The NHGRI Genomic Data Science Analysis, Visualization and Informatics Lab-space (AnVIL)

Michael C. Schatz[1], Anthony Philippakis[2], on behalf of the AnVIL project team[3]

[1]Departments of Computer Science and Biology, Johns Hopkins University, Baltimore, MD
[2]Broad Institute of MIT and Harvard, Cambridge, MA
[3]City University of New York, Harvard, Oregon Health & Sciences University, Penn State, Roswell Park Cancer Institute, University of California Santa Cruz, University of Chicago, Vanderbilt, Washington University.

AnVIL

http://anvilproject.org

Abstract

The traditional model of genomic data sharing – centralized data warehouses such as dbGaP from which researchers download data to analyze locally – is increasingly unsustainable. Not only are transfer/download costs prohibitive, but this approach also leads to redundant siloed compute infrastructure and makes ensuring security and compliance of protected data highly problematic.

The NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-Space, or AnVIL, inverts this model, providing a cloud environment for the analysis of large genomic and related datasets. By providing a unified environment for data management and compute, AnVIL eliminates the need for data movement, allows for active threat detection and monitoring, and provides elastic, shared computing resources that can be acquired by researchers as needed. AnVIL provides access to key NHGRI datasets, such as the CCDG (Centers for Common Disease Genomics), CMG (Centers for Mendelian Genomics), eMERGE (Electronic Medical Records and Genomics), as well as other relevant datasets.

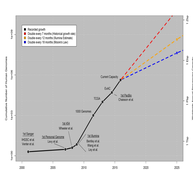
The platform is built on a set of established components that have been used in a number of flagship scientific projects. The Terra platform provides a compute environment with secure data and analysis sharing capabilities. Dockstore provides standards based sharing of containerized tools and workflows. Bioconductor and Galaxy provide environments for users at different skill levels to construct and execute analyses. The Gen3 data commons framework provides data and metadata ingest, querying, and organization.

AnVIL provides a collaborative environment for creating and sharing data and analysis workflows for both users with limited computational expertise and sophisticated data scientist users. It provides multiple entry points for data access and analysis, including execution of batch workflows written in WDL, notebook environments including Jupyter and RStudio, Bioconductor packages for building analysis on top of AnVIL APIs and services, and will offer Galaxy instances for interactive analysis. It will be possible to integrate additional analysis environments through standard APIs.

Long-term, the AnVIL will provide a unified platform for ingestion and organization for a multitude of current and future genomic and genome-related datasets. Importantly, it will ease the process of acquiring access to protected datasets for investigators and drastically reduce the burden of performing large- scale integrated analyses across many datasets to fully realize the potential of ongoing data production efforts.

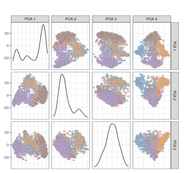
Why AnVIL?

Data



Scale,
Integration,
Sharing & Reuse

Computation

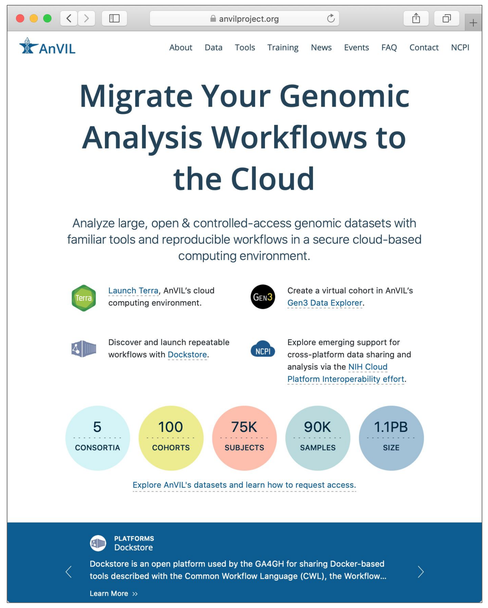


Simplicity,
Reproducibility,
Security

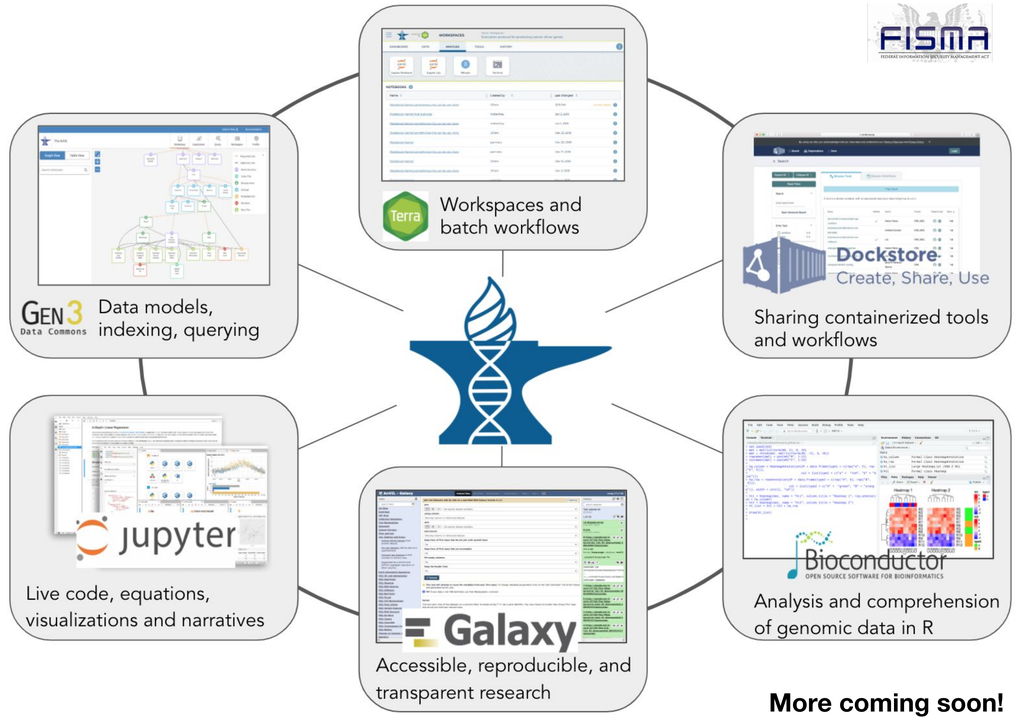
Users

Democratization,
Collaboration,
Discoveries

System Architecture



Implemented on Google Cloud Platform
Primary data storage costs covered by AnVIL,
user private data and compute billed directly through Google



More coming soon!

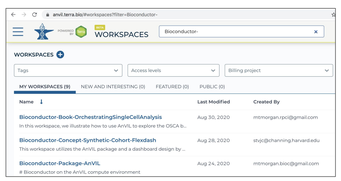
Tools Available

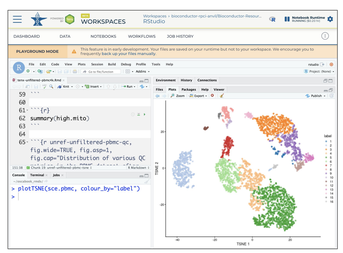
Running on Google Cloud Platform (GCP), the analysis solutions of AnVIL are scalable and manageable, uncapping previous limits of biomedical scalability. Linking billing to user accounts enables AnVIL users to monitor and scale their own analysis.

AnVIL uses the Terra analysis engine to launch it's analysis solutions. These analysis solutions are run within safe workspaces, providing regulated access to high value NHGRI data sets and protecting user loaded data as well.

AnVIL users will have access to the commonly used bioinformatic tools, including batch processing pipelines like the Workflow Description Language (WDL) using the Cromwell scaling engine all the way to downstream interactive analysis with Jupyter and R Studio with Bioconductor support.

Future additions to the AnVIL analysis ecosystem will include the Galaxy workbench, the Genome Browser support by UCSC, and additional community sourced analysis solutions.





Featured Workspaces

GATK Best Practices for Germline SNPs & Indels
This is a fully reproducible example of Processing For Variant Discovery, HaplotypeCaller(GVCF, and Joint Discovery workflows based on GATK Best Practices. Launch Workspace

GATK Best Practices for Somatic CNV Discovery
This workspace contains an example of the somatic copy number variation workflow, representing the Variant Discovery portion of the Somatic CNV Discovery pipeline. Launch Workspace

GWAS Pipeline Using Hail
This workspace serves as a basic tutorial for using Hail, a python-based package containing additional data types and methods for working with genomic data. Launch Workspace

inferCNV Tumor Single-Cell RNA-Seq Analysis Pipeline
The inferCNV workflow compares RNA from tumor samples with corresponding "normal" samples to identify evidence for copy number variations in tumors. Launch Workspace

Optimus Pipeline for Analysis of 3' Single-Cell Transcriptomic Data
The Optimus pipeline processes 3-prime single-cell transcriptome data from the 10X Genomics v2 (and v3) assay. Launch Workspace

Data Availability

Consortium	Cohorts	Subjects	Samples	Files	Size (TB)
1000 Genomes	1	3,202	3,202	3,202	52.35
CCDG	69	62,214	62,214	81,893	911.92
CMG	28	8,793	7,354	14,531	22.50
eMERGE	1	277	1	277	5.52
GTEx (v8)	1	979	17,382	41,844	157.64
Total	100	75,465	90,153	141,747	1,149.93

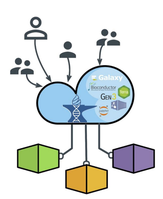
AnVIL is a repository for open and controlled access datasets. Dataset access is controlled in adherence to NIH Policy and in line with the standards set forth in the individual consents involved in each cohort.

AnVIL provides three types of data access: (1) Open Access - No restrictions; (2) Controlled Access - Access is granted by dbGaP data access; (3) Consortium Access - Datasets are accessible to consortia members under the consortium data sharing agreement.

Getting Started

This MOOC is designed to be a brief introduction to cloud computing and the AnVIL platform. Learn how this new resource can help you access data, scale your computing resources, and democratize access to genomic data science.

Genomic Analysis on AnVIL



https://leanpub.com/universities/courses/jhu/anvil-intro