

Analysis Tools

29OCT2021



Vincent Carey and Anne O'Donnell Luria



Overview



- Core components
 - Batch computing: Terra, Dockstore
 - Interactive computing: Jupyter Notebooks, RStudio/Bioconductor, Galaxy
- Basic Science
 - T2T Workflows and Analysis
- Clinical Science
 - Polygenic risk scores (PRS)
 - seqr
 - AHA Assessment
 - PharmCAT
- Extending AnVIL
 - Available technologies
 - Future directions

12:35-1:50

Session 1: Breakout rooms

Data submission and consortia engagement

Moderators: Dr. Adam Resnick (Children's Hospital of Philadelphia) and Ms. Valentina Di Francesco (NHGRI)

12:35-12:40

Moderator introductions

12:40-12:55

AnVIL presentation:
*Dr. Brian O'Connor (Broad) and
Dr. Frederick Tan (Carnegie)*

12:55-1:40

Discussion

1:40-1:50

Prepare breakout report

Analysis tools

Moderators: Dr. Marylyn Ritchie (University of Pennsylvania) and Dr. Ken L. Wiley, Jr. (NHGRI)

12:35-12:40

Moderator introductions

12:40-12:55

AnVIL presentation:
*Dr. Vincent Carey (HMS)
and Dr. Ira Hall (Yale)*

12:55-1:40

Discussion

1:40-1:50

Prepare breakout report

Breakout room: Analysis tools

Moderators: Dr. Marylyn Ritchie and Dr. Ken Wiley, Jr.

Dr. Nadav Ahituv

Dr. Joshua (Josh) Akey

Dr. Mark Craven

Dr. Sean Davis

Dr. Barbara Engelhardt

Dr. James Knight

Dr. Anshul Kundaje

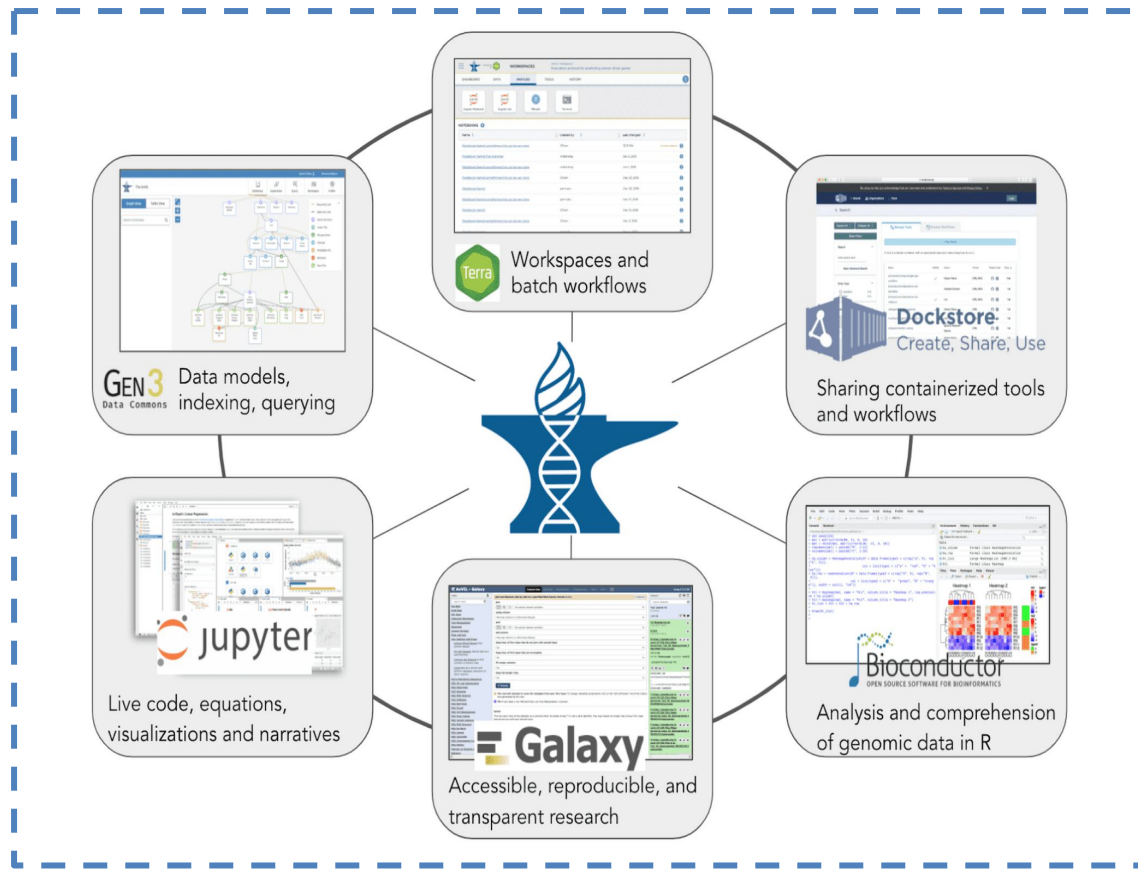
Dr. Karen Miga

Dr. Adam Phillippy

Dr. Timothy (Tim) Reddy

Dr. Chunhua Weng

Building a Secure Federated Data Ecosystem



FedRAMP

FedRAMP certified
1 ATO



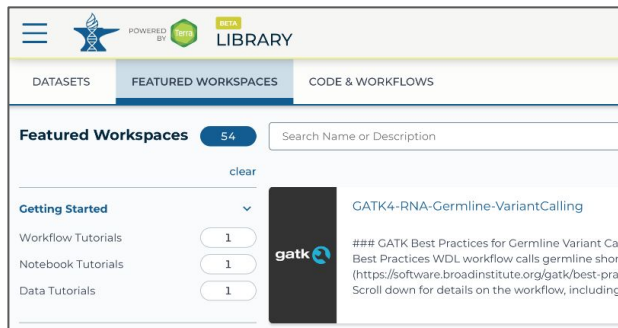
Implemented on Google Cloud Platform

Core Components Overview

Thousands of tools provided by:



Featured workspaces showcase reproducible, ready to run workflows and notebooks



Metrics: Tools & Workflows

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

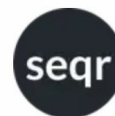
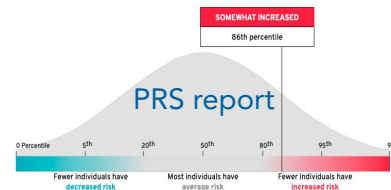
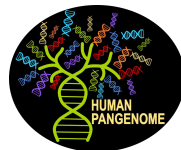


AnVIL toolsets by scale and purpose

Interactive



Create, Share, Use



Large Scale /
Batch

Basic Science

Clinical

The Workspace - The fundamental unit in Terra

Workspace: the fundamental unit in Terra

The screenshot shows a Terra Workspace page. The top navigation bar includes 'DASHBOARD', 'DATA', 'NOTEBOOKS', 'WORKFLOWS', and 'JOB HISTORY'. The main content area is divided into two columns. The left column, titled 'ABOUT THE WORKSPACE', contains a description of the workspace, a list of two parts (exploring phenotypes and testing genetic associations), and notes on data. The right column, titled 'WORKSPACE INFORMATION', displays metadata such as creation date, last updated, submissions, access level, and Google Project ID. Below this, it lists owners and tags (1000 Genomes, GWAS, Jupyter Notebooks, WDLs). At the bottom, it shows the Google Bucket name and location.

ABOUT THE WORKSPACE

This workspace reproduces the fundamental steps in a genome wide association study (GWAS), using 1,000 Genomes Project¹ (phase 3) genotypes and simulated phenotypes.

The analysis is structured in two parts:

1. Explore phenotypes and population structure (Jupyter Notebook - Hail/Python)
2. Test for genetic associations using mixed-models and generate summary visualizations (WDL workflow)

The output of the notebook (part 1) serves as the input to the workflow (part 2).

Instructions for applying the analyses presented in this workspace on your own data are provided in the penultimate section of this documentation.

Notes on data in this workspace

To demonstrate an analysis that could be run on typical whole genome sequence data, this workspace provides mock phenotype data generated from publicly available 1000 Genomes phase 3 genotypes. Phenotypes have been simulated based on individual genotypes and known associated loci for multiple complex traits. The [GCTA software](#)⁶ was used with lists of causal variants and an estimate of narrow sense heritability⁸ for each phenotype.

Traits and sources for causal variants

- a. BMI: GIANT-UKBB meta-analysis²
- b. Fasting glucose: MAGIC³
- c. Fasting insulin: MAGIC³
- d. Waist-to-hip ratio: GIANT-UKBB meta-analysis²
- e. Height: GIANT-UKBB meta-analysis²
- f. HDL: MVP⁴
- g. LDL: MVP⁴
- h. Total cholesterol: MVP⁴
- i. Triglycerides: MVP⁴

WORKSPACE INFORMATION

CREATION DATE 3/19/2020	LAST UPDATED 9/27/2021
SUBMISSIONS 2	ACCESS LEVEL Reader
GOOGLE PROJECT ID amp-t2d-op	

OWNERS

amanning@broadinstitute.org
tmajaria@broadinstitute.org
bshifaw@broadinstitute.org

TAGS

1000 Genomes GWAS
Jupyter Notebooks WDLs

Google Bucket

Name: fc-d5eb5311-1c3a-4c8f-84c5-...
Location: US (multi-region)
[Open in browser](#)

Collaboratively access and organize data, launch tools and run analyses

Featured workspaces: see how science gets done

The screenshot shows the Terra Library page. The top navigation bar includes 'DATASETS', 'FEATURED WORKSPACES', and 'CODE & WORKFLOWS'. The main content area is titled 'Featured Workspaces' and displays a list of featured workspaces. Each workspace entry includes a title, a date, a description, and a 'gatk' logo. The workspaces listed are: GATK4-RNA-Germine-VariantCalling (Jul 14, 2021), TRUST4 (Jun 25, 2021), Peat-Demo (May 18, 2021), and Intro-to-HCA-data-on-Terra (Apr 29, 2021). The right column shows a list of categories with counts: Workflow Tutorials (1), Notebook Tutorials (1), Data Tutorials (1), WDLs (38), Jupyter Notebooks (24), Hail (2), Bioconductor (2), GATK (19), Cumulus (1), Spark (3), GWAS (2), Exome Analysis (2), Whole Genome Analysis (3), Fusion Transcript Detection (1), and RNA Analysis (8).

Featured Workspaces 54

Search Name or Description Sort by most

Getting Started

- Workflow Tutorials 1
- Notebook Tutorials 1
- Data Tutorials 1

Analysis Tools

- WDLs 38
- Jupyter Notebooks 24
- Hail 2
- Bioconductor 2
- GATK 19
- Cumulus 1
- Spark 3

Experimental Strategy

- GWAS 2
- Exome Analysis 2
- Whole Genome Analysis 3
- Fusion Transcript Detection 1
- RNA Analysis 8

GATK4-RNA-Germine-VariantCalling Jul 14, 2021

GATK Best Practices for Germline Variant Calling in RNAseq
Best Practices WDL workflow calls germline short variants (SNPs/Indels) from RNAseq data using GATK v4.1 and related tools. Detailed description of the workflows is available in [Gatk's Best Practices Document](https://software.broadinstitute.org/gatk/best-practices/workflow?id=11164). Scroll down for details on the workflow, including input and output descriptions and requirements.

TRUST4 Jun 25, 2021

TRUST4
A fully reproducible example workflow for immune repertoire reconstruction.
Complete documentation for TRUST4 is available on the [TRUST4 repository](https://github.com/liulab-).

Peat-Demo May 18, 2021

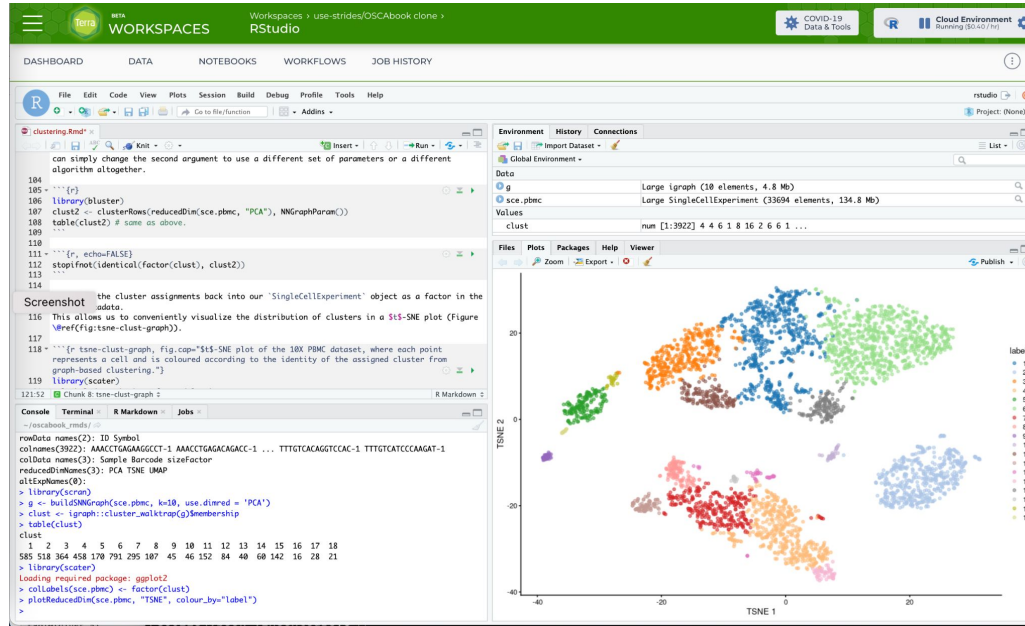
Demo of how to use [Peat (external link)](https://broad.io/peat) to save overhead by grouping jobs into fewer WDL scatter branches. To compare scatter with and without Peat, this workspace has two simple demo workflows using WDL scatter, one with and one without using Peat.

Intro-to-HCA-data-on-Terra Apr 29, 2021

Exploring Human Cell Atlas single-cell data
This tutorial workspace is a step-by-step guide to importing, accessing, and analyzing standardized cell-by-gene count matrices (Loom format) from the Human Cell Atlas (HCA) [Data Portal] (https://data.humancellatlas.org/) using community-supported single-cell analysis tools.

Public showcase workspaces for users to discover ready-to-run analyses

Terra: RStudio + Bioconductor



RStudio: analysis environment preferred by the R community.

- Machine learning, statistical computing, and visualizations

Bioconductor: tools and modules for the analysis and comprehension of high-throughput genomic data, implemented in R

- 1,903 software packages available

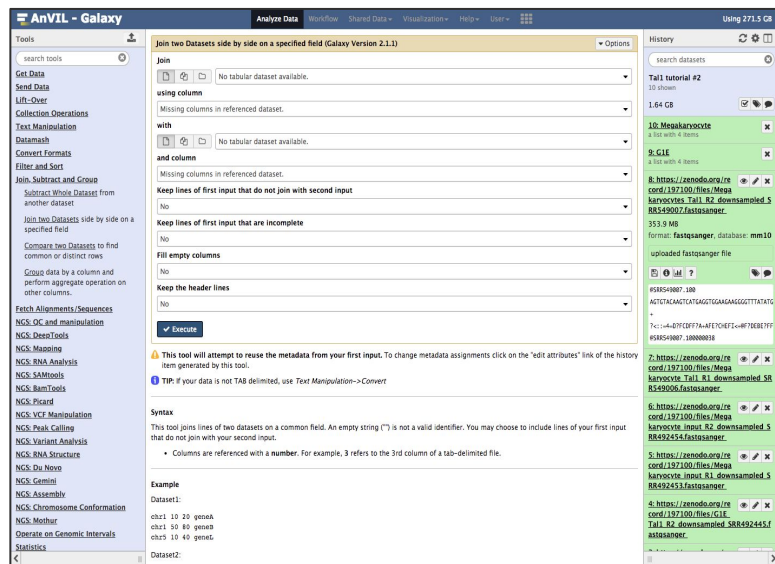
AnVIL provides a robust well tested RStudio environment with the latest Bioconductor release integrated

Galaxy: Cloud-scale & flexible analysis



Galaxy

- Accessible, reproducible, integrative science with thousands of tools
- Large, active community of users and contributors
- World-wide training network with materials, educating thousands every year
- Avoid data downloads
- Use Galaxy without quotas



Web-based analysis environment for running analysis tools and building workflows for users with no programming expertise

Telomere-to-Telomere (T2T) Analysis Workflows

☰ README.md

WDLs for T2T Variants

This directory contains the WDL files used for large-scale short-read

Data ingestion

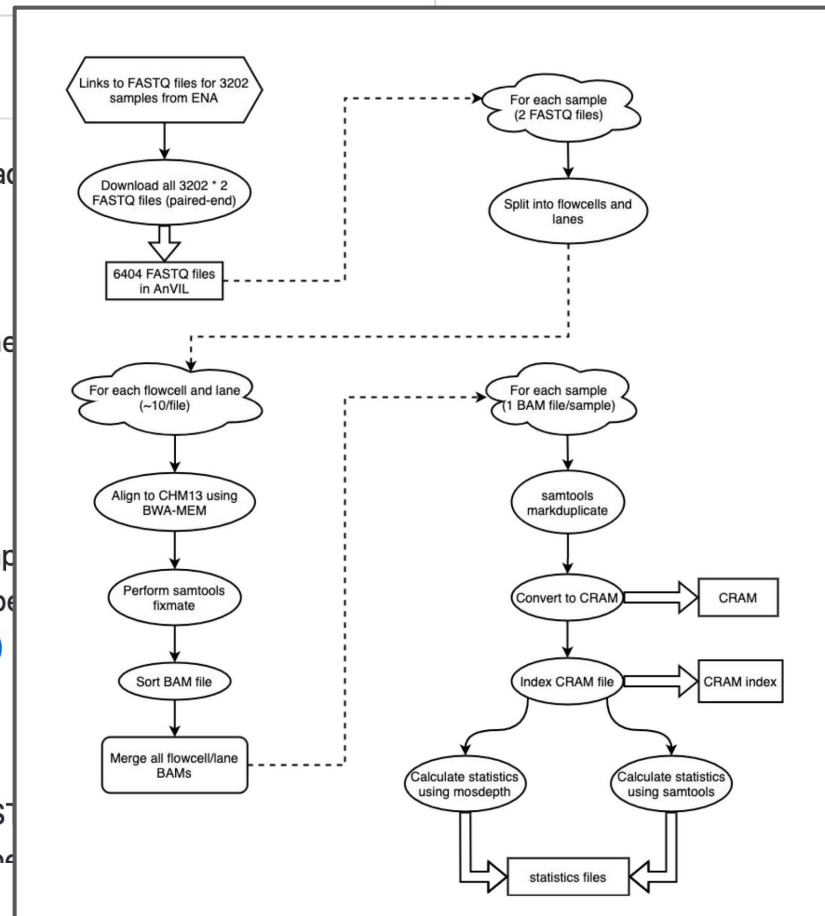
- `wdls/download_aspera.wdl` : Downloads FASTQ files from the Archive (ENA), given accession numbers

Read alignment

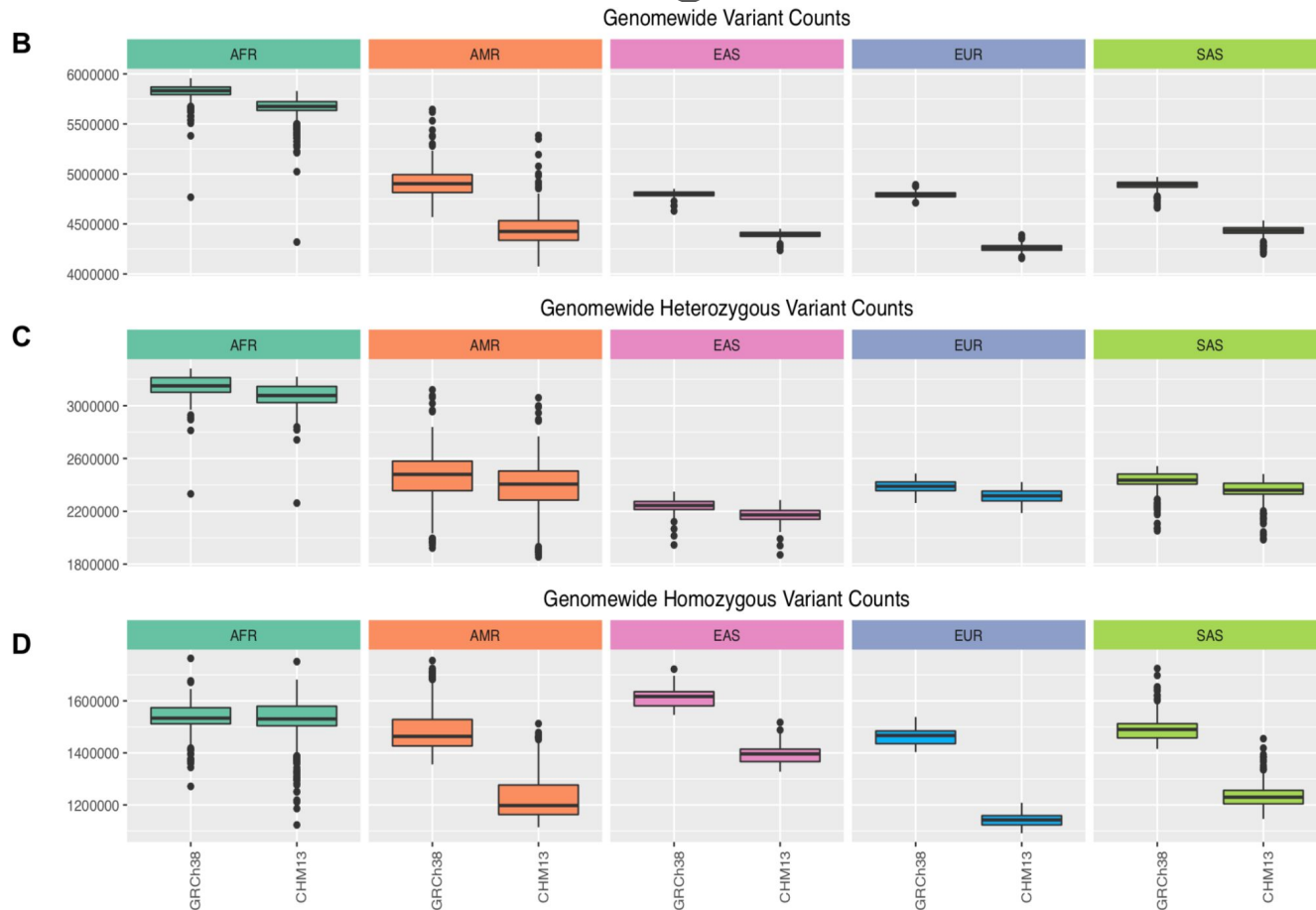
- `wdls/t2t_alignment.wdl` : Given a reference FASTA file, sample FASTQ files, BWA index, and dedup distance (default = 100), perform alignment as described in [Aganezov, Yan, Soto, Kirsche, Zarate, et al. \(2021\)](#)

Variant calling

- `wdls/haplotype_calling_chrom.wdl` : Given a reference FASTA file, sample index and dict, a sample CRAM (plus corresponding index), the

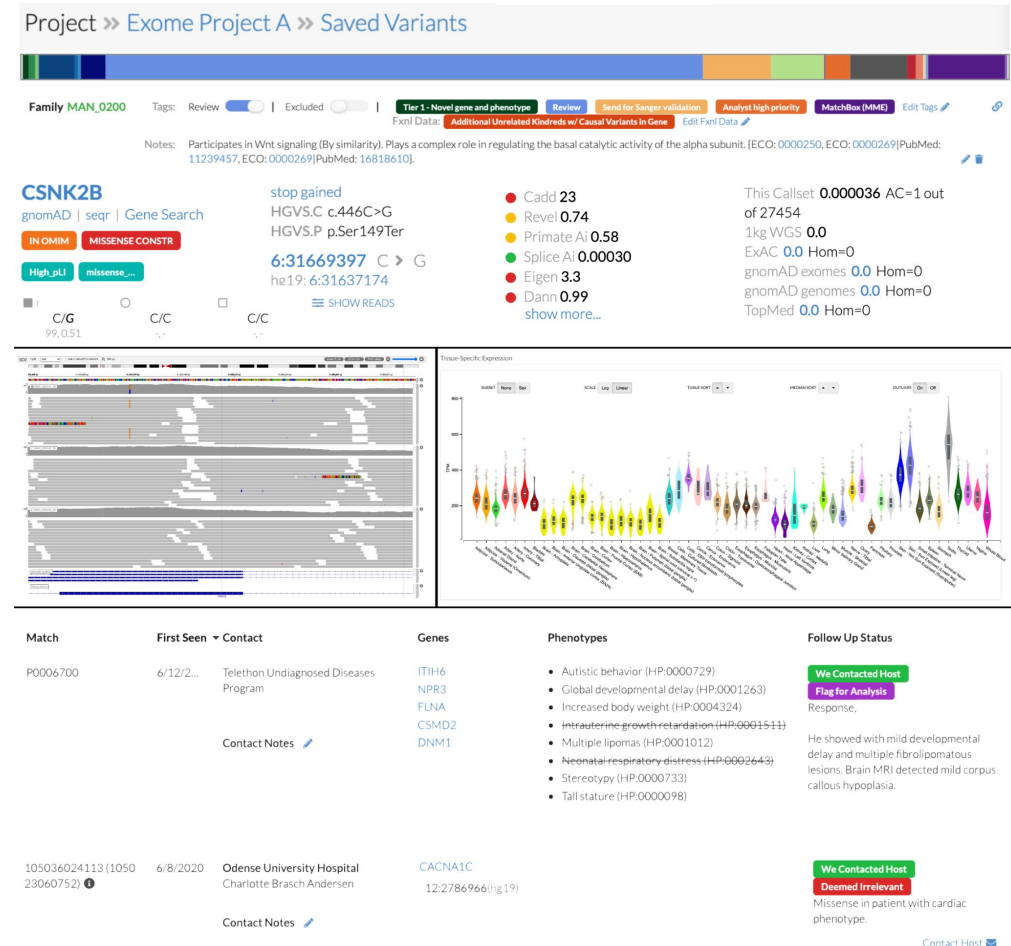


Re-analyzing diversity in 3,202 samples



Clinical Genomics: seqr - Mendelian Inheritance

- Open source software on AnVIL for collaborative exome and genome analysis
- Accepts joint called vcf input file
- Supports collection of extensive metadata
- Matchmaker Exchange node
- Ongoing development ideas
 - CNV (exome) and SV (genome) data loading and SNV-SV compound het searching
 - Improving representation of mito and STR variants
 - Integration of RNA-seq data for analysis
 - Support ACMG variant classification
 - Expansion of Matchmaker Exchange capacity
 - Increased support for cram to variant files to loading in seqr



Clinical Genomics: Jupyter Notebook for PRS

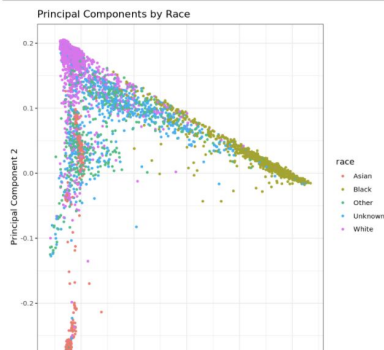
```
In [19]: #Determine the extent to which known risk factors are enriched in those with CAD
#Table 1
tabone=CreateTableOne(c('age', 'male', 'dm', 'currentsmok', 'htn'),data=dat,strata=c('cad'),factorVars=c('male', 'currentsmok', 'htn', 'dm'))
print(tabone,quote = F,digits=1)
#We see that individuals with CAD tend to be older, male, and more likely to smoke or have been diagnosed with diabetes or hypertension
```

	Stratified by cad			
	0	1	p	test
n	12446	644		
age (mean (SD))	48.28 (14.95)	63.12 (8.15)	<0.001	
male = 1 (%)	5081 (40.8)	417 (64.8)	<0.001	
dm = 1 (%)	401 (3.2)	190 (29.5)	<0.001	
currentsmok = 1 (%)	283 (2.3)	31 (4.8)	<0.001	
htn = 1 (%)	2712 (21.8)	595 (92.4)	<0.001	

```
In [20]: # For the genetic data, we compute genetic ancestry using principal components
addmargins(table(dat$race))
round(prop.table(table(dat$race))*100,1)
# Our population is 80% white, 6% Black, and 4% Asian
```

Asian	Black	Other	Unknown	White	Sum
482	784	493	833	10498	13090
3.7	6.0	3.8	6.4	80.2	

```
In [21]: # We can plot genetic ancestry of individuals
q <- ggplot(dat,aes(x=PC1,y=PC2,color=race,fill=race)) + geom_point(size=1)+theme_bw()
q = q+xlab('Principal Component 1')+ylab('Principal Component 2')+ggtitle('Principal Components by Race')
q
```

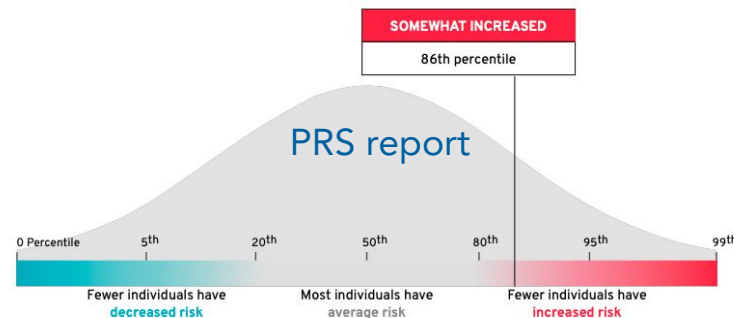


POLYGENIC SCORE REPORT

Coronary Artery Disease (CAD)

Source: CAD Genetic Score Report, Color

Date:
Name:
DOB:



Your polygenic score is in the **86th percentile** of the population. This means that out of every 100 people, your score is higher than 86 of them, but lower than 13. It does NOT mean that you have a 86% chance of developing CAD. This means that your genetic background places you at **somewhat increased risk** to develop the disease. In the U.S., up to 5% of individuals develop CAD by age 50, and up to 25% develop CAD by age 80.

AHA/AnVIL Working Group

Name	URL	Brief Description
PGS Catalog	http://www.pgscatalog.org/	An open database of polygenic scores and the relevant metadata required for accurate application and evaluation.
PRS-RS	https://www.medrxiv.org/content/10.1101/2020.04.23.20077099v2	Improving reporting standards for polygenic scores in risk prediction studies
LDPred	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4596916/	Bayesian PRS that estimates posterior mean causal effect sizes from GWAS summary statistics
PRSice	http://www.prsice.info/	Polygenic Risk Score software for calculating, applying, evaluating and plotting the results of polygenic risk scores (PRS) analyses.
PLink	https://academic.oup.com/gigascience/article/4/1/s13742-015-0047-8/2707533	selects variants below a p-value limit and within a linkage disequilibrium range
SBayesR	https://www.nature.com/articles/s41467-019-12653-0	Improved polygenic prediction by Bayesian multiple regression on summary statistics
WC-2d	https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1006493	Winner's Curse Correction and Variable Thresholding Improve Performance of Polygenic Risk Modeling Based on Genome-Wide Association Study Summary-Level Data
LDpred-func	https://www.biorxiv.org/content/10.1101/375337v3	LDpred-func: incorporating functional priors improves polygenic prediction accuracy in UK Biobank and 23andMe data sets
GCTA	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3014363/	GCTA: A Tool for Genome-wide Complex Trait Analysis
iCARE	https://journals.plos.org/plosone/article/related?id=10.1371/journal.pone.0228198	A Tool for Individualized Coherent Absolute Risk Estimation (iCARE)
Beagle	https://pubmed.ncbi.nlm.nih.gov/3010085/	A One-Penny Imputed Genome from Next-Generation Reference Panels
AlloMap	https://www.caredx.com/allomap/	AlloMap Heart is panel of 11 informative genes and 9 controlled genes. AlloMap Heart detects changes in gene expression associated with acute rejection and provides an actionable score.
CHD Risk	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4608367/	Genetic Risk, Coronary Heart Disease Events, and the Clinical Benefit of Statin Therapy
Breast Cancer Risk	https://pubmed.ncbi.nlm.nih.gov/30554720/	Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes
CAD-South Asians	https://www.jacc.org/doi/full/10.1016/j.jacc.2020.06.024	Validation of a Genome-Wide Polygenic Score for Coronary Artery Disease in South Asians
lcWGS	https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-019-0682-2	Low coverage whole genome sequencing enables accurate assessment of common variants and calculation of genome-wide polygenic scores
cPRS	https://www.sciencedirect.com/science/article/pii/S0002929720302329	Combined Utility of 25 Disease and Risk Factor Polygenic Risk Scores for Stratifying Risk of All-Cause Mortality

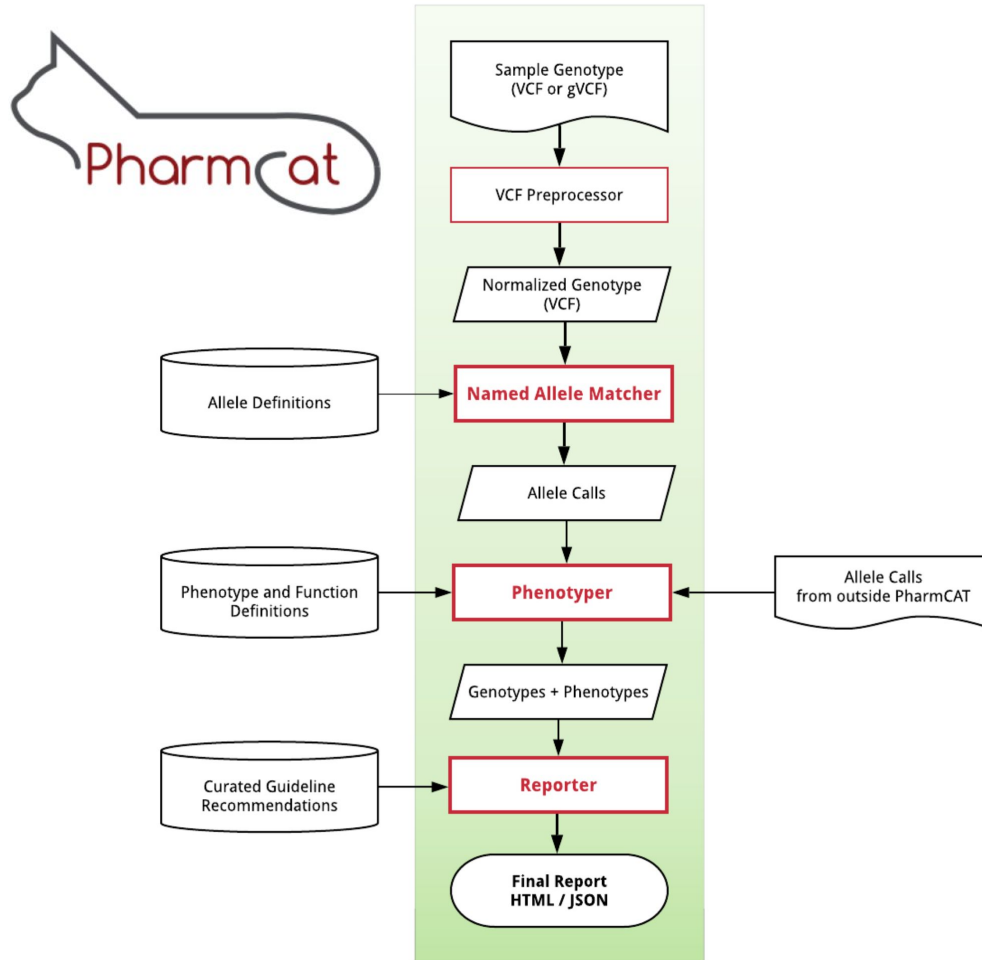


American
Heart
Association®



- Clinical genomics.
- Interviewed a panel of clinical scientists.
- Recommended the prioritization of polygenic risk score (PRS) calculation and pharmacogenetics as our initial focus areas.
- Focus groups of PRS calculation experts led to the identification of the 17 tools listed here.
- Conducted focus groups to discuss resources for pharmacogenetics (Casey Overby Taylor, JHU)

Clinical Genomics: PharmCAT (Coming Soon!)



What does PharmCAT do?

- PharmCAT is used to support clinical decision making
- Utilizes CPIC/PharmGKB guidelines for gene-drug pairs
- Gives prescription recommendations based on genetic variants

How does PharmCAT do what it does?

- VCF file is provided by the user
- Named Allele Matcher matches variant information with Allele definitions and gives a diplotype call for each gene
- The Reporter takes output from the Named Allele Matcher and gives prescription recommendations

Extending AnVIL

- *Bring your own tools and workflows*
 - Either by registering them in Dockstore, or by uploading your own custom WDL to Terra
- *Build on top of the AnVIL APIs*
 - All of the components of the AnVIL provide APIs
 - We will be providing a unified, stable API endpoint for the AnVIL with OpenAPI documentation
 - We are building API wrapper libraries in Python and R, largely generate from the OpenAPI specification but curated
 - See the repo: <https://github.com/anvilproject>
- *Adding new web applications*
 - We are defining standards to allow a containerized web application to be hosted inside AnVIL
 - Leveraging standards container orchestration (Kubernetes) for complex applications



Future directions

- *Integration of third-party applications*
 - Goal: empower app developers with streamlined integration and verification process
 - In process: create official Terra App Dev and Terra App Security guides for third-party developers
- *Machine learning*
 - Harmonized datasets for training and testing models
 - Optimized software libraries with GPU support for efficient processing
 - Advanced visualization capabilities to inspect and debug
 - Model Zoo - Make code and pretrained models available to the genomics community
- *Basic Sciences*
 - More diverse assays, e.g. ENCODE, Roadmap Epigenomics, IGVF, dGTEx
 - Large scale multi-omic integrations
- *Clinical genomics*
 - Diversity of data types, e.g. eMERGE (genomics, medical records, image analysis, etc)
 - Disease associations, clinical reporting, treatment guidelines