National Advisory Council for Human Genome Research
February 12, 2024
Concept Clearance for RFA

Enhancing Reuse of NHGRI Data Assets (R03)

Purpose

The purpose of this concept is to encourage the genomics research community to leverage data sets, analysis tools and other resources available through NHGRI's Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL) platform for innovative secondary analysis projects. AnVIL makes available data, tools and resources generated primarily by NHGRI-funded programs and initiatives. Research projects funded under this RFA will augment the value and impact of these resources for advancing biomedical research and increase the effectiveness of AnVIL in catalyzing scientific discoveries.

Background

The NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space is a cloud-based genomic data sharing and analysis platform. As stated in NOT-HG-19-024, AnVIL is the designated data repository for datasets generated by NHGRI programs, but it also hosts data from initiatives funded by other Institutes at the National Institutes of Health (NIH), or by other agencies that support human genomics research. AnVIL facilitates sharing, integration, and computing on both unrestricted and controlled access datasets. By providing a unified environment and workspaces for data management, access and computing, AnVIL alleviates the need for downloading or transferring large datasets across systems, provides a FISMA Moderate system that allows for data security, threat detection and data access audits, and offers elastic, shared computing resources that can be utilized by researchers on demand. AnVIL also provides a variety of training materials, including videos and featured workspaces to easily reimplement published analysis workflows, and plays a crucial role in democratizing data access, particularly for researchers lacking local infrastructure for data storage and computation.

Since its inception, AnVIL has ingested over ten petabytes of data, with more than half of this already accessible by researchers, and the rest becoming available in the future. With the new NIH Data Management and Sharing (DMS) policy effective since January 2023, the volume of NHGRI's data assets housed on AnVIL will increase even more rapidly. Given the richness and complexity of these data assets, both the AnVIL External Consultant Committee (ECC) and participants of the AnVIL workshop hosted by NHGRI in October 2021 recommended exploring avenues to actively engage the broader genomics community in leveraging the AnVIL platform's resources This initiative hence aims to foster innovative reuse of these data assets and other resources to catalyze their utilization for secondary research purposes.

Proposed Scope and Objectives:

This concept proposes a single RFA using the NIH Small Research Grant Program (R03) mechanism to support projects that propose secondary research use of AnVIL-hosted data (listed at https://anvilproject.org/data) and analysis resources in AnVIL workspaces. Contingent upon including at least one existing AnVIL data set, research proposals may utilize data uploaded by the applicant to AnVIL workspaces. Examples of secondary research uses that are in scope under this RFA include, but are not limited to:

i. Utilization of individual datasets for scientific research projects that are different in scope from the original research scope of the data submitters.
ii. Studies analyzing the combination of two or more AnVIL datasets.
iii. Analyses combining AnVIL-hosted data and tools with data and tools available through other NIH and non-NIH cloud platforms and other resources.

Each R03 award under the proposed RFA will be for a two-year non-renewable project period with a budget up to $120,000 total costs/year, including cloud costs. Applicants will be expected to utilize cloud cost control approaches and will be encouraged, but not required to use the NIH STRIDES initiative. Although the R03 funding mechanism is specifically to support small research grants, projects supported under this initiative may produce preliminary results for larger research projects grants and catalyze additional scientific discoveries.

Relationship to Ongoing Activities

This initiative complements the AnVIL program and amplifies the value of the data, analysis tools and other resources made available by NHGRI- and NIH-funded programs and initiatives, such as the Center for Common Disease Genomics (CCDG), Center for Mendelian Genomics (CMG), the Electronic Medical Records and Genomics (eMERGE) Network, the Genomics Research to Elucidate the Genetics of Rare Diseases (GREGoR) consortium, and the Genotype-Tissue Expression Project (GTEx). Award recipients will also provide valuable feedback to NHGRI on the utility of the data and analysis resources provided by AnVIL. In a broader context, this initiative is part of a trend of funding opportunities that aim to catalyze scientific discovery and maximize the value of NIH data investments by promoting secondary research on large datasets that are available through NIH projects and cloud platforms (e.g., RFA-RM-23-015 and PAR-23-075).

Mechanism of Support/Funds Anticipated

- R03 awards will be capped at $120K total costs/year for two years and will be non-renewable.
- NHGRI plans to make 6-8 awards per year in FY25, FY26, and FY27, with the following estimated budget: $720K-$960K in FY25; $1,440K-$1.920K in FY26; $1,440K-$1.920K in FY27; $720K-$960K in FY28.