National Advisory Council for Human Genome Research

September 16, 2019

Concept Clearance for FOA

Collaborative Data Integration and Analysis of Polygenic Risk Scores (PRS) from Populations of Diverse Ancestry

Purpose

NHGRI proposes a new initiative to collaboratively generate and refine PRS for populations of diverse ancestry by integrating existing datasets with genomic and phenotype data for a range of complex diseases and traits. The primary goals are to: 1) leverage genetic diversity to improve the applicability of PRS across diverse populations and for a broad range of health and disease measures; and 2) optimize the integration of large-scale, harmonized genomic and phenotype data to facilitate collaborative analysis, dissemination of PRS-related data, and development of related resources.

Background

PRS developed using large-scale genomic data from epidemiological studies are rapidly becoming linked to clinical implications, such as identifying individuals who would particularly benefit from modification of coronary heart disease risk factors. Recognizing the timeliness of this topic, NHGRI recently released Funding Opportunity Announcements for the Electronic Medical Records and Genomics (eMERGE) Network to support a clinical implementation study of genomic risk assessment and management, based on validated (and preferably peer-reviewed) PRS. Currently available scores, however, show poorer risk prediction in non-European populations due to vast under-representation of non-European ancestry (EA) populations in the underlying GWAS data. Methodological approaches cannot reliably eliminate the biases due to underrepresenting non-EA data, underscoring the importance of increasing the genetic diversity of the data on which PRS prediction is based.

This initiative proposes to narrow the gap between the generation and use of PRS in EA and non-EA individuals by generating and refining PRS that are more applicable to diverse populations. Specifically, this initiative proposes analysis and integration of extant GWAS datasets from genetically diverse populations, which has the potential to improve the clinical value of PRS in non-EA populations.

Several characteristics of PRS analysis are particularly amenable to collaboration. The rapid translation of findings to the clinical setting incentivizes data sharing and the development of approaches to facilitate data integration. Additionally, synergy is likely to be gained in studying multiple phenotypes with a similar approach. Also, summary statistics can facilitate integration of data that may not be feasible to share at the individual level. Though not the primary goal of this program, the approaches to data integration, phenotype harmonization and collaborative analyses developed in this consortium may serve as an example for related studies of complex disease genomics.

The scientific challenges and opportunities to be addressed by this consortium include:

1. Standardizing genomic and phenotype data, and mapping to existing ontologies.
2. Developing consensus approaches to incorporate ancestry information into PRS.
3. Developing PRS for diverse populations and refining PRS to improve risk prediction.
4. Optimizing the integration of data for PRS analyses according to FAIR (findable, accessible, interoperable, reusable) practices and policies. Data sharing practices will

be developed to enable analyses across the consortium; data sharing and outreach to the scientific community is also expected.
5. Disseminating PRS results publicly in the form of summary statistics datasets, algorithms, publications, and related resources.
6. Validating PRS in clinical settings by engaging ongoing genomic medicine efforts.
7. Identifying secondary uses of the data related to health and disease research and making data available for such uses, consistent with participants' consent.

Proposed scope and objectives

The above research opportunities would be addressed by a PRS consortium comprising 3-5 PRS Centers (PRSC) and 1 Coordinating Center (CC).

A PRSC would consist of a group of investigators representing one or more participating cohorts that would provide extant genomic and phenotype data from these cohorts and the necessary scientific expertise to analyze the data in collaboration with consortium members. A minimum threshold for genetic diversity would ensure inclusion of large numbers of diverse participants across the consortium. For each PRSC, aggregate and extant data from all cohorts participating in that PRSC must meet one of two criteria: 1) include up to 20,000 genotyped and phenotyped participants, with at least one non-EA group with a minimum of 10,000 genotyped and phenotyped participants; or 2) include at least 20,000 genotyped and phenotyped participants, with at least 50% of participants from non-EA populations. Non-EA populations include those designated as minority ethnic or racial groups according to the NIH policy on Inclusion, including Hispanic or Latino, American Indian or Alaska Native, Asian, Black or African American, or Native Hawaiian or Other Pacific Islander. Applications meeting these minimum criteria would further be given priority according to the following criteria: a) >50,000 genotyped and phenotyped participants across all participating cohorts; b) large numbers (≥10,000) of non-EA participants; c) broad phenotype information (multiple health and disease measures) available to be standardized and shared within the consortium; and d) commitment to data sharing of individual-level data both within and outside the consortium.

The CC would be responsible for overall logistical coordination and data science leadership, working closely with the PRSCs. The data science will be driven by the need to develop and implement efficient approaches to data harmonization, integration and analysis that adhere to FAIR principles. Additionally, the CC would develop approaches to disseminate PRS and related information to the broader scientific community and work with the PRSCs to ensure timely and user-friendly dissemination. The CC would also work with the NHGRI Genomic Data Science Analysis, Visualization and Informatics Lab-space (AnVIL), a scalable and interoperable resource leveraging a cloud-based infrastructure, to facilitate data sharing and analysis within the consortium. Ethical, legal and social implications (ELSI) are inherent to integrating heterogeneous datasets and generating PRS data which may differentially impact individuals of diverse ancestry. The CC would also provide and convene ELSI expertise to address these and other ELSI issues that arise during the conduct of this research. Recognizing that the funded PRSCs may not capture all available diverse cohorts, the CC would also identify and invite researchers representing high priority cohorts to participate as affiliate members and provide limited genotyping and analysis support for them if needed.

The details of data sharing within the consortium and with the scientific community would be established through consortium-wide data sharing agreements. Where possible, data sharing would be facilitated by the AnVIL. At a minimum, the data to be shared with the scientific community would include: the genetic variants contributing to the PRS; other covariates used to derive the PRS; aggregate sample sizes and ancestry information, and

other information needed to apply the consortium-derived PRS to an external dataset. To maximize synergy with ongoing efforts, the consortium would adopt and adapt existing standards for data sharing, genomic and phenotype harmonization, and other data standards as feasible, working with organizations such as the Global Alliance for Genomics and Health, with phenotype standardization efforts such as the Human Phenotype Ontology and Monarch, and with data models such as the Observational Medical Outcomes Partnership.

In years 1-2, funded PRSCs would work together and with the CC, AnVIL, NHGRI, and other resources to integrate genomic and phenotype data for collaborative analyses. Investigators would maximize the sample size and genetic diversity of available data, working to address challenges presented by potential differences across cohorts in availability of clinical data, data use limitations, informed consent, and availability of summary statistics vs. individual level data. The consortium would agree upon a set of health and disease measures to analyze; harmonize health and disease measures, harmonize genomic information, including imputation as needed; select common metrics for refining and improving PRS-based prediction, and develop ways to integrate ancestry measures into the analysis. In years 2-4 investigators would analyze genomic and phenotype data across the consortium to generate and refine PRS models related to multiple health and disease measures and update these PRS models with new data. In years 4-5, investigators would continue to disseminate results and refine PRS models based on community input.

Information on the performance of the PRS developed and refined by the consortium, as well as the scores themselves, would be useful for clinical implementation efforts that are not directly funded by this initiative. Regular outreach is planned throughout the project period to efforts such as eMERGE to ensure that the work of this initiative has a high likelihood of adoption by clinical implementation studies.

Relationship to ongoing activities

This initiative would build on existing efforts by expanding access to genetically diverse datasets and facilitating consensus-based analyses of PRS that are applicable to diverse populations. It differs from eMERGE, which focuses primarily on clinical implementation of existing PRS, by generating and refining PRS scores based on non-EA genetic data that have not yet been systematically analyzed for this purpose. It differs from the All of Us Research Program, whose resource building has minimal emphasis on data analyses. Members of these and other large common disease consortia, such as the International 100K Cohort Consortium, International Common Disease Alliance, Centers for Common Disease Genomics, Population Architecture using Genomics and Epidemiology, Human Health and Heredity in Africa, and Trans-Omics for Precision Medicine, would be eligible to apply for this initiative. It is likely that new cohorts would arise during the course of this initiative, and these cohorts, as well as existing cohorts not funded through a PRSC, would be invited to participate as affiliate members.

Mechanism of support

Two RFAs are proposed, one to solicit 3-5 PRSCs and the second to solicit 1 CC. Both would use the U01 (Research Project--Cooperative Agreements) activity code. The cooperative agreement mechanism would facilitate the alignment of consortium progress and priorities with those of NHGRI in this rapidly-moving scientific area.

Funds anticipated

NHGRI would commit approximately $33.5M over 5 years from FY21-FY25 to these RFAs to support 3-5 PRSCs and the CC.