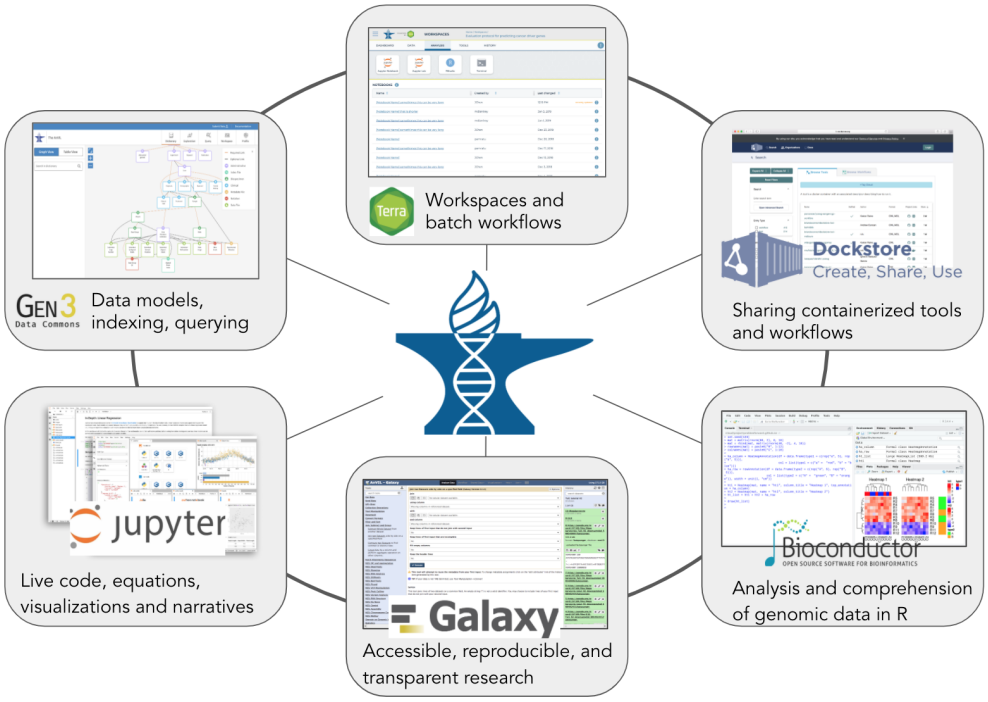
Future Direction of AnVIL Workshop

October 29, 2021



NHGRI Analysis Visualization and Informatics Lab-space

http://anvilproject.org

Table of Contents

Introduction to AnVIL **4**

Leadership Team and Organization **5**

Working Groups **8**

Summary of Accomplishments **9**

Scientific Accomplishments and Selected Publications 11

Data submission and consortia engagement **13**

Current AnVIL Data Submission Process 13

Improvements to the Data Submission Process 14

Current Consortia Engagement 15

Clinical Genomics Data Ingestion 16

Optimizing Consortium Engagements 16

Quality Assurance, Quality Control, and Data Harmonization 17

Quality Assurance 17

Quality Control 18

Data Harmonization 19

Data model 19

Gen3: Management, analysis, harmonization, and sharing of large datasets 20

Data Available in AnVIL 22

Data Ingestion Roadmap 22

Analysis tools **24**

Dockstore: Registry of Tools and Workflows 24

Jupyter Notebooks: Transparent Code, Visualizations, and Narratives 25

RStudio: Interactive Machine Learning, Statistical Computing, and Visualizations 26

Bioconductor: Community-driven Interactive Genomics with R and RStudio 26

Galaxy: Accessible, Reproducible, and Transparent Genomic Science 28

Clinical analysis tools 30

Polygenic Risk Score (PRS) 30

Seqr 31

PharmCAT 32

Extending AnVIL Tool Sets 32

Integration and Deployment of 3rd party applications 33

Development and Testing 33

Platform Security 33

Maintenance and User Support 34

Addressing Challenges in AnVIL 34

Expanding AnVIL Beyond Genomic-based analysis 35

Infrastructure **36**
Portal 36
Current Terra State 37
Terra Multi-cloud Integration 38
DUOS: Semi-automated Authorization and Management of Human Subject Data 40
NIH Cloud Platform Interoperability (NCPI) 41
Outreach and training **45**
Outreach Team Design 46
Summary of Accomplishments 47
Usage 47
Support 48
Content 50
Collaborations 55
Future Directions 57
Vision for the Next 5 Years 57
Three Complementary Teams 58
Strategic Initiatives 58

I. Introduction to AnVIL

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 the traditional 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.

The AnVIL is a federated cloud platform built from 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 standard-based sharing of containerized tools and workflows. The Gen3 data commons framework provides data and metadata ingestion, querying, and organization. Jupyter, R/Bioconductor, and Galaxy provide environments for users to construct and execute analyses using a diverse set of tools for basic and clinical research.

The AnVIL platform provides a collaborative environment for creating and sharing data and analysis workflows for both users with limited computational expertise and sophisticated data scientist users alike. AnVIL 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 Galaxy instances for interactive analysis. It is also possible to integrate additional analysis environments through standard APIs. Finally, and perhaps most importantly, 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), GTEx v8 (Genotype-Tissue Expression Project), as well as other relevant datasets.

The AnVIL is also a founding member of the NIH Cloud Platform Interoperability (NCPI) Effort. NCPI's goal is to establish and implement guidelines and technical standards to empower end-user analyses across participating cloud platforms and facilitate the realization of a trans-NIH, federated data ecosystem. Among other goals, these efforts make it possible to easily compare datasets stored in the AnVIL with datasets stored in other NIH cloud platforms, e.g. genetic variants or gene expression patterns from adult donors in AnVIL/CCDG and AnVIL/GTEx compared to those from pediatric cancer patients stored in the KidsFirst CAVATICA platform. The AnVIL has been a key partner in developing the enabling technologies for this include the NIH Researcher Authentication System (RAS) to manage single sign-on and user access credentials across platforms, the GA4GH Data Repository Service (DRS) to allow data objects to be referenced in a single, standard way regardless of where they are stored, the Fast Healthcare Interoperability Resources (FHIR) for describing data and an API for exchanging electronic health records (EHR), and The Portable Format for Bioinformatics (PFB) to bring search results and synthetic cohorts into workspace environments.

II. Leadership Team and Organization

AnVIL is funded through cooperative agreement awards to the Broad Institute (#5U24HG010262) and Johns Hopkins University (#5U24HG010263) with additional partners at the University of Chicago, Roswell Park Cancer Institute, the University of California at Santa Cruz, Penn State University, Washington University, Oregon Health and Sciences University, Harvard Medical School, Vanderbilt University, City University of New York, and Yale University. Below is the list of investigators and co-investigators from these institutions.

The Broad Institute of MIT and Harvard

- Anthony Philippakis, Co-Program Director

  Dr. Philippakis is the chief data officer of the Broad Institute of MIT and Harvard, co-director of the Eric and Wendy Schmidt Center, and Program Director of the Broad and Broad-associated teams. He trained as a cardiologist at Brigham and Women's Hospital, with a focus on rare genetic cardiovascular diseases. At the Broad Institute, he is the founding director of the Data Sciences Platform, an organization of over 200 software engineers and computational biologists that develops software for analyzing genomic and clinical data. In addition to his roles at the Broad Institute and Brigham and Women's Hospital, Philippakis is a venture partner at GV, focusing on machine learning, distributed computing, and genomics. Philippakis received his M.D. from Harvard Medical School and completed a Ph.D. in biophysics at Harvard. As an undergraduate, he studied mathematics at Yale University and later completed the Part III (equivalent to M.Phil) in mathematics at Cambridge University.

- Brian O'Connor, Co-Investigator
  Dr. O'Connor joined the Broad in March of 2020 as a Principal Investigator in the Data Sciences Platform (DSP). He works on a wide variety of cloud compute and interoperability projects in the DSP including Terra, NCI's Cancer Data Aggregator, and NHGRI's AnVIL. While at the Ontario Institute for Cancer Research (OICR), Dr. O'Connor created the Dockstore project, a platform for sharing tools and workflows used widely in the community including within AnVIL. Dr. O'Connor serves as co-chair of the AnVIL Technical working group and co-chair of the NCPI Systems Interoperability working group where he works on interoperability standardization across the NIH. Dr. O'Connor is also the co-chair of the Global Alliance for Genomics and Health (GA4GH) Cloud Work Stream where he works on cloud standards for the larger genomics community.

- Anne O'Donnell Luria, Co-Investigator
  Dr. Luria is a member of the Clinical Genomics Working Group and a member of the NHGRI Center for Mendelian Genomics.

Johns Hopkins University

- Michael C. Schatz, Co-Program Director
  Dr. Michael Schatz serves as the Program Director of the JHU and JHU-associated teams. He has deep expertise in bioinformatics, comparative genomics, advanced sequencing technology, and high-performance computing. He began his career in genomics at the Institute for Genomic Research in

2001, completed his Ph.D. in computer science and bioinformatics at the University of Maryland (2005-2010), and began his faculty career at Cold Spring Harbor Laboratory in 2010. In 2016, he moved into his current position as Bloomberg Distinguished Professor of Computer Science and Biology at Johns Hopkins University. Schatz is a key thought leader in cloud computing and genomics, starting in 2009 when Schatz authored the first paper in PubMed to use cloud computing for genomics for an application called CloudBurst (Schatz, Bioinformatics, 2009). Since 2018, Michael Schatz has been named a Web of Science Highly Cited Researcher (Top 1% of all published researchers). Dr. Schatz now chairs and is an active participant in several working groups, as well as oversees the entire JHU-based team.

- Jeffrey Leek, Investigator
  Dr. Leek serves as co-chair of the AnVIL Outreach working group. Jeff is a professor of Biostatistics and Oncology at the Johns Hopkins Bloomberg School of Public Health and co-director of the Johns Hopkins Data Science Lab. His group develops statistical methods, software, data resources, and data analyses that help people make sense of massive-scale genomic and biomedical data. As the co-director of the Johns Hopkins Data Science Lab he has helped to develop massive online open programs that have enrolled more than 8 million individuals and partnered with community-based non-profits to use data science education for economic and public health development. He is a Fellow of the American Statistical Association and a recipient of the Mortimer Spiegelman Award and Committee of Presidents of Statistical Societies Presidential Award. He co-chairs the Outreach working group and coordinates the efforts of the broader Outreach Team.

- Enis Afgan, Investigator
  Dr. Afgan serves as a key driver of Galaxy on AnVIL.

- Casey Overby Taylor, Investigator
  Dr. Taylor serves as a key driver of Clinical Genomics on AnVIL.

- Kasper Hansen, Investigator
  Dr. Hansen serves as a key driver of Bioconductor on AnVIL.

The University of Chicago

- Robert Grossman, Investigator
  Dr. Grossman oversees all AnVIL activity at the University of Chicago, especially the development of Gen3 and co-leads the NCPI Systems operability working group. Dr. Grossman is the Frederick H. Rawson Professor of Medicine and Computer Science. He is also the Jim and Karen Frank Director of the Center for Translational Data Science, Co-Chief of the Section of Computational Biomedicine and Biomedical Data Science in the Department of Medicine, and the Chief Research Informatics Officer for the Biological Sciences Division at the University.

Roswell Park Cancer Institute

- Martin Morgan, Investigator
  Dr. Morgan serves as a key driver of RStudio and Bioconductor on AnVIL.

University of California at Santa Cruz

- Benedict Paten, Investigator
  Dr. Paten oversees all AnVIL activity at UCSC, especially the development of Dockstore and the Portal working group team.

Penn State University

- Anton Nekrutenko, Investigator
  Dr. Nekrutenko serves as a key driver of Galaxy on AnVIL and as a chair for NIH Cloud Platform Interoperability (NCPI) Outreach working group.

Washington University

- Ting Wang, Investigator
  Dr. Wang serves as a key driver of the AnVIL Data Processing Group.

Oregon Health and Sciences University

- Jeremy Goecks, Investigator
  Dr. Goecks serves as a key driver of AnVIL APIs.

- Kyle Ellrott, Investigator
  Dr. Ellrott serves as a key driver of AnVIL APIs.

Harvard Medical School

- Vince Carey, Investigator
  Dr. Carey serves as a key driver of RStudio and Bioconductor on AnVIL.

Vanderbilt University

- Robert Carroll, Investigator
  Dr. Carroll is the co-lead of the Phenotype Working Group and co-lead of the NCPI FHIR working group.

City University of New York

- Levi Waldron, Investigator
  Dr. Waldron serves as a key driver of RStudio and Bioconductor on AnVIL.

Yale University

- Ira Hall, Investigator
  Dr. Hall serves as a co-lead of the Data Processing Group.

Carnegie Institution for Science

- Frederick Tan, Investigator
  Dr. Tan serves as a co-chair for the AnVIL Outreach working group.

Working Groups

Day-to-day operations within AnVIL are organized around 8 major working groups (**Table 1**). Each of these groups is chaired by 2 AnVIL team members and participation typically varies from approximately 8 to 35+ additional team members. It is notable that several of the working groups are co-chaired by one member from each of the two AnVIL awards. A summary of the progress of the working groups is distributed on a monthly basis, with contributions from all teams. Overall, there is very close communication and interactions between the teams, with extensive joint technical developments, planning, outreach, and other activities. A notable example of this collaboration has been the Galaxy/AnVIL deployment in which Galaxy Team members at JHU, Penn State, and OHSU directly participated in the Terra sprint reviews at Broad to discuss requirements and technical progress for the new Kubernetes based deployment.

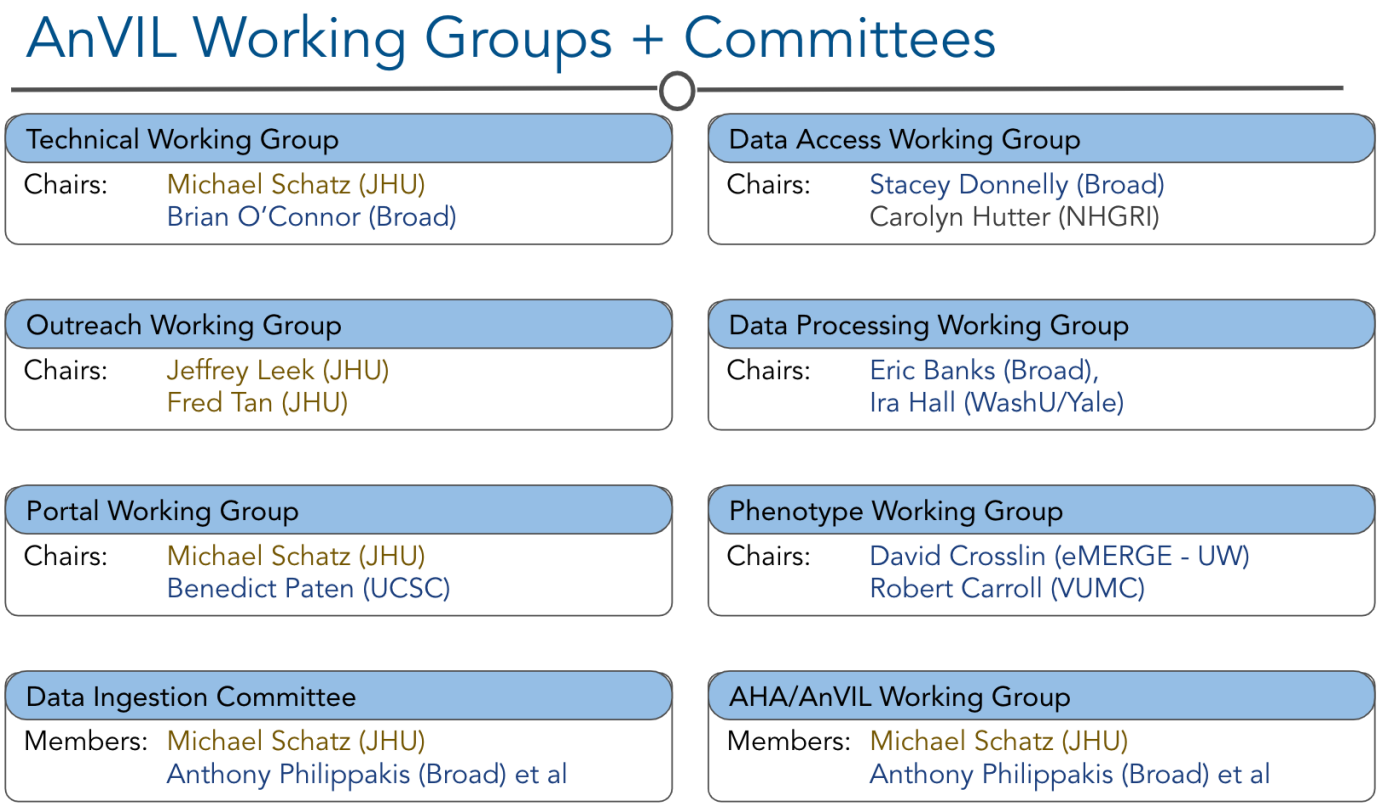


Table 1. Summary of AnVIL Working Groups.

III. Summary of Accomplishments

AnVIL launched in the fall of 2018 and has substantially matured in the number of tools, datasets, users, and features available. Our major accomplishments include:

- ***The ingestion of over 300,000 samples (3.87Pb of data) into AnVIL*** spanning 7 major consortiums: 1000 Genomes, CCDG, CMG, eMERGE, GTEx, HPRC, and T2T. We are currently in the process of harmonizing the QC and variant calling for these datasets, which will allow for the development of more accurate variant calling, more accurate allele frequency calculation & imputation, and ultimately improved power to discover variants associated with disease. Over the past year, we have begun to engage with and onboard 30 studies including the aforementioned consortiums. The ingested cohorts can be interactively searched and explored in our enhanced AnVIL data dashboard https://anvilproject.org/data.

- ***The deployment of Terra, for workspaces, interactive and batch computing.*** The functional unit of Terra is the workspace, each equipped with a Google Cloud bucket where data generated by a workflow analysis and notebook files are stored by default. Within a workspace, users can launch batch analysis jobs or one of several interactive computing environments, including Jupyter Notebooks, R/Bioconductor, and Galaxy. There are currently 273 public workspaces and 47 featured workspaces demonstrating a variety of widely used analysis tasks.

- ***The deployment of Gen3 within AnVIL***, for the management, analysis, harmonization, and sharing of large datasets. This brings new capabilities to search, explore, and develop synthetic cohorts from the tens of thousands of samples that are already loaded into AnVIL, thus increasing the value of the investment NHGRI and NIH have already made into generating these data.

- ***Release of Galaxy within AnVIL/Terra***, bringing nearly ten thousand genomics analysis tools into AnVIL within an easy to use graphical user interface. We presented the first public release of this capability at ASHG 2020 where we demonstrated Galaxy running within AnVIL to perform a GWAS analysis on human variant calls.

- ***Enhanced capabilities for Dockstore*** to share, explore, and manage reproducible workflows in the widely used Workflow Description Language (WDL), Common Workflow Language (CWL) and Galaxy Workflow specifications. This currently houses 1,040 reproducible tools and workflows for genome assembly, variant discovery, transcriptome analysis, and a variety of related tasks. Dockstore also greatly simplifies the process for researchers to deploy new tools and workflows within AnVIL.

- ***Enhanced capabilities for Jupyter notebooks*** to utilize persistent disks so that analysis code and results will be more robust to system failures and user disconnections.

- ***Deployment of RStudio available within Terra.*** The infrastructure is built to support current versions of R / Bioconductor, and adopts the 'all of Bioconductor' containerization strategy we use in Jupyter notebooks.

- ***The development and deployment of the Bioconductor AnVIL packages*** to enhance the user experience of AnVIL from within R, allowing programmatic access to all elements of the Terra and Gen3 environments (e.g., tables, buckets, cloud utilities for resource management). This builds on the 2,042 other software packages available through Bioconductor that are available in AnVIL.

- ***AnVIL's Portal*** (http://anvilproject.org) has been established to serve as a "meta-portal" to each of the AnVIL components. It also hosts the data dashboard (http://anvilproject.org/data) and a variety of training guides, FAQs, and other resources to support PIs, researchers, and other analysts to use the AnVIL for basic sciences and clinical research. Furthermore, the NCPI website has been deployed and hosted on the AnVIL portal at https://anvilproject.org/ncpi.

- ***Several successful outreach events*** reached thousands of participants each from the NHGRI Genome Sequencing Program, the Bioconductor community, ISMB, BOSC, GCC, and beyond. We recently launched AnVIL Outreach Office Hours as a means for AnVIL researchers to discuss their issues.

- ***We launched a new Genomic Data Science Community Network*** that will help to democratize access to AnVIL through strategic outreach to support educators and researchers at Historically Black Colleges and Universities (HBCUs), Minority Serving Institutions (MSIs), Tribal Colleges and Universities (TCUs), and Community Colleges (CCs). We currently have partnered with 27 faculty members across these diverse institutions.

- ***Developed an initial catalog of clinical genomics tools*** in collaboration with the American Heart Association (AHA) that is being used for prioritizing the deployment of clinically-oriented tools within AnVIL. These include tools for assessing the pathogenicity of individual variants, polygenic risk score calculators, pharmacogenomics-related analysis tools, and other clinically relevant capabilities. We have also integrated seqr as an analysis platform for Mendelian diseases.

- ***The launch of major efforts through the NIH Cloud Program Interoperability (NCPI) program*** to increase usability across cloud platforms through seamless user authentication (RAS), a flexible generic interface to data repositories (DRS), and initial support for standards and APIs for exchanging electronic health records (FHIR). AnVIL also hosts the newly established NCPI web presence at https://anvilproject.org/ncpi/. Furthermore, researcher use cases were developed within the NCPI effort which span the four NCPI stacks (AnVIL, BD Catalyst, CRDC, and Kids First). These are being continuously refined and are designed to guide interoperability development.

- ***The piloting of the Data Use Oversight System (DUOS)*** across six NIH ICs to semi-automate and efficiently manage compliant sharing of human subjects data. This will substantially streamline the access of controlled data to authorized researchers applying for access to these data.

- ***Egress-free Release of the GTEx V8 Protected Data*** within a secure environment that allows free download of the data to authorized users. This will enable authorized users to avoid the ~$15,000 egress fees that are currently required to download the raw protected data. We currently estimate at least 37 groups have downloaded these data, saving over $500,000 in egress fees in the last year.

Scientific Accomplishments and Selected Publications

- The AnVIL leadership has written a perspective manuscript available in bioRxiv (posted April 23, 2021) that is currently under review at Cell Genomics (Schatz, Philippakis, *et al*, 2021, bioRxiv). This manuscript documents the broader scientific achievements and contributions of the AnVIL project.

- The Dockstore team has published an article outlining its community platform for sharing reproducible and accessible computational protocols (Yuen *et al*, 2021, Nucleic Acids Research).

- The Galaxy team has published a paper detailing accessible, reproducible and collaborative biomedical analyses (Jalili *et al*. 2020, Nucleic Acids Research). Additional information and a tutorial for running Galaxy on Terra within AnVIL is available on the AnVIL Portal.

- The Bioconductor team has published an article detailing single-cell analysis with Bioconductor (Amezquita *et al*, 2020, Nature Methods). The code for this analysis is published to an AnVIL/Terra workspace using the Bioconductor AnVILPublish package.

- As an example of how other researchers are building on the AnVIL platform, researchers at the Baylor College of Medicine and colleagues published a structural variant analysis algorithm called Parliament2 (Zarate *et al*, 2020, GigaScience). The algorithm is a consensus SV framework that leverages multiple best-in-class methods to identify high-quality SVs from short-read DNA sequence data at scale. The manuscript discusses how the algorithm is available in multiple forms, including as a WDL for use on the AnVIL. This WDL is now being used with the TopMed Consortium to identify SVs in more than 100,000 human genome datasets.

- AnVIL Cloud Credits awardee Dr. Tychele Turner and her research group have published a paper detailing the development of ACES (Analysis of Conservation with an Extensive list of Species, Padhi *et al*, 2021, Bioinformatics), a computational workflow allowing users to query DNA elements of interest, such as enhancers promoters or exons, and returns BLAST hits of reference genomes, a multiple sequence alignment file, a graphical fragment assembly file, and a phylogenetic tree file.

- The Telomere-to-Telomere (T2T) Consortium has made extensive use of AnVIL to assess human genetic variation across thousands of globally diverse genomes. Their new T2T-CHM13 human reference genome includes gapless assemblies for all 22 autosomes plus Chromosome X, corrects numerous errors, and introduces nearly 200 million bp of novel sequence containing 2,226 paralogous gene copies, 115 of which are predicted to be protein coding (Nurk et al, 2021, bioRxiv). The newly completed regions include all centromeric satellite arrays and the short arms of all five acrocentric chromosomes, unlocking these complex regions of the genome to variational and functional studies for the first time. Importantly, the T2T consortium has demonstrated that the T2T-CHM13 reference genome universally improves the analysis of human genetic variation through the alignment and reprocessing of 3,202 short read datasets within the AnVIL (Aganezov *et al*, 2021, bioRxiv). These results also informed the

analysis is several of the T2T companion papers, including a discussion within the overall description of the T2T-CHM13 reference genome (Nurk et al, 2021, bioRxiv) and a discussion of variation within satellite regions and other repetitive regions in the newly resolved components of the reference genome (Altemose et al, 2021, bioRxiv); (Hoyt et al, 2021, bioRxiv).

- As a second example of consortiums using AnVIL, the Centers for Mendelian Genomics (CMG) recently published a perspective on the status of the project to the medRxiv preprint server (Baxter *et al.* 2021, medRxiv). The publication documents how the CMGs have deposited over 15,025 exomes and 707 genomes to 39 AnVIL workspaces along with accompanying metadata for each sample, including sample-, subject-, family-, discovery- and sequence-level information. The manuscript also provides recommendations for accessing CMG data through AnVIL, and highlights how data sharing empowers and expedites solving rare disease.

- A third example of the scientific work in progress on the AnVIL is a recent commentary from the Human Pangenome Reference Consortium (HPRC) discussing the need for a Human Pangenome Reference Sequence (Miga and Ting, 2021, Annual Review of Genomics and Human Genetics). Crucially, the manuscript explicitly names the AnVIL as the platform for sharing these results with the wider scientific community. Already, the phase I dataset consisting of data from 30 human samples sequenced with multiple short and long read datatypes is already available in the AnVIL.

We anticipate many future publications in the near future as more and more researchers come to rely on the AnVIL to support their research.

IV. Data submission and consortia engagement

As of October 1, 2021, 291,301 subjects spanning 3.9 petabytes of data have been ingested into AnVIL and stored within Terra Workspaces, a 3X-fold increase since October 2020 (**Figure 1**). The vast majority of these samples are stored as whole genome sequencing data sequenced with the Illumina sequencing platform to approximately 30x coverage. A variety of QC metrics are computed for each dataset, and when available, phenotypic data associated with each sample are organized using harmonized project specific ontologies.

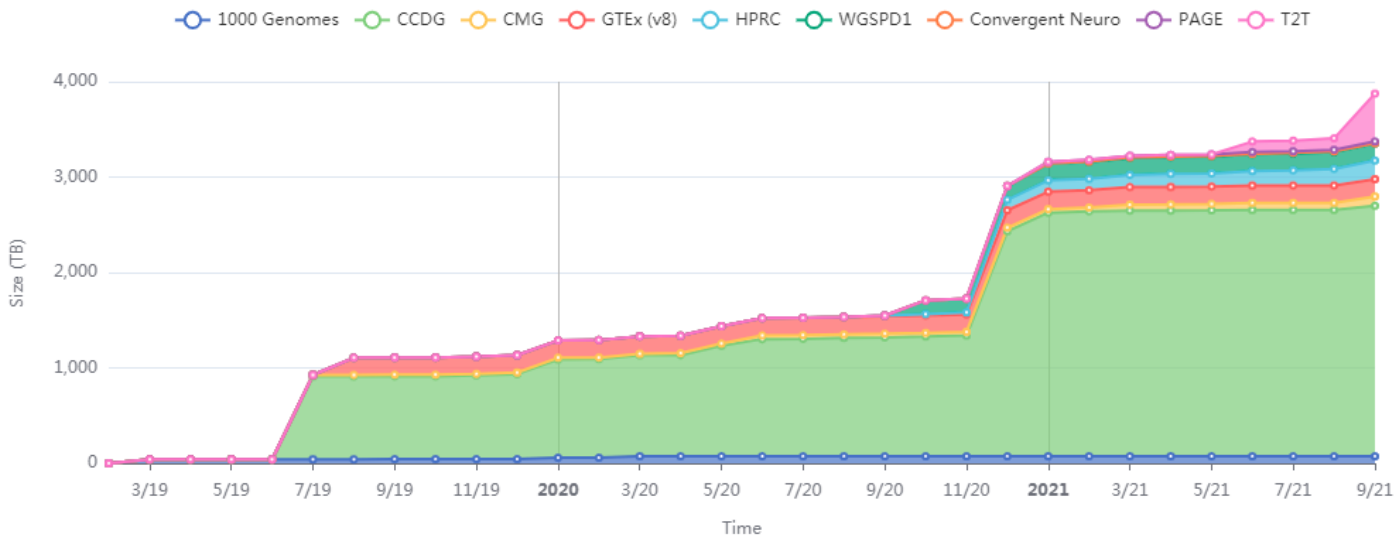


Figure 1. Volume of Data Ingested into the AnVIL over the past 2 years.

Current AnVIL Data Submission Process

Once a consortium/group is approved to deliver data to AnVIL, the data submission/ingestion team presents a kick-off meeting with the consortium/group providing an overview of AnVIL capabilities and highlighting the data ingestion process. The data submission/ingestion team then schedules bi-weekly followup meetings with the constoria to review AnVIL requirements for data submission, track progress, and answer any data submission / ingestion questions.

The current data submission path is as follows:

1. ***Register with dbGaP/Obtain required approvals*** - There are a few pathways in which a consortia/group can come into AnVIL where it was either explicitly stated in grant that AnVIL is the data repository or the investigator reached out to their program officer to seek approval to index their controlled-access data into AnVIL. Once it's approved that the data can be stored in AnVIL, the team reaches out to begin communication between the ingestion team and the submitters as described above.

2. ***Set up data model*** - Two common data models have been used for most studies submitting data into AnVIL - the rare disease data model and the common disease data model. The base of both models

includes subject, sample, and assay metadata tables. For researchers working in rare disease genomics, two additional tables are part of the model that stores familial structure (family table) and variants of interest (discovery table). Moving forward, AnVIL is adopting a minimum data model in an effort to ensure all AnVIL data is useful to the shared research community.

3. ***Prepare data for submission*** - In addition to preparing the data files for the data table, the genomic object files also need to be organized and prepared for submission. Genomic object files need to be accompanied by an object file metadata table (assay table). Object files need to follow the AnVIL file naming convention. To maximize the value of AnVIL-hosted data and minimize batch effects in cross-project analyses, AnVIL strongly encourages the submission of functionally equivalent-compliant genome and exome sequencing data aligned to GRChB38.

4. ***Ingest data into AnVIL*** - Once data and corresponding data tables are prepared, data submitters will deposit their datasets into AnVIL-owned Google Cloud buckets.

5. ***QC ingested data*** - AnVIL Data Processing Working Group has created a genomic evaluation tool for whole genome data available to the research community on Dockstore. Quality control metrics for genome and exome sequencing, with editable thresholds, are generated from this tool.

Improvements to the Data Submission Process

The AnVIL team has recognized the need for automation, self-service, and data wrangling to improve the data submission and ingestion process. To support these improvement aims, the AnVIL data submission/ ingestion team, in collaboration with the data processing, phenotype, and API teams, have created a critical path for data submission and planned the following improvements:

1. ***Creating standard instructions for data submission*** - To align data submitters to improved AnVIL data submission processes, we created a Data Submitters Persona on AnVILproject.org and outlined all steps required to deposit data into AnVIL. To accompany the automated data submission process and minimize the burden on data submitters, the AnVIL data submission/ingestion team, in conjunction with the Broad User Education team, created standard instructions that are displayed in the template workspace created by the automated process described below. These instructions will also be available as a Terra support article. To accompany the article, the two teams will work together to create videos that will help provide additional support on running the gsutil command line and the QC notebook (as described in step 3).

2. ***Automating the data submission process*** - AnVIL submission/ingestion team created automated tooling to programmatically create workspaces, authorization domains, and template workspaces for data submitters.

3. ***Providing data submission quality assurance*** - In addition to automating the data submission process, we have, in collaboration with the AnVIL Phenotype and AnVIL Apps/API teams, created Jupyter notebook based submission validation tools allowing data submitters to check the deposits they have made for completeness and compliance to AnVIL data submission requirements.

4. ***Creating a standard data model*** - To improve data FAIRness, we have adopted a standard data model for AnVIL based on the Terra interoperability model. This conceptual model is the minimum amount of data required for all data submissions, regardless of data type. Data that was ingested under a different data model will be reviewed and transformed to map to the AnVIL data model. To complement the data model, we also created a standard data model dictionary and templates for data submitters to review.
    - Subject, Sample, Family, Discovery data tables required
    - Model addresses many relationships for subjects with samples, discoveries, and phenotypes
    - Use of concept codes that link to established ontologies (HPO, OMIM, UBERON) to enhance use with other datasets and promote standardization
    - Generalizable to other genomic disease research datasets coming in to AnVIL

5. ***Developing data ingestion pipeline*** - The aforementioned data submission process was the first iterative step in creating a data ingestion pipeline that strongly and efficiently automates the data submission from start to finish, including automated QC. The Broad is developing a data ingestion team that will develop this pipeline and provide support and feedback to data submitters.

6. ***Supporting self-service*** - Feedback from our data submitters indicates a self service option in the AnVIL UI would allow efficient data submission. Therefore, we imagine a future in which the data submitter can initiate data submission in the AnVIL Ul which would be connected to the data ingestion pipeline and automated QC.

Current Consortia Engagement

Current constoria engagement begins at the time of consortia application and onboarding. The AnVIL data submission/ingestion team meets with newly approved consortia to define AnVIL requirements, inform consortia data policy, track data submission progress, deliver timelines for data ingestion, and review any data issues found in the submission process. During onboarding, the AnVIL Outreach team is introduced as a learning resource for consortia members. Many of our consortia members are participating in AnVIL Outreach activities including AnVIL Cloud Credits Program, AnVIL Deep Pilots, and AnVIL Office Hours which allows AnVIL teams to collect feedback from new and existing users on how they are using the platform and how we can improve AnVIL to make their research easier. In addition to these hands-on engagements, the AnVIL Data Ingestion, Processing, and Phenotype Working Group meetings have been opened to include consortia members. Consortia members are encouraged to attend the working group meetings to inform future development and ask questions regarding that particular step of the AnVIL data process.

Over the past two years, the AnVIL data submission/ingestion team has been working with members of the new and ongoing consortia to prepare, review, and ingest data generated by their programs into AnVIL. For the next year, AnVIL will engage with teams from the eMERGE PRS, PRIMED, TARN, Genomic Answers for Kids, and GREGoR consortia to facilitate ingestion of the data they will collect and to understand how AnVIL will store the data they will generate through analyses.

Clinical Genomics Data Ingestion

The eMERGE network has been engaged in large-scale genetic research in support of developing clinical genomic tools. AnVIL has been actively working to bring in eMERGE's retrospective studies to provide expansive joint calls to the network so that polygenic risk score (PRS) method development could happen in preparation for eMERGE IV. In eMERGE IV, clinical samples will be received at the Broad Institute starting in Q1 2022, genotyped on Illumina's Genomic Diversity Array (GDA), and then made available through AnVIL along with simple sample and imputed single sample variant call files (vcf). Imputation and PRS reports will be done through Terra workflow launcher (WFL) and then published to Vanderbilt's R4 Portal. To date, we are actively bringing in 8 retrospective studies, which include imputed GWAS, PGRNseq, and eMERGEseq datasets. This is expected to be completed by the end of 2021.

Optimizing Consortium Engagements

We are planning several activities to optimize consortium engagements to provide more efficient and scalable services:

1. ***Standardize AnVIL onboarding materials*** - We recently created a standard kickoff slide deck that standardizes the presentation and onboarding for consortia members. This slide deck provides a general overview of AnVIL capabilities and outlines the prerequisites for data submission. In addition to providing standard onboarding material, we designed the slides so that any AnVIL team member in Data submission/ingestion and Outreach working groups could present the slides to consortia.

2. ***Streamline consortia applications to deposit in AnVIL*** - Currently, consortium must complete several webforms to describe the data and metadata in their project. These forms are partially redundant, which increases the burden on consortium members and can lead to inconsistent reports. We are currently reviewing these webforms with the goal of removing any redundancy of the AnVIL application process to provide a more optimized experience for data submitters.

3. ***Create an information pipeline between dbGaP and AnVIL*** - Currently, data submitters have separate study registration and data submission processes for dbGaP and AnVIL. We are now in active discussions with dbGaP to optimize this process. To start, we have built and continue to build web agents that can pull study information directly from dbGaP rather than asking submitters to provide all dataset attributes twice.

4. ***Distribute consortia requests across AnVIL working groups*** - As researchers advance through the data lifecycle from data registration, data ingestion, and QC, to batch and interactive analysis, and data

dissemination, they will need the support of different groups within AnVIL. We are optimizing the handoff between such teams and providing new opportunities for researchers to engage with many groups at once through our recently established AnVIL Office Hours program and the AnVIL Discourse webforum (https://help.anvilproject.org/).

5. ***Create and publish learning opportunities and events*** - The data ingestion team works closely with the outreach and portal teams to publish training materials and other documentation on the data ingestion process. This helps empower users to self-serve as questions arise and inform them of the overall process from data registration through ingestion and QC.

Quality Assurance, Quality Control, and Data Harmonization

Quality Assurance

By recognizing the need for quality measures within AnVIL and engaging in process improvement steps, we have taken the first step in creating a quality assurance process for data submission/ingestion and data maintenance. We are also actively engaging our data contributors for feedback on how to improve data quality. The AnVIL Data Access and Policy working group is driving the Quality Assurance effort with input from the Data Processing working group.

To implement the first phase of quality assurance policy, we have created data versioning and change control plans. The versioning plan allows data submitters to submit new or updated data to AnVIL through the data submission process. All new and/or updated data is added to a new workspace that contains only the data from that submission. The versioning plan also includes guidance for data submitters on how to manage participant withdrawal. This plan ensures that the data are maintained for analyses and those analyses are replicable. Versioned workspaces will have the same name as the original workspace and be appended with a version number. Versioned workspace dashboards will include version number in the dataset attributes. For the change control plan, we need to address minor changes to data, such as changes to phenotype or metadata. These changes will be captured in a change log file that is stored in the workspace cloud bucket and noted in the workspace dashboard "About" section. In addition to dataset versioning, AnVIL will also apply method versioning when QC tools are run by the AnVIL data processing team by including the DOI to track the workflow and list it in the QC output data table.

Quality Control

The AnVIL Data Processing working group created a quality control package that evaluates the data quality for genomics sequencing data based on aggregate sequencing metrics shown below (**Figure 2**). Default AnVIL cutoffs for WGS and WES metrics including pass/fail/no QC statuses were created and WGS thresholds were applied to the whole genome data. Out of 136,825 genomes, 1.6 percent failed based on the application of 5 metrics (mean coverage, contamination, chimeras, and read 1 and 2 base mismatch rate). The QC aggregator collected these metrics and outputted them into a summary tsv, which AnVIL made available in the CCDG workspaces as a QC results data table (**Figure 3**). This type of tooling for comparing AnVIL QC batches/workspaces with a focus on managing change control has been implemented and made available.

It is important to highlight that while the AnVIL team will provide these QC tools to submitters, AnVIL is not responsible for running these tools or covering the cloud compute costs as the consortium is responsible for assessing their own datasets. However, the AnVIL Data Processing Working Group will work with the Broad User Education team to create a featured workspace to showcase this tool and provide clear instructions to AnVIL data submitters.

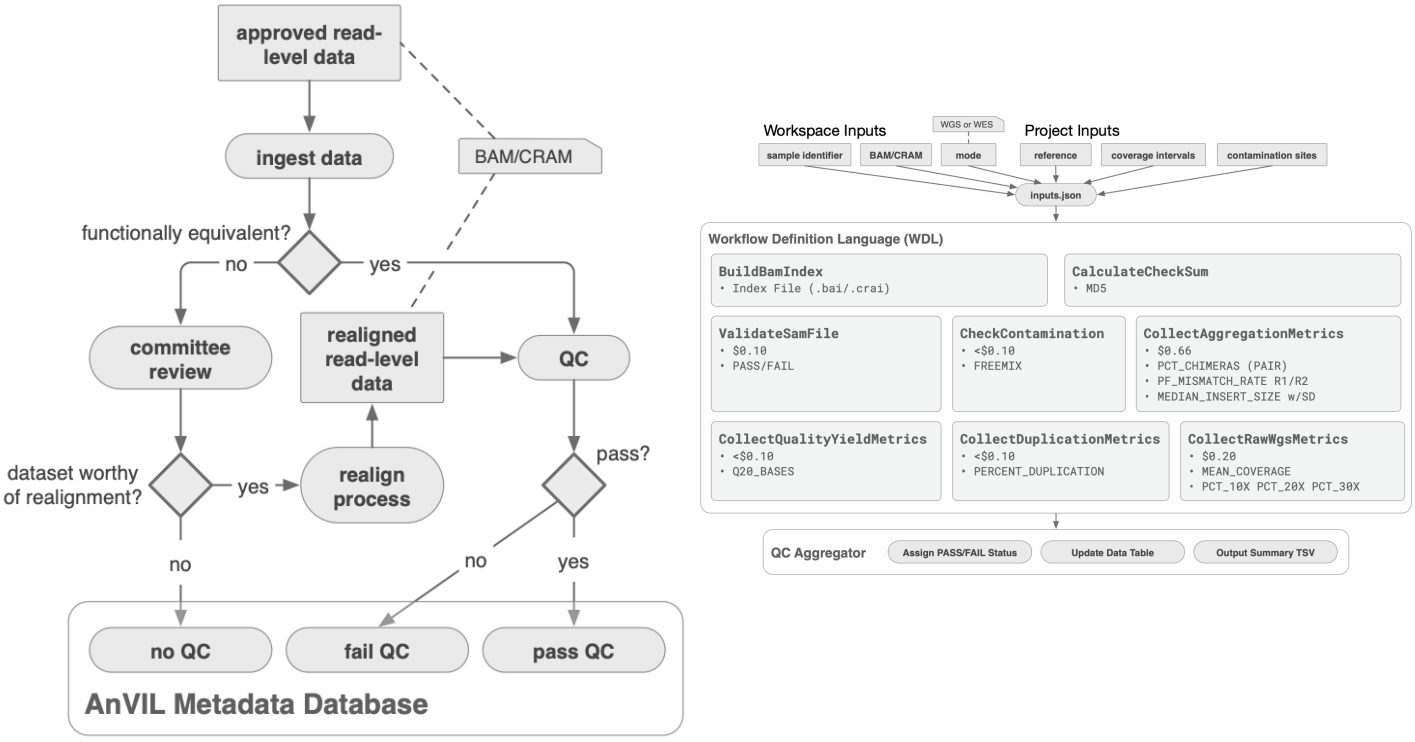


Figure 2: AnVIL Quality Control Process workflow (left) and definitions (right).

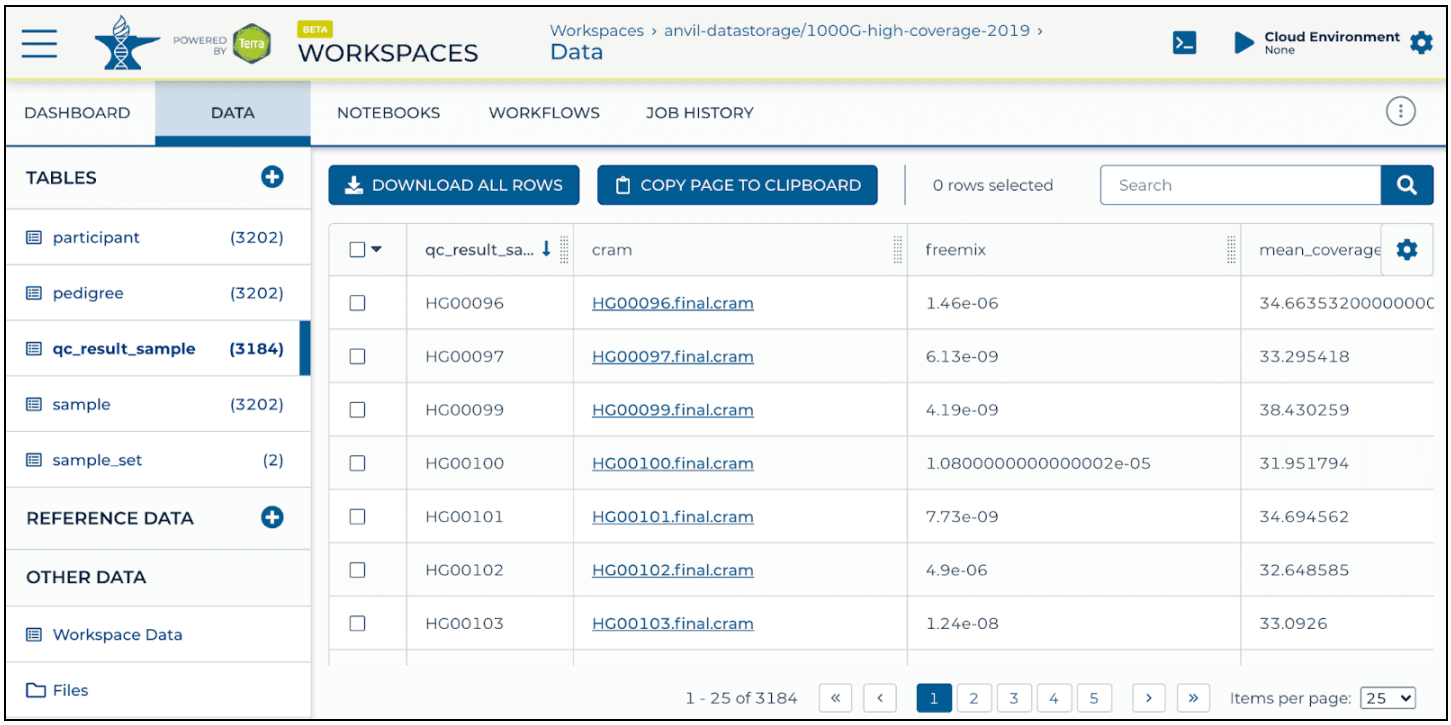


Figure 3. QC Results in 1000 Genomes workspace.

Data Harmonization

To maximize data usefulness to researchers, data coming in from disparate sources requires harmonization. AnVIL has embraced data harmonization by creating a data model to align all data submitted and partnering with Gen3 for data indexing, sharing, and harmonization across platforms.

Data model

As discussed in the data submission section, AnVIL has adopted the Terra Interoperability Model (**Figure 4**) which captures the minimum amount of data required for AnVIL data submission, ingestion, and QC. By aligning all AnVIL data to a common data model, the data are more efficiently searched and shared thereby improving scientific analysis and discovery. The Terra Interoperability Model captures a common set of concepts and relationships for biomedical research intended to facilitate and encourage data sharing and reuse. Its purpose is to enable researchers to find highly connected biomedical data in a federated search space and support interoperability among datasets. While the Terra Interoperability Model is a conceptual model, the AnVIL data submission / ingestion team are working to create a template data model reference for data submitters that aligns with the minimum requirements of the model and ensures AnVIL data submitters can easily prepare their data for submission and QC.

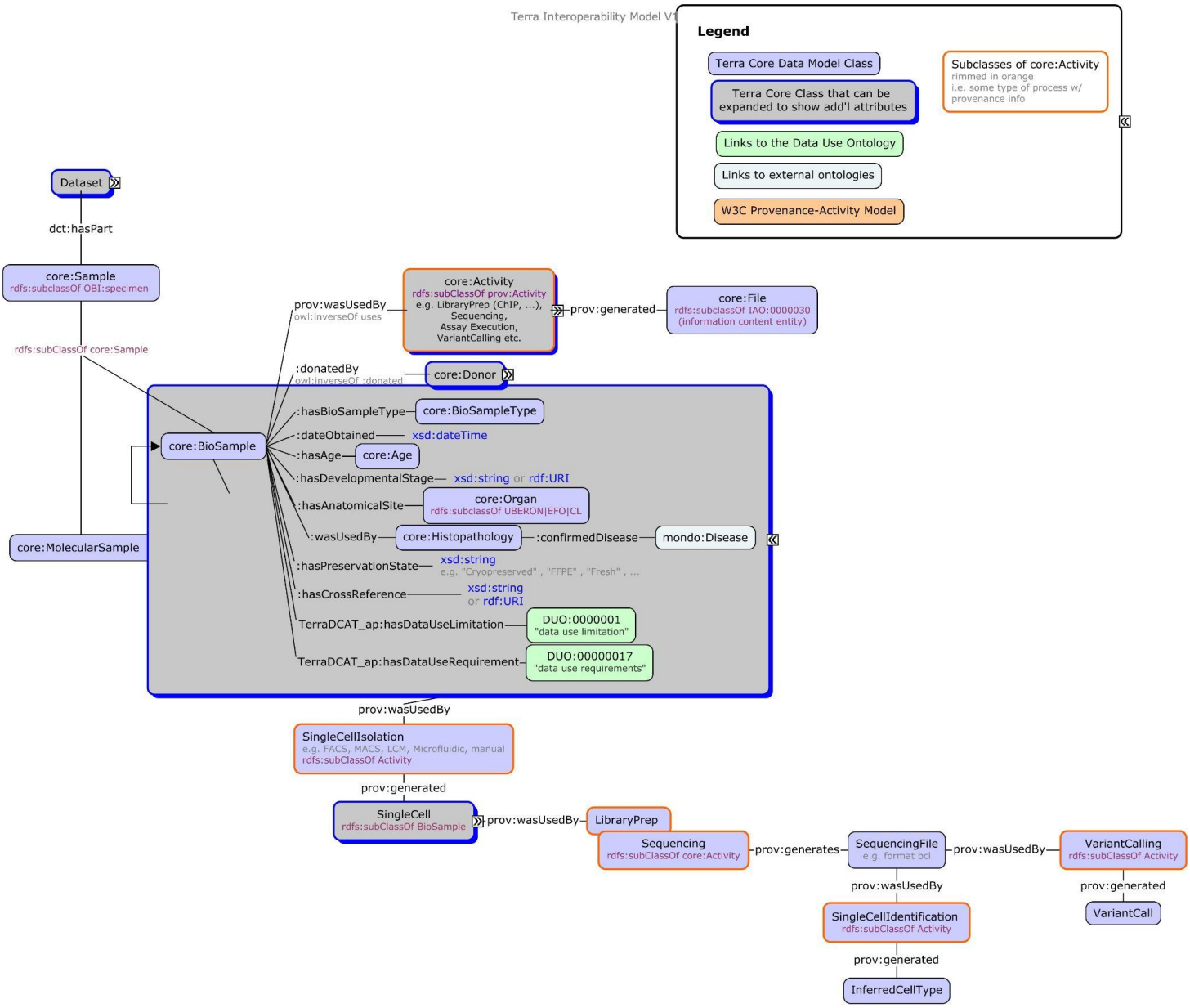


Figure 4. Terra Interoperability Model.

Gen3: Management, analysis, harmonization, and sharing of large datasets

Gen3 is a cloud-based software platform for managing, analyzing, harmonizing, and sharing large datasets. Gen3 is an open source platform for developing data commons. It accelerates and democratizes the process of scientific discovery, especially over large or complex datasets. Gen3 was deployed into production within AnVIL in June 2020, and since then many thousands of datasets have been indexed from the 1000 Genomes, GTEx, and CMG projects. The remaining cohorts in AnVIL from the CMG and CCDG projects are currently in processing.

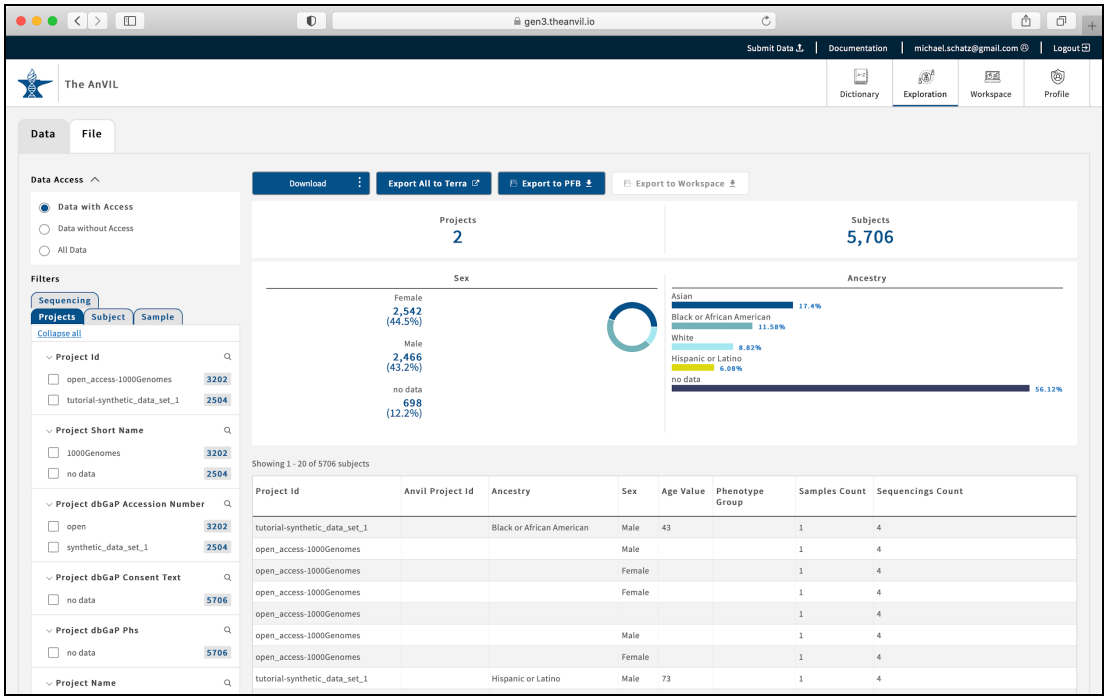


Figure 5. Gen3 Explorer Interface (https://gen3.theanvil.io/).

The Gen3 platform consists of open-source software services that support the emergence of healthy data ecosystems by enabling the interoperation and creation of cloud-based data resources, including data commons and analysis workspaces. Gen3 aims to accelerate and democratize the process of scientific discovery by making it easy to manage, analyze, harmonize, and share large and complex datasets in the cloud.

- **Indexing and Metadata Services.** Provides permanent digital IDs for data objects which can be used to retrieve the data, or query the metadata associated with the object via Gen3's API or GA4GH DRS. Tracks the locations and hash of every asset (file) in the data commons object store. Exports RESTful APIs for registering a new asset, and retrieving data for an existing asset. Other services are responsible for herding user submissions of metadata into the graph database. The submissions are quality controlled against the data dictionary to ensure all required fields are present and have appropriate data values. Also supports bulk export of the metadata into TSV or JSON formats.
- **Search.** High speed metadata seeking which responds to GraphQL search queries. The GraphQL service allows Commons operators and users to precisely query only the information they are most interested in from the metadata collections. The services translate the GraphQL search into the appropriate statements which are run against the PostgreSQL backend or graph database before being returned as a friendly JSON.
- **Auth Services.** Controls access to the metadata, submission, indexing, and data itself. Both authentication (AuthN) and authorization (AuthZ) support utilizes OpenID Connect flow (an extension of OAuth2) to generate tokens for clients and is adding support for GA4GH AAI and Passports. Can also provide tokens directly to a user. Clients and users may then use those tokens (JWT) with other Gen3 APIs to access protected endpoints that require specific permissions. Can be configured to support different Identity Providers (IDPs) for AuthN. At the moment, supported IDPs include Google, and Shibboleth supporting providers such as NIH iTrust and NIH RAS.

- **Gen3 UI**. An interactive website that allows users to explore, submit, and download data (**Figure 5**). Allows for interactive data exploration, search and cohort-building based on phenotypic variables and data types. Selected cohorts can then be exported into Terra for downstream processing.

Data Available in AnVIL

The AnVIL Data Dashboard (**Figure 6**) provides an overview of Terra workspaces to broadly summarize and discover datasets of interest. Datasets can be filtered by consortium, access control, or datatype as well as free text search within the workspace name or description, e.g. disease or phenotype of interest. Importantly, the cohort information displayed is strictly unprotected data so that any user, including those without AnVIL accounts, can discover what cohorts are currently available within AnVIL. Our expectation is that this will encourage new users to come to AnVIL as they discover the rich catalog of data available.

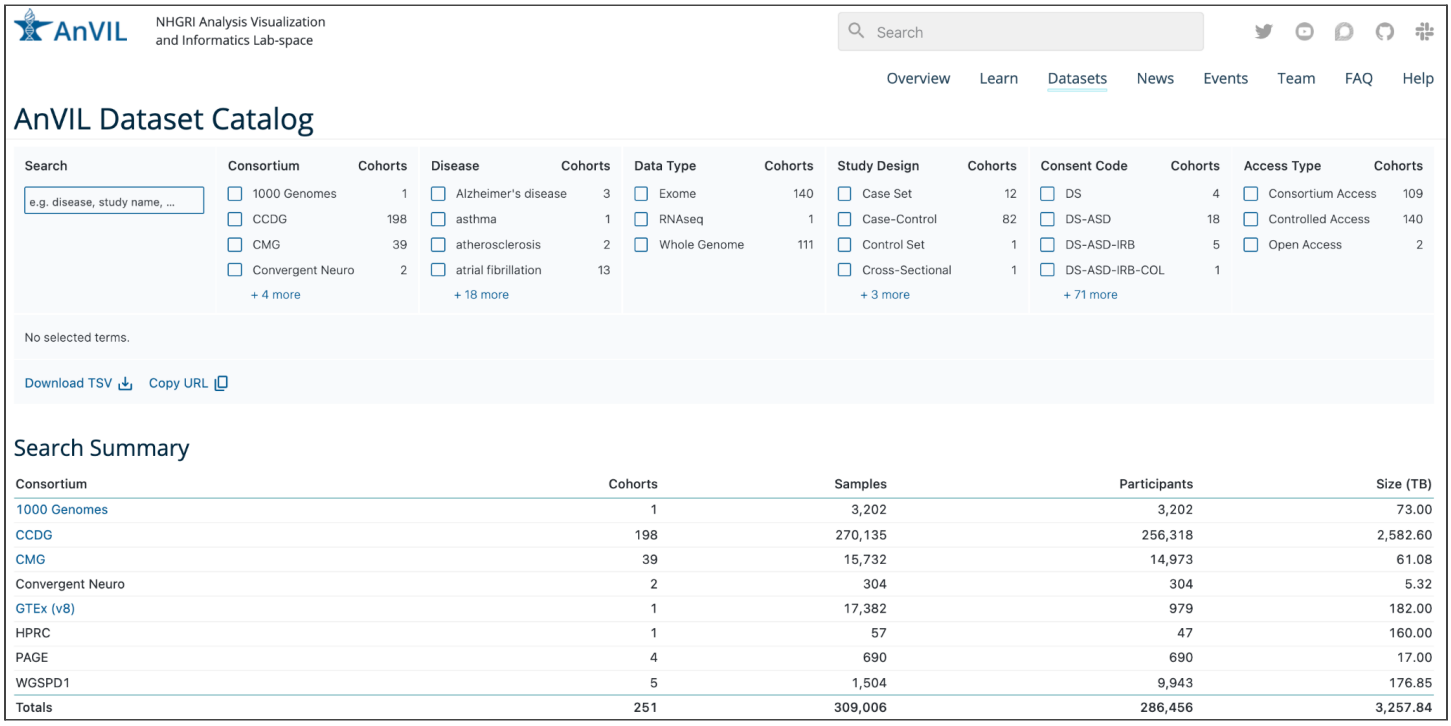


Figure 6. AnVIL Data Dashboard (https://anvilproject.org/data).

Data Ingestion Roadmap

At this time, AnVIL has mostly ingested whole genome and whole exome sequences as well as joint callsets. In the coming year, AnVIL will expand the number of whole genome and whole exome sequencing projects, as well as start hosting a wider variety of data types including ENCODE, dGTEx, CARD Dementia Long-Read project, and recount3. Our ingestion roadmap is displayed in **Figure 7.**

- ENCODE data features >18,000 functional genomic experiments, >2,000 CRISPR screens, and single cell data. Hg38 aligned and processed genomic data files will be copied from AWS into a public AnVIL workspace. In addition, AnVIL teams will work with ENCODE researchers to create a set of featured workspaces.

- Developmental Genotype-Tissue Expression (dGTEx) will kick-off later this fall and involve genomic characterization of postmortem tissues collected from the pediatric population.
- NIA is funding a long-read whole genome sequence project in 4,000 postmortem brain tissues to characterize large structural variation and repeat expansions in Alzheimer's disease and related dementias using the Oxford Nanopore platform. NIA investigators will work with UCSC to put the data files into AnVIL for data processing, joint calling, and data release.
- Finally, AnVIL is bringing in a valuable dataset generated by JHU called recount3, which consists of uniformly processed and quantified publicly available RNA-seq data from human and mouse totalling over 750,000 samples and close to 19,000 projects.

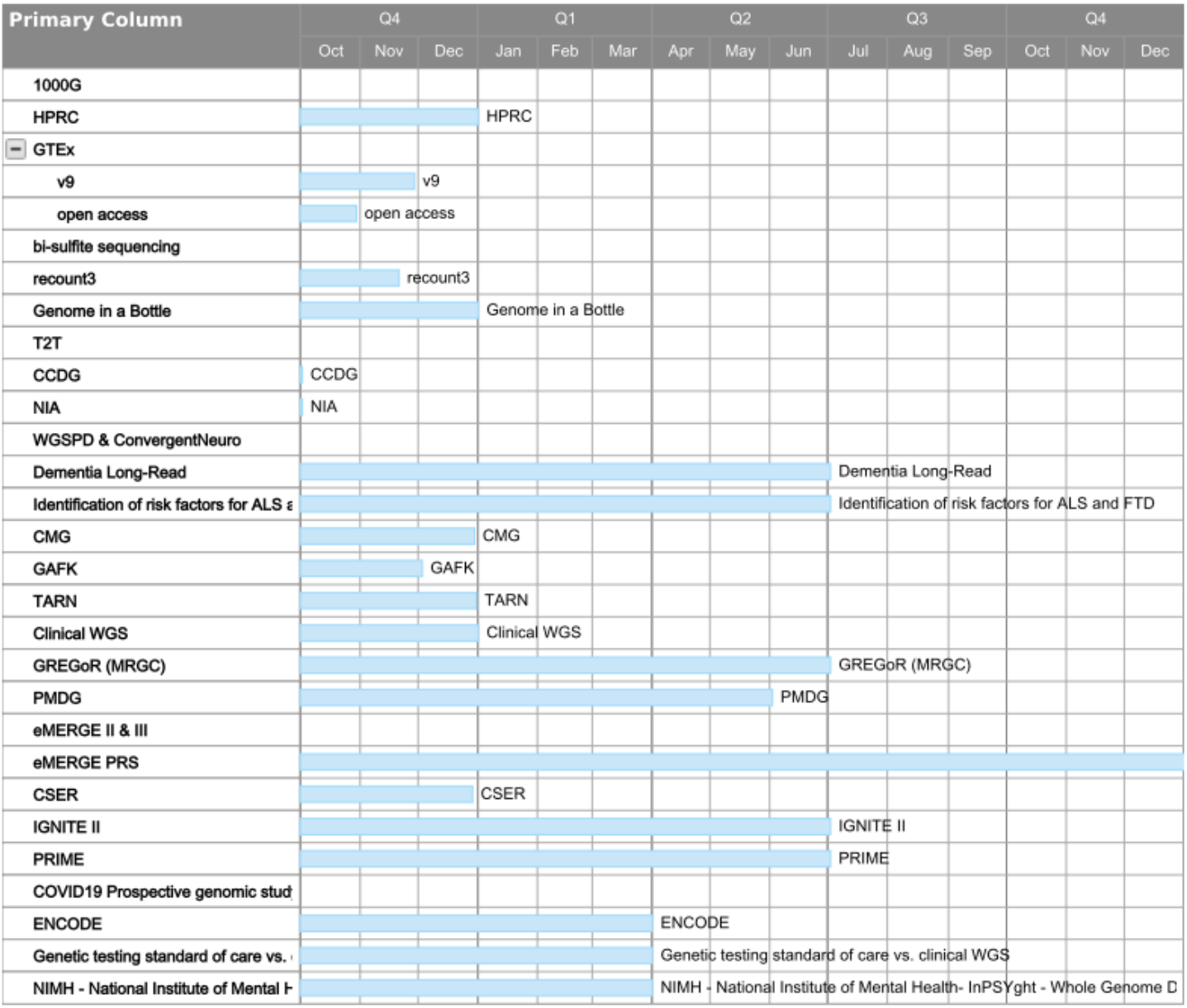


Figure 7. AnVIL 1 Year Data Roadmap.

V. Analysis tools

The AnVIL environment is a rich ecosystem of computational tools designed to enable researchers to perform batch analysis and interactive analysis. It includes tools that provide simple graphical user interfaces and lower-level tooling for programmatic access. The compute environment for AnVIL is built on the Google Cloud Platform (GCP) to enable massive scalability and capacity for users, as well as a robustly established FedRAMP-certified security perimeter. Users now have several options for analysis, ranging from large scale batch computing using WDLs executed in Terra that are stored within Dockstore, to more interactive environments using Jupyter notebooks, R/Bioconductor, and Galaxy.

Dockstore: Registry of Tools and Workflows

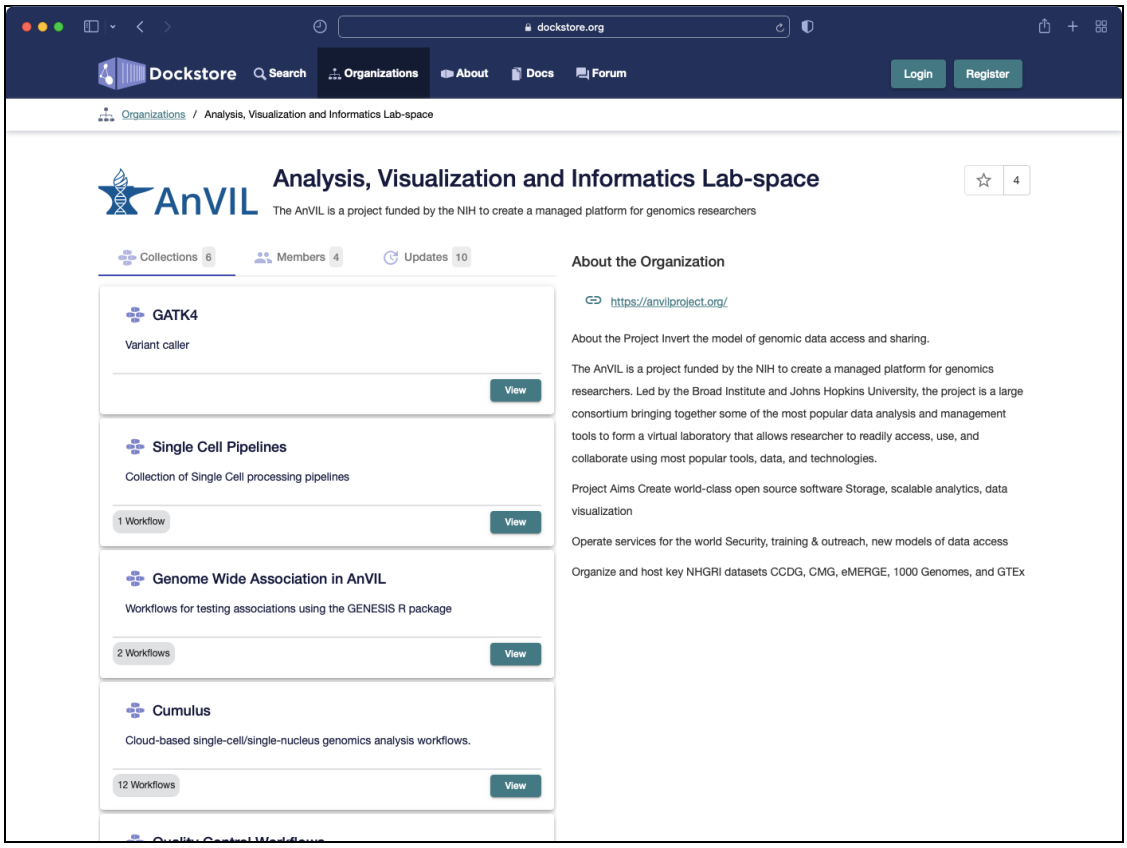


Figure 8. AnVIL collection of tools within Dockstore (https://dockstore.org).

The Dockstore (https://dockstore.org) concept is simple; it provides a place where users can share tools encapsulated in Docker and described with the Common Workflow Language (CWL), Workflow Description Language (WDL), or as Galaxy Workflows (**Figure 8**). This enables scientists, for example, to share analytical tools in a way that makes them machine readable and runnable in a variety of environments. Dockstore tools can be launched directly in AnVIL. While the Dockstore is focused on serving researchers in the biosciences, the combination of Docker with CWL, WDL, and Galaxy Workflows can be used by anyone to describe the tools and services in their Docker images in a standardized, machine-readable way.

Jupyter Notebooks: Transparent Code, Visualizations, and Narratives

Jupyter Notebook (https://jupyter.org/) is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text (**Figure 9**). Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more. Jupyter supports over 40 programming languages, including Python, R/Bioconductor, Julia, and Scala. Jupyter Notebooks are an open document format based on JSON that contain a complete record of the user's sessions and include code, narrative text, equations and rich output.

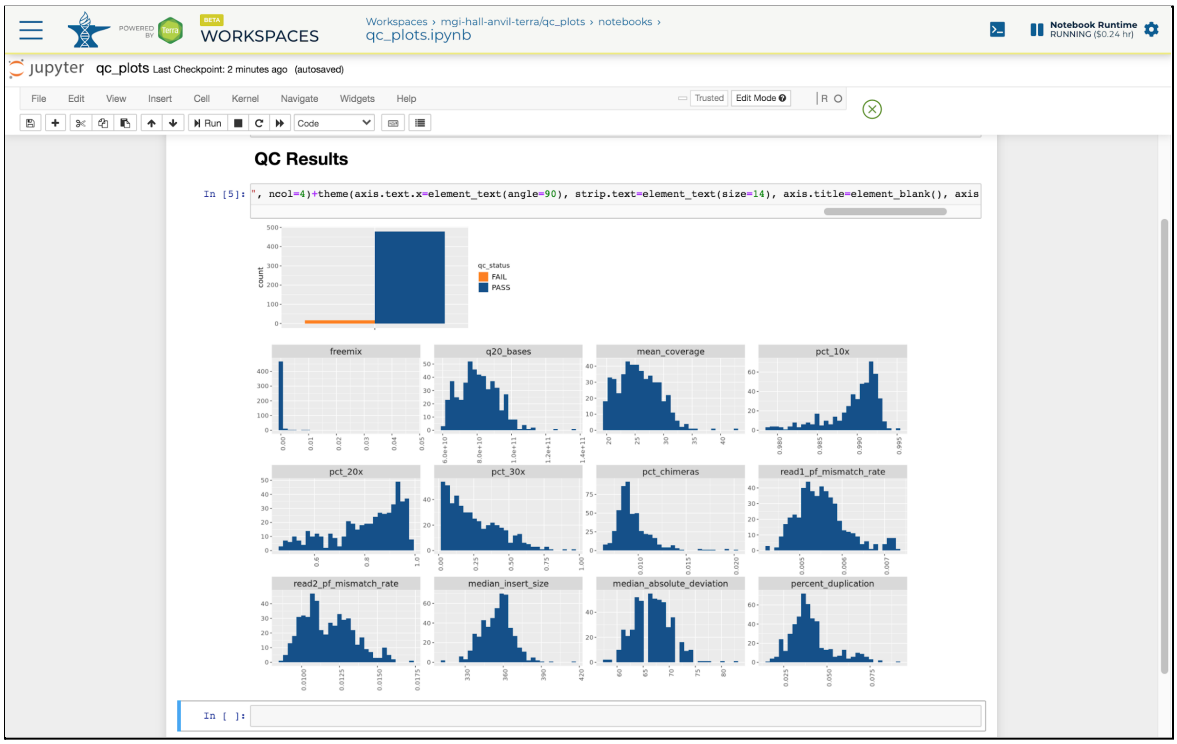


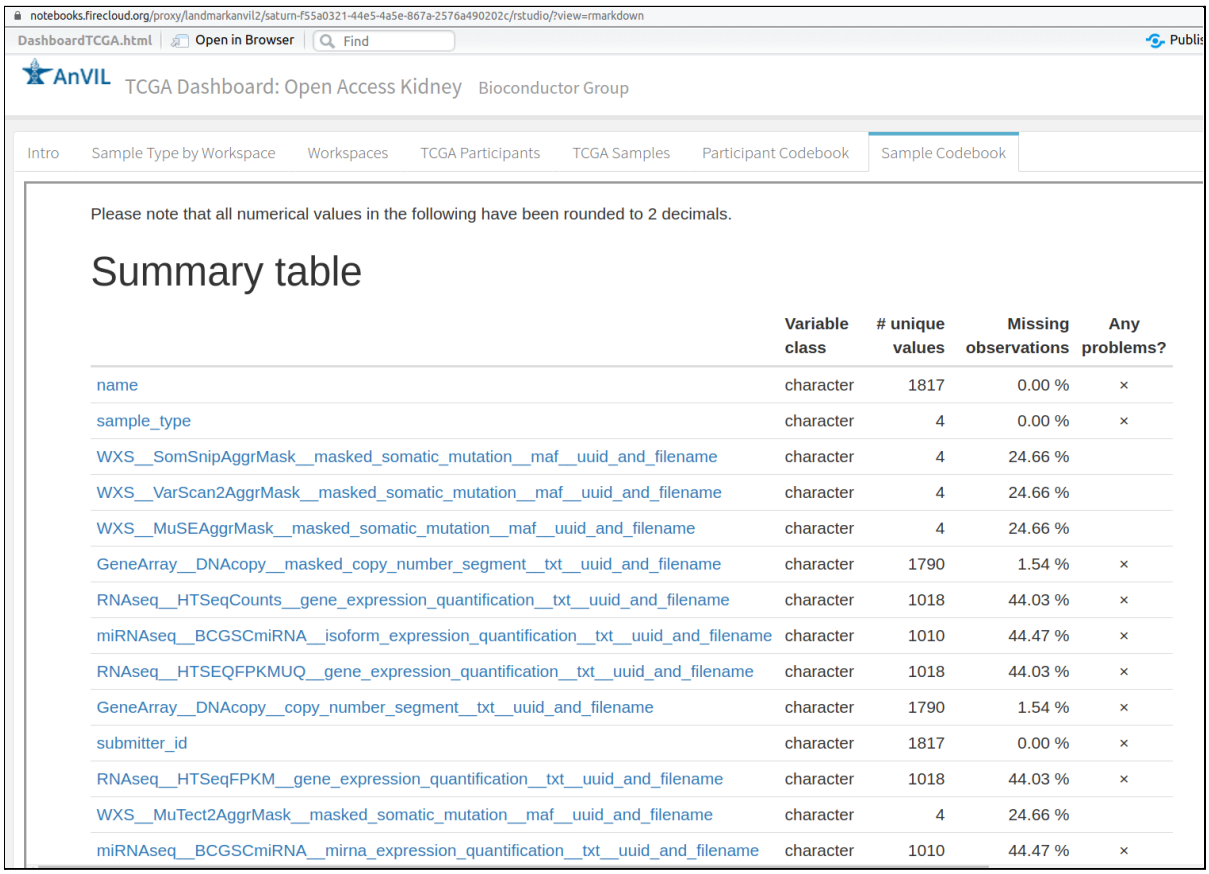
Figure 9. Jupyter Notebook in Terra showing read alignment QC metrics.

Recently, we enhanced the cloud environment (previously called "runtime") that we provide in Terra for running Jupyter Notebooks to support persistent disk storage. Previously, one of the limitations of our Notebook environment required users to manually save any outputs to a Google bucket (or other location of your choosing). For technical reasons, the storage space associated with the notebook was not guaranteed to remain available when not actively in use, and if users made certain configuration changes to the environment, the storage space was deleted and recreated from scratch.

Going forward, users have the option to use what's called a "detachable persistent disk" to store data that they plan to use as well as the outputs of any analyses they run in their notebooks. Persistent disks are analogous to virtual USB thumb drives; users plug them in when they want to do some work, then detach it when they are done, and keep it in their pocket until next time. In practice, the plugging in and detaching is done automatically by Terra. When users reconfigure or delete their environment, the system will automatically detach and reattach the persistent disk, or save it for later if they don't create a new environment right away.

RStudio: Interactive Machine Learning, Statistical Computing, and Visualizations

RStudio (https://rstudio.com/) is an integrated development environment for R, a programming language for statistical computing (**Figure 10**). R and its libraries implement a wide variety of statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and others. R is easily extensible through functions and extensions, and the R community is noted for its active contributions in terms of packages. Other strengths of R include advanced static and interactive graphics, and facile creation of graphical user interfaces for easy use of highly specialized packages.



	Variable class	# unique values	Missing observations	Any problems?
name	character	1817	0.00 %	×
sample_type	character	4	0.00 %	×
WXS__SomSnipAggrMask__masked_somatic_mutation__maf__uuid_and_filename	character	4	24.66 %	
WXS__VarScan2AggrMask__masked_somatic_mutation__maf__uuid_and_filename	character	4	24.66 %	
WXS__MuSEAggrMask__masked_somatic_mutation__maf__uuid_and_filename	character	4	24.66 %	
GeneArray__DNAcopy__masked_copy_number_segment__txt__uuid_and_filename	character	1790	1.54 %	×
RNAseq__HTSeqCounts__gene_expression_quantification__txt__uuid_and_filename	character	1018	44.03 %	×
miRNAseq__BCGSCmiRNA__isoform_expression_quantification__txt__uuid_and_filename	character	1010	44.47 %	×
RNAseq__HTSEQFPKMUQ__gene_expression_quantification__txt__uuid_and_filename	character	1018	44.03 %	×
GeneArray__DNAcopy__copy_number_segment__txt__uuid_and_filename	character	1790	1.54 %	×
submitter_id	character	1817	0.00 %	×
RNAseq__HTSeqFPKM__gene_expression_quantification__txt__uuid_and_filename	character	1018	44.03 %	×
WXS__MuTect2AggrMask__masked_somatic_mutation__maf__uuid_and_filename	character	4	24.66 %	
miRNAseq__BCGSCmiRNA__mirna_expression_quantification__txt__uuid_and_filename	character	1010	44.47 %	×

Figure 10. RStudio running in Terra. TCGA KIRC (renal clear cell) and KICH (chromophobe) cohorts were combined to survey data availability and quality using dataMaid and flexdashboard.

RStudio has been deployed within Terra, and therefore RStudio operates within the FedRAMP-certified security perimeter provided by Terra. The infrastructure is built to support current versions of R / Bioconductor, and adopts the 'all of Bioconductor' strategy we use in Jupyter notebooks. The RStudio image has been used to deploy important resources, including a fully computable version of the online book Orchestrating Single Cell Analysis with Bioconductor.

Bioconductor: Community-driven Interactive Genomics with R and RStudio

Bioconductor (https://bioconductor.org/) is a free, open source and open development software project for the analysis and comprehension of genomic data generated by wet lab experiments in molecular biology (**Figure**

11). Computational and statistical methods are continuously developed to interpret biological data. Many of these methods are developed by members of the Bioconductor community, and the Bioconductor project serves as a software repository for a wide range of statistical tools developed in the R programming language. Using a rich array of statistical and graphical features in R, more than 1900 Bioconductor software packages, 3200 exemplary experiments, and 50000 model organism annotation resources have been curated for use in genomic data analysis. The use of these packages requires only an understanding of the R language. As a result, R / Bioconductor packages, which include state-of-the-art statistical inference tools tailored to problems arising in genomics, are widely used by biologists who benefit significantly from their ability to explore and analyze both public and privately developed datasets. Many R / Bioconductor applications can be presented to users in a way that does not require advanced programming expertise, e.g., as 'shiny' applications with graphical interfaces.

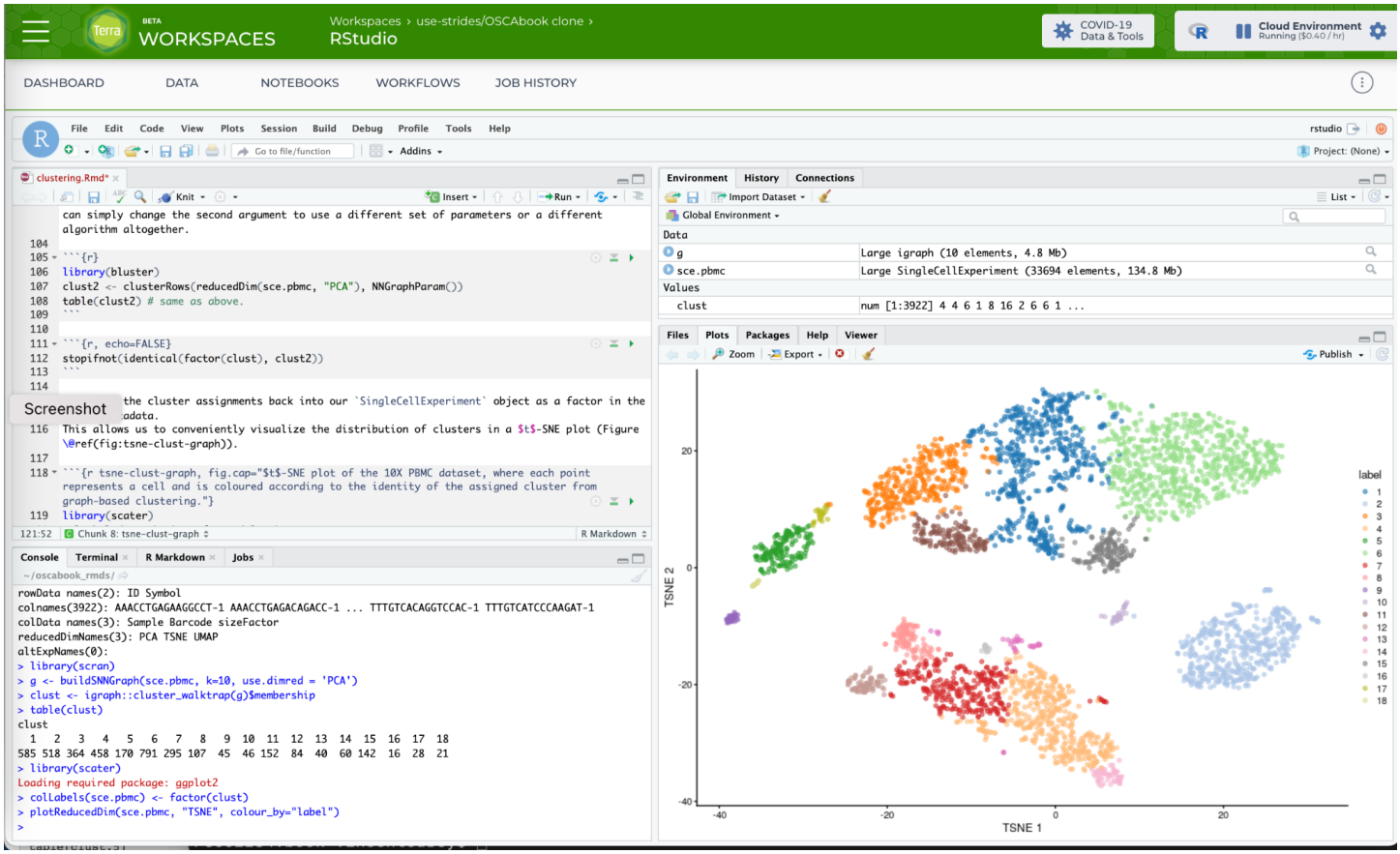


Figure 11. Bioconductor running in Terra performing a single cell RNAseq analysis.

Over the past few years, we have significantly enhanced the capabilities of interacting with AnVIL in Bioconductor through the development of several packages, workshops, and other materials (video, slideset). We have contributed to standardized, up-to-date notebook environments for computation using a novel strategy for enabling Docker-based installation of diverse R / Bioconductor packages. We continue to develop and maintain the AnVIL and Gen3 R packages to enhance user experience of AnVIL from within R, accessing all elements of the Terra environment (e.g., tables, buckets, cloud utilities for resource management). Recognizing the importance of engaging our own communities, we have contributed to workshops and training sessions at our annual conference, at the Bioinformatics Community Conference, and through other

venues; workspaces developed for this purpose embody principles of original scientific innovation via reproducible cloud computing (e.g., the publication-backed Bioconductor-Resource-Tumor_Only_CNV workspace for reliable analysis of tumor CNV/SNV without matching normal samples). Another paper sketched the relationship between Bioconductor and emerging GA4GH protocols TRS and WES in a comprehensive analysis of transcriptomic signature of immune infiltration in all TCGA tumor types. Maturing activities include notebook and workflow cost accounting (via the AnVILBilling package), automated documentation and training material transformation from R / Bioconductor formats to AnVIL workspaces (via the AnVILPublish package and the Bioconductor-Resource-OrchestratingSingleCellAnalysis workspace for comprehensive single-cell exploration), and fast binary R / Bioconductor package repositories for an enhanced experience when notebook configuration takes substantial time. We are working closely with Terra software engineers to coordinate evolution of AnVIL API components with the six-month release cadence (full version updating of all resources) observed by Bioconductor. We are also working on generalization of key user interfaces to data and compute resources by adapting components of the AnVIL package to function in the Microsoft Azure cloud environment.

Galaxy: Accessible, Reproducible, and Transparent Genomic Science

Galaxy (http://usegalaxy.org) is an open, web-based platform for performing accessible, reproducible, and transparent genomic science. It includes features for executing scientific workflows, data integration, and data and analysis persistence. A major aim of Galaxy is to make computational biology accessible to research scientists that do not have computer programming or systems administration experience. Although it was initially developed for genomics research, it now has broad support for biomedical research of many forms. There are more than 8,000 analysis tools available within Galaxy including those for gene expression, genome assembly, proteomics, epigenomics, transcriptomics, and a host of other disciplines in the life sciences.

Galaxy has been deployed within Terra, and therefore Galaxy operates within the FedRAMP-certified security perimeter provided by Terra. Since Galaxy can operate on large datasets requiring significant computing resources, it was not possible to deploy Galaxy through a single Docker container as was done for Jupyter notebooks or R/Bioconductor. Instead, we have leveraged an alternate cloud technology called Kubernetes that allows for automating deployment, scaling, and management of containerized applications. Furthermore, to maintain data security, each AnVIL/Galaxy user launches an independent instance of Galaxy within separate Kubernetes clusters (**Figure 12**). Accomplishing this required significant software engineering efforts from the Terra and Galaxy teams. From Terra, there has been a substantial extension to the underlying APIs to launch, manage, and monitor applications running within the Google Kubernetes Engine (GKE). From Galaxy, there was a coordinated effort to package Galaxy into a self-contained application that could be automatically initialized and robustly operated without a systems administrator as is normally required.

We are pleased to announce that these efforts have been successful, and that anyone can now launch Galaxy from within the Terra development environment. Prior to launching Galaxy, users can request multiple compute nodes with specific CPU and memory resources for processing their analysis jobs (**Figure 13**). Once Galaxy is running, AnVIL data (unprotected data and/or protected data that the user is authorized to view) can be imported into the user's instance of Galaxy using a newly developed import tool so that the full suite of

Galaxy tools can be run (**Figure 14**). This work builds extensively on the pyAnVIL python library for using AnVIL system components and the AnVILFS for file system representation from AnVIL workspaces.

Galaxy within AnVIL was publicly released in 2020 and was showcased at an ASHG workshop held on October 29, 2020, where participants learned to launch Galaxy from within Terra and perform a small GWAS analysis on human variant data. Additionally, on April 23, 2021, the AnVIL Outreach working group demonstrated the use of Galaxy within AnVIL at the Virtual Applied Data Science Training Institute (VADSTI), which was held as a series of training events highlighting data analytic skill sets required for big data analysis. Participants of this workshop learned how to launch Galaxy from within Terra, perform *de novo* assembly from sequencing reads, and identify novel sequence insertions compared to a reference genome. For those that were unable to attend either of these events, a video was created to guide users in launching Galaxy within Terra (https://anvilproject.org/learn/data-analysts/galaxy-gsg-video).

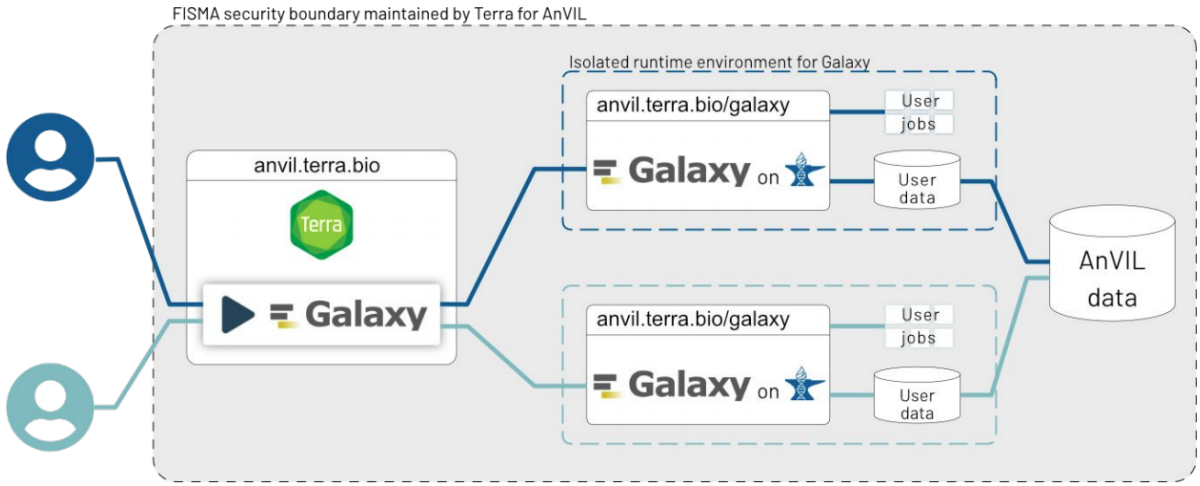


Figure 12. Galaxy deployment architecture.

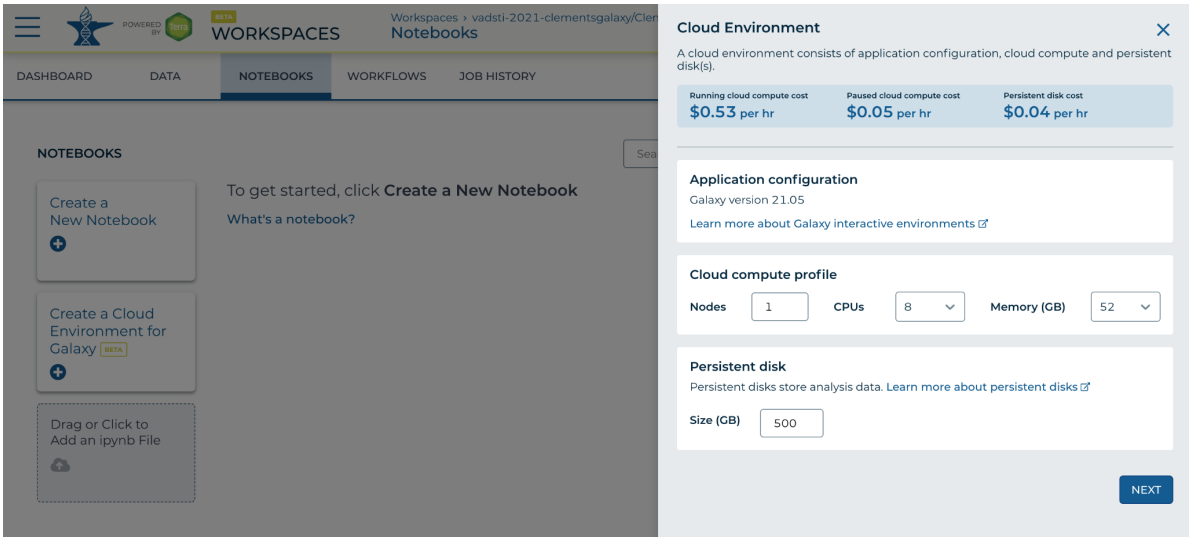


Figure 13. Launching Galaxy from within Terra Workspace.

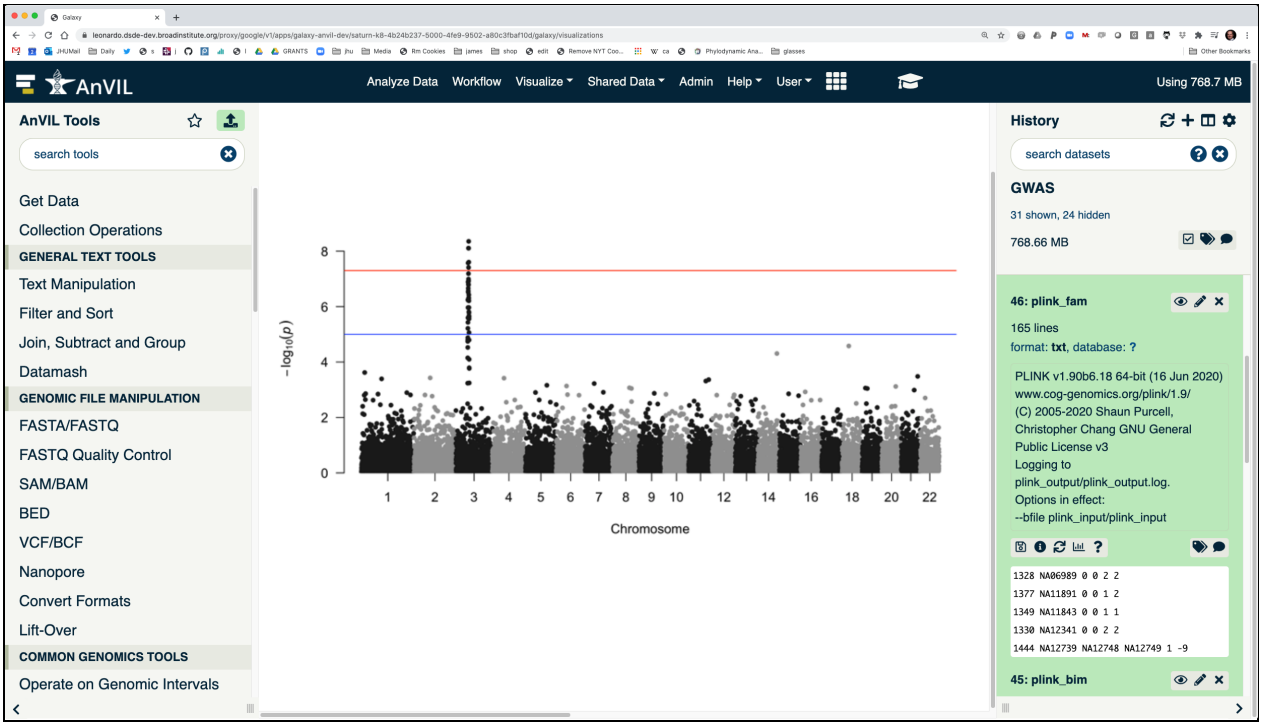


Figure 14. Running Galaxy within Terra on AnVIL.

Clinical analysis tools

With the core infrastructure for AnVIL now maturing, a major focus for the upcoming years is to increase the capabilities for clinical genomics research and analysis. This will build on our existing efforts with the CCDG and CMG to develop capabilities for recognizing pathogenic variants for common and rare diseases, but will also extend to other data types. For example, a major focus of the NHGRI eMERGE program is to perform joint analysis of genomics data with electronic medical record systems for genomic discovery and genomic medicine implementation research. Another important NHGRI clinical program is the CSER Consortium, which aims to rapidly advance the knowledge necessary to develop best practices for the implementation of genomic sequence data into clinical care. CSER is uniquely positioned to answer questions about the clinical implementation of genomic sequencing to meet its growing use in the clinical care of patients with diverse needs. Additionally, analysis tools designed for the examination of clinical data are being added to AnVIL and include those used for polygenic risk scores (PRS), Mendelian disease inheritance (Seqr) and genotype-based prescribing recommendations which can be used to inform treatment decisions (PharmCAT).

Polygenic Risk Score (PRS)

The onset of illness in humans is often associated with many factors that impact disease development, including behavior, environment and genomic variation of an individual. Polygenic diseases are those that stem from multiple genomic variants. Researchers and medical professionals can associate genomic variants with a polygenic illness by comparing the genomes of those with and without the disease. This work is made possible by analyzing vast genomic datasets and identifying genomic variants in common among those impacted by a polygenic disease. Using statistical models, researchers can distill an individual's genomic variants to a score that assesses their risk for disease development. This is the polygenic risk score, a relative

rating that describes how an individual's risk compares to a wider population. On AnVIL, researchers have a number of tools available via WDLs and Jupyter Notebooks, including analysis tools for imputation, data visualization and the generation of PRS reports (**Figure 15**).

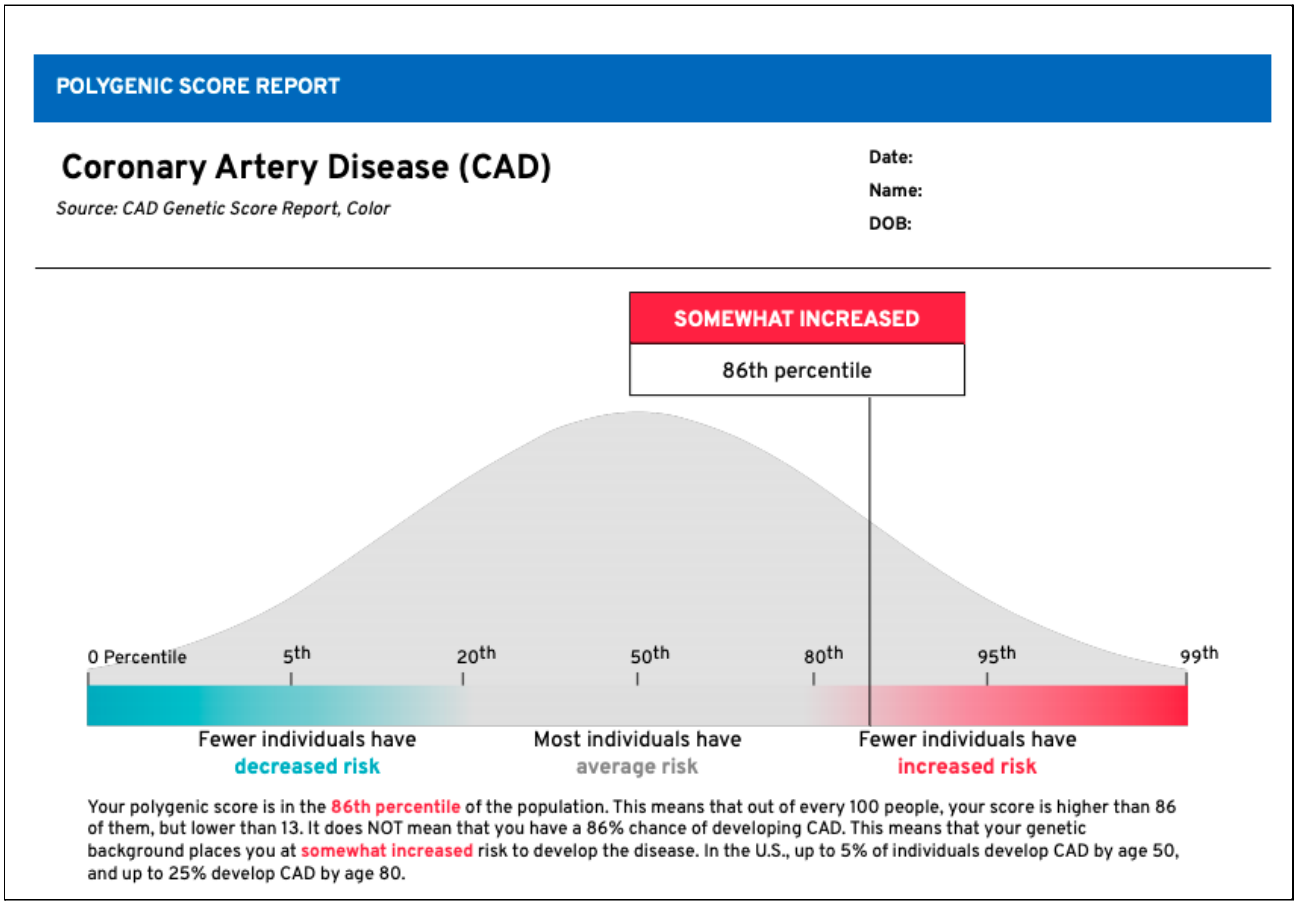


Figure 15. PRS Reporting in AnVIL.

Seqr

Next Generation Sequencing (NGS) is a powerful diagnostic and research tool for Mendelian diseases, but without proper tools, this data can be inaccessible to researchers. Seqr was developed to make research productive, accessible and user-friendly. This analysis platform for genomic data analysis for rare diseases is an intuitive browser-based system for analyzing rare disease exome and genome data on a familial basis. Seqr (**Figure 16**) is an open-source tool supported by the Translational Genomics Group at the Broad Institute that runs on Google Kubernetes Engine (GKE) and data is loaded using Google Dataproc Spark clusters.

The production instance of Seqr uses Postgres, which is a SQL database, to store project metadata and user-generated content (e.g., variant notes) and Elasticsearch to store variant callsets. With the onset of key NHGRI datasets being ingested to AnVIL such as Center for Mendelian Genomics (CMG), Clinical Sequencing Evidence-Generating Research (CSER), and Transamerican Autoimmune Research Network (TARN), the availability of Seqr will enable researchers to analyze and annotate their data with Seqr in the AnVIL platform and collaborate with other investigators as they choose.

Figure 16. Running Seqr within Terra on AnVIL.

PharmCAT

Pharmacogenomics is an emerging medical discipline that utilizes an individual's genomic information to support clinical decision making. Guidelines have been established with regards to gene-drug pairs from various groups around the world, allowing for prescription recommendations based on genetic variants. PharmCAT is a software tool that aggregates pharmacogenomic variants from genetic datasets in VCF data files to generate reports containing information on gene-drug pairs which are designed to support prescription decision making. The AnVIL team is currently working with the creators of PharmCAT to make this clinical analysis tool available within the AnVIL ecosystem.

Extending AnVIL Tool Sets

Currently, AnVIL supports thousands of tools which are available to researchers and are geared to manipulate and analyze biological data. As outlined above, Dockstore, Galaxy, Jupyter Notebooks, RStudio and Bioconductor all bring unique aspects that make AnVIL a powerful data analysis platform allowing multiple entry points for adding new components to AnVIL for clinical and basic research. For command line tools, these integrations can leverage our existing options (e.g. WDLs for batch workflows, Galaxy for interactive analysis, RStudio/Jupyter notebooks for executing code). We are also generalizing the kubernetes-based systems that we currently use for Galaxy to generally support additional third party applications, especially those with a graphical user interface. As a first example of how this can work, we are actively engineering this capability to deploy the UCSC Genome Browser within AnVIL so that users can use the powerful visualization capabilities on their own private datasets.

Integration and Deployment of 3rd party applications

As the AnVIL community continues to grow, the need for more diverse applications closely follows. The integration and deployment of third-party applications poses important challenges that can be grouped into 3 major categories:

1. Development and Testing
2. Platform Security
3. Maintenance and User Support

Development and Testing

One of the most important challenges related to 3rd party applications revolves around availability of developer resources from the analytical platforms (i.e., Terra) to complete the end-to-end testing and deployment work. There is an ever growing number of powerful biomedical and bioinformatics applications that can provide tremendous benefits to researchers. However, the number of possible applications that could be deployed in the AnVIL ecosystem far exceeds the number of software engineer resources available to complete the work. Consequently, a sustainable application onboarding process that leverages developer resources from the application team(s) themselves is necessary. Furthermore, app developers will also need easy access to a reliable and secure development platform where they can iterate on their code and verify correct deployments prior to a production rollout.

Summary:

1. Reliable test environment
2. Streamlined developer guide for app devs
3. Minimize workload placed on platform devs in order to build a sustainable onboarding process

Platform Security

Another critical challenge for 3rd party applications relates to the security requirements that need to be met in order to interconnect with Terra and other analysis platforms. Only by meeting the indicated security requirements set forth by the Broad Institute are apps allowed to launch within the Terra security boundary and access controlled/sensitive datasets. The Terra system, owned and operated by Broad Institute, Inc. (“Broad”), adheres to the National Institute of Standards and Technology (NIST) Special Publication (SP) 800-53 Moderate information security control standard. Terra hosts or may facilitate access to sensitive information including federal research data. As such, applications that wish to interconnect with Terra must meet certain information security compliance requirements based on the type of connection that needs to be made. The process to evaluate security compliance requires that the 3rd party application team submit a security packet for review by the Broad Institute’s Application Security team. The packet requires information security expertise to assemble, which may not always be available for that team. This poses a key challenge both in terms of available personnel as well as potential cost requirements. Lastly, the time needed to prepare the security packet can be lengthy, creating a roadblock and/or disincentive for teams with a tight timeline.

Summary:

1. Understanding the security requirements that need to be presented by the app dev team before the app can be supported
2. Streamlined security guide to create the submission packet
3. Resources and expertise needed to prepare the submission packet which potentially requires funding
4. Preparation of the security packet can be a time consuming process

Maintenance and User Support

Once an application is available in an analysis platform, there is a need to provide ongoing maintenance and user support. Without a proper system in place, a list of growing applications can quickly overwhelm the platform's available resources to provide adequate maintenance for each app. Necessary work includes, among others, regression testing, tracking of new versions and troubleshooting and addressing bugs. Similarly, there's a need to provide an active venue for users to submit help desk tickets and request support. This poses a challenge once again due to potential resource limitations and ensuring that response times are adequate for each application.

Summary:

- Create and fine-tune regression tests to ensure proper deployment of the app
- Tracking of app versions correctly
- Provide a venue for users to submit and receive help desk responses in a timely manner
- Maintenance and user support cannot all be supported directly by the analysis platform due to resource constraints

Addressing Challenges in AnVIL

In order to keep up with a growing list of user-requested apps, the goal is to empower app developers with a streamlined integration and verification process that promotes adoption. We plan on maintaining a Terra App Dev Guide that outlines requirements and shares tooling information to prepare and connect apps in Terra (focused on single-tenant apps deployed inside the Terra security boundary). We will also prepare an App Security Guide that focuses on the necessary security documentation needed to verify apps in Terra and meet NIST-800-53 moderate controls or equivalent. Lastly, the Terra help desk will be tasked with properly labeling user requests and redirecting tickets to app team contact(s) for resolution.

The motivations are 1) providing app developers a user-friendly, easy-to-follow process to test app connections in Terra, 2) expand the availability of powerful, strategically important apps that users can leverage for their research, and 3) minimizing support and upkeep needed from the Terra dev team to maintain a growing list of 3rd party applications. We need to implement an app vetting process that is both reliable and lightweight (where possible), and user/dev friendly to promote adoption.

In summary, the plan to solidify a sustainable 3rd party application onboarding process is as follows:

1. Official Terra App Dev Guide
2. Official Terra Security App Guide

3. Use of Terra Dev and a designated billing account to allow app devs to safely test connections and deployments
4. Focus on single-tenant apps launched within the Terra security boundary short- to medium-term
5. Iterate based on lessons learned to improve both the App Dev and Security guides, and also to further expedite security reviews where possible

Expanding AnVIL Beyond Genomic-based analysis

At this time, the catalog of tools available on the AnVIL platform is heavily geared towards genomic data analysis and data science. However, there are several upcoming opportunities for expanding the platform's capabilities. One major example involves the combinatorial analysis of medical records and associated patient genomic and non-genomic data working in partnership with the NHGRI Electronic Medical Records and Genomics (eMERGE) Network. While a number of ancillary genomic systems supporting pharmacogenomic data storage and reporting have been developed for use alongside electronic health records by various academic working groups and commercial organizations, an open-sourced solution is currently lacking. Another opportunity is radiological-based/image-based analysis. High resolution clinical image analysis remains a key mechanism whereby biopsies are tested for the presence of disease, including many types of cancer.

Related, we also anticipate the need for focused efforts on machine learning, with harmonized datasets for training and testing models, optimized software libraries with GPU support for efficient processing, and advanced visualization capabilities to inspect and debug the characteristics used for classification. This may include the establishment of a “Model Zoo” in which open source deep learning code and pretrained models are made available to the genomics community. Combining machine learning-based image analysis or medical records data with omics-based datasets, such as the AnVIL-hosted GTEx v8, offers an exciting possibility for expanding the utility of the AnVIL platform.

VI. Infrastructure

The AnVIL is a federated cloud platform designed to manage and store genomics and related data, enable population-scale analysis, and facilitate collaboration through the sharing of data, code, and analysis results. It includes a variety of graphical user interfaces along with RESTful interfaces and APIs for programmatic access in several popular programming languages. The compute environment for AnVIL is currently built on the Google Cloud Platform (GCP) to enable massive scalability and capacity for users within a robustly established FedRAMP-certified security perimeter authorized for the storage and analysis of controlled-access datasets. Within the AnVIL, users have several options for analysis and a rich data management ecosystem allowing researchers to search across large collections of data and build novel synthetic cohorts to empower new discoveries out of existing datasets. This section will focus on components of AnVIL including the Internet Portal, Terra, Gen3, and key efforts driving forward NIH Cloud Platform Interoperability (NCPI).

AnVIL Portal

The AnVIL Portal at https://anvilproject.org/ is the user-facing front-end of the AnVIL project. It serves to collect and disseminate a broad range of information and functionality to our user community. Information on the AnVIL Portal is discoverable by search engines, shareable in social media, tracked with Google Analytics, and easily updated by team members.

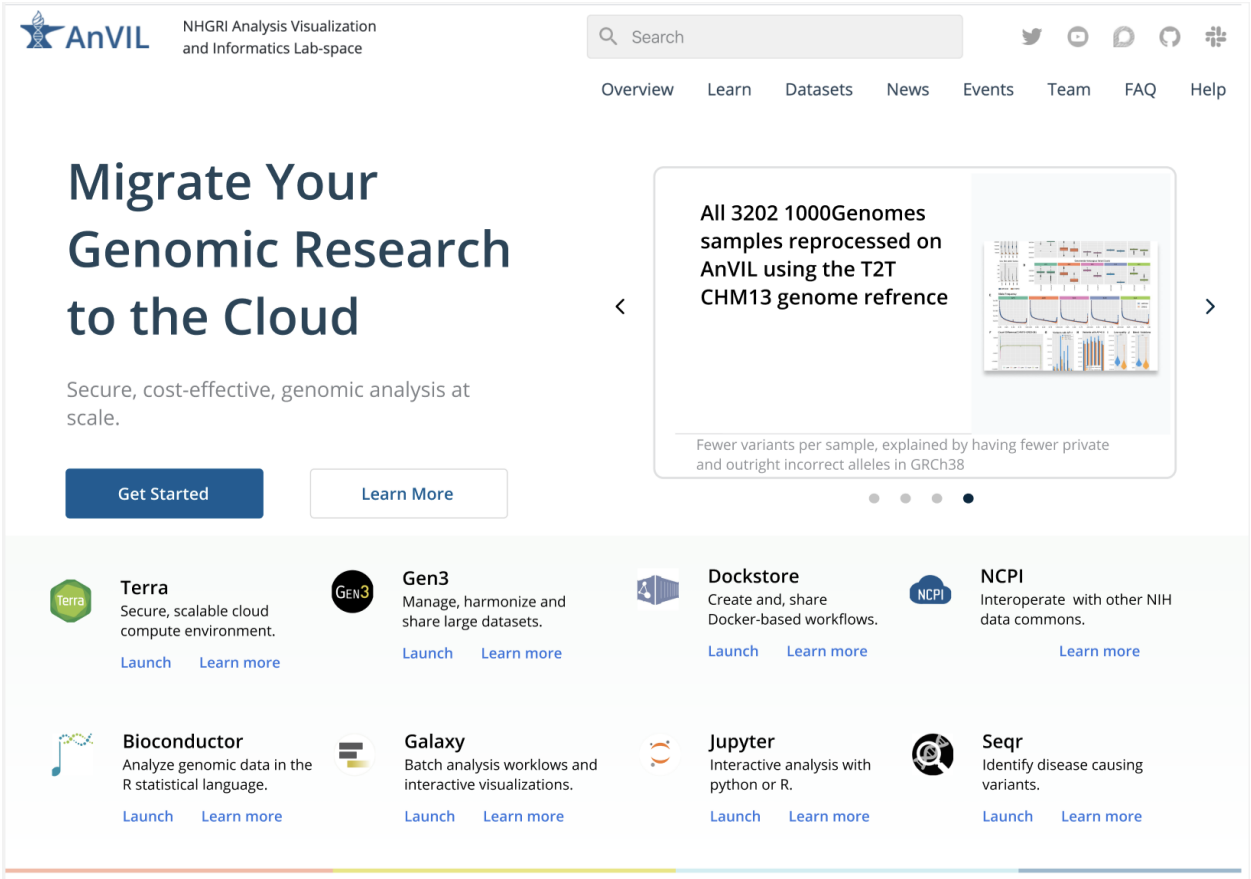


Figure 17. Homepage Redesign.

Home Page - The updated homepage (coming soon) demonstrates the type of science that can be accomplished on the platform and provides quick links to project components and tools (e.g, Terra, Gen3, Dockstore, Bioconductor, Galaxy) as well as links to the Twiter, YouTube, GitHub, Slack, and Discourse community forum (**Figure 17**).

Overview - The overview section details the general principles behind the AnVIL effort covering the benefits of reduced data storage and data egress costs, access to high-quality data sets, tools, and AnVIL's security profile.

Learn - The learn section provides overview and onboarding information and persona-specific tutorials for data submitters, data analysts, and investigators. The portal team continues to source, develop, incorporate, and link out to tutorials and guides demonstrating best practices in using AnVIL components and tools.

Datasets - The dataset section provides users with a faceted search capability where the entirety of the data available to AnVIL users can be queried. This tool is at the core of the AnVIL, as it provides access to the primary assets of the project. The Dataset Catalog is kept up to date as new data are ingested and continues to evolve by providing more information about each available study, as well as information on how to gain access to discovered datasets.

Cross-Site Search - The search utility provides a topical search engine implementation over the portal content and documentation from the Terra, Gen3, Dockstore, NCPI, Galaxy, and Bioconductor websites.

News and Events - In collaboration with the AnVIL Outreach team, the AnVIL Portal provides details on upcoming events and training opportunities for users of AnVIL, as well as in-depth summaries of recent publications and scientific achievements.

Tools Section - On the Portal team roadmap is a tools section that provides a comprehensive list of the algorithms available to run on the AnVIL platform's data, tools, and components.

The Portal team continues to evolve this information to provide a quicker ramp-up for new users, more effective search and access methods for data, and more detailed descriptions of how to utilize the many computational components of AnVIL.

Current Terra State

Terra (https://anvil.terra.bio/) is a cloud-native platform for biomedical researchers to access data, run analysis tools, and collaborate. Workspaces are the building blocks of Terra - a dedicated space where researchers and their collaborators can access and organize the same data and tools and run analyses together. Each workspace comes with a Google Cloud bucket where data generated by a workflow analysis and notebook files are stored by default. Workspaces also provide data tables for storing and maintaining structured data similar to a spreadsheet. By including links to the data's actual location in the cloud, the data table can link data files to

workspace tools. Finally, within a workspace, users can launch batch analysis jobs or one of several interactive computing environments, including Jupyter Notebooks, R/Bioconductor, or Galaxy.

Batch analysis in Terra uses the Workflow Description Language (WDL). WDL is a way to specify data processing workflows with a human-readable and -writeable syntax. WDL makes it straightforward to define analysis tasks, chain them together in workflows, and parallelize their execution. The language makes common patterns simple to express, while also admitting uncommon or complicated behavior and strives to achieve portability not only across execution platforms but also different types of users. Whether one is an analyst, a programmer, an operator of a production system, or any other sort of user, WDL should be accessible and understandable. WDLs can be stored, shared, and described in Dockstore. WDLs are then executed in Terra using the Cromwell compute engine running in the Google Cloud Platform.

Terra Multi-cloud Integration

Terra believes the best use of cloud architecture is to bring compute to data; therefore, Terra plans to integrate with data stored across different cloud objects stores and have Terra analysis tools also work across cloud platforms. Terra architecture is built and operated on the Google Cloud Platform. However, Terra has been building a cloud abstraction layer so the Terra components - tools, storage, and compute - can be multi-cloud. We see this as the gateway to bring Terra to more researchers who may not have the ability to be cloud-agnostic. Notably, earlier this year, Broad announced a partnership with Microsoft to adapt Terra within the Microsoft cloud Azure. Bringing Terra to Azure could lead to significant adoption of the software as Microsoft has more than 168,000 partners in health care and life sciences. Building Terra to function in a multi-cloud universe lowers barriers for researchers who are required by their institution to use a particular cloud platform. We anticipate multi-cloud functionality will attract more researchers to the cloud if storage, analysis, and tools can be run on the cloud of choice.

The first step of multi-cloud integration was the Terra Project Per Workspace milestone (**Figure 18**). The Project Per Workspace reorganized how Google projects were associated with AnVIL/Terra workspaces. In the classic Terra, a Google Project could have many workspaces which complicate a multi-cloud model. By transitioning to a single Terra Project Per Workspace, each workspace, including data, analysis tools, user authentication, and other artifacts, can be more easily assigned and managed within a single cloud platform, thus simplifying the management of workspaces across clouds.

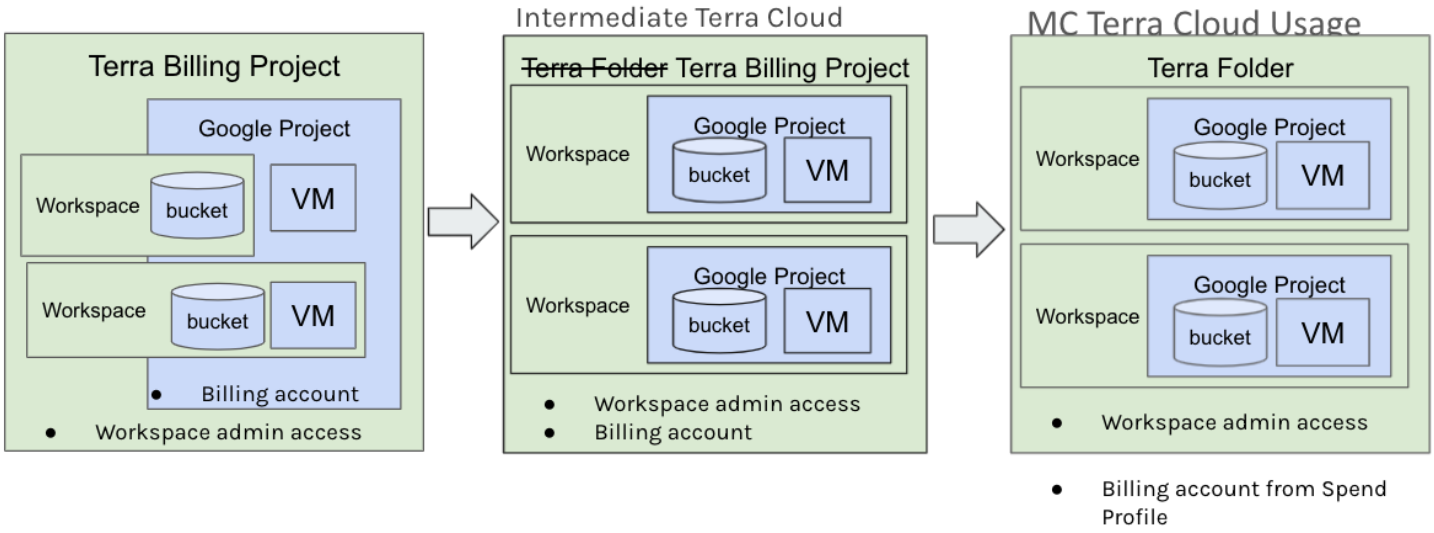


Figure 18. Updates to Terra Billing Project design to support the transition to a multi-cloud environment.

The next step of multi-cloud integration is to launch a preview of multi-cloud Terra on Azure with the ability for users to create a Terra profile linked to Azure billing project, upload and access data in Azure cloud, run best practices workflows that have been validated in Azure. This functionality is currently in development with the first prototypes to be available in 2022.

Current Data Access and Security

Terra, Gen3, and tools running within these environments, are FedRAMP compliant and operate in a FISMA-Moderate environment, and comply with all requirements set forth in NIST-800-53. This includes robust logging of data access, periodic audits, and monitoring for abnormal use patterns. AnVIL is organized into workspaces that are broken out by study registration and consent group mapping to ensure proper data access. In addition, each AnVIL workspace has an authorization domain (AD) to limit access of the workspace to only those researchers with the appropriate permissions to work with controlled data. AD protection follows a workspace when it is copied and only allows authorized access to both primary and generated data (**Figure 19**). Data access incidents are recorded and reported to the relevant parties immediately upon discovery.

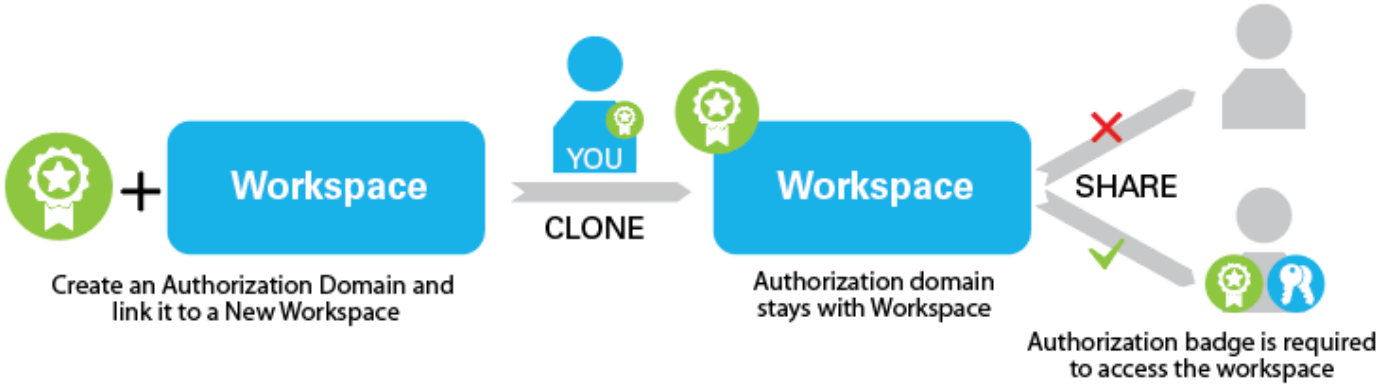


Figure 19. Data access limitation through authorization domain.

DUOS: Semi-automated Authorization and Management of Human Subject Data

Increasingly, a major challenge to data sharing is navigating the complex web of restrictions on secondary data use (**Figure 20**). Unfortunately, data use limitations are often described with unique language across the various consent forms in which they appear. Thus a DAC is left to attempt to interpret either the consent form or the original IRB interpretation to determine the official use limitations. On the other hand, researchers' data access requests are often narrative scientific proposals of varying levels of depth and specificity as to the research proposed.

Human subjects datasets often have complex and/or ambiguous restrictions on future use deduced from the original consent form, which must be respected when utilizing data. Previously, such data use restrictions were uniquely drafted across institutions, creating vast inconsistencies and requiring the investment of significant human effort to determine if researchers should be permitted to use the data. With support from our team members, the Global Alliance for Genomics and Health (GA4GH) published a solution for this ambiguous and inconsistent data sharing language in the form of their Machine Readable Consent Guidance.

As part of our efforts to enhance collaborative research by automating data use limitation alignment, we have developed the "Data Use Oversight System" (DUOS) to semi-automate and efficiently manage compliant sharing of human subjects data. The DUOS objective is two-fold, to enhance the data access committee's confidence that data use restrictions are respected while efficiently enabling appropriate data access.

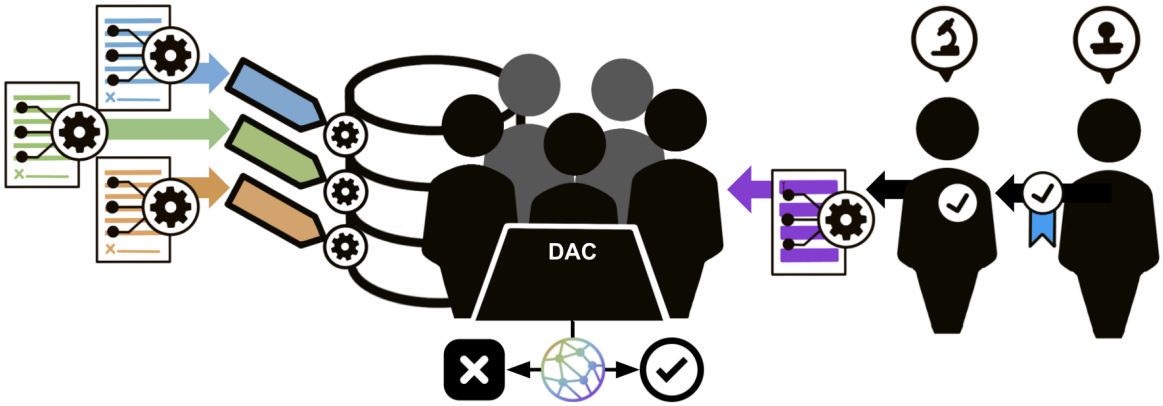


Figure 20. Overview of DUOS. DUOS leverages the Data Use Ontology (DUO) by enabling Data Depositors to describe their data use limitations with DUO terms, and Researchers to describe their research purposes with DUO terms. The result is that DACs using DUOS can compare data use limitations (left) and research purposes (right) using the same vocabulary of terms. DUOS can then enable an algorithm to compute the comparison of the data use limitations and proposed research in an attempt to replicate the decision the DAC would make (center).

To evaluate the feasibility of using machine-readable data use terms to interpret data use restrictions and access requests, AnVIL, NIH staff, and DUOS product manager conducted a pilot of DUOS with four NIH ICs (NHGRI, JAAMH, NHLBI, and NIAID). During the pilot, the ICs' datasets (N~600) were mapped to GA4GH DUO and registered in the data catalog. DACs composed of governmental and non-governmental data

custodians piloted the use of DUOS in Data Access Review (DARs) in its ability to structure use limitations and assess the accuracy of the DUOS algorithm. NHGRI and NIAID tested DUOS in real-time, NHLBI assessed the algorithm, and JAAMH will join the next pilot sprint once its DAC members have registered with DUOS. Within this pilot testing, the automated systems were found to be 87% to 90% concordant with the human DAC decisions, and in all discordant cases, the automated systems were more conservative than the human review. Importantly, the automated systems were 100% concordant across General Research Use (GRU) and Health/Medical/Biomedical Use (HMB) proposals. AnVIL, NIH staff, and the DUOS product manager will continue testing and discussing the feasibility of using DUOS for managing access to the datasets in AnVIL and managed by NIH staff with the ultimate goal of reducing bottlenecks and confusion and streamlining safe and compliant data access.

NIH Cloud Platform Interoperability (NCPI)

The NIH is the largest biomedical research agency in the world and home to many valuable data resources and platforms at the forefront of data science for data research. NIH-sponsored biomedical research is increasingly moving into cloud-based systems for large-scale data storage and analysis systems, with major cloud platforms established not only for NHGRI's AnVIL, but also for peer projects including NHLBI's BioData Catalyst (BDCat, https://biodatacatalyst.nhlbi.nih.gov), Common Fund's Gabriella Miller Kids First Pediatric Research Program (GMFK, https://kidsfirstdrc.org), and NCI's Cancer Research Data Commons (CRDC, https://datacommons.cancer.gov). Collectively, these platforms host almost 11 petabytes of genomic and related data that are currently accessible to researchers in cloud-based analysis platforms, and the scale of these datasets are growing quickly.

Yet, despite the enormous opportunity to cross-analyze data from these resources, researchers are faced with the daunting task of understanding the various technical interface differences between systems in order to analyze across them -- from a programmatic, user interface, and even policy perspective. As a result, there is great motivation for these systems to adopt consistent conventions and standards, enabling interoperability that facilitates researchers' ability to ask questions across the individual platforms.

The interoperability vision and accomplishments of AnVIL were not done in isolation but as part of a larger collaboration within the NIH. The NIH Cloud Platform Interoperability (NCPI, https://anvilproject.org/ncpi) effort was started in late 2019 with the goal of establishing and implementing guidelines and technical standards to empower end-users to analyze data across participating platforms and to facilitate the realization of a trans-NIH, federated data & compute ecosystem spanning AnVIL, BDCat, CRDC, and GMFK, along with strong ties to other NIH services such as dbGaP and the SRA. The NCPI Systems Interoperation working group has focused on leveraging the interoperability standards of Researcher Auth Service (RAS), GA4GH DRS and TRS, Fast Healthcare Interoperability Resources (FHIR), and conventions like PFB to address real-world scientific use cases.

AnVIL has been a leading advocate to advance the following critical interoperability components identified by the NCPI:

1. ***Researcher Auth Service (RAS)*** is an effort by the NIH's Center for Information Technology (CIT) to provide a common mechanism by which researchers can establish their identity and access data they are authorized to use across the systems outlined above. The RAS Application Programming Interface (API) allows seamless access to researchers for integrated data repositories. Using RAS facilitates access to both open and controlled datasets and repositories, eliminating the need to maintain multiple credentials for NIH-supported cloud platforms. RAS uses the OIDC/OAuth2 standards and leverages GA4GH Passports, providing a consistent way to describe datasets a researcher is authorized to access. Furthermore, RAS services offer increased protection via automated logging of data access. RAS uses open standards and protocols and provides integrating systems with many standards-based options for integration. RAS is part of the NIH CIT IAM General Support System (GSS) which is a Federal Information Security Management Act (FISMA) High system. As such, RAS adheres to NIST (National Institute of Standards and Technology) 800-53 and 800-57 guidelines pertaining to configuration management, least privilege, and cryptographic key establishment & management. The RAS API is defined in detail here: https://auth.nih.gov/docs/RAS/serviceofferings.html

   Currently, RAS is used for single sign-on across the AnVIL and other NCPI partners, as well as eRA Commons and other NIH websites. We are currently working on a deeper integration of RAS where the passport would also be used to authorize users to access certain protected datasets. This is currently managed through dbGaP "telemetry files", but this requires extensive manual processing and synchronization that does not scale to large numbers of users or protected datasets.

2. ***GA4GH DRS.*** The Global Alliance for Genomics and Health (GA4GH) is an international coalition formed to enable the sharing of genomic and clinical data. The GA4GH Data Repository Service (DRS) provides a generic interface to data repositories so data consumers, including workflow systems, can access data objects in a single, standard way regardless of where they are stored and how they are managed. The primary functionality of DRS is to map a logical ID to a means for physically retrieving the data represented by the ID. There are two styles of DRS URIs, Hostname-based and Compact Identifier-based, both using the drs:// URI scheme. The API defines the characteristics of those IDs, the types of data supported, how they can be pointed to using URIs, and how clients can use these URIs to ultimately make successful DRS API requests. The most recent version of this API is version 1.2 and is extensively documented here: https://github.com/ga4gh/data-repository-service-schemas

   Already each of the NCPI platforms offer DRS services for indexing their respective datasets for a total of more than 11 petabytes of data. Next steps are to deploy additional DRS services for other major NIH datasets, including within the Common Fund Data Ecosystem (CFDE) which includes the Human Microbiome Project (HMP), the 4D Nucleome, GTEx, and several other major projects.

3. ***Fast Healthcare Interoperability Resources (FHIR)*** is a standard describing data formats and elements (known as "resources") and an API for exchanging electronic health records (EHR). As patients move around the healthcare ecosystem, their electronic health records must be available, discoverable, and understandable across systems. Further, to support automated clinical decision support and other machine-based processing, the data must also be structured and standardized. The FHIR standard was

created by the Health Level Seven International (HL7) healthcare standards organization using a modern web-based suite of API technology, including an HTTP-based RESTful protocol, HTML and Cascading Style Sheets for user interface integration, a choice of JSON, XML or RDF for data representation, and Atom for results. One of its goals is to facilitate interoperation between legacy health care systems, to make it easy to provide health care information to health care providers and individuals on a wide variety of devices from computers to tablets to cell phones, and to allow third-party application developers to provide medical applications which can be easily integrated into existing systems. The FHIR standard is defined here: https://www.hl7.org/fhir/overview.html.

AnVIL completed a proof of concept FHIR server in Q1 2021. Using this proof of concept as a launch pad, AnVIL stood up an AnVIL FHIR server in Q3 2021 that has the capability to move data to the FHIR server, serve Terra files with existing AnVIL data, and perform basic searches. The next step for capability is to build out the existing FHIR app to import and search dataset metadata and build export to FHIR bundle or PFB. This work is ongoing and looking to complete Q4 2021.

Prior to releasing the FHIR server to the greater AnVIL community, we aim to test security, access, and capabilities to ensure that the FHIR development meets requirements of our security policies, data access requirements, and users. In order to test the current FHIR server and gather feedback from ideal users, the AnVIL FHIR server is accessible to the AnVIL Developer team for testing typical use cases - does FHIR work as expected and what could be improved. Feedback from the AnVIL Developer team will inform future development and eventual hand-off to Terra. We aim to test current data access security with the AnVIL developer team to ensure that data access boundaries enforced by Terra authorization domains are maintained in the FHIR server. We are also expecting the AnVIL FHIR testing to last until at least Q1 2022.

Next steps for advancing the FHIR standard within NCPI include alignment on research study and metadata representation and the identification of roadmaps for platforms around milestones services, use cases and limitations.

4. ***Portable Format for Bioinformatics (PFB)*** is an Avro-based file format that bundles schema, data, ontologies/controlled vocabularies, and pointers to data files in a single, serializable format that can be sent easily across systems and has the flexibility for different data models. PFBs are used to bring search results from hosted datasets into workspace environments that users can leverage for computational analysis. A Python library and command line interface to create, view and edit PFB files can be found here: https://github.com/uc-cdis/pypfb/#readme. More information on the PFB format and schema can be found in the pyPFB documentation.

   Currently, PFB supports different data models. We are currently working towards a PFB-lite file format for exchanging manifest information about research subjects in a cohort and associated BAM and CRAM files, which will improve interoperability. We are also working towards pre-computed PFB files with full clinical and phenotype data along with DRS references to BAM and CRAM files to improve the efficiency of data access.

5. ***Cross-Platform Search*** enables users of NCPI platforms to find relevant datasets across participating platforms. An NCPI Dataset Catalog hosted at https://anvilproject.org/ncpi/data summarizes datasets and currently lists 176 studies representing 689,301 participants in over 11 petabytes of data. The Dataset Catalog allows users to perform key-word searches as well faceted searches based on platform, focus or disease, datatype, study design and consent code. The intuitive search features of the NCPI Dataset Catalog gives a summary of the number of studies that fit the search parameters, the number of associated study participants and presents a list of study names and dbGaP IDs. Importantly, the search results study name links to a study description with a button for users to request access to the dataset via dbGaP.

   The next steps for advancing NCPI search capabilities involve the formation of a proposed Search Working Group to drive the development of use case driven search strategy. Specifically, this proposed working group will develop personas to support use cases, create guides and documentation of search components, define semantic maturity in data to enable search across all data, and build a roadmap to achieve these goals.

6. ***The NCPI Outreach and Training*** working group is tasked with spreading the word of participating NCPI platforms. Each of these platforms have a dedicated set of users who may not be aware of what peer-platforms are doing, the data they have and the tools they provide. The main mission of the NCPI Outreach WG is to prevent the formation of silos by providing unified access to key information and training resources associated with each NCPI platform. This mission has led to extensive updates to the NCPI portal at https://anvilproject.org/ncpi, including the aggregation of training and outreach materials, access to platform specific user support and social media communications, overviews of key technologies that enable interoperability, and the creation of the NCPI Dataset Catalog described above.

   In moving forward, the NCPI Outreach team will partner with the NCPI Systems Interoperation working group to highlight scientific use cases on the NCPI Portal and continue to underline developer resources and key technologies that make NCPI a success. Additionally, the NCPI Outreach team is working to document budgeting templates and other guides for managing cloud costs.

VII. Outreach and training

The AnVIL Project combines a FedRAMP-certified data management platform and scientific computing platform that is layered on top of Terra and Google Cloud Platform. This platform has a number of tools that are useful for scientists and analysts with a range of skill sets from basic visualization of data from small experiments, to massively parallel computing on hundreds of terabytes of data. The AnVIL Project is designed as a platform for the scientific community, but cloud computing is a paradigm shift for many researchers and the concepts, tools, billing, and data/people management can be intimidating for these users as they move to the cloud.

The mission of the AnVIL Outreach Group over the first funding period has been:

> *To develop scalable training, support, and incentives to make it easier for the scientific community to use the AnVIL Platform to share data, perform genomic analysis, and manage their computing.*

To tackle these challenges the AnVIL Outreach Team has adopted a number of principles to ensure we are able to support the broad range of potential AnVIL Users as they consider adopting the platform.

1. We use a persona focused model for developing user documentation and support. These personas include lab principal investigators, analysts, instructors, and consortia. For each, we develop content, activities, and support that are tailored to the unique needs of these audiences.
2. We use an agile content development approach developed at the Johns Hopkins Data Science Lab that allows us to quickly update and refresh course materials in response to changes to the underlying platform, tools, data, and policies.
3. We have both centralized and decentralized Outreach Activities that allow us to leverage common areas of need and interest among outreach communities while supporting the already existing user communities of the platforms and tools that comprise AnVIL.
    a. Our Central Team has been able to be responsive to new user needs - including additional user experience research, developing lightweight funding mechanisms, responding to requests from user communities, and supporting faculty at institutions with less computational infrastructure and support.
    b. Our Decentralized Team has allowed us to leverage the Outreach Efforts of our distinct user communities to support a broad range of users of the AnVIL Platform.
4. We develop our material using a common formatting guide with liberal open source licensing so our material can be repurposed and re-mixed to address different user needs.
5. We collaborate with the Infrastructure, Data Management, and Portal teams to ensure we get the best, most up to date information into the hands of users across different platforms.

We have adopted these strategies to overcome some important challenges to user engagement as the platform is being developed.

1. The platform is evolving over time with improvements to technologies, tools, data management, user management, and user experience. The Outreach Team has needed to adapt to often significant changes in the user experience during the first years of the project.

2. There are a number of distinct and complementary computational communities that are involved in the AnVIL Platform including Terra, Galaxy, Bioconductor, Dockstore, and Python/Jupter. Many of these communities already have their own education and outreach efforts which can be challenging for new users to navigate.
3. There are a large number of constituencies with competing demands for support and education that the Outreach Group needs to serve, from large scale consortia, to small labs, to clinical users, to individual data analysts.
4. Additional software engineering support is needed to implement user interface changes recommended by the Outreach team across the components that comprise the AnVIL including: Portal, Terra, Gen3, Dockstore, Bioconductor, and Galaxy.
5. There are significant mental, economic, and bureaucratic hurdles to users moving their analysis onto the cloud.

Using the approach described above we have been able to tackle these challenges to support users from many different communities as they move onto AnVIL.

Outreach Team Design

We have both centralized and decentralized Outreach Activities that allow us to leverage common areas of need and interest among outreach communities while supporting the already existing user communities of the platforms and tools that comprise AnVIL.

Our Central Outreach Team is based at Johns Hopkins School of Public Health at the Data Science Lab (DaSL, https://jhudatascience.org/). The DaSL is responsible for developing data science training and outreach materials that have reached more than 8 million learners around the world. The AnVIL Central Outreach Team is comprised of:

- Natalie Kucher (JHU, 100% AnVIL Effort), Project Manager
- Jeffrey Leek (JHU, 25% AnVIL Effort), MPI of AnVIL and Outreach Chair
- Frederick Tan (Carnegie, 50% AnVIL Effort), Educational Development Faculty
- Ava Hoffman (JHU, 70% AnVIL Effort, 30% GDSCN Effort), Educational Development Faculty
- Katie Cox (18% FTE Effort), Sarah Wheelan (13% Effort), Kai Kammers (10% Effort) (JHU), Supporting Educational Development Faculty
- Recently Open Searches for 2 Postdoctoral Teaching Fellows, Develop Content and Support Programs

The Decentralized Outreach Working Group is a crosscutting team with representation from Johns Hopkins University, Broad Institute, NHGRI, Gen3, Dockstore, Bioconductor, and Galaxy. This highly collaborative team works together to drive the diverse activities that further the working group's mission and objectives through biweekly Outreach Group meetings. At these meetings, work proceeds on multi-group efforts like the yearly ASHG workshops while updates are provided on individual group activities such as the Bioconductor Popup Workshops, Terra BioData Catalyst Training Webinars, and JHU AC2 Pilot Program.

Ongoing collaborations with other communities to address disparities in education and health through partnerships with Research Centers in Minority Institutions (RCMI) program, Social Determinants of Health (SDOH) Project, and the Genomic Data Science Community Network (GDSCN).

Summary of Accomplishments

Usage

A key indicator of the AnVIL platform's progress in the genomics data science community is its usage. The AnVIL Outreach team is building infrastructure to assess usage metrics, building programs to onboard users, and tracking several use cases to further the platform's mission.

Metrics

To assess usage on AnVIL, including the effectiveness of various outreach efforts, the Outreach Team has collaborated with Terra to regularly collect several important metrics (**Figure 21**). These usage metrics include several facets, such as logins, workspace creations and clones, and launches of interactive tools and workflows. Metrics are aggregated in a Terra workspace to allow interested parties to collaboratively analyze how usage patterns are impacted by training events, data availability, tool releases, and user interface changes. Access is available upon request to this workspace: https://anvil.terra.bio/#workspaces/anvil-outreach/metrics.

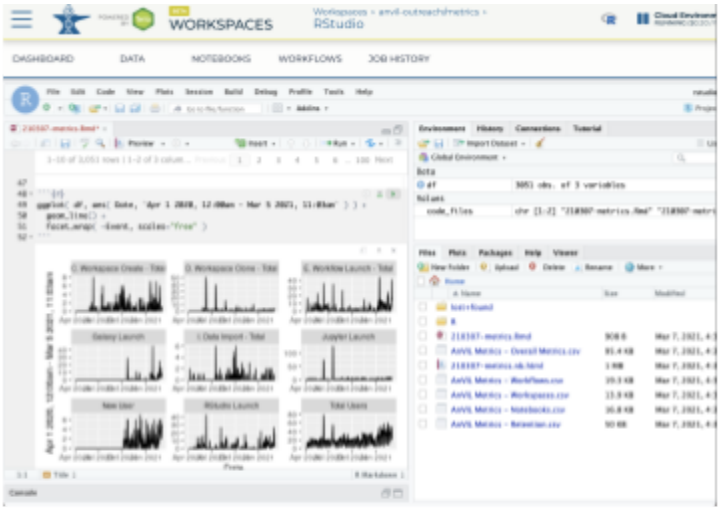


Figure 21. Screenshot of usage metrics shared through and analyzed using an AnVIL Workspace.

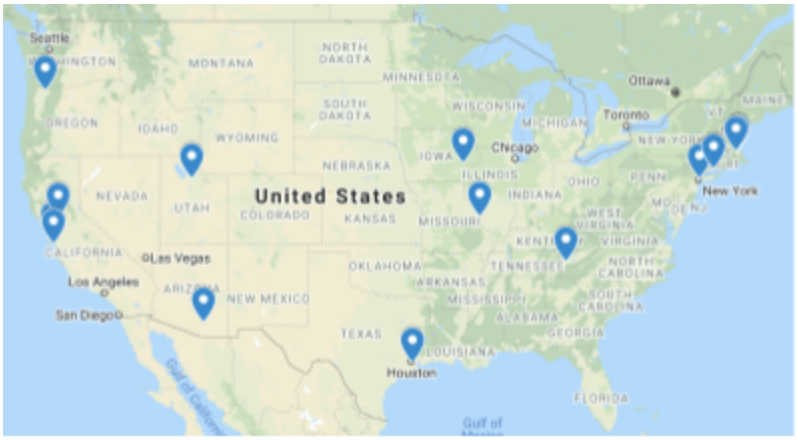


Figure 22. Geographical distribution of applications received for the AC2 Pilot Program.

AnVIL Cloud Credits (AC2) Pilot

Through the AC2 pilot, we have invited genomic researchers to submit proposals for using the AnVIL platform for large-scale data analysis with cloud computing credits supported by NHGRI through the NIH STRIDES program. This pilot program was announced in April 2021 via social media and the AnVIL Portal and resulted in 17 submitted applications from a diverse range of institutions across a two week period (**Figure 22**). An external review committee was established to rank the applications, and ultimately six groups were awarded $53,750. At this point, we are planning a second round of AC2 awards, with an RFA planned for November 2021, with a review period in December 2021 and awards distributed in January 2022. While the program is very new, already one of our AC2 awardees was able to publish the results of their efforts in a peer-reviewed journal (Padhi *et al*, 2021, Bioinformatics). Details from the original announcement can be found at https://anvilproject.org/news/2021/04/12/announcing-anvil-cloud-cost-credits-program.

Use Cases

We are test piloting a series of user studies sampling individual investigators, instructors, and research teams. While they perform their analysis on the AnVIL platform, we strive to collect in depth information about their user experience, including barriers, training gaps, and concerns in their transition to cloud based computation. These projects will provide key insights for the AnVIL platform and outreach teams on how to improve the platform to support the broad range of user communities as they transition their workflows to a cloud computing environment.

- ***DeepPilots | Johns Hopkins University*** | "A preliminary study of user experiences on the NHGRI AnVIL cloud platform" | $16,000
    - Research DeepPilots -- Makova Lab (Penn State), McCoy Lab (JHU), Wheelan Lab (JHU)
    - Teaching DeepPilots -- Bioconductor Popup Workshops, Introductory computational biology on AnVIL (Levi Waldron [CUNY])
    - Consortia DeepPilots -- GDSCN (http://gdscn.org), CSER (https://cser-consortium.org)
- ***AC2 Pilot Awardees***
    - Alex Greiner | The University of Iowa | Graduate Student, "Burden analysis of inherited cardiac arrhythmia genes in epilepsy" | $7,750
    - Melissa Suzanne Cline | UC Santa Cruz Genomics Institute | Principal Investigator, "Leveraging AnVIL and Terra for secure collaboration on genetic variant interpretation" | $5,000
    - Andrew Davidson | University of California | Graduate Student, "Comprehensive characterization of transposable element expression across human tissues" | $10,000
    - Anahita Khojandi | University of Tennessee-Knoxville | Associate Professor, "Deep Learning for Accurate Tissue-Specific Prediction of Gene Expression in Large Deeply-Phenotyped Population" | $10,000
    - Anshul Kundaje | Stanford University | Principal Investigator, "Deciphering cis-regulatory syntax of a transcription factor binding atlas with interpretable deep learning models" | $10,000
    - Tychele N. Turner | Washington University in St. Louis | Principal Investigator, "A k-mer based approach to assess copy number in PacBio HiFi data" | $10,000

Support

Events

The AnVIL Outreach team has hosted events for a variety of audiences as the platform has solidified and attracted a diverse set of users. These events include:

- ASHG 2019: Reproducible and Transparent Genomic Analysis with Galaxy
- ASHG 2019: An Introduction to Scalable Genomic Analysis with Hail
- MaGIC 2020: Massive Genome Informatics in the Cloud (MaGIC) Jamboree
- ISMB 2020: Finding and Analyzing Data in the Cloud with Gen3, Dockstore, Terra, and Galaxy
- BCC 2020: R / Bioconductor in the Cloud
- BCC 2020: Dockstore Fundamentals: Introduction to Docker and Descriptors for Reproducible Analysis
- BCC 2020: Reproducible Analysis in the Cloud with Dockstore and Terra

- BioC 2020: Cloud-based Genomics using Terra/AnVIL
- ASHG 2020: GWAS Analysis with Galaxy in AnVIL
- ASHG 2020: Find and Analyze Data in the Cloud with Gen3, Dockstore and Terra
- VADSTI 2021: Tools for Applied Data Science Using Cloud-Based Platforms
- Bioconductor Popup Workshops -- Spring 2021
    - Using R / Bioconductor in AnVIL
    - The R / Bioconductor AnVIL Package for Easy Access to Buckets, Data, and Workflows, and Fast Package Installation
    - Running a Workflow: Bulk RNASeq Differential Expression from FASTQ Files to Top Table
    - Single-cell RNASeq with 'Orchestrating Single Cell Analysis' in R / Bioconductor
    - Using AnVIL for Teaching R / Bioconductor
    - Reproducible Research with AnVILPublish
    - Managing Costs with AnVILBilling
    - Participant Stories
- BCC 2021: Introduction to the Terra/AnVIL Cloud-based Genomics Platform
- ASHG 2021: Structural variant discovery from long-read sequencing data on the cloud with Galaxy in Terra
- ASHG 2021: Reproducible Analysis of Human Pangenome Data using the AnVIL

Posters / Presentations

- Biological Data Science 2020 Conference
- T2T/HPRC Open Science 2020 Conference
- ISMB 2021: Modeling the computing requirements and costs for genomics analysis in the cloud
- Biology of Genomes (BoG) 2021 Conference
- GCC 2021: Using Galaxy File Source Plugins to Work with Remote Data
- GCC 2021: Streamlining accessibility and computability of large-scale genomic datasets with the NHGRI genomic data science Analysis, Visualization, and Informatics Lab-Space (AnVIL)

Many of the workshop agendas, recordings, and materials are available on the AnVIL Portal (https://anvilproject.org/events).

Office Hours

In September 2021, the Outreach Working Group hosted the first AnVIL Office Hours, offering users a place to communicate directly with the team about issues they encountered. Users shared their experiences getting started, described blocks to their work, and asked questions about how to move forward. The team provided possible solutions and guided users to resources that would help with their work, including a highlight of the AnVIL Discourse (https://help.anvilproject.org) where users can ask questions that come up in the future. The team aims to hold Office Hours on a regular basis going forward.

Discourse

The AnVIL Platform is composed of a federation of technologies. As such, it can be unclear to users whether a question is best answered by one support community or another. In mid-2021, we addressed this problem by launching a Discourse-based forum (https://help.anvilproject.org) (**Figure 23**). The Outreach Team helps

shepherd questions to the right community, alerting the relevant support teams that an AnVIL user needs assistance. Furthermore, as users post questions publicly, they can help answer other users' questions in addition to getting help from AnVIL support. This forum enables users to share their problems and solutions with other users, and build a community around using AnVIL. Technical support on Discourse is a collaboration between members of the JHU team and the Terra support team.

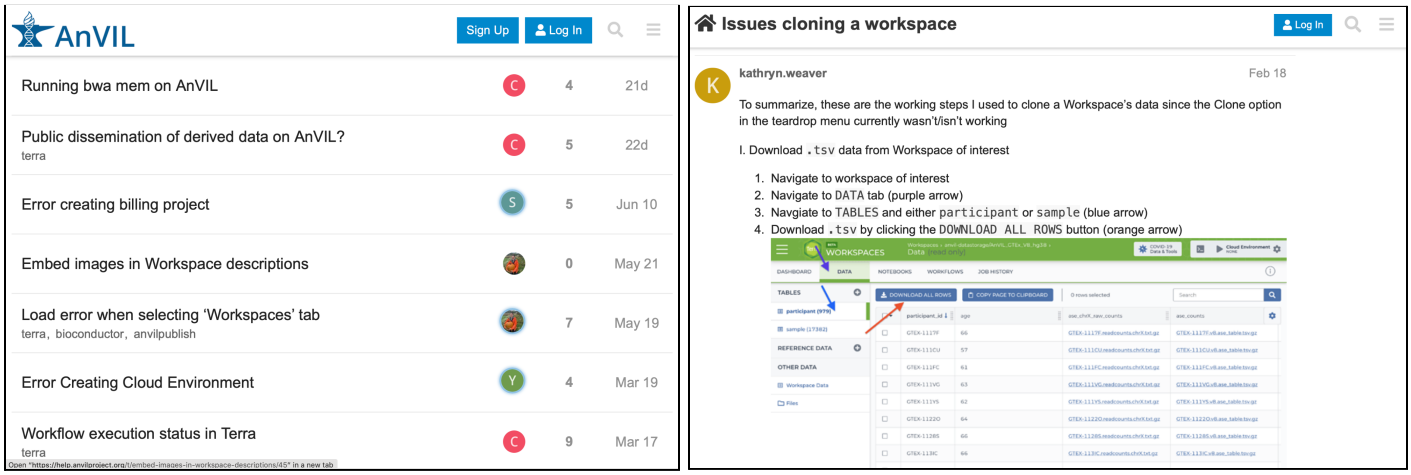


Figure 23. Screenshot of the AnVIL community support forum (left) and an example of an AnVIL user providing a quick tutorial on how to work with the GTEx workspace (right).

Content

Infrastructure

In order to keep pace with the rapid technological advancement and research breakthroughs in genomics and data science, the JHU Data Science Lab has developed outreach content templates for Bookdown (https://github.com/jhudsl/DaSL_Course_Template_Bookdown) and Leanpub (https://github.com/jhudsl/DaSL_Course_Template_Leanpub) to quickly and effectively deliver new material (**Figure 24**). In some cases, this means that material can be written once and published on the web as an HTML book and as a MOOC (Leanpub and Coursera). On the AnVIL Project, many groups are working quickly and simultaneously, leading to redundant training material. While this is useful in some cases (users benefit from material customized for their situation), we have also made it possible to author material using a common language that groups can customize (plain text Markdown). Continual platform innovations also rapidly render training material obsolete, but the templates have created a solution where changes can be published quickly and iteratively, making AnVIL more maintainable. In the spirit of collaboration, we have also worked to make documentation templates available to all groups working on the AnVIL Project.

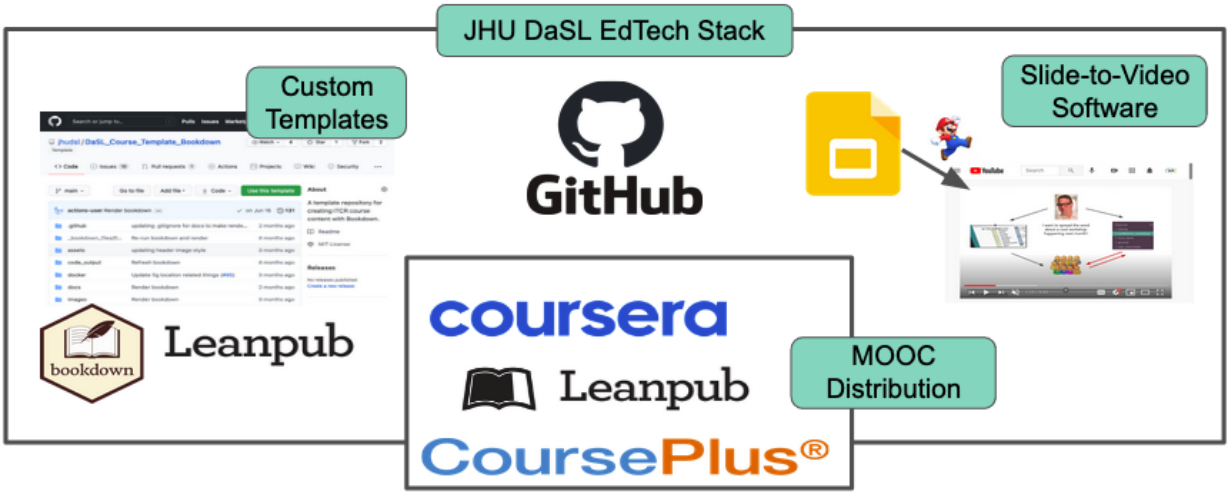


Figure 24. EdTech stack leveraged by JHU Data Science Lab for creating AnVIL outreach content.

The JHU tool "ari" (https://github.com/jhudsl/ari) converts Google Slides into videos with Google's Text-to-Speech engine (**Figure 25**). This means that the AnVIL Outreach team can collaboratively create YouTube videos by adding narration as Speaker Notes in Google Slides. This means that updating videos takes only minutes. It also increases accessibility of AnVIL content without extra effort, as closed captioning and language translation functionality are based on scripts and not third party speech recognition. As with the Bookdown and Leanpub templates, slide-to-video technology allows greater collaboration and sharing of tools among the JHU and Terra / Bioconductor / Galaxy teams.

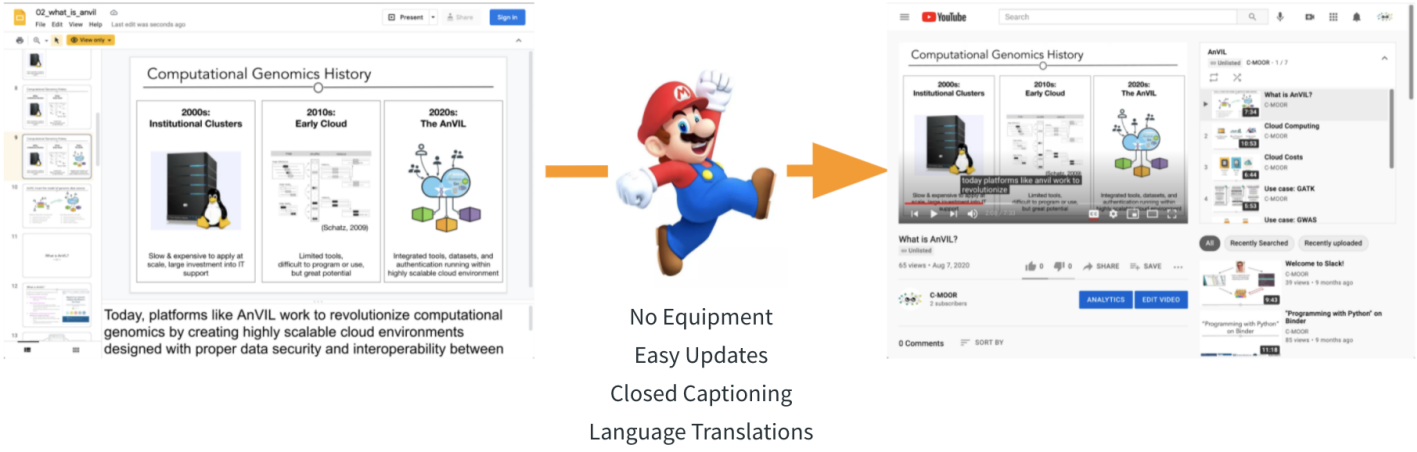


Figure 25. M"ari"o software converts Google Slides to accessible videos automatically in minutes.

Core Videos

We have created several core educational videos leveraging "ari" from the EdTech stack to add narration from speaker notes in Google Slides via Google's Text-to-Speech engine. These videos are available on the AnVIL YouTube channel and include (1) Why AnVIL? (3m 40s), which answers the questions "What is AnVIL?" and "How does it work?", (2) Staring Jupyter on the AnVIL Platform (4m 47s), which shows users how to start compute, create a notebook, and stop compute, (3) Starting RStudio on the AnVIL Platform (5m 37s), which shows users how to start compute, use RStudio to run a basic analysis, and stop compute, and (4) Starting

Galaxy on the AnVIL Platform (6m 59s), which shows users how to open Galaxy and start and stop compute (**Figure 26**). Importantly, these videos have been made using technology that makes them easy to maintain when the AnVIL platform changes, ensuring continuous support for users.

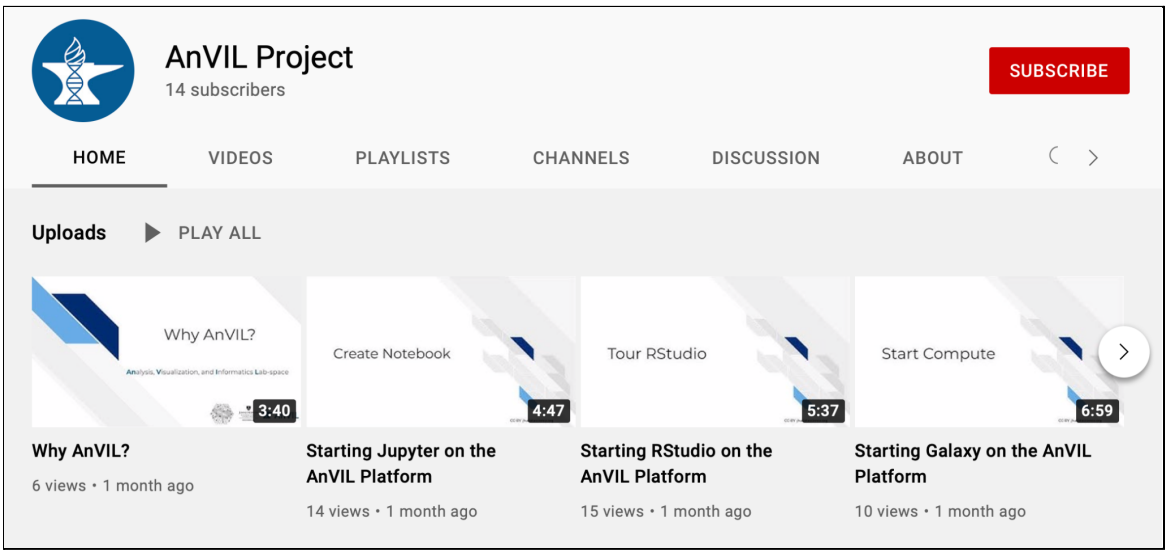


Figure 26. Screenshot of core videos on the AnVIL Project YouTube channel.

Bookdown Guides

Our Bookdown content holds a collection of guides to help new AnVIL users set up their accounts and start doing research on the AnVIL platform. By leveraging Bookdown, we have made a code-based, maintainable resource that is simultaneously visually appealing and user friendly. Our Getting Started Guide (https://jhudatascience.org/AnVIL_Book_Getting_Started/) is a detailed, step-by-step guide that focuses on three personas (principal investigators and lab managers, data analysts, and research consortia) and four main areas when working on AnVIL (Workspaces, tools, data, and workflows) (**Figure 27**).

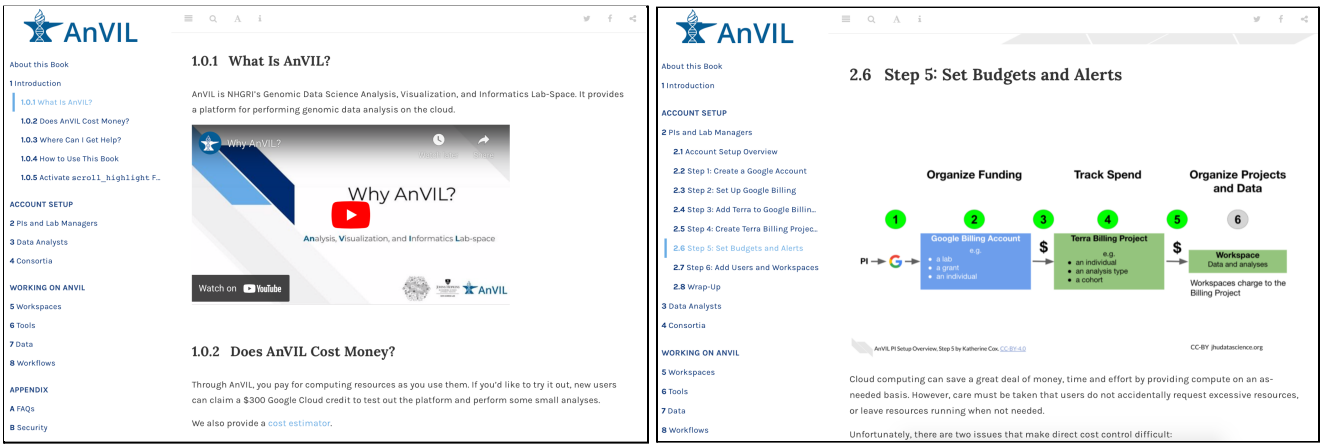


Figure 27. Screenshots of the AnVIL Getting Started Guide, which includes an introduction for new AnVIL users from three different personas (PIs and lab managers, data analysts, and consortia).

Our Teacher Guide (https://jhudatascience.org/AnVIL_Book_Teacher_Guide/) bookdown is intended to help teachers who are new to AnVIL (**Figure 28**). It helps these users set up their accounts and start developing content on the AnVIL platform for their courses. The Teacher Guide provides instructions for account setup,

developing teaching content, and offering and structuring classes (e.g class setup, student account setup, class ending, and a quick checklist). Like our Core Videos, these guides have been made using technology (Bookdown) that makes them easy to maintain when the AnVIL platform changes.

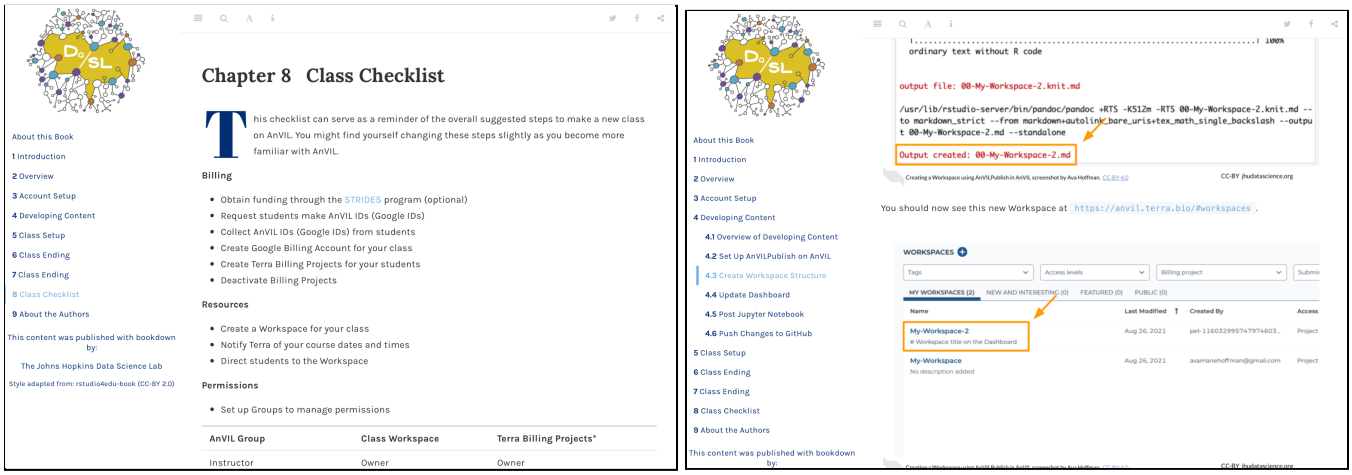


Figure 28. Screenshots of the AnVIL Teachers Guide, which helps instructors develop new course or workshop content on AnVIL.

Leanpub

The first MOOC that the Outreach Team published via Leanpub is an "Introduction to AnVIL" (http://leanpub.com/universities/courses/jhu/anvil-intro) (**Figure 29**). This course is designed to be a brief introduction to the AnVIL platform, cloud computing costs, and example workflows such as GATK, GWAS, and eQTL. The goal is to show how AnVIL can help researchers access data, scale computing resources, and democratize access to genomic data science.

Chapters

- What is AnVIL? [slides] [video]
- Cloud Computing [slides] [video]
- Cloud Costs [slides] [video]
- Use Case: GATK [slides] [video]
- Use Case: GWAS [slides] [video]
- Use Case: eQTL [slides] [video]

Figure 29. Screenshot of the opening page on Leanpub for the Introduction to AnVIL MOOC.

Exercises (via GDSCN)

SARS-CoV-2 Mutant Detection with Galaxy on AnVIL (CC-BY 4.0)

The SARS-CoV-2 Mutant Detection with Galaxy is a collection of curricular materials developed through the GDSCN (described below) specifically tailored for use on AnVIL (**Figure 30**). These materials are community

contributed, created in a maintainable format, and are reusable by many instructors on AnVIL. In collaboration with colleagues at the Morehouse University School of Medicine, we developed a set of lectures and a multi-step hands-on activity that focuses on the growing need for undergraduate students to learn cutting-edge concepts in genomics data science, including performing analysis on the cloud instead of a personal computer. The content focuses on several Genetics Core Competencies, including gathering and evaluating experimental evidence, including qualitative and quantitative data, generating and interpreting graphs displaying experimental results, critiquing large data sets and using bioinformatics to assess genetics data, and tapping into the interdisciplinary nature of science.

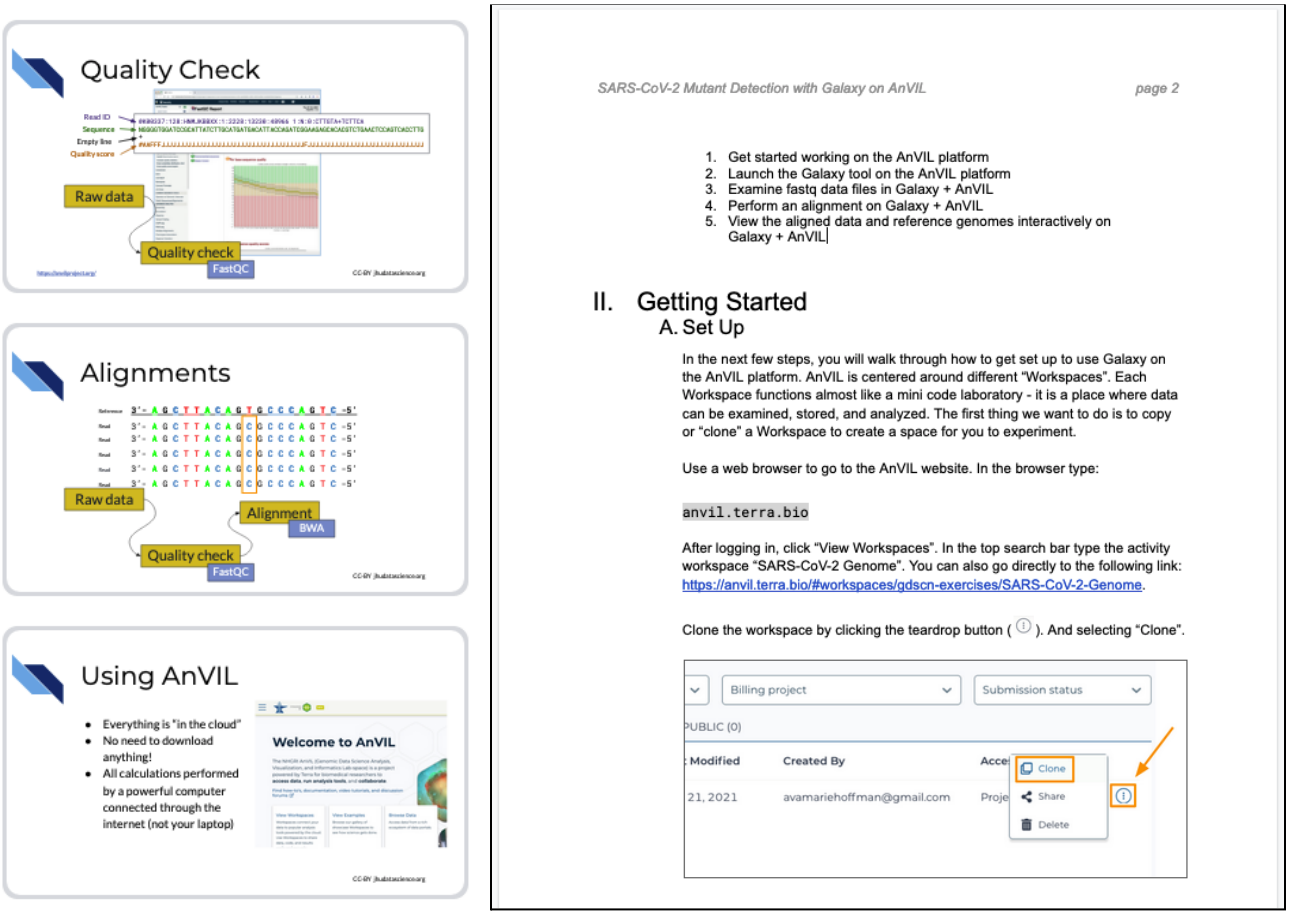


Figure 30. Lecture slides (left) and student lab activity (right) screenshots for the SARS-CoV-2 Mutant Detection with Galaxy on AnVIL Exercise.

The content aims to introduce a mutant detection bioinformatics pipeline based on a publicly available genetic sample of SARS-CoV-2, targeted toward an undergraduate biology student audience of 1-50 students with a laypersons' knowledge of genetics. Students are first introduced via lectures to the content, which will prepare them conceptually to do work on AnVIL. Lectures are prepared in both slide and video form, leveraging "ari" from the EdTech stack to add narration from speaker notes in Google Slides via Google's Text-to-Speech engine. This provides multiple modalities for instructors, who may choose to use the lecture scripts to prepare their delivery. Lecture modules (approximately ~15 minutes each) include (1) "What is a Variant?", which introduces the concept of genomic variants and and provides examples from life, (2) "Sequencing Revolution", which introduces the explosion of sequencing data and the technology that has made it possible, as well as

career paths in genomics, (3) “Alignments”, which dives into how alignments are performed and why they are important for variant/mutant detection, and (4) “Cloud Computing”, which helps students feel more comfortable with the concept of cloud computing and introduces its utility for genomics on AnVIL. We have also developed a ~30 minute introductory lecture to help students get oriented at the beginning of their lab period.

During the lesson, students work hands-on with the point-and-click Galaxy interface on AnVIL to check data, perform an alignment, and visualize their results. Short answer questions appear at each stage of the lab assignment, which can be assessed with a solutions key we have developed for instructors.

Open Case Studies (via GDSCN)

The Open Case Studies project (https://www.opencasestudies.org) is an education platform that provides self-contained, multimodal, peer-reviewed, and open-source guides for real-world examples for active experiences of complete data analyses, developed at the Johns Hopkins University Data Science Lab. Existing case studies focus on public health issues such as opioids, vaping behavior, mental health, obesity, dietary behaviours, and air pollution.

The scope of GDSCN was recently expanded to include development of five case study proposals covering a range of topics in data science including but not limited to genomics determined through collaboration with GDSCN faculty members. These proposals will have identified data sources and research questions, and will include research questions that can be explored at GDSCN institutions on the AnVIL platform.

Collaborations

Partnering with Health Disparities Communities

RCMI: As part of Howard University’s RCMI program, a Virtual Applied Data Science Training Institute (VADSTI) was created to advance education and research by providing training to support foundational pillars of data science, such as programming and analysis of large datasets. In support of this mission, the AnVIL Outreach Working Group presented a module of hands-on workshops titled “Tools for Applied Data Science Using Cloud-Based Platforms” (**Figure 31**). In this workshop, attendees were introduced to the AnVIL platform.

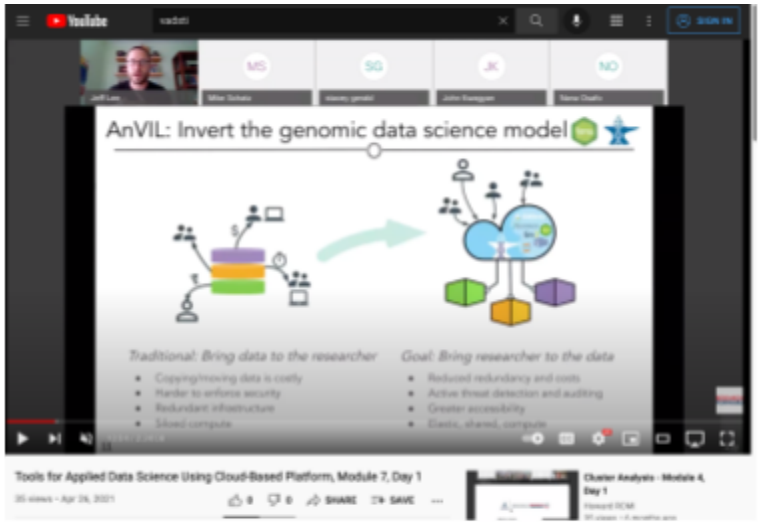


Figure 31. Two day hands-on training event using the AnVIL Platform for the VADSTI spring series.

SDOH: The National Institute of Minority Health and Health Disparities fund projects to assess the conditions in which people are born, grow, live, work and age. To accelerate the scientific understanding of how these factors impact human health, the PhenX project has been established to collect high-quality standard measures and generate datasets to better

understand the impact SDOH have on minority populations. The AnVIL team is in regular contact with SDOH researchers to support them as they onboard their data and analysts to the AnVIL Platform.

Partnering with GDSCN - The Genomic Data Science Community Network

Note: Funded through a separate contract to JHU (75N92020P00235)

There has been an explosion of measurement technologies that have dramatically decreased the cost of data collection and storage across the biomedical research spectrum from imaging, to genomics, to clinical and public health records. Huge amounts of accessible data mean that biomedical research is in a phase transition to becoming a leading data science. At the same time a range of new cloud-computing platforms make it possible for anyone with a web browser and an internet connection to access these data and begin to make scientific discoveries. This represents an incredible opportunity to bring new voices to biomedical research in a way that wasn't possible just five years ago.

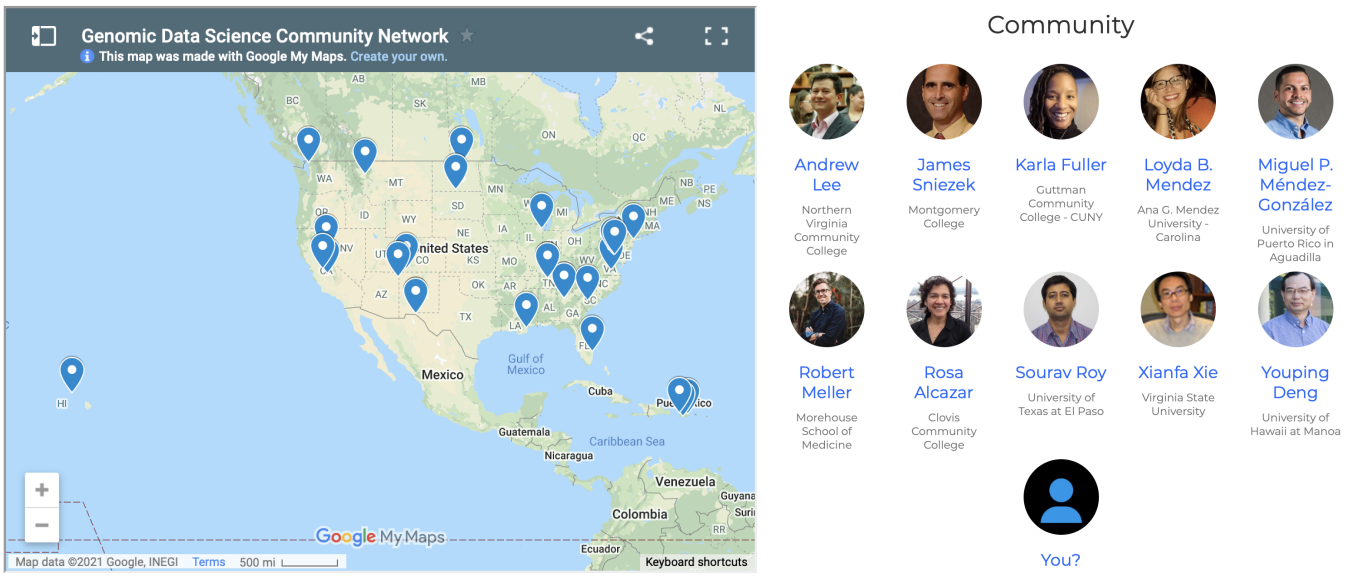


Figure 32. Geographical distribution of GDSCN member institutions (left) and profiles of several founding members (right).

Cloud-based data science platforms have the potential to democratize access to genomic data science, but as experience has consistently demonstrated, simply offering online courses or platforms does not increase engagement in traditionally underserved student populations. There is an urgent need to reach out to instructors at these institutions and provide them with the support they need to take advantage of these new resources coming online. This outreach must be designed to create opportunities that are useful for faculty at universities not traditionally engaged in genomic research such as Historically Black Colleges and Universities (HBCUs), Minority Serving Institutions (MSIs), Tribal Colleges and Universities (TCUs) and Community Colleges (CC).

Addressing this need, we launched the Genomic Data Science Community Network (http://www.gdscn.org) to define a strategy for advertising, training, and ongoing support to allow students and researchers from these institutions to learn about how to use cloud tools to perform biomedical data science and to leverage these technologies to engage in cutting edge research at their home institutions (**Figure 32**). The kickoff meeting was held in March 2021 to develop and share training materials on the AnVIL and an online training and research

symposium was held in May 2021. A third symposium is planned in November 2021 to share a modular course developed for implementation in the classroom. We hope this will serve as a launching point for additional collaborations around encouraging diverse participation in cloud data science.

Future Directions

Vision for the Next 5 Years

The Outreach Team has been increasingly engaged across the AnVIL Platform and in deep collaboration with other working groups. Our mission,

> *To develop scalable training, support, and incentives to make it easier for the scientific community to use the AnVIL Platform to share data, perform genomic analysis, and manage their computing,*

Has an ultimate goal which is:

> *To build a happy, big, diverse, vibrant AnVIL user community doing teaching & research.*

As the platform matures and the technology and data infrastructure hardens there will necessarily be a transition to more significant effort on the user focused side of platform development. This presents an exciting opportunity to expand the scope of the Outreach Team to a User Success model that supports the entire life cycle of researchers moving onto the AnVIL.

Our efforts so far have focused on the top of the "user engagement funnel" (**Figure 33**), which includes getting new users aware of the platform (MOOCs, Social Media, events like GDSCN & VADSTI), helping users evaluate the platform (MaGIC Jamboree, ASHG Hands on Event, Getting Started Guides), and improving our Outreach tech stack. We have also had a few pilot projects aimed at user intent to join the platform (AC2, DeepPilots), and user loyalty (Discourse Community).

Figure 33. User acquisition funnel to increase user engagement on AnVIL.

Going forward we have the opportunity to leverage the incredible work the AnVIL Infrastructure, Data, Portal, and Data Analysis Tools teams have accomplished to make AnVIL a platform with a user base that spans communities and provides significant computational and data resources to the genomics community broadly for basic and clinical research & education.

To do this we will continue to employ a Centralized/Decentralized approach that allows us to both accomplish core User Success activities and simultaneously support diverse computing communities. Here we envision a three component model for addressing User Success across the whole AnVIL user lifecycle.

Three Complementary Teams

Over the next five years we see the demand for the User Success Team growing significantly as the platform matures. We are already seeing the demand for supporting onboarding of consortia, ongoing recruitment of new users, supporting existing users, reducing economic and user barriers to success, and growing the AnVIL community.

To accomplish our ambitious goals for the platform we propose a team comprised of three components:

- **Content Developers**
    - Content developers will continue to produce and maintain open source, agile educational resources that adapt to changes in the platform, user experience, and policies and procedures for data management. They will deploy this content via open websites, user communities, massive online open courses, open source exercises, and templated workshop materials that can be leveraged by the other teams.
- **Engineering Team**
    - We envision a user experience engineering team who can rapidly prototype user experience interfaces to test with our users and help support the core Infrastructure team to scale these into AnVIL-wide user experiences. There is a particular need for testing and improvement to user interfaces around billing, cost control, user management, and data management - both to recruit new users and to support existing users on AnVIL.
- **Programmatic Support**
    - The programmatic team will focus on lower throughput, deeper interaction user support activities. For example, they will support user studies and grants programs (like AC2/DeepPilots), Community Building (like the Discourse Community and the Genomic Data Science Community Network), Consortia engagements (e.g. RCMI, GREGoR, All of Us), as well as hosting live in person or synchronous training events at conferences like BioC, GCC, and ASHG.

Strategic Initiatives

Just as with the first iteration of the AnVIL Project there are a core set of strategic challenges that we can meet by focusing on a few concrete user experience initiatives. These challenges are motivated by our early stage user engagement studies with DeepPilots and AC2 and include:

a. ***Cost Management*** - This is perhaps the single greatest barrier we have identified to adoption for users who are transitioning to cloud platforms. We will work to improve user interfaces, documentation, and support around cost management.
b. ***Team Management*** - This is the second largest barrier to adopting a scalable cloud computing platform - knowing who is on your team, what data they have access to, what they are doing and where their compute artifacts are located, and what their demands and costs are. We will work to improve user interfaces, documentation, and support around people management.
c. ***Data Management*** - Data onboarding including data management, tidying, documentation and ingestion are challenges particularly for consortia and large data generators. Similarly, individual research groups migrating their analysis to the cloud need help during the data life cycle of bringing, creating, deleting,

archiving, and sharing the data that they create. We will work to develop interfaces, tooling, documentation, and support that provide a more scalable approach to data ingestion and management.

d. ***User Studies*** - While we have identified some core user needs from our initial user studies, there is significant work to be done on studying the ways that people use AnVIL, pain points, and churn points where people chose to leave the platform. We will perform user studies to identify these pain points and use our Engineering Team to prototype solutions for scaling by the other teams.
e. ***Remixable Events*** - There is a dramatic increase in demand for events both for large audiences and small bespoke audiences. We propose to develop a set of remixable, off the shelf open source materials that can be used by the decentralized outreach teams to quickly spin up and host events - like the Bioconductor pop-up events from the initial phase of AnVIL.
f. ***Scalable Support*** - We will continue to invest in AnVIL Discourse and building the AnVIL community. Ultimately, support teams will be bandwidth limited as our user base grows. The most successful user communities (like the Bioconductor and Galaxy communities) rely on an engaged user base that actively gives back through discussion boards, events, and materials.
g. ***Scalable Resources*** - We will continue to develop agile, open source, free, easily maintainable materials touching on a variety of use cases that can be used by the broadest population of user communities.

Taken together we believe these efforts can help us accomplish the ambitious vision:

> *To build a happy, big, diverse, vibrant AnVIL user community doing teaching & research.*