NHGRI Analysis, Visualization, and Informatics Lab-space (AnVIL) Executive Summary

April 20, 2022

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 600,000 samples (4.56 Pb 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 289 public workspaces and 54 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 1,394 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 2,083 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 have launched two cohorts of the AnVIL Cloud Credits Program*** known as AC2 and AC3. Through this program, awardees are granted upto $10,000 in cloud credits to conduct cloud-based genomics research on the AnVIL Platform. In the pilot round AC2, 6 awards were made and resulted in a publication (Padhi *et al*, 2021, Bioinformatics) and a conference paper accepted to the Annual International Conference of the IEEE Engineering in Medicine and Biology Society in July 2022. The AC3 program expanded to 13 awards administered in March 2022.

- ***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. The Pharmacogenomics Clinical Annotation Tool (PharmCAT) is one of such tools which has already been implemented in Galaxy and is available through AnVIL. We have also integrated seqr (https://seqr.broadinstitute.org/) as an analysis platform for Mendelian diseases.

- ***The launch of major efforts through the NIH Cloud Platform 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/ as well as a catalog of datasets available across the participating platforms https://anvilproject.org/ncpi/data. Furthermore, researcher use cases were developed within the NCPI effort which span the five NCPI stacks (AnVIL, BioDataCatalyst, CRDC, and Gabriella Miller Kids First, NCBI). 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 a savings of over $500,000 in egress fees in the last year.

Scientific Accomplishments and Selected Publications

- ***The AnVIL leadership published a perspective article at Cell Genomics*** (Schatz, Philippakis, *et al*, 2022, Cell Genomics). This manuscript documents the broader scientific achievements and contributions of the AnVIL project.

- ***A technology feature has been published in Nature describing the role of Terra*** in sharing and analyzing genomic data in the cloud (https://www.nature.com/articles/d41586-021-03822-7).

- ***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.

- ***Also supported by the AnVIL Cloud Credits program, Dr. Anahita Khojandi and her group successfully published a conference paper*** within the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2022) on the modification of random forests for gene expression prediction associated with obesity.

- ***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*, 2022, Science). 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*, 2022, Science). 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*, 2022, Science) and a discussion of variation within satellite regions and other repetitive regions in the newly resolved components of the reference genome (Altemose *et al*, 2022, Science); (Hoyt *et al*, 2022, Science); (Gershman *et al*, 2022, Science); as well as an epigenetic study of previously unresolved sequences (Vollger *et al*, 2022, Science).

- ***As a second example of consortiums using AnVIL, the Centers for Mendelian Genomics (CMG) recently published a perspective on the status of the project*** (Baxter *et al*. 2022, Genetics in Medicine). 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.