2014/2015 Genome Sequencing Program Concepts

September 2014 Meeting of the National Advisory Council for Human Genome Research

OVERVIEW

We propose the following concepts based on the discussions at the July 28-29, 2014 NHGRI workshop: *Future Opportunities for Genome Sequencing and Beyond*. For the September 8, 2014 meeting of the NACHGR, we request consideration for Concept Clearance for Items I through III only. Items IV and V are provided for context, and will be considered for a future meeting.

I. Centers for Common Disease Variant Discovery —Purpose: to identify variants contributing to common diseases; to develop the means to do so comprehensively; to explore a range of example disease architectures and study designs; and to develop resources for multiple disease research communities and the wider biomedical research community.

Mechanism: Cooperative agreements; $60M of NHGRI funding in Year 1; seek funding partnerships to increase number of example diseases studied; four years; November 2015 funding.

II. Centers for Mendelian Genomics (CMGs)—Purpose: to identify the genomic bases of ('solve') ~300 Mendelian disorders that represent a broad range of phenotypes; to understand the genomic characteristics of Mendelian disorders as a class; to learn what it will take to solve all Mendelian disorders; and to develop and disseminate resources, methods, and tools to lay a foundation for solving all Mendelian disorders.

Mechanism: Cooperative agreements; $10M of NHGRI funding in Year 1; seek co-funding to increase number of disorders studied to at least ~400; four years; November 2015 funding.

III. Genome Sequencing Program Coordinating Center—Purpose: to support administrative and logistical functions, and coordinate and participate in certain analysis activities for I and II, and possibly other analyses that cut across multiple NHGRI-supported sequencing-oriented programs such as CSER, ENCODE, eMERGE, NSIGHT, and UDN. Activities would include tracking cost, production, project completion; and arranging/coordinating activities related to interactions with research and disease communities.

Mechanism: Cooperative agreement; $1.5M in Year 1; four years; November 2015 funding.

Potential future concepts:

IV. a. Producing High-Quality ("Gold") Genome Sequence— Purpose: to provide a well-selected (by population) set of 25-50 very high-quality human whole-genome sequences; to enable the detection of the vast majority of structural variants; and to improve the quality and utility of human genome references. Current cost is ~$250K each. In addition, refine the reference genomes for multiple primate species (see item V below).

Mechanism: Cooperative agreement for resource development; $2.5M in Year 1; three years; Funding in early FY 2016.

b. Improved Methods for Producing Gold-Quality Genome Sequences—Purpose: to improve methods and significantly reduce costs for producing and assembling gold-quality genome sequences. *This will be considered within the NHGRI Genome Technology Development Program.*

V. Comparative and Evolutionary Genomics – Purpose: to identify and foster high-priority projects of broad interest in comparative and evolutionary genomics, both as resources and to support analyses to address significant questions in genomics and genome evolution. These questions must be of broad significance, scope, and scale, that is, not something that could be supported elsewhere. The workshop identified two potential examples— one was to sequence sufficient branch-length representation in vertebrates to identify all conserved base pairs to single-base resolution; another was to refine the existing non-human primate references to enable reliable detection of lineage-specific sequence features. Other examples need to be encouraged.

Mechanism: Cooperative agreements; $2M in Year 1 (2-3 projects); three years; re-issue RFA at least once. *Bundle with the primate genome refinement with item IV.a;* Funding in early FY 2016.

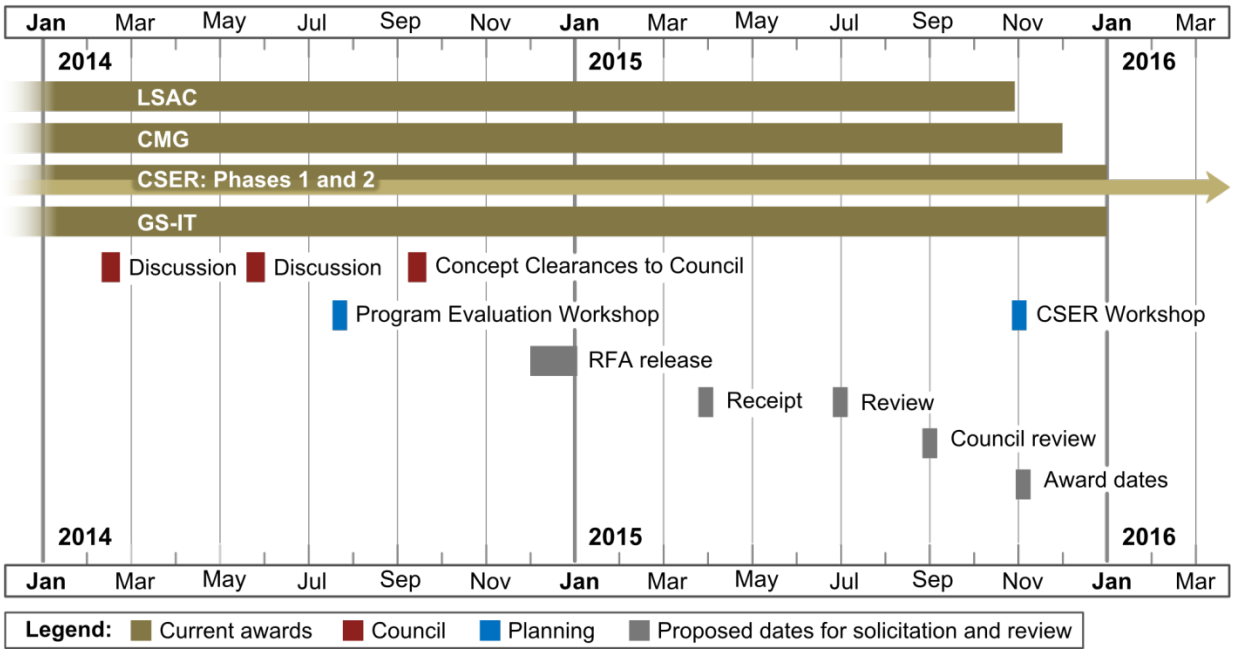
Fiscal Year 2015 estimated NHGRI Genome Sequencing Program Funding

Program	$M (with co-funding)	Cumulative Total
LSAC	**73**	**73**
CMG	**10 (+2)**	**85**
CSER	**15 (+2.7)**	**103**
GS-IT	**4**	**107**

Estimated Year 1 (Fiscal Year 2016) Funding of Proposed Concepts

Program	$M per year (without co-funding)	Cumulative Total
CD Variant Discovery	**60**	**60**
CMG	**10**	**70**
Program Coordinating Center	**1.5**	**71.5**
Potential future concepts:		
Gold Genome Production	**2.5**	**74**
Comparative and Evo. Genomics	**2**	**76**

Timeline for the Current Genome Sequencing and Analysis Program and Proposed Concepts



Background documents relevant to these concepts:

- **Report of the July 28-29, 2014 Workshop "The Future of Genome Sequencing and Beyond"; video of the workshop is available at http://www.genome.gov/27558042**
- **Presentations about the current Genome Sequencing Program: http://www.genome.gov/27556133 has links to presentations from the LSAC program from the February 2014 NACHGR meeting; http://www.genome.gov/27557112 has links to the presentations from the CMG and CSER programs from the May 2014 NACHGR meeting**

I. Common Disease Variant Discovery Centers (CDVD Centers)

Purpose

To fund a collaborative large-scale genome sequencing effort to identify rare risk and protective variants contributing to multiple common disease phenotypes; to explore a range of diseases with the ultimate goal of doing this for enough different disease architectures and study designs to understand general principles of how best to design rare variant studies for common disease; to better understand the genomic architecture underlying inherited disease; and to develop resources for multiple disease research communities and the wider biomedical research community. NHGRI intends that this program have as an explicit deliverable at least one comprehensive whole-genome sequencing study for a common disease, which we anticipate will become practical soon due to advances in the state-of-the-art. The program will necessarily include the need for some analysis and interpretation of variants, although this will be limited to available computational approaches.

Background

A comprehensive understanding of the genomic variants underlying inherited disease phenotypes is a key goal for biomedicine, with consequences for understanding diseases affecting billions of people. It is widely anticipated that a comprehensive account of variants affecting risk for, and protection from, disease phenotypes will afford improved diagnosis, prognosis, management, and even development of new treatments for inherited diseases. Moreover, there are large gaps in our basic understanding of the relationships between genotype and phenotype. The importance of this topic, and the current state of knowledge, were explored in an NHGRI Workshop on July 28-29, 2014 (a draft workshop report is being provided to the NACHGR as background; the final report will be posted on the NHGRI Web site).

For practical reasons relating mainly to technological approaches and needs for obtaining samples, this Concept will concentrate on variant discovery in common diseases; a separate Concept will focus on discovery of variants underlying Mendelian disease. We recognize that these two actually comprise a spectrum of disease architectures.

With regard to common disease variant discovery, genome–wide association studies have been effective in identifying common variants, but genome sequence data are required to identify the full spectrum of the genetic contributions to common diseases, particularly rare variants. Very large sample sets may be needed both to detect rare variants and to enable analyses such as clustering identified variants by gene function. Current estimates suggest that as many as 25,000 cases and 25,000 controls could be required for each disease study focused on analysis of protein-coding regions. Analysis of noncoding sequence could require even larger numbers, depending on assumptions about the effect size of noncoding variants. Comprehensive identification of contributing variants is difficult to define but could be indicated by a fall-off in the number of new alleles identified and/or the number of genes implicated. At least a few such very large, comprehensive studies are needed initially, to have sufficient data to test ideas about sample number requirements. NHGRI is uniquely positioned to undertake studies that are as complete as possible given the state-of-the-art, with adequate power to detect rare variants.

At the current time, producing sufficient whole-genome sequence (WGS) data for a comprehensive study is very costly and analysis methods are less well-developed for noncoding sequence. For these reasons, large common disease studies to date have mostly used whole-exome sequencing approaches,

but this is changing as DNA sequencing costs decline. NHGRI has the opportunity to explore how to design, implement, and perform analyses for WGS studies, recognizing that high-quality WGS data sets for common diseases will constitute important resources for a number of communities (including those developing methods to interpret noncoding variation). Variants in noncoding sequence will tend to be more difficult to interpret (e.g., smaller effect size; less clear assignment to gene function), and it is possible that even more samples will be required to adequately power such studies. The desirability of producing at least one high-quality WGS common disease dataset will need to be balanced against the need of the program to explore enough examples to understand a range of disease architectures and project designs.

Proposed Scope and Objectives

This Concept proposes to support work of a scale and scope that will be as comprehensive as possible for several diseases. It proposes to address several common diseases comprising a range of disease architectures that encompass, for example, differences in the number, population frequency, type, and effect size of underlying variation and that likely correlate with features of the disease phenotype, such as severity and age of onset. We propose that at least six to ten architecturally diverse disorders be explored comprehensively over the course of the program. A variety of study designs will also be encouraged including creative designs that may be more efficient (for example, choosing appropriate populations for study, in which some variants are enriched in frequency). Given NHGRI's unique role in propelling advances in large-scale sequencing, efforts proposed in this Concept will be confined to questions that can only be answered at very large scales. NHGRI intends that this program have as an explicit deliverable of at least one comprehensive whole-genome sequencing study for a common disease.

In addition to data production, innovative analyses will be required to provide useful insights into the major questions about common disease genomic architecture outlined above. Such analyses will be done in collaboration with multiple research communities (see *Mechanism of Support*).

Foundational goals and objectives

In addition to the scientific objectives above, the program has the following foundational goals:

- Developing durable data resources for the research community, including high-quality genome sequence and variant data for several large common disease studies and potentially a set of common controls.

- Disseminating knowledge through collaborations with multiple communities; in this case, each disease project will represent an opportunity to exchange and disseminate knowledge about design, methods, analyses, and limitations of genomic sequencing approaches.

- Making technological innovations to optimize cost and quality and expand to novel applications (e.g., non-standard samples, genome refinement, etc.).

- Developing data handling and analysis platforms compatible with advances in DNA sequencing and computer hardware and software.

- Improving and comparing genome analysis methods and informatics tools in the context of a research network within and across NHGRI programs.

- Serving as a locus for the development and refinement of policies and practices about genomic data deposition.

Additional features that will aid in attaining the main goals above include:

- Production and incorporation of other data modalities (e.g., transcriptomes and epigenomes), coordinated with other NHGRI functional genomics efforts.

- Projects that are compelling and broadly useful, but may not be directly related to studies of a specific disease (e.g., population variation studies).

- Administrative supplements for pilot collaborations with scientists outside the grantee institution to develop, for example, methods for analysis of noncoding sequence, developing functional validation of variants, etc.

- Cost savings on high-volume "commodity" sequencing, independent of technology development, including subcontracting of data production.

- To encourage collaboration, dissemination, and achievement of NHGRI programmatic goals, NHGRI proposes to hold back a significant proportion of funds each year to be awarded as supplements to support high-priority collaborations and cutting-edge projects. These could be used for the pilot collaborations described above as well as for NHGRI-identified priorities to fulfill goals of the Institute. See *Funds anticipated* below.

Identifying specific projects

This program will use multiple ways to undertake new projects.

- The applicants would propose fully-realized projects (including samples, etc.) within their applications sufficient for at least the first year of work, leaving capacity for new projects. Proposed projects that are related (i.e., similar diseases and/or phenotypes), will be coordinated by NHGRI before funding.

- NHGRI would solicit new projects for Year 2 and beyond though a reviewed mechanism (X01s), that may be proposed by or together with a funded grantee, or by other investigators. Criteria for selection will include the public health significance of the disease, diversity of the study population, whether doing the project will widen the number of disease architectures being explored, scientific merit (e.g., power or quality of phenotyping), sample availability, availability of other funding or other resources that will increase the likelihood of project success, and consistency of consents with data release policies.

- Project workshops in collaboration with scientific disease communities.
- Collaboration with other NIH institutes/centers.
- NHGRI-identified programmatic priorities.

In all of these instances, the work will also have to be reconciled with program goals, for example the need to explore a range of disease architectures and project designs.

Relationship to ongoing activities

The proposed program is conceptually situated in the center of many important ongoing activities funded by NHGRI and others. It will be unique in its focus on variant discovery, and the scale required to undertake comprehensive projects.

NHGRI-funded efforts:

Centers for Mendelian Genomics (CMG): CDVD and CMG will be managed as part of the same overall program, sharing a common Coordinating Center and external advisors.

Clinical Sequencing Exploration Centers (CSER): CSER will continue to have distinct aims and a separate Coordinating Center, with CSER pursuing general lessons at smaller scale about how and when to implement genome sequencing in a clinical setting and sharing those insights with data producers to enable clinically relevant technologic developments.

The Electronic Medical Records and Genomics (eMERGE): eMERGE seeks to acquire genomic sequence information in a set of clinical samples integrated with electronic medical records. NHGRI will look for opportunities for interaction, including whole genome sequencing, to encourage a virtuous cycle between discovery and the clinic. NHGRI's programs in newborn sequencing (NSIGHT) and undiagnosed diseases (UDN) will also be important potential participants in this regard.

Functional genomics program: Areas of interaction between variant discovery projects and functional genomics projects are likely to arise, especially in the area of downstream analysis of noncoding variants identified in this program. This will be of interest to the ENCODE and FunVar initiatives.

Other efforts:

Within NIH, NHGRI has had productive collaborations with a number of other institutes, including NIA, NCI, NIDDK, and NHLBI. Going forward, NHGRI aims to encourage additional, direct collaborations with the NHGRI Genome Sequencing Program.

Other efforts nationally and worldwide provide opportunities for understanding the genomic basis of common disease, including large cohort studies such as Genomics England, Million Veteran Program, UK Biobank, and UK10K. These prospective studies will eventually accumulate sufficient outcome data to power discovery for specific diseases, and will be critical for establishing risk associated with any variant(s). Their cohort designs will complement CDVD, which will concentrate on case/control designs

to attain power as quickly as possible. The CDVD program will maintain ongoing awareness of large cohort studies and actively seek opportunities for coordination, collaboration, and replication. One particular area for coordination is on development of common controls. The CDVD program will also maintain vigilance of ongoing individual disease studies to promote synergy and avoid duplication.

Mechanism of support

As in previous incarnations of NHGRI programs in large-scale genome sequencing, this program will need maximum flexibility to respond to ever-changing genomic technologies and the changing landscape of opportunities in the field. In the next four years, however, even more flexibility will be needed to accommodate rapid expansion and contraction of capacity due to co-funding, whether derived from NIH Institutes/Centers or the grantees' ability to leverage additional funding, as well as flexibility around the program goals to accommodate the aims of co-funders. Flexibility will also be needed in identifying and choosing projects and in consolidating and assigning projects to attain comprehensiveness in large studies. Withholding some block of funding at the start of each award year is one means by which NHGRI will increase administrative flexibility.

A cooperative agreement mechanism (e.g., U) will be required to provide the degree of staff involvement needed for this level of coordination and flexibility, while still fostering the scientific aspects required for success of the program.

A Genome Sequencing Program Coordinating Center will be needed to facilitate collaboration and joint analyses between CDVD and CMG, as well as other projects (see Concept III). The complexity and size of this program will require ongoing advice and guidance from an external scientific panel (ESP), which will provide guidance to the larger program (including at least this program, the Centers for Mendelian Genomics (Concept II), and the Coordinating Center).

This program requires that a small number (2-4) of awards be made both because of the amount of coordination that will be needed and because the program aims to pursue very large projects.

Funds anticipated

We propose providing $60M per year for four years for this program. As above, we expect that this will entail one or two large whole genome common disease studies, and six to ten large whole exome studies, selected across a range of disease architectures. Without the benefit of clear evidence on how many disorders will need to be explored (indeed, a clearer answer to this question is likely to be one of the products of the proposed effort), this number reflects our judgment about the *minimum* number of projects required for a credible effort to address the Concept goals and scope. Following this principle, cost considerations included:

- A comprehensive discovery effort could require obtaining genome sequence information from roughly 50,000 individuals or more for each disease (but see below).

- NHGRI's best available information about **current** (three month average retrospective) cost on existing DNA sequencing platforms is that a fully-loaded total cost per whole exome is $400, with an additional $50 for the automated portions of data handling and analysis.

- Based on our best current knowledge, we believe comparable costs for the new Illumina platform costs will be $2000 per whole-genome sequence next year (whole-exome sequences are not supported), plus $600 for automated data handling and storage.

We therefore project, conservatively, that next year a ~50,000-sample WGS study will require $130M for production and automated data analysis, and a ~50,000-sample WES study about $23M. Based on these conservative cost considerations, one WGS study and six whole exome studies will cost about $260M. Note that these total costs only include costs for data production and the automated aspects of data handling and production. This will result in a high-quality sequence with called SNPs and some other variants called.

This amount is not sufficient, however, to allow the general scientific or foundational goals of the program to be met (which will require non-automated analyses, technology development, etc.), nor does it allow any sustained collaborations with representatives of different disease communities (e.g., through project management, participation in consortia), an aspect that will be emphasized in this program. Based on NHGRI's tracking of cost and production data within large production centers, an additional amount on top of production costs, about 20% of overall funding, will be required to allow these additional program features to be accomplished in the context of a pipeline that is largely composed of whole exome studies. With the simplifying assumption that the additional costs are essentially the same whether a large study is whole exome or whole genome, this would translate to 2-4 funded centers each receiving on average $2-4M per year for all the non-production activities, including project leadership/management, custom analyses, technology implementation and incremental development, changes to analysis pipelines, foundational objectives, etc.

Based on all these considerations, we estimate ~$294M total, or ~$74M per year, will be required to have high confidence that the minimum goals of this Concept will be achieved. We propose, however, to provide significantly less than that---$60M per year, or ~80% of the program as estimated above---because we believe that several factors will effectively leverage NHGRI funding, either by adding to it or by decreasing costs. These include:

- Cost: Production and data storage costs are likely to decrease. If they decrease by half every two years, the estimate of the number of projects could increase (perhaps 1 or 2 large WGS and 10 whole exome studies, although the relative cost of the two project types may change).

- Study design: Not all studies may require so many samples, due to more efficient designs or more advantageous genetic architecture. In addition, if common controls can be developed, the number of samples needed per study will decrease significantly.

- Co-funding and other funding collaborations. As described below, we aim to provide mechanisms that will encourage co-funding and addition of outside funds.

As with previous iterations of the NHGRI genome sequencing program, we intend to evaluate the project cost on a continuous basis. The cooperative agreement mechanism allows the flexibility to make annual funding adjustments while ensuring that higher level program goals are met.

Leveraging NHGRI funding

If additional, non-NHGRI funds can be found for more example projects to be undertaken, it will enhance the chances of success in attaining the overall program goals, and each additional disease study will increase the impact of the program.

NHGRI will seek co-funding for this program, targeting the interests of NIH Institutes/Centers and other funding groups in the particular diseases being studies. Seeking co-funding is not simply to leverage funding; it will also improve and further disseminate the science by encouraging interactions with different constituencies that have extensive expertise about the biology of the diseases under study and who are committed to long-term follow-up studies beyond the genome sequencing efforts.

Furthermore, NHGRI will develop a means to incent grantees to find other sources of funding that enhance these projects, either by adding whole projects that fit with the program goals or by funding portions of ongoing projects. The ability to leverage additional resources for these projects will be one of multiple criteria for determining the issuance of mid-year supplements to grantees. As noted above, funds for such supplements will be made available by withholding a significant proportion of the funds designated for the program at the start of each award year. The specific amount to be withheld will be determined at the time of funding, and may vary from year to year.

II. Centers for Mendelian Genomics (CMGs)

Purpose

This Concept Clearance proposes to renew the Centers for Mendelian Genomics Program (CMG Program). NHGRI intends that the renewed program will: 1) identify the genomic bases of as many Mendelian disorders as possible using genome-wide sequencing at scale, and 2) bring the field forward toward the goal of solving all Mendelian disorders in the foreseeable future by enabling and coordinating with researchers worldwide.

Background

A comprehensive understanding of the genomic bases of Mendelian disorders will: 1) inform human biology and pathophysiology; 2) enhance our understanding of disease mechanisms; and 3) define diagnostic and therapeutic strategies for a broad range of both rare and common diseases. Achieving such goals is essential to two of NHGRI's strategic areas – understanding the biology of genomes and understanding the biology of disease. The completion of the Human Genome Project and advancement of genome technologies have made it feasible to find the genomic bases underlying all human Mendelian disorders in the foreseeable future. In November 2011, NHGRI initiated the CMG Program in collaboration with NHLBI. To date, by collaborating with investigators worldwide, the Centers for Mendelian Genomics (CMGs) have discovered the genomic bases of ('solved') over 160 Mendelian disorders. In addition, these disease gene discovery efforts have also resulted in discoveries of novel phenotypes of more than 120 previously solved disorders (i.e., phenotype expansion). Over 100 publications have been produced based on this work so far. In making these discoveries, the CMGs have: 1) demonstrated the power of genome sequencing at scale for solving Mendelian disorders; 2) revealed the extent of pleiotropy and genetic heterogeneity underlying Mendelian disorders; 3) developed tools for phenotype collection, storage, and analysis; and 4) developed innovative methods for solving Mendelian disorders.

Based on current estimates (CMGs and Online Mendelian Inheritance in Man, OMIM®, including personal communications), genetic changes underlying approximately 3,600 Mendelian disorders have been reported to date. In addition, approximately 3,700 Mendelian disorders remain unsolved, and more Mendelian disorders are reported each year. Thus, much work remains to be done in order to solve all Mendelian disorders.

Proposed scope and objectives

The main objectives of the next iteration of the CMG Program are: 1) to solve a large number of Mendelian disorders at the funded centers, and 2) to enable and coordinate with other programs and researchers to solve more Mendelian disorders.

Solving ~300 Mendelian disorders

NHGRI intends that this program, over time, solve as many Mendelian disorders as possible. This is an idealized goal, but the number should be sufficient to accomplish a number of things in addition to solving many individual diseases. This effort will solve a spectrum of Mendelian disorders---anticipating

that there is a range that spans, e.g., different types of variants (single nucleotide vs structural variant; coding vs. noncoding); allelic, locus, or genic complexity; somatic mosaicism; *de novo* single gene diseases; corresponding differences in the ease of solving these---in order to inform what approaches, methods, and scale would be effective to solve all Mendelian disorders. As a by-product of this, the program will uncover information about phenotype expansion, that is, new phenotypes caused by alleles already known to underlie a Mendelian disorder. This information is valuable in understanding the range and magnitude of the larger program goals; in addition it can have an impact on how clinicians categorize such disorders, and may ultimately aid diagnosis.

For this iteration of the program, we aim to solve at least an additional 300 Mendelian disorders (on top of the projected 300-plus for the current program). We base this number on the progress that the current CMGs have made thus far and the funds available (see below). We judge that this number is sufficient to continue to enable the CMG to attain its larger goals stated above.

In addition to these major goals, there are two new features of the proposed program. First, while whole-exome sequencing (WES) is expected to continue to be effective, whole-genome sequencing (WGS) will be used in some cases (e.g., where discovery based on exome sequencing failed to yield insight) in order to examine the entire genome. Second, while we expect that function assays following up from variant discovery to continue to be done largely by outside collaborators, we intend to support small-scale function assays at the funded centers with a small portion of the program funds in order to provide validation, to begin to understand mechanism, and more generally to explore productive links between “discovery” and “function” aspects of other NHGRI programs.

Enabling discoveries to be made elsewhere

In addition to the scientific goals, the magnitude of the effort described above will enable it to be a resource for the field, by achieving the following:

- Developing methods and tools. Currently, the success rates in solving Mendelian disorders (‘solve rates‘) at the CMGs average around 40% , and the ability to achieve such solve rates is not yet widespread. The grantees are expected to improve solve rates, efficiency, and costs by developing and refining: 1) approaches (genotype driven, phenotype driven, combination with non-sequencing genomic methods, etc.); 2) study designs; 3) methods and tools for collection, storage, and analysis of phenotype information; 4) sequence data production, and 5) data analyses of difficult genomic regions, such as repeat expansions, CNVs, fusions, etc.

- Outreach and coordination. Solving all Mendelian disorders will require unprecedented coordination and collaboration worldwide. Reaching out to clinicians and researchers and participating in the International Rare Diseases Research Consortium (IRDiRC) will continue to be necessary activities to pursue. Given the rarity of Mendelian disorders, project coordination is highly necessary to help accelerate discoveries. More specifically, we expect that coordination should be achieved by: 1) sharing lists of disorders under investigation, and 2) matching of samples or candidate disease genes of the same disorder.

- Dissemination. Dissemination of data, tools, and methods will continue to be an important aspect of the program. Sequence data are expected to be released in multiple forms (e.g., sequence in dbGaP, public posting of allele counts and causal allele information, etc.)

Relationship to ongoing activities

The NIH Common Fund's Undiagnosed Diseases Network (UDN) aims to diagnose rare diseases using sequencing and other genomic methods. While both UDN and the CMG Program will solve rare diseases, the CMG Program is a discovery program that mainly studies existing samples by sequencing at scale, rather than making diagnoses of newly enrolled patients. An outcome of this program will be an increase of the number of diagnosable rare disorders. Conversations have been initiated to explore the potential of coordination and collaboration between the two programs.

The NHGRI funded Clinical Sequencing Exploratory Research (CSER) studies may not be able to make molecular diagnoses of all enrolled patients. The CSER and CMGs have established a channel of communications to pass undiagnosed and appropriately consented patients, particularly patients suspected of having Mendelian disorders, to the CMGs for discovery research.

Mechanisms of Support

As with the current CMG Program, we propose to use a cooperative agreement mechanism due to the size and complexity of the effort, and the need for flexibility to accommodate changes in technology, knowledge, and opportunities for collaborations within and outside of the program, including with the ongoing Clinical Sequencing Exploration Research (CSER) program and the CDVD centers described under another Concept presented today. Up to three awards will be made with available NHGRI funds. The program announcement will be open to all applicants and will not be restricted to existing CMG awardees.

The complexity and size of this program will require ongoing advice and guidance from an external scientific panel (ESP). It is likely that the ESP will provide guidance to the larger program (including at least this program, the Centers for Common Disease Variant Discovery (Concept I), and the Coordinating Center.

Funds anticipated

We propose a total NHGRI funding of $40M for four years.

- Approximately $35.4M total cost is estimated for solving ~300 Mendelian disorders. This estimate is based on the current CMGs' total spending so far on small scale sample solicitation; phenotype collection and evaluation; sample processing and QC; WES; data analysis; publications; and coordination and dissemination activities outside of the CMG Coordinating Center. In addition, it assumes continued improvement of efficiency and costs, and that WES will still be the main sequencing method for discovery. Approximately $1.4M is included for WGS of a small number (~210) of samples at the cost of $6600 per genome. The exact number of WGS to be performed will depend on how the technology matures, availability of promising cases where WES has failed to discover the underlying genetic causes, overall cost saving and improvement of efficiencies, and availability of other potential funding sources. NHGRI will continue to seek co-funding for this program. The discovery goal will go up if co-funding is available. Currently, NHLBI provides $2M annual funds for the Program. If NHLBI commits to

the same level of funding for the next iteration, then we anticipate that the combined funds would support ~400 Mendelian disorders to be solved.

- Based on lessons learned from the current CMGs and the growing need for worldwide coordination, NHGRI expects the next iteration to give more emphasis to sample solicitation, outreach, and dissemination. With the proposed Genome Sequencing Program Coordinating Center being responsible for some of the activities currently carried out by the CMG Coordinating Center, some other coordinating activities will be absorbed into individual CMGs. The proposed budget includes more FTEs and associated costs for sample solicitation, coordination, outreach, dissemination, and training, for a total of approximately $4.6M.

- NHGRI expects the grantees to find ways to reduce overall costs in order to provide for a small number of function assays. No budget has been estimated for this activity.

III. Genome Sequencing Program Coordinating Center (GSPCC)

Purpose

The GSPCC will be responsible for administrative, logistical, coordination, analysis, and outreach activities that arise separately within, or cut across, the CDVD and CMG programs. Given the connection of these programs with others, the GSPCC will also have a role in coordinating other activities that may arise across programs.

Background

The current NHGRI Genome Sequencing Program is large and complex, requiring tracking of production data and extensive logistical coordination of large projects (e.g., multiple weekly conference calls, project documents, meetings). Many large programs at NHGRI already have coordinating centers to help manage these tasks. Moreover, coordinating centers are well-positioned to participate in analyses that span efforts of multiple grantees; both the CDVD and CMG program Concepts explicitly envisage high-level goals that will require integration of data across multiple grantees.

The GSPCC will, of necessity, need to work closely with other program grantees and with NHGRI program staff. A key criterion for this role will be success in leading large, diverse, and dispersed consortia.

Proposed scope and objectives

The GSPCC will:

- Work with NHGRI staff to track throughput, cost, data deposition, and project completion information.

- Maintain websites that provide information to the community about the activities of the individual programs, including causal variant information, project completion status, etc.

- Organize and take minutes for multiple conference calls required to coordinate complex consortia. This will include posting on an internal web site documents for the network of funded researchers and tracking key action items. Currently, a single large project consortium can require multiple conference calls per week. As the number of large projects expands (as planned in the Concepts above), this need will grow. This role may evolve into one that coordinates large projects that will have multiple interested parties (grantees, co-funding institutes, communities, etc.).

  Help organize meetings of the research consortium, as needed. In addition, the GSPCC will help organize workshops between the program grantees and the community (e.g., for outreach; and for project selection; see Concept I, Project Selection).

- Have an important role in analysis and project development activities. In the implementation of Concept I, it is likely that large disease projects will be split among awardees. A successful GSPCC will be helpful in certain aspects of integrating the results, including helping to reconcile

differences in analyses (for example, different variant call sets on the same data), or in looking for ways to synergize across sample sets with similar phenotypes, or in leading the discussion about, and developing sound criteria for, whether or not a project is complete. This will also give the GSPCC a role in project design, including sample selection. The GSPCC will have a role in all studies and publications that are produced by the consortium as a whole, while noting that some individual studies may be the sole responsibility of a single production center.

- Be encouraged to identify other analyses that would cut across multiple grantees within one program or multiple programs. These could include design of arrays based on consortium data, quality assessments, allele frequency analysis, and other analyses.

- Aid in and coordinate the development of universal controls for common disease rare variant studies. Development of such universal controls will require analysis across all the WGS data produced by the overall program (including any produced by the CMGs) and outside of it. The scope of this task could potentially be large; we therefore aim to begin with tasking the GSPCC with developing design considerations for developing a universal control set, based on the projects that will be undertaken by the production centers.

- In an outreach role, be helpful in summarizing and communicating to the community general "lessons learned" about how to use genome sequencing to find rare variants. These could include for example, standard operating procedures, power/design considerations, technology limitations, knowledge/technology gaps, etc.

- Will not act as a data coordinating center (i.e., one that consolidates data to serve to the consortium), but this role may be considered in the future.

- Will have to select what analyses it pursues carefully, and leverage resources within the consortium and any resources available independently to the GSPCC. This may place limits on the GSPCC doing large global data analyses.

Relationship with ongoing activities

The CSER program has already established a coordinating center, which it depends on. That coordinating center will continue together with the CSER program.

The CMG program also has a coordinating center function which is carried out by one of the sequencing center grantees. Some of its current responsibilities will be separated out and will be undertaken by the GSPCC.

Mechanism of support

We propose a cooperative agreement (U) mechanism for four years. The GSPCC will need a high level of flexibility to adapt rapidly to new projects and consortia. No applicant will be able to fully anticipate the range of projects that will be undertaken by the successful applicants to Concepts I and II.

The GSPCC will be part of the program research consortium with the programs in Concepts I and II, and will be managed by NHGRI staff with the help of an external scientific panel.

Funds anticipated

We estimate that the logistical and administrative functions will require up to three FTEs at a bachelor's or master's degree level. We estimate that analysis functions (including project design, analysis, coordination, outreach, and other activities described here) would require up to two PhD- or postdoc-level FTE's with informatics expertise included. We believe that the effort will be complex enough that it will require 10-15% time from a senior investigator. In order to do analyses, the GSPCC must have computational infrastructure.

With overhead, and allowing some funds to stimulate a creative response to the analysis and outreach functions called for, and considering budget limitations, we estimate that a total of **$1.5 M** will be needed in the first year. This amount may need to grow in out-years if, for example, creation of a universal control set is shown to be feasible.