National Advisory Council for Human Genome Research

September 16-17, 2019

Concept Clearance for RFA

Mendelian Genomics: Solving the Unsolved

Purpose:

Despite tremendous progress towards understanding the molecular basis of Mendelian disease, a significant proportion of known or suspected Mendelian disorders remain unsolved following standard whole exome sequencing. This concept clearance proposes two Funding Opportunity Announcements that will form a consortium aimed at significantly increasing the proportion of Mendelian disorders with an identified genetic cause through enhanced data sharing, collaboration, and an increased focus on the application of new technologies, sequencing strategies, and analytical approaches.

Background:

According to the Online Mendelian Inheritance in Man (OMIM) catalog, over 3700 of the ~20,000 genes encoded in the human genome have been conclusively associated with a Mendelian disorder or trait. A large proportion of these discoveries have been made within the context of the NHGRI Centers for Mendelian Genomics (CMGs). This consortium was launched in 2012, and renewed in 2016, with the goal of identifying the underlying genetic basis of as many Mendelian disorders as possible. In addition to the nearly 1800 novel gene/phenotype associations that can be directly attributed to the CMG program, many more such discoveries were facilitated through the dissemination of data, methods, and other tools for Mendelian gene discovery that were developed by the CMGs. Indeed, over the past decade Mendelian gene discovery using whole exome sequencing has become much more commonplace.

Despite the great success of the CMG program, several challenges remain in the field of Mendelian genomics. One is the high rate of unsolved cases – i.e. known Mendelian phenotypes/disorders for which whole exome sequencing did not identify an underlying genetic cause. Despite a stringent selection process aimed at choosing cases with the highest likelihood of being solved, approximately one third of the cases sequenced by the CMGs had no identified candidate gene, and only half of the cases analyzed had enough supporting evidence to be classified as truly solved. Some of these cases might be solved with re-analysis or by identifying additional individuals with the same underlying condition, but many of them may reflect cases of more complex genetic inheritance, mutations that are located outside of coding regions in functional genomic elements, structural variants, or other mechanisms that are not amenable to being identified using whole exome analysis. Because the novel gene discovery rate was the primary metric of success for the CMG program, these more complex cases were not exhaustively investigated.

An additional challenge is the need for more scalable approaches to functional validation of candidate variants, both to prioritize among a list of potential candidates, and to confirm suggestive variant/phenotype associations. Once a candidate variant(s) is identified, additional experiments must often be done to raise the association from the tier 2 (i.e. suggestive) to tier 1 (i.e. confirmed) level, including demonstrating that the candidate

variant does in fact impact protein function in relevant pathways or recapitulating the phenotype in a model organism. These types of corroborative data are necessary before researchers can begin additional investigation into the underlying biology of the disease, and before the findings can be translated into diagnostic testing.

One of the most significant challenges in this field are the barriers to data sharing. Data sharing will be especially important for solving cases that were not solved after initial whole exome sequencing, not only because it can assist in identifying additional related cases, but also because it allows other researchers to apply novel analytical approaches that may prove successful. The restrictive data use limitations and limited phenotype information of many historic samples hinder data sharing and make the data less useful for further analysis. Traditionally challenges have also included limitations to the platforms for responsible sharing of participant and patient data. There is a need to move the field towards increased, but still responsible, data sharing, focusing on samples consented for broad data sharing and on the development of an infrastructure that will enable secure, but accessible, sharing among researchers.

In order to address these challenges, this concept proposes a new Mendelian genomics program that places a stronger focus on solving these unsolved, more complex cases. The program will be built upon a foundation of broad data sharing and cross-consortium collaboration and will include opportunities for outside investigators to participate in analysis and validation efforts through the use of flexible opportunity funds.

Proposed Scope and Objectives:

The components of this initiative and their major objectives are described below.

Mendelian Genomics Centers (MGCs)
The Mendelian Genomics Centers will:

- Use state-of-the art sequencing technologies to uncover novel gene/phenotype associations underlying Mendelian disorders and traits
- Employ a variety of strategies to identify individual cases or cohorts of individuals with Mendelian disorders that are consented for general research use through controlled access, and for which detailed clinical and phenotypic data and metadata are available to be shared
- Significantly increase the solve rate for Mendelian disorders and traits by applying new technologies or sequencing strategies, developing novel analytical approaches, and working collaboratively with the other MGCs to solve cases that are not initially solved using standard whole exome sequencing
- Develop scalable approaches for functional follow-up in order to move findings towards tier 1 (confirmed) status
- Play an active role in working groups focused on specific cross-consortium activities, such as variant interpretation, functional follow-up, international data sharing, or other working groups that may arise
- Undertake outreach and education efforts on behalf of the program, empowering the broader research community to conduct Mendelian gene discovery

Data Coordination Center (DCC)
The Data Coordination Center will facilitate Mendelian gene discovery both within the Mendelian Genomics Centers and the broader scientific community by:

- Working with the MGCs to establish standards for genomic data, phenotype data, and appropriate metadata, in accordance with FAIR principles

- Receiving data, phenotype data, and other metadata from the MGCs, and ensuring that quality controlled, complete data and variant level information are submitted to AnVIL, ClinVar, Matchmaker Exchange, and other resources, as appropriate
- Tracking ongoing collaborations and samples being investigated by the MGCs, to ensure transparency and facilitate coordinated analyses across cohorts
- Developing a portal for disseminating the findings
- Overseeing and administering the Mendelian Genomics Opportunity Fund program, which will provide flexible seed funding to investigators outside of the Mendelian Genomics program to carry out validation studies needed to move candidate variants from suggestive to high confidence associations
- Developing the capacity to share sequence and phenotype data from unsolved and other more complex cases (such as phenotype expansions) in a manner that fits within the larger ecosystem of Mendelian and rare disease research – i.e. an Unsolved Cases Clearinghouse
- Coordinating logistical aspects of the consortium, such as organizing working groups, meetings, and consortium-wide communication

Relationship to Ongoing Activities:

Data generated by the Mendelian Genomics Centers is anticipated to be shared using AnVIL, the NHGRI funded NIH designated data repository. The Data Coordination Center will work with AnVIL to deposit data, that has been appropriately quality checked, and to develop a portal with which the scientific community can interact with the data and discoveries of the MGCs in a meaningful way.

The NIH Common Fund supports two ongoing programs related to rare disease genetics, though neither is specifically focused on Mendelian disease. The Undiagnosed Diseases Network (UDN) aims to improve the level of diagnosis and care for patients with undiagnosed diseases using a variety of complementary approaches, which include sequencing. The Gabriella Miller Kids First Pediatric Research Program (Kids First) has the goal of identifying the underlying genetic pathways involved in structural birth defects and childhood cancers. Like most other rare disease sequencing efforts, both Kids First and UDN have some proportion of cases for which the genetic basis of the disorder is not easily identified using whole exome sequencing. These cases could be pulled into the Unsolved Cases Clearinghouse in order to facilitate collaboration across these related consortia.

Mechanism of Support:

Due to the complexity of the program and the need for programmatic oversight and flexibility to achieve agreed-upon consortium goals, this program will use a cooperative agreement mechanism.

Funds Anticipated:

NHGRI will commit approximately $60M over 5 years to support 3-5 MGCs and one Data Coordination Center.

	FY21	*FY22*	*FY23*	*FY24*	*FY25*	*5 yr. total*
MGCs	$9 M	$9 M	$9 M	$9 M	$9 M	$45 M
DCC	$2.25 M	$3 M	$3 M	$3 M	$3.75 M	$15 M
Opp. Funds	(0)	($750K)	($750K)	($750K)	($750K)	($3 M)
total	$11.25 M	$12 M	$12 M	$12 M	$12.75 M	$60 M