Development and Curation of a Universal Human Genomic Variant Database

Submitted on 09/27/11 to the
National Human Genome Research Institute
of the National Institutes of Health
in response to PAR-11-095
Genomic Resource Grants for Community Resource Projects (U41)

Proposed Project Period: 7/01/12 – 6/30/15

Principal Investigators
Heidi L. Rehm, PhD
Christa Lese Martin, PhD
Robert L. Nussbaum, MD

Executive Committee
Sherri Bale, Andy Faucett, Madhuri Hegde, David Ledbetter, Christa Lese Martin,
David Miller, Robert Nussbaum Heidi Rehm, Erik Thorland, Patrick Willems

TABLE OF CONTENTS

PROJECT SUMMARY:

The centralization of data on human genomic variation is a critical step in accelerating research within the field of genomic medicine. Such centralization within a single database will not only enable more efficient approaches to data analysis, but will also ensure the use of a uniform set of standards across the many communities contributing to the database and using it for research and clinical applications. To address this critical need, we have convened a broad group of stakeholders who will develop the appropriate standards for the submission of both genotypic and phenotypic information into the public domain and a process for expert curation. This work will build upon the success of the International Standards for Cytogenomic Arrays (ISCA) consortium, which has already made major advances in establishing such a repository for structural variation. However, critically absent is the larger body of curated sequence-level variation that still remains in fragmented and poorly annotated environments or is simply inaccessible with data held in proprietary clinical laboratory databases. Most of the available sequence information on disease-causing genes comes from the clinical laboratory evaluation of affected individuals. Creation of a centralized database would allow us to harness the collective experience of multiple laboratories to support evidence-based curation of structural and sequence-level variants. The vast majority of the clinical laboratories in the US have agreed to provide access to their data, recognizing that their ability to interpret variants will be much improved if larger bodies of data are available. Furthermore, the primary laboratories involved are all at the forefront of integrating next generation sequencing technology into clinical use, including large disease gene panels, whole exome and whole genome strategies, allowing us to capitalize on these innovative sources of data. However, getting this data into a common repository will require considerable support. Furthermore, by creating an infrastructure to facilitate communication between experts in each disease area, we will enable consensus driven evaluation of evidence for variants and expert level curation to achieve a clinical grade database. Access to all data and evidence on human genomic variants will be maintained by ClinVar at the National Center for Biotechnology Information and the state of curation of the variants (e.g. uncurated, single-source curation, expert-level curation or practice guidelines) will be marked allowing components of the centralized database to be used for different applications from basic science research to clinical decision support.

RELEVANCE:

Hundreds of thousands of disease-causing variants have been identified in patients with disease, yet only a small fraction of that data, and the interpretation of it, is accessible to researchers and clinicians. This project will serve to collect and organize genomic data from many sources into a free and publically accessible environment and enable expert curation of that data for use in improving healthcare and biomedical research.

SPECIFIC AIMS

Human genomic variation underlies almost all human disease. Technological advances will soon make whole genome sequencing commonplace in the medical care environment, sparking an expansion of both basic and clinical research. In time, millions of individuals with and without disease phenotypes will undergo whole genome sequencing. At this time, however, our ability to detect DNA variation has greatly surpassed our ability to interpret the clinical significance of the variants detected. Correlating variation in individual whole genomes with clinical phenotype remains extraordinarily challenging due to the lack of publicly available and carefully curated information on gene variants. Collections of disease-associated variants that are publicly available are often subject to inconsistent standards and uncertain accuracy. In contrast, collections of disease-associated variants have been carefully curated from patient populations by individual clinical laboratories, but are currently sequestered within each of these laboratories and unavailable to the community. To address this challenge, we propose to create a clinical grade variant database that concentrates the knowledge and curation capabilities of our community into a single environment. This resource will enable the robust and systematic collection, curation, and sharing of human genomic variation for the benefit of the growing community of medical genomics researchers. Such an effort has already begun within the structural variant community, and this project will build upon its successful foundation to incorporate sequence-level variation. The specific aims to achieve this goal are:

1) Develop standardized formats for acquisition and submission of clinical genomic variation datasets. Develop a data element dictionary with broad input from the community to support the consistent description, annotation, and clinical classification of genomic variants (structural and sequence level) as well as standardized phenotypic data. Data standards will ensure the uniformity and integrity of the data and facilitate data transfer across all systems and community use.

2) Coordinate the submission of variant and phenotypic data into ClinVar, a universal centralized database at NCBI. Obtain structural and sequence-level genomic variants with clinical interpretations from clinical laboratories in the US, including 36 labs that have agreed to contribute sequence-level data as well as over 150 laboratories that are already involved in the International Standards for Cytogenomic Arrays (ISCA) Consortium gathering structural variant data. The clinical molecular laboratories chosen for this initial development period are highly experienced in sequencing diagnostics for single gene and gene panel testing, but are also at the forefront of integrating next-generation sequencing technologies into their services including whole exome and whole genome sequencing. As the necessary infrastructure is created and tested for long-term sustainability, submission support will be expanded to other laboratories. In addition, existing locus-specific databases from research or clinical laboratories will be integrated into the centralized database with their locus-specific curators joining the system.

3) Implement sustainable expert clinical level curation systems of human genomic variants. Develop protocols for expert clinical curation of structural and sequence-level variants utilizing a system of expert curators to achieve scalable solutions for ongoing curation of human genomic variation. We will build upon the structural variant curation system developed as part of the ISCA project and develop model curation projects for sequence variants in eight disease areas in which the participating researchers and laboratories have extensive expertise. The curation systems developed in this Aim will guide the standardization of variant classification in Aim 1.

At the conclusion of this project, we will deliver a fully functional resource for the ongoing collection, curation, and sharing of human genomic variation. This unprecedented resource will immediately improve the delivery and effectiveness of healthcare by providing a universal clinical genomics resource that integrates and supersedes current databases that are proprietary and/or difficult to access. In addition, through our Access and Dissemination Plan, we will engage a broad audience (e.g., clinical and research laboratories, clinicians, and patient advocacy groups) in the continued development of this database to ensure its success. Using this strategy, this high-quality database will accelerate progress across a wide spectrum of basic and clinical research, improving our understanding of human genomic variation.

RESEARCH STRATEGY - OVERVIEW

The crux of this application is the formation of an international consortium of cytogenomic, molecular and clinical geneticists based in diagnostic laboratories, research laboratories, NCBI, and clinical facilities to create a **universal**, **curated**, **database** of **structural and sequence-level genomic variants linked to phenotype. This database will be both freely and publicly accessible,** and will use new and existing software specifically designed for data submission and curation purposes. Accomplishing this goal requires 1) **software and database infrastructure** and 2) **human resources**.

The **software and database infrastructure** needed for creating and curating the Universal Human Genomic Variant Database, called ClinVar, is shown in Figure 1 and will be described in detail in the application. Briefly, standardized genotype and phenotype data will be submitted from participating laboratories, curated at various levels via an expert clinical level curation system, and stored in a publicly accessible database.

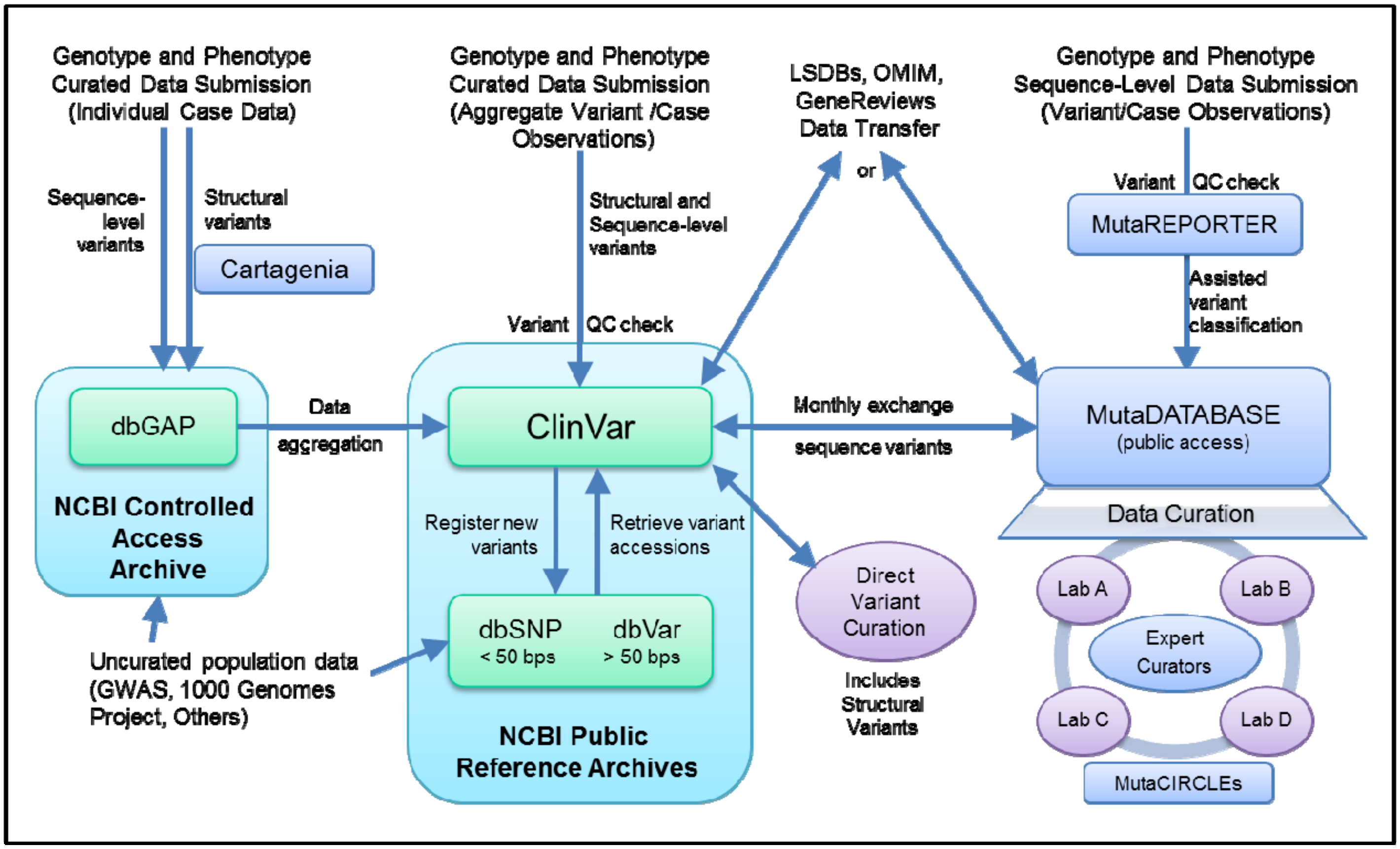

Genotype and Phenotype Curated Data Submission (Individual Case Data)
Sequence-level variants
Structural variants
Cartagenia
Genotype and Phenotype Curated Data Submission (Aggregate Variant /Case Observations)
Structural and Sequence-level variants
Variant QC check
LSDBs, OMIM, GeneReviews Data Transfer
or
Genotype and Phenotype Sequence-Level Data Submission (Variant/Case Observations)
Variant QC check
MutaREPORTER
Assisted variant classification
dbGAP
NCBI Controlled Access Archive
Data aggregation
ClinVar
Monthly exchange sequence variants
MutaDATABASE (public access)
Data Curation
Register new variants
Retrieve variant accessions
dbSNP < 50 bps
dbVar > 50 bps
NCBI Public Reference Archives
Direct Variant Curation
Includes Structural Variants
Uncurated population data (GWAS, 1000 Genomes Project, Others)
Lab A
Lab B
Expert Curators
Lab C
Lab D
MutaCIRCLEs

Figure 1. Overview of infrastructure and data flows

The **human resources** required to create and curate ClinVar, will be organized into five working groups: an overarching **Policies, Standards, and Sustainability Workgroup (PSSW)** that **sets** policies and standards, through interfacing with four additional workgroups that will **propose** and **implement** these policies and standards, each focusing on its own assigned area: 1) **Sequence Variants**, 2) **Structural Variants**, 3) **Phenotyping**, and 4) **Engagement, Education and Access**.

The PSSW is tasked with:

1. Developing **policies** surrounding data submission and defining the attributes of genotype and phenotype data that are to be collected and submitted to the database.

2. Defining **standards** for the classification of structural and sequence variants (genotype) using the clinical, biological, biochemical, and pathological features (phenotype) reported with the variants, including quality control.
3. Identifying strategies for the long term **sustainability** of the variant database.

The four focused workgroups will propose and implement the policies and standards of the PSSW while carrying out their own mandates as follows:

1. The **Sequence Variant Workgroup** will organize the data submission and curation of sequence-level variants including overseeing all model curation projects on individual genes or sets of disease genes, as well as the larger scale volunteer curation of other genes.
2. The **Structural Variant Workgroup** will organize the continued data submission and curation of structural variants.
3. The **Phenotype Workgroup** will define and implement approaches to the collection of clinical data for annotating genetic variants.
4. The **Engagement, Education, and Access Workgroup** will reach out to professional and patient groups in order to:
    a. encourage submission of data
    b. explain the various consent options including the opt-out process
    c. assist laboratories with IRB approval when required
    d. engage with professional organizations and laboratory groups to encourage all clinicians and laboratories to submit data
    e. partner with patient advocacy groups to increase the submission of patient phenotypes
    f. educate the research community about the resource and how to use it.

The organizational structure of the various workgroups constituting the human resources for this proposal is outlined in Figure 2 and will be described in more detail in the application that follows.

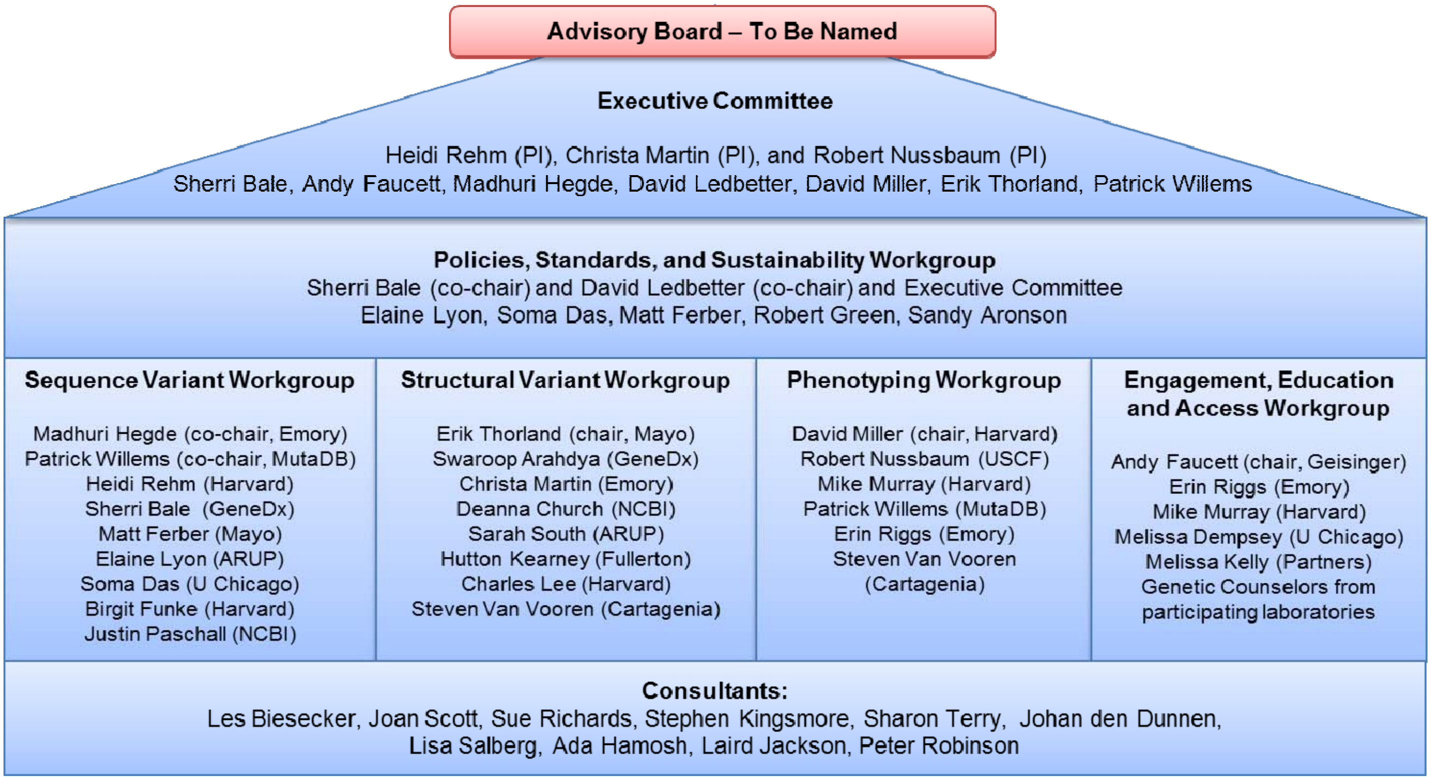

Advisory Board – To Be Named
Executive Committee
Heidi Rehm (PI), Christa Martin (PI), and Robert Nussbaum (PI)
Sherri Bale, Andy Faucett, Madhuri Hegde, David Ledbetter, David Miller, Erik Thorland, Patrick Willems
Policies, Standards, and Sustainability Workgroup
Sherri Bale (co-chair) and David Ledbetter (co-chair) and Executive Committee
Elaine Lyon, Soma Das, Matt Ferber, Robert Green, Sandy Aronson
Sequence Variant Workgroup
Madhuri Hegde (co-chair, Emory)
Patrick Willems (co-chair, MutaDB)
Heidi Rehm (Harvard)
Sherri Bale (GeneDx)
Matt Ferber (Mayo)
Elaine Lyon (ARUP)
Soma Das (U Chicago)
Birgit Funke (Harvard)
Justin Paschall (NCBI)
Structural Variant Workgroup
Erik Thorland (chair, Mayo)
Swaroop Arahdya (GeneDx)
Christa Martin (Emory)
Deanna Church (NCBI)
Sarah South (ARUP)
Hutton Kearney (Fullerton)
Charles Lee (Harvard)
Steven Van Vooren (Cartagenia)
Phenotyping Workgroup
David Miller (chair, Harvard)
Robert Nussbaum (USCF)
Mike Murray (Harvard)
Patrick Willems (MutaDB)
Erin Riggs (Emory)
Steven Van Vooren (Cartagenia)
Engagement, Education and Access Workgroup
Andy Faucett (chair, Geisinger)
Erin Riggs (Emory)
Mike Murray (Harvard)
Melissa Dempsey (U Chicago)
Melissa Kelly (Partners)
Genetic Counselors from participating laboratories
Consultants:
Les Biesecker, Joan Scott, Sue Richards, Stephen Kingsmore, Sharon Terry, Johan den Dunnen,
Lisa Salberg, Ada Hamosh, Laird Jackson, Peter Robinson

Figure 2. Overview of Executive Committee, workgroups and consultants

1. Rationale for the community resource

Understanding the biological and clinical significance of human genomic variation requires extensive investigation of both normal and disease populations. Several centralized, public databases of human variation in the general population now exist, including the integrated Database of Genomic Variation (DGV) [1] and Database of Genomic Structural Variation (dbVar) [2], and the Single Nucleotide Polymorphism Database (dbSNP) [2], which have already cataloged variation in thousands of individuals. In contrast, there is no single, comprehensive, freely accessible database of variation from disease populations. Most genomic variation databases from disease populations, so-called "locus-specific databases", or LSDBs, have several drawbacks that limit their utility: 1) size: most disease or gene-centric databases are very small, and many genes/genomic regions are not yet represented; 2) fragmentation: the data may be located in multiple, independent databases, and submission is often confined to a limited number of research laboratories; 3) lack of standardization: data submission and content is not standardized to allow comparison of data submitted across laboratories; 4) lack of curation: data is deposited without any, or with an incorrect, interpretation of the functional or clinical significance; and 5) access: the data is often not widely and freely available to the research and clinical communities.

Most sequencing of disease-associated genes occurs within clinical genetics laboratories, and many new genetic technologies, including genome-wide copy number analysis and whole genome/exome sequencing, are rapidly making their way into routine clinical use as primary diagnostic tests. Capturing these data from clinical laboratories presents a unique opportunity (and challenge) to collect, standardize and make available the large sets of high-quality data generated through the course of routine patient care. In this proposal, we will populate a standardized and centralized database of human genomic variation at NCBI, called ClinVar, integrating pathogenic and benign genomic variation of all types to enable the robust assessment of the association of variants to disease. Assisting NCBI with the ClinVar project through a dedicated and funded effort will be critical to its success, given that clinical laboratories historically do not submit data into databases without assistance, support, and incentives for the process.

Human genomic variation can be divided into two main categories: structural and sequence-level variation. Structural variation includes Copy Number Variants (CNVs), which are defined as deletions or duplications of large genomic segments [≥1 kilobase (kb) up to megabases in size] [1]. Numerous investigations in normal populations have shown that more than 20% of the human genome is subject to structural variation [1, 3].The great majority of these CNVs are less than 100 kb in size, common, and not associated with overt disease [4]. Rare, larger CNVs (>400 kb in size) that occur mainly *de novo* have been identified as a major cause of birth defects, intellectual disability, autism, and other neurodevelopmental disorders [5-7]. The primary clinical genetic laboratory test for this group of patients is now a chromosomal microarray (CMA), which has expanded upon the utility of the G-banded karyotype. The second major category of human genomic variation is sequence-level variation, defined as intragenic variants limited to a single gene. More than 5,000 genetic diseases due to defects in a single gene (referred to as monogenic diseases) have been identified (Online Mendelian Inheritance in Man, omim.org). In the United States alone, clinical diagnostic tests performed in CLIA-certified laboratories are available for more than 1,000 of these monogenic diseases (www.genetests.org).

With several hundred thousand clinical genetic tests now performed each year, there is an urgent need to capture this human genomic variation data in a centralized, standardized high-quality curated database that is freely available to the community. Such a resource would not only benefit clinical testing used for patient care, but would also provide invaluable data for researchers investigating human genomic variation, enabling them to build new knowledge from a robust foundation.

2. Description of the resource to be generated

The goals of this three year project are to build upon the experience and infrastructure developed by the International Standards for Cytogenomic Arrays (ISCA) Consortium (launched through an American Recovery and Reinvestment Act Grand Opportunity grant entitled "CNV Atlas of Human Development") and to develop similar clinical data sharing processes and infrastructure for sequence-level variants. This project will support the ongoing collection of a large and comprehensive number of whole genome and gene-specific variation datasets with associated phenotypic findings. Ultimately, all genomic variation and phenotypic data will be submitted to and accessible through the ClinVar database within NCBI. ClinVar will serve as the centralized database for this clinical dataset. All data will be freely accessible from ClinVar including the provision of efficient download capabilities to enable use in secondary tools and genomic analysis platforms. This project will result in a curated structural and sequence variant database from thousands of clinical cases from disease populations around the world. An ongoing community data-sharing effort will provide extensive, high-quality data to the basic science and clinical communities for use in exploring the biological and clinical consequences of human genomic variation.

3. Community support for proposed resource

As outlined in the strategic plan of the NHGRI [8], a substantial focus of genomics research will be its integration into clinical medicine. In anticipation of these developments, there is a well-recognized need for a centralized and curated genomic variation database. As described below, and in the Approach section, we have garnered broad support for data submission from 36 laboratories, as documented in the 72 included letters of support. The creation and use of such a database will allow standardization across laboratories for interpretation of the clinical significance and functional impact of genomic variants. This effort will in turn provide support for clinical genomics research and for increased quality of patient care. Indeed, the creation of such a database has been identified as one of seven critical projects for advancing the field of personalized medicine, as defined by thought leaders gathered at the Banbury Conference Center in 2010 [9].

The success of the ISCA Consortium project is indicative of the overwhelming enthusiasm for shared, curated clinical datasets that can be used to improve our understanding of genomic variation and its contribution to disease. As a result, the ISCA Consortium has quickly grown to a membership of more than 800 individuals, including laboratories, clinicians, genetic counselors, and other scientists. The online ISCA database of structural variation, which is displayed through dbVar and the ISCA website (www.iscaconsortium.org), is accessed daily by members. In the same way, there is an overwhelming community need for the availability of sequence-level variant data through a centralized database. As whole genome sequence analysis is increasingly incorporated into clinical research and clinical care, such large datasets will be critical to providing accurate diagnoses and prognoses. We have received letters of support from senior researchers supporting our efforts and documenting the extraordinary value of such a resource (see letters of support from Drs. Altshuler, Biesecker, Church, Eichler, Green, Kohane, Kingsmore, Lander, Lifton, Scherer, Seidman, Weiss, and many others).

4. Anticipated impact of the resource for biomedical research

The National Institutes of Health have funded grants that will lead to more than 70,000 patients having their genomes sequenced in 2012; more will be sequenced through other funding sources (Francis Collins, personal communication). The task of curating these genomes will be extremely challenging due to the limited availability of informative and accurate existing resources. In addition, clinical laboratories are poised to begin offering CLIA-certified whole genome sequencing (WGS) services, though the lack of a publically available, centralized, curated set of data correlating genotype and phenotype makes the use of WGS services in a clinical context extremely difficult. As such, our clinical research and medical care community will benefit greatly from the development of a human genomic variant database with clinical grade curation, arguably one of the most important resources needed by our genomics community today.

Impact on clinical laboratories: Although genomic variation has been identified as a major contributor to human disease, there are substantial gaps in our knowledge regarding the biological significance and clinical impact of individual variants on most disease phenotypes. There is an urgent scientific need for much larger, high-quality datasets from both normal and disease subjects [4]. As the technology for disease-targeted and genome-wide variant detection and assessment becomes widespread in clinical settings, there is a unique and timely opportunity to capture and mine large datasets generated through the course of routine patient clinical care. The availability of genomic variation data in a central database with expert curation is an invaluable resource that will significantly accelerate the understanding of variants. Ultimately, this data will enable clinical laboratories to write more informative patient reports leading to improved patient care.

Impact on the basic science community: We anticipate that this database will house data from hundreds of thousands of individuals with various phenotypes and will be invaluable in differentiating between benign and pathogenic variants and in defining genotype/phenotype correlations. In addition, the data infrastructure system we are proposing will allow powerful new research strategies to be designed which will aid in the identification of novel genes, allow more detailed characterization of known disease genes, and discover other genomic contributors to human disease. Gaining access to the spectrum of genomic variants causing disease will better define the functional activities of proteins and the biological pathways in which they operate. Only through the availability of large datasets will such studies be possible, resulting in a transformative impact on human variation research.

Impact on clinical research: The development of data standards and the facilitation of genotypic and phenotypic data submissions from clinical genetics laboratories will allow widespread data sharing for clinical research studies. The availability of such data will vastly improve the recruitment of patients and families into disease-based registries, clinical research studies, and clinical treatment trials that utilize genomic information.

Impact on clinical care and public health: Our limited knowledge regarding the potential clinical impact of genetic variants causes substantial uncertainty in the interpretation of genetic laboratory results, making it difficult to counsel families regarding the potential causal relationship of a variant to a disease phenotype. Inaccurate and uninformed results can limit the utility of genomic medicine and even harm patients when care is based upon incorrect interpretations of genetic data. Clear evidence of this critical problem was recently demonstrated through whole genome sequencing studies of normal individuals harboring variants that were reported to be pathogenic, yet the predicted phenotype was inconsistent with their current state of health [10]. A large, centralized database of genotype and phenotype information will greatly accelerate our ability to interpret laboratory results and will be invaluable in improving patient care. High-quality data will be released publicly on an ongoing basis through public genome browsers and commercial vendors, with efforts to develop “clinician-friendly” user interfaces to allow searches of data on cases similar to their own patients. Furthermore, both patients and society will benefit from the knowledge obtained from this project: families faced with a genetic abnormality will receive improved and more accurate genetic counseling, and detailed information about the etiology of specific diseases could lead to more targeted treatment approaches.

5. Rationale for developing a new database with support for multiple interfacing tools

As diagrammed in Figure 1, existing databases at NCBI containing genomic variation include dbSNP, dbVar, and dbGaP. Both dbSNP and dbVar are publically accessible databases that support accessioning, defining and tracking of variation against the genomic reference for sequence-level (<50 bases) and structural (>50 bases) variation, respectively. dbGaP allows a restricted access environment for datasets that are considered identifiable [11]. None of these databases serve the goal of ClinVar, which will support the capture of clinical assertions and the evidence behind them for all genomic variants in an unrestricted environment. Although ClinVar will be responsive to future community needs, it will not immediately provide a data de-identification and curation interface. As such, we will use two software systems, Cartagenia BENCH and MutaREPORTER (integrated with the MutaDATABASE project), which have each been optimized for use with structural and sequence-level variant data, respectively; there is a need for both tools.

Cartagenia BENCH captures genomic variants directly from raw data analysis software, and allows phenotypic information to be imported directly from local and hospital laboratory information management systems (LIMS), internal lab databases, or referrer portals. In this way, both genotype as well as phenotype submission are automated and require minimal effort from laboratory technicians, genetic counselors and referrers – an essential quality to guarantee a continued and consistent flow of routine findings into the ClinVar databases. This process also allows the data to be submitted in a controlled way, supporting de-identification and accessioning, while combining genotype with phenotype information. The MutaDATABASE project is an international effort supporting open access to a standardized and centralized database with the goal of placing all variant data from human disease genes into a freely accessible resource (www.mutadatabase.org), a mission consistent with ClinVar. To date, there has been significant interest and support for the MutaDATABASE project with more than 80 investigators and lab directors on a large advisory panel, over 100 gene curators signed up to curate over 500 genes (see Appendix), and over 14,000 variants already deposited in the database. We intend to harness this interest and support for the project and make use of the well-developed MutaREPORTER software designed to support community curation of gene variants. However, all data in MutaDATABASE, regardless of the current state of curation, will be submitted into ClinVar as the ultimate repository for the data. There is already ample precedent for databases at NCBI receiving data from, and collaborating with, non-NCBI groups including the Consensus Coding Sequence Project, Human Genome Variation Society, OMIM, ISCA, Human Gene Mutation Database, and Database of Genomic Variation.

6. Project components, including core technologies needed to develop the resource

The data to be included in the universal database created through this project will be acquired from genome-wide copy number analysis as well as both gene-specific and genomic sequence analyses. Data will be generated through multiple technologies, including chromosomal microarrays (comprised of oligonucleotide and/or SNP probes), targeted Sanger or next generation sequencing, whole exome and whole genome sequencing. Data submission will occur either directly to ClinVar, or via multiple software systems linked to ClinVar, including dbGaP, Cartagenia BENCH, and MutaREPORTER (see Figure 1). These software systems have been designed to support and automate the data submission process for busy clinical laboratories, in order to increase broad participation. Cartagenia BENCH, a software system currently used by the ISCA Consortium, has built an ISCA specific version of their platform to allow for “one-click” submission of genotype and phenotype data to the current ISCA database. This system automates the de-identification process, allows for the conversion of free-text phenotype information into Human Phenotype Ontology (HPO) terms, and associates the genotype and phenotype information for a particular case. The MutaREPORTER software enables automated submission of variants by laboratories using MutaREPORTER for variant assessment. MutaREPORTER also allows communication among collaborating labs and curators working on the same gene(s) through a community group system called MutaCIRCLEs. Data will be made publically available through several resources supported within NCBI, including ClinVar, dbVar, dbSNP, and dbGaP, as well as through MutaDATABASE for sequence-level variants.

7. Preliminary data supporting the approach

We will refer to the ISCA Consortium frequently throughout this grant because it is both an example of the type of resource we wish to construct and because we intend to leverage the success and existing infrastructure of ISCA to jumpstart the construction of our expanded resource. The history of the ISCA Consortium is directly relevant to this proposal because it demonstrates the creation of a publicly accessible database containing clinical data that would not have been possible without grant funding.

Dr. Martin is a founding member, along with Dr. David Ledbetter, of the ISCA Consortium. Other active ISCA leaders that are contributing to this project include: Faucett, Miller, Thorland, Aradhya, D. Church, South and Van Vooren. The development of this consortium was initially supported by a grant from the American College of Medical Genetics Foundation with the goals of improving the quality of patient care related to clinical array testing, developing standard guidelines for array interpretation and reporting, and enhancing CNV research opportunities through the development of a shared public database. An NIH Recovery Act Grand Opportunities grant (CNV Atlas of Human Development, 5 RC2 HD064525, Ledbetter, PI and Martin, Co-Investigator) was subsequently funded to support the development of the ISCA database and to make the data publicly available through an interface with NCBI.

A. Organizing key curators and data contributors: To establish the organization of the ISCA Consortium, there were a series of three small meetings, which included key personnel on the grant, prominent researchers and clinicians in the field, and database and bioinformatics specialists. A similar approach has already been initiated to organize the sequence-level variant component of this proposal. The combined structural and sequence-level variation personnel have been organized as shown in Figure 2.

B. Establishing standards and existing databases: The initial infrastructure for data submission for structural variants was put into place through the work of the ISCA Genotype and Database Committees (now combined into the Structural Variant Workgroup). This includes a data submission template for structural variant genotype and phenotype information with standardization of terms (see Appendix). A similar approach has begun for sequence-level variants, and a preliminary data element dictionary and data submission template has been established (see Appendix).

The structural variant database created by the initial work of the ISCA Consortium is housed at NCBI through dbGaP and dbVar. The study at dbGaP (www.ncbi.nlm.nih.gov/projects/gap/cgibin/study.cgi?study_id=phs000205.v2.p1) contains prospective data that includes raw genotype data along with phenotype information, and access is restricted to authorized users. Data limited to identified variants exists in dbVar. Two studies are available to the public in dbVar: nstd37 contains ISCA data submitted by the clinical laboratories and nstd45 contains the ISCA curated dataset. The ISCA Consortium has successfully recruited over 28,000 submissions to this public resource. However, ~28% of variants in the database are still classified as variants of unknown significance, demonstrating the need to expand upon this effort through additional data collection and evidence-based curation. The curation process will be described in detail in Specific Aim 3.

A sequence-level variant database is currently housed by the MutaDATABASE project (www.MutaDATABASE.org). This effort has enlisted broad support for data submission and curation as described in the aims of the grant; more than 14,000 variants in over 197 genes have already been deposited into the database. MutaDATABASE also has initiated collaborations with an extensive international group of laboratories. More than 100 curators have already volunteered to oversee curation of over 500 genes (see Appendix). MutaDATABASE is also collaborating with other existing databases such as the Leiden Open Variation Database (LOVD) [12] and the Human Gene Mutation Database (HGMD) [13]: all publicly-accessible HGMD variants have been integrated into MutaDATABASE. Existing curators for many locus-specific databases, including LOVD databases, have agreed to transfer the current content of their databases into MutaDATABASE, curate data within the MutaDATABASE system, and mutually exchange data between both databases on a regular basis. The software to support data curation, MutaREPORTER, which will be provided free of charge to curators, has been developed and is currently in use by laboratories to upload and curate variants in MutaDATABASE. All data collected in MutaDATABASE will be shared on an ongoing basis with ClinVar. In addition, the Cartagenia BENCH platform and MutaREPORTER are being aligned to optimize a streamlined workflow for both structural and sequence-level variation.

C. Developing ClinVar: At the time of this grant submission, ClinVar has not been launched into public use. However, significant planning and data collection has already begun and an abstract defining the system has been submitted for presentation at the ASHG/ICHG meeting in October 2011. Staff members at NCBI are developing the infrastructure and specifications for the system as described in Aim 2. A community of stakeholders has been gathered representing the major US clinical laboratories, and this group is interacting by conference call every other week to inform policies, standards and sustainability of ClinVar. Most members of this grant proposal are involved in that effort, along with additional laboratories and stakeholders. Data is being collected by NCBI with current submissions obtained from over 50 sources involving 3,211 genes. Over 38,000 variants have been collected from these sources with over 25,000 variants classified according to pathogenicity. Currently, the majority of these variants are derived from existing databases integrated with NCBI (21,629 variants from OMIM and GeneTests/GeneReviews). In summary, efforts are underway to build ClinVar as a centralized, curated public repository for all human genomic variation. Similar work was initiated by the ISCA Consortium for structural variants, and with this proposal, structural and sequence-level variation will be combined in a single database**.**

APPROACH

In the aims below, we will outline the steps necessary to develop standards and define content for the ClinVar database (Aim 1), facilitate submission of data into the database (Aim 2), and demonstrate model examples of evidenced-based variant curation (Aim 3). These coordinated efforts will lead to the development of a freely and publicly accessible clinical grade resource for interpreting the clinical consequence of genomic variants on human health. See figure 3 for an overview of activities.

Aim 1: Develop standardized formats for acquisition and submission of clinical genomic variation datasets.

Genotype and phenotype data standards will be developed to ensure the uniformity and integrity of the data and to facilitate de-identification and data transfer to the ClinVar database. Such standards are necessary for data quality, confirmation, and classification of structural or sequence variants as pathogenic, benign, or of uncertain clinical significance.

A Policies, Standards, and Sustainability Workgroup (PSSW) will be chaired by Drs. Sherri Bale and David Ledbetter. Membership on the PSSW will include the Executive Committee, those individuals who are leading one of the model curation projects, a researcher in translational genomics (Dr. Robert Green), an ELSI expert (Ms. Joan Scott), and an IT specialist (Mr. Sandy Aronson). Other individuals who are serving on the other workgroups, as well as our consultants, will be included as needed. Figure 2 shows the complete organization of key personnel, workgroups, and consultants.

The PSSW will meet in person a minimum of twice per year throughout the three years of the grant. To minimize travel costs, one meeting will be held at a dedicated time, while additional meetings may be held in conjunction with the national meetings of the American College of Medical Genetics and the American Society of Human Genetics. Additionally, there will be a monthly conference call for all committee members and relevant consultants.

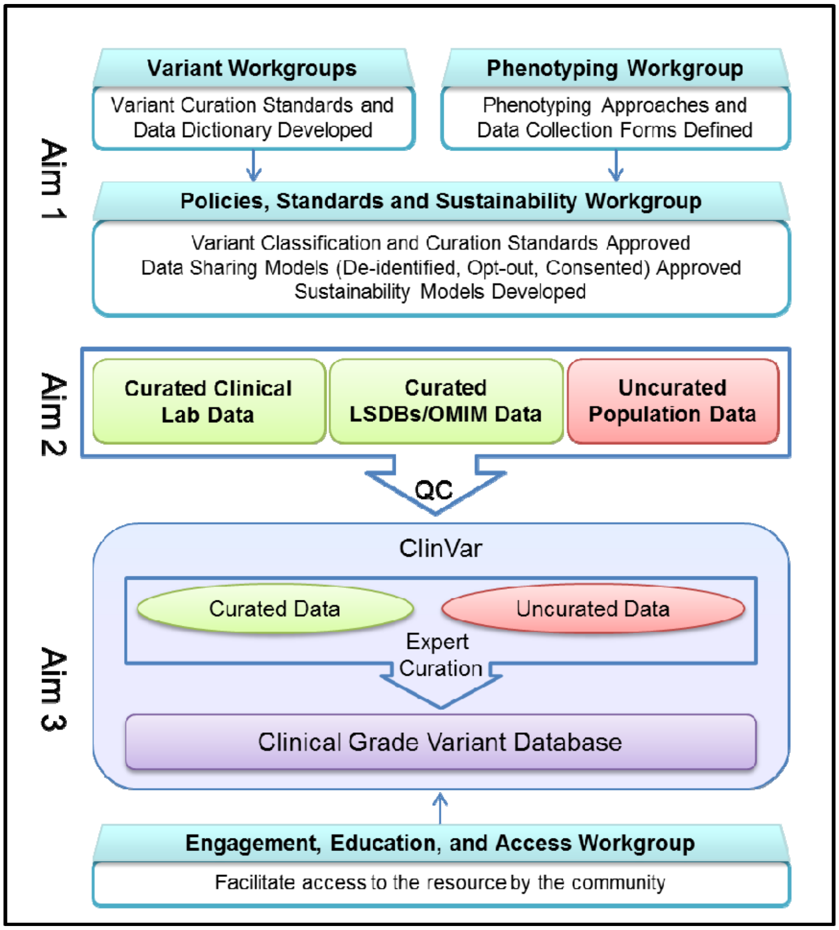

Aim 1
Variant Workgroups
Variant Curation Standards and Data Dictionary Developed
Phenotyping Workgroup
Phenotyping Approaches and Data Collection Forms Defined
Policies, Standards and Sustainability Workgroup
Variant Classification and Curation Standards Approved
Data Sharing Models (De-identified, Opt-out, Consented) Approved
Sustainability Models Developed
Aim 2
Curated Clinical Lab Data
Curated LSDBs/OMIM Data
Uncurated Population Data
QC
Aim 3
ClinVar
Curated Data
Uncurated Data
Expert Curation
Clinical Grade Variant Database
Engagement, Education, and Access Workgroup
Facilitate access to the resource by the community

Figure 3. High-level overview of proposed activities

1A. Policies

Of the several responsibilities of the PSSW, the first to be addressed will be development of the policies surrounding data acquisition, submission, and public access. This task will include defining the mechanisms for different types of data sharing that insure the protection of patient privacy. It will build upon existing policies, as defined below, which were developed by the ISCA Consortium. The policies will be followed by both the structural and sequence-level efforts.

Data will be of two kinds: 1) data that are not considered identifiable and can be submitted into a public database and 2) data that are potentially or clearly identifiable. In the first case, individual variants or small sets of variants (e.g. compound heterozygous variants in a single gene or independent calls from cytogenomic microarrays) will be considered non-identifiable; however, large datasets (e.g., genome-wide raw data associated with chromosome microarrays, whole genome or exome sequencing or large next generation sequencing panel tests) will be considered identifiable. In these cases of potentially identifiable data, data will be submitted into dbGaP, a controlled access database at NCBI. For submission of such datasets, the laboratory will need to have an IRB protocol in place to manage at least an opt-out model, or possibly full consent, depending on the evolving policies of patient privacy and protection.

The consenting process for patient data collection will be tailored to fit the risk for a breach of privacy associated with the data collection approach. The specific consenting details will be adjusted based on the potential for personal identification of the patient via their data. In the initial phase of collecting sequence-level variants, data will be obtained from laboratories primarily on patients who have undergone single gene mutation analysis or gene panel testing. The patient data will consist primarily of a limited phenotypic dataset provided at the time of test order and will not prompt a new consenting process for the patient. For data that could be considered identifiable, other approaches, as described below, will be supported.

Opt-Out Consent Model: Currently, the ISCA Consortium successfully uses an opt-out method of consent for the submission of raw data files to dbGaP. This method will be utilized for data collection of all large genomic datasets from structural or sequence-level assessments. The opt-out model was developed by the Collaboration, Education and Test Translation (CETT) Program of the NIH Office of Rare Disease Research [14]. It was reviewed and approved by the NIH Office of Human Subjects Research (OHSR) and has currently been approved by 5 academic IRBs for use in the ISCA Consortium. The opt-out model is considered appropriate for these purposes for several reasons: 1) Re-contacting patients to obtain formal consent is not possible for most clinical laboratories and not feasible for the scope of this project. Often, patient contact information is not provided to clinical testing laboratories, making it impossible to contact patients directly. Further, managing the full consent process of thousands of potential participants around the world would hinder the collection of a robust dataset. 2) There is little risk to participants. Though the raw data files generated during genomic testing are theoretically identifiable, this would be highly unlikely in practice. The raw data files are kept under controlled access, and only researchers with IRB approved protocols may apply for and be granted access to this data. 3) There are ample opportunities for patients to learn of the project and opt-out of participation. Descriptions of the project and the opt-out process in lay-friendly terms are on test requisition forms, clinical result reports, and on participating laboratory websites. Patients have the opportunity to opt-out of participation by simply checking a box on either the requisition or test result and returning it to the laboratory, calling a toll-free number, or submitting a short form through the laboratory website.

The opt-out model also provides a process by which researchers can re-contact participants with multiple layers of protection to reduce possible identification. Interested researchers may contact the submitting laboratory to discuss potential research opportunities; the laboratory then contacts the referring clinician, who contacts the patient directly. If the patient is interested in the opportunity, he/she can then contact the researcher directly and consent to the study.

The recent publication by the Department of Health and Human Services of an *Advanced Notice of Proposed Rule Making (ANPRM)* raises several issues about the use of data collected in clinical situations for future research (test results linked to clinical information) [15]. The ANPRM proposes that consent may be required, but that a general consent may be possible. The ANPRM also raises the issue that advances in genetic technologies may make de-identification difficult and may make clinical data re-identifiable. Furthermore, ANPRM recognizes the barriers posed by multiple individual IRB consents for multisite consortia such as the one proposed here. There is a strong recommendation that multisite studies be covered by a single lead IRB, with the IRBs at other sites consenting to defer to the lead IRB's review process. In this project, we will monitor efforts of the ANPRM and hold discussions guided by our ethics consultants, with the laboratory, research, and patient advocacy communities about how to continue to support robust data submission from many different sites and still provide adequate protections.

Opt-In Consent Models: Other phases of the project will pursue the collection, submission or linkage of clinical data outside of the testing process and associating that data with a genotype (ranging from one variant to full genomic datasets). These data engender a greater potential risk for breach of privacy based upon deeper clinical datasets and the need to engage more directly with patients outside of a testing process. In these cases, we will work directly with researchers and patient advocacy groups, who will interface with the patients, to construct appropriate consent processes. In addition, we will encourage clinical laboratories offering whole genome or exome studies to offer an opt-in consent process to allow full use of these datasets. Such a process is already being developed in Dr. Rehm's Partners Laboratory for Molecular Medicine as well as in the Geisinger Health System enabling the use of CLIA-certified whole genome sequencing analyses in conjunction with patient re-contact in order to obtain robust phenotypic data.

1B. Data to be collected

The PSSW will also define which information is collected. The types of data to be collected include both genotype and phenotype data.

For structural variation, the project will focus on genome-wide data from individuals with developmental or intellectual disabilities and/or congenital anomalies. A standard format for genotype data, along with standardized variables, has been established by the ISCA Consortium. All CNVs from an individual will be captured in the ClinVar database, with the number of CNVs ranging from 0-20+ variants per patient. Each CNV will be classified into one of 5 categories (pathogenic, likely pathogenic, uncertain, likely benign, or benign) [17]. Phenotype information, using a standardized HPO terminology, will also be incorporated with the genotype information.

For sequence variation, the ad-hoc ClinVar Community Call Project (a group of approximately two dozen stakeholders including molecular diagnostics laboratory directors, researchers, and staff of the ClinVar/NCBI program) has developed a preliminary set of required and optional variables to be provided with submission of sequence-level variants in individuals referred for diagnostic, carrier, or pre-symptomatic testing for inherited disorders or from healthy cohorts. A preliminary data element dictionary has been developed (see Appendix) and two groups have submitted data. However, the data dictionary will be refined through real-time use and subsequent versions will be released based upon that experience and continued input from the community. Due to the different internal databases maintained by each collaborating laboratory, a minimum list of required variables sufficient to unequivocally define each variant and its phenotypic source must be determined. Additionally, a complete set of data elements will be defined to enable the richest dataset possible. This dataset will include clinical assertions, data on variant frequency in cases versus controls, detailed phenotypic information, variants found in cis/trans, publications, segregation data, functional data, analytical method, source of data, and many other attributes as defined in the data dictionary (see Appendix). Although not yet initiated, we will also work on defining the overlapping, yet distinct requirements for annotating the evidence associated with somatic changes identified in tumor samples.

For both sequence and structural variation, the submitting clinical laboratories will be asked to include their pathogenicity call for each variant, as well as information about how they came to that interpretation (i.e. population studies, family segregation studies, in silico analyses, functional studies, etc.). Variants identified in patients with clinical disease, but which are interpreted by the submitting laboratory to be benign, will also be collected. Each individual observation submitted to the database will include an attribution showing the laboratory that submitted the data, as well as the year in which it was submitted to the database, documenting the point in time when their clinical assertion was made. Each submission will also record the method used to identify the variant. All data entry fields to be used will ensure capture of all potentially available data on a variant as well as encourage laboratories to capture the data going forward for fields they do not capture initially. Outputs from the database will include individual observations of each variant, summary lists showing frequency of individual variants, and overview tables of all variants.

Phenotype data, in contrast to the core elements of variant data, are very heterogeneous and vary significantly among genes/disorders. Phenotypic observations may be more subjective and variable and dependent on the observer/reporter. It will therefore be necessary to define the origin of phenotypic data at the time of entry. We anticipate the following sources of clinical data, in order of increasing likelihood of providing accurate and comprehensive phenotype data:

1. Disease scope of test
2. Overall diagnosis or testing indication provided by ordering clinician
3. Clinical data collection form, using a predefined ontology specific to a disease area, provided at the time the test is ordered or before reporting test results (see samples in the Appendix for Noonan Spectrum Disorders and Hearing Loss)
4. Data mining of electronic medical records (EMR)
5. Retrospective entry or linkage of phenotype data sets to genetic records through consented patients (supported by clinician researchers or patient support organizations who harbor patient data)

Phenotypic data collection will occur through multiple mechanisms (see “Sources of Phenotypic Data” in Aim 2), and a policy on data collection will guide that process. Clinical laboratories submitting minimal data, such as test indications, will be able to upload retrospective data in batches. This type of data can easily be uploaded as a “.csv” or comma-separated value file by the submitting lab and/or the project’s IT support will assist labs in converting their data into a usable format. More detailed phenotypic data will be collected through Clinical Data Collection Forms and web access to MutaDATABASE, Cartagenia, and ClinVar. Laboratories and clinicians possessing more detailed phenotypic information will be able to submit data by Case Report Forms which will be tailored to each disease entity/gene using HPO terminology. Laboratories and clinicians submitting data can also access the disease-specific HPO-based ontology (approved by a gene curator) through MutaREPORTER as a web-based interface for data submission.

1C. Standards

Another critical mission of the PSSW will be to define the **nomenclature** and **standards** for the classification of structural and sequence variants (genotype) using the clinical, biological, biochemical and pathological features (phenotype) reported with the variants, including quality control. This task will be carried out in conjunction with the curation and evidence-based systems developed in Aim 3, which will help guide the standardization of variant classification.

To facilitate a more efficient classification of structural variants, the ISCA Consortium recognized that standardization of array design was a first step in this process. Through a consensus process, the ISCA group recommended that microarrays should have a minimum resolution to detect imbalances of at least 400 kb throughout the genome [16]; this recommendation was echoed in subsequent ACMG guidelines for the interpretation and reporting of postnatal constitutional copy number variants [17].

For development of the nomenclature and evidence-based criteria for variant classification, the PSSW, along with the Structural and Sequence Variant Workgroups, will meet to develop these criteria. For example, these discussions will address how to incorporate various types of evidence into the classification of variants including, case versus control frequencies, proband clinical data, family history, parental studies, segregation analysis, *in silico* analytic methods, etc. There are currently several publications with proposed standards, however they have not yet gained wide acceptance. These include:

- The ACMG standards and guidelines for interpretation of sequence variations [18],
- The recommendations proposed for classifying cancer variations [19]
- The recommendations proposed by the Clinical Molecular Genetics Society (http://cmgsweb.shared.hosting.zen.co.uk/BPGs/Best_Practice_Guidelines)
- The ACMG standards and guidelines for interpretation and reporting of postnatal constitutional copy number variants [17]

Our group has developed a consensus proposal for variant classification based upon these guidelines. At the start of funding, we will solicit additional feedback on this proposal through a survey to all US clinical laboratories using contact information from the GeneTests registry. We will also engage with other communities for wider feedback including the American College of Medical Genetics (ACMG), the College of American Pathologists (CAP), 1000 Genomes project [20], and the Human Variome Project (HVP) [21] (see letters of support from Wayne Grody-ACMG, Stanley Robboy-CAP, David Altshuler-1000 Genomes, and Richard Cotton-HVP). After a consensus is reached, we will work with both the ACMG Laboratory Quality Assurance Committee (Dr. Rehm is a member) to update their standards and guideline document (see letter of support from Sue Richards, Chair, ACMG Lab QA Committee), as well as with regulatory agencies such as CAP and CLIA to request that the standards be enforced through the clinical laboratory accreditation process (see letter of support from Stanley Robboy, President of CAP). The standards will also be posted on the ISCA Consortium, ClinVar, and MutaDATABASE project websites as well as an integrated website for this project which will be developed.

The Phenotyping Workgroup will define, review, and modify the standards for collection, storage, curation, and access to phenotypic data. To facilitate uniform phenotypic data collection, phenotypes, diseases and clinical terms will be described using standard ontologies wherever possible, though accommodation must be made for conditions that have no standardized term. An ontology is a computational representation of a domain of knowledge based upon a controlled, standardized vocabulary for describing entities and the semantic relationships between them. The terminology for phenotype ontology will come preferentially from two sources: the Human Phenotype Ontology (HPO) [22] (http://www.human-phenotype-ontology.org/index.php/hpo_home.html), as established by Dr. Peter Robinson, a consultant to this project [22], and SNOMED CT (Systematized Nomenclature of Medicine--Clinical Terms; http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html). Entries from UMLS, MeSH, OMIM, ICD-9 and ICD-10 will also be acceptable.

The HPO was initially developed using information from Online Mendelian Inheritance in Man (OMIM), which is relevant to the vast majority of disorders for which clinical testing, and therefore genotype-phenotype data, is available. The HPO contains over 50,000 annotations to hereditary diseases, and is available for download or can be browsed using PhenExplorer (www.human-phenotype-ontology.org/PhenExplorer/PhenExplorer). SNOMED-CT is a comprehensive clinical terminology, originally created by the College of American Pathologists (CAP) and, as of April 2007, owned, maintained, and distributed by the International Health Terminology Standards Development Organisation (IHTSDO), a not-for-profit association in Denmark. The NLM is the U.S. Member of the IHTSDO.

As proof-of-concept for this approach, PhenExplorer is already used by the ISCA Consortium for data entry into dbVar and dbGaP, and this program will also be adopted by the ClinVar database. Secondly, electronic versions for capturing essential phenotype information for both prenatal and postnatal assays have been implemented into the Cartagenia BENCH data submission tools. Our prior experience working with Cartagenia on the ISCA project helped define the standards for data collection. For example, we dealt with issues of multiple phenotypic codes being submitted for the same patient and learned that phenotype annotations for patient records should never be coded with multiple codes for the same phenotypic trait: rather, if multiple terms from the ontology define the same concept, the ontology itself should be updated. Additionally, Cartagenia has developed algorithms to automatically detect HPO classifiers from free text, such as Reason For Referral (RFR) text, and patient case report descriptions. This Cartagenia algorithm includes "ICD9 expansion", a feature which allows frequently used ICD9 codes to automatically be mapped to HPO classification codes. This functionality has improved the quality of data collected, as ICD-9 codes were frequently the only information available. Secondly, this feature has allowed us to analyze which phenotypic descriptions are actually used by clinicians and referrers, information that we have used to modify the phenotype forms and interfaces, allowing for their increased utility. We will continue to focus on expanding the utility of these applications in the phenotype data collection process. Copies of the phenotype forms developed for the ISCA Consortium are included in the Appendix.

MutaREPORTER also uses HPO and gene-specific common features derived from PhenExplorer. A set of phenotype fields has been developed for all human disease genes through an automated process using OMIM as a source of data to define related HPO terms for each gene. While the HPO terms extracted from OMIM provide an excellent resource for detailed phenotyping of human genetic disorders, the automated nature of the process can result in omission of appropriate features or incorrect attribution of other features. For these reasons, the HPO-based ontologies will be a starting point, and expert curators will modify as appropriate. For example, authors of GeneReviews will be asked to review the Human Phenotype Ontology and Clinical Data Collection Forms for their condition/disorder and revise as appropriate. Recognition of such participation can be provided within the GeneTests website as an incentive. Expert curators will also be asked to provide guidance on development of case report forms. Also, for challenging situations, such as similar terms that create confusion within the ontologies, there is an online issue tracker for HPO so that a member of the HPO team can respond to requests from curators. Peter Robinson, consultant to this project, has agreed to oversee this effort and states in his letter of support "it is important that all efforts to collect phenotypic data for genetic disorders adhere to a common language to enable the robust sharing and understanding of these powerful datasets".

1D. Sustainability

A third task of the PSSW will be to identify strategies for the long term sustainability of ClinVar. Building upon the efforts of the ad-hoc ClinVar Community Project's Sustainability Committee (Sherri Bale, Stephen Kingsmore, David Margulies, David Ledbetter), the PSSW will work to 1) anticipate the uses of the database by the larger community, 2) suggest an organizational structure for the database that is amenable to its future growth and development, 3) determine what possible sources of revenue might be available for the long term maintenance of the database and its ongoing curation, and 4) interact with other types of organizations (patient support groups, other not-for-profit entities, professional organizations, for-profit public entities, etc.) to see how they may be supportive of the database initiative going forward.

Aim 2: Coordinate the submission of variant and phenotypic data into ClinVar, a centralized database at NCBI.

The ultimate goal of this grant is to support NCBI's ClinVar database as the primary site for deposition, curation, and maintenance of all human genomic variation. In his letter of support, Jim Ostell from NCBI states: "NCBI is happy to work with this collaboration both to provide access and tools to use the public data at NCBI, and to incorporate the data they produce and their recommendations for its structure into the public resources at NCBI." ClinVar has been developed by NCBI to "provide a freely accessible, public archive of reports of the relationships among human variations and phenotypes along with supporting evidence" (http://www.ncbi.nlm.nih.gov/clinvar/intro). ClinVar will support the submission of variant observations by laboratories, either as summarized variant data or individual case data, with each submitter explicitly acknowledged. The system will also enable the submission of assertions made regarding the clinical significance of variants as well as the evidence to support those assertions. ClinVar will present the data for individual users, as well as support laboratories and other organizations that want to efficiently incorporate the data into their own applications. Human variations are reported to the user as sequence changes relative to an mRNA, genomic, and protein reference sequence. Genomic sequences will be represented in RefSeqGene/Locus Reference Genomic (LRG) coordinates, as well as in chromosome coordinates. ClinVar tracks disease associations through the use of standardized medical terminologies from sources including SNOMED CT, GeneReviews, Genetic Home Reference, Office of Rare Diseases, MeSH, OMIM, and HPO.

2A. Data Submission Methods

Aim 2 will support the submission of numerous sources of genetic and phenotypic data into centralized locations. Several data flows will be supported depending on the type of variants, the level of identifiability of the data, and the choice of the submitter for where they want to submit their data. Figure 1 shows a diagram of the overall data flow. The major systems used for data submission, phenotype enrichment, and data curation are described in the paragraphs following, including details on functions and quality control checks.

For structural variants, the infrastructure for data submission and storage has been established through dbGaP and dbVar at NCBI. Multiple array vendors, including Agilent Technologies, BlueGnome Ltd, and Oxford Gene Technology, have also incorporated the standardized data form into their software such that genotype data is in the proper format for direct submission to the ISCA database. In addition, the ISCA Consortium selected the software company, **Cartagenia**, to provide another source for laboratories to use to facilitate the transfer of data to the NCBI database. Cartagenia utilizes a flexible and extensive web-based database and software platform for managing genotype and phenotype data for patients and study subjects, which has been adapted to the needs of the ISCA Consortium through the creation of the ISCA BENCH platform. ISCA BENCH includes features such as security measures of accountability and HIPAA-compliance, de-identification of patient identifiers, and transparent linkage of genotype and phenotype information. In addition to genotype information, Cartagenia has incorporated the ISCA one-page phenotype form (see Appendix), which uses HPO terms to describe the patient’s phenotype, into the ISCA BENCH platform. The Cartagenia software has been installed in multiple laboratories and automatically performs the de-identification and transfer of the data from the clinical laboratory to the ISCA database at NCBI. For cases with the appropriate consent mechanism, raw data files are included in the submission to dbGaP. The raw data includes two files with the following information: 1) the intensity values from every probe on the microarray and 2) normalized $\log_2$ ratios from the array. The raw data files are stored in a controlled-access database at dbGaP. Once quality control checks are completed at dbGaP, dbGaP transfers the CNV data (excluding raw data) to dbVar for public access. Once the ClinVar database at NCBI is fully developed, data flow will incorporate ClinVar as the primary archive of curated clinical assertions, while dbVar will continue to provide accessions for the primary variant definitions linked to phenotypic context.

For sequence-level variants, data can either be submitted directly to NCBI for import into ClinVar or use the **MutaREPORTER** software which enters data into MutaDATABASE. In turn, MutaDATABASE will have bidirectional monthly data exchange with ClinVar. This will allow laboratories to choose either site for submission, yet enable curators of the data to have full access to all data. Likewise, we will support the integration of existing locus-specific databases into MutaDATABASE and/or ClinVar, and engage the curators of those databases to use MutaREPORTER software (provided free to curators) for their curation activities.

The MutaREPORTER software is a Flex-based Web 2.0 application that includes a genome browser with comprehensive, user-customizable, track-based views of the genomic sequence, gene structure, protein sequence, and conservation. Available tracks are arranged in a track library and can be rearranged, added, and resized. The genome browser allows automatic numbering of genomic, cDNA, and protein locations, and shows the localization of variants in the genome. For batch imports a standardized form with defined data fields has been prepared in collaboration with NCBI/ClinVar. Information can be submitted from various sources, including other variant databases and laboratories (Figure 1). IT support will be provided (through partial funding of the submitting lab staff or though centralized support staff) to assist the collaborating laboratories with extracting, converting, and submitting their data. A variant QC step checks whether the wildtype base exists in the reference sequence and whether Human Genomic Variation Society (HGVS) nomenclature has been used appropriately to describe the variant. The system also ensures that a minimum set of data (defined below) is included in each submission. The software provides facilitated access to prediction tools including PolyPhen, SIFT, CONDEL, and splice site prediction tools. The software enables communication among collaborating laboratories and curators working on the same gene(s) through a system called MutaCIRCLEs. Curators are given free access to the software while others will need to purchase the software for a licensing fee of $1,000 per year. We will ensure, however, that license fees do not inhibit data submission. As such, we will allow laboratories to send data in spread sheets to curators who will enter the data into the system themselves, or laboratories can directly submit data to NCBI for import into ClinVar which will flow into MutaDATABASE for curation support. In addition, we are under discussion with MutaBASE about providing a scaled down version of MutaREPORTER, like ISCA BENCH, which can be provided to laboratories free of charge to support de-identification and data submission.

Both the Cartagenia and MutaREPORTER software applications create unique IDs for case submission, with links to laboratory IDs which are maintained only by the submitting lab. This system enables re-contact under the opt-out model described in Aim 1 and ensures that data are not sent in duplicate from laboratories. ClinVar will also prevent duplicative variant submission from multiple databases tracking published variants by connecting all data to PubMed IDs; duplicate entries will be avoided if any PubMed IDs match.

2B. ClinVar Quality Control and Versioning

Structural and sequence variants submitted to ClinVar are mapped to reference sequences and annotated according to HGVS standards. ClinVar assigns a version number to all data submissions, so when submitters update their records, the previous version is retained for review. Updates in content will happen on a daily basis with semi-annual or annual changes in data definitions and backward compatibility in data reporting. The system also ensures that a minimum set of data is included in each submission. The level of confidence in the accuracy of assertions of clinical significance depends in large part on the supporting evidence, so this information, when available, is collected and visible to users.

In addition, several quality control checks have been built into the system: a check that the variant description conforms to HGVS nomenclature; identification of conflicts among submitters and the published literature or prior submitted data; validation of the phenotype relative to controlled vocabularies and current understanding of gene to phenotype relationships. The quality control checks will be implemented in a manner to minimize barriers to data submission yet ensure high quality data. As such, certain checks will prevent data submission, whereas others will allow data submission but render a warning output to the submitter. For example, variant descriptions that are not resolvable on a reference sequence will not be allowed, whereas discrepancies in variant classification will be provided in a discrepancy report. This latter quality control check of the variant classification system has been in place for several years in support of the ISCA project and has been highly successful in improving the quality of structural variant interpretations. Approximately 5-10% of variants submitted to the ISCA database have been flagged to the submitting laboratory due to a discrepancy in variant classification. Thus, this type of curation process has already demonstrated enhanced quality control for clinical reporting.

2C. User Support for Software Applications

Cartagenia, funded in part by this grant, will work directly with individual laboratories wishing to submit retrospective data and/or, prospective data, from their routine diagnostic workflow. Cartagenia will also provide support for bulk data transfer automation, training, and setup of the variant and phenotyping submission software packages, as well as general maintenance and user assistance.

The MutaREPORTER software is supported by MutaBASE, a joint venture between GENDIA (www.gendia.net), a genetic diagnostic network, and Genohm (www.genohm.com), a bioinformatics company. All genetic data questions regarding the software will be dealt with by employees of GENDIA (Patrick Willems and Kristel De Boulle), who will be funded in part by this grant, while all technical IT questions will be dealt with by Genohm. Genohm has a large track record in user support service and their CSO (Martijn Devisscher) will be responsible for all IT related problems and user support service together with Hilde Dierckx, both funded in part by this grant. A collaboration is also under discussion between MutaBASE and Cartagenia to streamline support for this project through a common process.

NCBI also has a robust system for registering questions from users and tracking responses. For example, the RefSeq/RefSeqGene group receives about 25 requests each week, and responds within no more than 2 business days. Users often take time to acknowledge the rapidness and helpfulness of the responses. FAQs are generated as appropriate. ClinVar has already established the infrastructure for extending that service (clinvar@ncbi.nlm.nih.gov) and is committed to providing the same level of service for this resource (personal communication, D. Maglott).

2D. Minimum Data Set and Data Quality Thresholds

The ClinVar, ISCA, and MutaDATABASE groups have been working closely together to ensure consistent alignment of data fields enabling seamless and robust data exchange. Detailed data dictionaries have been developed to support submission systems and a minimum dataset has been agreed upon. The minimum dataset includes an unequivocal genomic location, a minimum amount of phenotypic information, the method used to detect variant, zygosity, the submitter source, and for clinical labs, the clinical assertion about each variant. In addition, given the increasing amount of DNA sequencing data from next generation sequencing platforms, it is clear that an analytical quality threshold will need to be applied to ensure that variant submissions represent accurate variant calls. Although the community has not yet come to consensus on a universal standard, several of the investigators on this grant (Drs. Bale, Das, Ferber, Hegde, Lyon, Rehm, Thorland, South, Kearney, Aradhya, and Martin) are involved in professional efforts to define these parameters and will ensure their incorporation into the standards for the ClinVar and MutaDATABASE projects. We anticipate setting a minimum threshold for data inclusion as well as collecting specific quality thresholds on variant submissions to document the confidence level for each variant.

2E. Sources of Variant Data

Variant data will be submitted to or viewable from a variety of sources including clinical laboratories, research laboratories, locus-specific databases, OMIM, HGMD, GeneReviews and dbSNP. For the structural variant project, there are currently over 150 laboratories participating in the ISCA Consortium and data collection will continue with these laboratories as well as new laboratories. We anticipate enrolling approximately 42,000 more cases during the 3 year funding period based upon the rate of accrual to date.

The sequence-level project will begin with data submitted by the clinical laboratories listed in Table 1. To date, 36 laboratories have agreed to submit data from over 160,000 cases. Most laboratories have only been queried for submission to the 8 model curation projects defined in Aim 3. As such, we anticipate a much larger data set once the project expands to other gene sets. As soon as the 8 model projects are initiated, work will be expanded to solicit and support data for all gene variation, including both somatic and germline changes. The initial target for expanded gene sets will be to focus on supporting the over 100 curators who have agreed to submit and curate data on over 500 genes through the MutaDATABASE project (see Appendix).

For existing general mutation databases, several models of data transfer will be supported. For OMIM and GeneReviews, all variant data has and will continue to be submitted into ClinVar on a routine basis. All dbSNP entries will be viewable through ClinVar and MutaDATABASE, while maintained in dbSNP.

For locus-specific databases (LSDBs), existing datasets will be imported into ClinVar. In addition, several LSDB owners have agreed to shift their projects to operating within either MutaDATABASE or ClinVar as a primary site of function, and we will continue to encourage as many databases as possible to merge with this effort. The curators for those databases would then curate data using MutaREPORTER software or submit updated interpretations directly to ClinVar. Alternatively, other LSDBs plan to maintain a separate database, yet are willing for data to be exchanged on a regular basis (e.g. InSiGHT for hereditary colorectal cancer – see letter of support from Finlay Macrae). The latter approach may be necessary in disease areas where identifiable clinical information is being maintained with genetic data (e.g. Parent Project Muscular Dystrophy). All models of data sharing or transfer will be supported as long as they increase the community's access to data.

2F. Format and Sources of Phenotypic Data

The phenotypic data submission format will consist of general variables such as age/date of birth, gender, race/ethnicity, type of testing (symptomatic vs. asymptomatic), proband versus family member, and identification of the origin of clinical data. More disease-specific and/or gene-specific data will follow the format

Table 1. Laboratories agreeing to submit sequence-level data

Laboratories	**HCM**	**Noonan**	**HCrC**	**Metabolism**	**DevDelay**	**CMD**	**PTEN**	**ZEB2**	**Other**	**Cases**
Ackerman Lab, Mayo	1000								(LongQT)	**1000**
Alfred I Dupont Hospital for Children		488			138					**626**
All Children's Hospital St. Petersburg					TBD					**TBD**
ARUP		121	TBD	500	670		179		3800	**5270**
Athena Diagnostics		TBD								**TBD**
Baylor Medical Genetic Laboratories		TBD			17000					**17000**
Boston University		TBD			TBD					**TBD**
Children's Hospital Boston		TBD								**TBD**
Children's Hospital of Philadelphia					623			8		**631**
Children's Mercy Hospital, Kansas City, MO					TBD			100	604	**704**
Cincinnati Children's Hospital				538						**538**
City of Hope Molecular Diagnostic Laboratory					TBD					**TBD**
CureCMD						475				**475**
Detriot Medical Center					TBD					**TBD**
Emory University		395	195		700	253	255	80	8283	**10161**
Fullerton Genetics Laboratory					TBD				TBD	**TBD**
GeneDx	2018	2300		727	400		4023		TBD	**9468**
Genomic Medicine Institute, Cleveland Clinic							TBD			**TBD**
Greenwood Genetics		695			220		275			**1190**
Henry Ford Hospital				27						**27**
InSiGHT			25000							**25000**
LabCorp/Correlagen	1000	TBD	TBD						5500	**6500**
Mayo Clinic			9000	1450	945					**11395**
Mt. Sinai School of Medicine		193								**193**
Nationwide Children's Hospital		475			TBD					**475**
Nemours Biomolecular Core, Jefferson Medical College		348								**348**
Oregon Health Sciences University					TBD					**TBD**
Partners Laboratory for Molecular Medicine	3900	2426	10						53117	**59453**
Quest Diagnostics			TBD	TBD	TBD					**TBD**
Transgenomics	1000									**1000**
University of Chicago					3215			46	5904	**9165**
University of Nebraska Medical Center		124			TBD					**124**
University of Oklahoma		107								**107**
University of Sydney					720					**720**
Women and Children's Hospital					100					**100**
Wayne State University School of Medicine					TBD					**TBD**
Cases Per Disease Area	**8918**	**7672**	**34205**	**3242**	**23911**	**728**	**4732**	**234**	**77208**	**160850**

of SNOMED-CT terms and the terminologies for that gene/disorder will be identified through the ontologies of HPO that have been generated based on OMIM. Data collection formats, such as Case Report Forms, are described in Aim 1. Examples of such forms can be found in the Appendix, one being used for the ISCA Consortium and two others being used by individual laboratories for specific disorders. For example, phenotype forms are provided for hearing loss and Noonan syndrome spectrum; these forms will be modified for this project to conform to the ontologies being developed in Aim 1.

There will be three main sources of phenotypic data: laboratories, clinicians, and patient groups. Collection of phenotypic data on vast numbers of patients faces two main challenges. First, there are consent issues, as addressed in Aim 1, and the rules for consent may vary among different local IRBs representing a challenge to acquiring detailed data. Second, the level of detail for this type of data scales proportionately with the effort that individuals (i.e. clinicians) must make to provide the time and expertise to identify and consent patients, acquire phenotypic data (e.g., through exam or chart review), enter/record data in a database, and curate the data. Clinicians who recognize the value of this resource may be more willing to contribute to this effort, and the Engagement, Education, and Access Workgroup (EEAW) will attempt to increase participation in this way. We recognize the challenge of sustaining the level of community effort to achieve deep phenotyping across all genes/diseases. However, we also recognize that even limited phenotypic datasets will be extremely useful in assessing whether variants are pathogenic even before more detailed genotype-phenotype studies are attempted through richer datasets.

Collecting a minimal amount of phenotypic data, such as test indications from a clinical laboratory, is more tractable, and this type of data is already being collected and will be submitted. Clinical laboratories performing chromosomal microarray (CMA) and testing for the eight model curation projects outlined in this grant will provide this type of phenotypic data at a minimum as proof-of-principle for this mechanism of data capture; the same mechanism can be modified and expanded to accommodate additional laboratories across many other genes/disorders. Additional laboratories will also be recruited through the efforts of the EEAW. From our experience, the level of phenotypic data provided to clinical laboratories is minimal, and consists mainly of a test indication and/or overall diagnosis. Some laboratories will have additional information, but in most cases these data will not require curation by an expert clinician since the submitting laboratories are merely relaying the test indication to the database.

Although this level of data appears to be minimal, it still has value because many more cases can be collected this way. Our efforts at data collection for the ISCA Consortium have illustrated this point when compared to the data collection model of the DECIPHER database. The DECIPHER database has attempted to collect very detailed phenotypic data on all patients through a full consent process; however, as a result, this database has achieved many fewer database entries. Since the ISCA Consortium has allowed the collection of a minimal amount of phenotypic data, such as a test indication from a clinical laboratory, the ISCA database already has at least five times the number of data entries compared to DECIPHER and still provides a valuable resource by offering users the opportunity to observe how variants are classified by other experts in the community. For this reason, collection of all laboratory data will be allowed as long as there is at least a disease scope or test indication provided.

Curated data from locus-specific mutation databases and OMIM is expected to yield minimal phenotypic data as well, such as a test indication or syndrome diagnosis, similar to what is expected from clinical laboratory data. However, this data will still be valued towards the overall goals of this project.

Phenotypic data entry by clinicians offers the best opportunity for detailed phenotypic information, and can be collected at the time of test ordering (preferred) or after test results are available. In ISCA the genetic counselors working in the EEAW were able to increase the submission of phenotypic information by directly contacting clinicians who frequently submitted samples. A similar effort will be employed in this project. Clinician information collection will be best achieved at the point of care (POC), and we will facilitate this process by providing standardized clinical data collection forms. We will also assist laboratories in getting the necessary IRB protocols in place to enable robust data collection and submission by the variety of mechanisms discussed in Aim 1. POC data entry will be facilitated by Case Report Forms that can be downloaded from our websites, including the ISCA, ClinVar, and MutaDATABASE websites.

Although clinicians will be the best source of clinical phenotypic data, we anticipate that participation by busy clinicians will be variable and therefore unreliable as the sole source of detailed phenotypic data. Therefore, we are also engaging several patient groups to assist in the submission of phenotypic data from patients as well as the linkage of existing patient registry data to genetic data. We will integrate this mechanism as part of the model curation projects. We will work with patient organizations, such as UNIQUE, the HCM Association, CureCMD (Congenital Muscular Dystrophies), and Patient Crossroads to integrate datasets. (See letters of support from 6 patient advocacy groups.) For example, Patient Crossroads, an online patient registry service hosting multiple genetic disorders (http://www.patientcrossroads.com/), has a demonstration project with dbVar for Congenital Muscular Dystrophy where they are linking mutation results from clinical laboratories with phenotypic information from the patient registry; the combined datasets are submitted to dbVar. Because this service uses open source software, this model can be applied more broadly to a number of different disorders. Engaging the patient groups will not only result in a "standardized" input of data (if guided appropriately), but also harness the most motivated cohorts of the process, the patients and their families.

2G. Sustainability of Variant and Phenotypic Data Submission

To encourage submission of data and assist laboratories with the various processes necessary to initiate the project at their individual institutions, the Engagement, Education, and Access Workgroup (EEAW) will develop a section of the project website, host webinars, and develop educational materials for clinicians and laboratories. The workgroup will also partner with patient advocacy groups to increase the submission of phenotypic data. UNIQUE, a patient support organization for individuals with rare chromosome disorders, considers helping individuals without a specific advocacy organization as part of its mission. We will work with UNIQUE and the NIH Office of Rare Disease Research Global Patient Registry to encourage collection of genotypic and phenotypic data for individuals without a “specific” disease registry.

We have had discussions with several groups about how to promote the submission of data by clinical laboratories on a continuing basis. Laboratory directors and laboratory genetic counselors need to understand that submitting their data to a combined database is a powerful method of quality control (QC). For example, as part of the ISCA Initiative on Medical and Payer issues for Array-based Cytogenomic Testing (IMPACT) Workgroup, we discussed this issue with Margaret Piper of the BlueCross/BlueShield Technology Center. She indicated that insurance medical directors are concerned about the variability of laboratory report interpretations related to genomic variation and would strongly support efforts to improve QC by submitting data to centralized databases. Preliminary discussions with the College of American Pathologists (CAP) have also indicated that CAP would consider a similar effort to improve QC (see letter of support from Stanley J. Robboy, M.D., FCAP, President of CAP). Dr. James Ostell, Chief of the Information Engineering Branch of NCBI has also indicated that the NCBI Genetic Testing Registry (GTR) would consider adding a field to indicate which laboratories are contributing data to public databases, which may influence customer choice of testing laboratories and allow market forces to stimulate data submission.

Most importantly, through the development of the eight model curation projects described in Aim 3, we have found that laboratories are most willing to participate in data submission and shared curation when there is a clear organized multi-institutional effort to gather and curate data among experts. Therefore, we will continue to organize these collaborative groupings to help facilitate robust sharing and curation. We anticipate that these efforts will then continue in a self-sustaining manner as laboratories realize the added benefit of an expert community and data-sharing model. This result has already been observed through the experience of the ISCA Consortium.

Aim 3: Implement sustainable expert clinical level curation systems of human genomic variants.

In conjunction with the development of classification standards (Aim 1) and the submission of data into a single accessible environment (Aim 2), tremendous effort will be focused on curating the data, including classifying variants with respect to their role in human health and disease. Data curation will be defined at one of four different levels (Figure 4):

- **Uncurated:** Variants without clinical categorization
- **Single-source curation:** Clinical categorization by a single laboratory or other entity
- **Expert-level curation:** Evaluation by an expert panel
- **Guideline-level curation:** Publication of consensus guidelines by professional organizations

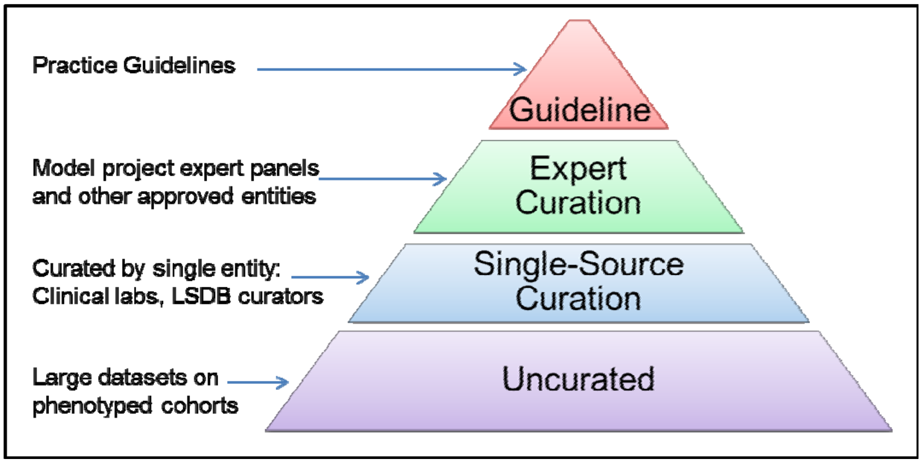

Practice Guidelines
Guideline
Model project expert panels and other approved entities
Expert Curation
Curated by single entity: Clinical labs, LSDB curators
Single-Source Curation
Large datasets on phenotyped cohorts
Uncurated

Figure 4. Levels of data curation for clinical classification of variants

The different levels of curation will be described and clearly noted in ClinVar, enabling users to filter the data based upon the level of curation. In this way, the system will support diverse types of usage. For example, use of variant information for research applications requires different levels of confidence in the significance of variant classification than information used for clinical care decisions. It should be noted that levels of curation are not to be confused with actual variant classification categories such as pathogenic, benign, or variant of unknown significance. For example, a variant could be classified as a variant of unknown significance by a professional organization (i.e., guideline level) demonstrating a high level of confidence that the significance of the variant is unknown. More detailed descriptions of various levels of curation are described below:

A) Uncurated

Data submitted without a clinical classification, such as data from the 1000 Genomes project [20], or other large scale sequencing efforts would be considered "uncurated."

B) Single-source Curation

Structural or sequence-level data submitted from clinical genetic laboratories will have clinical assertions associated with each variant and will therefore be considered curated at the single-source level. In addition, data submitted from most locus-specific databases will also be considered to be curated at this level.

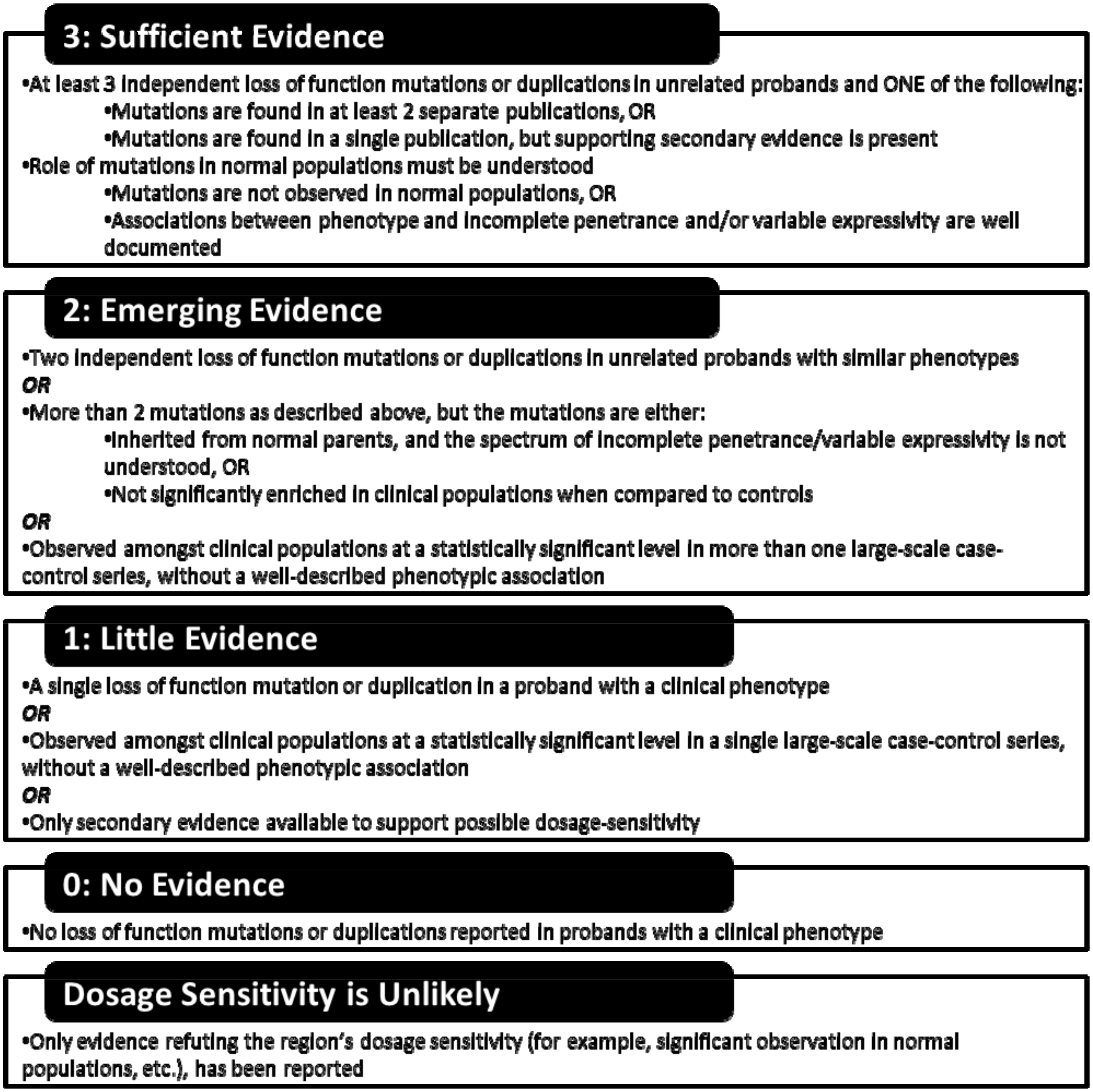
In an effort to improve the accuracy and standardization of single-source curation, the ISCA Consortium, in partnership with NCBI, has developed several algorithm-based curation tools and integrated them into the data submission process. The goal of this effort is to improve the quality of data entering the database while providing immediate value to the submitter by performing validation and consistency checks on the data. Data curation is performed at three levels: 1) check that the variant is defined within valid regions of genomic sequence, 2) identify regions where the laboratory has reported different clinical calls for similar observations, and 3) compare the submitted calls against an expert curated list of regions with established clinical significance. The results are presented to the submitter via an authenticated web page that includes a genome overview, a genome browser and a tabular report of the regions to be reviewed. The submitter examines each identified "conflict" region and has the option to click on the variant and change the call, or chose to retain the original call and ignore the conflict warning. For these cases, the submitter is asked to provide a reason for their decision. Upon completion of the review, the submitter finalizes the submission, which will incorporate a history of any changes applied. The editing history of each event is retained in a central database to allow for full review at a later date. This curation tool provides a user-friendly system by which the information submitted to the ISCA database can be reviewed.

Similar laboratory-level curation tools will be developed for sequence-level data. For example, laboratories submitting variants through the MutaREPORTER software can use MutaREPORTER to facilitate the assignment of an appropriate category of pathogenicity based upon the user's entry of evidence for the variant. This algorithm will reflect the broad consensus developed for variant classification rules as described in Aim 1. Given that some variants may not meet a strict set of definitions for variant classification, a user will be able to override the facilitated classification system and choose the appropriate category. If the user chooses a category inconsistent with other data in the system, this discrepancy will be returned to the user for consideration. All evidence and indicated reasons for clinical classification will be retained by the software and viewable at the variant level.

C) Expert-level Curation

In order to improve our understanding of disease relevance for individual variants (e.g., those that are "known" pathogenic, benign, etc.), expert-level curation of aggregated data is necessary. Individual laboratories may come to different conclusions regarding the clinical consequences of the same variant, and the expert-level review process can provide a method of resolution [23]. Furthermore, as evidence available regarding particular variants changes over time, clinical classifications made at the time of data submission may require re-evaluation [24]. Expert level curation will allow for "real time" integrated analysis of data from multiple submitting laboratories with currently available evidence for clinical assertion.

To address these needs for the structural variant community, the ISCA Consortium has established an Evidence-Based Review Committee (EBRC) which will become part of the Structural Variant Workgroup for this project. This committee is charged with collecting and evaluating the evidence necessary to determine the pathogenicity and clinical impact of particular structural variants, as well as reviewing clinical categorization conflicts for structural variants within the database. A rating system was developed to quantify the available evidence for standardized decision-making regarding clinical classifications (Riggs et al., submitted; see Appendix). Since CMA can detect losses and gains of genomic material, each genomic region is given two independent ratings: a loss of function rating to address deletions and loss of function mutations resulting in haploinsufficiency and a triplosensitivity rating to address whole gene duplications. Loss of function and triplosensitivity ratings range from 0 to 3, with increasing levels of evidence suggesting that dosage sensitivity results in a particular phenotype (Figure 5). Both the loss of function and triplosensitivity ratings, along with all supporting evidence, are being recorded in a web-based database customized for our evidence-based review process (http://www.atlassian.com/software/jira/), and are being made available to the public for review and comment (http://www.ncbi.nlm.nih.gov/projects/dbvar/ISCA). To incorporate new and emerging evidence and to ensure that the evidence-based recommendations remain up-to-date, each region will be re-evaluated on a periodic basis.



3: Sufficient Evidence

•At least 3 independent loss of function mutations or duplications in unrelated probands and ONE of the following:
- •Mutations are found in at least 2 separate publications, OR
- •Mutations are found in a single publication, but supporting secondary evidence is present

•Role of mutations in normal populations must be understood
- •Mutations are not observed in normal populations, OR
- •Associations between phenotype and incomplete penetrance and/or variable expressivity are well documented

2: Emerging Evidence

•Two independent loss of function mutations or duplications in unrelated probands with similar phenotypes

OR

•More than 2 mutations as described above, but the mutations are either:
- •Inherited from normal parents, and the spectrum of incomplete penetrance/variable expressivity is not understood, OR
- •Not significantly enriched in clinical populations when compared to controls

OR

•Observed amongst clinical populations at a statistically significant level in more than one large-scale case-control series, without a well-described phenotypic association

1: Little Evidence

•A single loss of function mutation or duplication in a proband with a clinical phenotype

OR

•Observed amongst clinical populations at a statistically significant level in a single large-scale case-control series, without a well-described phenotypic association

OR

•Only secondary evidence available to support possible dosage-sensitivity

0: No Evidence

•No loss of function mutations or duplications reported in probands with a clinical phenotype

Dosage Sensitivity is Unlikely

•Only evidence refuting the region's dosage sensitivity (for example, significant observation in normal populations, etc.), has been reported

Figure 5. Evidence-based rules for the classification of structural variants

Those regions with the strongest evidence supporting dosage sensitivity (i.e., those with loss of function or triplosensitivity ratings of 3) may be considered clinically “pathogenic,” while those with the strongest evidence refuting dosage sensitivity (i.e., those assigned to the “Dosage Sensitivity Unlikely”) category may be considered clinically “benign.” These regions will be brought to the Structural Variant Workgroup for consideration for inclusion in the ISCA Consortium curated list of “known” pathogenic and benign regions (“ISCA Consortium curated dataset”) (http://www.ncbi.nlm.nih.gov/dbvar/studies/nstd45/).

As part of the curation process, a conflict report is generated, using the data from the evidence-based process, to identify conflicts in structural variant clinical categorization reporting. In general, calls overlapping by >50% in adjacent tiers are considered “in conflict.” The comparisons, however, are directional: for example, a large region classified as “pathogenic” may not be in conflict with smaller overlapping regions classified as “benign”, but a large region classified as “benign” would generate conflicts with smaller contained “pathogenic” regions. The minimum region of overlap would generate a new conflict region definition, which is reported back to the Structural Variant Workgroup for review. The workgroup also incorporates frequency data from the submitted structural variants with the evidence-based data described above to resolve these conflicts.

This model of expert-level curation will be modified for use for sequence-level variants. Initial efforts will focus on expert-level curation of variants in ~50 genes included in the eight model projects being used to develop an expert curation process for sequence-level variants (see tables below). After the models are developed, the effort will expand to other gene sets. Expert curators have been identified for each model sequence-level project as noted in the tables below. In order to avoid duplication of curators with other locus-specific databases; existing curators from other variant databases have been invited to join this project. Each staff curator will review data being submitted from many sources. Curators operating within the MutaREPORTER system will have access to a communication tool called MutaCIRCLES, which allows groups of clinical laboratories, researchers, and curators to discuss variant data (Figure 1). This system is important given that detailed phenotypic data, segregation data, control data and other evidence may not be submitted during variant submission due to labs harbouring this data in unstructured ways. This type of information will be shared and discussed though the MutaCIRCLES communication platform, with these communications archived.

Before laboratory or LSDB data submission, the staff curator (also called MutaADMINISTRATOR) will evaluate each lab’s or curator’s system of variant classification and decide how to translate their data set to conform to the agreed-upon standards. Once the data is in the system, it will be ready for review. Difficult variants will be discussed across all curators and laboratories for consensus agreement. The staff curator will lead communication among the experts within MutaCIRCLES, and documentation of curation efforts will occur in the database. Algorithms will be developed to alert a curator when variants exist in the database with discrepant classifications. Variants with inconsistent classifications will first be queued for evaluation, then variants of unknown significance, then missense variants with limited documented evidence. Other automated rules will also be developed to comb the data for discrepant pieces of evidence. For example, if a variant has been classified as pathogenic for a rare disease but has frequency information from controls documenting common occurrence in the general population, this variant would be flagged for review.

In addition to the ISCA evidence-based review project, the eight other demonstration projects described below will focus on the following areas: 1) Hypertrophic cardiomyopathy, 2) Noonan syndrome spectrum disorders, 3) Hereditary colorectal cancer, 4) Developmental delay, 5) Inborn errors of metabolism, 6) Congenital muscular dystrophies, 7) *PTEN* associated disorders, and 8) Mowat-Wilson Syndrome. Tables listing the key components for each project are included below.

Other groups not included in these initial demonstration projects can request that their clinical classification process be reviewed by our Executive Committee for assignment as an “Expert Curation” process. Approval would be based upon assurance that the developed standards for classifying variants are being adhered to and that experts on these particular gene sets and diseases are involved in the review process.

For expansion of the curation projects, the Sequence Variant Committee will meet regularly to review the diseases and gene sets for which clinical testing is offered. This workgroup will initiate new projects, giving higher priority to those diseases and gene sets for which multiple laboratories offer testing. Expert researchers, clinicians, laboratory directors and patient advocacy groups will be contacted to participate in the curation process in a similar manner as for the initial eight projects.

Diseases	Hypertrophic Cardiomyopathy
Clinical synopsis	1/500 incidence; dominant inheritance; typically adult onset cardiac disease with high risk for sudden cardiac death (SCD)
Clinical utility of testing	Identifying risk for SCD, particularly in asymptomatic family members; Delineating sarcomere disease from storage disease which may be treatable (ERT for *GLA* mutations)
Minimum gene set	*MYH7, MYBPC3, TNNI3, TNNT2, TPM1, ACTC1, MYL2, MYL3, PRKAG2, LAMP2, GLA*

Existing LSDB	Cardiogenomics (last updated in 2006) – will be replaced with this effort
Project director	Heidi Rehm (Harvard)
Expert curators	Heidi Rehm (Harvard), Birgit Funke (Harvard), Christine Seidman (Harvard), Jon Seidman (Harvard), Wendy Chung (Columbia), Chris Semsarian (Australia), Michael Ackerman (Mayo)
Curation staff member	Melissa Kelly (Harvard-Partners)
Patient advocacy groups	HCM Association (HCMA President Lisa Salberg - see letter of support))
Contributing labs	Harvard-Partners, GeneDx, Transgenomics, Correlagen (LabCorp)
Cases to contribute	>3000 (HP), >2000 (GeneDx), >1000 (Transgenomics), >1000 (Correlagen), 1000 (Ackerman)
Variants to contribute	>1500 unique variants from Harvard-Partners; data TBD from other labs
Phenotyping approaches	Clinical lab data submission (retrospective as well as improved collection with standardized form use); discuss project with HCMA to associate existing >3000 consented patient records with genotypic data

Diseases	Noonan Syndrome, Cardio-Facio-Cutaneous Syndrome, LEOPARD Syndrome, Costello Syndrome
Clinical synopsis	Facial dysmorphology, short stature, cardiac defects, motor delay, bleeding diathesis. Autosomal dominant or de novo inheritance
Clinical utility of testing	Disease management; family planning
Minimum gene set	*PTPN11, SOS1, RAF1, KRAS, SHOC2, BRAF, MAP2K1, MAP2K2, HRAS*
Existing LSDB	SOS1 LOVD database (39 variants)
Project director	Sherri Bale (GeneDx)
Expert curators	Sherri Bale (GeneDx), Bruce Gelb (Mt. Sinai School of Med), Marco Tartaglia (Italian network for RASopathies), Amy Roberts (Harvard), Martin Zenker
Curation staff members	Brad Williams (GeneDx), Lisa Vincent (GeneDx)
Patient advocacy groups	Noonan Syndrome Support Group (Wanda Robinson - see letter of support))
Contributing labs	Baylor, Athena, Greenwood, Boston University, Emory, Children's Hospital of Boston, U of Oklahoma, Harvard-Partners, ARUP, GeneDx, Nationwide Children's, Nemours/Alfred I Dupont, Mt Sinai School of Medicine
Cases to contribute	>7400 (includes 10 of the above labs)
Variants to contribute	>7700 variant observations (includes 10 of the above labs)
Phenotyping approaches	Clinical lab data submission (retrospective as well as improved collection with standardized form use)

Diseases	Hereditary Colorectal Cancer (HCrC)
Clinical synopsis	~150,000 new colon cancer cases annually, ~12,000 due to a germline gene variant. Age of onset is less than 50 yrs., with aggressive cases noted in teen years. Usually autosomal dominant;
Clinical utility of testing	Identifying risk for CrC, particularly in unaffected family members; prophylactic surgery and increased monitoring decreases mortality in affected families.
Minimum gene set	*APC, MUTYH, MLH1, MSH2, MSH6, and PMS2*
Existing LSDB	InSiGHT (see collaborative letter from Finlay Macrae)
Project director	Matthew Ferber (Mayo)
Expert curators	Matthew Ferber (Mayo), International InSiGHT curation team

Curation staff member	Kim Schahl and Brittany Thomas (Mayo Clinic)
Contributing labs	InSiGHT, Mayo Clinic, Harvard-Partners, Emory, Quest, Correlagen (LabCorp), ARUP
Cases to contribute	>25,000 (InSiGHT), >9000 (Mayo Clinic), Other labs TBD
Variants to contribute	>5000 (InSiGHT), >2000 (Mayo Clinic) unique; data TBD from other labs
Phenotyping approaches	Clinical lab data submission (retrospective as well as improved collection with standardized form use).

Diseases	Inborn Errors of Metabolism
Clinical synopsis	Various metabolic enzyme deficiencies associated with a varied clinical spectrum 1/500 combined incidence; autosomal or X-linked recessive inheritance
Clinical utility of testing	Confirming disease identified through NBS, testing family members for carrier or affected status.
Minimum genes set	*GALT, BTD*, *ACADM*, *ACADVL*, *LCHAD*, *SLC22A5*, *OTC*, *ASS1*, *ARG, PAH*
Existing LSDB	ARUP databases for *BTD, GALT, SLC22A5*
Project directors	Rong Mao (ARUP), Elaine Lyon (ARUP),
Expert curators	Rong Mao (ARUP), Elaine Lyon (ARUP), Barry Wolf (Henry Ford Health System), Nicola Longo (University of Utah)
Curation staff member	Anna Gardner (ARUP)
Contributing labs	GeneDx, Cincinnati Children's, Quest, Henry Ford Hospital, ARUP, Mayo
Cases to contribute	>500 (Cincinnati Children), >700 (GeneDx), >500 (ARUP), >1000 Emory , >1450 (Mayo), >80 (Henry Ford Hospital)
Variants to contribute	>350 (ARUP), >350 (GeneDx); >400 Emory; data TBD from other labs
Phenotyping approaches	Clinical lab data submission (patient information collected at time of testing available from ARUP and GeneDx) possible collaboration with advocacy groups, functional studies may be possible through Dr. Longo's research laboratory

Diseases	Developmental delay (Rett/Angelman/Early infantile epileptic encephalopathy)
Clinical synopsis	Incidence of ~1 in 8,500 (females with Rett) and 1 in 12,000-20,000 (Angelman syndrome and EIEE), different modes of inheritance including X-linked, dominant and imprinting disorder that share overlapping features that include mental retardation, developmental delay, seizures and microcephaly
Clinical utility of testing	Diagnosis, management implications and accurate recurrence risk estimates
Minimum gene set	*ARX, CDKL5, MECP2, UBE3A*
Existing LSDB	RettBase (*MECP2* and *CDKL5*) [not standardized with other locus-specific databases]; LOVD (*UBE3A*) [uncurated]; None for *ARX*
Project director	Soma Das (University of Chicago)
Expert curators	Soma Das (University of Chicago), Ping Fang (Baylor College of Medicine), Madhuri Hegde (Emory University), Mike Friez (Greenwood Genetics). Additional curators for specific genes include: John Christodoulou for the *MECP2* and *CDKL5* genes (University of Sydney and current curator for RettBASE), Jozef Gecz and Cheryl Shoubridge for *ARX* (University of Adelaide) and Simon Ramsden for *UBE3A* (Central Manchester University)
Curation staff member	Melissa Dempsey (University of Chicago)

Contributing labs	Baylor, University of Chicago, Emory, Greenwood Genetics, GeneDx, Mayo, Alfred I. duPont Hospital, All Children's Hospital St. Petersburg, ARUP, Boston University, Children's Mercy Hospital, Children's Hospital of Philadelphia, City of Hope, Detroit Medical Center, Nationwide Children's Hospital, Nebraska Medical Center, Oregon Health Sciences University
Cases to contribute	> 25,000 cases (combined labs) that can be broken down as follows: >17,000 (Baylor College of Medicine), >3,700 (University of Chicago), ~700 (Emory University), >400 (GeneDx), >900 (Mayo), ~140 (Alfred I. duPont Hospital for Children), >650 (ARUP), >600 (CHOP), numbers TBD from remaining labs
Variants to contribute	> 2,000 variants (combined labs) that can be broken down as follows: >680 (Baylor College of Medicine), >390 (University of Chicago), >50 (Emory), >200 (Greenwood), >50 (GeneDx), >90 (Mayo), >40 (Alfred I. duPont Hospital for Children), >20 (All Children's Hospital St. Petersburg), >20 (ARUP), >20 (Children's Mercy Hospital), >60 (CHOP), >30 (City of Hope), >10 (OHSU)
Phenotyping approaches	Clinical lab data submission with diagnoses and indication for testing; improved collection with standardized form use; contacting patient support groups, such as the Angelman syndrome foundation, to help with collection of clinical information.

Diseases	Congenital Muscular Dystrophy
Clinical synopsis	1 in 50,000-200,000; muscle degeneration presenting at or near birth
Clinical utility of testing	Identifying risk and confirming diagnosis for CMD particularly in family members, Clinical trials enrollment through TREAT-NMD
Minimum gene set	*COL6A1, COL6A2, COL6A3*, *ITGA7*, *FKTN*, *FKRP*, *POMGNT1*, *POMT1, POMT2*, *SEPN1*, *LARGE, LAMA2*
Existing LSDB	CureCMD
Project director	Madhuri Hegde (Emory)
Expert curators	Madhuri Hegde (Emory), Christin Collins (Emory), Carsten Bonnemann (NINDS), Anne Rutkowski (CureCMD)
Curation staff member	Alice Tanner, Ephrem Chin (Emory)
Patient advocacy groups	CureCMD
Contributing labs	Emory, CureCMD registry which draws from several laboratories
Cases to contribute	Emory (253), CureCMD (475)
Variants to contribute	>475 unique variants from contributing labs
Phenotyping approaches	Clinical lab data submission (retrospective as well as improved collection with standardized form use);

Diseases	*PTEN* Hamartoma Tumor Syndrome, Cowden Syndrome, Banayan-Riley-Ruvalcaba syndrome, Macrocephaly-Autism
Clinical synopsis	1 in 200,000 for Cowden Syndrome; dominant inheritance; benign and malignant tumor growth
Clinical utility of testing	Identifying risk for malignancy and confirming diagnosis for PHTS, CS, Autism, particularly in family members
Minimum gene set	PTEN
Existing LSDB	None
Project director	Madhuri Hegde (Emory)

Expert curators	Madhuri Hegde (Emory), Bradford Coffee (Emory), Patrick Willems (MutaDatabase), Charis Eng (Cleveland Clinic)
Curation staff member	Alice Tanner, Ephrem Chin (Emory)
Contributing labs	Emory, GeneDx, Greenwood Genetics, University of Chicago, ARUP
Cases to contribute	Emory (255), GeneDx (4023), Greenwood Genetics (275); ARUP:TBD
Variants to contribute	>250 unique variants from contributing laboratories
Phenotyping approaches	Clinical lab data submission (retrospective as well as improved collection with standardized form use)

Diseases	Mowat-Wilson Syndrome
Clinical synopsis	unknown incidence; dominant inheritance but usually occurs *de novo*, germline mosaicism has been reported; multiple congenital abnormality syndrome
Clinical utility of testing	Establish a diagnosis and end the diagnostic odyssey, determine recurrence risk for family
Minimum gene set	*ZEB2*
Existing LSDB	None
Project director	Madhuri Hegde (Emory)
Expert curators	Madhuri Hegde (Emory), Lora Bean (Emory), Carol Saunders (Children's Mercy Hospitals and Clinics), Margaret Adam (University of Washington)
Curation staff member	Alice Tanner, Ephrem Chin (Emory)
Contributing labs	Emory Genetics Laboratory, Children's Mercy Hospitals and Clinics, U of Chicago, Children's Hospital of Philadelphia
Cases to contribute	80 (Emory), 100 (Children's Mercy), 46 (U of Chicago)
Variants to contribute	77 (Emory), 40 (Children's Mercy), 5 (U of Chicago)
Phenotyping approaches	Clinical lab data submission (retrospective as well as improved collection with standardized form use); Consultation with Dr. Adam as needed.

D) Guideline-level curation

Variant classifications endorsed by professional societies will be considered the highest level of curation. Some examples of variants with guideline-level curation would be those in the *CFTR* (Cystic Fibrosis Transmembrane conductance Regulator) gene that have been recommended for inclusion on carrier screening panels by the American College of Medical Genetics [25]. Very few variants will exist in this curation level but those that are published within professional guidelines will be noted as such.

ACCESS AND DISSEMINATION PLAN

Submitted data, including variants, phenotype information, clinical assertions, and evidence supporting such clinical assertions will be publicly accessible through the ClinVar website, currently housed within NCBI (http://www.ncbi.nlm.nih.gov/clinvar/). ClinVar will develop visualization tools to suit users' needs as appropriate, and will work with interested individuals, laboratories, and others on other methods to incorporate the data into their routine work flows. Through support of its other databases, such as dbVar and dbGaP (where ISCA data has been deposited to date), NCBI has demonstrated its willingness to work with the community to develop the tools necessary to effectively utilize data. In response to queries by ISCA Consortium members, NCBI instituted several formatting and display changes to their browser, provided the ISCA data in an easily downloadable format for use in other genome browser or vendor applications, and adapted an existing software program for use in structural variant curation. This commitment to community access will be extended to ClinVar.

Although the data in ClinVar will be readily accessible through the NCBI website, alerting the community to its existence and involving them in its continued development will be critical to optimizing its success. Given the substantial breadth of this project, there are a variety of stakeholder groups to be engaged. Without ongoing data contribution from the community, ClinVar will fail to achieve the goal of collecting large datasets of human genomic variation. Therefore, in Year 1, the project will focus on setting up collaborations for data submissions with clinical laboratories, research laboratories, clinicians, and patient advocacy groups. Reaching out to this varied group will require multiple strategies, and will be the focus of the Engagement, Education, and Access Workgroup (EEAW). This group will be modeled after a similarly-focused workgroup within the ISCA Consortium, and will use the most successful engagement strategies from the ISCA project to guide future efforts. Through our work with the ISCA Consortium, we have found that laboratories require, at a minimum, technical support to be able to navigate the IRB process and facilitate data submission. Therefore, the infrastructure supported by this grant will be critical to the success of the ClinVar database.

Website for Communication and Engagement

One of the most critical and wide-reaching engagement tools is a web presence. The EEAW will develop a website unique to this project that will serve as a “home” for its various stakeholders. Modeled after the ISCA Consortium website (www.iscaconsortium.org), which has had over 3,600 unique visitors within the past six months, the website for this project will include a description of the primary aims/goals, the purview and membership of each working group, and contact information for core project leaders. Users will be encouraged to register with the site for access to additional material; the registration process will be free and simple (requiring only basic contact information as is done for ISCA), and will allow us to easily track user numbers. This registration process will also allow us to contact website users via email blasts to alert them to important new developments, such as learning opportunities, research collaboration requests, project deadlines, volunteer needs, etc. This type of newsworthy information will also be displayed in a dedicated “News” section of the site.

Communication with our user community will be a key feature of the website. The experience of the ISCA Consortium demonstrated what worked well and what did not work well. Multiple different communication options were necessary to capture the needs of a wide variety of users. As stated above, contact information for core project personnel will be prominently displayed; a generic “help” email address will also be implemented, allowing users to send questions/requests to a general address if they are unsure how they should be personally directed. These queries will be routed to key project personnel (e.g. website staff, coordinators for the structural and sequence portions of the project, etc.) for appropriate distribution. Telephone-based user support will also be supported for those who are interested in this method of communication. We will also foster communication between community members through a web forum for those who wish to pose questions, opportunities, etc. directly to the larger community. In ISCA, we found many participants were reluctant to post forum queries with their name; therefore, we developed a system for them to remain anonymous. A similar process will be used whereby questions can be sent to the project staff and posted anonymously. A “Frequently Asked Questions” page will also be featured on the site.

The site will also serve as a hub for various online tutorials, webinars, etc. for our different stakeholder groups. The ISCA Consortium successfully used these web-based tactics to engage clinical and laboratory communities. These groups often do not have schedule flexibility to be able to attend live events, so the availability of pre-recorded webinar content or easily accessible online tutorials has been key to the success of ISCA. Furthermore, in an effort to educate clinicians and submitting laboratories about the process of data submission, the ISCA Consortium developed separate pre-recorded webinar content, each focusing on the key interests of the different groups. In conjunction with this, several live question and answer sessions were scheduled. Users were able to watch the informational content at their convenience and then ask questions live at a time of their choosing. The question and answer sessions were also recorded, allowing those unable to attend live to still access the information. Questions were also taken via email and phone before, during, and after the session for those unable to participate. All recorded content is available on the ISCA website (https://www.iscaconsortium.org/index.php/articlesabstracts/82-announcement2*).* This type of multi-faceted strategy designed to accommodate the availability of the ISCA community will be employed to engage the larger community of this project. In addition to sessions like the ones described above, sessions will be developed to discuss data curation, phenotype data collection, use of software tools, ClinVar, and other topics as proposed by the community. Sessions will also be tailored to suit different stakeholder groups (e.g. test requisition forms for laboratories, phenotype forms for clinicians, and joining/engaging patient registries for patient advocacy groups).

The website will also support educational modules for the continuing enrichment of users. The "Virtual Case Conference" format developed by the ISCA Consortium will be continued, allowing members to post challenging or unusual cases that may present during the course of routine clinical care. This feature has been popular among the ISCA community, with each posting generating an average of 185 views. This educational tool will easily be adapted to include cases regarding all types of genomic variation. Other potential educational modules include: information regarding privacy protections; related databases and how to locate and access them; patient advocacy group registries and how to locate them; different types of cases in the databases and whether re-contact is possible; how to re-contact individuals in the databases for research studies, and examples of successful research projects using the databases.

Practice-based tools may also be developed and hosted on the website as needed/desired by the community. For example, in response to requests from ISCA members for assistance with appeals to insurance companies for coverage of CMA, the ISCA Consortium developed an online toolkit providing a letter of medical necessity template, sample wording, and a catalog of useful literature references. Since posting this resource 5 months ago, this toolkit has had over 2,000 hits on the ISCA website. Therefore, the content of the website will very much be the result of community input; as requests are received, solutions will be developed to reflect the needs of the larger group.

The website will also house downloads needed for data visualization/utilization. For example, the ISCA website has an entire section dedicated to downloads needed for array analysis, including tracks displaying ISCA data in the two most recent genome assemblies, design files for the ISCA standard array design, software data interpretation tools, and quality control information. This approach will be extended to support any similar needs for sequence-level variants that are not already provided by NCBI and other genomic resource providers. The website will also provide links to many other existing resources that may be relevant to users.

Other Communication Plans

Although electronic communication will be a key tool for community engagement, face-to-face communication will also be important. Therefore we will maintain a strong presence at professional society meetings and through project-specific conferences to provide various groups the opportunity to interact with one another, which often facilitates the development of new ideas or approaches. The ISCA Consortium hosted ancillary events at several professional society meetings, including those of the ACMG and ASHG; they also maintained exhibit booths at these meetings for informal question and answer sessions with meeting participants. Presence at these meetings introduced the ISCA project to genetics professionals such as clinical geneticists, genetic counselors, laboratory geneticists, and researchers. This level of exposure yielded new members and additional support for the project. These same strategies will also be employed for the extended project. Once the project is well-established, a project-specific conference will take place to update community members on current progress and to elicit feedback on future goals. This strategy was undertaken by the ISCA Consortium, which held its first public conference in January 2010, attracting over 200 participants. This conference allowed for an open dialogue between ISCA Consortium personnel and database users, resulting in concrete changes to the database. For example, a decision was made to standardize the color-coding scheme for annotating losses and gains across multiple publicly available CNV databases, including ISCA, dbVar, DGV, DECIPHER, Ensembl, and the UCSC Genome Browser. Representatives from all databases agreed to use red for losses and blue for gains, making the interpretation and comparison of data from these various sources much easier and reducing the likelihood of errors.

We will also engage the clinical and research communities through various professional publications. This mechanism allows for exposure of the project's goals, directions, and accomplishments to specific professional groups. The ISCA Consortium has utilized this strategy, particularly in relation to the engagement of genetic counselors: an article was published in *Perspectives in Genetic Counseling* describing the importance of databases and the submission of phenotypic data [26]. A proposed article for the *Journal of Genetic Counseling* about the role of genetic counselors in ISCA and the importance of clinicians working with laboratories to increase submissions to databases has also been accepted for review. Additional articles discussing the output of the ISCA database [27] and recommendations for CMA use [16] have been published, while invited reviews on the curation process (for *Clinical Genetics*) and the phenotype data collection process (for *Human Mutation*) are being developed. Publications of standards, as they are developed, as well as research and clinical activities and outcomes resulting from the database, will continue to be pursued to demonstrate our collective expertise in genomic variation.

Collaborating with Patient Advocacy Registries

The patient advocacy community will also play a critical role in this project particularly for integrating phenotypic data with the collected genotype data. In addition to continuing to work with existing groups, such as UNIQUE (the rare chromosome disorder support group), Duchenne Connect, and Simons Variations in Individuals Project (SVIP) Connect, we have already engaged several other patient support organizations eager to work with us (see letters of support from the Genetic Alliance, the HCM Association, CureCMD, UNIQUE, Patient Crossroads, and the Noonan Syndrome Support Group) and we will continue to involve others. We will work with the patient advocacy community to provide input on clinical data collection forms discussed in Aim 1 and circulate them with the patient community. In previous projects, we have found that when patients share these information sheets with clinicians, the clinicians are much more likely to fill them out. We will use a similar campaign as the project is extended across disease areas. Many of the patient advocacy organizations have existing long-term registries with rich phenotypic information. We will explore with our patient advocacy partners how to link registries and databases and make researchers aware of these resources. We will also connect researchers and clinicians with the patient advocacy groups to discuss the type of information that would be the most useful and how to collect the information in a way that meets the needs of researchers. In supporting Duchenne Connect, CureCMD, and Simons VIP Connect, we found that data contributed by participants was very accurate. All three patient registries verified the accuracy of critical data to gain the acceptance and use by the research community. In this project we will explore the development of similar processes for all of our patient advocacy registry partners. We have worked with UNIQUE and the NIH Office of Rare Disease Research Global Registry effort to develop resources and registries for individuals where there is not a “critical mass” for a targeted advocacy group or registry. In this project we will explore how to capture phenotypic data from these families without a specific registry or support group and make it available to the research community. We will also work together to encourage families to register in a specific patient registry or a “global registry” if a targeted registry is not available for their condition or variant.

Outreach to Research Community

Beginning in Year 2, the EEAW will reach out to the basic science and clinical research communities to describe the project and how the databases can be used for research. Webinars and other educational programs will be developed for the research community to describe: (1) how to use ClinVar and related databases, (2) the various types of phenotypic data collected and available in ClinVar or linked databases, (3) how to locate patient advocacy registries and databases, (4) information about the NIH Office of Rare Diseases Research global patient registry, (5) how to work with patient advocacy groups to obtain more extensive phenotypic information, (6) how to re-contact individuals in the various databases for potential research projects, and (7) a Question & Answer feature with posted responses from project experts. The EEAW will hold focus groups with researchers and the educational programs will be modified based on feedback received. The EEAW will engage the research community by: (1) ancillary sessions targeting the research community during the ASHG and the ACMG meetings and other related research meetings, (2) email “blasts” to various research community members about the project linking them to the project website for additional information, (3) booths at research meetings to advertise and explain the project, and (4) editorials and announcements in research journals about the databases and the project. Surveys will be developed to evaluate the effectiveness of the database for supporting clinical and research activities as well as how well our support and educational infrastructure is working for communicating with the community.

Work Process

The EEAW will use biweekly conference calls for most of its work. We found this process to be very productive in ISCA. The various educational and outreach efforts will be discussed as a group and assigned to one or more of the EEAW members to develop. Draft products will be reviewed by the full workgroup.

OVERALL PROJECT TIMELINE, MILESTONES AND EXPECTED DELIVERABLES

Project Activities	Year 1				Year 2				Year 3			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Meetings												
Initial executive committee meeting	■											
Workgroup meetings	■	■	■	■	■	■	■	■	■	■	■	■
Annual workshop meeting		■				■				■		
Aim 1												
Review ISCA policies and use or modify for sequence-level variant/phenotype collection	■											
Solicit stakeholder feedback for variant classification nomenclature and standards	■											
Develop disease-specific ontologies and phenotype collection forms for diseases	■	■	■	■	■	■	■	■	■	■	■	■
Aim 2												
Support structural variant data submission	■	■	■	■	■	■	■	■	■	■	■	■
Organize model projects for sequence variants	■											
Secure and input data for model projects		■	■	■	■	■	■	■	■	■	■	■
Secure and input data for all other genes			■	■	■	■	■	■	■	■	■	■
Aim 3												
Expert data curation of structural variants	■	■	■	■	■	■	■	■	■	■	■	■
Define curation methods for sequence variants	■											
Customize software to support sequence variant curation		■										
Expert data curation of sequence variants			■	■	■	■	■	■	■	■	■	■
Access and Dissemination												
Engage and educate community about ClinVar	■	■	■	■	■	■	■	■	■	■	■	■

ADMINISTRATION AND MANAGEMENT

Executive Committee:

Heidi Rehm (PI), Christa Martin (PI) and Robert Nussbaum (PI)
Sherri Bale, Andy Faucett, Madhuri Hegde, David Ledbetter, David Miller, Erik Thorland, Patrick Willems

A. Organizational structure and staff responsibilities

The project will involve five workgroups overseen by the principal investigators and additional executive committee members as diagrammed in Figure 2. Drs. Rehm and Martin will oversee the sequence and structural variant projects respectively. The Executive Committee contains representatives from both the structural variant initiatives (Martin, Ledbetter and Thorland) the sequence variant initiatives (Rehm, Bale, Hegde and Willems) as well as three with equal involvement in both activities (Nussbaum, Miller and Faucett). The committee will meet by conference call on a monthly basis with in-person meetings occurring twice per year. Updates will be provided on all key areas at each meeting as defined by each workgroup below. The executive committee will ensure that the databases have been structured and configured to operate as specified and it will also ensure that all points of decision-making will be cued up for the PSS

Workgroup. Overall leadership and division of responsibilities among the principal investigators will be distributed as described in the Multi PI leadership plan.

Of the five workgroups, the Policies, Standards, and Sustainability Workgroup will have deliverables that impact all of the activities of the grant, being charged with key decision-making, policy and standards development. As such, this workgroup will involve the entire executive committee, the directors of all model curation projects (adding Drs. Lyon, Das, and Ferber), as well as representatives and consultants from a variety of disciplines. Drs. Bale and Ledbetter, co-chairs of the workgroup, will be responsible for making sure key decisions are being made to keep all activities on track.

The Sequence Variant Workgroup will be chaired by Drs. Hegde and Willems. Dr. Hegde will be responsible for overseeing all sequence level model curation projects and Dr. Willems will be charged with overseeing the larger scale unfunded curation of other genes through distribution of activities. The directors of each model curation project will give a bi-monthly summary report of activities regarding variants deposited and curated and progress on efforts to include clinical data. Dr. Hegde will share those reports with the executive committee. Dr. Willems will provide bi-monthly updates on data being deposited into MutaDATABASE and Donna Maglott will provide bi-monthly updates on data deposited to NCBI's ClinVar.

The Structural Variant Workgroup will be chaired by Dr. Thorland. Dr. Thorland will be responsible for organizing the curation and evidence-based efforts and making sure that the goals of this committee are completed in accordance with the project timeline. He will also report the progress of this committee back to the Executive Committee. Other members of this Workgroup include Drs. Aradhya, Martin, Kaminsky, Church, South, and Kearney who bring various areas of expertise to the group, including copy number variation and bioinformatics. The Structural Variant Workgroup will also oversee efforts to foster the evolution of the infrastructure needed for clinical laboratories to easily submit data to ClinVar to ensure the success of this project.

The Phenotype Workgroup will be chaired by Dr. Miller. Dr. Miller will be tasked with defining baseline approaches to the collection of clinical data for annotating genetic variants. He will work closely with all of the model curation projects to gauge progress on clinical data collection and report progress on a bi-monthly basis to the whole executive committee.

The Engagement, Education, and Access Workgroup will be chaired by Mr. Faucett. Mr. Faucett will be responsible for overseeing initiatives to engage, educate and train the community with regard to the resource being developed and how best it can serve the broader community. Dr. Faucett will also be in charge of working with our ELSI representative, Joan Scott, to ensure that the activities of the grant are sensitive to the ethical, legal and social perspectives of the community.

B. Scientific Advisory Board

A Scientific Advisory Board (SAB) will be chosen in collaboration with the NHGRI program office, if funding is awarded. The SAB will consist of senior members of the genetics community. It will include representation from those individuals who have led successful community resource initiatives as well as those who represent academic leaders who will make use of the resource and understand the broad needs of the genetics communities in both research and healthcare. Direct expertise and/or established liaisons with patient advocacy as well as ethical, legal and social issues will be represented. As appropriate, some of the named consultants may be appointed to the SAB. The SAB will meet twice a year with the executive committee, once by conference call and once at the in-person annual meeting. Scientific Advisors will also be welcomed to participate in any workgroups or aspects of the project. The Executive Committee (EC) will provide updates to the SAB on the progress of all grant initiatives and the SAB will advise the EC on priorities and direction of the projects as needs in the community change over time and ensure cost-effective and time-efficient approaches to completing projects. Meeting agendas will be organized by the PIs with input solicited from the SAB and EC membership.

C. Progress Reporting

Progress reporting by each workgroup will be led as described above. Dr. Rehm will supervise the collection of progress reports from workgroups and summarize progress for presentation at SAB and annual meetings. In addition, she will oversee the generation of progress reports for submission to NHGRI for annual grant renewal.

5. BIBLIOGRAPHY

1. Iafrate, A.J., L. Feuk, M.N. Rivera, M.L. Listewnik, P.K. Donahoe, Y. Qi, S.W. Scherer, and C. Lee, Detection of large-scale variation in the human genome. Nat. Genet., 2004. 36(9): p. 949-51.
2. Sayers, E.W., T. Barrett, D.A. Benson, E. Bolton, S.H. Bryant, K. Canese, V. Chetvernin, D.M. Church, M. DiCuccio, S. Federhen, M. Feolo, I.M. Fingerman, L.Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D.J. Lipman, Z. Lu, T.L. Madden, T. Madej, D.R. Maglott, A. Marchler-Bauer, V. Miller, I. Mizrachi, J. Ostell, A. Panchenko, L. Phan, K.D. Pruitt, G.D. Schuler, E. Sequeira, S.T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T.A. Tatusova, L. Wagner, Y. Wang, W.J. Wilbur, E. Yaschenko, and J. Ye, Database resources of the National Center for Biotechnology Information. Nucleic Acids Res, 2011. 39(Database issue): p. D38-51.
3. Sebat, J., B. Lakshmi, J. Troge, J. Alexander, J. Young, P. Lundin, S. Maner, H. Massa, M. Walker, M. Chi, N. Navin, R. Lucito, J. Healy, J. Hicks, K. Ye, A. Reiner, T.C. Gilliam, B. Trask, N. Patterson, A. Zetterberg, and M. Wigler, Large-scale copy number polymorphism in the human genome. Science, 2004. 305(5683): p. 525-8.
4. Itsara, A., G.M. Cooper, C. Baker, S. Girirajan, J. Li, D. Absher, R.M. Krauss, R.M. Myers, P.M. Ridker, D.I. Chasman, H. Mefford, P. Ying, D.A. Nickerson, and E.E. Eichler, Population analysis of large copy number variants and hotspots of human genetic disease. Am. J. Hum. Genet., 2009. 84(2): p. 148-61.
5. Sebat, J., B. Lakshmi, D. Malhotra, J. Troge, C. Lese-Martin, T. Walsh, B. Yamrom, S. Yoon, A. Krasnitz, J. Kendall, A. Leotta, D. Pai, R. Zhang, Y.H. Lee, J. Hicks, S.J. Spence, A.T. Lee, K. Puura, T. Lehtimaki, D. Ledbetter, P.K. Gregersen, J. Bregman, J.S. Sutcliffe, V. Jobanputra, W. Chung, D. Warburton, M.C. King, D. Skuse, D.H. Geschwind, T.C. Gilliam, K. Ye, and M. Wigler, Strong association of de novo copy number mutations with autism. Science, 2007. 316(5823): p. 445-9.
6. Cook, E.H., Jr. and S.W. Scherer, Copy-number variations associated with neuropsychiatric conditions. Nature, 2008. 455(7215): p. 919-23.
7. Baldwin, E.L., J.Y. Lee, D.M. Blake, B.P. Bunke, C.R. Alexander, A.L. Kogan, D.H. Ledbetter, and C.L. Martin, Enhanced detection of clinically relevant genomic imbalances using a targeted plus whole genome oligonucleotide microarray Genet. Med., 2008. 10(6): p. 415-29.
8. Green, E.D. and M.S. Guyer, Charting a course for genomic medicine from base pairs to bedside. Nature, 2011. 470(7333): p. 204-13.
9. Tonellato, P.J., J.M. Crawford, M.S. Boguski, and J.E. Saffitz, A national agenda for the future of pathology in personalized medicine: report of the proceedings of a meeting at the Banbury Conference Center on genome-era pathology, precision diagnostics, and preemptive care: a stakeholder summit. Am J Clin Pathol, 2011. 135(5): p. 668-72.
10. Bell, C.J., D.L. Dinwiddie, N.A. Miller, S.L. Hateley, E.E. Ganusova, J. Mudge, R.J. Langley, L. Zhang, C.C. Lee, F.D. Schilkey, V. Sheth, J.E. Woodward, H.E. Peckham, G.P. Schroth, R.W. Kim, and S.F. Kingsmore, Carrier testing for severe childhood recessive diseases by next-generation sequencing. Sci Transl Med, 2011. 3(65): p. 65ra4.
11. Mailman, M.D., M. Feolo, Y. Jin, M. Kimura, K. Tryka, R. Bagoutdinov, L. Hao, A. Kiang, J. Paschall, L. Phan, N. Popova, S. Pretel, L. Ziyabari, M. Lee, Y. Shao, Z.Y. Wang, K. Sirotkin, M. Ward, M. Kholodov, K. Zbicz, J. Beck, M. Kimelman, S. Shevelev, D. Preuss, E. Yaschenko, A. Graeff, J. Ostell, and S.T. Sherry, The NCBI dbGaP database of genotypes and phenotypes. Nat Genet, 2007. 39(10): p. 1181-6.
12. Fokkema, I.F., P.E. Taschner, G.C. Schaafsma, J. Celli, J.F. Laros, and J.T. den Dunnen, LOVD v.2.0: the next generation in gene variant databases. Hum Mutat, 2011. 32(5): p. 557-63.
13. Stenson, P.D., M. Mort, E.V. Ball, K. Howells, A.D. Phillips, N.S. Thomas, and D.N. Cooper, The Human Gene Mutation Database: 2008 update. Genome Med, 2009. 1(1): p. 13.
14. Faucett, W.A., S. Hart, R.A. Pagon, L.F. Neall, and G. Spinella, A model program to increase translation of rare disease genetic tests: collaboration, education, and test translation program. Genet Med, 2008. 10(5): p. 343-8.
15. Department of Health and Human Services, Human Subjects Research Protections: Enhancing Protections for Research Subjects and Reducing Burden, Delay, and Ambiguity for Investigators. Federal Register, 2011. 76(143): p. 44512-44531.

16. Miller, D.T., M.P. Adam, S. Aradhya, L.G. Biesecker, A.R. Brothman, N.P. Carter, D.M. Church, J.A. Crolla, E.E. Eichler, C.J. Epstein, W.A. Faucett, L. Feuk, J.M. Friedman, A. Hamosh, L. Jackson, E.B. Kaminsky, K. Kok, I.D. Krantz, R.M. Kuhn, C. Lee, J.M. Ostell, C. Rosenberg, S.W. Scherer, N.B. Spinner, D.J. Stavropoulos, J.H. Tepperberg, E.C. Thorland, J.R. Vermeesch, D.J. Waggoner, M.S. Watson, C.L. Martin, and D.H. Ledbetter, Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. Am. J. Hum. Genet., 2010. 86(5): p. 749-64.
17. Kearney, H.M., E.C. Thorland, K.K. Brown, F. Quintero-Rivera, and S.T. South, American College of Medical Genetics standards and guidelines for interpretation and reporting of postnatal constitutional copy number variants. Genet Med, 2011. 13(7): p. 680-5.
18. Richards, C.S., S. Bale, D.B. Bellissimo, S. Das, W.W. Grody, M.R. Hegde, E. Lyon, and B.E. Ward, ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. Genet Med, 2008. 10(4): p. 294-300.
19. Plon, S.E., D.M. Eccles, D. Easton, W.D. Foulkes, M. Genuardi, M.S. Greenblatt, F.B. Hogervorst, N. Hoogerbrugge, A.B. Spurdle, and S.V. Tavtigian, Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. Hum Mutat, 2008. 29(11): p. 1282-91.
20. 1000 Genomes Project Consortium, A map of human genome variation from population-scale sequencing. Nature, 2010. 467(7319): p. 1061-73.
21. Ring, H.Z., P.Y. Kwok, and R.G. Cotton, Human Variome Project: an international collaboration to catalogue human genetic variation. Pharmacogenomics, 2006. 7(7): p. 969-72.
22. Robinson, P.N., S. Kohler, S. Bauer, D. Seelow, D. Horn, and S. Mundlos, The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am J Hum Genet, 2008. 83(5): p. 610-5.
23. Cotton, R.G., A.D. Auerbach, J.S. Beckmann, O.O. Blumenfeld, A.J. Brookes, A.F. Brown, P. Carrera, D.W. Cox, B. Gottlieb, M.S. Greenblatt, P. Hilbert, H. Lehvaslaiho, P. Liang, S. Marsh, D.W. Nebert, S. Povey, S. Rossetti, C.R. Scriver, M. Summar, D.R. Tolan, I.C. Verma, M. Vihinen, and J.T. den Dunnen, Recommendations for locus-specific databases and their curation. Hum Mutat, 2008. 29(1): p. 2-5.
24. Giardine, B., J. Borg, D.R. Higgs, K.R. Peterson, S. Philipsen, D. Maglott, B.K. Singleton, D.J. Anstee, A.N. Basak, B. Clark, F.C. Costa, P. Faustino, H. Fedosyuk, A.E. Felice, A. Francina, R. Galanello, M.V. Gallivan, M. Georgitsi, R.J. Gibbons, P.C. Giordano, C.L. Harteveld, J.D. Hoyer, M. Jarvis, P. Joly, E. Kanavakis, P. Kollia, S. Menzel, W. Miller, K. Moradkhani, J. Old, A. Papachatzopoulou, M.N. Papadakis, P. Papadopoulos, S. Pavlovic, L. Perseu, M. Radmilovic, C. Riemer, S. Satta, I. Schrijver, M. Stojiljkovic, S.L. Thein, J. Traeger-Synodinos, R. Tully, T. Wada, J.S. Waye, C. Wiemann, B. Zukic, D.H. Chui, H. Wajcman, R.C. Hardison, and G.P. Patrinos, Systematic documentation and analysis of human genetic variation in hemoglobinopathies using the microattribution approach. Nat Genet, 2011. 43(4): p. 295-301.
25. Watson, M.S., G.R. Cutting, R.J. Desnick, D.A. Driscoll, K. Klinger, M. Mennuti, G.E. Palomaki, B.W. Popovich, V.M. Pratt, E.M. Rohlfs, C.M. Strom, C.S. Richards, D.R. Witt, and W.W. Grody, Cystic fibrosis population carrier screening: 2004 revision of American College of Medical Genetics mutation panel. Genet Med, 2004. 6(5): p. 387-91.
26. Faucett, W.A., The International Standards for Cytogenomic Arrays (ISCA) Consortium and its Genetic Counseling Workgroup Make Progress for Families and Genetic Counselors. Perspectives in Genetic Counseling, 2011: p. 4-5.
27. Kaminsky, E.B., V. Kaul, J. Paschall, D.M. Church, B. Bunke, D. Kunig, D. Moreno-De-Luca, A. Moreno-De-Luca, J.G. Mulle, S.T. Warren, G. Richard, J.G. Compton, A.E. Fuller, T.J. Gliem, S. Huang, M.N. Collinson, S.J. Beal, T. Ackley, D.L. Pickering, D.M. Golden, E. Aston, H. Whitby, S. Shetty, M.R. Rossi, M.K. Rudd, S.T. South, A.R. Brothman, W.G. Sanger, R.K. Iyer, J.A. Crolla, E.C. Thorland, S. Aradhya, D.H. Ledbetter, and C.L. Martin, An evidence-based approach to establish the functional and clinical significance of copy number variants in intellectual and developmental disabilities. Genet Med, 2011. 13(9): p. 777-784.

6. PROTECTION OF HUMAN SUBJECTS

6.1. RISKS TO THE SUBJECTS

6.1.a. Human Subjects Involvement and Characteristics:

The proposed research in this application meets the definition set forth in the Department of Health and Human Services regulation "Protection of Human Subjects" (45 CFR Part 46, administered by OHRP) as human subject research. In addition, sections of this project meet the definition of clinical research but not that of a clinical trial.

NIH requires education on the protection of human research participants for all individuals identified as Key Personnel. Up-to-date human subjects research certification will be provided should this proposal be selected for funding.

Collection of Data from Clinical Tests:

The goal of this project is to contribute clinical testing results to a common public database to improve understanding of genomic variation and the interpretation of results (benign or pathogenic). This "digital biobank" of samples are initially collected for clinical care to aid in the clinical interpretation of test results. They can be made available for research purposes in a HIPAA-compliant manner to stimulate additional discovery. If the future research use of the data is consistent with the Oct. 16, 2008 OHRP "Guidance on Research Involving Coded Private Information or Biological Specimens" [http://www.hhs.gov/ohrp/humansubjects/guidance/cdebiol.pdf], the NIH can make the clinical data available for research purposes and publicly display these data through the databases maintained by the National Center for Biotechnology Information (NCBI). The following Privacy Safeguards will be in place for all publicly shared data:

- All data samples will include the name of the testing laboratory and a laboratory provided random code for the individual sample.
- No HIPAA identifiers for the tested individual or information regarding the ordering physician will be included in the data submission.
- NCBI will never be provided with the keys to the codes.

This group is developing policies to facilitate the widest access for these data. At this time the group's policy will follow the following principles with regard to the submission of datasets that may be considered identifiable (whole or large genomic analyses):

1. Participating clinical testing laboratories will follow de-identification procedures as defined within the NIH GWAS Policy [http://grants.nih.gov/grants/gwas/gwas_ptc.pdf].
2. The group will develop procedures and informational documents, consistent with the previous points, for notification to the tested subjects or their legally authorized representative that:
   a) De-identified clinical data will be submitted and stored at the NIH for future distribution for research purposes
   b) Informational materials are available about future research (including the potential benefits and risks of research) to patients or their surrogates at the point of clinical care and subsequently through the NCBI resource
   c) There is a process for individuals to decide that their clinical data will not be submitted to NIH for research sharing, i.e. an "opt-out" procedure, and for their individual data to be removed from future distributions should they decide to opt-out at a later date. The materials for each of the participating laboratories will be shared and discussed with NIH as they are developed.
      a. All educational and informational materials and laboratory results reports from each of the consortium participants will contain language similar to the opt-out model provided below:

Sample language for the clinical submission form:

The requested clinical information is important for interpreting your patient's test result. Collected clinical information and test results will be included in a HIPAA-compliant public database as part of the National Institute of Health's effort to improve diagnostic testing and understanding of this disorder. Access to de-identified information allows researchers, clinicians, and other stakeholders to find relationships between genetic changes and clinical symptoms. Confidentiality of each sample will be maintained. Patients may withdraw consent for use of their data at any time by contacting our laboratory at XXX-XXX-XXXX. Refusal for inclusion in public de-identified databases may be indicated by checking this box. _____ If the box is not marked, consent is implied.

Sample language for the report form*:
_____retains patient samples indefinitely for validation, education purposes and/or research. Submitted clinical information and test results are included in a HIPAA- compliant public database as part of the National Institute of Health's effort to improve diagnostic testing and understanding of this disorder. Access to de-identified information allows researchers, clinicians, and other stakeholders to find relationships between genetic changes and clinical symptoms. Confidentiality of each sample is maintained. Patients may withdraw consent for the storage of their sample and use of their data by contacting our laboratory at XXX-XXX-XXXX.

***On the lab report this example includes the issue about sample retention and data retention/posting. You may want to discuss these separately, but this is an example that combines them and may save space.

3. Genomic data that group members contribute to NIH will include:
   a) Files containing the name and description of the specific test used by the contributing laboratory
   b) File containing structural or sequence-level variants per sample
   c) A combined list of genomic variation observed for all samples included in the data deposit, i.e. an aggregated list of all the clinically significant findings reported ("positives" and "VOUS")
   d) Identification of all known polymorphisms identified by the test
   e) Phenotype data at a minimum to include "affected" or "unaffected" for disease. In cases where parents or relatives of the affected individual are also tested, the relationships will be provided.
   f) The laboratory's protocols used to report findings (thresholds). Tests with both positive and negative findings will be included.
   g) All PubMed IDs and other sources used as references in the test interpretation for annotation

Participating clinical laboratories will follow the following procedures:

1. Clinical information will be requested at time of sample submission. An online clinical information submission form will be available to the clinician as part of the test-ordering process. The form will include information about opt-out of the deposition of data into the resources at NCBI and refer to the additional information available from (2) below.
2. Information about the submission of de-identified molecular results and clinical information being contributed to public databases will be posted on clinical testing laboratory websites along with test ordering information.
3. The clinical testing laboratory will discuss samples with a positive finding or a finding of unknown significance with the ordering clinician and additional clinical information may be requested from the clinician.
4. The policy on the posting of de-identified information will be discussed directly with the ordering clinician when contact is made for item (3). The availability of an opt-out option will be fully described for the clinician to appropriately advise the patient or the patient's guardian.
5. Tested samples with a negative result will be returned to the clinician without direct contact by the laboratory.

6. All test reports will include a statement about the submission of de-identified molecular and clinical information to the resources at NIH along with information about how to opt-out.
7. All test reports will include a statement about the desire to collect both clinical and molecular data along with instructions for how to provide such information.
8. De-identified, coded data will be securely transferred to the databases at NCBI under appropriate data security protocols.

The following procedures will be followed at NCBI to ensure appropriate use of the de-identified clinical and genomic data:

1. Authorized researchers present data access requests, including a brief research proposal, through the dbGaP data access process as for other genomic variation studies in dbGaP. NICHD already has provided the Data Access Committee (DAC) function for the prior structural variation project, which follows the policies and procedures established for GWAS datasets. We will pursue update of the Data Use Agreement if funding is awarded.
2. Researchers publish research results from the data. Published work forms the basis of future clinical decisions by the community per usual professional practices.
3. In the event a researcher wishes to locate specific individuals whose samples are in dbGaP to invite them to participate in an IRB-approved research protocol, the researcher must (and can only) contact the submitting testing laboratory. In turn, the testing laboratory can contact the ordering physician and convey the request. The ordering physician can then determine whether they feel it is appropriate to contact the patient with the request. The researcher is never given the patient’s contact information.

   The researcher’s contact information is given to the patient, mediated by the ordering physician. If the patient decides to participate in the research study, they contact the researcher directly and the informed consent process for the research protocol is conducted at that time.

6.1.b. Sources of Materials

This project is utilizing data generated from clinical or research based genomic testing of patient samples. As such, no new biological samples will be collected for the purpose of developing the database resource.

6.1.c. Potential Risks

The major risks for this project involve the release of identified clinical and genotype data. Multiple protections are built into the program to reduce this risk. There are no other significant risks associated with this project. Data submitted to the database will be de-identified.

6.2. ADEQUACY OF PROTECTION AGAINST RISKS

6.2.a. Recruitment and Informed Consent

Patient data will use an opt-out process for most large or whole genomic datasets. Data submitted from certain sources may have been obtained with full consent by an external collaborator (e.g. researcher or patient support organization). Multiple levels of protection will be in place as described in the section above describing data collection for the clinical cases.

6.2.b. Protection Against Risk

Our study personnel are experienced and well trained in human subjects research and collecting clinical phenotype data in order to minimize any potential risk to the family. Family data, both molecular and clinical, will be numerically coded, kept confidential and in locked filing cabinets or in HIPAA compliant electronic files. In all cases, personal identifiers will be kept physically separate from genotype and phenotype data.

6.3. POTENTIAL BENEFITS OF THE PROPOSED RESEARCH TO THE SUBJECTS AND OTHERS

We find that participation in studies such as the one proposed can provide psychological benefits, and families usually state that they appreciate being able to contribute to the advancing scientific knowledge and helping others. There are no significant risks to subjects in return for a study designed to yield answers to important clinical and biological questions about genomic variation in human health and disease.

6.4. IMPORTANCE OF THE KNOWLEDGE TO BE GAINED

The importance of understanding the contribution of genomic variation to human health and disease cannot be overstated; however, the role of individual variants, including structural and sequence-level variants, is often not clear in terms of contribution to phenotype. This project aims to develop a common environment to share genomic data and enable the robust assessment of human genomic variation.

7. INCLUSION OF WOMEN AND MINORITIES

Because we will be identifying patients for participation in this project primarily through clinical laboratories with diverse patterns of sample receipt, the planned distribution of subjects will be inclusive with regard to gender, minorities, and ethnic groups. Data from one of the large contributing laboratories shows a racial and gender distribution consistent with the US population. This project will not directly recruit patients but instead liaison with researchers and patient support organizations that may in turn consent patients for data submission to our resource. Through these relationships, we will ensure that patients are being recruited without exclusion of women or ethnic/racial minorities, unless disease-specific biases exist in certain situations (e.g. X-linked diseases). The Targeted/Planned Enrollment Table contains recruitment estimates for each category based on the referral pattern for clinical testing from the Harvard-Partners site based upon 60,000 probands tested. These data are likely to be representative of all sites.

8. TARGETED/PLANNED ENROLLMENT TABLE

The targeted enrollment table contains a minimum dataset based upon the data provided by many of the initial 36 laboratories who have agreed to participate in the eight model curation projects (see Table 1) as well as the estimated additional 42,000 cases to be recruited for the structural variant project. Race distribution is as described above in the “Inclusion of Women and Minorities” section.

Targeted/Planned Enrollment Table

This report format should NOT be used for data collection from study participants.

Study Title: Development and Curation of a Universal Human Genomic Variant Database

Total Planned Enrollment: 202,850

TARGETED/PLANNED ENROLLMENT: Number of Subjects			
Ethnic Category	**Females**	**Males**	**Total**
Hispanic or Latino	6,193	6,193	12,386
Not Hispanic or Latino	95,232	95,232	190,464
Ethnic Category: Total of All Subjects *	101,425	101,425	202,850
Racial Categories			
American Indian/Alaska Native	526	526	1,052
Asian	6,805	6,805	13,609
Native Hawaiian or Other Pacific Islander	179	179	358
Black or African American	9,348	9,348	18,697
White	84,567	84,567	169,134
Racial Categories: Total of All Subjects *	101,425	101,425	202,850

* The “Ethnic Category: Total of All Subjects” must be equal to the “Racial Categories: Total of All Subjects.”

9. INCLUSION OF CHILDREN

Individuals under the age of 21 are considered children according to the US Department of Health and Human Services guidelines for the conduct of human subject research. All sites submitting data accept samples from both adults and children and therefore data from all ages will be incorporated into the study.

10. VERTEBRATE ANIMALS

No vertebrate animals will be used in this project.

11. SELECT AGENT RESEARCH

Not applicable; no agents or toxins identified to have the potential to pose a biologic threat to public health and safety will be used in this project.

12. MULTIPLE PI LEADERSHIP PLAN

The rationale for having multiple PIs for this submission is that of complementary expertise and experience on the part of the PIs. This project represents a unique interface between the clinical laboratory and human research communities, with the most critical components of leadership required in the clinical laboratory sciences in order to organize and gain cooperation from this community. In addition, clinical genetic laboratories generating genomic data are split into two specialties, molecular genetics and cytogenetics. As such, success for this project is best achieved by having leadership in both of these specialties as represented by Drs. Rehm and Martin respectively. In addition, Dr. Nussbaum brings overall experience and long-standing leadership in research, administration, and medical genetics.

Heidi L. Rehm, PhD, FACMG, is Assistant Professor of Pathology at Brigham & Women’s Hospital and Harvard Medical School and an ABMG-certified clinical molecular geneticist. She is well-recognized in the research community for her long-standing contribution to the study of inherited hearing loss as well as in the clinical laboratory community having built a translational clinical laboratory from its inception in 2002. Recognition for her leadership and administrative skills has been shown by her appointed and elected roles including Director of the Laboratory for Molecular Medicine at the Partners Healthcare Center for Personalized Genetic Medicine, Director of the Clinical Molecular Genetics training program at Harvard Medical School, and past elected president of the New England Regional Genetics Group. She is also recognized as a thought leader in her fields of work having been invited to give over 36 presentations in national and international venues in the past 4 years on her work ranging from hearing loss and cardiomyopathy research, laboratory technology development, healthcare IT and personalized medicine as well as numerous advisory board and committee membership roles for both academic and commercial activities. She has longstanding experience interacting with NCBI on their RefSeqGene, Genetic Testing Registry and ClinVar projects. She also has significant experience in healthcare IT and software support for housing genetic variant knowledge having co-developed the GeneInsight Suite with her IT colleagues at Partners Healthcare which is now a commercial product in use in several clinical laboratories and being integrated into electronic health record environments.

As the lead principal investigator for this project, Dr. Rehm will be responsible for leading monthly meetings of the executive committee and annual meetings of all workgroup members. She will oversee all aspects of the sequence-level variation projects. She will also be responsible for the integration of all projects related to this proposal and will be the primary contact point for interfacing with large external organizations including NCBI, ACMG, CAP, and the FDA. She will also be the contact PI who will be responsible for all administrative issues and communication with the NIH and NHGRI. While Partners Healthcare will be the prime grantee, subcontracts will be established with nine other institutions and organizations and Dr. Rehm will oversee the distribution of funds necessary to support all activities associated with the project. Funding distributions have already been agreed upon and are well delineated in the budget justifications included in the application.

Christa Lese Martin, PhD, FACMG is Associate Professor of Human Genetics at Emory University School of Medicine and an ABMG-certified clinical cytogeneticist. She has a successful track record for laboratory-based genomics and molecular cytogenetics research of neurodevelopmental disabilities having successfully led multiple NIH- and private foundation-funded research projects. Her leadership and administrative organizational skills have been well developed as Operations Director of Emory Genetics Laboratory (EGL) and Co-Director of the Cytogenetics Laboratory within EGL. She has over 20 years of experience in the area of copy number variation and human genetic disease and her leadership in this field is now well recognized as co-founder of the International Standards for Cytogenomic Arrays (ISCA) Consortium, together with David Ledbetter, PhD, a resource now used by laboratories around the world.

As joint principal investigator Dr. Martin will oversee the structural variant portion of the project including directing the existing coordinating center for the International Standards for Cytogenomic Arrays (ISCA) Consortium. Dr. Martin will monitor the overall progress of the project, set project priorities, and supervise personnel. She will also co-chair the Executive Committee with Drs. Rehm and Nussbaum.

Robert Nussbaum, MD, FACP, FACMG is Professor of Medicine and Neurology and a senior member of the Institute for Human Genetics at the University of California, San Francisco School of Medicine. He is board certified (ABMG) in both Clinical Genetics and Clinical Molecular Genetics. In addition to practicing medical genetics for over 30 years, he has run a productive research laboratory that has made important contributions both to basic and translational research in human genetics. For example, his lab not only isolated the gene in which mutations cause Lowe syndrome by positional cloning, he and his colleagues also developed both biochemical and molecular testing for the condition and, as Chief of the Genetic Disease Branch in intramural NHGRI, established the first CLIA-certified laboratory anywhere in the intramural NIH so that he could provide diagnostic testing, carrier detection, and prenatal diagnosis for this condition. He also established and oversees a curated database of mutations for this condition. As an expert on Parkinson disease genetics, he has participated in large collaborative efforts to establish phenotypic characterization of disease, such as the distinction between Parkinson disease, Parkinson disease with dementia, and diffuse Lewy body disease. Finally, he established two specialty clinics at UCSF, one in cardiovascular genetics, the other in complex hereditary cancer syndromes, and has been obtaining, interpreting, and applying molecular diagnostic results for patients and physicians in these two areas for the past six years.

As joint principal investigator Dr. Nussbaum will monitor the overall progress of the project and ensure that the project is meeting the needs of the clinical and research communities as they evolve. He will also interface with senior leadership at NIH to ensure that the project is conforming to the goals and priorities of NIH and NHGRI. He will also co-chair the Executive Committee with Drs. Rehm and Martin.

Resolution of Disputes

Drs. Rehm, Martin and Nussbaum share many common professional interests and in organizing the project have had no disputes regarding the overall project goals and methods to achieve them. Therefore, no significant dispute is expected to arise during the course of the proposed project. In the very unlikely event that a dispute does arise, the three PIs will seek input from the NHGRI program office and then discuss the issue in person to attempt to reach an immediate and satisfactory resolution to the matter in question. If disagreement remains among the three PIs, Dr. Nussbaum will be asked to make the final decision based on his long-standing well-respected senior leadership in the field. If a dispute exists between the PIs and the NHGRI program office, the PIs will follow a similar approach and only if necessary, engage NIH's Dispute Resolution process.

Responsibility for regulatory compliance and human subjects protection

All three principal investigators will be responsible for ensuring that the appropriate systems are in place to guarantee institutional compliance with US laws, DHHS and NIH policies throughout the project at their respective institutions and at the subcontract. This includes the obtainment of all human subject approvals and ensuring the ongoing protection of human subjects.

13. CONSORTIUM/CONTRACTUAL ARRANGEMENTS

This proposal involves the **Brigham and Women's Hospital (BWH),** acting on behalf of the **Partners Center for Personalized Genetic Medicine (PCPGM)** as the primary/grantee institution and **ARUP Laboratories**, **Children's Hospital Boston**, **Emory University**, the **Geisinger Clinic**, **GeneDx**, the **Mayo Clinic**, the **MutaDATABASE Foundation (Belgium)**, the **University of California San Francisco**, and the **University of Chicago** as consortium/subcontract institutions. All institutions involved in this proposal have agreed to the following statements:

"The appropriate programmatic and administrative personnel of each institution involved in this grant application are aware of the PHS-NIH consortium grant policies and are prepared to establish the necessary inter-institutional agreements consistent with those policies. The Cooperating Institution certifies it has implemented and is enforcing a written policy of conflicts of interest consistent with the provisions of 42 CFR Part 50, Subpart F & 45 CFR Subtitle A, Part 94. If a conflict is identified by the Cooperating Institution during the period of the award contemplated under this agreement, the Cooperating Institution will report to the Prime Awardee the existence of the conflict, including the grant title, PI (if different from the investigator with the financial interest) and the specific method the Cooperating Institution adopts for addressing the conflict (managing, reducing, or eliminating it). The Cooperating Institution will rely on the Prime Awardee to report the existence of the conflict to NIH."

ARUP Laboratories

The consortium with ARUP Laboratories will be directed by **Dr. David K. Crockett** and will have as co-investigators **Drs. Elaine Lyon, David Crockett, Sarah South and Rong Mao.** In addition to particpating as a core laboratory submitting both structural and sequence-level variants, ARUP will also be responsible for directing a model curation project covering genes for metabolic disorders.

Children's Hospital Boston

At Children's Hospital Boston, **Dr. David Miller** will serve as consortium PI and he will be responsible for coordinating and executing strategies to collect and integrate phenotypic data into the database.

Emory University

The consortium with Emory University will be directed by **Dr. Christa Lese Martin**, joint principal investigator of this proposal. Emory University will serve as both the coordinating center for the International Standards for Cytogenomic Arrays (ISCA) Consortium (the structural variation portion of the overall project), as well as a core laboratory contributing both structural and sequence-level variants to the database. In addition, Emory University will contribute significant effort to the sequence-level variation portion of the study in which **Dr. Madhuri Hegde** will be a key investigator including directing several model curation projects. Emory University will also coordinate the efforts of several key consultants to the overall project (i.e., Cartagenia and Dr. Hutton Kearney).

Geisinger Clinic

The consortium with the Geisinger Clinic will be headed by **Mr. W. Andrew Faucett** and have as a co-investigator **Dr. David Ledbetter**. Geisinger Clinic will direct and coordinate the activities of the Education, Engagement and Ethics Workgroup (Aim 4 of the project proposal). Mr. Faucett will direct the activities of multiple genetic counselors from each of the participating cytogenetic and molecular laboratories, and along with other Geisinger Clinic staff, will coordinate the efforts of several key consultants to the overall project who will assist in efforts to increase the collection of phenotypic information by clinicians and increase the contribution of phenotypic information to public patient registries.

GeneDx

Dr. Sherri J. Bale will serve as the consortium PI at GeneDx, which will serve as one of the Core Laboratories for the submission of structural and sequence variation. Dr. Bale will also oversee one of the model curation projects on genes involved in Noonan spectrum disorders. Further, **Dr. Swaroop Aradhya** will serve as co-investigator and will serve on the curation group for the ISCA evidence-based genotyping group and assist the classification group for sequence variants (non-CNV).

Mayo Clinic
The Mayo Clinic consortium will be led by **Dr. Matthew Ferber** and **Dr. Erik Thorland**. Mayo Clinic will serve as one of the Core Laboratories for the submission of structural and sequence variation. Dr. Ferber will also direct one of the model curation projects for inherited colon cancer genes and Dr. Thorland will direct the Structural Variant workgroup.

University of California San Francisco
The consortium with UCSF consists of **Dr. Robert Nussbaum** in support of his leadership role as joint principal investigator.

University of Chicago
Dr. Soma Das will direct the consortium with the University of Chicago, which will participate as a core lab to contribute sequence-level variants to the database. Dr. Das will also direct one of the model projects on EIEE/Rett/Angelman syndromes.

MutaDATABASE Foundation
Dr. Patrick Willems, will serve as the consortium PI for the MutaDATABASE Foundation. He and his team will participate in and oversee efforts to enable a robust data centralization and curation process for genetic data that will ultimately be deposited in ClinVar. He will manage interactions with thousands of gene curators and data submitters sending data into MutaDATABASE and ensure the transfer of data from MutaDATABASE into ClinVar.

Foreign Work
MutaDATABASE is a resource that is a foreign component to this domestic U41 application. The decision to engage the work performed by the MutaDATABASE Foundation is based on two major reasons. The MutaREPORTER software to support the automated submission and assisted community curation of variants in a single system with visualization in a genome browser has not been developed anywhere in the US. The MutaDATABASE is a standardized universal database that is exponentially growing since its start 2 years ago, currently harboring data on more than 14,000 variants possibly associated with genetic disease. Dr. Willems has already engaged over 500 gene curators willing to work on the project, and over 100 labs willing to submit data to the database. The MutaDATABASE advisory board consists of more than 80 renowned scientists and lab Directors. Although the ultimate repository for data will be ClinVar at NCBI in the USA, curation infrastructure is not yet in place at NCBI, thereby requiring an alternate approach to data curation offered.

14. RESOURCE SHARING PLAN

The investigators fully endorse the National Institutes of Health's (NIH) goals of sharing unique research resources arising from NIH-funded research within the scientific community. Genomic and phenotypic data will be deposited in the Database of Genotypes and Phenotypes (dbGaP) and other databases at NCBI. We agree that the existing mechanisms and protocols for data release already established at NIH will be employed in the dissemination of our data.

The investigators will establish a policy and infrastructure for distributing research materials and transfers of material and such procedures will be conducted in a manner consistent with NIH guidelines with respect to the availability of research tools and in accordance with federal laws and regulations. For the purposes of any material arising from this grant, when providing or licensing materials to for-profit organizations, the investigators will distinguish the use of materials for internal research use from the right to use such materials for commercial development and sale. Transfers of material to not-for-profit entities generally are conducted using a material transfer agreement (MTA), which contain terms no more restrictive than a Simple Letter Agreement.

The investigators understand and firm support that technology arising from NIH-funded research should remain available and accessible to the research community. When licensing technology covered by pending patent applications or issued patents, such agreements include a reservation of rights provision, which ensures that such technology can be used both by the investigators and by third parties for educational and academic research purposes.

The proposed project has been designed with an aim of developing the appropriate resources to facilitate and foster continued collection of genomic variation and phenotype data for a permanently accessible registry of genetic variation. In addition, we will support software to facilitate the collection, validation and de-identification of the data prior to upload to the dbGaP and other repositories at NCBI. We will work collaboratively with investigators evaluating clinical genomic variation data in developing standardized genotype and phenotype nomenclature so that linkages between these data repositories will facilitate investigative and clinical usage. We will also develop a follow-up system to evaluate the long term phenotypic effects of genomic variation findings of uncertain clinical significance.

The investigators recognize that rights and privacy of subjects who participate must be protected at all times. Therefore, care will need to be taken to ensure that data shared will be free of identifiers that would permit linkages to individual research participants and variables that could lead to disclosure of individual subjects. Since the initial design of the proposed study incorporates data sharing, the proposal more readily and economically establishes adequate procedures for protecting the identities of participants and therefore a useful data set with appropriate documentation.

Patients are permitted to opt-out of having their de-identified data stored. There will be an option on the clinical requisition form for patients to opt-out of this storage of data. Patients that we wish to gather full phenotypic information on for follow-up purposes will be asked to sign a separate consent form.

In general, the data will be evaluated using accepted statistical and scientific principles and methods to ensure that the risk of re-identification is very small. The methods and procedures used will be documented per the HIPAA requirements.

15. LETTERS OF SUPPORT

A copy of an e-mail from Dr. Lisa Brooks of the NHGRI to Dr. Heidi Rehm, the contact PI of this proposal, is included in the following pages. Dr. Brooks' letter confirms that the NHGRI will accept, review and consider this proposal at the funding level requested, which is greater than $500,000 in direct costs in each of the three years of the proposed project period.

In addition, the following individuals have provided letters of support and commitment to this project proposal:

Leading scientists who support the need for this grant

David M. Altshuler, M.D., Ph.D.	Co-Chair, 1000 Genomes Project Deputy Director and Chief Academic Officer The Broad Institute of Harvard and MIT
George M. Church, Ph.D.	Director, Harvard NHGRI-Center of Excellence in Genomic Science Professor, Harvard Medical School and The Broad Institute
Evan E. Eichler, Ph.D.	Professor of Genome Sciences, University of Washington Investigator, Howard Hughes Medical Institute
James P. Evans, M.D., Ph.D.	Editor-in-Chief, Genetics in Medicine Professor, University of North Carolina at Chapel Hill
Robert Green, M.D., M.P.H.	Associate Director, Partners Center for Personalized Genetic Medicine Associate Professor, Harvard Medical School
Isaac S. Kohane, M.D., Ph.D.	Co-Director, Harvard Center for Medical Bioinformatics Professor, Harvard Medical School
Eric S. Lander, Ph.D.	President and Director, The Broad Institute of Harvard and MIT Professor, Harvard Medical School and MIT
Richard P. Lifton, M.D., Ph.D.	Professor and Chair of Genetics, Yale University Investigator, Howard Hughes Medical Institute
Stephen Scherer, Ph.D.	Director, Centre for Applied Genomics in the Hospital for Sick Children Professor, University of Toronto, Canada
Christine E. Seidman, M.D. and Jonathan G. Seidman, Ph.D.	Professors, Harvard Medical School Investigators, Howard Hughes Medical Institute
Scott T. Weiss, M.D.	Director, Partners Center for Personalized Genetic Medicine Professor, Harvard Medical School

Professional organization leaders who will collaborate on the project

Richard G.H. Cotton, Ph.D., D.Sc.	Scientific Director, Human Variome Project Director, Genomic Disorders Research Centre, University of Melbourne
Wayne Grody, MD	President, American College of Medical Genetics Professor, University of California Los Angeles
James M. Ostell, Ph.D.	Chief, Information Engineering Branch National Center for Biotechnology Information, NIH
Roberta A. Pagon, M.D.	Principal Investigator, GeneTests; Editor-in-Chief, GeneReviews Professor, University of Washington
Stanley J. Robboy, M.D., FCAP	President, College of American Pathologists Professor and Vice Chair for Diagnostic Pathology, Duke University
Patrick J. Willems, M.D., PhD.	Principal Investigator, The MutaDATABASE Foundation

Consultants

Leslie G. Biesecker, M.D.	Senior Investigator, National Human Genome Research Institute, NIH Chief, Genetic Disease Research Branch Director, Physician Scientist Development Program
Johan T. den Dunnen, Ph.D.	Board Member, Human Genome Variation Society Professor, Leids Universitair Medisch Centrum
Ada Hamosh, M.D., M.P.H.	Scientific Director, Online Mendelian Inheritance in Man Clinical Director, McKusick-Nathans Institute of Genetic Medicine Professor, Johns Hopkins School of Medicine
Laird Jackson, M.D.	Professor of Genetics, Drexel University College of Medicine
Stephen F. Kingsmore, M.B., Ch.B., D.Sc..	Director, Center for Pediatric Genomic Medicine Children’s Mercy Hospitals and Clinics, Kansas City, MO
Sue Richards, Ph.D.	Director, Clinical Molecular Genetics Professor, Oregon Health & Science University
Peter N. Robinson, M.D., M.Sc.	Founder, Human Phenotype Ontology Professor, Institut für Medizinische Genetik Universitätsmedizin Charité, Berlin
Joan A. Scott, M.S., CGC	Executive Director, National Coalition for Health Professional Education in Genetics
Lisa Salberg	Founder and CEO, Hypertrophic Cardiomyopathy Association
Sharon Terry	President and CEO, Genetic Alliance

Patient advocacy and registry leaders who will collaborate on this project

Kyle Brown	CEO, Patient Crossroads
Wanda Robinson	President and Director, Noonan Syndrome Support Group, Inc.
Anne Rutkowski, M.D.	Co-Founder and Vice Chairman, CureCMD
Beverly Searle, Ph.D.	CEO, UNIQUE Rare Chromosome Disorder Support Group

See also Letters from Consultants Sharon Terry and Lisa Salberg

Laboratory directors and gene experts who will contribute data and/or be an expert curator

Margaret Adam, M.D.	Associate Professor, University of Washington Clinical Geneticist, Seattle Children’s Hospital
Sherri J. Bale, Ph.D.	Managing Director, GeneDx
Pinar Bayrak-Tordemir, M.D., Ph.D.	Medical Director, Molecular Genetics and Genomics Section ARUP Laboratories
Professor John Christodoulou, AM, M.B.B.S., Ph.D.	Head, Centre for Rett Syndrome Research, Professor, University of Sydney, AUSTRALIA
Soma Das, Ph.D.	Director, Molecular Diagnostic Laboratory, University of Chicago
Charis Eng, M.D., Ph.D.	Chairman & Director, Genomic Medicine Institute Professor, Cleveland Clinic
Ping Fang, Ph.D.	Co-Director, Medical Genetics Laboratories, Baylor College of Medicine
Gerald L. Feldman, M.D., Ph.D.	Medical Director, Molecular Genetics Diagnostic Laboratory Wayne State University School of Medicine
Matthew J. Ferber, Ph.D.	Co-Director, Molecular Genetics Laboratory, Mayo Clinic
Michael J. Friez, Ph.D.	Director, Diagnostic Laboratories, Greenwood Genetic Center

Julie M. Gastier-Foster, Ph.D.	Director, Cytogenetics/Molecular Genetics Laboratory Nationwide Children’s Hospital (Columbus, OH)
Vicky L. Funanage, Ph.D.	Director, Molecular Diagnostic Laboratory Alfred I. duPont Hospital for Children
Cheryl Shoubridge, Ph.D.	Head, Molecular Neurogenetics Women’s and Children’s Hospital (Adelaide, AUSTRALIA)
Bruce D. Gelb, M.D.	Director, Child Health and Development Institute Professor, Mount Sinai School of Medicine
Madhuri Hegde, Ph.D.	Scientific Director, Emory Genetics Laboratory, Emory University
Julie R. Jones, Ph.D.	Director, Molecular Diagnostic Laboratory, Greenwood Genetic Center
Ruth Kornreich, Ph.D.	Laboratory Director, Molecular Genetics Laboratory Mount Sinai School of Medicine
Elaine Lyon, Ph.D.	Medical Director, ARUP Laboratories
Professor Finlay Macrae, M.D.	Honorary Secretary, International Society for Gastrointestinal Tumours
Rong Mao, M.D.	Medical Director, Molecular and Genomics Section ARUP Laboratories
Kristin G. Monaghan, Ph.D.	Director, DNA Diagnostic Laboratory, Henry Ford Hospital
O. Thomas Mueller, Ph.D.	Director, Biochemical and Molecular Genetics All Children’s Hospital (St. Petersburg, FL)
Devin Oglesbee, Ph.D.	Co-Director, Biochemical Genetics Laboratory, Mayo Clinic
Simon Ramsden, FRCPath	Regional Genetics Laboratory Services Central Manchester University Hospitals, UK
Amy Roberts, M.D.	Cardiovascular Genetic Division, Children’s Hospital Boston and Harvard Medical School
Juan-Sebastian Saldivar, M.D.	Director, Molecular Diagnostics City of Hope National Medical Center (Duarte, CA)
Benjamin A. Salisbury, Ph.D.	Vice President, Clinical Genetics, Transgenomic, Inc.
Warren G. Sanger, Ph.D.	Director, Human Genetics Laboratory, Univ. of Nebraska Medical Center
Avni Santani, Ph.D.	Scientific Director, Molecular Genetics Laboratory, Children’s Hospital of Philadelphia
Carol Saunders, Ph.D.	Director, Molecular Laboratory, Pediatric Genome Center Children’s Mercy Hospitals and Clinics (Kansas City, MO)
Katia Sol-Church, Ph.D.	Director, Biomolecular Core Laboratory, Jefferson Medical College
Catherine A. Stolle, Ph.D.	Director, Molecular Genetics Laboratory Children’s Hospital of Philadelphia
Charles Strom, M.D., Ph.D.	Senior Medical Director of Genetics, Quest Diagnostics, Inc.
Sharon F. Suchy, Ph.D.	Director, Division of Inherited Metabolic Disorders, GeneDx
Jack Tarleton, Ph.D.	Director, Fullerton Genetics Laboratory Mission Hospital (Asheville, NC)
Marco Tartaglia, Ph.D.	Section Director, Molecular Medicine, Istituto Superiore di Sanità, Italy
Edwin Trautman, Ph.D.	Director of Clinical Bioinformatics, Correlagen Diagnostics, Inc. (LabCorp)
Patrick J. Willems, M.D., Ph.D.	Director, GENDIA
Martin Zenker, PD Dr. med.	Professor, Institute of Human Genetics, University of Magdeburg, Germany
Kejian Zhang, M.D., M.B.A.	Director, Molecular Genetics Laboratory, Cincinnati Children’s Hospital

Contractual arrangements

Patrick J. Willems, M.D., PhD. Director, MutaBASE

APPENDIX – See included CD