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

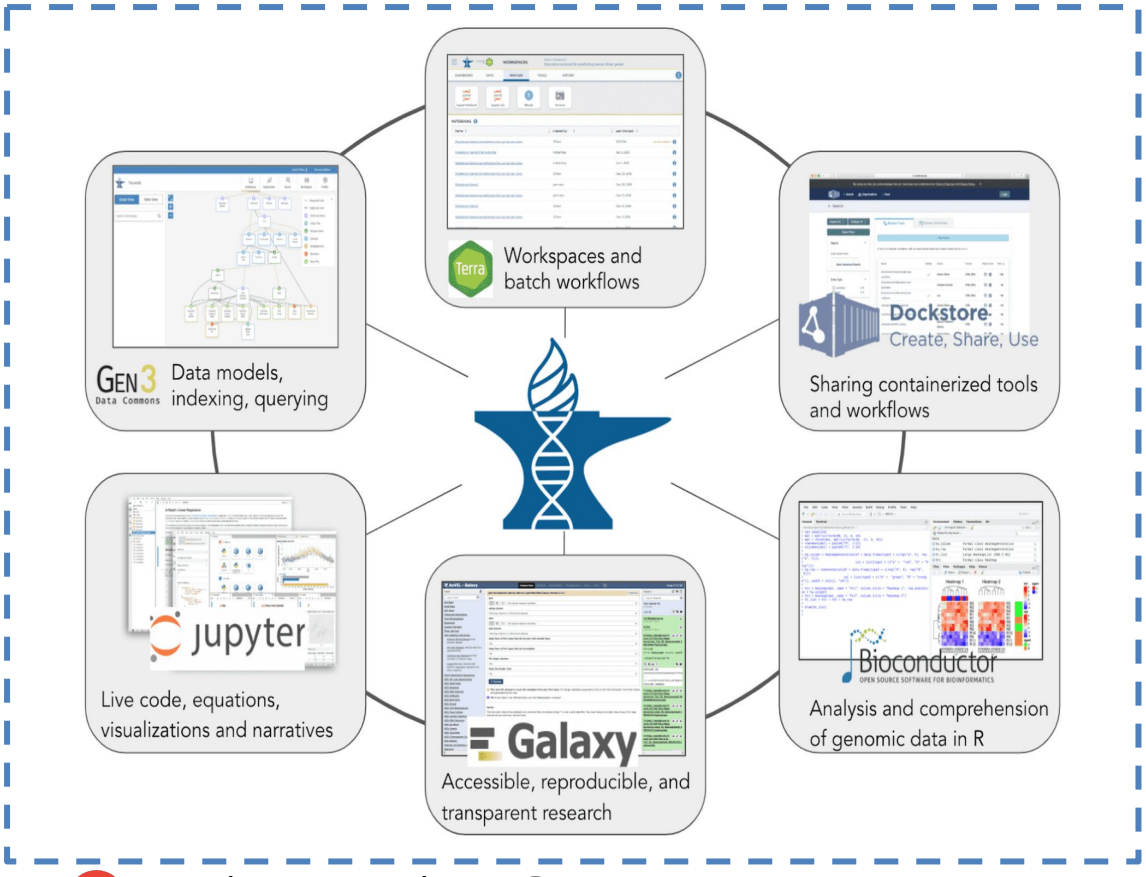
Data submission and consortia engagement		Analysis tools	
Moderators: Dr. Adam Resnick (Children's Hospital of Philadelphia) and Ms. Valentina Di Francesco (NHGRI)		*Moderators: Dr. Marylyn Ritchie (University of Pennsylvania) and Dr. Ken L. Wiley, Jr. (NHGRI)*	
12:35-12:40	Moderator introductions	12:35-12:40	Moderator introductions
12:40-12:55	AnVIL presentation: *Dr. Brian O'Connor (Broad) and Dr. Frederick Tan (Carnegie)*	12:40-12:55	AnVIL presentation: *Dr. Vincent Carey (HMS) and Dr. Ira Hall (Yale)*
12:55-1:40	Discussion	12:55-1:40	Discussion
1:40-1:50	Prepare breakout report	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

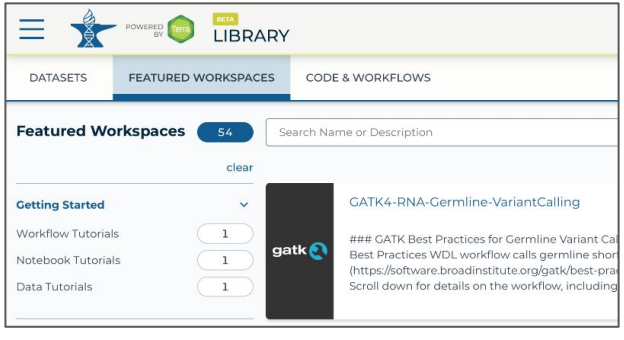
Thousands of tools provided by:

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Dockstore
Create, Share, Use

Galaxy

Featured workspaces showcase reproducible, ready to run workflows and notebooks



Metrics: Tools & Workflows

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

AnVIL

AnVIL toolsets by scale and purpose

Interactive

Large Scale / Batch

Basic Science

Clinical

RStudio

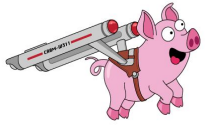
jupyter

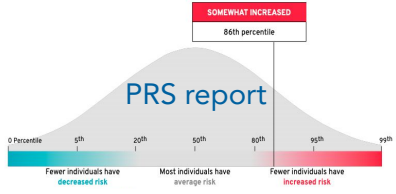
Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

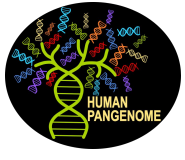
Galaxy
PROJECT

{wdl}

Dockstore
Create, Share, Use





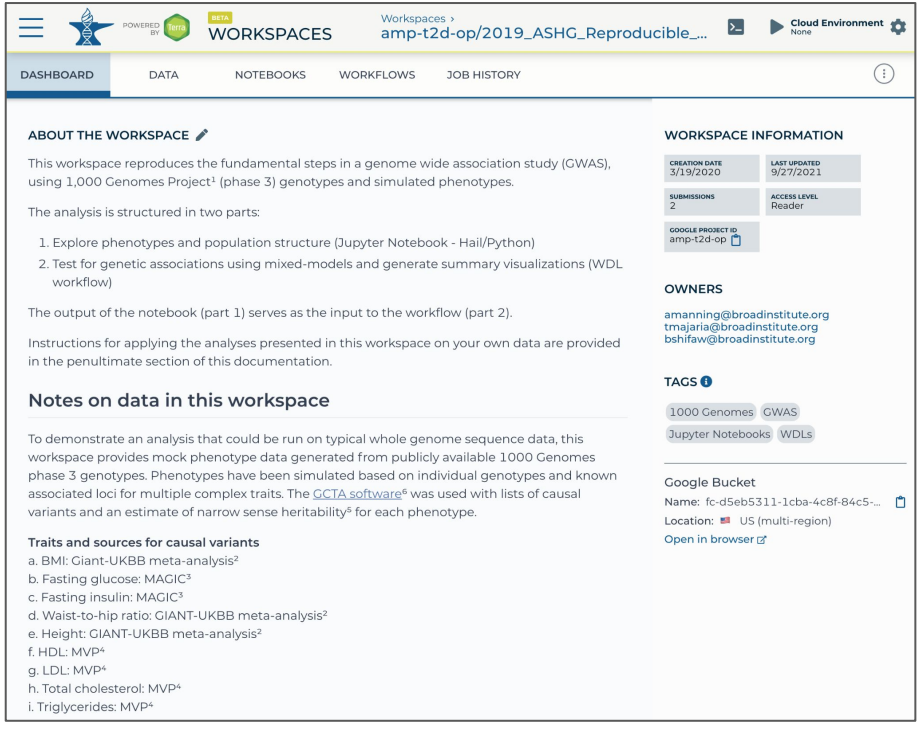


Pharmcat

seqr

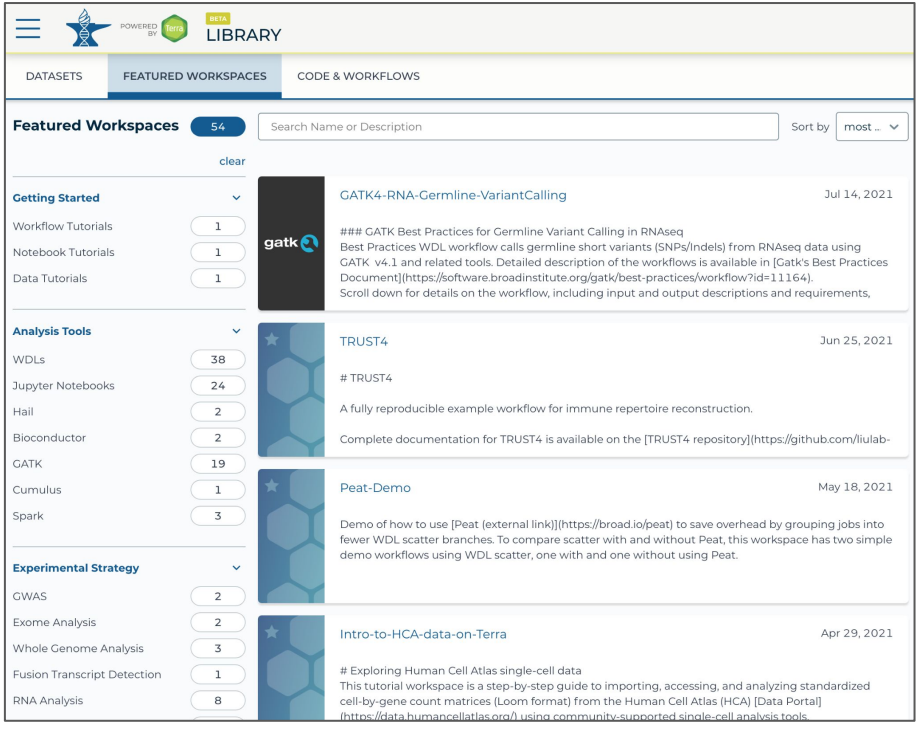
The Workspace - The fundamental unit in Terra

Workspace: the fundamental unit in Terra



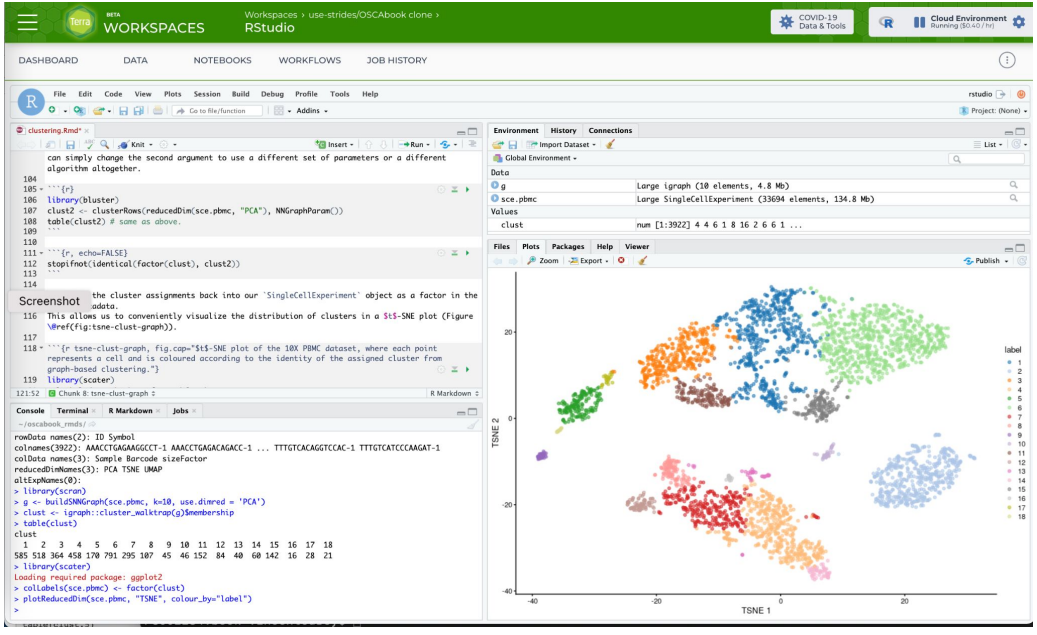
Collaboratively access and organize data, launch tools and run analyses

Featured workspaces: see how science gets done



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

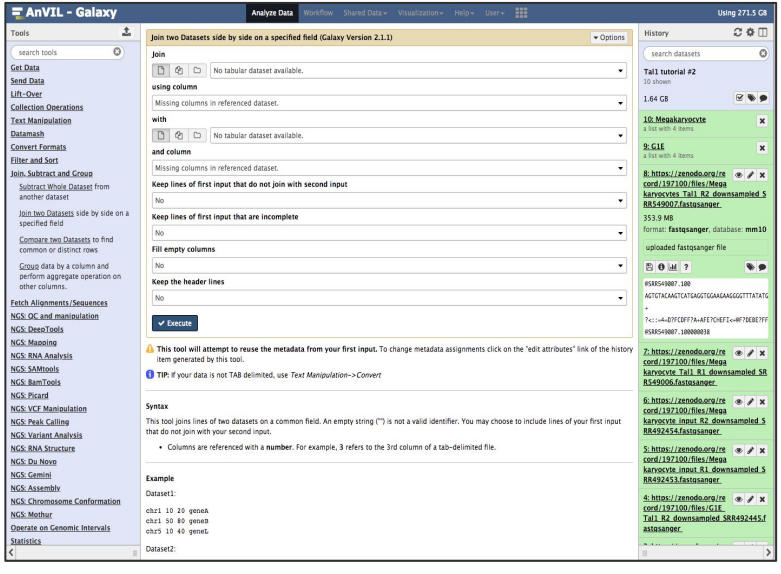
Bioconductor: tools and modules for the analysis and comprehension of high-throughput genomic data, implemented in R

- 1,903 software packages available

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

Galaxy: Cloud-scale & flexible analysis

- 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

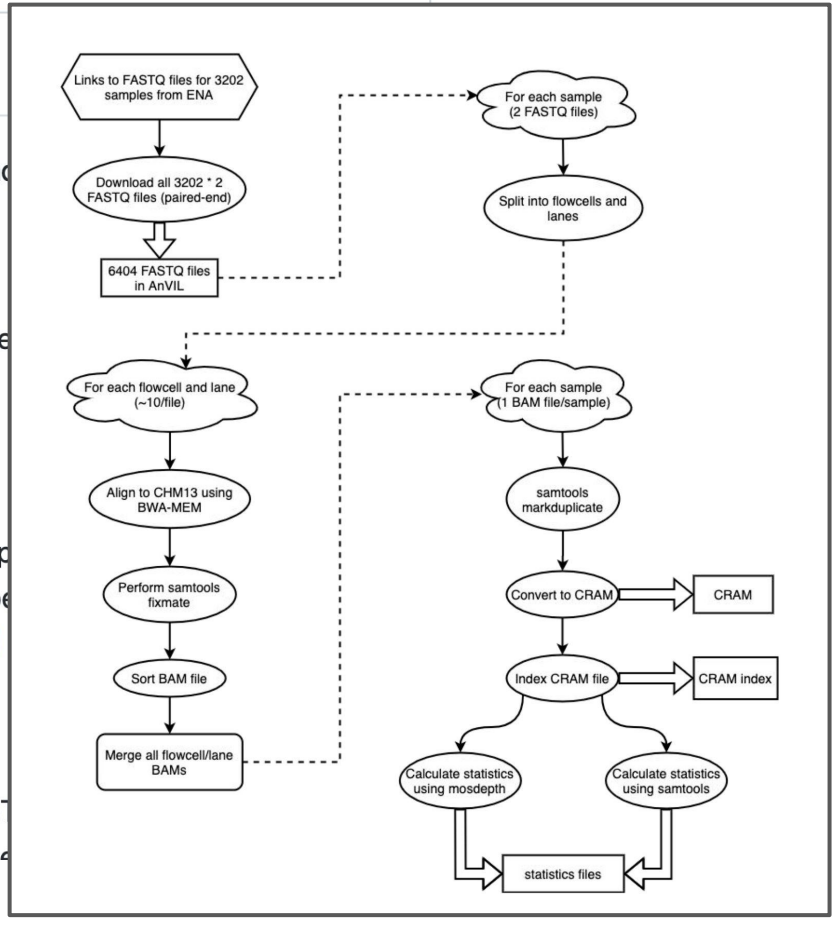
- `wdls/download_aspera.wdl` : Downloads FASTQ files from the Archive (ENA), given accession numbers

Read alignment

- `wdls/t2t_alignment.wdl` : Given a reference FASTA file, samp FASTQ files, BWA index, and dedup distance (default = 100), pe described in Aganezov, Yan, Soto, Kirsche, Zarate, *et al*. (2021)

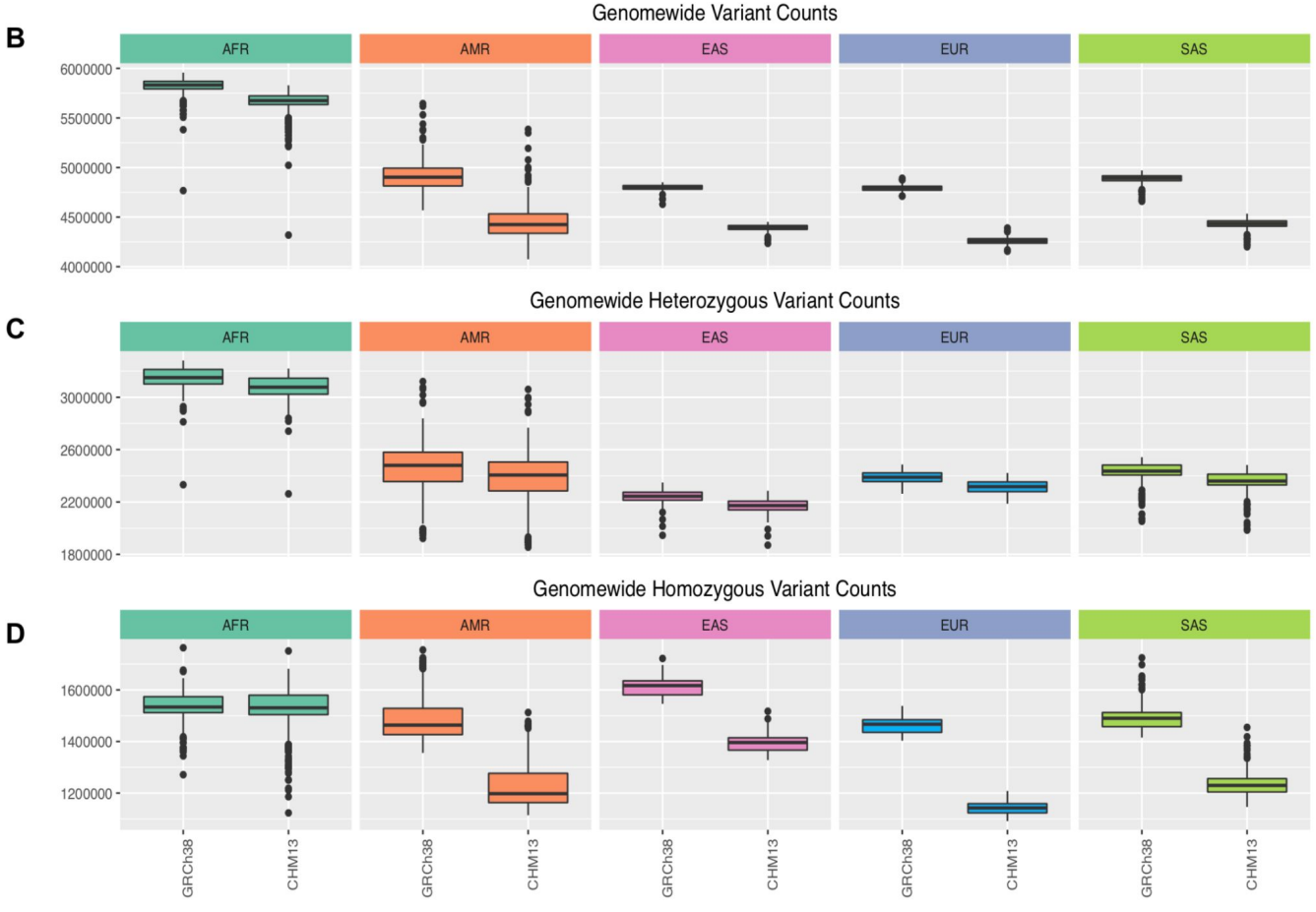
Variant calling

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



github.com/schatzlab/t2t-variants

Re-analyzing diversity in 3,202 samples



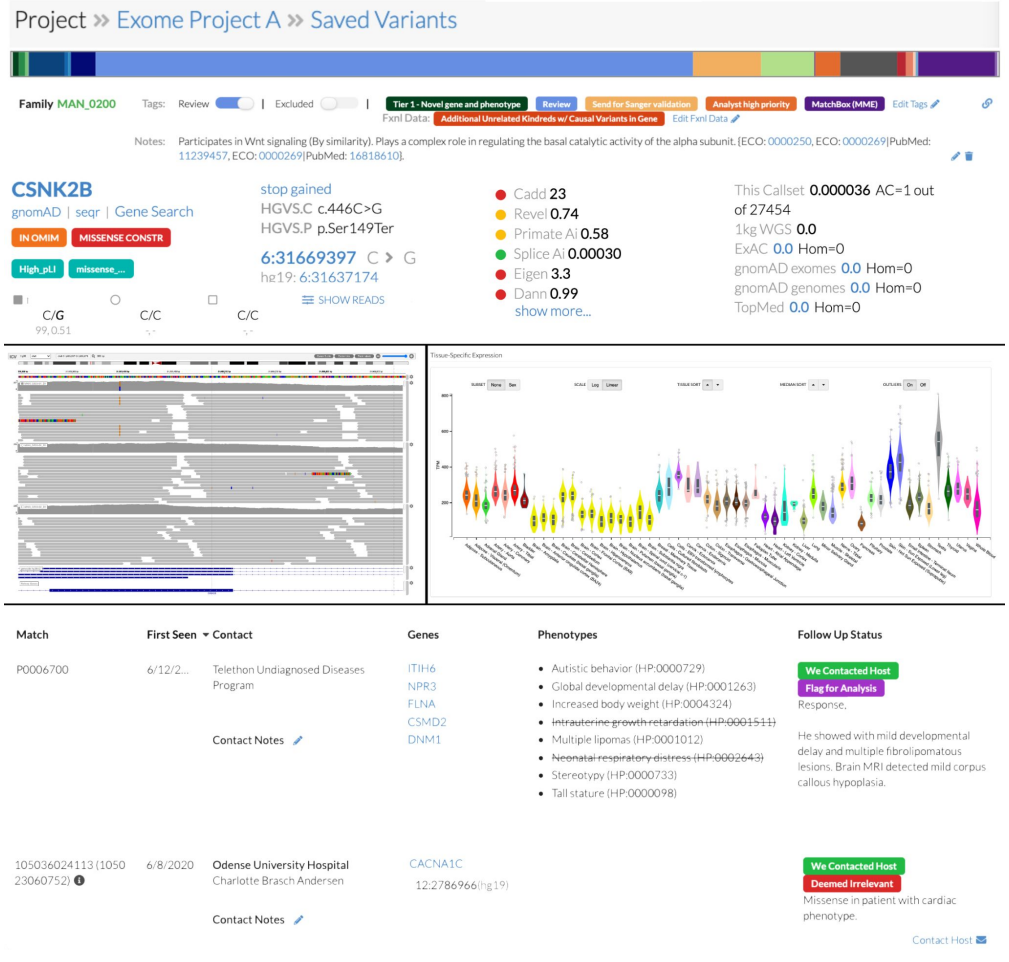
Aganezov, *et al.*, 2021 bioRxiv

Powered by AnVIL

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



Pais et al., medRxiv, 2021

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
```

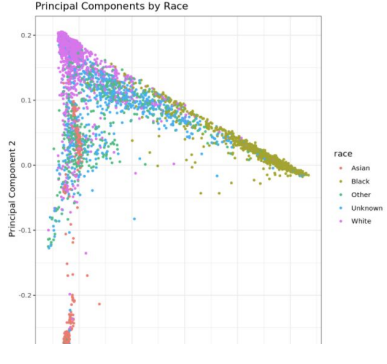
```
                    Stratified by cad
                     0             1             p      test
  n                  12446           644
  age (mean (SD))    48.28 (14.95) 63.12 (8.15)  <0.001
  male = 1 (%)        5081 (40.8)   417 (64.8)   <0.001
  dm = 1 (%)           401 ( 3.2)   190 (29.5)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

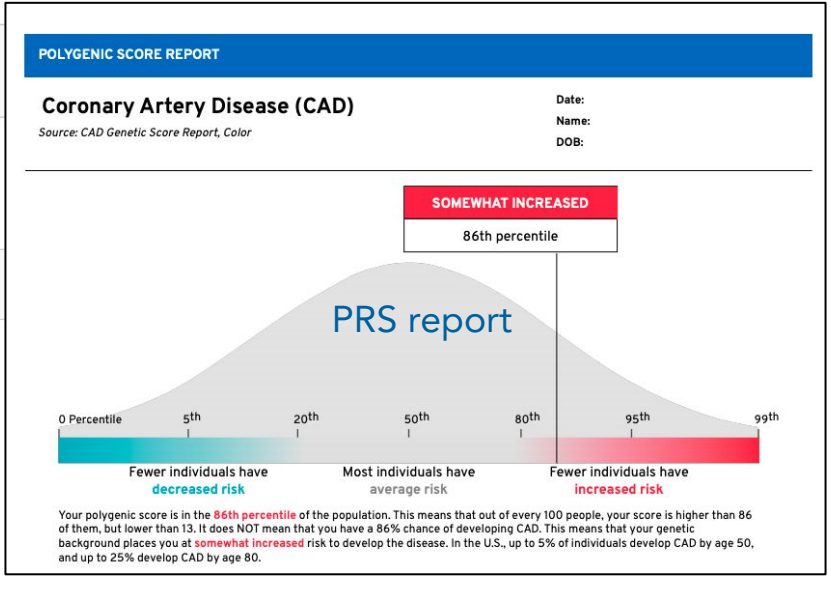
In [20]:
```
# For the genetic data, we compute genetic ancestry  using principal components
addmargins(table(dat$race))
round(prop.table(table(dat$race))*100,1)
# Our population is 80% white, 6% Black, and 4% Asian
```

```
Asian   Black   Other Unknown   White     Sum
  482     784     493     833   10498   13090

Asian   Black   Other Unknown   White
  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
```





AHA/AnVIL Working Group

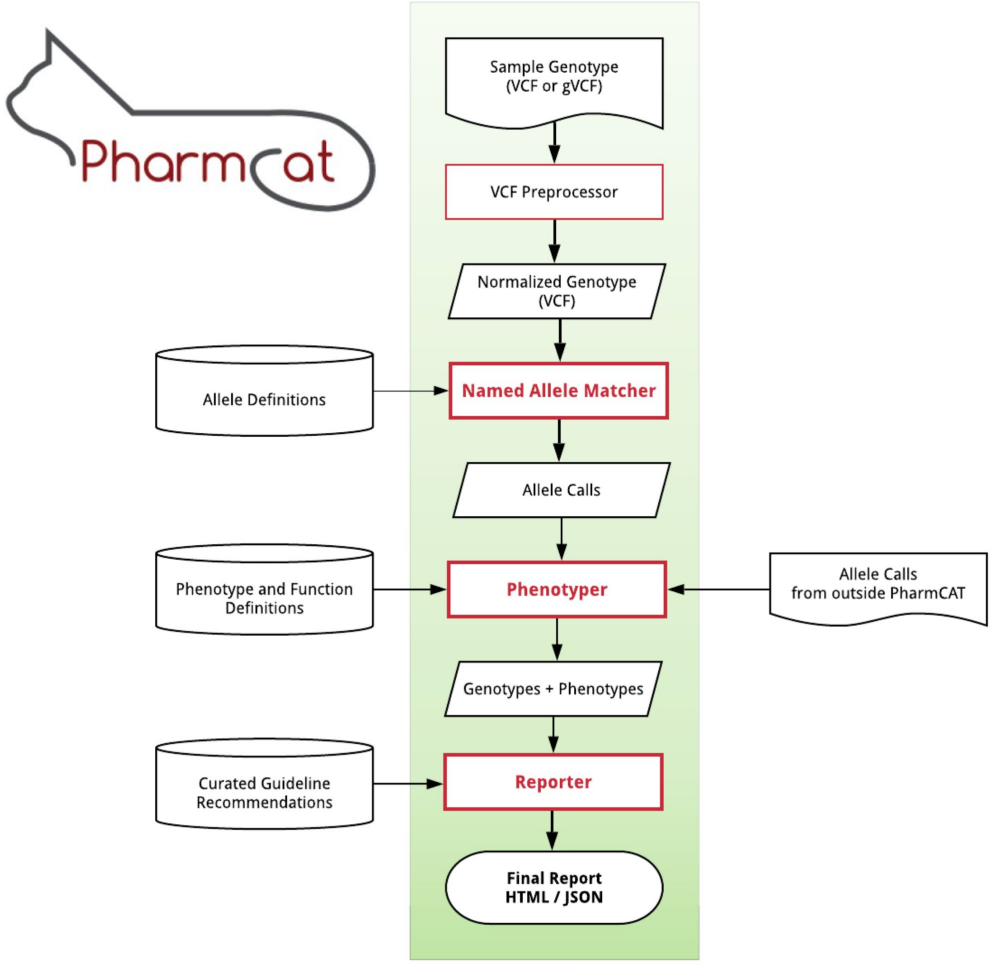
Name	URL	Brief Description
PGS Catalog	http://www.pgscatalog.org/	An open database of polygenic scores and the relevant metadata required for accurate application and evaluation.
PRS-RS	https://www.medrxiv.org/content/10.1101/2020.04.23.20077099v2	Improving reporting standards for polygenic scores in 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-funct: 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/30100085/	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.

AnVIL

- 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?

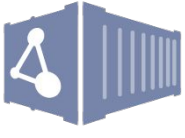
- PharmCAT is used to support clinical decision making
- Utilizes CPIC/PharmGKB guidelines for gene-drug pairs
- Gives prescription recommendations based on genetic variants

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



Dockstore
Create, Share, Use

kubernetes

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