National Advisory Council for Human Genome Research (NACHGR)

February 22, 2021

Concept Clearance for FOAs

Molecular Phenotypes of Null Alleles in Cells (MorPhiC) Phase 1

v. 02092021

Purpose:
A long-term goal of this proposal is to develop a consistent catalog of molecular and cellular phenotypes for null alleles for every human gene, using *in vitro* systems, to be made available for broad use by the biomedical community. In this Concept, we propose a Phase 1 of this effort that will focus on a target subset of 1000 protein coding genes. Specifically, this initial effort will develop standard methods to create null alleles and to measure their effects, assess the scale limitations of such methods, develop common data formats, and establish "use cases" for this catalog. Phase 1 will also inform a potential Phase 2.

Background:
The 2020 NHGRI Strategic Vision laid out a set of "bold predictions for human genomics by 2030" that included: "The biological function(s) of every human gene will be known; for non-coding elements in the human genome, such knowledge will be the rule, rather than the exception." Further, systematically obtaining genome-wide information about gene functions is critical for elucidating the roles and relationships of genes and regulatory elements in pathways and networks, which is one of the compelling genomic research projects outlined in the NHGRI Strategic Vision.

A catalog of molecular and cellular phenotypes of null alleles (e.g., created with CRISPR) assayed in multicellular contexts (e.g. iPSC-derived organoids, or other systems) would provide wide-ranging insights into gene function, filling in a gap between more direct molecular readouts, such as RNA expression, and whole-organism phenotypes. Importantly, such data would provide a foothold for understanding the mechanisms through which genes act to produce phenotypes (including human diseases) and aid in understanding pathways and gene regulatory networks.

A key idea for this Concept is that the data be *consistent* (i.e., generated using standardized assays yielding similar depth and breadth of information). A coordinated approach will yield data that will be more readily compared between genes and integrated with other functional and phenotypic data, making it a better substrate for computation and modeling. This provides advantages over efforts that aggregate existing data from many diverse sources.

Another key idea is that the catalog will provide a well-defined "minimal set" of information for all genes and will, in that respect, be *comprehensive*. We recognize that it is not yet possible to be comprehensive in all aspects of gene function (e.g., knowledge about all alleles in all cell types derived from all potential assays). As a result, this Concept is focused on exploring how best to derive an important subset of this information: molecular and cellular phenotypes in multicellular *in vitro* systems for null alleles for protein coding human genes.

The utility of null alleles —generally, those producing no functional protein— is well-established: the null is useful as a basis for interpreting other alleles, including regulatory alleles that may have weaker effects (e.g. lower penetrance and expressivity), and gain-of-function alleles. A null allele is also useful for interpreting alleles of other genes that produce

similar or "opposite" phenotypes, and interpreting effects of other perturbations (e.g., small molecule or environmental exposures). While null alleles are the focus of this Concept, other alleles may need to be assayed in some cases (e.g., if the null allele is lethal).

A catalog of information about null alleles and their associated molecular and cellular phenotypes could also serve as a reference for assimilating additional data about human gene function. Towards that end, it will be important to ensure that this catalog is effectively integrated with other basic and clinical genomics resources.

Proposed Scope and Objectives:
Phase 1 is proposed as a four-year project with three components that will explore the feasibility of developing a catalog comprising all human genes. The focus areas will include: demonstrating the scalability of the needed technologies; developing methods for creating the alleles (e.g., investigating consistency, throughput, and compatibility with assay systems); developing high-throughput yet informative assays; developing data standards and quality control elements; and assessing the utility of the data. Phase 1 will focus on coding genes, but ultimately the catalog will include RNA genes.

Overall, this initial phase will focus on identifying and addressing barriers to the development of a complete catalog. Data production activities will test multiple approaches for assaying specific molecular and cellular phenotypes, establishing what will work while still producing data that can be used internally for comparison of approaches as well as by the community. Phase 1 activities should confront challenges to scale-up, including the prospect for cost decreases and throughput increases for assays, and also address barriers to interpretation (e.g., tissue specificity, pleiotropy, cell lethality), identifying ways to overcome these challenges. Ideally, Phase 1 will suggest the best assays and structure for a potential Phase 2. All components of the program will work together to address issues related to common aspects (e.g., standards, analysis, and reproducibility).

One potential outcome of Phase 1 may be that it is not feasible to achieve the scale and consistency to move to a Phase 2. Even so, the data and analyses performed in Phase 1 will substantively advance our knowledge of gene function.

The components described below will operate as a consortium:

I. Data Production Centers: System and Assay Development to explore, develop, and compare approaches; (4-6 awards)

- Produce null alleles and high-throughput cellular and molecular assays in *in vitro* human system(s) of choice. Multicellular systems (e.g., organoids) are preferred. We propose a target of 1000 protein coding genes, with the idea that there should be some overlap in genes tested between Data Production Centers to facilitate comparison across assays and production groups.

- Lead consortium prioritization of genes to study. Prioritization should be based on one or more of the following: genes with no known function; genes associated with a disease or mouse knockout phenotype but with unknown mechanism; subset chosen to capture a breadth of functions (tissue specificities or restrictions; protein class; etc.).

- Develop metrics and quality standards; standardize allele and assay validations.

II. Data Analysis and Validation Centers to evaluate and ensure utility of the generated data; (2-3 awards)

- Lead specific analyses that raise and address key analytic issues (e.g., imputation, pleiotropy, pathways/networks, inference of cell type, interpretation of alleles).
- Identify and investigate data and design issues (e.g., validation, sampling/statistical analyses of data, phenotype descriptions and ontologies, integration of data, and utility for different "use cases").
- Identify additional community resource deliverables.

III. Data Resource Center to receive, annotate, and present data for consortium and public use; (1 award)

- Integrate data from Data Production Centers and host secondary data/tools from Data Analysis and Validation Centers; data wrangling.
- Develop approaches to enable community use of the data.
- Integrate consortium activities with other data types and projects, including those looking at specific loci or specific genomic variant effects.
- Serve as a logistical coordinating center.

Relationship to Ongoing Activities:
NHGRI funds other efforts to characterize functional genomic elements (ENCODE); and the impact of genomic variation on function (IGVF). NHGRI also has an ongoing interest in tissue-specific gene expression and regulation (such as the Common Fund GTEx effort). This Concept complements these activities. For example, consistent null molecular and cellular phenotypes will provide a basis for interpreting non-coding alleles found in "variant X molecular phenotype" studies. These data will also improve our ability to interpret variant association studies (e.g., the NIH-funded Genome Sequencing Program, Mendelian Genomics Research Consortium, disease-focused sequencing studies at other institutes) by enabling better inference between variant, gene, molecular/cellular, and anatomical/physiological phenotypes.

NHGRI also has a direct interest in human gene "knock outs" expressed through its support of the Common Fund Knockout Mouse program, and gnomAD. These projects, and other similar efforts, would be enhanced by information about molecular/cellular phenotypes of null alleles.

Mechanism of Support/Funds Anticipated:
NHGRI will use cooperative agreement mechanisms and commit approximately $10M/year total cost for 4 years beginning in FY2022 for a total of $40M.

- 4-6 Data Production Centers (UM1; averaging $1.4M/year total cost each)
- 2-3 Data Analysis and Validation Centers (U01; $0.5M/year total cost each)
- 1 Data Resource Center (U24; ramp up to $1.5M/year total cost)