RESEARCH STRATEGY

A. SIGNIFICANCE AND OVERVIEW

Since the completion of the Human Genome Project, the National Human Genome Research Institute (NHGRI) launched several highly successful research projects that pursued comprehensive large-scale genotyping, genome sequencing, and omics profiling. The knowledge gained from these projects provides the foundation for our current understanding of human genome biology and has ushered in the era of genomic medicine. As sequencing is becoming more accessible and less costly, many other projects of similar scale are in progress that will provide further insights into human biology and clinically relevant knowledge. Some of the ongoing projects such as ENCODE, Centers for Mendelian Genetics, Centers for Common Disease Genetics, and eMERGE are on track to generate petabytes of data. However, the infrastructure required to compute on datasets of this size and the specialized expertise required for its effective use are beyond reach of most researchers.

To address this need, we will build the NHGRI Genomic Data Science Analysis, Visualization, Informatics Lab-space (AnVIL) resource. The AnVIL's foundation will be built on the Seven Bridges Core Infrastructure (SB-CI), an interoperable data and compute cloud platform (**Fig. 1**). The AnVIL will host both unrestricted and controlled-access data, tools for their analysis and visualization. Outreach and training targeted at genomic researchers and clinicians with specific levels of expertise will empower them to make discoveries and derive clinically relevant knowledge from these data.

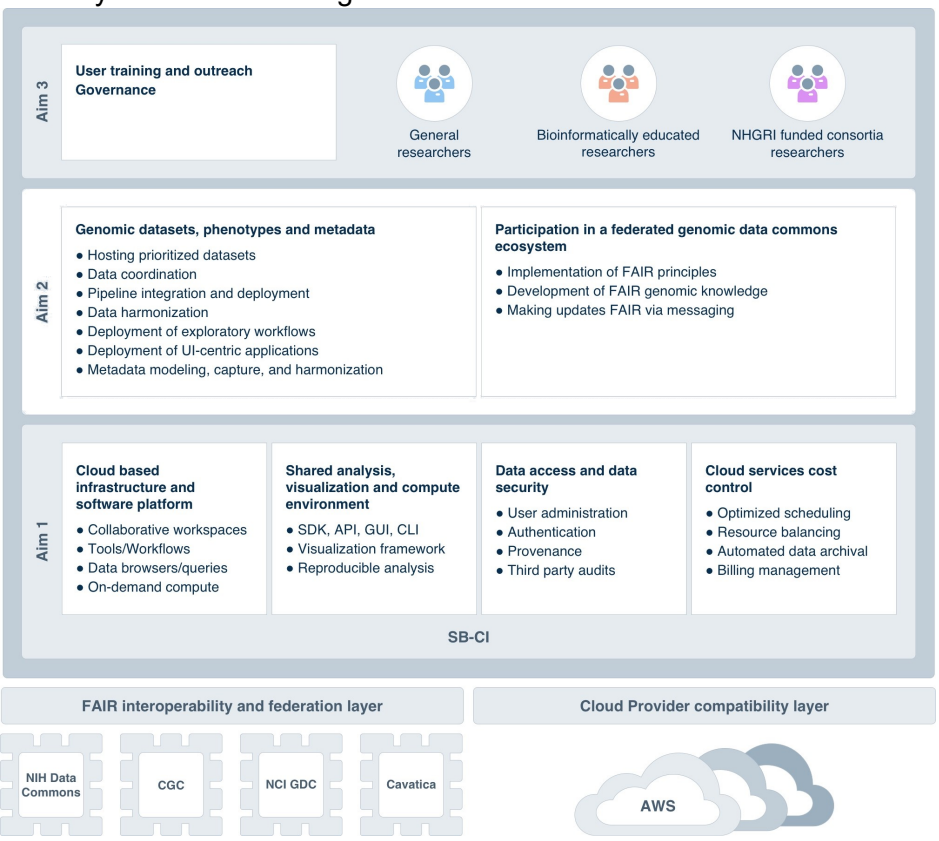


Figure 1. The AnVIL will be built on top of the Seven Bridges Core Infrastructure (SB-CI). Users will have secure access to data and compute workspaces with the capability to perform exploratory analyses (**Aim 1**). The main components of AnVIL include a robust data management system, capability for users to develop their tools (including visualization tools) via an easy to use software development kit, interoperability with other data commons, and strong mechanisms for cloud services cost control. NHGRI designated datasets will be carefully onboarded to the platform after ensuring data harmonization and metadata modelling/mapping to an interoperable ontology as described in **Aim 2**. Users will be at the core of AnVIL, as it will support a wide range of researchers with varied levels of computational expertise as described in **Aim 3**.

To accomplish the AnVIL project, we have assembled the AnVIL Resource Consortium (ARC), an academic/private industry consortium that includes accomplished senior and junior investigators from both industry and academia, including Dr. Aleksandar Milosavljevic from Baylor College of Medicine, Dr. Isheeta Seth from Seven Bridges Genomics, Dr. Joel Rozowsky from Yale University, and Dr. Alex Pico from Gladstone Institutes. Their extensive experience and record of accomplishments in data analysis and coordination for major consortia, in developing community data standards, and building open-source software and software infrastructure for genomic research will ensure AnVIL's success. To meet the current and future needs of the whole research community, ARC will implement an agile governance structure (**Fig. 2)** that includes funders and external stakeholders.

Recognizing that an optimal academic/private industry partnership formula lies at the core of the successful AnVIL, our proposal combines complementary strengths of academic and private

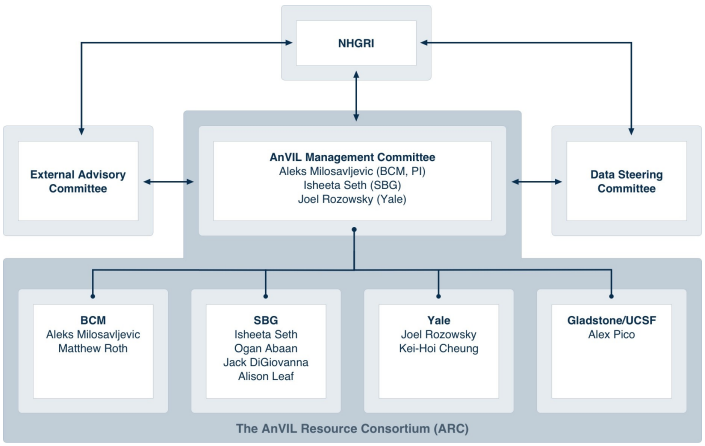


Figure 2. Proposed governance structure of the AnVIL.

industry participants in a uniquely synergistic way on the basis of data FAIRness and tool interoperability and portability.

Building on the SB-CI will provide the AnVIL with a suite of frameworks and tools for scalable and reproducible genomic data analysis and visualization on the cloud and for collaboration via secure user-controlled workspaces. To ensure robust user adoption of the AnVIL resources, we will engage the users during development, collect feedback for continuous improvements to address evolving needs, and provide training tailored to specific types of users.

Along with the technical difficulties in data management, researchers face the challenge of managing the high costs associated with access, storage, and compute for these data. SB-CI provides extensive infrastructure to address this need, including the SB-CI admin panel, a single point of control for managing resources across multiple cloud providers, and the SB-CI data management system that automatically moves dataset files between “hot”, “warm”, and “cold” storage classes to minimize cost.

In consultation with stakeholders and the research community, the AnVIL will prioritize datasets for hosting and ensure their utility by providing tools for their analysis and visualization. The AnVIL platform is well suited for community-based development of portable analysis and visualization tools. Seven Bridges pioneered and co-founded the Common Workflow Language (CWL) specification in 2014 that is now an emerging GA4GH standard for describing tools and workflows to ensure reproducibility. Seven Bridges’ Rabix (Reproducible Analysis for Bioinformatics) software development kit for CWL is an open-source development and execution toolkit that wraps tools in the CWL for execution either locally or on the cloud. Seven Bridges’ Data Cruncher manages cloud deployment of open-source Jupyter Notebooks, a multi-language platform for reproducible exploratory data analysis and visualization that supports both R and Python, the two most commonly used languages in bioinformatics.

To ensure that the hosted datasets are accessible to the entire research community and are not siloed, the AnVIL will be built on the principles of Findable, Accessible, Interoperable and Reusable (FAIR) data as they emerge from the community. By implementing FAIR principles, the AnVIL will immediately become an integral part of the emerging federated Data Commons.

In summary, by effectively addressing major obstacles between researchers and the wealth of data waiting to be utilized to its full potential, the AnVIL will empower researchers to make discoveries from large genomic datasets, extract clinically actionable knowledge from these datasets, and apply the knowledge to improve human health.

B. INNOVATION

Our proposal addresses a critical community need for a cloud-based infrastructure and software platform for collaborative analysis of genomic datasets. In this proposal we break new ground in several ways, empowering the community while contributing significantly to the success of NHGRI’s mission to advance genomic research and medicine. Some innovative highlights include the following:

Seven Bridges Core Infrastructure (SB-CI). SB-CI is a flexible framework of modular software services that support efficient data transfer and comprehensive data security. SB-CI provides the means for scalable, reproducible, and cost-effective bioinformatics analyses on the cloud, and supports collaboration via secure user-controlled workspaces. The admin panel provides a single point of control for managing resources and accounts, monitoring usage statistics, and generating reports about usage.

Data Browser. A visual query builder to quickly find and access genomics datasets stored on the platform.

The Data Browser provides seamless data navigation across different storage classes using metadata uploaded in RDF format from GenboreeKB or other sources.

Reproducible and portable analyses. Common Workflow Language (CWL) specification for ensuring reproducibility across analyses is now an emerging GA4GH standard for describing tools and workflows. A CWL description of a tool captures the command used to invoke the tool on the Unix terminal, required file types, parameter settings, resources, and the Docker image location for portability. Rabix (Reproducible Analysis for Bioinformatics) is an open-source development and execution toolkit for CWL. Rabix helps researchers develop new tools/workflows locally and troubleshoot rapidly before paying for analysis on the cloud.

Data Cruncher for managing portable Jupyter Notebooks. Data Cruncher manages Jupyter Notebooks within workspaces, allowing a user to share a Notebook by inviting collaborators into their workspace. The Notebooks are executed inside a container-based environment, ensuring reproducibility. The Notebooks are portable to other local or cloud environments with minimal modifications assuming data and library dependencies are met.

Visualization Software Development Kit (SDK). The kit will allow researchers to create, modify, publish, and maintain their custom-built visualization Apps. The Apps can then be published to a workspace and used to visualize data from files and other sources available in that workspace.

GenboreeKB. The GenboreeKB is a free open source metadata and knowledge modeling, processing, validation, and hosting platform built on top of the document-oriented MongoDB database. GenboreeKB is built in API-centric fashion and on Linked Data principles for distributed deployment and cross-cloud interoperability.

Virtual Biorepository. To capture data from biosample repositories and any associated clinical information from Electronic Health Records (EHR) systems, we will deploy a distributed Virtual Biorepository built on GenboreeKB. This will empower projects such as eMERGE.

Nucleating a FAIR Commons of computable genomic knowledge. The ClinGen Allele Registry (CAR) serves as a critical “network-adapter” linking each variant to information within and beyond ClinGen about the variant. We will extend this model to collate information and computable knowledge about genetic variants, genomic elements, and relevant networks and pathways using the FAIR and innovative WikiPathways, thus creating a FAIR Commons of computable genomic knowledge.

Making updates FAIR via messaging. To ensure timely automated reruns upon data updates and propagation of information about genetic variants and genomic elements within the AnVIL and across the federated Data Commons, we will deploy the Apache Kafka messaging system that emerged as the backbone for update propagation within the LinkedIn social network.

In combination, these innovations will open a new way toward an open and inclusive ecosystem of data and computing resources, and will alleviate bottlenecks on the road to discovery while democratizing genomic research.

C. APPROACH

C.1. SPECIFIC AIM 1: To build the cloud-agnostic AnVIL platform for analysis and visualization of high-volume genomic data, secure collaboration, and economical scalable computing.

C.1.a. Scientific premise of Aim 1

With the increasing need for data democratization within the research community, researchers need to have access to a secure and robust infrastructure where data is collocated with scalable storage and compute systems. The overarching goal of the AnVIL is to provide researchers access to valuable NHGRI datasets within a secure collaborative environment where they can upload and analyze their data alone or in conjunction with the hosted datasets.

To utilize these datasets to their full potential, the AnVIL will need to interoperate with other data commons to ensure that further knowledge can be gained by combining these datasets. Ultimately, the success of the AnVIL will depend on the ease by which researchers can access and analyze controlled or unrestricted data

from multiple datasets, regardless of their supporting infrastructure. To accomplish this goal, the AnVIL will be built on a cloud-agnostic foundation made interoperable with other data commons. Interoperability will be based on compliance with community-developed FAIR standards for data and analytical tools/workflows. The AnVIL will thus enable researchers to analyze their data in combination with multiple datasets regardless of where they are hosted. By doing so, the AnVIL will empower researchers to make discoveries, answer complex scientific questions, and gain insights that cannot be answered by analyzing datasets in isolation.

We propose to use the existing Seven Bridges Core Infrastructure (SB-CI), a flexible framework of modular software services, as the foundation for the AnVIL. SB-CI has been used to build similar platforms such as the National Cancer Institute (NCI) Cancer Genomics Cloud (CGC) and the Children's Hospital of Philadelphia's Cavatica platform. Importantly, SB-CI will also be used to build Seven Bridges' full-stack solution for the NIH Data Commons Pilot Phase. Reuse of existing infrastructure provides researchers immediate access to secure workspaces for collaboration and data transfers, as well as a suite of frameworks and tools for scalable, reproducible, and cost-effective bioinformatics analyses on the cloud. The SB-CI (**Fig. 3**) will provide several core capabilities for the AnVIL's foundation out of the box, including: a graphical user interface (GUI) and an application programming interface (API); scalable, reproducible, and portable analyses using CWL, Rabix, and Docker; sophisticated administrative, monitoring, access and usage controls; and shared, secure workspaces. Using SB-CI ensures that the AnVIL can be developed in a time and cost-effective manner, and will provide a shared interoperability layer with other NIH data resources built on SB-CI, such as the CGC, Cavatica, and the emerging NIH Data Commons.

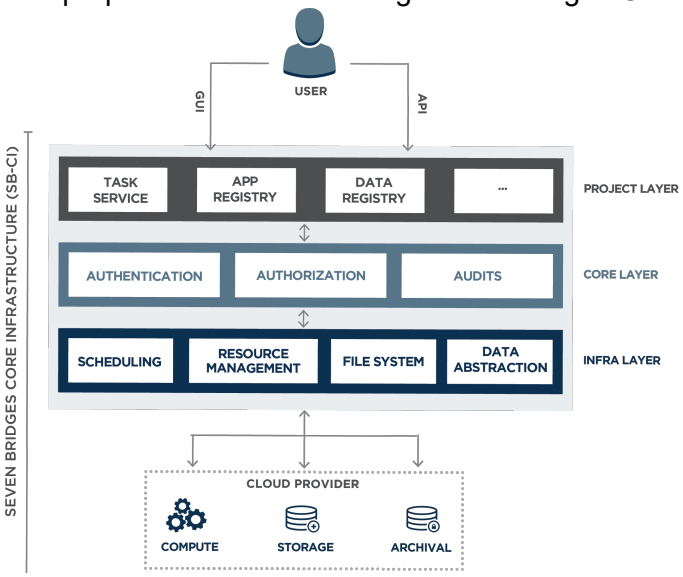


Figure 3. Cloud agnostic compute infrastructure of SB-CI. Conceptual overview of the SB-CI. Users can interact via a web-interface i.e. GUI or a programmatic interface i.e. API. The project layer refers to the workspaces. Services are functionally arranged into Interaction, Security, and Management layers. Compute instances, file storage, and archiving services are provided by the cloud service and managed and organized by the SB-CI.

The AnVIL will offer the following capabilities which will be described in **Aim 1**: 1) an easy-to-use platform for researchers with diverse levels of expertise, 2) access to highly relevant community-generated genomic data (unrestricted or controlled access), 3) optimized tools for analyses, 4) sophisticated visualization tools, and 5) a cost control mechanism for compute and storage costs on the cloud. In addition, we will incorporate user feedback and support continual improvements through our established agile software development best-practices.

C.1.b. Preliminary Studies for Aim 1

The underlying cloud infrastructure and shared analysis and computing environment for the AnVIL will benefit from Seven Bridges' engineering expertise gained over the last 8 years building similar cloud-agnostic platforms for public projects as well as commercial use. Below is a brief description of the platforms that have been built using SB-CI, as well as other relevant experiences related to setting community standards and other major programs.

- **NCI Cancer Genomics Cloud (CGC)** – Launched in February 2016, the CGC is a Seven Bridges project built with SB-CI core technologies that provides user authentication and authorization via dbGaP for accessing controlled datasets, including The Cancer Genome Atlas (TCGA), Clinical Proteomic Tumor Analysis Consortium (CPTAC), Therapeutically Applicable Research To Generate Effective Treatments (TARGET), and The Cancer Imaging Archive (TCIA), totaling nearly 2 petabytes. CGC was initially one of the three pilot projects awarded in 2014 by the US National Cancer Institute (NCI) for developing cloud-based solutions to improve data accessibility and usability. To date, the CGC has been adopted by more than 2,500 researchers from academic, nonprofit, and commercial sectors. In September 2017, the NCI awarded an extension for all three cloud pilots to move to NCI Cloud Resources. Seven Bridges, as a

subcontractor to Leidos, will provide enhancements to CGC in support of the wider NCI Data Commons framework compliance. The CGC is being developed continuously and can be extended to perform most types of bioinformatics analyses on the cloud using public as well as private datasets. Extensive documentation is available through our CGC Knowledge Center, and training materials are regularly released to the public[2].

- **Cavatica** – Launched in 2017, Cavatica is a pediatric diseases platform built by Seven Bridges in collaboration with the Children's Hospital of Philadelphia (CHOP), Childhood Brain Tumor Tissue Consortium, and the Pacific Pediatric Neuro-Oncology Consortium. Cavatica empowers collaborative dataset curation and analysis, ensuring data harmonization is maintained at ingress, during subsequent queries, and upon output. Seven Bridges has been responsible for integrations, cloud IT, and leading the development and collaborative activities to best apply the collaborators' specialized biomedical expertise. In addition, Seven Bridges has spearheaded efforts to perform multi-center harmonization of data to enable powerful cross-dataset queries within Cavatica using its technologies like Data Browser and Sonar. In August 2017, CHOP received an NIH Common Fund grant to develop the Kids First Pediatric Data Resource Center, where Cavatica and Seven Bridges will play a central role.
- **NIH Data Commons Pilot Phase** – In November 2017, Seven Bridges and its partners Repositive, Elsevier, and the U.S. Department of Veterans Affairs (Team FAIR4CURES) were awarded a contract to build one of the NIH Data Commons full stack solutions during the initial pilot phase. The FAIR4CURES team will collaborate with other awardees to establish community-supported FAIR guidelines and metrics, open standard APIs, cloud-agnostic computation architecture, workspaces, data indexing, search, regulatory workflows for data access, and regulatory compliance. These functional elements will be supported by the community expertise of all FAIR4CURE members; the infrastructure and engineering expertise provided by Seven Bridges; and a shared deep core tenet of community and user engagement.
- **NCI PDX Data Commons and Coordination Center** – The NCI awarded Seven Bridges and The Jackson Laboratory a U24 grant to build the PDX Data Commons to accelerate translational research using patient-derived tumor xenograft (PDX) datasets. This joint initiative will establish a PDX Data Commons and Coordinating Center to support PDXNet, a collaborative network that coordinates large-scale testing for preclinical therapeutic cancer drugs in PDX trials. Seven Bridges will leverage its CGC model to host a PDXNet portal for storing, harmonizing, and disseminating PDX data; searching for PDXs models based on data and metadata attributes; and co-analysis of PDX cohorts together with other large-scale NCI datasets, such as TCGA and TARGET. This showcases the flexibility and suitability of the CGC codebase to different consortia where synergy between disparate data and projects is needed in order to generate novel insights.
- **Participation in genomics standards working groups** – Seven Bridges is actively involved in efforts to define and standardize genomics methods to support federation and interoperability. In 2014, Seven Bridges co-founded the CWL - a community-developed specification to describe reproducible computational workflows. Seven Bridges has been working to develop this standard and foster the adoption of CWL by developing easy to use open source CWL development tools for production environments. In collaboration with the Global Alliance for Genomics & Health (GA4GH), Seven Bridges has worked to harmonize cloud infrastructures by developing open standards APIs. Moreover, Seven Bridges collaborated with the FHIR leadership to develop a FHIR enabled point of care prototype for bringing genomic information to physicians and patients to support decision making. Seven Bridges is a major contributor to the development of the BioCompute Object[3], a standard design to facilitate FDA regulatory review of NGS-related submissions. The FDA, in collaboration with two CGC users, converted our workflows into BioCompute objects, demonstrating our ability to quickly support new standards as they arise.
- **Department of Veteran's Affairs Million Veteran Program (MVP)**. Seven Bridges has a Collaborative Research And Development Agreement (CRADA) with the VA MVP. The overarching aim of the project is to establish an automated execution framework that has the ability to distribute data, metadata, and compute jobs across on-premises and cloud resources on the basis of access permissions of the datasets involved. The overarching deliverable is a pilot hybrid cloud system that demonstrates the above, showcasing distribution of tasks in a permissions-aware way. We are extending our current reproducible and portable workflow technologies to support seamless integration with existing IT resources, such as GenISIS, a FISMA-appropriate hybrid cloud solution requires. Our teams successfully collaborated with the VA MVP team to successfully implement the MVP alignment and variant calling workflow using Rabix and CWL.

C.1.c. Approach to Aim 1.

C.1.c.i. Development of a cloud-based infrastructure and software platform

With the ever-increasing data storage and computation needs of the community, researchers need a robust cloud-based infrastructure that will be scalable with respect to data compute and storage. We also understand that due to technical or economical convenience, there may be a need in future to redeploy the AnVIL on other cloud providers than the one the AnVIL will be initially built on. ARC's choice of SB-CI as the foundation for the AnVIL is a cloud-agnostic solution, currently powering the deployments of Seven Bridges platforms on Amazon Web Services (AWS) and Google Cloud Platform (GCP). There are ongoing engineering efforts bringing SB-CI to Alibaba Cloud, and analyzing Microsoft Azure and several other cloud providers for compatibility requirements. Furthermore, SB-CI provides the capability (Volumes API) that enables easy data exchange between disparate cloud providers without the need for costly platform deployments. Thus, ARC is confident that AnVIL can rapidly expand to other cloud service providers. While the final decision about the cloud provider for the AnVIL will be made in consultation with the NHGRI program officials, ARC proposes AWS as the initial cloud provider because it has been successfully used by Seven Bridges to build other data commons, such as the NCI Cancer Genomics Cloud, and would incur no data transfer costs when cross-accessing data between CGC and AnVIL. It is worth noting that the FAIR4CURES NIH Data Commons will be built on AWS too, as well as GCP, thus providing AnVIL with additional datasets through interoperability.

The existing SB-CI will also provide the AnVIL with scalable computing and storage systems. The SB-CI utilizes dynamic allocation of computational instances in order to scale the execution of thousands of bioinformatics pipelines at the same time. The provisioning of computation is completely transparent to the user and depends on the capacity of the cloud provider. When the system reaches capacity or a user reaches their limit, the computation jobs are queued until capacity returns. The SB-CI uses a scheduling algorithm based on the tool that needs to be executed within a task, as well as the currently available resources. The algorithm also assigns a compatible instance to each tool such that each tool within a workflow is allocated sufficient CPU processing power and memory. In addition, the scheduling algorithm automatically optimizes instance usage by executing multiple tools within a workflow in parallel even when a single instance is used for the task. The SB-CI allows researchers to select a specific instance type for their tools and workflows, providing technically capable researchers the flexibility to further optimize their workflows for time or cost. Researchers can also choose the amount of storage associated with the instance through Amazon Elastic Block Storage (EBS). Amazon EBS is well-suited for storage of data that must remain quickly accessible and will likely undergo granular updates over time.

The AnVIL will provide multiple data storage systems for diverse data access types, including data that is used frequently for immediate computing, as well as for data used less frequently. Long term archiving will also be available on the AnVIL to reduce unnecessary storage cost (**Fig. 4**). The decision to move data between the three storage tiers will be made based on data usage. For long-term storage, the SB-CI utilizes the storage and scaling capabilities of AWS S3 and Google Cloud Storage coupled with an internal abstraction layer that caches common operations on files to improve portability, responsivity, and eliminate access costs. The SB-CI offers users the ability to archive data through Amazon Glacier. Storage on Amazon Glacier is significantly cheaper than storage on Amazon S3, resulting in a cost-conscious solution for infrequently accessed files. As an additional cost control mechanism for the AnVIL, we

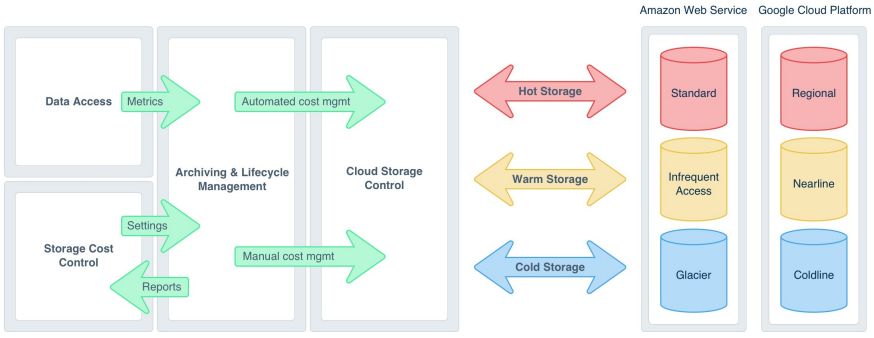


Figure 4. Automated and manual archiving system flow. Based on data usage or user action, data will be moved between the cloud provider's storage classes. Frequently accessed data will be in "hot storage," ensuring high availability. Less frequently access data will be kept in "warm storage" at a reduced cost, while data accessed only rarely will reside in "cold storage" at the lowest storage cost. Automated archiving and restoring will be done in such a way as to lower total storage cost over time. Users will be able to manually configure this process to trigger at certain points or to not trigger at all.

propose to develop a system to track data usage in order to automatically manage archiving, hence making cost management of data storage easier for users. This system would automatically archive users' uploaded input files and output files from workflows that go unused for 1) a pre-determined amount of time built in the policies for the archiving system, or 2) based on user action. The AnVIL will also manage the archiving of hosted data based on monitoring the usage and relevance of datasets within the research community with time. In essence, the data management system will also be used to automatically move dataset files between "hot", "warm" and "cold" storage classes, which will further reduce the cost of overall data storage for researchers.

C.1.c.ii. Development of a shared analysis and computing environment

The proposed AnVIL will provide researchers with a shared analysis and computing environment that is shielded from the complexities of their computational environment's underlying infrastructure. We envision that users will primarily interact with their own workspaces that sit on top of the storage and computation infrastructure. From the workspace, users will be able to access the hosted datasets, as well as easily and securely upload and annotate their own data. Data will be collocated with analytical tools to perform cost effective secondary analysis and interactive exploratory (visualization) analyses on the cloud. Seven Bridges has extensive experience building and maintaining similar cloud-based biomedical data platforms such as the CGC and Cavatica platforms.

User Administration via Admin panel

User administration on the AnVIL will be managed via an "admin panel" that will allow system administrators to control access to the platform and individual datasets. The admin panel will serve as the centralized point of control for the entire AnVIL, and allow administrators to manage resources, accounts, and access to data. In addition to controlling access to the platform, AnVIL administrators will be able to monitor usage statistics and generate reports about user storage and computation.

Compute workspace

AnVIL researchers will have access to secure collaborative workspaces to perform their computational analysis. These workspaces will serve as containers for data files, analysis workflows, and results (**Fig. 5**). Workspaces will facilitate secure collaboration with the ability to control fine grained permissions for multiple collaborators depending on the need. Collaborators can be assigned different levels of permissions: admin, copy, execute, read, or write. For example, users with "read" permissions can only view the contents within that. In contrast, users with "execute" permissions can run new analyses. Within a workspace, researchers with appropriate permissions may create, view, or re-run analyses and explore both the raw data and results. This allows research to be carried out collaboratively among a diverse group of researchers with varied expertise, while protecting potentially sensitive data. Researchers will be shielded from the complexities of the underlying computational infrastructure and have FAIR access to data and the bioinformatics tools/workflows. In addition, Seven Bridges already

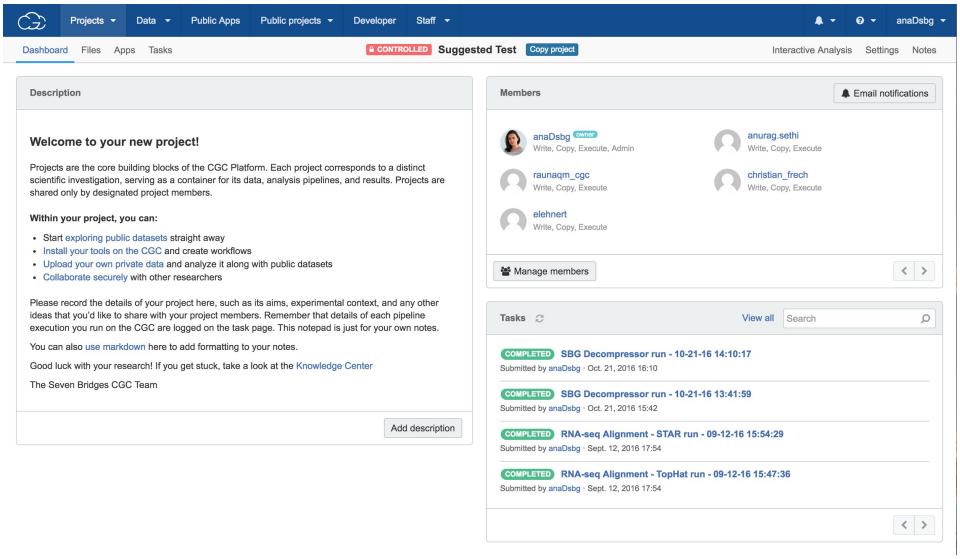


Figure 5. An example of a user's workspace on the Seven Bridges' Cancer Genomics Cloud. The workspace has sections showing name, description, members, and compute tasks at different stages. The AnVIL will use the same core technology to realize the concept of a secured workspace. Workspaces can be shared with other researchers by adding them as members to a workspace with granular control over permissions.

has documentation that details the process of creating and managing projects, user permissions, and how to execute tasks and workflows within a workspace.

Data transfer

A key component of the AnVIL will be to enable a diverse group of researchers of varying level of technical expertise to upload and download large data easily. To support researchers with varying degrees of expertise, we will provide data upload and download functions through the command line, graphical user interface, and API transfer mechanisms all of which enable data to be annotated using metadata manifest files. The SB-CI already offers these capabilities for reliable data transfer. In addition, the SB-CI provides researchers with the ability to access data on their personal cloud storage (on AWS or GCP) using the Volumes API. Researchers can access their private storage via API or GUI, making their data available for computation immediately, without the need to duplicate or move data to a different location. This will allow researchers to work with data that they already store on the cloud without incurring additional ingress costs.

Reproducible and scalable computational analyses

The AnVIL will provide researchers with an environment where they can upload and analyze their private data, as well as can generate meaningful insights by analyzing their data in combination with the datasets hosted on the AnVIL. Users will have the option to choose from a number of optimized ready to use tools and workflows, curated by ARC. Our team has extensive experience in optimizing computational pipelines for the research community and making them portable on the cloud for reproducible analysis. The SB-CI currently hosts 270 popular tools and workflows for typical secondary and tertiary analysis of genomic data. In addition, the teams from Baylor and Yale both have extensive experience developing and deploying pipelines for consortia, as described in Aim 2. For the AnVIL, the Yale team with contribution from Baylor will be responsible for curating our team's existing pipelines and tools on the AnVIL platform, making it easier for researchers to find the tools that fit their analysis.

The AnVIL users will also be able to develop their own tools and workflows on the platform by using the user-friendly, open source Software Development Kit (SDK) developed by Seven Bridges. The SDK wraps tools in the Common Workflow Language (CWL), a specification for ensuring reproducibility across analyses. A CWL description of a tool captures the command used to invoke the tool on the Unix terminal, the tool's required file types, parameter settings, and resources. CWL is also compliant with metadata inheritance and JavaScript expressions that enable dynamic typesetting of command line parameters and allocated resources based on different attributes of an input file including its metadata. To further guarantee reproducibility and ensure the tool can run, the actual tool binary is also captured using a software container technology, Docker, in conjunction with CWL. Thus, a CWL description contains the location of a Docker image ("container snapshot") that packages the bioinformatics tool along with its software dependencies. This container is automatically fetched and its contents dereferenced when the tool is run, ensuring that tools described with CWL can always run reproducibly. For instance, a CGC user used the SDK to deploy their own software for identifying patient-specific tumor neoantigens called CloudNeo on the platform[4].

Since 2014, the CWL specification has been adopted by several prominent research organizations (currently including Intel, NCI Genomic Data Commons, Institute of Systems Biology, Broad Institute, Harvard Chan School of Public Health, Institut Pasteur, Wellcome Trust Sanger Institute, and more) and is the emergent Global Alliance for Genomic Health (GA4GH)'s standard for describing tools and workflows. Based on the experiences gained from having multiple platforms in production, Seven Bridges has learned many valuable lessons that have improved the implementation of the CWL specification.

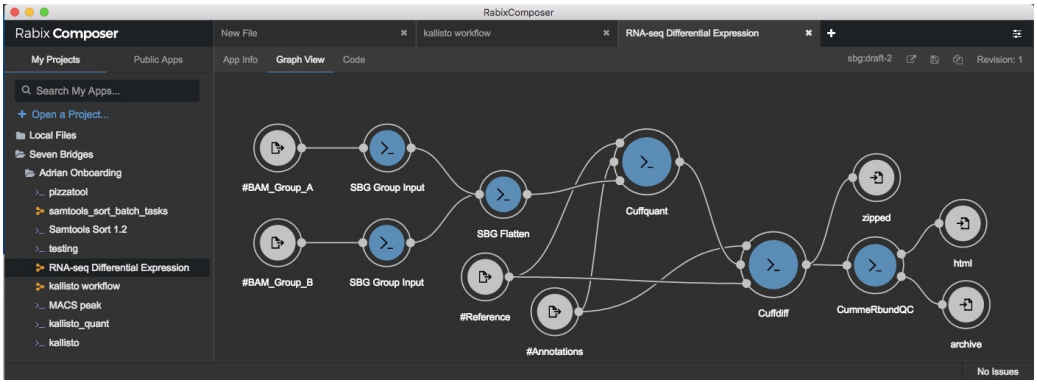


Figure 6. Rabix Composer, a local integrated development environment for CWL tools and workflows. Rabix Composer can sync tools and workflows between local and cloud environments to perform genomics analyses at scale.

Seven Bridges has developed an open-source CWL development and execution toolkit called Rabix (Reproducible Analysis for Bioinformatics), which helps researchers develop new tools/workflows locally and troubleshoot rapidly before paying for analysis on the cloud. Rabix currently consists of two components: Rabix Composer and Rabix Executor. The Rabix Composer is a standalone integrated development environment (IDE) with rich visual and text-based editors for rapidly describing CWL tools and workflows. It can be used as a standalone application or in sync with cloud platforms (**Fig. 6**). The Rabix Executor runs CWL workflows both in a local environment and on cloud infrastructures, allowing researchers to test workflows locally before moving to the cloud. AnVIL users will have access to the Rabix toolkit, which will further enhance workflow development.

In addition to the existing tools and workflows hosted by the Seven Bridges Platform, ARC proposes to migrate the workflows, developed by the Baylor and Yale groups, to uniformly process ENCODE and exRNA datasets that are highly relevant for the research community. This ensures that user uploaded data can be made interoperable with ENCODE or exRNA datasets by processing their data with the exact workflows used within these consortia. We expect that the NHGRI consortia members and researchers will share additional workflows developed through initial pilot projects, and ARC will assist in further optimizing these workflows if needed. We anticipate that AnVIL will become the go-to-place for the community to share the latest workflows. These activities will provide assessments and recommendations for best combinations of tools to process experimental datasets generated by the NHGRI funded programs. ARC will work with the NHGRI program officials and the research community to continuously populate the public apps gallery with highly relevant, optimized, and validated tools and workflows. Our detailed plans for bringing new workflows to the AnVIL will be further discussed in Aim 2.

Seven Bridges maintains interoperability with upstream and downstream services through a rich set of tools and APIs built in anticipation of a complex data creation and processing ecosystem. These tools include:

- a RESTful API, command line tool, and library bindings for major languages (Python, R, and Java, with Go coming soon);
- a command line interface tool, suitable for rapid control, query and scripting;
- a filesystem mounting tool that makes the files within a workspace available locally;
- a fast, reliable, and secure command line file uploader able to parse tabular metadata formats;
- the ability to connect cloud storage, such as S3 buckets, to mediate data exchange;
- interoperability with the existing apps by wrapping them as a standard tool on the platform;
- support for user-configured workflow automation.

The design philosophy for SB-CI is embracing standards rather than devising its own. In this way SB-CI has been successfully integrated with multiple analytics providers, such as OmicSoft Array Suite, Station X's GenePool, and SolveBio, as well as customized to automate complete workflows spanning raw data input from sequencers and LIMS systems to third party analysis and report generation.

Tertiary data analysis and visualization

One of the key differentiators of the AnVIL platform will be the capability for researchers to perform custom interactive and exploratory analyses on the cloud without the need to download or move data to another environment. A feature of the SB-CI, Data Cruncher addresses this need. Scripts can be written inside a JupyterLab notebook environment running within the workspace, and are executed inside a container-based

environment, ensuring reproducibility between runs. These notebooks can easily be shared with collaborators by providing them access to the workspace, as well as interchanged between Data Cruncher and other local or cloud environments with minimal modifications assuming data and library dependencies are met.

In addition to Data Cruncher, we will make the following interactive analysis tools readily available on the AnVIL: a **VCF Benchmarking** service to benchmark the performance of variant calling pipelines, and a fully-featured **Genome Browser** to visually inspect aligned reads and variants within BAM files. Seven Bridges will integrate the **Genome Browser** with **GenboreeKB.**

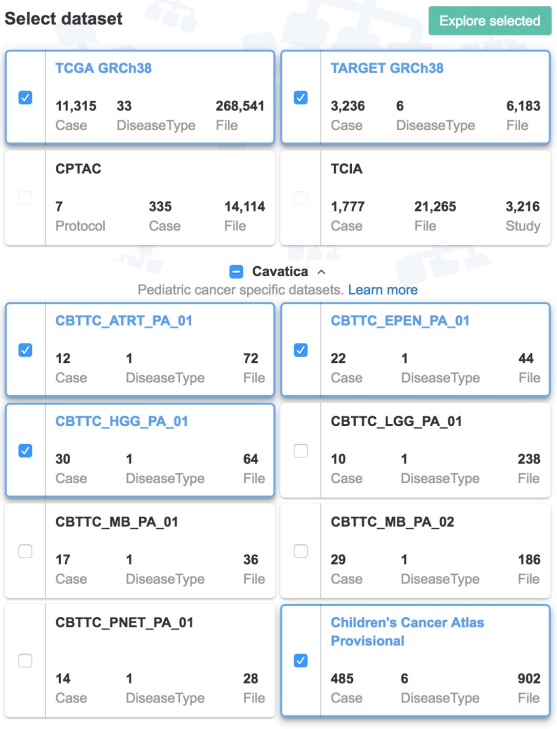


Figure 7. Data overview shows the landscape of the available data on the Platform. In this case, users can search across multiple datasets on the Cavatica platform. For the AnVIL, users will have a similar data overview with the ability to choose multiple datasets to perform cross query searches.

Visualization will be a core component of the AnVIL. In addition to an open specification and mature SDK for tool onboarding and pipeline development, Seven Bridges is actively working towards deploying a Visualization SDK as part of SB-CI. Researchers will be able to create, modify, publish and maintain their custom-built Visualization Apps. Visualization Apps will be built in JavaScript and able to use any of the numerous data visualization libraries widely available (D3.js, Processing.js, Chart.js, Highcharts, Plotly.js, etc.) Visualization Apps will be contained in and fully executing inside the user's browser, with data sourced from the AnVIL. Apps can then be published to a workspace and used to visualize data from files and other sources available in that workspace. In addition, select members can push apps to global *Visualization App Store,* from which apps can later be simply copied by other users to their respective workspaces.

Data management system

The AnVIL platform will host a data management system that will enable users to easily perform queries on datasets, including cross-dataset query for datasets. Access to individual datasets hosted on the AnVIL will be managed by implementing policies to govern controlled and unrestricted data access. These policies will be developed in collaboration with the dataset owners and the NHGRI officials. Seven Bridges has experience hosting multiple datasets on the CGC and Cavatica platforms. The CGC hosts two public genomic datasets, TCGA and TARGET, as well as the proteomics dataset, CPTAC, and imaging dataset, TCIA. Cavatica hosts 30 datasets related to pediatric cancer and other pediatric diseases, and these datasets can be queried together if described by a compatible schema, highlighting the importance of metadata harmonization for interoperability. To manage access to these datasets, two user access control mechanisms are used: 1) discretionary access controls, where the owner of the dataset or project can manage permissions of other users, and 2) attribute-based access control, where parts of the data is restricted.

Seven Bridges platforms use sophisticated data annotation and query mechanisms to make data usable for the researcher. Features such as the Data Browser, Case Explorer, and Dataset Overview are the favorites among users and demonstrate Seven Bridges expertise in making data usable. Based on our experience with making data usable, the AnVIL's dataset management system will include the following three major components:

1. **Data catalogue / Data overview:** Data overview will provide researchers the ability to quickly look at the landscape of all data available on the AnVIL platform (e.g., **Fig. 7**). Users will be able to select multiple datasets from the data overview. The data overview will present information about the key descriptors such as cohort demographics, data types collected, etc. for each dataset hosted on the AnVIL. The key descriptors will be identified in coordination with the dataset owners. Details of the datasets that we will propose to host on AnVIL are stated in Specific Aim 2.
2. **Data Browser:** The Data Browser (**Fig. 8**) was developed initially for the NCI Cancer Genomics Cloud (CGC) for users to query TCGA data. Since then, the Data Browser has been adopted as the preferred

method to quickly find and access genomics data on the CGC and has been deployed on several other Seven Bridges' platforms. Seven Bridges will leverage the existing data browser technology and redeploy on the AnVIL with refinements specific for the hosted datasets. The underlying framework of the data browser will be adjusted depending on which datasets are approved by the NHGRI program officer, DSC and EAC to be hosted on the AnVIL.

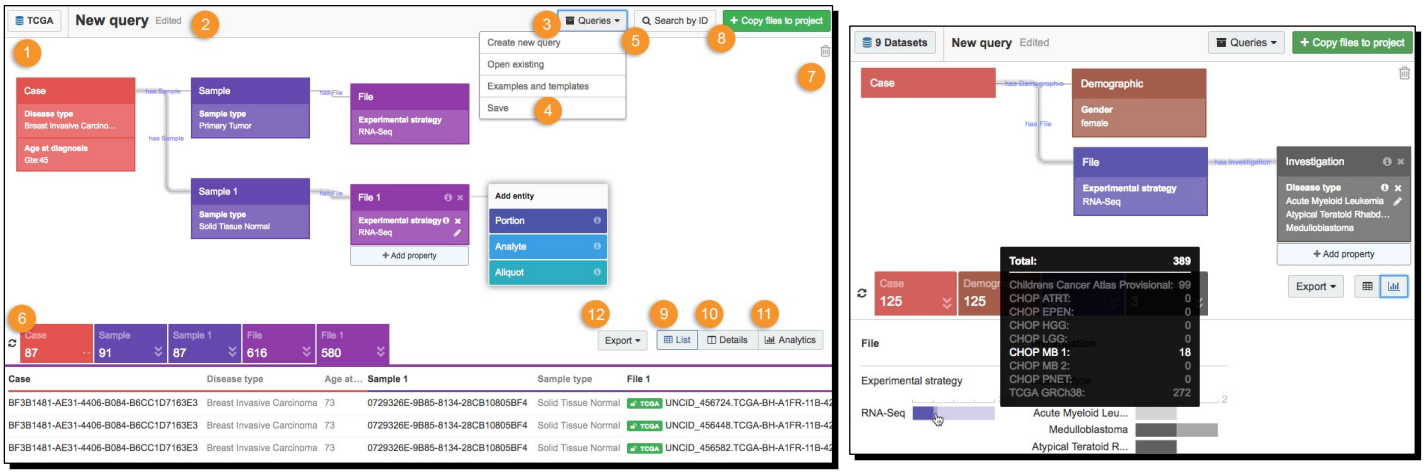


Figure 8. The Seven Bridges Data Browser. (A) The data browser is a visual query builder to quickly find and access genomics datasets stored on the platform. Queries are built using a point-and-click interface. Main control elements include: (1) name of active dataset; (2) query name; (3) and (4) actions such as creating a new query and loading/saving existing queries; (5) search a dataset entry by ID; (6) count feature quantifying the scope of the data returned by your query; (7) delete all previous queries and open a new query canvas; (8) copy files that satisfy your query to your desired analysis project; (9) List View, displaying UUIDs for entities matching query; (10) Details View, displaying selected entity with inbound and outbound connections; (11) Analytics View, displaying distributions of matching entities as graphical bar charts; (12) export query results as a CSV, JSON, or XML file. **(B)** Aggregate query results on the CAVATICA data browser. Researchers can perform queries across multiple datasets to gain insights otherwise not possible by analyzing an individual dataset in silo.

3. **Upload and share a dataset, metadata harmonization:** Data harmonization will be performed with strict adherence to the FAIR principles as they're developed by the scientific community. The AnVIL platform will facilitate interoperability with genomic resources in other data commons following FAIR principles. The detailed approach on metadata harmonization and modeling is discussed in Aim 2.

Workspace publishing

The AnVIL will provide a collaborative environment where researchers can share their datasets and workflows with the broader scientific community. Users can build their tools using the SDK, and subsequently publish these tools. A mechanism will be developed where the researcher can publish all materials relevant to their workspaces, including data, workflows, notebooks to the general public and/or users of the AnVIL. This also involves a "workspace freeze" so that further modification to the workspace is not allowed to prevent unintended modifications to the data, tools, or workflows.

The AnVIL will also provide researchers the flexibility associated with publishing workspaces. Researchers who want to share their workspace to facilitate the peer-review process will be able to provide anonymous access to journal editors for reviewing all the data, tools, analyses, and visualizations (through data cruncher) associated with the manuscript. The user credentials will automatically expire within a preset time period, and will have to be regenerated for each journal submission. Going through findings in full transparency will be critical to enable reproducibility of science. Once the manuscript is published, the researchers will also be able to publish their workspace and make all their data, tools, and workflows accessible to the entire AnVIL community but no further modifications to the workspace will be allowed once the workspace is published to preserve its accuracy to the published findings. Such capabilities will help lower the replication crisis in science because each analysis in a scientific publication will be visible to the entire research community in a reproducibility centric environment.

C.1.c.iii. Cloud services cost control

The performance and cost competitiveness of the cloud services offered by the AnVIL will be carefully monitored and users will be provided with tools to estimate and monitor costs. Seven Bridges will leverage the existing SB-CI's mechanisms to monitor and estimate compute and storage costs including:

- On demand computation, thus starting virtual machines only when-needed;
- Optimized scheduling algorithm that efficiently reuses provisioned resources;
- AWS spot instances handled by SB-CI job-try mechanisms, where large cost savings can be utilized;
- Per second billing by AWS for greater compute cost savings

For storage cost control, we will use deduplicated storage, where the actual physical file is only stored once on the S3 bucket. The same storage is utilized while file can appear on many users and on many logical locations. In addition, as described earlier, we will develop an archiving system for automatic management of data storage based on policy and usage patterns. We will have built in generic policies for automated archiving, but users will also be able to set policies that can trigger archiving.

ARC will develop a price calculator tool for AnVIL with a simple user interface for users to estimate costs related to computing, storing, and data transfer for their specific use case. Giving users the ability to estimate costs will support cost-conscious work on the platform.

To manage the billing system for the AnVIL platform, ARC proposes to use a billing group entity, similar to what Seven Bridges provided for the CGC. Users will be able to view a detailed breakdown of their spending on computation and storage, and a monthly invoice will be generated for per user account. We will pass through all AWS costs directly to the user with no markup. Additionally, project administrators will be able to monitor, moderate, and restrict access to compute resources by placing credit limits on researchers. This provides administrators with greater control to manage compute and storage costs. The final billing model will be established after consultation with the NHGRI staff and the External Advisory Committee (EAC). Directly supported payment methods requiring no additional development for AnVIL are checks, wire transfers, credit cards, and the upcoming support for virtual credits issued by the third party.

C.1.c.iv. Data security and access

ARC recognizes the need for the AnVIL to be compliant with regulatory frameworks, including logging, auditing, versioning, and quality control processes which are needed to enable compliant use of human genomic information on the cloud. Seven Bridges already has gained the *NIH Trusted Partner* status and *FISMA Moderate Authority to Operate* for the Cancer Genomics Cloud. ARC will ensure controlled and secure access of to data and cloud resources on the AnVIL.

Data security

Secure data transfer to and from the AnVIL is a key concern given that the platform will host sensitive patient data. Seven Bridges has demonstrated the ability to provide a mature infrastructure for handling sensitive data on the cloud through the SB-CI, Cavatica, and CGC. Data is uploaded to an encrypted storage bucket. At the end of the data life-cycle, a strict data purging policy ensures that all data is safely deleted if it is no longer needed on ephemeral storage or when an authorized user chooses to delete data on the platform. The SB-CI is also secured by having all AWS computation instances run within Virtual Private Clouds (VPC), logically isolated networks within the AWS cloud that are kept only minimally open for the necessary external and internal access. In addition, users can choose to isolate all computation resources through an "Instance Lockdown" mode that disables any access during the computation, even by platform admin staff. Best security practices regarding tenancy, such as AWS Architecting for HIPAA, are followed. We propose to employ these security features for the AnVIL and ensure that all data transfers on the AnVIL platform will be secured end-to-end using industry-standard AES-256 TLS connections.

The security and compliance of AnVIL will be guaranteed through Seven Bridges' comprehensive framework to ensure end-to-end security and control of genomic data in the cloud. This framework covers three main areas:

- **Data Privacy:** Ensuring that all sensitive data are kept safe during its full lifecycle. This includes data encryption (in storage and in transit) and secure user authentication.
- **Platform and Infrastructure Security and Auditing:** Ensuring that the software platform and its underlying infrastructure (server and network) support the secure architecture.
- **Security Controls:** Ensuring security of the system by implementing administrative, technical, and other security controls.

SB-CI leverages the broad spectrum of built-in compliance and security features provided by industry-leading cloud infrastructure providers, including trusted cloud providers Amazon Web Services and Google Cloud Platform. Data is encrypted in storage, in transit, and in ephemeral storage. At the end of the data lifecycle, a strict data purging policy ensures that all ephemeral data are safely deleted.

Data access

SB-CI contains components for providing user authentication and authorization for controlled data access based on dbGAP credentials and permissions. ARC will use the methodology established for the CGC to integrate the SB-CI with dbGAP and other evolving authentication mechanisms. We will also make the AnVIL user authentication system interoperable with any new system and processes for regulatory workflow for data access control policies as they emerge from the NIH. In addition, because of Seven Bridges' participation in the NIH Data Commons Pilot Phase, our team will be able to quickly adapt, implement and integrate with any new security standards developed by the Data Commons to provide seamless interoperability of the AnVIL with the NIH Data Commons.

User access will be governed by secure authentication and support for further access controls, such as client-encryption, integration with a client's single-sign on solution, and integration of external key management (such as AWS KMS), with two-factor authentication on the roadmap. All data and software that users add to the AnVIL will be secured for their access, but can be selectively shared with collaborators using a spectrum of granular permissions. Beyond these measures, the AnVIL will secure access to production and development environments through Virtual Private Networks. For administrators, audit trails and logs will be available for a range of activities on the AnVIL, including upload, download, analysis, and querying of data, sharing of workflows, and API calls. Each analysis is fully replicable, and each data asset can be traced back to the analysis (and inputs, parameters, workflow) or upload mechanism and the user that produced it.

Third party audits

The Seven Bridges Information Security Team conducts regular security audits of the platform and organization based on industry standard frameworks such as ISO 27001 and NIST 800-53 Rev 4. Seven Bridges has completed ISO 27001 certification and we currently hold an ATO from the NIH at a the FISMA Moderate level for our CGC Platform. We are prepared to undergo a FISMA assessment and other annual third-party assessments as required to confirm regulatory compliance.

Pitfalls and Alternative Strategies

A significant risk to progress on the development of the AnVIL are prolonged development cycles that result from unclear product specifications and inadequate planning. To achieve clear product specifications, we will utilize our established project management practices which include focused meetings to understand product requirements. This will be followed by detailed designing and prototyping before investing in scaled up development effort. We also understand there may be delays in data and/or workflow availability due to issues with access approvals or consensus. We will avoid this problem by engaging in best practices of project management and frequent meetings to reach consensus quickly.

In addition, a common pitfall in the software development process is a disconnect between development and deployment. Even if the solution passes extensive testing in the testing environment, inadequate simulations can often lead to issues when the solution is deployed in the user environment. We will mitigate this risk by using our experience with continuous deployment cycles. Any deployments on the AnVIL will be tested in an environment that closely reflects a production environment, taking into consideration network traffic, compute stress, and storage traffic. This testing will minimize interruptions for AnVIL users.

C.2. SPECIFIC AIM 2: To enable analysis and visualization of cloud-hosted genomic datasets and integrate AnVIL with the federated Data Commons ecosystem.

C.2.a. Scientific premise for Aim 2

The key to the AnVIL's initial adoption will be the availability of most important datasets and of the analysis and visualization tools to unlock their value to the community. To ensure AnVIL's early adoption, ARC will establish at the outset a regular process to prioritize the datasets to be hosted on the AnVIL. To enable discoveries from hosted datasets and ensure early adoption of the resource by a diverse set of users, we will deploy standardized pipelines, shareable exploratory workflows and interactive analysis and visualization tools to support use scenarios of high interest for each audience. To ensure that the hosted datasets are not siloed, AnVIL will be built on FAIR principles. The principles will be extended beyond project-specific data to also include research resources such as biosamples and computable genomic knowledge, thus extending interoperability between research and clinical domains (see **Fig. 9**).

C.2.b. Preliminary Studies for Aim 2

Developing data standards, metrics, and pipelines for ENCODE, modENCODE and NIH Roadmap Epigenomic data. Yale team has a 10-year history of data analysis for the ENCODE and modENCODE projects[5] . Rozowsky previously led data analysis for mod/ENCODE[6, 7] and has extensive experience in developing advanced pipelines for all ENCODE and modENCODE datatypes including DNAseq, RNAseq, and ChIP-seq. Milosavljevic lead data analysis and coordination for the NIH Roadmap Epigenomics project and has extensive record of accomplishment in genome sequencing, ChIP-seq, and DNA methylation mapping. The teams helped develop standards for all the functional genomic data generated by the consortia. The standards for analyzing ChIP-seq experiments were jointly published[8].

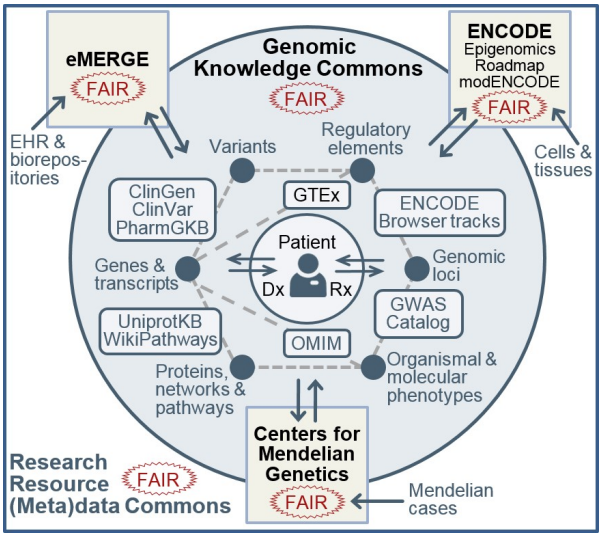


Figure 9. A vision of a genomic data and knowledge Commons built on FAIR principles. In addition to applying FAIR principles to project-specific data, the same principles will be extended to research resources such as EHR-linked biosamples and to computable clinically actionable knowledge, thus creating interoperability between the research and clinical domains.

Pipelines for RNA-Seq data. The Yale team has extensive expertise with both long and short RNA-Seq data. *RSEQtools*[9], a computational package enables identification of splice sites and gene models[9]. Aggregation and Correlation Toolbox (*ACT*) is a general-purpose tool for comparing genome signal tracks[10]. Database of Annotated Regions with Tools (*DART*) package contains tools for identifying unannotated genomic regions that are enriched for transcription, as well as a framework for storing and querying this information[11]. *FusionSeq*, pipeline detect transcripts that arise due to trans-splicing or chromosomal translocations[12]. Our tools have been adopted by several major consortia, including long RNA-Seq analysis tools by mod/ENCODE[6, 7] and short RNA-Seq pipelines by the Extracellular RNA Communication Consortium (ERCC)[13].

Pipelines for ChIP-Seq data. The Yale team helped develop the ENCODE ChIP-seq data processing pipeline. Dr. Rozowsky developed PeakSeq[14], a versatile tool for identification of TF binding sites and a standard peak calling program used by the ENCODE and modENCODE consortia for ChIP-Seq datasets[14]. We also played an important role in developing the IDR method for determining reproducibility of ChIP-seq experiments. We also developed a new peak caller, MUSIC[15], for multiscale decomposition of ChIP signals to enable simultaneous and accurate detection of enrichment at a range of narrow and broad peak breadths.

Pipelines for whole-genome bisulfite sequencing and DNA methylation analysis. As part of the Roadmap Epigenome project, Milosavljevic laboratory developed and benchmarked[16] algorithms for the analysis of

whole genome bisulfite sequencing data and studied the role of DNA methylation in gene regulation[17], cellular differentiation[18], and genomic stability[19].

Pipelines for mapping structural genome variation from whole-genome profiling data. Milosavljevic laboratory has an extensive record of accomplishments in the analysis of structural variation from whole-genome profiling data[19-21].

Deployment and performance analysis of whole genome sequence analysis pipelines on commercial cloud platforms. In collaboration with the Human Genome Sequencing Center at Baylor College of Medicine, Milosavljevic laboratory performed pioneering analyses of performance of whole genome sequencing data analysis pipelines on commercial cloud computing platforms[22, 23].

Centers for Mendelian Genetics (CMG). The Yale group is part a site of the CMG and provides bioinformatics expertise to the consortium.

Developing metadata standards and harmonizing metadata for major consortia. Milosavljevic and Cheung have led multiple national and international metadata standardization and collaborated on community-based ontology development[18] and data interoperability development[13]. Milosavljevic led the effort to harmonize ENCODE and the NIH Roadmap Epigenomics data[24] and drafted epigenomic metadata standards currently adopted by the International Human Epigenome Consortium (IHEC). Cheung and Milosavljevic developed ontology-driven data standards and performed semantic data integration for the Extracellular RNA Communication Consortium (ERCC)[13] and extended Gene Ontology by adding ExRNA related terms and relationships[18]. Cheung chaired the BioRDF Task Force of the W3C Semantic Web for Health Care and Life Science Interest Group and is currently leading data modeling efforts for the Human Immunology Project Consortium (HIPC) where he mapped ontology terms to the data templates that are used for data submission to ImmPort[25]. Cheung's metadata-ontology mapping work has been integrated into CEDAR's NCBI-compatible templates for metadata submission to the NCBI BioProject, BioSample and Sequence Read Archive (SRA). These templates are used for entering metadata as part of the data submission process and include 36 metadata entry fields (11 for BioProject, 15 for BioSample and 10 for the SRA), five of which are constrained by ontologies.

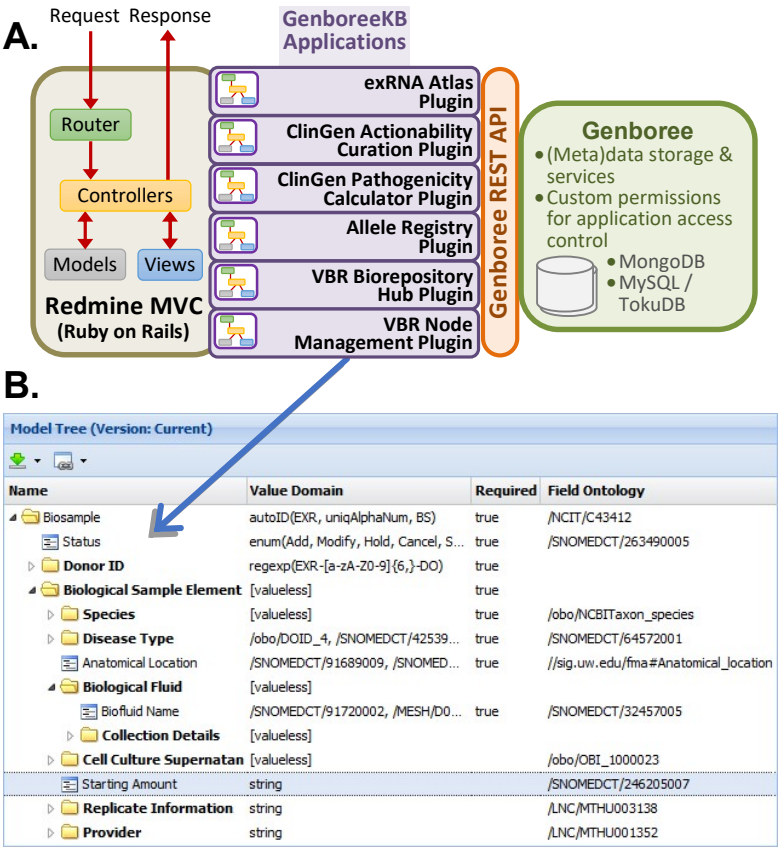


Figure 10. (A) Overview of the GenboreeStack. Each application plugin employs multiple *GenboreeKB* document models for their entity metadata. **(B)** Illustrative document model for the *Biosample* entity used by the *VBR Node Management* plugin**.** The domains of document values are ideally ontology-backed, although non-ontology options are available.

GenboreeKB platform. GenboreeKB, a component of the GenboreeStack (**Fig. 10**), is a metadata and knowledge modeling, processing, validation and hosting platform. It embodies over ten years of experience in metadata and knowledge modeling by the Milosavljevic group. GenboreeKB is built on top of the document-oriented MongoDB database. Applications are built on top of GenboreeKB using the Genboree plugin model (**Fig. 10**). GenboreeKB provides APIs that allow interoperability and fine-grained access to structured data.

Genboree Plugins: Virtual Biorepository Hub and Nodes. GenboreeKB is built in an API-centric way to support distributed deployment: different instances can be hosted across clouds, providing access to locally hosted (meta)data via uniform APIs to applications such as data

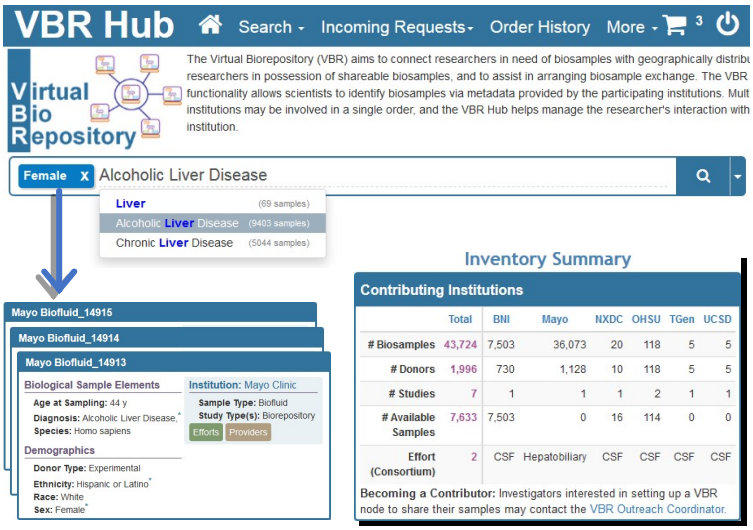
indexers and GenboreeKB hubs for unified search across distributed instances. Most recently, we employed this capability to integrate over 40,000 clinically annotated biosamples from six institutions within the Virtual Biosample Repository (**Fig. 11A**). VBR consists of a Hub and multiple Nodes hosted locally at participating institutions or on the cloud. VBR connects researchers in need of biosamples with specific clinical profiles with the institutions and clinicians that have shareable biosamples with associated clinical information.

Genboree Plugins: exRNA Atlas Portal. The exRNA Atlas Portal is a metadata-driven data portal for the NIH Common Fund ExRNA Communication Consortium implemented as a Genboree plugin (**Fig. 11B**). Similar custom metadata-driven portals can be implemented for any metadata-annotated omic dataset.

Genboree Plugins: ClinGen Pathogenicity Calculator. Developed for inclusion in the Clinical Genome (ClinGen) Resource, the Calculator automates reasoning about pathogenicity of genetic variants according to community-developed guidelines, documents evidence for conclusions, thus helping resolve discordant conclusions, and is configurable to accommodate any rule-based reasoning about variants[26].

Genboree Plugins: The ClinGen Allele Registry (CAR). The CAR serves as a critical "network-adapter" for translating between allelic representation standards, and for linking each allele to information within and beyond ClinGen about the variant[26]. The registry is designed to serve ClinGen curators and individual users via a web interface and also applications such as curation tools via robust, reliable and well-documented APIs. Rather than being a centralized up-to-date snapshot of all variant information—an increasingly unrealistic proposition in the Personal Genome Era—the Registry enables instantaneous high-bandwidth registration of new variants, providing dereferenceable canonical URIs for variants, thus enabling just-in-time linking of data and knowledge about both previously known and new variants.

A.



B.

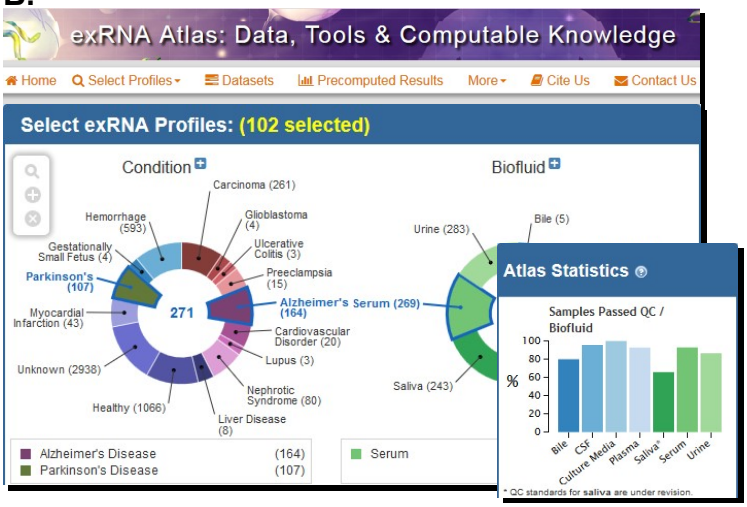


Figure 11. (A) Virtual Biorepository (VBR) Hub. **(B)** exRNA Atlas Portal.

Conveying clinically actionable knowledge through EHR systems as part of ClinGen/eMERGE collaboration. As part of our ongoing ClinGen/eMERGE collaboration we designed a framework for the delivery of information to Laboratory and EHR systems about the actionability of genes and pathogenicity of variants via the HL7 InfoButton standard[27, 28].

Integration between Cytoscape and Jupyter Notebooks. Cytoscape is now the de-facto standard network visualization tool in bioinformatics community. Since 2006, Alex Pico's team from Gladstone has contributed to the core development and outreach of Cytoscape, including App Store, which averages over 1000 downloads per day across 330 apps. Cytoscape itself is downloaded 18,000 times per month and launched 4000 times per day. Pico team is contributing and supporting R and python libraries for interacting with Cytoscape via Jupyter Notebooks[29].

WikiPathways, developed by the Alex Pico group, captures the collective knowledge represented in biological pathways. Web services are documented by OpenAPI and resources are annotated to FAIR standards. The number of metabolite nodes has recently doubled[52].

API-based integration of a GenboreeKB plugin for the exRNA Atlas with WikiPathways. As part of the ExRNA Communication Consortium (ERCC) effort to advance data linking within an interoperable ecosystem of web applications[13], Pico and Milosavljevic integrated GenboreeKB-hosted exRNA Atlas dataset and tools with WikiPathways and Cytoscape capabilities via APIs, allowing interpretation and visualization of "omic" analyses in the context of networks and pathways[30].

C.2.c. Approach to Aim 2.

C.2.c.i. Genomic datasets, phenotypes and metadata

Selection of datasets for hosting. Working with the Data Steering Committee (DSC), ARC will establish criteria and a process for selecting datasets to host on the AnVIL. Outreach efforts (Aim 3) will engage the community nominating datasets. Nominated datasets will be scored based on immediate usability, complementarity with other hosted datasets, long-term value to the research community, and effort required to bring the datasets to FAIR standards. The scored criteria along with additional considerations such as NHGRI program priorities will help the Committee prioritize datasets for hosting. We anticipate selecting two new project/datasets per year in Years 3, 4 and 5. Lists of candidates, nominated and currently hosted datasets will be reviewed at least quarterly. Similar process will be established for tools (as described below under **Community tool promotion**). We propose to initially on-board the datasets from the following three active, complementary, and highly visible projects that are representative of the diversity of the NHGRI portfolio.

1. Electronic Medical Records and Genomics (eMERGE) is a premier genomic medicine project that serves as a model for the UK Biobank, the PMI's All Of Us cohort, and the US Department of Veterans Affairs' Million Veterans Program (MVP). eMERGE has already sequenced a panel of 109 actionable genes (including the "ACMG 59" set) in 14K individuals. Current data also includes 4K WES and 2K WGS datasets under protected access in dbGaP, 12K exome chip profiles, and 83K GWAS profiles. By the end of 2019, eMERGE will have collected EHR-linked biospecimens for 135K individuals. Assuming eMERGE continues to accumulate WES and WGS data at its current pace, we extrapolate that the project will accumulate 12K WES exomes and 6K WGS genomes by the end of 2019. When genome sequence information is coupled with informative EHR records, a new "virtual" reverse-genetic approach called Phenome Wide Association Studies (PheWAS)[31] becomes possible. PheWAS utilizes typically up to 100 phenotypes inferred algorithmically from EHR records from the general population to detect clinical associations of these phenotypes with specific variants. PheWAS scales to more study participants by orders of magnitude than traditional GWAS because GWAS requires highly selected and thoroughly phenotyped cohorts. On the other hand, EHR-based phenotypes cannot harness the full potential for discovery from the vast amount of genomic information on the other end of the association equation. One exciting new approach addressing this "phenotyping bottleneck" is metabolomic profiling. MVP and TOPMed are already applying Metabolon technology to profile thousands of individuals for correlative analysis with genomic information. Because they address the key "phenotyping bottleneck" in an effective way, we anticipate that metabolomic and other omic profiling will become more widespread during the course of the AnVIL project.

Return of incidental findings in eMERGE. An additional aspect of eMERGE is the return of "incidental findings" of pathogenic variants within actionable genes in sequenced patients. This step illustrates the power of genomics to affect clinical care or cause change in behavior to prevent disease for which one may have a predisposition. As part of the ClinGen project we have developed tools and knowledge base of gene actionability that informs the selection of genes targeted by eMERGE for incidental findings. As part of our ongoing ClinGen/eMERGE collaboration we designed a framework for the delivery of information to Laboratory and EHR systems about the actionability of genes and pathogenicity of variants via the HL7 InfoButton standard[27, 28].

***2. Centers for Mendelian Genomics (CMG*)** cross-links the basic and medical components of the NHGRI portfolio by revealing human gene function and identifying clinically relevant genes. In contrast to eMERGE's focus on common diseases within a relatively healthy population, CMG focuses on rare—typically severe—Mendelian disease phenotypes that segregate within families. While eMERGE pursues relatively shallow EHR-based phenotyping and genome sequencing, CMG focuses on deep phenotyping and deep genomic analysis for causative variants. While smaller in sample and sequencing volume, CMG will provide a testing ground for

deep genomic characterization, including analysis of structural variation from whole-genome sequencing data, an area where Milosavljevic has significantly contributed in the past[19-21]. Genome sequencing data from the Centers for Mendelian Genetics is in the form of raw reads (BAM files) as well as called variants (VCF files) are available under protected access in dbGaP.

Metabolomic Molecular Phenotyping in CMG and eMERGE. Mendelian disease studies also increasingly rely on metabolomic profiling using untargeted mass spectrometry. Not surprisingly, metabolomic profiling was first used to diagnose inborn errors of metabolism[32]. Application of this first (outside of cancer) clinically deployed "omic" molecular phenotyping technology is now expanding to undiagnosed diseases via the NIH Common Fund Undiagnosed Disease Network (UDN) and, as mentioned above, is being used by eMERGE-like projects such as MVP and also by TOPMed. We therefore anticipate that metabolomic profiling will play an increasingly important role in discovery and diagnosis. Baylor is pioneering application of metabolomic profiling in diagnosis of Mendelian disorders and Milosavljevic is collaborating with Dr. Elsea, the Senior Director of the Biochemistry Laboratory at Baylor (*see Letter of Support from Dr. Elsea*) on analyzing thousands of metabolomic profiles already collected from patients. Because accurate interpretation of metabolomic profiles requires pathway knowledge, metabolomic profiling will also serve as an early testing ground for the deployment computable pathway knowledge to inform the interpretation of molecular phenotypes[33].

3. Encyclopedia of DNA Elements (ENCODE) has become the premier genome biology project within the NHGRI genome biology portfolio. It recently absorbed the datasets from NIH Roadmap Epigenomics and modENCODE projects. ENCODE contributes annotation of regulatory elements such as enhancers to help interpret genetic variation within regulatory regions. This information has helped interpret GWAS variants within regulatory elements that associate with common diseases. The information about regulatory elements provided by ENCODE will also be essential for interpreting whole-genome sequencing datasets from eMERGE and to a somewhat lesser degree CMG (as Mendelian traits tend to be caused by protein coding variants). Moreover, modENCODE (also now part of ENCODE) will facilitate biological validation through animal models. With BCM's extensive experience in leading Roadmap Epigenomics and integrating it within ENCODE and Yale's expertise in leading the development of ENCODE, modENCODE datasets, we do not foresee major difficulties on-boarding these datasets to the AnVIL and deploying that tools that will help inform interpretation of genetic variants in regulatory regions. Data that will be incorporated into ARC from the ENCODE and modENCODE projects will maintain the file formats adopted by the ENCODE DCC for the various data types for both raw read data (BAM files) as well as processed output files (BED region files and BIGWIG signal tracks) for the functional genome data including RNA-Seq, ChIP-Seq (TF and chromatin marks), DNase hypersensitivity and ATAC-seq assays. The read data is publicly accessible in the SRA archive. As part of the associated metadata, we will also include the quality control metrics from ENCODE that we helped develop.

Storage estimates. We estimate that cloud storage will dominate the overall cloud cost. Because the future directions of eMERGE are currently being formulated, exact estimates of sequencing data cannot be made. Storage for the compressed data adds up to slightly less than 0.5PB for compressed eMERGE data alone assuming up to 15GB in BAM file storage per exome at clinical-grade coverage and up to 60GB in BAM file storage per genome. Multiplying by two to allow for derived non-compressed information and intermediate results, we estimate that 1.0 PB will suffice for eMERGE data. An additional 1PB will be reserved for the ENCODE and CMG datasets. Assuming the projects continue during Years 2-5 of AnVIL, we plan to secure an additional 2PB storage for that period, bringing the total to 4PB for the initial projects for the whole period.

Data coordination for new data generation projects. In addition to hosting independently generated datasets, the AnVIL will also be available to serve as a data coordination center for new data-generating projects. This may entail development of new data and metadata models and workflows for processing, analysis, and visualization. Although in the rest of this aim we do not refer explicitly to data coordination for new data generation projects, the combination of capabilities and processes described will be deployed toward both data coordination for new data generation projects and for the hosting of independently generated datasets.

Pipeline integration and deployment

To analyze their own datasets in conjunction with those already hosted on the AnVIL, the users will have to

process their datasets using standardized pipelines. Building on our extensive experience in data processing and coordination, we will compile and deploy the standardized pipelines using CWL and SB-CI tools described in Aim 1. In addition to being deployed on the AnVIL, the pipelines will be Dockerized and made available on widely accepted platforms such as Docker Hub.

Benchmarking of published and user-supplied analysis pipelines. Using our extensive experience in pipeline development, relevant analysis pipelines will be identified and benchmarked against an agreed set of agreed gold-standard datasets. The selected benchmarking datasets as well as documented tool performance will be shared via public AnVIL workspaces. To test new tools against the benchmarks, the tool developers will be able to port their tools onto the AnVIL using SB-CI's SDK and copy these published benchmarking workspaces on to their own private workspaces. When desired, the tool developer would also be able to publish their own workspaces on the AnVIL so that all their benchmarking analysis is made public, thus promoting their tool to the community. The Yale and BCM teams will review benchmarking results and use them to help select tools for optimization by the Seven Bridges team and promotion to the status of AnVIL public apps.

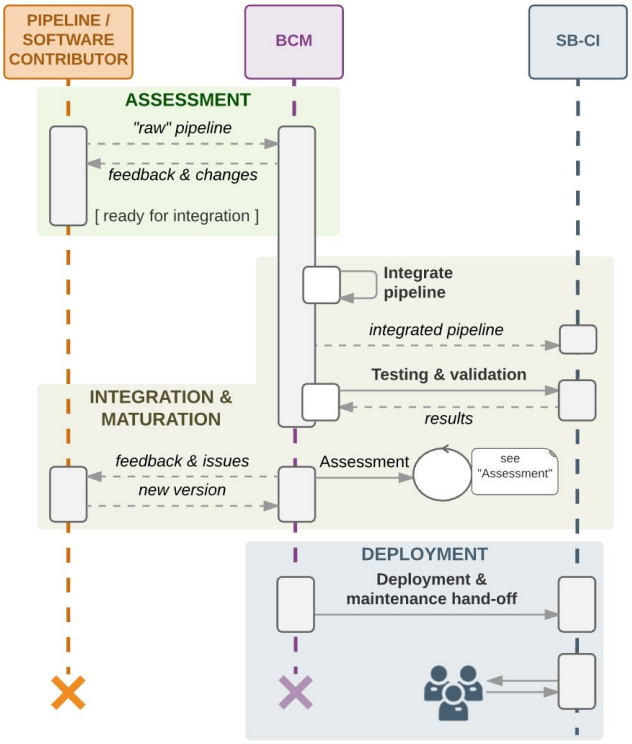


Figure 12. Pipeline assessment, integration and maturation, and deployment.

Pipeline development and maturation. The Seven Bridges Platform currently host more than 270 popular tools and workflows for standard genomics analyses. Over the past five years, Genboree stack hosted over 30 pipelines for genomic, epigenomic, and, metagenomic analysis. The latter also includes key pipelines developed by Yale group. Building on this experience, we defined a pipeline development and maturation process as illustrated in **Fig. 12**. As indicated in the figure, the pipelines will typically be developed in collaboration with developers ("Contributors" in **Fig. 12**) and made available to various consortium users. We will extend the extensive pipeline development experience of ARC to accelerate development of minimal viable products, shorten user-feedback release cycle during pipeline maturation, and maintain a stable schedule of regular improvements after deployment. All developed pipelines will be deployed on the AnVIL thus ensuring high-bandwidth, low-latency access to cloud-hosted high-volume data.

Ensuring pipeline FAIRness. To ensure FAIRness of pipelines, their portability and reusability, and the capture of procedural and execution details pertinent to specific computation environments, we will continue to document all pipelines using CWL. In addition, CWL is also compatible with metadata inheritance leading to an increase in the FAIRness of data generated using tools and workflows described in CWL. As part of the NCI Cloud Pilot program, Seven Bridges developed their fourth-generation open source SDK (Rabix) for enabling users to generate CWL-documented pipelines.

Data standardization and harmonization. To improve data interoperability as well as lower storage/bandwidth cost, community data standards for final and intermediate outputs of data processing pipelines will be identified and implemented. To further improve the ability of users to compare and reuse datasets from different sources, the AnVIL may perform data harmonization by re-processing raw data, e.g., re-running of variant calling pipelines to harmonize variant calls on datasets from various NHGRI and NIH genotyping and genome sequencing studies, or re-running RNA-Seq pipelines to map sequence reads from multiple expression studies to the same human reference genome and consistently detect transcribed variants and splicing events. The data harmonization projects will be identified and prioritized based on cost/benefit analysis and unique scientific opportunities.

Deployment of exploratory workflows

Hosting of JupyterLab Notebooks via Data Cruncher. The automated pipeline-based data processing is often followed by a less structured, more exploratory and interactive data analysis and visualization. Jupyter Notebooks provide means to capture data exploratory workflow in executable, reproducible and shareable form. Using the SB Data Cruncher, an integrated cloud provisioning solution for Jupyter Notebooks (**Fig. 13**), users of the AnVIL will be able to share existing workflows, modify custom input/outputs and workflow parameters or even develop their own workflows. The workflows will be stored the user's workspace for future use and sharing. The Notebooks of wide utility will be curated for readability and robustness and optimized for speed of execution.

Community tool promotion

Leveraging AnVIL resources, we will organize outreach and training activities (**Aim 3.**) to engage the research community around tool development and sharing on the AnVIL. Hackathons will help develop ad-hoc Notebook prototypes of potentially high utility. We will establish a "tool promotion" process (**Fig. 13**) where the most successful *ad hoc* Notebooks will be selected for either curation for wide use or as working prototypes for the development of pipelines or UI-centric applications. A list of tools in each category—most contributed by the community—will be maintained and reviewed regularly.

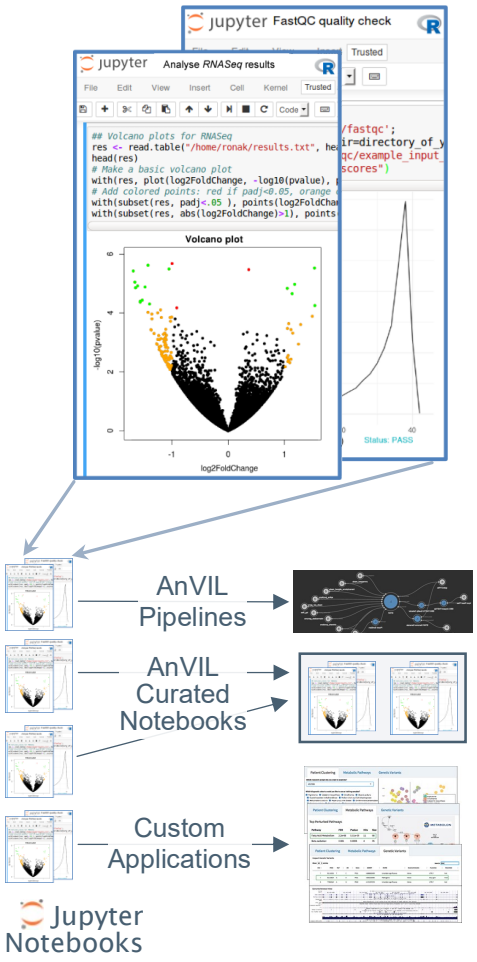


Figure 13. Prototyping and exploratory analysis via Jupyter Notebooks can lead to AnVIL Pipelines, Curated Notebooks, or Custom Applications.

Deployment of UI-centric applications of highest relevance for the hosted datasets

The AnVIL will provide visualization of genomic data within genome browsers, including both the UCSC genome browser and IGV, as well as an innovative and more interactive SB Genome Browser. For visualization of multi-layered datasets that include transcriptomic, epigenomic and other quantitative "omic" layers of information we will deploy multi-dimensional data visualization tools from the community. Building on our experience with integrating WikiPathways and Cytoscape, we will enable visualization of "multi-omic" perturbations in network and pathway contexts. For example, the Gladstone team has developed interactive pathway viewing widgets that can accept parameters to highlight specific genes, proteins and metabolites based on AnVIL datasets. We will also leverage the latest web and automation support from Cytoscape to provide script-based workflows (e.g., in R or Python) and JavaScript elements (buttons, links and viewers) to send AnVIL data to a running instance of Cytoscape, enabling data exploration, integration, and visualization.

We will prioritize applications that best unlock the value of hosted datasets. For example, for projects such as eMERGE, CMG, MVP, UDN and TOPMed that already combine or are likely to combine genomic sequencing and metabolomic profiling of patient samples, it will be of interest to develop an application that helps identify variants causing metabolomic perturbations in affected patients. One effective visualization tool for the pathways is Metabolync, a Cytoscape app built by Metabolon. Metabolync extends the capabilities of the Cytoscape by allowing the user to directly import additional study data via APIs.

For the purpose of tracking genome sequencing data, all the variants would be registered with the ClinGen Allele Registry using the existing high-bandwidth registration service, allowing the integration of up-to-date variant-specific information across projects (see **Fig. 14**, below). By virtue of the ClinGen Allele Registry, variant centric information will be integrated to other key resources that help in interpretation of genetic

variants, e.g. Exome/Genome aggregation consortium, ClinVar, dbSNP, and PubMed (**Fig. 14**).

For some metabolomic disorders, single analyte scores alone are not sufficient to clearly identify a metabolic perturbation. Advanced analyses address this situation by performing "metabolomic matchmaking" to identify clusters of patients with highly specific yet similar metabolic pathway perturbations and then examine whether they have pathogenic variants in the same gene or pathway. The number of metabolite nodes in WikiPathways has recently doubled[52] indicating the beginning of a virtuous cycle of the growth of computable knowledge that will improve ability to interpret the data. ENCODE tracks will help inform the interpretation of any variants discovered by WGS that fall into regulatory regions. As the size of metabolomic profiles within the AnVIL and federated across the Data Commons grows, so will be the potential to make new discoveries by metabolomic and other types of "molecular phenotypic matchmaking" in the context of the growing Commons of FAIR computable knowledge.

Metadata modeling, processing, and harmonization

(Meta)data Identification. Unique GUID / DOI identifiers will be issued for both the datasets as well as metadata documents describing them. When assigning these identifiers, we will follow the same process we developed for hosting the TCGA dataset as part of the NCI Cloud Pilot project. Because the standard ways of assigning GUIDs within the NIH Data Commons are still being developed by the FAIR Standard WG, being participants in this group, we will contribute toward the development of the standard and will implement it in the course of this project.

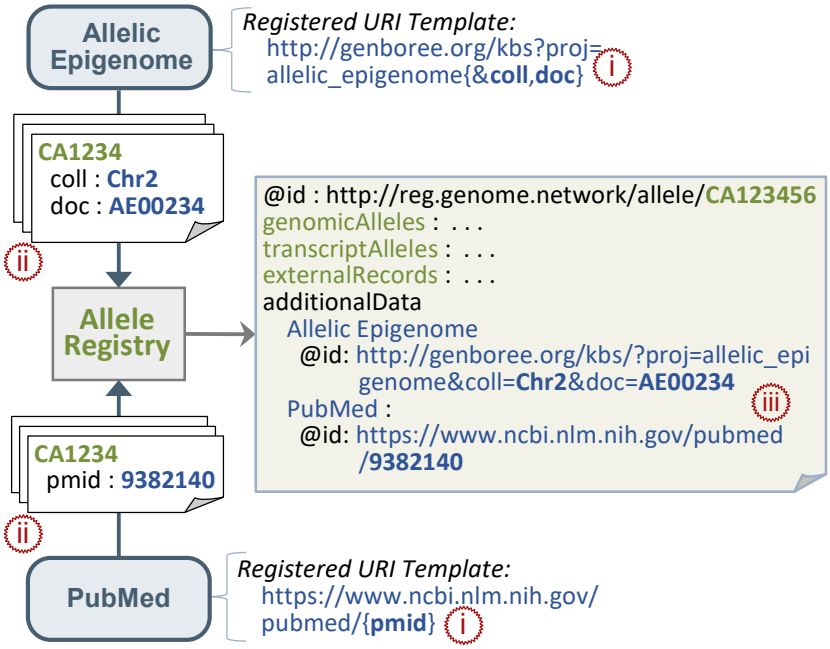


Figure 14. Proposed Allele Registry services will allow *novel* sources of variant information to link their (meta)data to registered alleles. This open system comprised of: **(i)** information source registration via RFC6570 URI Templates (Levels 1-3) and **(ii)** declaration of allele↔n-tuple association. Unlike dedicated import efforts by the allele Registry team, the sources themselves can drive knowledge linking. Thus, in addition to the ExAC, dbSNP, gnomAD, and other external data *already* present in the Allele Registry, a growing knowledge corpus that includes this additional information becomes available to the wider scientific community **(iii)**. Depicted are illustrative examples of two sources—"Allelic Epigenome"[1] and PubMed—adding links from alleles (CA Id) to their own documents, and the result for consumers of variant knowledge.

Metadata modeling and processing on the Genboree KB platform. The evolution of GenboreeKB will continue during the course of this project to support increased automation required for scaling to tens to hundreds of thousands of biosample and EHR record data for controlled sharing as part of eMERGE and related massive sequencing projects. GenboreeKB will adopt relevant emerging standards within the community such as tracking of evidential support via the SEPIO model developed by the Monarch project[34]. GenboreeKB will continue to be freely available as part of the GenboreeStack open-source package, enabling academic research groups to perform data coordination.

Metadata standardization. Metadata will be standardized to ensure data FAIRness and participation in a federated genomic Data Commons ecosystem. We will leverage Dr. Cheung's previous work to create AnVIL-compatible metadata templates with links to ontologies such as SNOMED CT (https://www.snomed.org/snomed-ct) and Human Phenotype Ontology[35] that are relevant to AnVIL datasets (e.g., eMERGE[36]). We will expand our ontology mapping approach to include common data elements (CDE's) linked to ontology concepts such as those in LOINC[37]. In addition to the ontology mapping tools (Annotator and Ontology Recommender) that we have previously used, we will explore the use of natural language processing (NLP) tools including those based on UMIA [38] to

facilitate lexical and semantic mapping of CDE's. UMIA-based tools like cTAKES[39] can incorporate UMLS[40] vocabularies--particularly relevant to eMERGE--into the NLP annotation pipeline. To curate the mapping results, we will use the browsing tools and application programming interfaces (APIs) provided by common data element repositories. For example, NCI and other NIH institutes have developed CDE repositories (e.g., caDSR[41]). We also will make use of the NIH CDE Repository[42] that provides an integrated access to CDE's created by different NIH institutes and projects including PhenX[43]. We will also pursue continual improvement of our metadata modeling and curation processes to improve efficiency and quality. Being both a CEDAR member and experienced in GenboreeKB modeling, Dr. Cheung will be in an ideal position to integrate elements of the CEDAR submission pipeline with high-throughput processing and validation features of the GenboreeKB system to allow the populated metadata templates to be ingested at high throughput while minimizing human curation. Upon validation, all the metadata will be exported via GenboreeKB to the SB platform in the form of RDF-serializable JSON-LD documents for hosting. In addition, the AnVIL will also capture clinical information from biorepositories, sample-associated EHR records and other sources using the VBR Virtual Biorepository, also built on the GenboreeKB platform.

Metadata modeling process. For new projects, harmonization of existing datasets, as well as for the purposes of Virtual Biorepository, we will perform metadata modeling using GenboreeKB. Every document collection within GenboreeKB conforms to a specific document model that enforces constraints such as controlled vocabularies, ontologies, or CDEs that are specified within the model. The modeling projects will precede and overlap with metadata capture (described below) and will consist of the following three stages. 1. Development: through communication with the (meta)data provider, we will extract information about their metadata content; ideally there is already an existing schema in some form, but in our experience, this is often not the case; thus, typically, metadata model needs to developed using GenboreeKB document modeling tools. 2. Deployment and metadata *content* capture for model testing: the proposed model is instantiated in GenboreeKB, becoming accessible through generic GenboreeKB UI and through domain/context-specific APIs that wrap generic GenboreeKB REST APIs; if the model is not mature, it is tested by capturing metadata content (as described below). 3. Maturation and archiving: once mature, an approved model is released by exporting to AnVIL platform and deployed for metadata capture (as described below).

Metadata capture process. Metadata capture will be performed in the following three stages (with some variations depending on whether this involves a new project, harmonization of existing datasets, or Virtual Biorepository). 1. Metadata development: the metadata will either be pulled from the external source or uploaded by the external source using available tools; typically, several metadata validation and refinement cycles and communication with the source will be required before the metadata is ready for deployment. 2. Metadata deployment and maturation: if the metadata is untested with the target audience, it will be exposed to users and software via domain/context-specific UIs and APIs for usability testing; this enables specialized program portals, novel external software, existing external software to access and test the metadata; upon successful usability testing, the metadata will be deployed via the SB platform. 3. Deposition to AnVIL for permanent distribution and archiving: once the metadata are mature they will be transformed from its non-LD structured form to an appropriate JSON-LD/RDF representation for deposition on the SB Platform; a GenboreeStack plugin provides Transformation API services for these purposes and a growing number of Templates that define various transformations.

Capture of pre-experimental metadata and clinical data using Virtual Biorepository. To support projects such as eMERGE, the AnVIL will capture clinical information from EHRs and other sources and metadata about biosamples using the Virtual Biorepository that is built on the GenboreeKB platform and already contains over 40,000 registered biosamples with linked clinical records from six institutions. This information is often managed locally albeit in often unstructured and non-standard ways. Means are needed to make subsets of this information FAIR to enable collaborative projects while putting data owners in control of where the data is located and how and with whom it is shared. As part of this project, we will incrementally extend VBR to meet the needs of the project by empowering data owners to populate their VBR nodes (cloud-hosted or hosted on local hardware).

C.2.c.ii. Participation in a federated data and knowledge commons

Implementation of FAIR principles. The approached described above will enable the AnVIL to serve as a

nucleation point of the federated data ecosystem by providing access to data and knowledge based on *FAIR principles*. As members of the GA4GH and NIH Data Commons working groups that focus on various aspects of FAIRness we will contribute to refining operational definitions based on the principles and implementing them. *Findability* will be ensured through the use of unique dereferenceable URIs for data resources and—when appropriate–GUIDs coupled with the structured metadata descriptors of biosamples and data developed using standardized Ontologies and Common Data Elements, as described above. This will enable indexing by bioCADDY[44], web search engines, and navigation via the SB Data Browser and, in case of the Virtual Biorepository by the VBR Hub. *Accessibility* will be ensured by exposing the (meta)data for access via Open REST APIs using GUIDs and dereferenceable URLs. Access will be subject to authentication and authorization through SB Workspaces, as described above. For *Interoperability,* the metadata will be delivered as JSON-LD documents that are serializable as RDF. Data will be delivered using standard file formats such as VCF. To ensure *Reusability*, metadata modeling (as described above) will include detailed provenance information such as inputs and processing steps described using used to either physically process material samples or algorithmically derive processed datasets using CWL-described workflows. Scientific evidence for computationally derived conclusions will be tracked using Monarch Initiative's SEPIO model or other emerging GA4GH and NIH Data Commons standards.

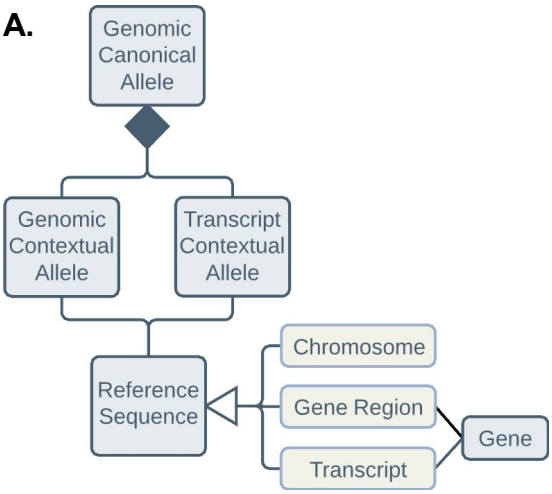


Figure 15. ClinGen Allele Registry data model, originally developed by the ClinGen Data Model Working group. The model indicated the conceptual entities, rather than actual implementation. Alleles are canonicalized during registration via HGVS representation. Following canonicalization, contextual alleles can be generated procedurally and are not explicitly stored.

Going beyond the FAIRness of data itself we extend application of FAIR principles to resources such as EHR records and biosamples and to the computable genomic knowledge (as illustrated in **Fig. 9**). This will be essential for lasting impact of the AnVIL as genomic knowledge is not only produced during data analysis, but also it also informs the analyses, creating a virtuous cycle of knowledge growth.

Making the knowledge about genetic variants FAIR. Much of the data and knowledge about genetic variants--beyond that "siloed" within databases-- is disjointed and not FAIR. For example, it is currently not possible to collate information about an indel across multiple sources without performing an "Extract-Transform-Load" process that includes a de-duplication step, typically alignment against the same reference assembly. Even that approach typically fails to recognize identity of common genetic variants such as indels since indel alignments are not unique. Thus Findability (the "F" in FAIR) of different bits of information about any particular genetic variant is currently limited. To address this problem, we will integrate variant knowledge by employing ClinGen Allele Registry services to deliver dereferenceable URIs for canonical variants (**Fig. 14** and **Fig. 15**, above). The Registry itself delivers basic information about the allele as RDF-serializable JSON-LD. A "source registry" will provide URL patterns for linking to additional sources of information about the registered variants, some hosted as GenboreeKB document collections (**Fig. 14**). The GenboreeKB collections will be accessible via Open APIs and will provide layers of variant information in the form of RDF-serializable JSON-LD documents.

Making the knowledge about other genomic features FAIR. Knowledge about other genomic features such as regulatory elements, non-coding RNA species etc. suffer from the same problem as variants. Using the Allele Registry as a model, we will address it by developing and deploying additional Registry services for those elements and by enabling layering of FAIR knowledge from sources distributed within the federated Knowledge Commons.

Building and using computable and FAIR network and pathway knowledge. Network and pathway knowledge increasingly provides the context for visualizing and interpreting experimental or natural perturbations--such as mutations--within biological systems. Very often such knowledge is not FAIR and sometimes not even accessible for review and critical examination. WikiPathways addresses this problem by

making it contents FAIR and accessible in standard formats such as RDF[52]. Moreover, by providing the community means to contribute new pathway knowledge, WikiPathways feeds the virtuous cycle of knowledge creation where the increasing amounts of stored knowledge lead to its larger utility and faster growth through crowdsourcing. Through the contribution of Alex Pico, the leader of the WikiPathway project, the AnVIL will partake in this virtuous cycle. Of particular interest will be the pathways and networks that will help interpret genetic and metabolomic perturbations that will be increasingly detected in large scale sequencing projects such as eMERGE and CMG. By providing dereferenceable URIs for genomic elements such as regulatory elements and by linking them to regulated genes (via GTEx-derived regulatory networks, for example) and to variants, we will be creating FAIR computable knowledge about functional effects of variants that may be used for predicting their phenotypic effects.

Making evidential support for conclusions FAIR. In addition to the knowledge itself, proper interpretation of conclusions, particularly for actionable information in clinical setting will also require that evidential support be FAIR, whether the conclusions are reached by a human or an algorithm. Building on the capabilities of the Pathogenicity Calculator, we will extend support for FAIRness of evidential support by adopting the SEPIO ontology developed by the Monarch Initiative that was also adopted by the GA4GH and is now implemented by ClinGen. Using SEPIO, we will extend transparent reasoning beyond genetic variants to also include evidential support for computationally derived inferences about regulatory and other genomic elements.

Making updates FAIR via messaging. Genomic data and knowledge is in an ever-accelerating flux. It will therefore become increasingly important for the relevant inferences from the data to be updated without much delay, particularly those leading to clinically actionable knowledge.

Rapid and automated update propagation is a key process within social network sites such as LinkedIn, the original developer of the Kafka messaging system that supports update propagation. The open-source version Apache Kafka is currently being tested within ClinGen as a means of propagating updates about actionability of genes and pathogenicity of genetic variants. Building on this experience, towards Y3 or Y4 of the grant period, we would have gathered enough information to deploy a similar notification system within the AnVIL as well as provide a connecting point with applications and web services outside the AnVIL such as ClinGen.

We anticipate the largest number of initial use cases to revolve around the propagation of updates about genetic variants. For example, a “listener” program may tune into the stream of updates about any evidence for or against pathogenicity of a variant seen in a patient. The ”listener” may employ the Pathogenicity Calculator to detect any change in status from “Unknown Significance” at initial diagnosis to Pathogenic or Benign in light of newly available evidence. For example, additional information about a positive PheWAS[31] association with disease from eMERGE may provide evidence for pathogenicity while a sufficiently high allele frequency within a healthy population may provide evidence that the variant is benign.

We note that the messages are not typically intended to contain full information, only the information required for the recipient to identify messages of interest and request any additional information via a REST API from the original source. This is where we come back to FAIR principles since the additional information must be Accessible, Interoperable and Reusable. Update propagation may therefore serve as one of the key “killer apps” for the federated Data Commons, providing “use scenarios” to guide its development and end-user utility for wide adoption.

Pitfalls and Alternative Strategies

The utility of the resource will depend on the ability of the tools to produce meaningful results for the users. To minimize the risk that wrong tools are prioritized, we will help recruit accomplished experts into the External Advisory Committee and regularly seek their input. We will also identify categories of tools relevant for AnVIL-hosted datasets and will maintain “top lists” for each category.

It may not be possible for us to identify the pipelines that would exactly reproduce the results obtained by eMERGE, CMG and ENCODE consortia from raw reads. To maximize the value generated by AnVIL’s efforts, we will evaluate the impact of these differences and the benefits and cost of reprocessing using FAIR and reproducible pipelines before deciding where to allocate resources.

C.3. SPECIFIC AIM 3: To implement a targeted outreach and training strategy and responsive governance.

C.3.a. Scientific Premise for Aim 3

To ensure effective adoption of the AnVIL resource, a comprehensive social media outreach and intensive training strategy will be led by BCM with participation from Yale, Gladstone, and Seven Bridges. The AnVIL's database and tool deployments will be accompanied by online and workshop training on dataset exploration and integrative analyses that incorporate their own data**.** Being one of the major centers of genetics and genome research, BCM will provide an opportunity for in-person testing of training materials. The outreach and training will be tailored to diverse researchers, including novice users, advanced users with programming experience, and contributors of data and tools. An agile governance structure will ensure responsiveness to evolving priorities of users, funders and other stakeholders.

C.3.b. Preliminary Studies for Aim 3

Managing user base for production level cloud based platforms. The Seven Bridges private industry biomedical data analysis platform, the Seven Bridges Platform, is used by more than 6,500 researchers from leading pharmaceutical, biotechnology, and clinical diagnostic labs throughout the world, manages nearly 16 petabytes of highly diverse data, hosts more than 270 bioinformatics apps and pipelines, and averages more than 50 running compute instances each second.

Personalized and enterprise-level support. Seven Bridges provides context-specific training to ensure resource adoption through ~20 collaborative academic research projects across the world, trained in the use of its platform, SDK, and API. (See letters of support from CGC users). In addition, Seven Bridges provides 24/7 help desk support and personal guidance on how to initiate, optimize, manage, share, and scale analyses using various datasets. Seven Bridges' extensive online library of training and educational materials, tutorials, white papers, and videos cover all skill levels and a wide range of topics, such as workflow execution and optimization. Baylor/Yale/Gladstone has developed dozens of online tutorials, use cases, and videos for the exRNA Atlas, Virtual Biorepository, exceRpt, Allele Registry, and Pathogenicity Calculator (**Aim 2**) created through their Roadmap, ERCC, ENCODE and ClinGen participation.

Workshops and hackathons. Training workshops that educate users on methods and tools for "big data" analyses are essential for researcher adoption[45-47]. Baylor has developed fifteen in-person training workshops for ERCC (8), Epigenome Roadmap (5), and ClinGen (2). Attendees (>1,000 total) included graduate students, postdocs, basic, translational, and clinical faculty, software developers, and research staff. Domain expertise spanned genomics, transcriptomics, and epigenomics, as well as biomarker and therapeutic experts.

Baylor workshops. ERCC workshops brought together experimental biologists and developers to encourage developer-user exchange and advance tool development, and incorporated hands-on data analysis exercises, including data (download/upload) and sample (Virtual Biorepository) access, and data analysis using dozens of tools and online resources including exRNA Atlas, exceRpt, WikiPathways, DESeq2, and BioGPS. Epigenome Roadmap workshops were 2-day hands-on events with real-time data analysis using web-based tools: MACS, LIMMA, TopHat, CuffLinks, Cuffdiff, and Spark, etc. and integrative analyses using reference epigenomes produced by Roadmap, ENCODE, and IHEC. Baylor's "Introduction to Epigenome Analysis" workshop attracted 200 researchers (2013 ASHG). ClinGen workshops (ASHG 2015, and Curating the Clinical Genome 2017) described how ClinGen collects assertions on genes and variants and develops approaches and bioinformatics resources (Allele Registry, Pathogenicity Calculator) to enable assessment and curation processes. Highly interactive use cases used real data to train >150 researchers and clinicians on procedures and online tools to curate genes and variants of clinical significance. Post-workshop surveys were administered, and feedback used to guide improvements.

Seven Bridges workshops and hackathons. Hackathons provide unique, highly collaborative learning opportunities and are highly valued by researchers and developers[48, 49]. In 2016, we received a NIH BD2K grant to organize a hackathon on cancer data analysis using open source tools. We also curated an Applied

Knowledge Exchange Session on collaborative and reproducible genomics at a global scale at the Intelligent Systems for Molecular Biology meeting (bit.ly/AKES16 Applied Knowledge Exchange Session; ISMB 2016). We have also hosted numerous workshops at Stanford, Weill Cornell Medical Center, Harvard, and MIT, and have conducted multi-day workshops internationally. Topics included the SB-CI SDK for wrapping user-specific tools, and workflows and the SB-CI API for scaling and automating their analysis on the platform.

Supporting community initiatives. DREAM Challenges crowdsource data analysis solutions for fundamental questions in systems biology and translational medicine. The Seven Bridges Cancer Genomics Cloud hosts resources and facilitates the submission process for both the SMC – RNA DREAM Challenge[50] and the GA4GH – DREAM Workflow Execution Challenge[51]. For each of these challenges, Seven Bridges created detailed tutorials and webinars to help participants get started.

Consortia meetings. Baylor has planned eight ERCC semi-annual meetings (~125-325 researchers). We worked closely with NIH and ERCC PIs to create agendas, select oral/poster presentations, invite keynote speakers, organize panel/roundtable discussions, and arrange review meetings with external scientific advisors. Online Google docs captured meeting/workshop registration, abstract submission, and attendee feedback to guide improvements. Baylor interacted directly with hotel management to secure all facilities.

C.3.c. Approach to Aim 3

C.3.c.i. Outreach and Community Engagement

The AnVIL community must have an active voice early in the development process to ensure alignment with community needs. Outreach will leverage Seven Bridges 3,000 registered users, 4,188 LinkedIn followers, and 2,255 Twitter followers. Beta access (Q2 of Y1) to select users will be identified in consultation with NHGRI stakeholders, EAC, and DSC, with educational workshops at key institutions, and smaller scale events at major meetings to increase awareness. Impact will be assessed by user feedback and new user enrollment immediately after such events. However, long term adoption will be assessed by the number of active (daily, weekly, monthly) users over time, which better reflects integration of the AnVIL into their research and discovery process. Metric review of task failure rates (user vs platform errors), and the actual source (Docker error, cloud infrastructure, etc.) will improve AnVIL usability and reduce user- and platform-related errors.

The AnVIL portal. ARC will develop the AnVIL portal (login required for non-public information) to serve as the central informational resource, leveraging our experience in creating the Cancer Genomics Cloud, Cavatica, and ERCC portals. The AnVIL portal will include ARC-generated training materials, use cases, resource links, blog, and AnVIL user-generated publications, etc. Portal usage metrics (tools/pipelines accessed, data uploads/downloads, user contributed tools/pipelines, top portal pages visited, etc.) will be monitored using Google analytics and reported quarterly to NHGRI. We will establish an AnVIL Facebook page and Twitter account to publicize events and developments. Twitter engagement will be measured by number of followers and clicks, likes, replies, and retweets. User behavior data will inform training events and resource additions.

Workshops and training materials. Workshops and training materials will be customized to AnVIL target audiences: researchers (non-programmers), bioinformatically trained biologists (use Jupyter, re-run pipelines), content providers (extend AnVIL, contribute tools). ARC will develop an extensive collection of tutorials, white papers, and videos (AnVIL YouTube channel) to train new users on basic features, and to describe for advanced users porting workflows and using the API to automate and perform large-scale analyses. We will host one 2-day on-site workshop during Y1 and at least two 2-day training workshops (one every six months), developed in parallel with dataset and tool deployments (**Aim 1** and **Aim 2**). Training events will be widely publicized through the portal and social media channels. Tool and workflow usage will inform workshop themes.

User pilot projects. ARC believes that collecting user feedback is the best way to improve production platforms. To incentivize the initial use and generate feedback we will offer users free credits for pilot projects. Free credits encourage users to run analyses risk-free and explore all aspects of the AnVIL without the burden of incurring costs. Based on Seven Bridges' success with free user credits on the CGC, we propose $10,000 of user credits per year towards first-time AnVIL users. On the CGC, users were given $100 each and offered

$1,000 in additional credits if they agreed to provide feedback to Seven Bridges. Therefore, following the CGC model, we will provide 100 beta users with $100 credits to cover their compute costs. With each subsequent year, we will continue providing user credit for general use ($100 for new users), while providing a small group of users with up to $1,000 in credits to carry out a pilot project. The selection of users getting $1,000 will be performed in a competitive manner. We would request proposals from the community for researcher-initiated projects that can be performed on the AnVIL. The proposals will request basic information about project requirements related to data and metadata, workflows, and new tools. A working group will be established from ARC researchers, EAC, DAC, and NHGRI representatives. After review, projects that are most impactful for the AnVIL's mission and deliver the most value for a dataset or research objective will be selected for funding based on the imputed compute/storage needs up to $1,000. **Table 1** shows a yearly breakdown of number of users budgeted for pilot projects. We understand the actual number of users may be higher than budgeted, however, we will seek to provide additional user credits by leveraging our cloud provider relationships.

AnVIL annual meetings. Prior to the AnVIL kickoff meeting, ARC will reach out to NHGRI, EAC, and DSC to solicit input on content and identify needs and expectations. ARC Management Committee (AMC) will present ARC goals, plans, and administrative and logistical responsibilities, and outline the AnVIL execution plan. We anticipate AnVIL annual meetings to initially include ~50 attendees participating in oral presentations and working group breakout sessions. Review sessions will assess ARC progress, AnVIL performance metrics, user adoption, and ongoing development plans. As the number of AnVIL users grows we anticipate expanding AnVIL meetings to accommodate a larger user base.

Table 1. Number of users supported for pilot projects per year

Year	*# users supported with $100 credit each*	*# users supported with $1000 credit each*
Year 1	100	0
Year 2	80	2
Year 3	60	4
Year 4	40	6
Year 5	20	8
Total:	300	20
Total # of users supported with credits for pilot projects:		320

AnVIL scientific and technical seminar series. ARC will recruit AnVIL community researchers to present at a monthly webinar on their experience with the AnVIL. This outreach program will begin in year two and be modelled off the ERCC scientific webinar series (20 presentations, >560 attendees to date) wherein presentations are recorded and posted on exRNA.org portal (presenters may opt out). A monthly webinar provides a "soft-touch" approach to build awareness, identify training needs, and recruit tool contributors and workflow developers. Speakers will be recruited via Seven Bridges registered users, social media followers, and early AnVIL adopters.

Continuous improvement of AnVIL by incorporation of user feedback.

Qualitative feedback will be collected on AnVIL platform features by SB-CI ("Leave Feedback" and "Get Support" buttons), direct email, and feedback received internally from ARC members. Qualitative feedback will be captured, prioritized, tracked, and resolved in an internal issue tracking software supported by Seven Bridges' existing Customer Support. Quantitative platform metrics will be collected and tabulated in concordance with the mechanisms developed and implemented for NCI's Cancer Genomics Cloud Pilot; namely: weekly platform uptime, number of critical issues, number and categorization of issues received via qualitative feedback, number of platform-related failures, number of issues per platform category (misused tools/pipelines, invalid data, insufficient instance capacity, etc.), average first response time over all issues, and time to full issue resolution. Workshop surveys will be used to improve training events and materials.

Customer Support. The AnVIL will be supported through Seven Bridges' Customer Support Team that operates in multiple time zones with a typical response time of <24 hours including weekends and holidays. The team uses Atlassian's issue tracking software, JIRA, to ensure timely responses, escalation, and generation of regular reports, and has responded to more than 3,000 customer inquiries over the last 5 years. During the beta release period, ARC members will be supported through Seven Bridges' existing Customer Support architecture. In the production period, a dedicated Customer Support Representative will resume those operations, operating from within the customer support but collaborating closely with AnVIL personnel.

Pitfalls and Alternative Strategies. It may be that pilot project users need more resources than anticipated to

learn how to use the AnVIL. We observe a bimodal distribution of credit use on the CGC, such that a significant percent of users spend less than $20 to get started, and leave the rest of their credits unused. To extend the credits to more users we would discuss with EAC tier credits: a beginner $20 credit upon signup and an additional $80 only after the beginner credits are exhausted. Therefore, we could increase the effective distribution and use of the user credits. Also, workshops and hackathons may not allow enough time for less advanced users to explore all datasets or tools of interest, or for more sophisticated users to sufficiently evaluate technical features and capabilities. We may therefore need to extend workshops, add an additional trainer or, alternatively, provide follow up webinars for those needing additional assistance.

C.3.c.ii. Governance

Overview & Rationale. The AnVIL Management Committee (AMC) for the proposed project is composed of Aleksandar Milosavljevic PhD (PI), Isheeta Seth, PhD (co-I); and Joel Rozowsky, PhD (co-I).

Rationale. The proposed ARC brings together academic and private industry partners to deliver on the goals of the NHGRI AnVIL initiative. Herein we describe the Leadership Plan by which the AMC will govern the scientific and administrative direction of the proposed research.

Communications Plan. The AMC will communicate weekly by tele- or videoconference to discuss project execution and progress including administrative, fiscal, and regulatory responsibilities. At the beginning of the project, AMC meetings will be more frequent, and members will jointly oversee a kickoff meeting along with key personnel from BCM, Yale, and Seven Bridges to discuss project objectives, milestones, tasks, roles, and responsibilities. As the project progresses, we will hold monthly cross-team progress meetings that will allow all ARC members to receive progress updates and discuss/brainstorm any impediments. Each individual team will hold more frequent meetings and scrums to arrange development cycles. Also, ad-hoc cross-team project teams led by the product owner/internal stakeholders will assemble based on specific deliverables as needed. Dr. Milosavljevic will be responsible for submitting reports and initiating all communication with NIH. Any agreements and/or presentations made during meetings with NHGRI staff, External Advisory Committee (EAC), and/or Data Steering Committee (DSC) will be approved by the AMC (see *Decision Making and Conflict Resolution*). AMC members will meet with their respective teams routinely to evaluate progress and milestones. They will discuss any obstacles and act immediately to remove them. Such working meetings will lead to clarifying processes, and help the AMC determine if additional resources are needed to accomplish a specific goal of the project. These working meetings frequently result in the generation of new ideas, further driving success of the project. Additional one-to-one communications (e.g., between key personnel, laboratory, and administrative personnel) will take place on an ad hoc basis. There will be an assigned scribe in every meeting to document notes and action items which will be kept in a shared repository (Google Drive) and communicated through an ARC Wiki. For intra and inter development teams' communication we plan to establish a Slack group.

Scientific Direction. AMC will directly supervise the administrative, technical and scientific responsibilities of the grant on a day-to-day basis. AMC members understand each other's specific contributions to this project and share a commitment towards meeting the goals and aims of the AnVIL project. ARC recognizes the need to consult with experts in policy development and bioethical/ethical, legal and social implications (ELSI). To meet this need ARC will regularly consult with ELSI experts (see *Letter of Support from Prof. Amy McGuire, JD, the Director of the Bioethics Center at Baylor*). This being a U24 award, there will be extensive involvement of NHGRI staff in the scientific direction of the project. Much of this scientific governance will come from interactions with the various working groups specified in the RFA. We anticipate that many of the decisions on scientific direction will be made in those forums. All decisions made will be approved by the AMC (see *Decision Making and Conflict Resolution*).

Budgets. Drs. Milosavljevic, Rozowsky, and Seth each maintain sole authority and responsibility over the budgets for their components of the grant, and can rebudget within and between budget categories to meet unanticipated needs. At any time during the award period the AMC can request additional funding from NHGRI for a current budget period to meet increased costs for specific tasks that are within the scope of the work but were unforeseen when the application was submitted. In such a scenario, the AMC will present a case to the leadership team and the budget will be modified once the leadership team is in agreement and has received

NIH/NHGRI's prior approval, as required. Progress, tasking, and scheduling will be reviewed during the weekly AMC calls. AMC will evaluate current and estimated spending monthly. Any budget changes resulting from NIH/NHGRI-initiated actions post-award will be reviewed by the AMC, and all further decisions will be made by the AMC. In any scenario, the leadership team will adhere to NIH policies and seek approval from NIH when required.

AMC Decision Making and Conflict Resolution. The AMC developed the proposed Research Plan and Governance Plan. We therefore do not foresee any disagreements among the three institutions that would negatively affect the proposed research strategy. However, should an unforeseen conflict arise, the AMC expects to reach consensus through constructive discussion and carefully considering the pros and cons of specific actions towards the AnVIL's goals. Conflict resolution will take place at the regularly scheduled AMC progress meetings, or for immediate matters, the AMC will meet by ad hoc teleconference to resolve urgent conflicts. If the unlikely scenario that AMC cannot reach a consensus and resolve an issue, at the request of any AMC member, the NHGRI Program Officer in charge of the project will be consulted for advice.

Roles and Responsibilities. The AMC will be responsible for execution and implementation of the project, ensuring timely achievement of the milestones. Since AMC members belong to different institutions, each will ensure that systems are in place to guarantee institutional compliance with US law and DHHS and NIH policies, including regulations on human subjects research, data, and facilities. The entire leadership team will participate in drafting and finalizing reports to the NHGRI on project status. Any discussions with the NIH/NHGRI will be led by Dr. Milosavljevic; however, any and all actions must be approved by the AMC (see *Decision Making and Conflict Resolution*). Each AMC member will lead their respective institution's effort while coordinating with the other two institutions to align timelines and evaluate progress through weekly teleconferences. Weekly cross-team meetings will be held to update all participants of Team ARC and collect feedback on ongoing work.

Dr. Milosavljevic (BCM)(PI) will be responsible for overall coordination, administrative, and fiscal management of the program and for communicating with NHGRI program officials in consultation with Drs. Seth and Rozowsky. Drs. Seth (SBG) and Rozowsky (Yale) will be responsible for fiscal, administrative and scientific oversight of all the activities to be performed by their respective groups in alignment with the overall ARC vision as proposed in the specific aims. Drs. Rozowsky, Milosavljevic, and Seth will appoint members from the Yale, BCM, and Seven Bridges team, respectively, who will support the community engagement, scientific development, and scientific communication strategy for the AnVIL.

Intellectual property and Publication. The AMC will comply with NIH policies and expectations on resource sharing for data, tools, and algorithms. Any intellectual property developed prior to or independently of the AnVIL will belong solely to the developer. The intellectual property developed by one party as part of the AnVIL will belong solely to that party, subject to any applicable provisions of the resulting AnVIL agreement. AMC will be responsible for generating publications regarding the AnVIL infrastructure and services and are expected to equally contribute to the effort. Publication authorship will be based on the relative scientific contributions of the PD/PIs, key personnel, and staff. AMC members will be authors on any publications resulting from this project as long as they remain on the grant.

D. DATA/SOFTWARE TRANSITION

Transition Plan for Transferring AnVIL-associated Data and Software

Seven Bridges will ensure that all data and software resources provided by the AnVIL remain available without interruption to the research community. All hosted data generated by the AnVIL will be stored using secure cloud resources, such as encrypted AWS S3 buckets. Upon the termination or expiration of this cooperative agreement, ARC will work with any designated NIH-funded and/or managed resource projects to provide access to the hosted data. This process may include, for example, establishing a process to copy the data to a separate public cloud, or transferring data bucket ownership to the appropriate parties. Any open source software built as part of the AnVIL will be made available on a publicly accessible website (e.g., the Seven Bridges Github at https://github.com/sbg). Seven Bridges will additionally provide any support needed regarding installation and usage of software developed as part of the AnVIL.

E. PROPOSED TIMELINES AND MILESTONES

Tasks
Deliverables

Proposed timeline and milestones for the AnVIL	Y1				Y2				Y3				Y4				Y5			
	2018		2019				2020				2021				2022				2023	
Task	Q 1	Q 2	Q 3	Q 4	Q 5	Q 6	Q 7	Q 8	Q 9	Q 10	Q 11	Q 12	Q 13	Q 14	Q 15	Q 16	Q 17	Q 18	Q 19	Q 20
Specific Aim 1: Build cloud-agnostic AnVIL platform.																				
Develop comprehensive platform specifications																				
Collect requirements: NHGRI, EAC and DSC																				
Prioritize datasets, tools, workflows for AnVIL beta																				
Present specifications to EAC																				
Develop cloud-based infrastructure																				
Deploy cloud-agnostic SB-CI on AWS																				
Beta release of AnVIL to select users (NHGRI, EAC, DSC)																				
Production release of AnVIL platform to research community																				
Establish shared analysis, visualization, and compute env.																				
Deploy admin panel for user administration																				
Demonstrate collaborative compute workspaces																				
Deploy visualization SDK																				
Demonstrate data management system																				
Demonstrate third party publishing																				
Cloud services cost control																				
Develop the price calculator																				
Demonstrate data use monitoring for archiving decisions																				
Demonstrate billing system (compute & storage cost)																				
Data access and data security																				
Demonstrate dbGaP authentication integration																				
Establish user admin functions, provenance, audits																				
Perform third party audits on data security processes																				
Specific Aim 2: Provide data, tools and FAIRness.																				
Develop metadata ontologies/dictionaries for dataset onboarding																				
Establish AnVIL interoperability with Data Commons ecosystem																				
Implement FAIR principles																				
Development of FAIR genomic knowledge																				
Deployment of updates messaging system																				
Genomic datasets, phenotypes and metadata																				
Onboard beta test datasets																				
Onboard prioritized production datasets																				
Deploy prioritized tools and workflows																				
Deployment of exploratory workflows																				
Deployment of UI-centric applications																				
Specific Aim 3: Drive AnVIL adoption.																				
Years 1-5: Plan annual AnVIL meeting																				
Outreach: email, Twitter, Facebook, LinkedIn, webinar series																				
Training workshops and hackathons (+ surveys)																				
Governance: ARC management reports to NHGRI																				