National Advisory Council for Human Genome Research
February 10, 2020
Concept Clearance: Consortium for Understanding the Impact of Genomic Variation on Genome Function

Purpose:

Understanding how genomic variation affects genome function to influence phenotype remains one of the central challenges in biology. To address this challenge, NHGRI proposes a major new program that will leverage emerging genomic technologies to (1) systematically perturb the genome and assess the impact of genomic variation on genome function and phenotype, (2) identify where and when genes and regulatory elements function, with increasing resolution, (3) advance network-level understanding of the influence of genome function on phenotype, (4) develop and test innovative predictive models of the impact of genomic variation on genome function, and (5) generate a resource of data, tools, and models to advance future community investigation of this important area of biology.

Background:

Research efforts such as NHGRI's Genome Sequencing Program, the 1000 Genomes Project, Genome-Wide Association Studies, and NHGRI's Genome Technology Program have led to key breakthroughs in our understanding of genomic variation and the role it plays in health and disease. Genomic analyses of populations or individuals to identify disease-associated genomic variants is now routine, but the greatest challenges and opportunities lie in defining the role of genomic variants in influencing phenotype, including disease (for example, identifying the specific causal variants that determine phenotype, identifying the cell types that are sites of action for particular variants' functions, identifying the genes that are affected by particular variants, and measuring molecular effects such as changes in gene expression or predicted protein products). Establishing causal relationships between individual variants and disease risk is hampered by a lack of mechanistic understanding; similarly, understanding the clinical relevance of variants is stymied by the already overwhelming and ever-growing number of variants of unknown significance (VUSs). For these reasons, the development of resources, methods, and approaches that would help gain a more complete understanding of the relationship between genomic variants and phenotypes was identified as a priority area at the NHGRI Strategic Planning workshop "From Genome to Phenotype: Genomic Variation Identification, Association, and Function in Human Health and Disease."

The program proposed here will utilize emerging experimental and computational approaches to examine how the genome functions, how genome function shapes phenotype, and how these processes are influenced by genomic variation. Specifically, the program will involve a consortium of investigators involved in five components (Functional Characterization Centers, Mapping Centers, Regulatory Network Projects, Predictive Modeling Projects, and a Data Coordinating Center; see **Appendix 1, Figure 1** and additional details below). Synergies among the different components will be leveraged to generate a durable community resource of data, software, tools, methods, and standards that will be developed by consortium investigators.

This program builds upon and weaves together threads from years of genomics research into genome function and gene regulation (see **Figure 2**). Efforts such as the Encyclopedia of DNA Elements (ENCODE) and Roadmap Epigenomics Mapping Consortium (REMC) succeeded in the systematic identification of functional elements in the human genome. However, available technology largely limited these projects to profiling bulk tissue samples and/or cell lines. Functional elements are known to have specific effects based on cell type, state, and fate, and significant work remains to link the activity of specific functional elements to specific cell types. The most recent phase of ENCODE piloted the use of functional element characterization centers for refining newly-developed technologies and establishing best practices. Other groups have performed similar exploratory work for characterizing both coding and non-coding genomic regions. The NHGRI Genomics of Gene Regulation (GGR) project aimed to develop better methods to construct predictive, accurate gene-regulatory network models, while the goal of the NHGRI Non-Coding Variants (NoVa) program was to predict which variants (from a set known to be statistically associated with a phenotype) were most likely the relevant causal ones; however, additional work is needed to build on insights from those programs using the latest technology and data.

Indeed, there is now the opportunity to deploy these ever-improving approaches to systematically probe the effects of genomic variation on genome function and phenotype. These include strategies for perturbation of genomes and epigenomes at scale that enable the systematic study of the consequences of genomic variation. Emerging multi-omic mapping methods that profile genomes at single-cell resolution within intact tissues can relate functional elements and gene activity to cell- and tissue-level phenotypes. Although it is not feasible to experimentally probe all genomic variants of interest in all cell types of interest, advances in artificial intelligence and machine learning offer increasingly powerful means to model gene-regulatory networks and predict the impact of untested genomic variants and groups of variants.

The goals of this program are well-aligned with the emerging NHGRI strategic plan. One theme is the importance of functionally characterizing whole-genome sequences, transcriptomes, epi-transcriptomes, and epigenomes; this program will both identify and characterize candidate functional elements. Another theme is to establish the role(s) of all genes and regulatory elements in pathways, networks, and phenotypes; this program will test the interactions of candidate functional elements and model them within networks. Yet another theme is to establish the functional consequences of any genomic variant affecting human health and disease; this program will systematically characterize disease-associated variants and develop predictive models of function for yet-untested variants. The expected short-term outcomes from this program are summarized in **Appendix 2**.

Proposed scope and objectives:
This program will support a consortium of researchers that will aim to advance our understanding of the relationships among genome function, genomic variation, and phenotypic consequences. Research is expected to be done in mammalian systems such as cells, tissues/organs, organoids, or intact organisms with a preference for human studies that are properly consented for data sharing through open, unrestricted access in public databases. Specific objectives will be addressed by each of five components (see **Appendix 1**):

1) **Improve our understanding of the relationships between sequence variation and genome function by systematically collecting data on the effects of genomic variation on molecular, cellular, and organismal phenotypes.** Specifically, **Functional Characterization Centers** may:
    - Systematically apply existing state-of-the-art high-throughput genomic perturbation methods (for example, massively parallel reporter assays, genome editing, epigenome editing, high-throughput protein mutagenesis)
    - Assay impact of genome variation on molecular, cellular, or organismal phenotypes
    - Develop and refine novel or emerging generalizable high-throughput approaches for characterization of functional genomic elements
    - Test non-coding and/or protein-coding functional elements and variants
    - Investigate the interaction of genomic elements and variants, with a focus on cis effects.

2) **Improve the ability to determine where and when regulatory elements and genes are active at single-cell resolution.** Specifically, **Mapping Centers** may:
    - Establish multi-omics pipelines for high-throughput mapping of functional regions of the mammalian genome at single-cell resolution
    - Use state-of-the-art methods to identify biochemical marks (for example, transcription, open chromatin, and epigenomic marks) while preserving information about biological and/or spatial context
    - Apply mapping pipelines to identify candidate regulatory elements, cell types, and element-cell type associations in samples of high value to the consortium and research community.

3) **Enhance ability to define the role of specific sequences in regulating gene networks and establishing phenotypes.** Specifically, **Regulatory Network Projects** may:
    - Develop and apply systematic multi-omic data collection (examples include functional characterization and/or mapping technologies) to measure changes in the activity of genes and regulatory elements in biological or disease systems

- Employ assays with spatial and/or temporal resolution to capture information during biological changes (for example, change in cell fate or cell state)
- Develop and apply analytical approaches to identify gene regulatory networks using functional data from combinations of genomic assays
- Investigate the interactions of genomic functional elements and genomic variants in cis and trans
- Identify network-level relationships among genomic variants, functional elements, and phenotypes in specific systems and develop generalizable paradigms/approaches.

4) **Develop computational approaches to model and predict relationships among genomic variation, genome function, and phenotype.** Specifically, **Predictive Modeling Projects** may:
   - Develop, apply, and experimentally test predictive models. Examples include, but are not limited to: a) the impact of genomic variation on function and phenotype; b) the location and function of elements in specific cell and spatial contexts; c) interactions of genomic variants; d) systems-level effects (for example, a genetic effect acting on one gene could cause mis-regulation of another gene that is directly connected to phenotype)
   - Identify the highest-priority samples, approaches, and experiments for the consortium to study in order to conduct the most informative experiments
   - Create tools (for example, visualization tools) to enable inferences about genome function
   - Provide analytical expertise and support for consortium-wide efforts.

5) **Establish a resource to enable future studies of genomic variation, genome function, and phenotype by organizing data, methods, and software and distributing them to the research community.** Specifically, the **Data Coordination Center** may:
   - Provide community access to data, software, and resources from all components of this program as a resource of high utility and value to the research community
   - Make data available in a form ready for advanced machine-learning, artificial intelligence, and other computational approaches
   - Work with other consortia and international groups to facilitate data integration, joint analysis, and sharing of standards and best practices
   - Organize and facilitate consortium activities, including convening working groups and outreach activities.

All components will be expected to contribute to consortium-wide activities, including:

- Develop standards for data and metadata, and establish data quality metrics

- Contribute all data, metadata, analyses, software, and other products to the Data Coordination Center and appropriate repositories in appropriate formats
- Participate in the planning, implementation, and analysis of consortium-wide or component-wide projects, typically through working groups
- Share best practices and lessons learned
- Take part in annual meetings
- Contribute to outreach efforts.

Synergies are anticipated both within and across components through a managed, open consortium structure (see **Figure 1**).

- Managed consortium means NHGRI program staff will be actively involved in coordinating research and ensuring that policies are followed
- Open consortium means scientists that are not funded by this program could join as affiliate members, agree to abide by consortium rules, and work on consortium projects
- A consortium structure provides the opportunity for groups to learn from each other and improve their approaches
- Within-component coordination will ensure that the generated data maximally covers biological space, while avoiding unnecessary redundancy
- Across-components coordination will allow for the development of an experimental plan with better statistical validity and ensure that the data generated supports predictive modeling
- Coordination between the Data Coordination Center and the data collection groups will lead to interoperable metadata and data that are uniformly processed in a reproducible manner with portable pipelines that can be used by the broader research community, consistent with the expected outcomes outlined in **Appendix 2**.

Importantly, consortium management will be most flexible in the more research-oriented projects (for example, Regulatory Network Projects), allowing the investigators to retain scientific autonomy, while tighter management will be used for the data-production activities (for example, Functional Characterization Centers and Mapping Centers), data-sharing activities (for example, Data Coordinating Center), and cross-consortium activities.

Relationships to ongoing activities (see **Figure 3**):
NHGRI will continue to support research and technology development in functional genomics and genomic variation via the NIH parent R01, R21, and SBIR/STTR announcements, the Computational Genomics and Data Science announcements, Novel Genomic Technology Development, Variation, Function and Disease funding opportunities, and the Centers of Excellence in Genomic Science (CEGS) program.

This program will build upon advances made by previous and existing functional genomics consortia, including ENCODE, REMC, and the International Human Epigenome Consortium (IHEC). It will differ from these other efforts in (1) its increased emphasis on testing how genomic variation influences genome function and phenotype, (2) the inclusion of analyses of protein-coding regions, and (3) a single-cell emphasis.

This program will rely heavily on, and will also inform, single-cell and spatially-resolved multi-omics methods, similar to those used in the ongoing NIH Common Fund Human BioMolecular Atlas Program (HuBMAP) and Human Cell Atlas. In contrast to those programs, which focus on using single-cell technologies to identify and map different cell types within the context of whole tissues and organisms, this program will leverage single-cell multi-omic and spatio/temporal methods to probe genome function as it relates to genomic variation and biological activity.

This program will provide key data that will inform efforts to understand the role of genomic variants in human health and disease. Both rare and common variants identified as being associated with disease by NIH-funded projects (for example, NHGRI Centers for Mendelian Genomics and Genome Sequencing Program; Common Fund Gabriella Miller Kids First and Undiagnosed Diseases Network (UDN); and NHLBI Trans-Omics for Precision Medicine (TOPMed)) may be prioritized within this program (for example, experiments by the Functional Characterization Centers, Mapping Centers, and Regulatory Network Projects; analyses from Predictive Modeling Projects). The findings from this proposed program, if successful, could improve future experimental and/or analysis efforts for disease association and genomic medicine studies. Results from this program could be integrated as part of the evidence base in variant resources such as The Clinical Genome (ClinGen) Resource.

Mechanisms of support:

1) **Functional Characterization Centers (7-10 Awards) Max 15M TC/year**
   - UM1 (Research Project with Complex Structure - Cooperative Agreements)
   - Up to $1.5M DC/year; project period of 5 years
2) **Mapping Centers (3-5 Awards) Max 8M TC/year**
   - UM1 (Research Project with Complex Structure - Cooperative Agreements)
   - Up to $2M DC/year; project period of 5 years
3) **Regulatory Network Projects (5-7 Awards) Max 7M TC/year**
   - U01 (Research Project - Cooperative Agreements)
   - Up to $750K DC/year; project period of 5 years
4) **Predictive Modeling Projects (6-8 Awards) Max 5M TC/year**
   - U01 (Research Project - Cooperative Agreements)
   - Up to $500K DC/year; project period of 5 years
5) **Data Coordination Center (1 Award) Max 5M TC/year**
   - U24 (Resource-Related Research Projects--Cooperative Agreements)
   - Up to $4M DC/year; project period of 5 years

Funds anticipated: NHGRI intends to commit approximately $40M total costs per year for 5 years from Fiscal Years 2021-2025 for this program. The number of awards is contingent upon NIH appropriations and the submission of a sufficient number of meritorious applications.

Appendix 1. Overview of Program Components

Component	Objective	# Awards	TC/year
1) Functional Characterization Centers	To characterize the effects of genomic variation on genome function	7-10	$15M
2) Mapping Centers	To identify where and when regulatory elements and genes are active	3-5	$8M
3) Regulatory Network Projects	To explore the role of functional genomic elements in regulating networks and influencing phenotypes	5-7	$7M
4) Predictive Modeling Projects	To develop computational approaches to model and predict relationships among genomic variants, genome function, and phenotype	6-8	$5M
5) Data Coordinating Center	To enable future studies by organizing data, methods, and software and distributing them to the community	1	$5M

Appendix 2. Expected Outcomes

Data Resource:

- “AI-ready” data resource structured to enable use in artificial-intelligence and machine-learning approaches
- Database of all genomic variants tested, reporting the functional and phenotypic effects of variation in a rich metadata context (for example, assays, biological system, phenotypic tests, etc.), organized to enable community use in artificial-intelligence- and machine-learning-based efforts
- Catalogs of regulatory elements and gene expression with single-cell resolution in organs, tissues, and cell types relevant to disease, placing some genes/elements in regulatory networks and pathways
- Predicted effects of untested genomic variants through application of new computational approaches
- All raw and processed data and metadata (uniformly processed data when possible).

Tools, Models, Methods, and Standards:

- Computational approaches and software for modeling and predicting effects of genomic variants on genome function and phenotype
- Approaches to identify network/pathway information about genes and elements
- Models predicting effects of genomic variants on gene-regulatory network models
- Data standards, quality metrics, and protocols
- Methodological advances resulting from large-scale implementation of emerging technologies, potentially resulting in improved performance, increased throughput and efficiency, and decreased costs.

Figure 1 Synergies Within the Program

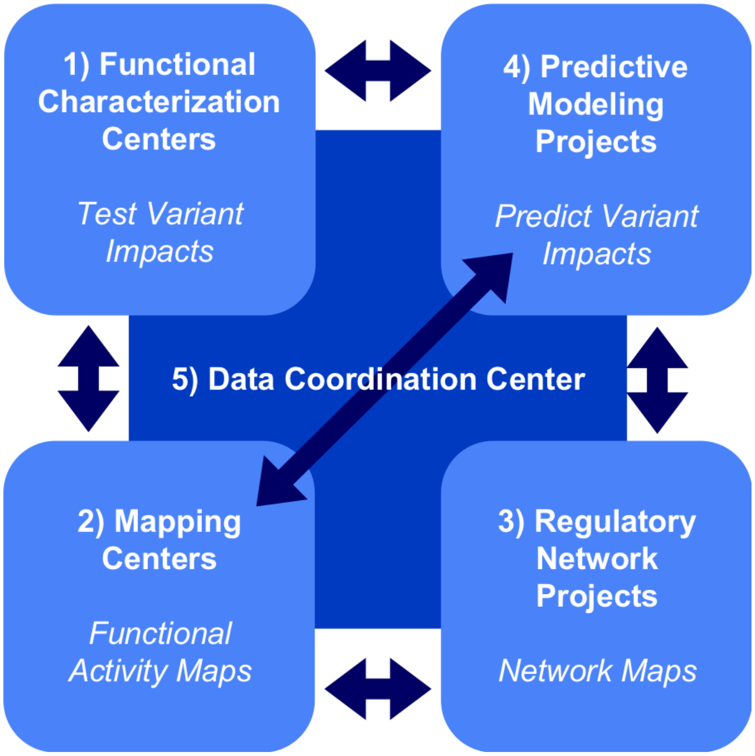
1) Functional Characterization Centers
Test Variant Impacts
4) Predictive Modeling Projects
Predict Variant Impacts
5) Data Coordination Center
2) Mapping Centers
Functional Activity Maps
3) Regulatory Network Projects
Network Maps

Figure 1: Synergies Within the Program.
Program components are represented as numbered boxes. Arrows indicate interactions. Data will flow to the coordinating center to be available to the broader research community. (Interactions within program components, though important, are not represented here.)

Figure 2 Foundational Projects

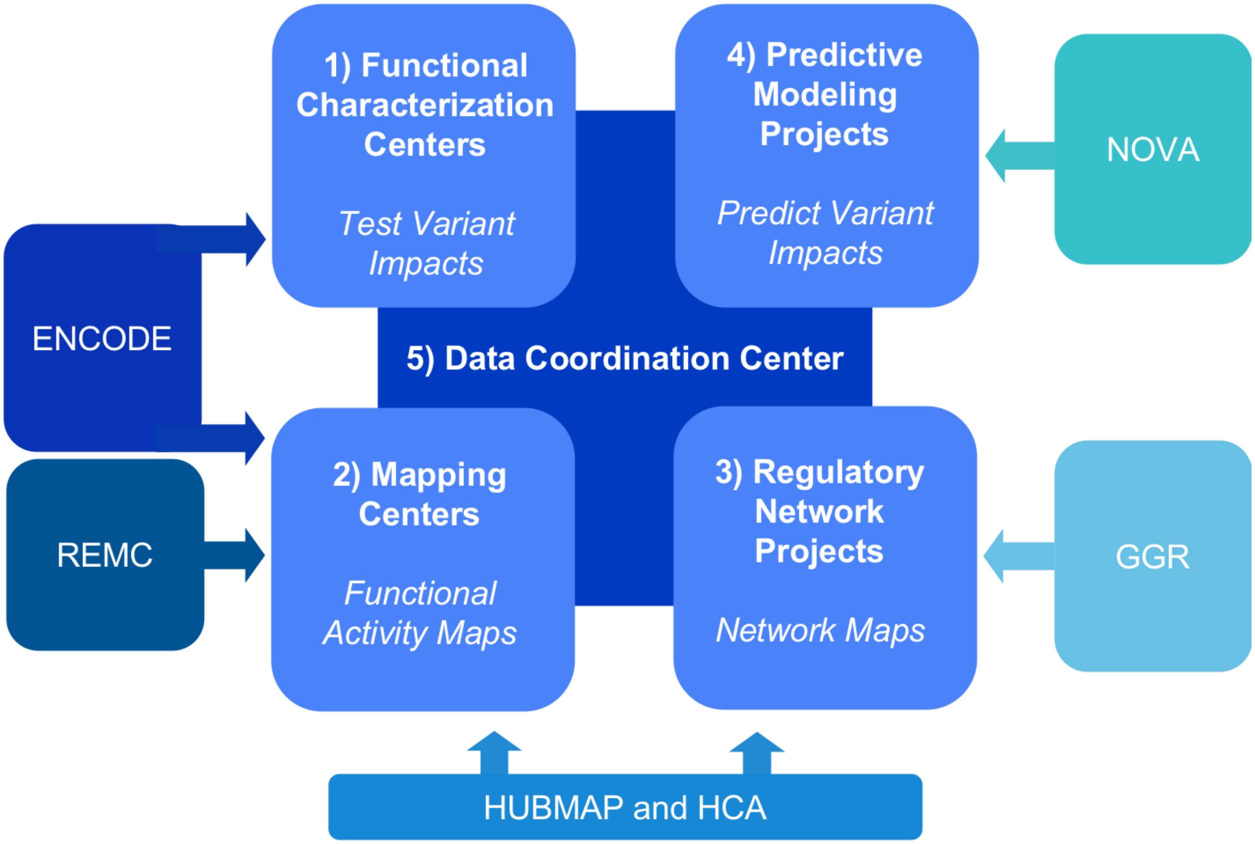

Figure 2: Foundational Projects.
Program components are represented as numbered boxes. External projects are represented as named boxes outside the central box. Arrows indicate foundational work that forms a basis for activities in this program. Abbreviations: ENCODE - Encyclopedia of DNA Elements; REMC – Roadmap Epigenomics Mapping Consortium; HuBMAP – Human BioMolecular Atlas Program; HCA – Human Cell Atlas; NoVa – Non-Coding Variants Program; GGR – Genomics of Gene Regulation.

Figure 3 Program Interactions

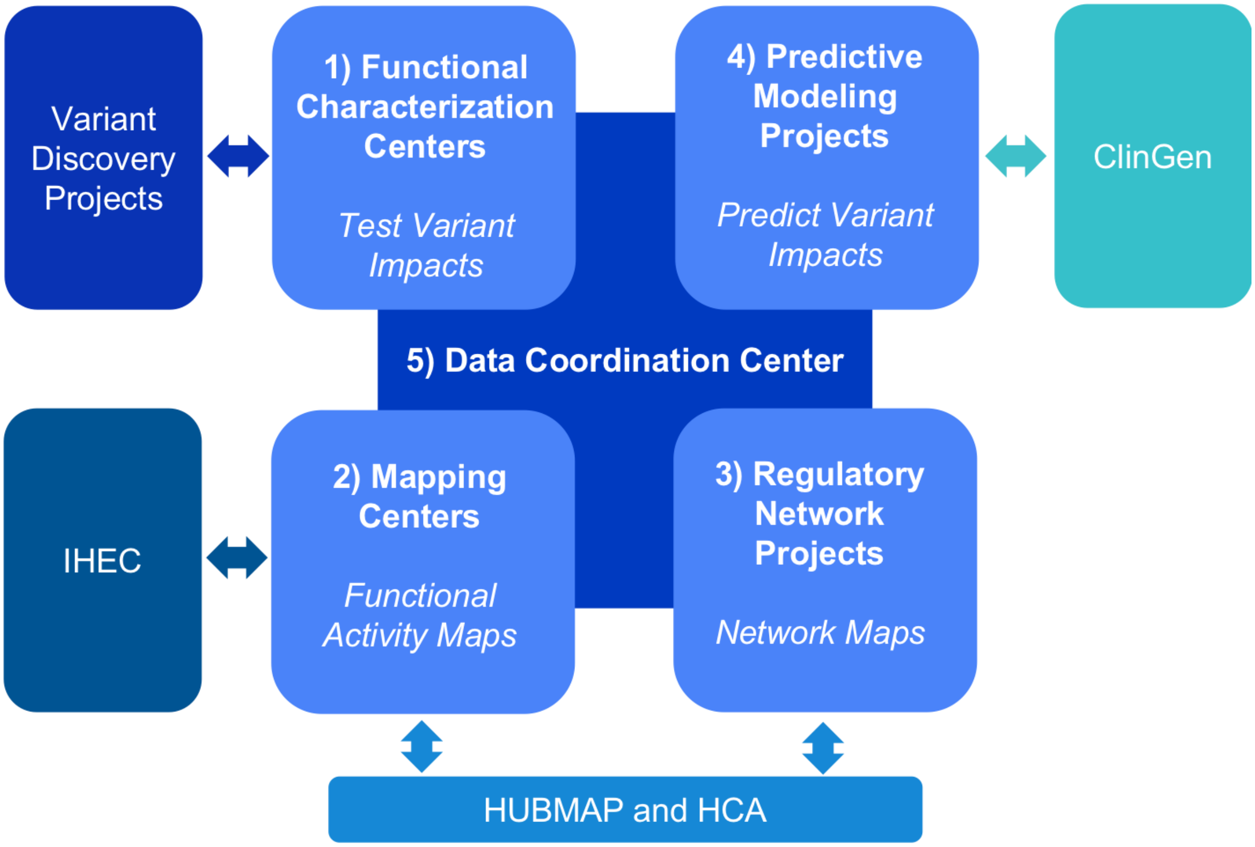
Variant Discovery Projects
1) Functional Characterization Centers
Test Variant Impacts
4) Predictive Modeling Projects
Predict Variant Impacts
ClinGen
5) Data Coordination Center
IHEC
2) Mapping Centers
Functional Activity Maps
3) Regulatory Network Projects
Network Maps
HUBMAP and HCA

Figure 3: Program Interactions.
Program components are represented as numbered boxes. External projects are represented as named boxes outside the central box. Arrows indicate expected interactions between program components and other entities. Abbreviations: IHEC – International Human Epigenome Consortium; HuBMAP – Human BioMolecular Atlas Program; HCA – Human Cell Atlas; ClinGen – Clinical Genome Resource.