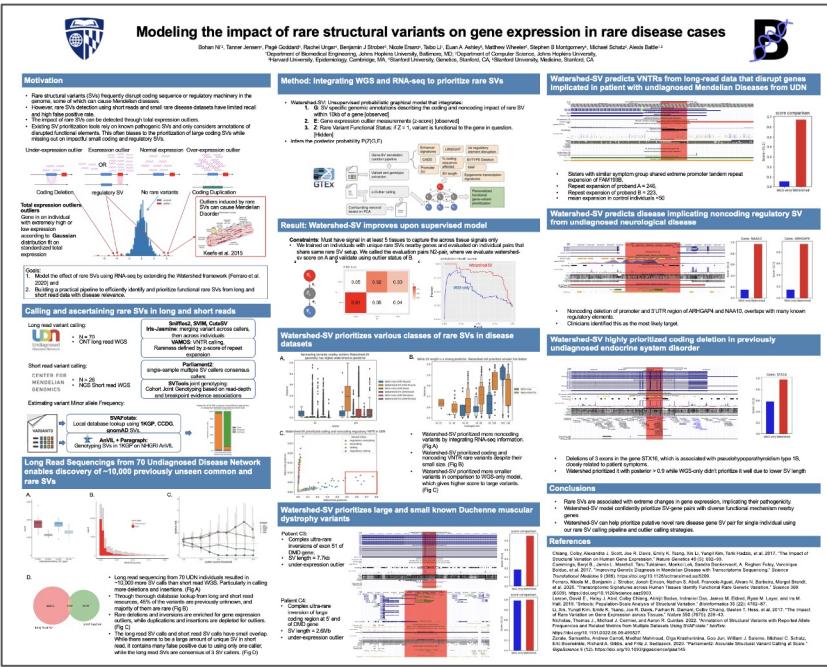
Jasmine and Iris: population-scale structural variant comparison and analysis
Kirsche et al. (2023) *Nature Methods*. <https://doi.org/10.1038/s41592-022-01753-3>

Ongoing Projects



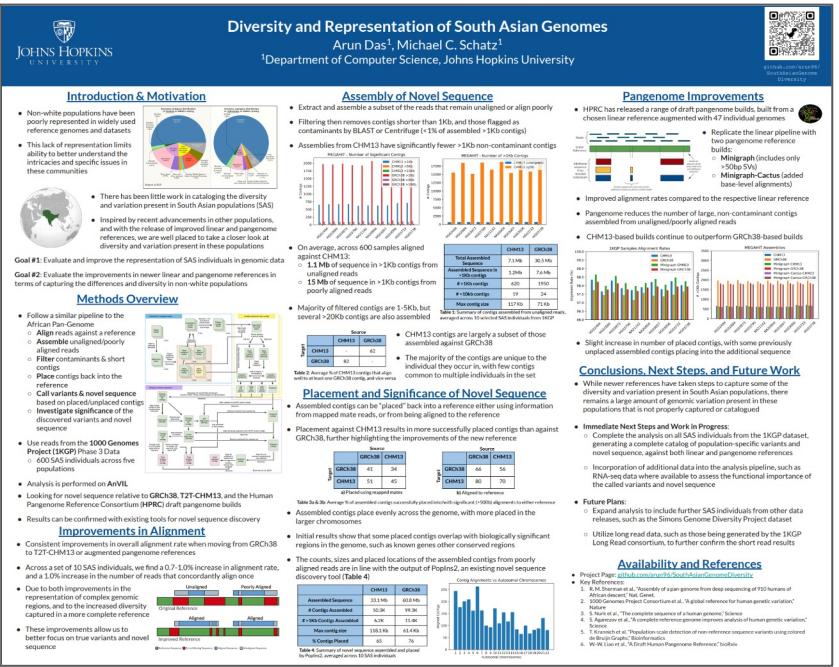
Bohan Ni

Rare Variant Analysis



Arun Das

South Asian Pan-Genome



Diversifying Genomics



Perspective

Diversifying the genomic data science research community

The Genomic Data Science Community Network¹

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 Latino, 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 petabytes 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 (Khouri 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,

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

Corresponding authors: rosa.alcazar@covccollege.edu, amelia.howard@med.cornell.edu, jstevenort@msis.edu, sroy@utep.edu

Article published online before print. Article first published online: 10 August 2022. Article © 2022 The Author(s). *Journal compilation* © 2022 Cold Spring Harbor Laboratory Press; ISSN 1088-9051/22; www.genome.org





Learning More!

anvilproject.org

Migrate Your Genomic Research to the Cloud

Secure, cost-effective genomic analysis at scale.

[Get Started](#) [Learn More](#)

Terra Collaborate in Terra, AnVIL's secure, scalable, cloud compute environment. [Launch](#) [Learn More](#)

Gen3 Manage, harmonize, and share large datasets. [Launch](#) [Learn More](#)

Dockstore Create and share Docker-based workflows. [Launch](#) [Learn More](#)

Galaxy Run batch analysis workflows and interactive visualizations. [Learn More](#)

Bioconductor Analyze genomic data in the R statistical language. [Learn More](#)

Jupyter Run interactive analysis with python or R. [Learn More](#)

Access diverse, open and controlled access, cloud-hosted datasets

450+ Cohorts 4+ Petabytes 600+ thousand Participants

CMG Centers for Mendelian Genetics [Learn More](#) [Datasets](#)

CCDG Centers for Common Disease Genomics [Learn More](#) [Datasets](#)

GTEX The Genotype-Tissue Expression Project [Learn More](#) [Datasets](#)

anvilproject.org

Learn

[Introduction](#) [Data Analysts](#) [Investigators](#) [Data Submitters](#)

Getting Started with AnVIL

The AnVIL platform is an AnVIL-supported data commons running on the Google Cloud Platform (GCP). AnVIL enables researchers to analyze high-value open and controlled access genomic datasets with popular analysis tools in a secure cloud computing environment.

AnVIL uses Terra as its analysis platform, Gen3 for data search and artificial cohort creation, and Dockstore as a repository for Docker-based genomic analysis tools and workflows.

In addition to Docker-based analysis workflows, AnVIL supports popular interactive analysis tools such as Jupyter notebooks, Bioconductor, RStudio, and Galaxy.

By operating in the cloud, AnVIL users can scale analyses from a single computer to thousands and securely share data, workflows, and reproducible results with collaborators and colleagues.

About AnVIL's Documentation

AnVIL's training materials curate, and augment existing component and tool documentation, and show how to use AnVIL's parts together to accomplish the goals of AnVIL's different user personas.

To complement this onboarding and introductory section, the AnVIL team is in the process of developing persona-specific guides and tutorials. For example see the guides for data analysts, investigators, developers, instructors, and data contributors.

New User Onboarding

The following is a guided walk-through of the AnVIL / Terra documentation with a focus on onboarding and preparing new users to run genomic analyses in the cloud.

This section covers:

1. Setting up and linking user accounts.
2. Obtaining access to AnVIL data.
3. An overview of Terra workspaces.
4. An overview of cloud compute costs and setting up billing.

Setting Up and Linking User Accounts

All you need is a Google account to register with Terra and browse

Cell Genomics

Volume 2 Number 1 January 12, 2022

CellPress

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

Discourse: <https://help.anvilproject.org/>

Office Hours & Events: <https://anvilproject.org/events>

Paper: Schatz, Philippakis et al. (2022) *Cell Genomics*. doi: 10.1016/j.xgen.2021.100085

AnVIL Team



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, Qiang 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

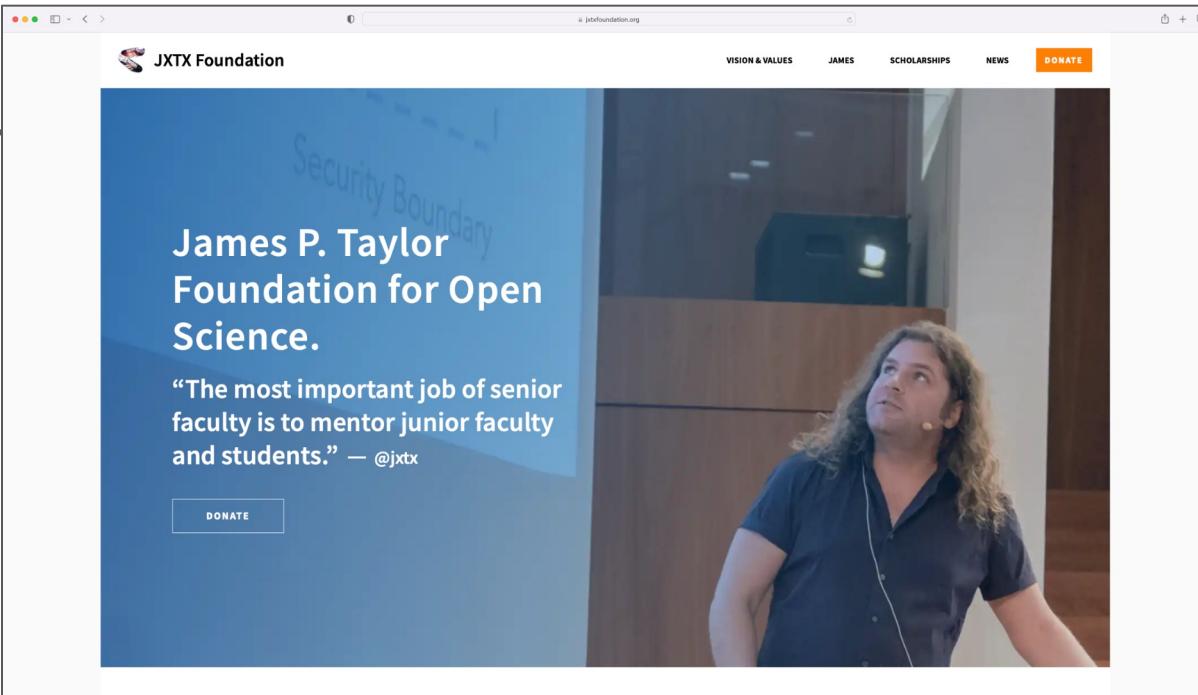
Carnegie Institution for Science

Frederick Tan

City University of New York

Levi Waldron, Sehyun Oh, Ludwig Geistlinger



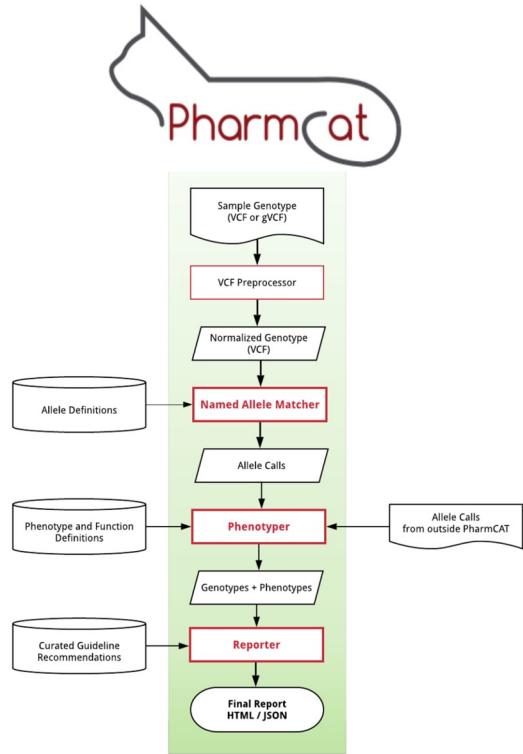


Thank you!

<http://schatz-lab.org>



PharmCAT: Pharmacogenomics Clinical Annotation Tool



The screenshot shows the AnVIL Galaxy interface for running the PharmCAT tool. The search bar contains 'pharmacat'. The main panel displays the 'Full PharmCAT pipeline' workflow. A message indicates that a VCF file is required, but none is currently available. Below this, options for generating TSV files and reporter JSON reports are shown, each with a 'No' radio button selected. At the bottom right is a 'Execute' button. A note at the bottom states: 'PharmCAT is a tool to extract all CPIC guideline gene variants from a genetic dataset (represented as a VCF file), interpret the variant alleles, and generate a report.' A citation for the tool is listed at the bottom left: Klein, T. E., & Ritchie, M. D. (2017). PharmCAT: A Pharmacogenomics Clinical Annotation Tool. *Clinical Pharmacology & Therapeutics*, 104(1), 19–22. <https://doi.org/10.1002/cpt.928>

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 Network¹

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.]

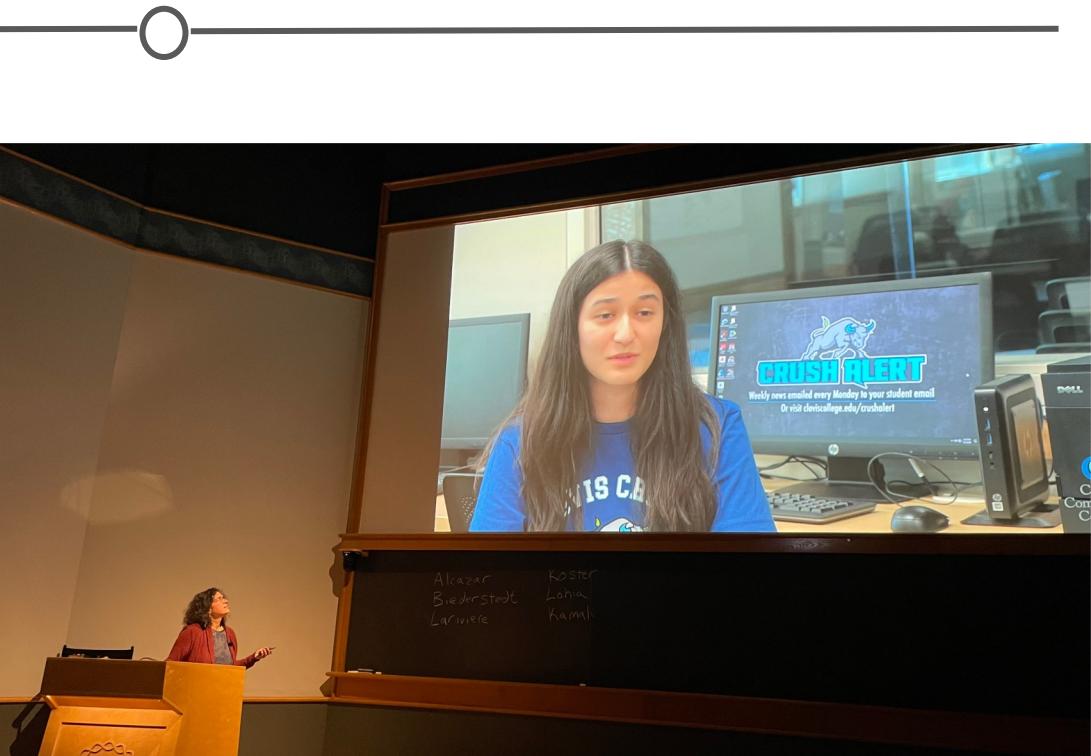
Foundations for justice in genomic data science

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 Latino, 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 petabytes 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 (Khouri 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, racial, 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



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

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

Corresponding authors: rosa.alcazar@cloviscollege.edu, jstevenort@hccs.edu, sroy@atgc.edu

Article published online before print. Article first published online in Genome Research, available under a Creative Commons License (Attribution 4.0 International), as described at <https://creativecommons.org/licenses/by/4.0/>. Freely available online through the Genome Research Open Access option.