HHS Public Access

Author manuscript

Cell Genom. Author manuscript; available in PMC 2022 February 22.

Published in final edited form as:

Cell Genom. 2022 January 12; 2(1): . doi:10.1016/j.xgen.2021.100085.

Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space

Michael C. Schatz[1,2,*], **Anthony A. Philippakis**[3,*], **Enis Afgan**[1], **Eric Banks**[3], **Vincent J. Carey**[4], **Robert J. Carroll**[5], **Alessandro Culotti**[3,6], **Kyle Ellrott**[7], **Jeremy Goecks**[7], **Robert L. Grossman**[6], **Ira M. Hall**[8], **Kasper D. Hansen**[9], **Jonathan Lawson**[3], **Jeffrey T. Leek**[9], **Anne O'Donnell Luria**[3], **Stephen Mosher**[1], **Martin Morgan**[10], **Anton Nekrutenko**[11], **Brian D. O'Connor**[3], **Kevin Osborn**[12], **Benedict Paten**[12], **Candace Patterson**[3], **Frederick J. Tan**[13], **Casey Overby Taylor**[14], **Jennifer Vessio**[1], **Levi Waldron**[15], **Ting Wang**[16], **Kristin Wuichet**[5]
AnVIL Team

[1]Department of Biology, Johns Hopkins University, Baltimore, MD, USA

[2]Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA

[3]Broad Institute of MIT and Harvard, Cambridge, MA, USA

[4]Harvard Medical School, Harvard University, Cambridge, MA, USA

[5]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA

[6]Center for Translational Data Science, University of Chicago, Chicago, IL, USA

[7]Biomedical Engineering, Oregon Health & Science University, Portland, OR, USA

[8]Yale School of Medicine, Yale University, New Haven, CT, USA

[9]Department of Biostatistics, Johns Hopkins University, Baltimore, MD, USA

[10]Department of Biostatistics and Bioinformatics, Roswell Park Comprehensive Cancer Center, Buffalo, NY, USA

[*]Correspondence: mschatz@cs.jhu.edu (M.C.S.), aphilipp@broadinstitute.org (A.A.P.).

CONSORTIA

The members of the AnVIL team are Anne O'Donnell Luria, Alessandro Culotti, Alexander Baumann, Andrew Rula, Anthony A. Philippakis, Anton Kovalsy, Brian D. O'Connor, Clare Bernard, Derek Caetano-Anollés, Eric Banks, Geraldine A. Van der Auwera, Jonathan Lawson, Justin Canas, Kaan Yuksel, Kate Herman, M. Morgan Taylor, Marianie Simeon, Michael Baumann, Qi Wang, Robert Title, Ruchi Munshi, Sushma Chaluvadi, Valerie Reeves, William Disman, Salin Thomas, Allie Hajian, Elizabeth Kiernan, Candace Patterson, Namrata Gupta, Trish Vosburg, Frederick J. Tan, Ludwig Geistlinger, Levi Waldron, Marcel Ramos, Sehyun Oh, Dave Rogers, Frances McDade, Mim Hastie, Nitesh Turaga, Jeffrey T. Leek, Kasper D. Hansen, Alexander Ostrovsky, Alexandru Mahmoud, Dannon Baker, Dave Clements, Enis Afgan, Jennifer Vessio, Katherine E.L. Cox, Keith Suderman, Nataliya Kucher, Sergey Golitsynskiy, Stephen Mosher, Samantha Zarate, Casey Overby Taylor, Michael C. Schatz, Sarah J. Wheelan, Kai Kammers, Vincent J. Carey, Ana Stevens, Carolyn Hutter, Christopher Wellington, Elena M. Ghanaim, Ken L. Wiley, Jr., Shurjo K. Sen, Valentina Di Francesco, Denis Yuen, Jeremy Goecks, Kyle Ellrott, Brian Walsh, Luke Sargent, Vahid Jalili, Anton Nekrutenko, John Chilton, Lori Shepherd, Martin Morgan, B.J. Stubbs, Ash O'Farrell, Benedict Paten, Benton A. Vizzier, Jr., Charles Overbeck, Charles Reid, David Charles Steinberg, Elizabeth A. Sheets, Julian Lucas, Kevin Osborn, Lon Blauvelt, Louise Cabansay, Noah Warren, Brian Hannafious, Tim Harris, Robert L. Grossman, Radhika Reddy, Eric Torstenson, Kristin Wuichet, Robert J Carroll, M. Katie Banasiewicz, Ting Wang, Haley J. Abel, Jason Walker, and Ira M. Hall.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2021.100085.

[11]Department of Biochemistry and Molecular Biology, The Pennsylvania State University, State College, PA, USA

[12]UC Santa Cruz Genomics Institute, UC Santa Cruz, Santa Cruz, CA, USA

[13]Department of Embryology, Carnegie Institution, Baltimore, MD, USA

[14]Departments of Medicine and Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA

[15]Department of Epidemiology and Biostatistics, City University of New York Graduate School of Public Health and Health Policy, New York, NY, USA

[16]Department of Genetics, Washington University of St. Louis, St. Louis, MO, USA

SUMMARY

The NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL; https://anvilproject.org) was developed to address a widespread community need for a unified computing environment for genomics data storage, management, and analysis. In this perspective, we present AnVIL, describe its ecosystem and interoperability with other platforms, and highlight how this platform and associated initiatives contribute to improved genomic data sharing efforts. 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. By inverting the traditional model of data sharing, the AnVIL eliminates the need for data movement while also adding security measures for active threat detection and monitoring and provides scalable, shared computing resources for any researcher. We describe the core data management and analysis components of the AnVIL, which currently consists of Terra, Gen3, Galaxy, RStudio/Bioconductor, Dockstore, and Jupyter, and describe several flagship genomics datasets available within the AnVIL. We continue to extend and innovate the AnVIL ecosystem by implementing new capabilities, including mechanisms for interoperability and responsible data sharing, while streamlining access management. The AnVIL opens many new opportunities for analysis, collaboration, and data sharing that are needed to drive research and to make discoveries through the joint analysis of hundreds of thousands to millions of genomes along with associated clinical and molecular data types.

INTRODUCTION

The last 20 years have seen tremendous growth in human genomics, with millions of human genomes sequenced so far and many millions more to be sequenced in the near future.[1,2] These data, combined with ever-growing amounts of single-cell and functional genomics data, electronic medical records, and other biomedical data, have the potential to substantially enhance our understanding of the basic processes for healthy life as well as to revolutionize the treatment of disease. This research will be accomplished, in part, by aggregating and synthesizing data using new computational, statistical, and machine-learning methods, combined with new high-throughput experimental methods that can systematically evaluate large numbers of candidate relationships. However, reaching these ambitious goals requires us to embrace new paradigms for computational research where

cloud computing plays a central role; there is simply no other way to effectively share and analyze data at these scales.

In this perspective, we report the development of the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL) to enable large-scale data management, sharing, and analyses in a cloud-computing environment. We reflect on the past, present, and future of genomic data sharing and analysis, providing perspective on how the AnVIL meets current challenges and advances these efforts in order to enable population-scale analyses on federated datasets and facilitate collaboration. To start, we briefly describe some of the major ongoing initiatives to promote data sharing in genomics and highlight some of the major limitations of the traditional data sharing model where datasets are copied across institutions. We then present an inverted form of data sharing, where instead of copying data to multiple institutions, researchers connect to remote datasets via centralized cloud platforms, and describe how this can enhance analysis, collaboration, and data sharing. In the following sections, we describe the AnVIL system architecture and the datatypes and data models used by the AnVIL and discuss some of the communities that are engaging with the AnVIL. We then describe some of the key interoperability technologies available, especially the Global Alliance for Genomics and Health (GA4GH; https://www.ga4gh.org) standards, which enable researchers to seamlessly transition across cloud platforms.[2] In the final section, we present our outlook on the future of genomic data sharing and analysis. The AnVIL portal is publicly accessible at https://anvilproject.org.

HISTORY OF GENOMICS DATA SHARING

Genomics has become a central component to the study of many facets of biology and medicine.[3,4] Across ancestry analysis,[5,6] disease and trait associations,[7,8] developmental biology,[9,10] and many other fields, large-scale genome and genomics sequencing has grown tremendously over the past few decades, driven in large part by the technological improvements that have substantially decreased the cost and time required for sequencing.[11] For example, the National Institutes of Health (NIH) National Human Genome Research Institute (NHGRI) Centers for Common Disease Genomics (CCDG) and Centers for Mendelian Genomics (CMG) programs seek to identify the genetic components of many major common and rare diseases through the sequencing of more than one hundred thousand genomes.[4] Internationally, several major genomics projects are in progress, such as the establishment of the UK Biobank with genetic and clinical data from more than 500,000 volunteers from across the UK.[12] Additional nation-scale initiatives include Iceland's DeCode project, Finland's FinnGen project, China's Genome Sequencing Archive, the Korean Reference Genome Database, the Saudi Human Genome Program, and the Integrated Biobank of Luxembourg, which together provide hundreds of thousands of human genomes and related data.[13–17] Furthermore, GA4GH hosts a catalog of genomic data initiatives aimed at aggregating global resources for sharing clinical and genomic data.[2]

The scale of these projects opens many new opportunities for discovery that would not otherwise be possible, especially for detecting weak associations with rare variants that can only be measured over large cohorts.[18] However, this scale of sequencing also introduces major new technical challenges that require overhauling how genomics and genomics data

science are performed. Most urgently, it has become increasingly impractical to perform genomics research by replicating datasets across institutional computing clusters, leading us to reconsider how genomics data can be shared and analyzed.[19]

Because the power of genomics is often only realized through large-scale data aggregation, genomics has developed a strong tradition for collaborative research and the open sharing of data. Most famously, this tenet was codified by the global leaders of the Human Genome Project in 1996 as the "Bermuda Principles," where they agreed that all human genomic sequence information generated by the project should be made freely available and entered into the public domain within 24 h after generation.[20] These principles were established to maximize the benefit of the data to society, especially as private companies during this era were beginning to apply for patents around human gene sequences.[21] These core principles were later extended in 2009 by the "Toronto Agreement," which established the rules for sharing data pre-publication,[22] and later in 2015 in the United States by the NIH Genomic Data Sharing Policy, which requires all large-scale sequencing data funded by the NIH to be openly shared.[23] Complementing the efforts by the funding agencies, many major scientific journals now require data to be deposited into public databases before papers can be published, especially journals serving the genomics community. *Cell Genomics* requires that datasets and code be made publicly available earlier, at manuscript submission.[24–26]

In response to these requirements for data sharing, several large repositories have been established for storing and sharing genomics data. For high-throughput sequencing data, the NIH National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) along with international partners at the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) European Nucleotide Archive and the DNA Data Bank of Japan (DDBJ) have formed the International Nucleotide Sequence Database Collaboration (INSDC). This collaboration has emerged as the largest publicly available repository of sequence data, with over 50 petabases (Pbp) of data currently available through multiple cloud providers and institutional servers.[27–30] Within the United States, the closely related NCBI Database of Genotypes and Phenotypes (dbGaP) was developed to archive and distribute genomics and related data from studies that have investigated the interaction of genotype and phenotype in humans.[31] This database currently manages access for 7,582 datasets in 1,232 studies, most of which are controlled access, where researchers must apply for access to the datasets to an NIH Data Access Committee (DAC) that evaluates if the research goals are consistent with patient consent forms and any constraints identified by the institutions that submitted the data.

However, as valuable as these and related databases have become, they are generally static resources that do not allow detailed analyses to be performed directly within these systems. Instead, researchers using these data and researchers involved in large sequencing efforts most often begin by downloading the data to an institutional computing cluster for analysis.

INVERTING THE MODEL OF DATA SHARING

The traditional model of genomic analysis has been centered around institutional computing clusters where researchers install and maintain their own suite of computational tools to

analyze the datasets that are stored directly within their data center. This model presents a high level of flexibility and control for an individual researcher, but the siloed nature of this model introduces several major barriers and inefficiencies. To start, this model leads to a redundant infrastructure where each institution establishes its own data center and creates major administrative inefficiencies where many of the same analysis tools must be deployed and maintained within each center. Software management tools like bioconda[32] or integrated analysis suites such as Bioconductor[33] aim to simplify such installations, but maintaining software remains a huge burden in aggregate considering the large number of data centers and users involved.

This model is particularly challenging for collaborative analysis, as it requires data to be copied from one data center to another, which becomes more difficult and costly as the datasets increase in size. For example, a moderately large project, such as the 1000 Genomes Project, which contains the CRAM files for 3,202 genomes in the extended collection,[5] is 73 TB and requires several days to make a single copy over typical institutional internet connections. Larger studies, such as the recent TopMed release that included whole-genome sequence data for 53,831 individuals[7] and is approximately 2 PB in size for the CRAM files, will require several weeks to several months to download a single copy of the dataset. Equally important, reproducibility is very challenging in such a paradigm as it becomes increasingly difficult to record the provenance of how files are created across systems. In extreme cases, incompatible or conflicting versions of a tool or a dataset could be used by different groups, leading to scientifically invalid results.

A much more scalable model for collaborative research is to invert the model of data sharing: instead of moving data to each researcher, researchers virtually move to the data through the use of cloud-computing resources[19,34,35] (Figure 1). This way, only a single copy of the data needs to be maintained, which can then be accessed and analyzed by any number of researchers. This model introduces substantial advantages, including reduced redundancy and lower costs for data storage and greater flexibility in computing resources. Notably, computing in the cloud is “elastic,” meaning that additional computational resources can be dynamically added to match the needs for the analysis to be performed at a given time. Crucially, these resources can also be scaled down after an analysis is complete to limit the costs involved. This model is also much more efficient to manage, as software only needs to be installed or updated in one location for all users to benefit. Finally, centralized services, especially intrusion detection and auditing, can be far more detailed to ensure data security for protected datasets.

Such web- and cloud-based resources have a strong and growing role in genomics, starting with ubiquitous and classic examples such as the NCBI BLAST server[36] or the UCSC Genome Browser.[37] Another rich example is Galaxy,[38,39] an open, web-based computational workbench for performing accessible, reproducible, and transparent genomic science with features for executing scientific workflows, data integration, and data and analysis persistence. Even more recent is the NCI Cloud Pilots program, which supports three complementary cloud-based platforms that provide secure on-demand access to cancer datasets, analysis tools, and computing resources.[40] As valuable as these resources have proved to be, there is a need for wider analysis and data management capabilities that

can integrate data across multiple cohorts and multiple datatypes while providing very flexible analysis. Ideally, such a cloud-based system would offer everything possible from an institutional data center along with the additional benefits for scalability, elasticity, and collaboration that are afforded by a cloud platform. Furthermore, security is essential for human genetics research, and cloud systems offer enhanced capabilities for data encryption, logging, auditing, and intrusion detection that are not always available within institutional data centers, especially smaller clusters managed by individual research groups.

AnVIL SYSTEM ARCHITECTURE

In response to these needs, the AnVIL team, with the support of the NHGRI, have developed the Genomic Data Science AnVIL. 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(application programming interfaces) for programmatic access in several popular programming languages. The compute environment for the AnVIL is currently built on the Google Cloud Platform (GCP) to enable massive scalability and capacity for users within a robustly established security perimeter authorized for the storage and analysis of controlled access datasets. Specifically, the AnVIL is a Fe-dRamp-certified computing environment, and it complies with all requirements set forth in NIST-800–53. By providing a standardized method for security and risk assessment, the United States' government-wide program known as FedRAMP promotes the adoption of secure cloud services across the United States federal government.[41] This includes robust logging of access to data, periodic audits by third-party analysts, and monitoring for abnormal use patterns. We are also planning to extend the AnVIL to other cloud platforms to offer the most flexibility and capabilities for our users, especially in order to respect governmental guidelines that limit data sharing on certain cloud platforms because of privacy or security considerations.

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 new synthetic cohorts to empower new discoveries out of existing datasets. Similar to how a laptop or personal computer has multiple applications (e.g., web browser, email client, word processor, messaging client, etc.) running within a common operating system and file system, the AnVIL offers several analysis components that can be independently launched and yet interrelate to each other through a common file system and APIs (Figure 2). The analysis components are broadly characterized into 3 major categories: (1) those supporting data management querying, especially Gen3, (2) those supporting batch computing, especially through the use of the WDL on Terra and the closely related Dockstore[42] for sharing and distributing workflows, and (3) interactive computing using popular analysis suites such as R/Bioconductor, Jupyter Notebooks, and Galaxy. Through these components, more than 10,000 analysis tools and workflows are immediately available for a wide variety of analyses in genomics and beyond. This includes population-scale variant calling from genome sequence data with GATK or freebayes[43,44] including with the new Telomere-to-Telomere CHM13 reference genome,[45] gene expression analysis for both bulk and single-cell datasets,[46–48] methylation analysis,[49] COVID-19 viral genomics

analysis workflows,[50,51] and thousands more. Additionally, these components support reproducibility and reusability as methods and workflows that are deposited in Dockstore are assigned DOIs, cohorts (synthetic or designed) can be referenced via Terra workspace URLs, and we are developing technologies for versioning and publishing workspaces with DOIs that can scale to millions of files and petabytes of information.

Notes S1–S3 describe three example AnVIL workspaces for applications on human genome analyses and variant calling, gene expression, and *de novo* genome assembly.

Note S1 displays the workspace for germline variant calling using GATK4. Using this workspace, users are able to input raw sequencing data for one or more genomes, such as the standard 30× short-read whole-genome sequencing data used in many population- and clinical studies, and then the workflow will process all of the steps for alignment and variant calling in less than 1 day and for less than $5.00 worth of compute per sample. Interestingly, because of the highly scalable nature of cloud computing, processing additional samples, even hundreds or thousands of additional samples, will require approximately the same amount of wall-clock time, although costs will scale approximately linearly with the number of samples. In contrast, users performing similar analyses on their institutional clusters will be limited by the number of CPUs(central processing units) and RAM(random access memory) available, which are often limited to a few dozen or a few hundred at a time.

Note S2 displays the workspace for analyzing differential gene expression with Bioconductor's edgeR package. Using the interactive notebook environment, R/Bioconductor code and visualizations can easily be interleaved throughout the analysis, starting with quality control through the identification of statistically significant differentially expressed genes. This workspace reanalyzes a recently published gene knockdown dataset of the oncogene BACH1 to study how it promotes pancreatic cancer metastasis by repressing epithelial genes and enhancing the epithelial-mesenchymal transition.[52] Within a few minutes, any user can execute the R/Bioconductor code displayed in the workspace to identify and visualize the differentially expressed genes in the knockout cell lines.

Note S3 demonstrates how to perform *de novo* genome assembly and whole-genome alignment within Galaxy. The input data for the example are simulated short-read sequencing data in standard fastq format, which are then assembled using the SPAdes genome assembler within less than 1 min.[53] After the assembly, the assembled contigs are aligned to the reference genome using DNAdiff from the MUMmer package[54] to identify a novel insertion in the assembly. Finally, the sequence of the novel insertion is decoded into amino acids using transeq from the EMBOSS package[55] to reveal a message spelled out as English text. While these tools can be used for much larger genomes and much more sophisticated problems, we have found this to be a very effective classroom exercise because the students immediately know if they have followed the directions correctly if they see an interpretable message in English. This exercise is also appropriate for novices, as everything can be executed within the intuitive Galaxy interface without any command line or programming experience.

AnVIL portal: Entry into the AnVIL ecosystem

Already more than 15,000 users have used the AnVIL, and the number of users is rapidly growing. The initial entry point for AnVIL users is through the AnVIL portal (https://anvilproject.org). The portal provides unified entry to all of the available applications and datasets within the system as described below. In addition, the portal also contains a wide variety of training materials and announcements as well as a searchable catalog of the data that are loaded within the AnVIL. Currently, the AnVIL hosts data from >280,000 human genomes from >240 different cohorts spanning CCDG, CMG, the Electronic Medical Records and Genomics (eMERGE) Network, Genotype-Tissue Expression (GTEx),[56] and several other major NHGRI projects (Figure 2). In this view, only summary information is displayed so that any user can browse all of the datasets present even if they are not authorized to view the specific data files. This way, a user can learn what is available (e.g., all studies of a particular disease or phenotype) and, if necessary, can be directed to apply for authorization through the appropriate DAC (e.g., dbGaP or the consortium that maintains the data). The AnVIL also maintains a few critical open access datasets, most notably the widely used 1000 Genomes Project whole-genome sequencing dataset from a collection of diverse human samples,[5] including both raw data and harmonized variant calls from 3,202 samples.

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

Gen3 (https://gen3.theanvil.io) is an open-source cloud-based data platform for managing, analyzing, harmonizing, and sharing large datasets. It is based on a set of standards-based services with open APIs called “framework services” for the authentication, authorization, creation, and accession of FAIR data objects[57] and import and export of bulk clinical and phenotype data. In particular, it supports assigning persistent digital identifiers to data objects, assigning associated metadata, and accessing the data objects using the GA4GH Data Repository Service (DRS) standard, a generic interface allowing data access in a cloud-agnostic manner. Gen3 supports authentication and authorization management using OpenID tokens and interoperates with the NIH Research and Authorization Service (RAS). Framework services are also used by other large-scale genomics platforms, including NCI’s Cancer Research Data Commons, NHLBI’s BioData Catalyst, and the Kids First Data Resource. Framework services provide the basic scaffolding so that systems such as AnVIL can access data from other cloud-based platforms for genomic data and, in turn, make their data available to these platforms, assuming the appropriate policies supporting this interoperability are in place.

Gen3 also provides services for managing clinical and phenotype data and metadata using a graph database. Gen3’s Windmill service is an interactive website built over the graph database that allows users to explore, submit, and download data. The Windmill service allows for interactive data exploration, search, and cohort-building based on phenotypic variables and data types. For example, using Windmill, users can query across multiple sequencing projects (e.g., CCDG, CMG, and eMERGE) to create a synthetic cohort of patients fitting a certain set of inclusion criteria (e.g., based on gender, ethnicity, or disease status). Selected cohorts can then be exported into a Terra workspace for further processing (e.g., disease association, expression analysis, expression quantitative trait locus [eQTL] analysis, etc.). In this way, researchers can maximize the value of the data in the AnVIL by

enabling search and analysis over all relevant data to answer a particular research question, even if those data were originally generated from unrelated sequencing projects.

Terra: Access data, run analysis tools, and collaboratation in workspaces

Terra (https://anvil.terra.bio) is a cloud-native platform for biomedical researchers to access data, run analysis tools, and collaborate within the AnVIL. Workspaces are the building blocks of Terra—a dedicated space where collaborators can access and organize the same data and tools and run analyses together. Each workspace is associated with a cloud bucket where data can be stored, such as data generated by a workflow analysis[58] or notebook files for interactive computing. 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 links large-scale datasets to workspace tools. Finally, within a workspace, users can launch batch analysis jobs or one of several interactive computing environments, especially Galaxy, R/Bioconductor, and Jupyter Notebooks (as described below).

Batch analysis in Terra primarily uses the Workflow Description Language (WDL; https://openwdl.org). WDL is a specialized programming language to specify data processing workflows with a human-readable and -writable syntax. WDL makes it straightforward to define analysis tasks, chain them together in workflows, and parallelize their execution without retooling the application to run in a different computing environment. The language makes common patterns (scatter/gather, etc.) simple to express while also admitting uncommon or complicated behavior through conditionals and strives to achieve portability not only across execution platforms but also across different types of users. WDLs can be stored, shared, and described in Dockstore and executed in Terra using the Cromwell compute engine (https://cromwell.readthedocs.io), allowing for a reproducible analysis of even the largest cohorts with tens of thousands of samples.

Dockstore: Registry of tools and workflows

The Dockstore (https://dockstore.org) is another widely used platform where users can find, share, and use curated tools and workflows. Workflow content is encapsulated in Docker[59] and described using a workflow language. The use of Docker makes workflows in Dockstore reproducible by making them easy to run without user installation. Dockstore enables scientists to share analytical tools in a way that makes them machine-readable and -runnable in a variety of environments. Dockstore currently supports 4 workflow languages: the WDL, Common Workflow Language (CWL), Nextflow, and Galaxy Workflows (GWs). Dockstore currently contains 745 workflows in WDL that can be launched in Terra within a few clicks. As such, Dockstore provides one of the most straightforward entry points for users to add batch workflows to the AnVIL as it can work with any tool/workflow that can be encapsulated into a Docker container and executed on the command line.

Jupyter Notebooks: Transparent code, visualizations, and narratives

Jupyter Notebooks (https://jupyter.org) are widely used open-source web applications that allow users to create and share documents that contain live code, equations, visualizations, and narrative text. Uses include data cleaning and transformation, numerical simulation,

statistical modeling, data visualization, machine learning, and many other analyses. Jupyter supports multiple programming languages, including Python, R, 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. The familiar programming environment makes it easy for users to perform custom analysis of AnVIL data in a secure and collaborative research environment within Terra.

RStudio: Interactive machine learning, statistical computing, and visualizations

RStudio (https://rstudio.com) is an integrated development environment for R, a widely used programming language for statistical computing and visualization. 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 actively contributes many new packages. Other strengths of R include advanced static and interactive graphics and the facile creation of graphical user interfaces for easy use of highly specialized packages. R is supported in AnVIL through Jupyter Notebooks and a web version of RStudio that executes within Terra. The RStudio interface offers a complete Integrated Development Environment (IDE) for developing and executing code, supporting a windowed interface for displaying code, plots, data, and a console all at the same time.

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 with a focus on developing new computational and statistical methods to interpret biological data. Many of these methods are developed by members of the Bioconductor community,[33] 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 1,900 Bioconductor software packages, 3,200 exemplary experiments, and 50,000 model organism annotation resources have been curated for use in genomic data analyses. 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 publicly 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. The AnVIL/Bioconductor environment can be accessed within RStudio or Jupiter Notebooks and contains many important resources for the AnVIL, including a fully computable version of the online book Orchestrating Single Cell Analysis with Bioconductor.[47] A variety of methods for programmatically interacting with the AnVIL APIs are also available within the AnVIL Bioconductor package (https://bioconductor.org/packages/release/bioc/html/AnVIL.html).

Galaxy: Accessible, reproducible, and transparent genomic science

Galaxy (http://usegalaxy.org) is an open, web-based computational workbench for performing accessible, reproducible, and transparent genomic science that is used daily by

thousands of scientists across the world. There are more than 8,000 analysis tools available within Galaxy that are now accessible within the AnVIL including for variant calling and interpretation, chromatin immunoprecipitation sequencing (ChIP-seq) analysis, RNA-seq analysis, genome assembly, proteomics, epigenomics, transcriptomics, and a host of other analyses in the life sciences. To maintain data security, each AnVIL user runs within an independent Galaxy instance within Terra where they can import both unprotected data and the protected human genomics datasets they are authorized to access. This is accomplished using a newly developed capability to programmatically launch and administer Galaxy using Kubernetes and a new import tool allowing data to be added into a user's instance. An AnVIL user can thus use any available Galaxy tool to analyze or visualize data within the boundaries of a compliant, isolated, and secure environment. This marks a major advance, as AnVIL users can now leverage Galaxy for the analysis of protected human datasets, which is not possible with other public instances of Galaxy.

Extending the AnVIL capabilities

In addition to the components described above, we are considering many ways to extend the AnVIL to include new capabilities. The most straightforward approaches are to develop a new Docker-based WDL that can launch novel analysis tools and to wrap an analysis or visualization tool so that it can be executed within the Galaxy GUI. More sophisticated integrations are also possible using a variety of low-level APIs and resources. Recent efforts have focused on deploying new applications using Kubernetes (https://kubernetes.io), which can be used for managing very complicated software stacks on scalable infrastructure. Applications are deployed and managed in the Kubernetes cluster by Helm (https://helm.sh/) in the form of charts. In this design, a Helm chart translates an application's software stack into customizable Kubernetes manifests. This model, originally developed by the Galaxy Team to enable Galaxy's deployment within the AnVIL, can be replicated and extended to facilitate the integration of other platforms of varying complexity into the AnVIL. We also have several major additional components in development, including deploying seqr (https://seqr.broadinstitute.org) and the UCSC Genome Browser[37] within the AnVIL.

DATA ACCESS AND DATA USE

A key priority of the AnVIL is ensuring responsible data management, which includes secure access to the data in its cloud storage and compute environments. The AnVIL Data Access Working Group (DAWG) defines the methods used to securely control and grant access to controlled-access datasets hosted within the AnVIL and is testing improved processes for handling data access requests (DARs). The DAWG evaluates the data coming into AnVIL and considers downstream data access needs. For example, the DAWG generated the Consortium Guidelines for AnVIL Data Access (https://anvilproject.org/learn/data-submitters/resources/consortium-data-access-guidelines) to clarify expectations for the various consortia using the AnVIL to facilitate inter-consortium data sharing and access controls.

Importantly, the DAWG is leading a pilot of the Data Use Oversight System (DUOS; https://duos.broadinstitute.org/), a platform developed by the Broad Institute that aims to

expedite data access for researchers by facilitating and enhancing DAC's workflows.[60] The pilot currently includes multiple NIH DACs who are testing the system and providing feedback to further develop the DUOS software, most notably DUOS's DAR decision-support algorithm. This algorithm leverages the GA4GH Data Use Ontology (DUO; https://github.com/EBISPOT/DUO) to code both datasets' data use terms and researchers' proposed research contained within the DARs.[61] With both of these inputs in terms from the same ontology, the algorithm can assess if the proposed research is within the bounds of the data use terms and provide a recommended decision to the DAC. In the long term, the pilot will also provide powerful empirical and conceptual evidence of the feasibility of semi-automated approaches to data use oversight.

The DAWG is also refining the Library Card concept by which an institution can pre-authorize trusted researchers to make controlled DARs. This concept will leverage the GA4GH Passport Visa specification (https://github.com/ga4gh-duri/ga4gh-duri.github.io).[62] If implemented, the Library Card concept would reduce the steps required for researchers to submit a DAR while ensuring the researcher has the appropriate permissions to do so.

If successful, we believe DUOS and the Library Card concept will standardize and streamline the DAR process. As the number of requests for data increases in magnitude over the years, DUOS could ensure DAC members' time is reserved for fine-grained judgment of complex requests, and the Library Card could streamline the authorization of researchers. We hope that by pioneering implementations of the GA4GH DUO and Passports standards, the AnVIL will drive interoperable, ethical, and accelerated genomics research.

AnVIL COMMUNITY

The AnVIL is designed to support a broad range of user communities, from multi-institution consortia to individual research labs to computational tool developers and researchers at institutions without access to high-performance computing. Some needs of these communities are common—the ability to upload, manage, and share controlled-access protected data, the ability to do high-performance computation in either workflow or interactive environments, and the ability to develop training materials and share results with the broader community. However, the diversity of the AnVIL user base also requires satisfying specific needs of the constituent communities.

- Consortia and data generators: the primary needs of these groups include data ingestion, quality control, management, and sharing among consortium members and collaborators. We have developed a process for data ingestion and management on the AnVIL platform that supports consortia in sharing their data while ensuring user management and access via access groups following a consortium's data sharing and access guidelines. As of August 2021, the AnVIL contained over 200 datasets from NHGRI-sponsored projects, including the widely accessed GTEx version 8 data, which are also optionally available for direct download free of egress charges.
- Research groups and investigators: the primary needs of these groups include access to data, interactive and batch workflow computing environments, and

the ability to manage their data science projects. We have developed a user management system leveraging the Terra workflow and the workspace access management system. We have also partnered with STRIDES (https://datascience.nih.gov/strides) to support several pilot user education events with an eye toward scaling support to the broader research community. As of August 2021, the AnVIL has supported computation from more than 1,950 users running more than 775 workflows and launching more than 240 workspaces.

- Computational tool developers: tool developers need an environment where they can reproducibly test their genomic data science tools, integrate them into workflows, and share them with the broader community. The AnVIL supports several major avenues of deployment, including Docker containers to execute as WDL workflows, conda packages that can execute within Galaxy, and new Bioconductor packages. Notably, by leveraging existing data science tool developer communities, thousands of Bioconductor software packages and GWs are already integrated in the AnVIL environment.
- Under-resourced genomic data science communities: one of the biggest advantages of a fully cloud-based computational environment like the AnVIL is the ability to do high-performance computing from anywhere. Genomic data science with the AnVIL is accessible to anyone with a web browser and an internet connection, extending access to high-performance computing to communities that do not have local resources to support this kind of science. We have begun a collaboration called the Genomic Data Science Community Network (http://gdscn.org) with community colleges, historically black colleges and universities, and tribal colleges to support data-intensive genomic research and teaching using the AnVIL.

INTEROPERABILITY WITH OTHER CLOUD PLATFORMS

Freed from the constraint of needing to download data to local compute infrastructure, cloud-based research environments are becoming more widely used to streamline data access and focus on the analysis to be done. Across the AnVIL and 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 Cancer Research Data Commons (CRDC; https://datacommons.cancer.gov), for example, almost 8 PB of genomic and related data are currently accessible to researchers in cloud-based analysis platforms, and the scale of these datasets are growing quickly. Several commonalities exist across these platforms, aimed at streamlining usage by offering several widely used analysis systems such as R, RStudio, and Jupyter Notebooks. Furthermore, the Terra, Gen3, and Dockstore components of the AnVIL directly support multiple cloud platforms. One key distinguishing attribute of the AnVIL is its access to unique datasets: as the central cloud-computing platform for NHGRI, several major datasets are only available on the AnVIL. Additionally, the AnVIL platform offers a rich ecosystem of analysis tools for NHGRI-related research, including Galaxy, Bioconductor, seqr, and low-level APIs.

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 programmatic, user interface, and even policy perspectives. As a result, there is great motivation for these systems to adopt consistent conventions and standards to enable interoperability that facilitates researchers' ability to ask questions across the individual platforms.

The AnVIL project has pushed the interoperability envelope by piloting new technologies and adopting key standards and conventions from known standards bodies such as GA4GH. This was done to realize the vision of researchers using data and to compute across NIH cloud-based platforms seamlessly. The AnVIL's interoperability strategy focuses on 4 distinct areas: (1) data access, (2) portable analysis, (3) authentication and authorization, and (4) search and handoff between systems. For data access, the AnVIL has implemented the GA4GH DRS, which provides a consistent interface to data resources on cloud environments (both public and private) that enable data analysts and researchers to access data in a fashion that is agnostic to cloud service providers. This is possible because DRS Uniform Resource Identifiers (URIs), not actual data files, are passed between platforms. To enable portable analysis, the AnVIL supports both the WDLs and GWs through Terra and Galaxy, respectively. Each system allows researchers to write analysis tools and workflows that leverage Docker images, a popular containerization technology that facilitates portability. These workflows are shared through Dockstore, which itself supports the GA4GH Tool Registry Service (TRS), making it possible to share workflows among many different systems beyond the AnVIL. At the time of writing, the AnVIL and other NIH cloud platforms are working toward developing prototype implementations of additional GA4GH standards such as the recently ratified Task Execution Service (TES) and Workflow Execution Service (WES) standards. These technologies will bring exciting new capabilities where analysis jobs can be remotely launched and monitored so that users can easily distribute work across multiple cloud platforms. For authentication and authorization, the AnVIL uses the NIH Research Auth Service (RAS) from NIH's Center for Information Technology (CIT) to facilitate 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 the datasets a researcher is authorized to access. Furthermore, RASs offer increased protection via automated logging of data access.[62] The AnVIL has explored and is implementing search and data discovery through the Fast Healthcare Interoperability Resource (FHIR) standard, a data modeling language with an API specification focused on the interoperability of clinical and research data. Finally, the AnVIL has developed a search handoff mechanism between the AnVIL data discovery portal and Terra analysis environment using the Portable Format for Bioinformatics (PFB) file type.

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 and compute ecosystem spanning the tAnVIL, 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 DRS and TRS, conventions like PFB, and the auth services of RAS to address real-world scientific use cases. The real-world use cases include the joint analysis of datasets from Clinical Proteomics Tumor Analysis Consortium (CPTAC, hosted by CRDC), The Cancer Genome Atlas (TCGA, hosted by CRDC), and GTEx (hosted by the AnVIL) to address the causes and implications of expression variation in somatic tissues. Feasibility studies have also been conducted in analyzing both open and controlled access data by bringing 1000 Genomes Project data hosted by the AnVIL into the GMKF platform for co-analysis with the GMKF-hosted TARGET Neuroblastoma Project data with the aim of exploring differential expression across healthy and diseased samples. Similarly, another use case involving pooled analyses of the Pediatric Cardiac Genetics Consortium (PCGC, hosted by GMKF), PCGC's Congenital Heart Disease Biobank (hosted by BDCat), the Framingham Heart Study (hosted by BDCat), and the Jackson Heart Study (hosted by BDCat) on the AnVIL is currently underway. Finally, additional interoperability work is in development to expose the Telomere-to-Telomere CHM13 reference genome and the associated analysis workflows hosted by the AnVIL[63] to other cloud platforms to improve read mapping and variant calling.[45] The NCPI is using these demonstrations as the foundation for future work with the goal of expanding the interoperability between systems to all of the NCPI and beyond.

Beyond technical interoperability is the need for semantic interoperability. The AnVIL hosts data from a wide diversity of projects that contain differing levels of annotation, alternate ontologies, and even mismatched measurement units, sometimes even within the same project. These issues make analyzing phenotypic data in conjunction with the genomic data difficult even if the systems utilize the same APIs for data transfer. In developing a unified data transform for the AnVIL data dashboard, we encountered known problems with common metadata and phenotypic data elements, such as inconsistent or missing project disease mapping, subject disease affected status, consistent keys for specimen, specimen attributes, CRAM/BAM statistics, and other missing files.

Extending these data normalization efforts, researchers from the AnVIL project began mapping common elements to other NIH projects. One example of NIH investment in these standards is HL7 FHIR (http://hl7.org/fhir), a standard for representing, searching, and sharing clinical data. For data commons, FHIR can be seen as a target data model and staging database for data interchange efforts. With a common data model, protocol, and search mechanism for clinical attributes in place, the gap becomes a difference between metadata values and ontologies. Recent notices from NHGRI and the NIH at large have emphasized the importance of projects following standards for metadata formatting (https://grants.nih.gov/grants/guide/notice-files/NOT-HG-21-022.html), including the use of FHIR (https://grants.nih.gov/grants/guide/notice-files/NOT-OD-19-122.html). And while the original primary focus of FHIR development was on enabling the exchange of medical data, its formalization of record versioning, provenance tracking, and ontology mapping make it a useful platform for cross-project data interoperability. This technology is now being utilized

to provide more normalized data models that allow for queries across the AnVIL and other NIH projects, such as the Kid's FirstDataResource(https://kidsfirstdrc.org/).

OUTLOOK

The AnVIL launched just over 3 years ago. While there has been remarkable progress since then, there is still significant work required before the promise of this effort is fully realized. In the next stages of the AnVIL, we will focus on ensuring robustness and security for our core components while also working to develop many new capabilities in key scientific areas across human genetics and clinical genomics. For example, one major current focus is to integrate additional tools and workflows for clinical genomics, especially the calculation and utilization of polygenic risk scores[64] and pharmacogenomics analysis.[65]

More broadly, we are still in the earliest stages of the cloud transformation within the life sciences, and many institutions have already made significant investments into on-premises computing clusters and data centers that we cannot ignore. During this transition period, it is likely that cloud resources will be used for the largest analyses and collaborative research projects, but summarized data and institutionally generated private data will still be analyzed locally. As such, one of the key requirements for the AnVIL is that all of the major analysis components can be run locally: WDLs are increasingly used on institutional computing clusters, R/Bioconductor works equally well on a laptop or in the cloud, and Galaxy can be deployed on a laptop or within an institutional cluster as needed.

Another major consideration for cloud-based research is the costs involved. Even if the cost per genome or cost per sample is only a few dollars, once multiplied by tens or hundreds of thousands of genomes, the total costs for an analysis can quickly become a major expense. Moving large datasets can also be a major cost to consider; fortunately, some large datasets such as SRA are now mirrored in the cloud, reducing transfer times and associated costs. Additionally, researchers can also import their own datasets, which can scale to the petabyte level and beyond. Within the AnVIL, we also provide free egress for one of the most important datasets, the raw data for the widely studied GTEx dataset, by mirroring our cloud copy within an academic computing center so that authorized users can access it freely over Internet2 (https://anvilproject.org/news/2020/11/20/nhgri-anvil-now-supports-free-export-of-gtex-data).

While computing costs also play a major role for local computing, these costs are often amortized or supplemented through institution- or department-wide initiatives beyond individual research labs. Researchers currently have limited information for the expected costs of running analysis tools in the cloud, which challenge budgeting and prevent many researchers from adopting cloud solutions. In addition, software developers may not focus on optimizing costs for cloud environments, which increases the expense even when relatively simple optimizations are available. We and the entire genomics community must address this using all options available, including (1) educating users on the expected costs for different analyses and strategies for minimizing costs, (2) implementing additional technology safeguards (quotas or "bumpers") that will prevent users from uncontrolled spending, and (3) developing optimized tools and workflows to reduce costs. As cloud

costs are primarily a function of CPU time, RAM required, and storage space, this will include optimizations for decreasing computing time by leveraging parallel and vectorized computing instructions (e.g., AVX512 vectorization[66]) or advanced search strategies (e.g., learned index structures[67,68]), decreasing RAM requirements using more advanced data structures (e.g., Burrows-Wheeler transform,[69] Bloom filters,[70] or Sequence Bloom Trees[71]), and decreasing storage requirements by using compressed data formats (e.g., CRAM[72]), using optimized IO routines (e.g., fixed-length records instead of variable-length records[73]), and removing intermediate data. This will also include developing heuristics and approximation techniques that can often run substantially faster than more exhaustive approaches.[74,75]

With these considerations in mind, it is worth pointing out that providing access to tools, protected data, and computational resources via the AnVIL is a process of democratization. Previously, only well-funded institutions with extensive on-premises computing clusters and supporting information technology (IT) staff could afford access to high-powered computing of genomic data. While the cost of compute cannot be ignored, all researchers now have an equal opportunity to access elastic computational resources with minimal upfront costs. To ease the burden of cost projection, we are actively developing user guides and budgeting templates and making these resources available to the public. Additionally, the recently launched AnVIL Cloud Credits (AC2) program makes cloud-computing funds available to researchers and educators, including to faculty belonging to the above-mentioned GDSCN. Both AC2 and GDSCN are exciting pilot programs that we are keen to expand in the future.

Overall, the futures of the AnVIL and of cloud computing in genomics are very bright. We have several major initiatives underway to enhance our capabilities for basic science and clinical genomics, such as by integrating the tools and data from the Telomere-to-Telomere[63] and Human Reference Pan Genome projects to provide more comprehensive and diverse reference human genomes as well as major efforts with the American Heart Association (AHA), the eMERGE Network, and the Clinical Sequencing Evidence-Generating Research (CSER) programs to increase our capabilities for clinical genomics. We are particularly excited by how these datasets will allow for analysis of a broader range of genome variation, especially complex structural variants, and of additional functional genomics data types, including at the single-cell level, to develop a better understanding of the molecular basis of health and disease across diverse patient populations. Internally, we also have several major technical enhancements planned, such as offering multi-cloud support and enhanced support for deploying additional complex applications using kubernetes. In addition, NHGRI is broadening the support for the AnVIL by promoting it as a primary data sharing platform and/or a primary data analysis platform for several funding opportunities. Finally, these efforts are coupled with major training and outreach efforts to ensure everyone is aware of the platform and can use it for their research needs for many years to come.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This work is dedicated to the late James Peter Taylor, the Ralph S. O'Connor Professor of Biology and Computer Science at Johns Hopkins University, who was one of the original architects for the AnVIL and an ardent champion for open science (https://galaxyproject.org/jxtx). V.D.F., E.M.G., C.H., N.K., S.K.S., A.S., C.W., and K.L.W. provided substantial involvement and guidance for the project activities and contributed to the manuscript in their official roles as program coordinators for the NIH, NHGRI. The AnVIL is supported through cooperative agreement awards from NHGRI with co-funding from OD/ODSS to the Broad Institute (U24HG010262) and Johns Hopkins University (U24HG010263). The GDSCN is supported through a contract to Johns Hopkins University (75N92020P00235).

DECLARATIONS OF INTERESTS

A.A.P. is a venture partner at GV and has received funding from Intel, IBM, Microsoft, Alphabet, and Bayer. D.B., E.A., J.G., J.C., and A.N. are founders of and hold equity in GalaxyWorks, LLC. The results of the study discussed in this publication could affect the value of GalaxyWorks, LLC. These arrangements have been reviewed and approved by the Johns Hopkins University, Oregon Health & Science University, and The Pennsylvania State University in accordance with their respective conflict of interest policies. V.C. has financial interest in Amazon, NVIDIA, and AMD.

REFERENCES

1. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, Iyer R, Schatz MC, Sinha S, and Robinson GE (2015). Big Data: Astronomical or Genomical? PLoS Biol 13, e1002195. [PubMed: 26151137]
2. Rehm HL, Page AJH, Smith L, Adams JB, Alterovitz G, Babb LJ, Barkley MP, Baudis M, Beauvais MJS, Beck T, et al. (2021). GA4GH: International policies and standards for data sharing across genomic research and healthcare. Cell Genom. 1, 100029. [PubMed: 35072136]
3. Koboldt DC, Steinberg KM, Larson DE, Wilson RK, and Mardis ER (2013). The next-generation sequencing revolution and its impact on genomics. Cell 155, 27–38. [PubMed: 24074859]
4. Green ED, Gunter C, Biesecker LG, Di Francesco V, Easter CL, Feingold EA, Felsenfeld AL, Kaufman DJ, Ostrander EA, Pavan WJ, et al. (2020). Strategic vision for improving human health at The Forefront of Genomics. Nature 586, 683–692. [PubMed: 33116284]
5. Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, Corvelo A, Clarke WE, Musunuri R, Nagulapalli K, et al. (2021). High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. bioRxiv. 10.1101/2021.02.06/430068.
6. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, et al. ; Genome Aggregation Database Consortium (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581, 434–443. [PubMed: 32461654]
7. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, Taliun SAG, Corvelo A, Gogarten SM, Kang HM, et al. ; NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature 590, 290–299. [PubMed: 33568819]
8. Wainschtein P, Jain DP, Yengo L, Zheng Z, TOPMed Anthropometry Working Group; NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium; Cupples LA, Shadyab AH, McKnight B, Shoemaker BM, et al. (2019). Recovery of trait heritability from whole genome sequence data. bioRxiv. 10.1101/588020.
9. Tanay A, and Regev A (2017). Scaling single-cell genomics from phenomenology to mechanism. Nature 541, 331–338. [PubMed: 28102262]
10. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, and Rinn JL (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol. 32, 381–386. [PubMed: 24658644]
11. Goodwin S, McPherson JD, and McCombie WR (2016). Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. 17, 333–351. [PubMed: 27184599]
12. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, et al. (2015). UK biobank: an open access resource for identifying the causes of

a wide range of complex diseases of middle and old age. PLoS Med. 12, e1001779. [PubMed: 25826379]

13. Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, Besenbacher S, Magnusson G, Halldorsson BV, Hjartarson E, et al. (2015). Large-scale whole-genome sequencing of the Icelandic population. Nat. Genet. 47, 435–444. [PubMed: 25807286]
14. Sakaue S, Kanai M, Karjalainen J, Akiyama M, Kurki M, Matoba N, Takahashi A, Hirata M, Kubo M, Matsuda K, et al. ; FinnGen (2020). Trans-biobank analysis with 676,000 individuals elucidates the association of polygenic risk scores of complex traits with human lifespan. Nat. Med. 26, 542–548. [PubMed: 32251405]
15. CNCB-NGDC Members and Partners (2021). Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2021. Nucleic Acids Res. 49 (D1), D18–D28. [PubMed: 33175170]
16. Saudi Genome Project Team (2015). The Saudi Human Genome Program: An oasis in the desert of Arab medicine is providing clues to genetic disease. IEEE Pulse 6, 22–26.
17. Castellanos-Uribe M, May ST, and Betsou F (2020). Integrated BioBank of Luxembourg-University of Luxembourg: University Biobanking Certificate. Biopreserv. Biobank. 18, 7–9. [PubMed: 32069098]
18. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, and Hirschhorn JN (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat. Rev. Genet. 9, 356–369. [PubMed: 18398418]
19. Thorogood A, Rehm HL, Goodhand P, Page AJH, Joly Y, Baudis M, Rambla J, Navarro A, Nyronen TH, Linden M, et al. (2021). International federation of genomic medicine databases using GA4GH standards. Cell Genom. 1, 100032. [PubMed: 35128509]
20. Barranco C (2021). The Human Genome Project (Nature Research).
21. Gold ER, and Carbone J (2010). Myriad Genetics: In the eye of the policy storm. Genet. Med. 12 (4, Suppl), S39–S70. [PubMed: 20393310]
22. Toronto International Data Release Workshop Authors; Birney E, Hudson TJ, Green ED, Gunter C, Eddy S, Rogers J, Harris JR, Ehrlich SD, Apweiler R, et al. (2009). Prepublication data sharing. Nature 461, 168–170. [PubMed: 19741685]
23. National Institutes of Health (2014). Final NIH Genomic Data Sharing Policy. Fed. Regist 79, 51345–51354.
24. Powell K (2021). The broken promise that undermines human genome research. Nature 590, 198–201. [PubMed: 33568833]
25. MacArthur JAL, Buniello A, Harris LW, Hayhurst J, McMahon A, Sollis E, Cerezo M, Hall P, Lewis E, Whetzel PL, et al. (2021). Workshop proceedings: GWAS summary statistics standards and sharing. Cell Genom. 1, 100004.
26. Bahcall OG (2021). Genomics for all: Open, collaborative, pioneering. Cell Genom 1, 100008.
27. Leinonen R, Sugawara H, and Shumway M; International Nucleotide Sequence Database Collaboration (2011). The sequence read archive. Nucleic Acids Res. 39, D19–D21. [PubMed: 21062823]
28. Kodama Y, Mashima J, Kosuge T, Kaminuma E, Ogasawara O, Okubo K, Nakamura Y, and Takagi T (2018). DNA Data Bank of Japan: 30th anniversary. Nucleic Acids Res. 46 (D1), D30–D35. [PubMed: 29040613]
29. Harrison PW, Ahamed A, Aslam R, Alako BTF, Burgin J, Buso N, Courtot M, Fan J, Gupta D, Haseeb M, et al. (2021). The European Nucleotide Archive in 2020. Nucleic Acids Res. 49 (D1), D82–D85. [PubMed: 33175160]
30. Arita M, Karsch-Mizrachi I, and Cochrane G (2021). The international nucleotide sequence database collaboration. Nucleic Acids Res. 49 (D1), D121–D124. [PubMed: 33166387]
31. Tryka KA, Hao L, Sturcke A, Jin Y, Wang ZY, Ziyabari L, Lee M, Popova N, Sharopova N, Kimura M, and Feolo M (2014). NCBI's Database of Genotypes and Phenotypes: dbGaP. Nucleic Acids Res. 42, D975–D979. [PubMed: 24297256]
32. Gruning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, Valieris R, and Köster J; Bioconda Team (2018). Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat. Methods 15, 475–476. [PubMed: 29967506]

33. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5, R80. [PubMed: 15461798]
34. Schatz MC, Langmead B, and Salzberg SL (2010). Cloud computing and the DNA data race. Nat. Biotechnol. 28, 691–693. [PubMed: 20622843]
35. Langmead B, and Nellore A (2018). Cloud computing for genomic data analysis and collaboration. Nat. Rev. Genet. 19, 208–219. [PubMed: 29379135]
36. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, and Madden TL (2008). NCBI BLAST: a better web interface. Nucleic Acids Res. 36, W5–W9. [PubMed: 18440982]
37. Navarro Gonzalez J, Zweig AS, Speir ML, Schmelter D, Rosenbloom KR, Raney BJ, Powell CC, Nassar LR, Maulding ND, Lee CM, et al. (2021). The UCSC Genome Browser database: 2021 update. Nucleic Acids Res. 49 (D1), D1046–D1057. [PubMed: 33221922]
38. Goecks J, Nekrutenko A, and Taylor J; Galaxy Team (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 11, R86. [PubMed: 20738864]
39. Jalili V, Afgan E, Gu Q, Clements D, Blankenberg D, Goecks J, Taylor J, and Nekrutenko A (2020). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. Nucleic Acids Res. 48 (W1), W395–W402. [PubMed: 32479607]
40. Lau JW, Lehnert E, Sethi A, Malhotra R, Kaushik G, Onder Z, Groves-Kirkby N, Mihajlovic A, DiGiovanna J, Srdic M, et al. ; Seven Bridges CGC Team (2017). The Cancer Genomics Cloud: Collaborative, Reproducible, and Democratized-A New Paradigm in Large-Scale Computational Research. Cancer Res. 77, e3–e6. [PubMed: 29092927]
41. Taylor L (2014). FedRAMP: History and Future Direction. IEEE Cloud Computing 1, 10–14.
42. Yuen D, Cabansay L, Duncan A, Luu G, Hogue G, Overbeck C, Perez N, Shands W, Steinberg D, Reid C, et al. (2021). The Dockstore: enhancing a community platform for sharing reproducible and accessible computational protocols. Nucleic Acids Res. 49 (W1), W624–W632. [PubMed: 33978761]
43. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinformatics 43, 11.10.1–33. [PubMed: 25431634]
44. Garrison E, and Marth G (2012). Haplotype-based variant detection from short-read sequencing. arXiv, 1207.3907. https://arxiv.org/abs/1207.3907.
45. Aganezov S, Yan SM, Soto DC, Kirsche M, Zarate S, Avdeyev P, Taylor DJ, Shafin K, Shumate A, Xiao C, et al. (2021). A complete reference genome improves analysis of human genetic variation. bioRxiv, 2021.07.12.452063.
46. Li B, Gould J, Yang Y, Sarkizova S, Tabaka M, Ashenberg O, Rosen Y, Slyper M, Kowalczyk MS, Villani A-C, et al. (2020). Cumulus provides cloud-based data analysis for large-scale single-cell and single-nucleus RNA-seq. Nat. Methods 17, 793–798. [PubMed: 32719530]
47. Amezquita RA, Lun ATL, Becht E, Carey VJ, Carpp LN, Geistlin-ger L, Marini F, Rue-Albrecht K, Risso D, Soneson C, et al. (2020). Orchestrating single-cell analysis with Bioconductor. Nat. Methods 17, 137–145. [PubMed: 31792435]
48. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29, 644–652. [PubMed: 21572440]
49. Krueger F, and Andrews SR (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 27, 1571–1572. [PubMed: 21493656]
50. Lemieux JE, Siddle KJ, Shaw BM, Loreth C, Schaffner SF, Gladden-Young A, Adams G, Fink T, Tomkins-Tinch CH, Krasilnikova LA, et al. (2021). Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. Science 371, eabe3261. [PubMed: 33303686]
51. Baker D, van den Beek M, Blankenberg D, Bouvier D, Chilton J, Coraor N, Coppens F, Eguinoa I, Gladman S, Gruning B, et al. (2020). No more business as usual: Agile and effective responses

to emerging pathogen threats require open data and open analytics. PLoS Pathog. 16, e1008643. [PubMed: 32790776]

52. Sato M, Matsumoto M, Saiki Y, Alam M, Nishizawa H, Rokugo M, Brydun A, Yamada S, Kaneko MK, Funayama R, et al. (2020). BACH1 Promotes Pancreatic Cancer Metastasis by Repressing Epithelial Genes and Enhancing Epithelial-Mesenchymal Transition. Cancer Res. 80, 1279–1292. [PubMed: 31919242]
53. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–477. [PubMed: 22506599]
54. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, and Salzberg SL (2004). Versatile and open software for comparing large genomes. Genome Biol. 5, R12. [PubMed: 14759262]
55. Rice P, Longden I, and Bleasby A (2000). EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 16, 276–277. [PubMed: 10827456]
56. GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science 369, 1318–1330. [PubMed: 32913098]
57. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3, 160018. [PubMed: 26978244]
58. Reiter T, Brooks PT, Irber L, Joslin SEK, Reid CM, Scott C, Titus Brown C, and Tessa Pierce N (2020). Streamlining Data-Intensive Biology With Workflow Systems. GigaScience 10, giaa140.
59. Boettiger C (2015). An introduction to Docker for reproducible research. Oper. Syst. Rev. 49, 71–79.
60. Cabili MN, Lawson J, Saltzman A, Rushton G, O'Rourke P, Wilbanks J, Rodriguez LL, Nyronen T, Courtot M, Donnelly S, et al. (2021). Empirical validation of an automated approach to data use oversight. Cell Genom. 1, 100031.
61. Lawson J, Cabili MN, Kerry G, Boughtwood T, Thorogood A, Alper P, Bowers SR, Boyles RR, Brookes AJ, Brush M, et al. (2021). The Data Use Ontology to streamline responsible access to human biomedical datasets. Cell Genom. 1, 100028.
62. Voisin C, Linden M, Dyke SOM, Bowers SR, Alper P, Barkley MP, Bernick D, Chao J, Courtot M, Jeanson F, et al. (2021). GA4GH Passport standard for digital identity and access permissions. Cell Genom. 1, 100030.
63. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. (2021). The complete sequence of a human genome. bioRxiv, 2021.05.26.445798.
64. Torkamani A, Wineinger NE, and Topol EJ (2018). The personal and clinical utility of polygenic risk scores. Nat. Rev. Genet. 19, 581–590. [PubMed: 29789686]
65. Lauschke VM, and Ingelman-Sundberg M (2020). Emerging strategies to bridge the gap between pharmacogenomic research and its clinical implementation. NPJ Genom. Med 5,9. [PubMed: 32194983]
66. Darby CA, Gaddipati R, Schatz MC, and Langmead B (2020). Vargas: heuristic-free alignment for assessing linear and graph read aligners. Bioinformatics 36, 3712–3718. [PubMed: 32321164]
67. Kirsche M, Das A, and Schatz MC (2021). Sapling: Accelerating Suffix Array Queries with Learned Data Models. Bioinformatics 37, 744–749. [PubMed: 33107913]
68. Kraska T, Beutel A, Chi EH, Dean J, and Polyzotis N (2017). The Case for Learned Index Structures. arXiv, 1712.01208. https://arxiv.org/abs/1712.01208.
69. Langmead B, Trapnell C, Pop M, and Salzberg SL (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25. [PubMed: 19261174]
70. Chikhi R, and Rizk G (2013). Space-efficient and exact de Bruijn graph representation based on a Bloom filter. Algorithms Mol. Biol. 8,22. [PubMed: 24040893]
71. Solomon B, and Kingsford C (2016). Fast search of thousands of short-read sequencing experiments. Nat. Biotechnol. 34, 300–302. [PubMed: 26854477]

72. Hsi-Yang Fritz M, Leinonen R, Cochrane G, and Birney E (2011). Efficient storage of high throughput DNA sequencing data using reference-based compression. Genome Res. 21, 734–740. [PubMed: 21245279]

73. Langmead B, Wilks C, Antonescu V, and Charles R (2019). Scaling read aligners to hundreds of threads on general-purpose processors. Bioinformatics 35, 421–432. [PubMed: 30020410]

74. Rhyker Ranallo-Benavidez T, Lemmon Z, Soyk S, Aganezov S, Salerno WJ, McCoy RC, Lippman ZB, Schatz MC, and Sedlazeck FJ (2020). SVCollector: Optimized sample selection for cost-efficient long-read population sequencing. bioRxiv. 10.1101/2020.08.06.240390.

75. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, and Phillippy AM (2016). Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 17, 132. [PubMed: 27323842]

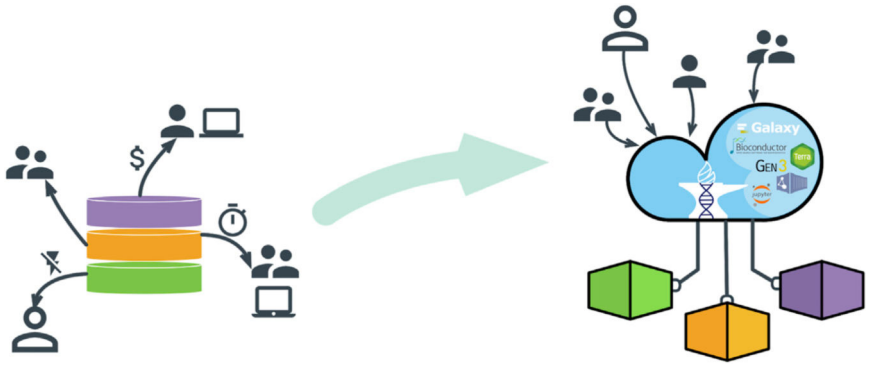
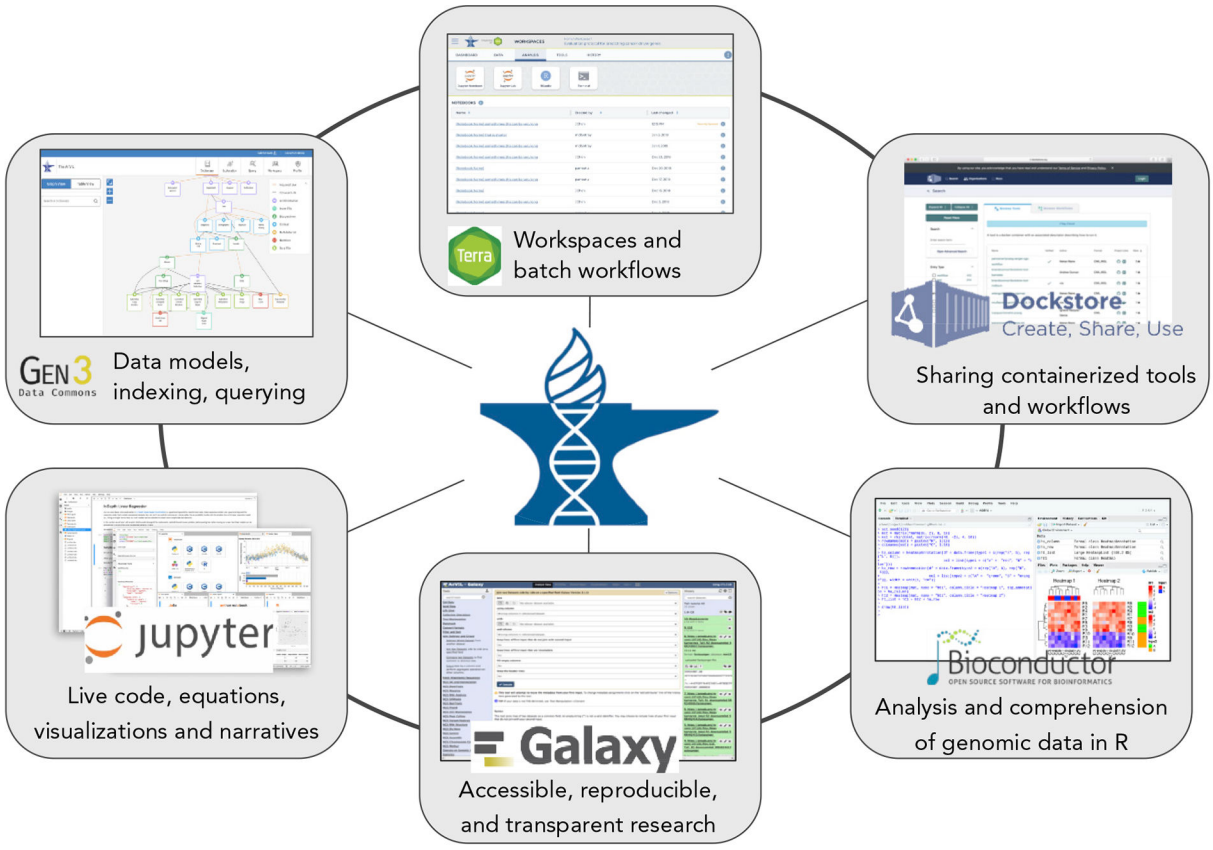


Figure 1. Inverting the model for data sharing

(Left) In the traditional model, project data (shown in purple, orange, and green) are copied to multiple sites where they are accessed by users on institutional computing clusters. Under this model, each institution must establish its own data center, and collaboration is achieved primarily through copying files between data centers. (Right) In the inverted model, users connect to a cloud-enabled resource such as the AnVIL to remotely access and analyze the data without copying. In this model, users virtually access a unified data center, allowing for deeper collaboration and sharing of the results.



Consortium	Datatypes	Cohorts	Samples	Participants	Size (TB)
1000 Genomes Project (1KGP)	WGS	1	3,202	3,202	72.98
Centers for Common Disease Genomics (CCDG)	WGS, WXS, Clinical Phenotypes	198	272,306	256,318	2,624.12
Centers for Mendelian Genomics (CMG)	WGS, Clinical Phenotypes	41	20,706	16,599	97.89
Convergent Neuroscience	WGS	2	304	300	5.32
Genotype-Tissue Expression (GTEx v8)	WGS, RNAseq	1	17,382	979	182.14
Human Pangenome Reference Consortium (HPRC)	Short & long-read WGS	1	57	47	223.47
Population Architecture using Genomics and Epidemiology (PAGE)	WGS	4	690	690	16.98
Telomere-to-Telomere (T2T)	WGS	1	3,202	3,202	571.64
Whole Genome Sequencing for Schizophrenia and Bipolar Disorder (WGSPD1)	WGS	5	9,588	9,575	177.36
Total		**254**	**327,437**	**290,912**	**3,971.91**

Figure 2. Overview of the AnVIL ecosystem

(Top) The AnVIL is a federated cloud environment for the analysis of large genomic and related datasets. The AnVIL is built on a set of established components that bring together widely used platforms. The Terra platform provides a compute environment with secure data and analysis sharing capabilities. Dockstore provides standards-based sharing of containerized tools and workflows. R/Bioconductor, Jupyter, and Galaxy provide environments for users at different skill levels to construct and execute analyses. The Gen3 data commons framework provides data and metadata ingest, querying, and organization.

(Bottom) The AnVIL has been used in a number of flagship NHGRI and other genomics projects. Summary of the genomics datasets available within the AnVIL as of December 2021, as shown at https://anvilproject.org/data. WGS, whole-genome sequencing; WXS, whole-exome sequencing.