National Advisory Council for Human Genome Research (NACHGR)
February 7, 2022
Concept Clearance for RFAs
Multi-Omics for Health and Disease

Purpose:

NHGRI proposes a collaborative initiative to advance the application of multi-omic technologies to study health and disease in diverse populations. By leveraging disease contexts where multi-omic approaches are expected to be most impactful, the proposed consortium will 1) explore the use of multi-omics, integrated with phenotypic and environmental exposure data, including social determinants of health (SDOH), to detect and assess molecular "profiles" associated with healthy and disease states; 2) leverage these association studies to develop generalizable data harmonization, integration, and analysis methods, as well as best practices and standards for the optimal application of multi-omics; and 3) create a multi-dimensional dataset that is available to the research community. While this program may provide some insights into disease etiology, its primary goal is to validate and enhance generalizable multi-omic approaches to identify meaningful biological changes related to health or disease.

Background:

Expansions in high-throughput technologies have increased access to molecular (or ‘omic) data generated at distinct levels within a biological network, such as genomic, epigenomic, transcriptomic, proteomic, and metabolomic levels. While single ‘omic analyses have produced valuable insights, recent studies have shown that integrative (or multi-omic) analysis approaches can improve the classification of disease into clinically relevant subgroups and potentially identify biomarkers of health or disease. Multi-omic analyses can also help define relationships among ‘omic data types to unravel biological networks regulating transitions from health to disease. For example, integrating genomics, transcriptomics, proteomics, and metabolomics from multiple anatomical locations was shown to improve the statistical power and accuracy of classification of chronic obstructive pulmonary disease cases[1]. In addition, the integration of transcriptomics, epigenomics and proteomics helped identify histone acetyltransferases as drivers of Alzheimer's disease[2].

Despite some successes such as the large datasets produced by NHLBI's TOPMed program, critical gaps prevent the routine application of multi-omics to disease studies. Production and standardization of multiple data types from the same sample remain key challenges, including inter and intra ‘ome variability, non-uniform content across platforms and assays, and lack of consensus quality assessment approaches and imputation practices. Computational methods to integrate, analyze, and interpret multiple ‘omes from the same sample, multi-omic data combined with phenotypic and environmental exposure data, and multi-modal data across diverse populations remain underdeveloped. Importantly, lack of prospective data collections adhering to agreed upon data standards has limited the number of widely available multi-omic datasets that have both well-described and harmonized metadata and multiple types of ‘omic data.

Through a prospective study design and collaborative collection and analysis of multiple ‘omic data in conditions where disease transitions and/or disease subtypes can be defined, this initiative is expected to produce consensus approaches, best practices, and standards that can be generalized across diseases and populations. It will also generate a standardized and harmonized dataset for general research use available through controlled-access processes as well as a portal for visualization. Ultimately, this program will enhance the utility of ‘omic technologies in understanding the biology of health and disease, setting the stage for further research into their future application in the clinic. This program aligns with recommendations offered by participants in the June 2021 NHGRI workshop on the topic.

Proposed Scope and Objectives:
This initiative will support a consortium to implement multi-omic technologies following the establishment of consensus approaches for the study design, participant recruitment, data production, and data analyses consistent with the state of the science and tailored to the diseases proposed for investigation. The consortium will comprise:

4-5 Disease Study Sites (DSS): Each site should propose a study focused on a disease area for which integrative multi-omics could be useful in elucidating disease etiology and course, such as: 1) relapsing diseases with difficult-to-predict exacerbations and remissions, 2) heterogeneous diseases with clinically important subtypes relevant to prognosis or treatment, or 3) diseases with distinct stages or transitions. Each DSS should enroll 150-200 participants experiencing disease and 75-100 participants without disease, similar in key demographics and exposures as the participants with disease, to be contributed to a pooled, consortium-wide comparator group. Participant phenotypes and environmental exposures, including SDOH, as relevant to the disease under study, should be well-defined using standard measures when available (e.g. PhenX Toolkit). Recurring assessments of all participants are expected, with measures collected at a minimum of three time points such as those corresponding to baseline (study entry), exacerbations, remissions, or treatments.

All study participants should give consent for broad data sharing and collection of multiple measures from multiple ‘omes taken at multiple time points. Re-consent of individuals participating in existing longitudinal studies for prospective data collection and sharing may be considered with a clear description of how they meet the requirements of this program.

A robust body of research has identified the scientific and ethical challenges with the overrepresentation of European ancestry groups in research, some of which include undiscovered genetic variation, inaccurate risk prediction tools, and inequity in the distribution of benefits from research[3,4]. To increase the diversity of genetic ancestries represented in this project, each DSS should aim to recruit a minimum of 75% individuals from ancestral backgrounds currently underrepresented in genomic research. To help maximize benefits from participation in this research, applicants are expected to establish recruitment, retention, and meaningful community engagement strategies, including outreach to racial and ethnic minority communities in the United States.

1-2 ‘Omics Production Centers (OPC): The OPC(s) will utilize state-of-the art high-throughput molecular assays to produce ‘omic data at multiple timepoints from samples (including tissues and cells, as needed) provided by participants enrolled by the DSSs. The OPCs should aim to produce: 1) genomics (WGS), 2) epigenomics (methylation arrays, ATAC-Seq, or ChIP-Seq), 3) transcriptomics (bulk RNA-Seq), 4) proteomics (targeted or untargeted; mass-spectrometry (MS), SOMAmer), and 5) metabolomics (targeted or untargeted; NMR spectroscopy, GC-MS, LC-MS). Other data types and assays will be considered with justification. While applicants proposing to produce high-quality data for all ‘omic data types are encouraged, applicants must propose a minimum of three data types, one of which is non-nucleic acid-based (proteomics, metabolomics, etc.).

1 Data Analysis and Coordination Center (DACC): All demographic, phenotypic, environmental exposure, and molecular data related to the participants enrolled by the DSSs and processed at the OPCs will be collected, catalogued, and maintained by the DACC. The DACC will be responsible for coordinating all consortium-wide activities, including: 1) building consensus on participant recruitment strategies and the choice of ‘omic data types and assays for each study; 2) developing the consortium-wide data analysis process; 3) liaising with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL) to facilitate data sharing; 4) establishing working groups to develop methods, best practices, and standards; 5) producing the standardized and harmonized multi-dimensional data set; 6) developing the data portal to visualize and make the data, metadata, and software open, findable, accessible, interoperable, and reusable (FAIR); and 7) rapidly disseminating consortium outputs to the broader scientific community.

The 2020 NHGRI strategic vision emphasizes that the promise of genomics can't be fully realized without a diverse genomics workforce. In order to enhance the excellence and inclusivity of the research environment, applicants are strongly encouraged to assemble study teams from diverse backgrounds, including individuals from underrepresented groups[5-6]. All applicants are also expected to demonstrate experience and expertise in multi-omics high throughput assays and in computational and statistical data integration and data analysis methods. The DSSs and DACC should also have expertise in participant recruitment approaches and community engagement. The first year of the program will be used to develop network-wide protocols for key aspects of the program, such as 1) recruitment strategy, including the characteristics of disease participants and pooled comparison group, 2) community engagement plan; 3) core phenotypic and environmental exposure measures; 4) types of assays; and 5) methods for biospecimen procurement, processing, and analysis.

All awardees are expected to contribute to protocol development, data integration, and analysis through a Steering Committee. The consortium will apply computational modeling and machine learning approaches, interpret molecular "profile" associations, explore gene networks, and assess causal relationships. It will also develop generalizable methods, best practices, and standards that address data harmonization, integration, and analysis challenges and gaps. Demonstration of the generalizability of these methods across different diseases and populations is expected. The consortium will use the resulting data to create a standardized and harmonized multi-dimensional data set that is freely available to the research community and adheres to NIH genomic data sharing expectations. This rich data set will include 1) persons of diverse backgrounds, including those from groups that are underrepresented in biomedical research, 2) participants with disease, 3) participants without disease, 4) harmonized and standardized data for all or most 'omes for each sample, 5) data from all time points, and 6) associated meta-data to facilitate links across data types. The consortium will also create a visualization and access portal that follows FAIR principles and is interoperable with resources such as HuBMAP, GTEx, and TOPMed.

Relationship to Ongoing Activities:
This initiative will complement NIH investments such as those funded by NHLBI (TOPMed), NCI (TCGA), NIA, NIMH, and the Common Fund (HuBMAP). This program is distinct from other efforts in its prospective enrollment and study design, collection of specimens at multiple time points, production of major 'omic data types from the same sample, and requirements for consent for future use and broad data sharing without data use limitations. Furthermore, NHGRI will contribute to and enhance the state of the science more broadly by emphasizing generalizable approaches and standards.

Mechanism of Support:
Three RFAs will solicit: 1) 4-5 DSS), 2) 1-2 OPCs, and 3) 1 DACC. The U01 (Research Project--Cooperative Agreements) activity code will be used to facilitate the alignment of consortium progress and priorities with those of NHGRI.

Funds Anticipated (Total Costs):
NHGRI will commit approximately $39M over 5 years from FY23-FY27 to these RFAs.