CONCEPT CLEARANCE FOR RFP/BAA (Contract)

National Advisory Council for Human Genome Research

February 6, 2017

The NHGRI Sandbox – A resource for Genomic Data Sharing and Analysis

Purpose

This new concept is to establish the NHGRI Sandbox for sharing genomic data and analysis resources in support of genomic research. A data "sandbox" co-locates data, storage and computing infrastructure with commonly used services and tools for analyzing and sharing data, to create an interoperable resource for the research community (RL Grossman et al, 2016). The NHGRI Sandbox will leverage a cloud-based infrastructure to democratize genomic data access by the broad scientific community, and facilitate integration and computing on and across large datasets generated by NHGRI programs. It will also provide a data exchange and computing platform for NHGRI consortium members.

Background

Decreasing sequencing costs have led to increasing amounts of genomic data to manage and analyze, bringing new challenges for researchers who lack high-performance computing infrastructure, technical expertise and resources in their labs and research institutions. Cloud-based storage and computing resources can alleviate those challenges by offering a scalable infrastructure to address storage and computing needs dynamically and in real-time, reducing the data management demands on local data centers and system administration resources. A cloud-based computing environment allows users to upload and run their own analysis tools, or utilize state-of-the-art, optimized-for-the-cloud genomic data analysis workflows and visualization tools where the data reside, therefore decreasing the need to download data to local systems, reducing the potential for data security incidents and network traffic.

Although the Sandbox resource will be available to any genomic researcher, whether funded by NHGRI or not, we anticipate that the users of the Sandbox will be primarily two types: the bioinformaticians and computational biologists who are familiar with cloud or high-performance computing, and researchers who do not have extensive coding experience and may want to use simpler tools and web interfaces provided by the Sandbox. Investigators of NHGRI research consortia are also expected to take advantage of the data and analysis tool sharing services of the Sandbox, not only within the consortium, but also for direct and rapid release to the scientific community of the resources generated, bypassing the time consuming submission to centralized data repositories (such as dbGaP and SRA) for public data redistribution.

NHGRI hosted an Informatics and Data Science workshop on Sept 29-30, 2016 to discuss how to strengthen the NHGRI computational genomics and data science program portfolio. The workshop was attended by approximately fiftyfive extramural scientists and a report is in preparation about the outcomes and the recommendations from the workshop. The establishment of a cloud-based resource that can be used by NHGRI funded programs to collect and share genomic data and analysis tools was recommended as a high priority for the Institute. The replication of data on multiple cloud service providers to offer choices and

promote competition was encouraged, as well as interoperability and coordination with ongoing genomic data commons projects at the NIH (see "Relationship to Ongoing Activities" below). Ultimately, this project should be considered as a pilot to explore and address the challenges of transitioning from the traditional centralized genomic data archival systems to distributed and federated data commons systems to facilitate genomic data access, sharing and computing at scale.

Proposed Scope and Objectives
The NHGRI Sandbox will provide:

- A cloud-based scalable storage and computing system for genomic data (including genome sequence data, transcriptomic, epigenomic and other data produced with sequencing technologies) generated by NHGRI programs and by other programs of interest to the genomic research community.
- Application Programming Interfaces (APIs) and interoperability with other data commons, such as the NCI Genomic Data Commons and the NIH Data Commons.
- Web interfaces and API access to datasets, analysis tools and workflows commonly used by the genomic research community.
- A shared analysis environment, such as a data exchange area for funded consortia, where data and analysis workflows are shared with consortium members, and data submission services to central data repositories (such as dbGaP, SRA or GEO) for long-term storage.
- A workspace where individual researchers can upload their own datasets and analysis tools for combined analyses of private and public data.
- Implementation and availability of commonly used genomic analysis workflows (e.g. for variant calling, RNA-Seq analysis pipelines) and sharing of their results.
- Co-location of NHGRI funded datasets and their metadata with other important datasets for genomic research, and metadata harmonization across funded programs.
- User authentication and authorization mechanisms and implementation of data security practices, as required by US Government policies.
- Auditing tools on the use of datasets, analysis tools and workflows to help determine when software services should be decommissioned and data should be moved from high performance storage to cost-effective archival systems.

Data access: The Sandbox must be able to provide access to both unrestricted and controlled access genomic data. In order to do so, the Sandbox will be required to establish a Trusted Partnership with dbGaP and use FISMA-Moderate, HIPAA compliant cloud service providers (private or commercial). It will also have to implement the requirements specified in the "NIH Security Best Practices for Controlled Access Data Subject to the NIH Genomic Data Sharing Policy". An investigator who wishes to use controlled access data in the Sandbox will have to undergo the already existing user authentication and authorization processes established for dbGaP users, including the eRA Commons user authentication process, and will have to submit a Data Access Request (DAR) for review and approval by the NHGRI Data Access Committee (DAC).

The Sandbox will also provide a data access, sharing and computing environment to NHGRI research consortia. Consortium members will be able to analyze consortium data, and share analysis tools and results with collaborators. For consortium investigators the Sandbox's user authentication and authorization processes will follow the data access and sharing requirements established by the consortium and NHGRI program staff.

Datasets: Datasets included in the Sandbox, their data types and file formats (e.g. BAM/CRAM files, VCF, phenotypic data, gene expression data, peak calls, signal tracks) will be determined by NHGRI extramural program staff in collaboration with investigators of the funded programs and other extramural investigators. A few datasets have been identified to start populating the Sandbox based on their importance and broad utilization by the genomic research community: 1000 Genomes, eMERGE, GTEx and ENCODE. Also, it is anticipated that the data from the Centers for Common Disease Genomics and the Centers for Mendelian Genomics will be provided via the Sandbox.

Relationship to Ongoing Activities

Some NIH ICs with extensive genomic research programs have established genomic data commons resources similar to the proposed NHGRI Data Sandbox. For example, the NCI Genomic Data Commons (GDC) and Cloud Pilots provide the cancer research community with data repositories that are unified in their ability to enable data sharing, integration, analysis and standardization of genomic and clinical data across cancer genomic studies. They also provide platforms to query and download datasets, APIs for secure data submission and data access, and analysis workflows that can be shared with collaborators. The NHLBI Trans-Omics for Precision Medicine (TOPMed) Program will build a data commons repository that collects WGS data, RNA-Seq, methylation, metabolomics, other "omics" and clinical outcomes data from NHLBI-funded studies, making the data available for analysis. The Precision Medicine Initiative (PMI) *All of Us Research Program* includes a Data and Research Center that will store and curate PMI data in a cloud-environment and make the data accessible for researchers through a dedicated analysis platform.

The risk of having non-interoperable IC-specific or initiative-specific data siloes should be alleviated by establishing collaborations with the emerging NIH Commons, which is a shared virtual ecosystem for digital objects that will follow the FAIR (Findable, Accessible, Interoperable, Reusable) principles. Both the NCI and NHLBI TopMed data commons are expected to be members of the NIH Commons, and the NHGRI Sandbox will be required to participate in this federated data system, and follow the NIH Commons compliance requirements that are currently being developed. Queries, analysis workflows and APIs that work across federated cloud-based systems are under development for retrieving and computing on data in different systems.

Funding Mechanism

To satisfy the NIH requirements for Trusted Partners of dbGaP, the Sandbox will be funded by a contract for seven years.

A contract funding mechanism also allows for the formulation of specific deliverables, reporting requirements, and close oversight by NHGRI staff. NHGRI staff will work cooperatively with the Sandbox's staff to decide on the datasets and analysis workflows to be deployed and supported, to develop metrics of usage of the resource, to manage user accounts, and to perform other activities as needed.

To promote responsible usage of the Sandbox and its cloud-based resources we anticipate that users will be responsible to pay for computing costs, storage of their own data and data egress.