A Unified Clinical Genomics Database

Submitted on 07/03/12 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: 4/01/13 – 3/31/14

Principal Investigators
David H. Ledbetter, PhD
Christa Lese Martin, PhD
Joyce A. Mitchell, PhD
Robert L. Nussbaum, MD
Heidi L. Rehm, PhD

INTRODUCTION

The initial review of this application acknowledged that a large amount of valuable data exists in databases outside the public domain and "a centralized database giving access to clinically informative variant information would have considerable value and be widely used." However, several appropriate criticisms were raised, and we have significantly modified the proposal in response. Our modifications are described below; specific changes are not indicated in the text, as they were too extensive.

Several reviewers raised concerns about **scalability and sustainability**, particularly related to curation. With regards to scalability, we emphasize that the 6 labs directly supported by this grant plan to submit many more variants than just those in the genes involved in the curation projects described in Aim 3. For sequencing, the 6 supported labs have agreed to submit over 73,000 variants in more than 700 genes during the course of testing ~250,000 cases. The additional 42 laboratories that have agreed to participate will add even more variants, greatly increasing the availability of data for research and clinical use. We have provided a more detailed explanation in Aim 2 of the development of an automated and efficient data submission process. Regarding curation, we are taking advantage of the expert curation that clinical laboratories are already doing as part of routine care, which is paid for by clinical revenues and not this grant. As described in Aim 3, multi-lab curation, not requiring grant funding, will also improve the quality of interpretations by enabling labs to compare their data and share analyses. In summary, even in the absence of the expert curation activities described in Aim 3, the collected data will provide an enormously useful curated dataset with built-in sustainability because it leverages ongoing clinical testing. With rapid movement toward genomic sequencing in patients, a clinical grade variant database drawing from many sources will be an absolute necessity to interpret the hundreds of thousands of rare variants found in each patient tested.

Reviewers also raised concerns about insufficient **IT effort** in building the interfaces between the laboratories and NCBI. In response, we recruited Dr. Joyce Mitchell, a leader in biomedical informatics, to join the team as a Principal Investigator, and we have included additional experienced IT staff. To better coordinate the IT activities and interaction with NCBI, we created a Bioinformatics and IT Workgroup, bringing together core project staff, collaborators at NCBI, and staff at the funded laboratories. We have also added funding for staff members with IT and genetics expertise to travel to laboratory sites to better engage them in the project and co-develop approaches for efficient data submission.

Concern was also raised regarding our ability to ensure **adoption of standardized terminologies and data formats**. We agree that a centralized variant database with standardized vocabulary and formats is necessary to effectively compare data across laboratories. Most laboratories have already adopted the basic standards developed by the ACMG; these are being further refined with participation of our group (See Aim 1). Furthermore, we have agreement from CAP, the primary accrediting body for clinical laboratories, to enforce the use of standards developed by the project (see letter of support from CAP).

Reviewers expressed concern about the lack of emphasis on genetic variants underlying **common diseases and pharmacogenetics**. The clinical utility of genetic testing for most common diseases with complex inheritance is still evolving. Therefore, we decided to focus initially on the rare variants with clear clinical validity and proven utility in widespread clinical use today. Pharmacogenetics is a special case as there is increasing adoption of pharmacogenetic testing in clinical laboratories. We therefore plan to work with the PharmGKB team (see letters of support from the PharmGKB leadership and Dan Roden, chair of the PGRN) to determine the best approach for inclusion and curation of pharmacogenetic variants in the centralized database at ClinVar.

In terms of **phenotype data collection**, reviewers asked why we had not proposed integration with electronic medical record (EMR) systems. Although we recognize the difficulty of robust phenotypic data collection, we do not think that focusing on EMR integration will be cost effective initially, as current EMRs often lack structured data relevant to the annotation of genetic variation. Instead, we plan to supply more robust front end tools for capture of the phenotypic information submitted by clinicians to clinical laboratories, as we have done successfully in the ISCA project. To enhance this information, we have added a patient registry enabling clinicians and patients (the most motivated participants) to contribute to data collection and providing researchers with a direct connection to patients for follow-up on phenotypes and genetic findings (See Aim 2B). EMR integration will be a future goal of the project.

Reviewers requested that we expand on how the data would address problems in **genomic medicine**. In response, we have expanded the section entitled "Anticipated impact of the resource on biomedical research."

Some concern was raised regarding the **investigators,** particularly in the area of informatics expertise and documented past history of working together. In response, we have recruited Dr. Joyce Mitchell and appointed Dr. David Ledbetter as Principal Investigators. Four members of this team have already been working closely together for over a year, including organizing a joint conference bringing the relevant experts together at the ISCA Conference in May 2012. The leadership team has also been meeting weekly by teleconference to plan the project. We feel strongly that this team is ideally suited to carry out the goals of the proposed project.

SPECIFIC AIMS

Genomic variation underlies almost all human disease. Technological advances are quickly making variant detection across the whole genome commonplace in the medical care environment, sparking an expansion of both basic and clinical research. At this time, however, our ability to detect DNA variation has greatly surpassed our ability to interpret the clinical validity of these variants, particularly given the lack of publicly available, carefully curated information on variants and phenotypic consequences. Although collections of some disease-associated variants are publicly available in Locus Specific Databases (LSDBs), the data are often subject to inconsistent standards and have limited accuracy. In contrast, carefully curated disease-associated variants from patient populations evaluated by individual clinical laboratories are often sequestered and unavailable to the community. To address these challenges, we propose to create a database of primarily clinical grade variation data across the genome that concentrates the knowledge and curation capabilities of our clinical laboratory community into a single environment. Capturing the large number of clinical genetic tests being performed on patients with disease phenotypes presents a unique opportunity to contribute to our understanding of the functional significance of human genomic variation, which will benefit the growing community of medical genomics clinicians and researchers. We have already initiated such an effort for the structural variant community by establishing the International Standards for Cytogenomic Arrays (ISCA) Consortium, and this proposed project will build upon the Consortium's successful foundation to incorporate sequence-level variation. The specific aims to achieve this goal are:

1) Develop a standardized infrastructure for data acquisition, submission and public access for a clinical genomic variation database. To support the consistent description, annotation, and clinical classification of genomic variants, a data element dictionary, already in use, will be refined by the Structural and Sequence Variant Workgroups. The Phenotyping Workgroup will refine data dictionaries for phenotypic information utilizing standardized formats, such as the Human Phenotype Ontology. These standards will ensure the uniformity and integrity of the data and facilitate data transfer across all systems leading to a sustainable resource. Policies regarding public access will balance clinical and research utility with patient privacy. To stimulate data acquisition, an Engagement, Education and Access Workgroup will develop educational materials aimed at clinicians, clinical genetics laboratories, and patients/families emphasizing the importance of high-quality genomic and phenotypic data for basic science, public health, and improved patient care.

2) Coordinate the submission of variant and phenotypic data into ClinVar, a unified database at NCBI.
a) **Submission of variant data into ClinVar.** A large consortium of highly experienced clinical laboratories has committed to submitting genomic data (from clinical cytogenomic microarray, single gene, gene panel, and whole exome/genome sequencing testing) and basic phenotypic data into ClinVar with a minimum estimated contribution of data from over 300,000 cases. A Bioinformatics and Information Technology Workgroup (BITW) will oversee data submission and the development of software bridges to current analysis programs in clinical laboratories to create simple, automated mechanisms to de-identify and reformat datasets for submission. Workflow-integrated solutions will enable scalability of the project, and will also ensure sustainability. **b**) **Submission of phenotype data into ClinVar.** Recognizing the importance of phenotypic information in variant interpretation and curation, we will develop point of care tools for clinicians, improving their ability to submit phenotype data to laboratories at the time of test ordering. In addition, we will develop an online patient registry using validated questions shown to have high levels of accuracy to capture enhanced phenotypic information from both clinicians and patients. This novel approach will also allow patients to indicate interest in research. Registered researchers will be able to identify cases for genotype-phenotype discovery efforts and clinical trials. The BITW will address issues inherent with linking phenotypic information to genotype information within ClinVar.

3) Implement sustainable expert clinical level curation systems for human genomic variants. There is a lack of consensus within the field regarding the most efficient way to curate genomic data. Therefore, we will investigate various approaches for expert clinical curation through demonstration projects in different disease areas, with the ultimate goal being the development of efficient, sustainable systems for the curation of genome-wide variation data. We will explore a combination of automated and expert-level curation efforts designed to achieve genome-wide scalable curation solutions.

At the conclusion of this project, we will deliver a fully functional, unified clinical genomics resource for the ongoing collection, curation, and sharing of human genomic variation, developed by extracting rich but underutilized sources of information buried in clinical laboratories and clinical encounters and making it publicly available. 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 both the database and patient registry to ensure their ongoing success. The unified clinical genomics resource we propose is critical to assessing the clinical validity of variants which, in turn, is essential and therefore highly complementary to the assessment of clinical utility, such as that being proposed under the Clinically Relevant Variants Resource initiative.

RESEARCH STRATEGY - OVERVIEW

Determination of the biological and medical significance of structural (copy number) and sequence-level variation in the human genome requires extensive investigation of both normal and disease populations. Technological advances are quickly making variant detection across the whole genome commonplace in the medical care environment, sparking an expansion of both basic and clinical research. In the clinical arena, whole genome structural and sequence analyses are rapidly being integrated into routine clinical care, providing a timely opportunity for capturing genomic variation from thousands of individuals during the course of clinical genetic testing (i.e., essentially "free" data as a byproduct of patient care).

The crux of this application is the formation of an international consortium of cytogenomic, molecular and clinical geneticists based in diagnostic laboratories, clinical facilities, and research laboratories to harness such genomic variation and phenotypic information into a **unified**, **curated, publically available database.** Data collected, submitted, and curated through this project will be housed within **ClinVar**, a new database being launched within the National Centers for Biotechnology Information (NCBI). We will utilize new and existing software specifically designed for data submission and curation purposes to develop an open resource for understanding genomic variation. In addition, a main objective is to build all aspects of the proposed structures with an eye to minimizing the need for costly upkeep, the ultimate goal being broad sustainability through routine use in the course of clinical care. This effort will be critical to provide clinicians and researchers with the information needed to interpret the enormous levels of variation present in human genomes. Such a resource would also benefit professional organizations seeking to define guidelines on how to use genetic information in clinically useful ways.

We have already initiated such an effort for the structural variant community by establishing the International Standards for Cytogenomic Arrays (ISCA) Consortium, and this proposed project will build upon the Consortium's successful foundation to incorporate sequence-level variation. Through this project, we will deliver a fully functional resource for the ongoing collection, curation, and sharing of human genomic variation. In addition, this novel resource is an innovative approach to acquiring the variant and phenotype data that is both essential and complementary to the recently proposed "Clinically Relevant Variants Resource: A Unified Approach for Identifying Genetic Variants for Clinical Use."

1. Rationale for the community resource

The current landscape of human genomic variation databases includes those focused on variation as it relates to normal populations, association with common diseases (e.g., variation from genome-wide association studies), pharmacologic responses, and limited sets of rare variation. Each of these types of databases provides vital information when trying to decipher the functional significance of human genomic variation; however, the study of variation as it relates to specific disease populations has the greatest potential to provide information on genotype-phenotype correlations amongst rare but highly penetrant traits, information that could be broadly applied to more common diseases. Effective studies of this type of variation require large datasets from both normal (control) and disease populations.

Several centralized, public databases of human variation in the normal population now exist, including the Database of Genomic Variation (DGV) [1], the 1000 Genomes Project [2], the Exome Sequencing Project [3] and the Single Nucleotide Polymorphism Database (dbSNP) [4]. In contrast, there is no single, comprehensive (including **both** sequence and structural variation), freely accessible database of variation from disease populations. Though freely accessible and curated databases of **structural variation** are in development, including our ongoing effort, the ISCA Consortium database, most publicly available **sequence-level genomic variation** from disease populations is either buried in publications or contained within diverse Locus Specific Databases (LSDBs), of which over 1,500 exist (http://www.hgvs.org/dblist/glsdb.html#). These 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: some data is deposited without any, or with an incorrect, interpretation of the functional or clinical significance [5]; and 5) access: the data is not easily accessible to the research and clinical communities, particularly for use in high-throughput genomic analyses.

Although some variant data from disease populations exists within the Online Mendelian Inheritance in Man (OMIM) database (www.omim.org), a catalogue of human phenotypes and genes, **only a small subset of gene variation** is typically documented within OMIM. The variants catalogued within OMIM are deliberately selected to represent exemplary types of mutations that have led to the understanding of phenotype, patterns of inheritance and mechanisms of gene dysfunction, not necessarily to document **all** variation within a gene. An additional resource for

human genetic variation in disease populations is the Human Gene Mutation Database (www.hgmd.org). This resource, however, is **not freely available**, and **only represents published data**, which is a much smaller proportion of the actual amount of variation being observed in clinical laboratories. Further, approximately 30% of mutations within these databases annotated from the literature are in non-standardized formats, and/or are labeled with incorrect clinical assertions [5]. One reason for this high error rate is that incorrect initial variant interpretations are only revised when an erratum to the original manuscript is published.

An alternative approach is the development of a consortium of clinical genetic testing laboratories vested in sharing variant data in a single, open access database. Such an approach would harness the wide phenotypic and genotypic variant spectrums observed in clinical laboratories, and take advantage of data already being generated during the course of routine clinical care. Historically, it has been difficult to identify and collect large groups of individuals with specific phenotypes, particularly with Mendelian genetic traits, from an unselected population. Clinical genetic testing laboratories, however, are uniquely poised to provide large volumes of variant data on affected populations, as these individuals constitute their primary sample base.

Currently, most clinical genetic testing is carried out on individuals affected with demonstrable phenotypes, including (but not limited to) neurodevelopmental disorders, congenital anomalies, cancers, or any number of single or multisystem phenotypes. Individual phenotypes may be rare (e.g., Tetralogy of Fallot) or relatively common (e.g., breast cancer). For all of these individuals, their specific phenotypes may suggest a genetic etiology, either due to presentation, age of onset, family history, etc., prompting their referral for genetic testing. These etiologies may ultimately be Mendelian, multigenic, multifactorial, or environmental, but the goal of the clinical genetic testing is to attempt to determine the genetic contribution. Dependent upon their specific phenotype, a given individual may undergo several different genetic tests, generating a more comprehensive picture of that individual's genomic variation. In addition, as phenotypic information is routinely captured by many clinical laboratories in order to properly process and interpret test results, basic phenotype information is already available for many clinical testing cases.

In this project we will leverage the hundreds of thousands of tests performed every year within clinical laboratories on affected patients with demonstrable phenotypes to build a database of clinical grade genomic variation. Initial efforts will focus on rare disorders; the majority of clinical genetic testing performed today is in this category, and this information has the highest immediate clinical utility in genomic medicine. We will expand our assessment of variation to include the entire human genome as whole exome and genome sequencing rapidly make their way into routine clinical use as primary diagnostic tests. **Unlike other database efforts, the goals of our effort are to support the submission of <u>all</u> variation observed in affected individuals during clinical testing into a publically available database, and to make this information freely and easily accessible to the community.** We intend to ensure the quality of the information through ongoing curation efforts. Although this project will support the collection of data on all human genetic variation, our initial curation efforts will focus on rare germline variation. Over time, and through partnerships with other groups, this resource will also be able to support the curation of somatic and common variation.

2. Description of the resource to be generated

The goals of this three year project are to support the submission of high-quality genome-wide variants and their associated phenotypes into the public domain, and develop methods and infrastructure to support evidence-based curation of the data. For the purposes of this project, we will define two main categories of human genomic variation: 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, 6]. The great majority of these CNVs are less than 100 kb in size, common, and not associated with overt disease [7]. 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 [8-10]. 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 [11, 12]. Currently, there are close to 100 laboratories participating in clinical proficiency testing offered by the College of American Pathologists (CAP) for chromosomal microarray analysis.

The second major category of human genomic variation is sequence-level variation, which we are defining 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 over 600 CLIA-certified laboratories are available for more

than 2,500 of these monogenic diseases (www.genetests.org). In addition, evaluation by whole exome and genome sequencing is also now clinically available.

We will build upon the experience and infrastructure we developed for the ISCA Consortium, a publically available database for structural variants. We will develop clinical data sharing processes and infrastructure for sequence-level variants similar to those developed for structural variation. This project will support the ongoing collection of a large and comprehensive number of gene-specific, whole exome and whole genome variation datasets with associated phenotypic findings. All genomic variation and phenotypic data will be submitted to and accessible through the ClinVar database within NCBI. Efficient download capabilities (FTP and API) will be available 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. We will also develop a patient registry containing enhanced phenotypic data, enabling researchers to search the database and identify potential research subjects.

Accomplishing these goals requires 1) **human resources** and 2) **software and database infrastructures**. As shown in Figure 1, the **human resources** required to create and curate this database will be organized into a Scientific Advisory Board, Executive Committee, Working Groups and Consultants. The six working groups will include: an overarching **Policies, Standards, and Sustainability Workgroup (PSSW)** that will set policies and standards, and a **Bioinformatics and Information Technology Workgroup (BITW)** that will develop and manage IT support for this project. These two groups will continuously interface with four additional workgroups that will propose and implement these policies and standards, with support and feedback from the BITW, each focusing on its own assigned area: 1) **Sequence Variants**, 2) **Structural Variants**, 3) **Phenotyping**, and 4) **Engagement, Education and Access**. Each part of this organizational structure is described below. See section entitled "Administration and Management" for more information about specific personnel.

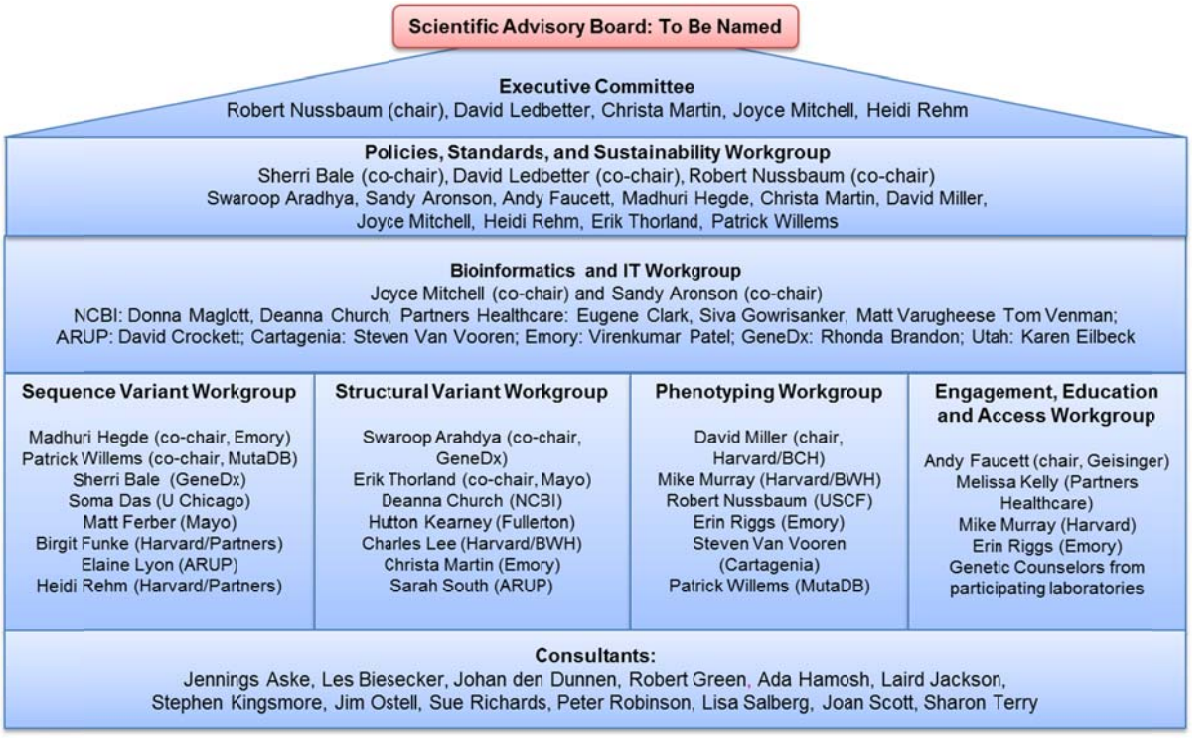


Figure 1. Organizational structure of workgroups and consultants

The **Scientific Advisory Board** will be appointed by the Principal Investigators in collaboration with the NHGRI program office. It will include representation from individuals who have led successful community resource initiatives as well as those who will use the resource and understand the broad needs of the genetics communities in both clinical and research applications. The SAB will assess the progress of the project and help establish priorities for this resource. Additional information is included in the section titled "Administration and Management".

The **Executive Committee** consists of the five project Principal Investigators [Nussbaum (chair), Ledbetter, Martin, Mitchell, and Rehm]. A detailed description of how this group will interact and their individual responsibilities can be found in the section titled “Multiple PI Leadership Plan”.

The **PSSW** is tasked with:

1. Data content, structure, and submission: The PSSW will develop **policies** surrounding data submission and defining the attributes of genotype and phenotype data that are to be collected and submitted to the database.
2. Data classification standards: The PSSW will define **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. Sustainability: The PSSW will identify strategies for the long term **sustainability** of the variant database.

The **BITW** is tasked with:

1. Data content, structure, and submission: The BITW will work closely with NCBI and the PSSW to refine the **data dictionaries** and basic content to be collected on variants and cases, as well as data submission formats. The BITW will also work with laboratories and genetic software vendors to ensure the accepted standards and data structures are included within systems used to support laboratories, and to develop **robust methods for data submission**.
2. Data curation systems: The BITW will evolve the curation tools developed for the ISCA Consortium to enable **curation of all types of genomic variation**.
3. Data security and access: The BITW will be responsible for working with NCBI, contributing laboratories, and genetic software vendors to ensure the security of the data with respect to access and **adherence to HIPAA standards** as well as other regulations imposed by IRBs and local, state and federal guidelines.

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

1. The **Sequence Variant Workgroup** will organize the submission and curation of sequence-level variants, including overseeing initial demonstration projects that will develop and test data submission and curation systems for sequence variation, as well as oversee the expansion of the database to include whole exome and whole genome sequencing data.
2. The **Structural Variant Workgroup** will organize the continued data submission and curation of genome-wide structural variants, as well as the eventual expansion to curation of structural variation detected through whole exome and whole genome sequencing.
3. The **Phenotyping Workgroup** will define and implement approaches for the standardization and enhanced collection of phenotypic data for annotating genetic variants.
4. The **Engagement, Education, and Access Workgroup (EEAW)** will reach out to laboratories, clinicians, researchers, and patient advocacy groups in order to encourage submission of genotype and phenotype data, explain the various consent options, assist laboratories with IRB approval when required, and educate the research community about the resource. This group will also oversee the development of a patient-centered registry, designed to serve as both a repository for detailed, patient-reported phenotype information, as well as a resource for researchers seeking to identify potential subjects.

Figure 2 shows an overview of the specific aims and activities described in this proposal. In **Aim 1**, the PSSW, guided by the Variant and Phenotyping Workgroups, will develop standardized data definitions and formats for genotype and phenotype data submission from clinical laboratories. The EEAW will engage and educate the laboratory, clinician, and patient communities to actively participate in data submission to ultimately benefit the quality of clinical care by standardizing the interpretation of genomic variation. **Aim 2** will focus on the increasingly automated submission of variant and phenotypic data from clinical laboratories into ClinVar. We will initiate data submission with the laboratories funded by this project (ARUP, Chicago, Emory, GeneDx, Harvard/Partners and Mayo) and expand to other contributing laboratories once efficient processes are tested and in place. Further, to facilitate the recruitment of patients into research studies and gain additional methods of phenotypic data collection, we will create an online patient registry. In **Aim 3,** we will develop sustainable methods for expert clinical level curation of genome-wide variation data, utilizing lessons learned from demonstration projects in several different

disease areas. An access and dissemination plan carried out by the EEAW will facilitate open access of the resource to the clinical and research communities.

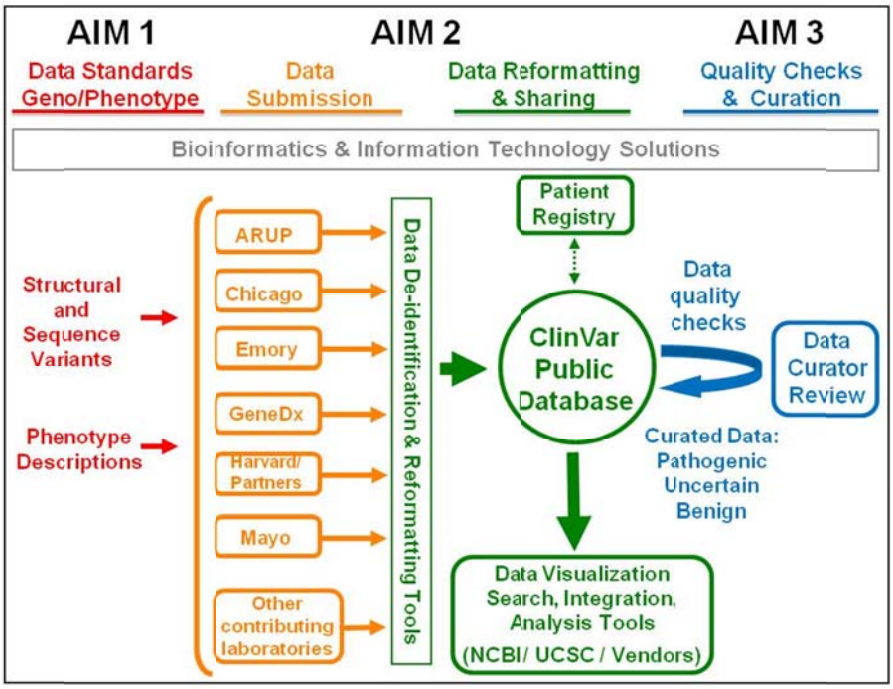


Figure 2. Process to develop the Unified Human Genomic Variant Database

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

The core public resource that we will build upon is ClinVar, a new database at NCBI (www.ncbi.nlm.nih.gov/clinvar). ClinVar was developed at NCBI to improve the management of medically important human variation in a centralized public environment. Although the official launch of ClinVar will occur in July 2012, data integration has already begun. In addition to the data originally submitted to NCBI's database of genotypes and phenotypes (dbGaP) by the ISCA Consortium, all variants from OMIM and GeneTests/GeneReviews, resources that are already supported by the NCBI infrastructure, have been integrated into ClinVar. Data from 83 locus specific databases and three clinical laboratories have also been submitted to ClinVar. However, to become a more robust database, hundreds of thousands of variants still need to be collected. In this project, we propose collecting this data from hundreds of clinical laboratories and over 1,000 locus-specific databases, emphasizing a critical need for the support to facilitate these submissions.

The data needed to populate this database will be acquired from genome-wide copy number analysis as well as both gene-specific and genomic sequence analyses performed during routine clinical assessment of individuals affected with a variety of phenotypes. Data will be acquired from testing performed using multiple technology platforms, including chromosomal microarrays (comprised of oligonucleotide and/or SNP probes), targeted Sanger or next generation sequencing, whole exome and whole genome sequencing. Several methods for data submission are supported and more will be built.

Due to the potential identifiability of certain types of data (such as raw cytogenomic microarray files or full exome and genome sequencing data sets), all data will be initially processed through dbGaP, a controlled access system. Sensitive datasets in their original forms such as those mentioned previously will remain under controlled access in dbGaP, but variant calls and clinical assertions that cannot be associated with an individual will be extracted from these datasets and transferred to ClinVar for public use (Aim 2).

Data will be made publically available through several resources supported within NCBI, including ClinVar, the database of genomic structural variation (dbVar), and dbSNP. This publically available data can then be used by interested parties to generate additional tools. For example, the ISCA Consortium database has been integrated into a number of data analysis software programs routinely used in clinical laboratories such as Affymetrix, Agilent, BioDiscovery, BlueGnome, Cartagenia, and Oxford Gene Technology. In addition, this data has been incorporated into publicly available tracks within the University of California Santa Cruz (UCSC) genome browser (http://genome.ucsc.edu/) and dbVar for easy visualization of the data. Researchers will also be encouraged to apply for access to the raw data files housed within dbGaP to facilitate genetic discoveries.

4. Preliminary data supporting the approach

We will leverage our success with the ISCA Consortium and use the existing infrastructure to jumpstart the construction of the expanded resource that will be developed in this project. 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. Drs. Ledbetter and Martin are founding members of the ISCA Consortium. In addition, other active ISCA leaders are also continuing to

contribute to this project, including Aradhya, Church, Faucett, Kearney, Miller, South, Thorland, and Van Vooren. The development of this consortium was initially supported by a grant awarded to Dr. Ledbetter from the American College of Medical Genetics (ACMG) 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. Drs. Ledbetter and Martin were subsequently awarded an NIH American Recovery and Reinvestment Act Grand Opportunities grant (CNV Atlas of Human Development, 5RC2 HD064525) to further support these efforts and the creation of the ISCA database to make the data publicly available through an interface with NCBI.

In the Research Approach section, we will detail projects already in progress as part of the ISCA Consortium effort to demonstrate how these activities relate to the aims of the proposed project. As a brief overview, we will highlight here some of the major accomplishments to date. The ISCA database, with elements housed within both NCBI's dbGaP and dbVar, now includes data from almost 35,000 cases analyzed with genome-wide cytogenomic microarrays and submitted by clinical laboratories. Standardized genotype and phenotype formats were developed for data submission and incorporated into array software analysis programs to facilitate data submission. The fourth version of this dataset has recently been made publically available (http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000205.v4.p2). The CNV calls from this dataset have also been released through NCBI's dbVar for use in genome-browsers, such as those available online (e.g., dbVar, ISCA, UCSC) and those provided within vendors' array analysis software (e.g., Affymetrix, Agilent, BlueGnome, Oxford Gene Technology). The ISCA dataset has already been used to produce one of the largest case-control CNV studies to date that focused on defining the pathogenicity of common CNVs across the genome [15].

Other major accomplishments of the ISCA Consortium include the standardization of a clinical array design, which is now available through multiple commercial vendors. The use of a standardized array format provides healthcare providers with the assurance that there is equal coverage related to minimum content and resolution across laboratories; many aspects of this design, including a minimum backbone resolution of 400 kilobases, were incorporated into the ACMG Practice Guidelines [11, 12, 16]. The Consortium also worked to standardize how CNVs are visually represented in genome browsers. Historically, gains and losses were represented by different color codes, which were chosen independently by the genome browser. During an ISCA Consortium meeting in January 2011, representatives from the major genome browsers (dbVar, ISCA, DECIPHER, DGV, and UCSC) established a standard for color coding: red for losses and blue for gains. Although this accomplishment may seem trivial, trying to analyze data from various sources that did not standardize the representation of gains and losses was very confusing for users and led to data misinterpretation; the unified color-coding consensus resolved these issues. The ISCA Consortium has also made significant advances in the curation of the ISCA database, including published guidelines for the evidence-based review of genes [17]. Finally, this group has engaged the clinical and research communities through educational materials such as instructional webinars defining the need for genotype and phenotype information and how to submit such data, all currently available on the ISCA website (www.iscaconsortium.org).

With the foundation of this database established for structural variation, we are poised to expand to the acquisition of sequence-level variation to develop a unified clinical genomic variant database. We recognize that with this expansion comes more challenges, but we are well positioned to accommodate the continued expansion of data. Recently, we merged the efforts of the structural and sequencing communities in the Second Annual ISCA consortium Conference held on May 21-22 in Bethesda, MD. Over 200 attendees, including clinical laboratory directors, physicians, genetic counselors and researchers, participated and discussed the current efforts of the ISCA consortium, as well as the merger of the ISCA consortium with the molecular/sequencing community. This meeting was one of the first efforts of the PIs of the proposed project to unite the structural and sequence variation communities to develop a unified database. Topics included data submission and curation, phenotyping, and community engagement and attribution. The majority of the presentations are available on the ISCA website at: https://www.iscaconsortium.org/index.php/isca-consortium-conferences-and-workshops/166.

As part of developing data dictionaries and the process of depositing sequence variants in ClinVar, three clinical laboratories have already worked with NCBI to submit subsets of their variant data. This process was successful and resulted in several thousand variants being submitted, which will be viewable in ClinVar once launched. However, significant funding will be required to support extending this effort to a larger community of laboratories.

5. Community support for proposed resource

As outlined in the strategic plan of the National Human Genome Research Institute (NHGRI) [13], 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 unified and curated genomic variation database. The success of the ISCA Consortium 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. The ISCA Consortium has quickly grown to a membership of more than 1,700 individuals, including laboratorians, clinicians, genetic counselors, and other scientists. The ISCA database of structural variation is accessed daily by members for clinical interpretation of structural variation, and identification of cases suitable for research initiatives.

In the same way, there is an overwhelming community need for the availability of sequence-level variant data through a unified database. As whole genome sequence analysis is increasingly incorporated into clinical research and clinical care, a unified variant database will be critical for clinicians to provide accurate diagnoses and prognoses. Though NCBI has developed the ClinVar infrastructure in response to this need, assisting NCBI with the population of the ClinVar database through a dedicated and funded effort will be critical to its success. Clinical laboratories, the primary generators of variation data from affected individuals, have historically not submitted data into databases without assistance, support, or incentives for the process. This observation is supported by a recent survey (https://www.surveymonkey.com/s/PHBMWWR) we performed of the approximately 100 US clinical sequencing laboratories registered with GeneTests, 32 of which responded. Of the respondents, 90% reported either never submitting data to the public domain, or only doing so through publications. Lack of time, resources, and an uncertainty regarding the actual submission process were cited as major barriers. However, when asked if they would be willing to submit data to a public database in the future, **100%** indicated that they would. By addressing time and resource barriers through the creation of automated, work-flow integrated tools and raising awareness through the efforts of the EEAW, this project will undoubtedly result in substantially increased data submission to ClinVar. As described below, and in the Approach section, we have already garnered broad support for data submission from 48 laboratories (see Appendix), as documented in the included letters of support and results of our recent laboratory survey. To date, only three laboratories have declined participation, though two of these laboratories have indicated the potential to join the effort in the future when the process is better established.

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: laboratories will use this database for the clinical interpretation of variants; clinicians can utilize the provided phenotype information associated with their patient's particular variant(s) when deciding a course of clinical action; and researchers can utilize both the raw patient data files as well as the variant spectrum for studying gene function and the potential impact of variants of unknown significance. 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, and many others).

6. 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; even more will be sequenced through other funding sources (http://www.genome.gov/27545796). 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. The task of curating these genomes is extremely challenging due to the limited availability of informative and accurate existing resources and standards for their interpretation. 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. This effort is anticipated to have a wide impact throughout the genetics community, including (but not limited to) professional societies, government agencies, patient advocacy groups, and other genotype/phenotype databases (see Table 1).

Impact on clinical laboratories: **CLIA-certified** laboratories have begun to offer whole exome and whole genome sequencing services on a clinical basis. However, the lack of a publically available, centralized, curated set of data correlating genotype and phenotype makes the use of these services in a clinical context extremely difficult. 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 unified database with expert curation is an invaluable resource that will enable clinical laboratories to write more informative patient reports

leading to improved quality of patient care. Such a resource is expected to result in a demonstrable impact on quality assurance for clinical laboratories, a potential benefit recognized by groups such as CAP and ACMG (see letters of support).

Impact on the basic science community: We anticipate that this database will house data from hundreds of thousands of individuals with various phenotypes. This will be invaluable in differentiating 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 that aid in the identification of novel links between genotypes and disease phenotype, allow more detailed characterization of known disease genes, and facilitate discovery of 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. We will foster open communication with existing databases to ensure that all efforts are collaborative and additive as opposed to duplicative (see letters of support from the Human Variome Project, OMIM, MutaDATABASE, LOVD, GeneReviews, and PharmGKB).

Table 1: Liaisons with other organizations. The listed key contacts represent people with whom we have successfully worked with to date and/or have commitments to work together for this project.

	Organization	Key Contact	U41 Liaison
Organizations, projects, and consortia with overlapping missions	NCBI	Jim Ostell	Joyce Mitchell
	Human Variome Project	Richard Cotton	Heidi Rehm
	OMIM	Ada Hamosh	Ada Hamosh
	HGVS/LOVD	Johan den Dunnen	Johan den Dunnen
	DECIPHER	Matthew Hurles	Christa Martin
	1000 Genomes	David Altshuler	Heidi Rehm
	MutaDATABASE	Patrick Willems	Patrick Willems
	GeneReviews	Roberta Pagon	Heidi Rehm
	PharmGKB	Russ Altman Terry Klein	Robert Nussbaum
	PGRN	Dan Roden	Robert Nussbaum
	NGHRI CSER	Brad Ozenberger	Robert Green
	NGHRI eMERGE	Rex Chisholm	David Ledbetter
	Clinically Relevant Variants Resource	NGHRI	Many
Professional societies	ACMG	Wayne Grody	Many
	CAP	Stanley Robboy	Sue Richards
	AMP	Mary Williams	Elaine Lyon
	ASHG	Mary-Claire King	Robert Nussbaum
Government agencies	CDC	Barbara Zehnbauer	Andy Faucett
	FDA	Elizabeth Mansfield	Andy Faucett
Broad focus patient advocacy groups	Genetic Alliance	Sharon Terry	Sharon Terry
	Global Rare Diseases Registry	Kyle Brown	Andy Faucett
	UNIQUE	Beverly Searle	Andy Faucett

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 and more accurate inclusion criteria for clinical trials. Additionally, partnering with existing patient registries and developing a global patient registry for those without existing registries will vastly improve the recruitment of patients and families into clinical research studies and clinical treatment trials that utilize genomic information (see letters of support from UNIQUE, Patient Crossroads, the Noonan Syndrome Support Group, and CureCMD).

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 [5]. 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. Professional societies and regulatory agencies will have access to the data needed to begin to develop guidelines for the use of genetic information on clinical care (see letters of support from ACMG, ASHG, and CAP). 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. Furthermore, the ClinVar database can form the basis for domain experts to develop standards surrounding the more common variants or classes of variants that demonstrate sufficient clinical utility for routine application in clinical care; we will collaborate closely with NHGRI's Clinically Relevant Variants Resource to ensure that data generated from this effort is utilized to this end (http://www.genome.gov/Pages/About/NACHGR/May2012AgendaDocuments /Clinically_Relevant_Variants_Resource_Revised_05%2015%2012.pdf). 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 [14].

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 genotype and phenotype data into the database (Aim 2), and support evidence-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 and allow for the discovery of novel links between genetic mutations and human disease.

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 and cataloging the clinical significance of human variation.

Though each of the different workgroups will be involved in the development of the standards and policies needed to complete this project, the Policies, Standards, and Sustainability Workgroup (PSSW) will oversee all efforts. The PSSW will be chaired by Drs. David Ledbetter, Bob Nussbaum, and Sherri Bale. Membership on the PSSW will include the Principal Investigators and chairs of each workgroup. Other individuals, who are serving on the other workgroups, as well as our consultants, will be included as needed. Figure 1 shows the complete organization of key personnel, workgroups, and consultants.

Defining nomenclature and standards for genotype and phenotype

A critical mission of the PSSW is 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, and to ensure integration of this standard into broad use. 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.

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 further define the criteria for structural variation and develop these criteria for sequence-level variation, Karen Eilbeck, PhD, will assist the group from her focus on the ontology of sequence variations (the Sequence Ontology, www.sequenceontology.org, already utilized in part by ClinVar) and standardized file formats [20, 21]. 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 that will be included in our considerations. These include:

- The ACMG standards and guidelines for interpretation of sequence variations [22],
- The recommendations proposed for classifying cancer variations [23],
- 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 [16].

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 process 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). Our group has come to consensus for sequence-level variant classification based upon existing guidelines and the categories that most laboratories are using today. At the start of funding, we will solicit additional feedback on this consensus through a survey to all US clinical

laboratories performing DNA sequencing using contact information from the NIH Genetic Testing Registry. Although the community has not yet come to consensus on a universal standard, several of the investigators on this grant (Drs. Aradhya, Bale, Das, Ferber, Hegde, Kearney, Lyon, Martin, Rehm, South, and Thorland) are involved in professional efforts to define these parameters. This will ensure their incorporation into the standards for the ClinVar database.

In addition to the American College of Medical Genetics (ACMG), we have also begun engaging with other communities for wider consensus on standards, including the College of American Pathologists (CAP), 1000 Genomes project [2], and the Human Variome Project (HVP) [24] (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 to update their standards and guideline document for sequence variants (see letter of support from Sue Richards, Chair, ACMG Lab QA Committee and Dr. Rehm is a member), as well as with regulatory agencies, such as CAP, who have agreed to enforce such standards through the clinical laboratory accreditation process. The standards will also be posted on the ISCA Consortium and ClinVar websites as well as sent to other collaborating projects for posting on their websites (MutaDATABASE project, Leiden Open Variation Database, Human Gene Variation Society, Human Variome Project, etc.).

In coordination with the Phenotyping Workgroup, the PSSW 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 already built into ClinVar comes preferentially from two sources: the Human Phenotype Ontology (HPO) [25] (http://www.human-phenotype-ontology.org/index.php/hpo_home.html), as established by Dr. Peter Robinson, a consultant to this project [25], and SNOMED CT (Systematized Nomenclature of Medicine--Clinical Terms; http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html). The HPO was initially developed using information from OMIM, which is relevant to the vast majority of disorders for which clinical testing 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 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. Though these two systems are already supported within ClinVar, entries using others, such as UMLS, MeSH, ICD-9 and ICD-10, will also be acceptable.

Using standardized vocabularies to build databases, such as the one proposed, allows for more efficient computational analyses of the data. As proof-of-concept for this approach, HPO terms are already used by the ISCA Consortium for data entry into dbVar and dbGaP. Electronic forms utilizing HPO terms organized by body system and displayed in a simple, check-box format have been developed for capturing essential phenotype information for both prenatal and postnatal assays (see Appendix). The ISCA Consortium has worked with Cartagenia (www.cartagenia.com), a web-based software and database platform, to develop a data submission tool incorporating these forms. The forms are also available for general use through the ISCA Consortium website (www.iscaconsortium.org). We have demonstrated that using these forms result in double the amount of usable phenotype information collected per case as compared to free text annotations included on test requisition forms [26]. 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 that allows frequently used ICD9 codes to automatically be mapped to HPO classification codes. The use of this algorithm has allowed us to mine over 13,000 HPO terms from over 6,000 different cases initially submitted with only free-text phenotype information [26].

This functionality has improved the quality of data collected, as ICD-9 codes were frequently the only information available. In addition, 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 [26]. We will continue to focus on expanding the utility of these applications in the phenotype data collection process. Additionally, we will develop disease-specific phenotype forms on an as-needed basis (utilizing standardized vocabulary) in order to facilitate the detailed phenotype information necessary for certain single gene or gene-panel tests. For example, a phenotype form that might accompany a sample for a hearing loss panel might include specific information regarding audiograms, temporal bone abnormalities, etc., information that would not typically be included on a more generalized form.

More recently, a clinician-friendly system for documenting patient phenotypes, called PhenoDB, has also been developed by Ada Hamosh's group for the Baylor-Hopkins Mendelian Sequencing Center (see letter of support from Dr. Hamosh). We will also evaluate this system in determining what approaches are most effective for gathering detailed, structured and standardized phenotypes on patients and how we can support their use in easy and efficient ways by physicians, laboratories and patients. The most efficient methods for collecting phenotypic data using a standardized vocabulary will be integrated into the online patient registry system, as discussed in Aim 2B below.

Peter Robinson, phenotype ontology expert and consultant to this project, has agreed to assist Dr. David Miller, Chair of the Phenotyping Workgroup, in 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." In addition, standards that have enabled both the Human Genome Project and the Electronic Medical Records (EMR) to be interoperable will be used whenever possible. From the genome perspective, these include the HGNC (assigns unique gene symbols to protein coding genes), RefSeq LRG (a curated database of transcripts and protein sequences) and HGVS (a nomenclature for description of sequence variants). From the EMR perspective, these include HL7 (a messaging standard to send and receive data between EMR components), LOINC (a universal terminology for identifying laboratory and clinical observations), RxNorm (a standardized vocabulary for drug names), and SNOMED CT (a standardized vocabulary of disease names and some symptoms). One ultimate goal of this project is to build the foundation for the future to allow clinical decision support within the EMRs for both providers and patients; ensuring that our database adheres to these standards from the beginning will facilitate this future goal.

Defining the data to be collected

The PSSW will also define which information is collected and will work closely with the BITW to ensure all systems and tools are developed to support the chosen content. 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 a broad spectrum of clinical phenotypes, including (but not limited to) neurodevelopmental disabilities and/or congenital anomalies. A standard format for genotype data, along with standardized variables, has been established by the ISCA Consortium (see Appendix). All CNVs from an individual are captured and submitted to NCBI, with the number of CNVs ranging from 0-20+ variants per patient. Each CNV is submitted with the initial clinical assertion of the submitting laboratory; this assertion must fall into one of the 5 ACMG categories (pathogenic, uncertain-likely pathogenic, uncertain, uncertain-likely benign, or benign). All clinical assertions are subject to further evaluation by an expert consensus process, and could be reclassified following discussions between the curation committee and the submitting laboratory [16]. Phenotype information, using a standardized HPO terminology, is also being incorporated with the genotype information. The most recent dataset from the ISCA Consortium (nstd37) is currently being transferred into ClinVar.

For sequence variation, a group of approximately two dozen stakeholders including members of our group and others representing 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 data element dictionary has been developed (see Appendix), and three clinical laboratories (ARUP, Correlagen, and Partners Healthcare) have begun submitting data as a pilot effort. However, the data dictionary will be refined through real-time use and continued incorporation of standardized terminologies as they evolve, 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 has been determined. Additionally, a complete set of data elements will be defined to enable the richest dataset possible. This dataset includes 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). 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. Variants identified in patients with clinical disease, but which are interpreted by the submitting laboratory to be benign, will also be collected. 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.

Although the initial stages of the project will focus on genes involved in highly penetrant phenotypes, the infrastructure will support the collection of all variation, including evidence-based data on common variants for complex traits, pharmacogenetic variants and somatic variation. For example, we have already begun discussions with PharmGKB on the integration of their pharmacogenetics dataset into ClinVar (see letter of support from Russ Altman and Teri Klein) and we anticipate working with other domain-specific efforts to centralize curated datasets and create a foundation of data to support the development of clinical guidelines for integration of variants into medical decision-making.

Phenotype data, in contrast to the core elements of variant data, are very heterogeneous and vary significantly among genes/disorders. 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. For more details, see Aim 2.

1. Overall diagnosis or testing indication provided by ordering clinician
2. General data collection forms, provided at the time the test is ordered or before reporting test results (see ISCA prenatal and postnatal forms in Appendix)
3. Disease-specific data collection forms, 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. Electronic use of PhenoDB
5. Retrospective entry or linkage of phenotype data sets to genetic records through consented patients (supported by patient registries)

Development of educational materials to stimulate data acquisition

In an effort to educate clinicians and submitting laboratories about the importance of participation and the process of data submission, the ISCA Consortium has developed educational materials, such as one-page handouts and pre-recorded webinar content, focusing on the key interests of the different groups. In conjunction with the webinars, 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. Similar strategies will be employed within this project to stimulate the acquisition of both structural and sequence-level variation data.

Patient protection policies for data collection

The PSSW will also address the 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 through our work for the ISCA Consortium. The policies will be adhered to 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. independent calls from cytogenomic microarrays or compound heterozygous variants in a single gene) 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. Though all data is initially processed through dbGaP before de-identified variant information is released publicly through ClinVar, potentially identifiable files will remain within dbGaP under controlled access. For submission of such datasets, the laboratory will need to have an IRB protocol in place to manage at least an opt-out process, 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 variant data, we will focus on patients who have undergone genetic testing through clinical laboratories. Initially, most submitted data will be derived from single gene mutation analyses, gene panel testing, or cytogenomic microarray testing. Data consisting of only variant information (including clinical interpretation) and basic phenotype information (age at time of testing, gender, phenotype provided at the time of testing utilizing standardized vocabularies) will be considered non-identifiable, and will therefore not require full consent or submission through an opt-out process. For data that could be considered identifiable, other approaches, as described below, will be supported.

Opt-Out Process: Currently, the ISCA Consortium successfully uses an opt-out process for the submission of potentially identifiable 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 [18]. It was reviewed and approved by the NIH Office of Human Subjects Research (OHSR) and has currently been approved by several academic IRBs for use in the ISCA Consortium. The opt-out process 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 process 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. However, while providing enormous protections to patients, multiple layers can also inhibit research and limit patient benefit. As a result, we will also develop a patient registry as described in Aim 2b that will allow a more robust and controlled communication process between researchers and patients.

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) [19]. 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.

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 a consent process to allow for the collection of additional phenotype information and expanded use of these full datasets. We have obtained agreement from several laboratories offering or planning to offer whole exome or genome sequencing services, including Emory, Geisinger, GeneDx and Partners Healthcare to support a full consenting process for data deposition to dbGaP as well as facilitate the registration of patients with undiagnosed disorders into the patient registry described in Aim 2b. We will approach additional laboratories if funded.

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

The ultimate goal of this grant is to populate NCBI's ClinVar database with clinical grade genomic data through the deposition, curation, and maintenance of all human genomic variation. In his letter of support, Dr. James Ostell, Chief of the NCBI Information Engineering Branch, 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". Because NCBI does not have the expertise to curate the data in ClinVar, our relationship will be mutually beneficial. Towards this collaborative effort, NCBI has agreed to have members of their team (Donna Maglott and Deanna Church) participate in the working groups,

particularly the BITW. 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 and basic phenotype information 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. All data relevant to a variant, with or without a clinical assertion, and whether curated or uncurated, will be searchable and viewable within ClinVar. A mock view of the screen that will be launched for ClinVar in July 2012 is shown in Figure 3.

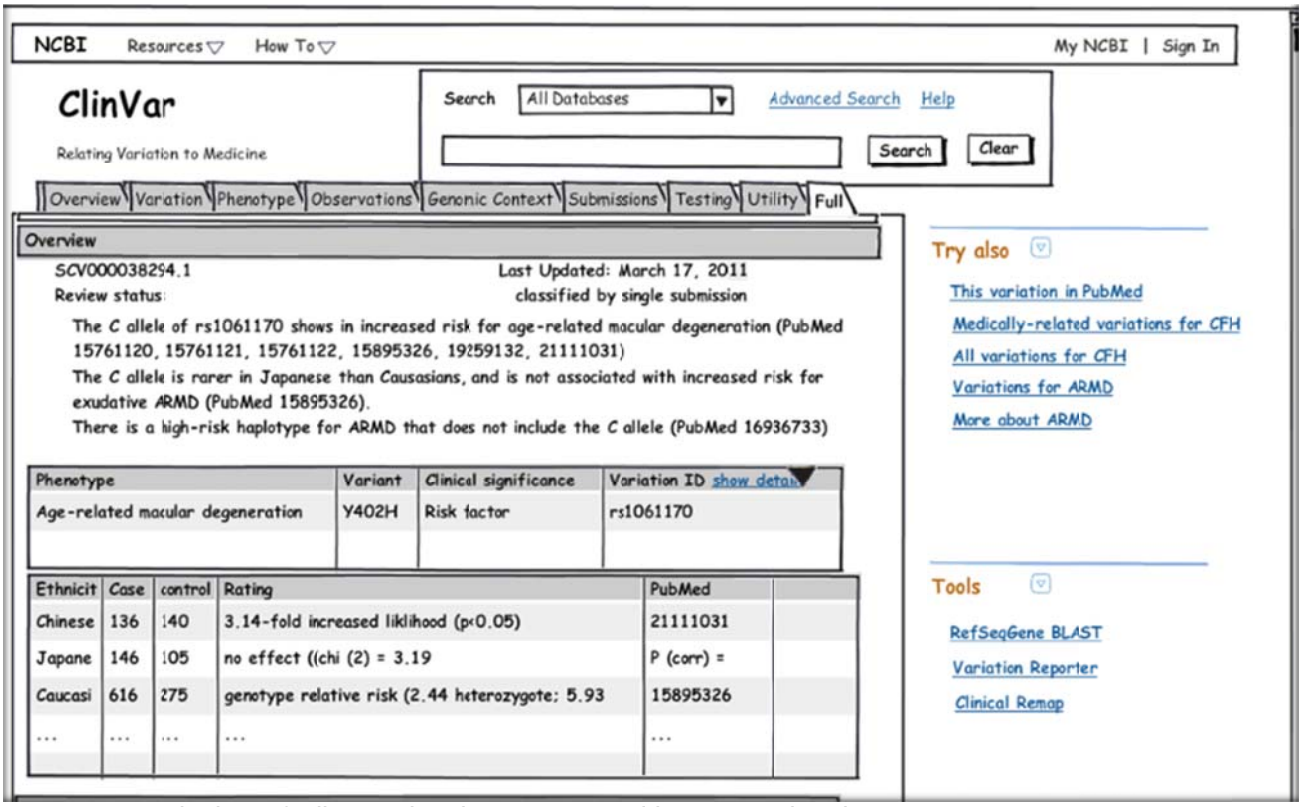


Figure 3. Mock view of ClinVar showing a screen with a curated variant.

2a. Submission of variant data into ClinVar

Data submission methods

Aim 2 will support the submission of numerous sources of genotypic and phenotypic data into a centralized location. Genotypic variations will be reported to the user as sequence changes relative to an mRNA, genomic, and protein reference sequence. Genomic sequences will be represented in chromosome coordinates, as well as in RefSeqGene/Locus Reference Genomic (LRG) coordinates. ClinVar will also track disease associations through the use of standardized medical terminologies as described in Aim 1. Several data flows will be supported depending on the type of variants and the level of identifiability of the data. Figure 4 shows a diagram of the overall data flow. The major systems that will be used for data submission and phenotype enrichment are described in the paragraphs following, including details on functionality and quality control checks.

For structural variants, the infrastructure for data submission and storage has been established through dbGaP and dbVar at NCBI. The ISCA Consortium worked with the software company Cartagenia to provide a more integrated source for laboratories to use to facilitate the de-identification and transfer of data to the NCBI database. Cartagenia utilizes a 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.

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.

In addition, multiple other array software vendors (such as Affymetrix, Agilent, BioDiscovery, BlueGnome, Oxford Gene Technology, and Roche Nimblegen) 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 at NCBI.

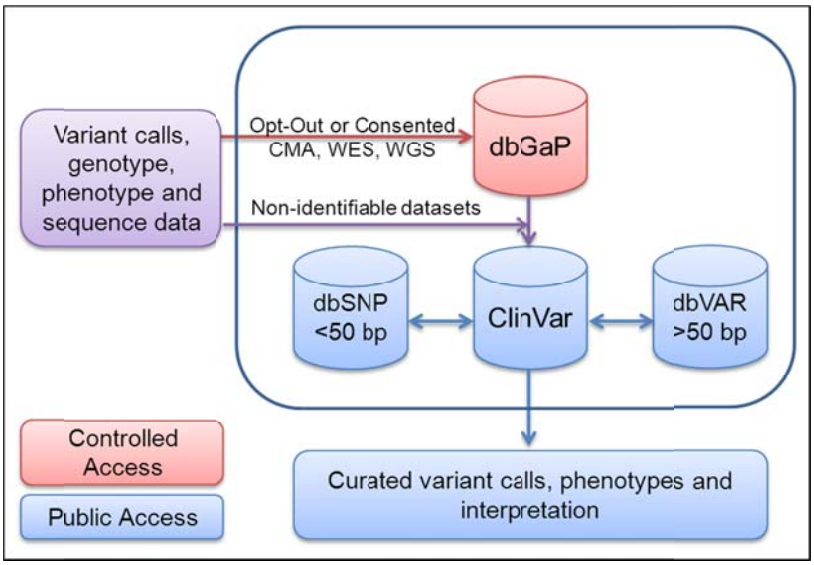


Figure 4. Overview of infrastructure and data flows.

For cases collected through an opt-out process or with consent, 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, and dbGaP will continue to house the raw data files under controlled access. In general, previously undescribed variants will be accessioned through dbVar (typically variants greater than 50 base pairs in size), or dbSNP (typically variants less than 50 base pairs in size).

For sequence-level variants, data will be submitted directly to NCBI for import into ClinVar. There are three major types of submissions currently being processed by ClinVar and supported by NCBI staff: 1) *Automated, with identification of conflicts and generation of reports for curatorial review.* On a daily basis, ClinVar imports data from dbSNP, NCBI's Gene, GeneTests, Genetic Testing Registry (GTR), and OMIM to maintain information on disease names, gene names, gene/disease relationships, variations subject to explicit testing (as defined in the GTR), and allelic variants. If for any reason an insert or update cannot be managed automatically, a report is generated for curatorial review. 2) *Semi-automated, with resource-specific processing*. Groups can provide regular submissions as e-mailed attachments in a standard format. These files are converted via source-specific scripts to our common data format, and processed automatically thereafter. 3) *Unique cases.* Submitters can contact ClinVar staff, describe what they want to submit, and then provide a submission. These unique processing events have resulted in refinements of the data model and enhancements to the loading tools for ClinVar.

The BITW will work with submitting laboratories, supporting vendors, and NCBI to define and facilitate data submission processes. Data collection will consist of a combination of gathering and submitting retrospective data as well as setting up systems for efficient ongoing future data submission. A senior IT developer and a genetics specialist (or an IT analyst with significant genetics experience, if available) will travel to the larger labs and conduct conference calls with smaller laboratories to catalog each institution's capabilities. Based on this information, the BITW will define submission protocols designed to be practical for all laboratories while at the same time yielding consistent datasets. The BITW will monitor the implementation of these protocols and coordinate the cross-site resolution of issues as they arise. Submission processes will account for retrospective data extraction and submission. Term mapping and format conversion protocols will be leveraged to account for differences between laboratories. The BITW will constantly work with all groups to iteratively improve submission protocols to enable increasingly automated submission of deeper datasets.

A bidirectional interface will also be built between the Leiden Open Variation Database (LOVD) system, which currently supports approximately 80% of existing locus-specific databases, and ClinVar to enable efficient bidirectional data transfer between both environments. Similar interfaces or regular data exchange policies will also be set up with other locus specific databases not using the LOVD infrastructure. We will engage the curators of those databases to consider migrating their primary site of data into ClinVar and using tools being developed in ClinVar for their curation activities. All systems used for data submission will enable the creation of unique IDs for case submission, with links to submitter IDs that are maintained by the submitting source. This functionality has already been built within Cartagenia and will be extended to other tools supported by the project. This system

enables re-contact under the opt-out model described in Aim 1 and ensures that data are not sent in duplicate from individual laboratories.

Sources of variant data

Variant data will be submitted to ClinVar from a variety of sources, including clinical laboratories, research laboratories, locus-specific databases, OMIM, GeneReviews, UniProtKB, PharmGKB and MutaDATABASE. 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 40,000 more cytogenomic array cases from the five laboratories supported by this grant performing cytogenomic microarray testing (ARUP, Emory University, GeneDx, Mayo, and Partners Healthcare) during the 3 year funding period based upon the rate of accrual to date. We are also actively working with Baylor and Signature Genomics, two other large volume cytogenomic microarray testing laboratories, to support the submission of their data into the repository. The sequence-level project will begin with data submitted by the six clinical laboratories supported by this grant and listed in Figure 2 (all of the laboratories listed above, with the addition of University of Chicago) and then expand to additional laboratories. After querying our six core clinical laboratories, we have estimates of over 250,000 sequencing cases, including over 73,000 curated variants in over 700 genes from these labs alone. To date, 48 laboratories have agreed to submit variant data. Adding in these other committed laboratories will provide an unprecedented amount of data to be submitted.

The bidirectional interface between the LOVD system and ClinVar will enable thousands of curated variants to be transferred into ClinVar. We will also support exchanges with non-LOVD locus specific databases (LSDBs). For example, we have agreed to exchange data with the InSiGHT database for hereditary colorectal cancer – see letter of support from Finlay Macrae). Over time, and as tools are built to support data curation within ClinVar, we anticipate curation groups directly gathering and curating data within ClinVar. However, initially interfaces and routine data exchanges may be necessary to capture certain locus specific data already housed within LSDBs. For example, bidirectional data exchanges may be necessary for databases where identifiable clinical information is being maintained with genetic data (e.g. Parent Project Muscular Dystrophy), in an effort to ensure that only appropriate, consented information is transferred. All models of data sharing or transfer will be supported as long as they increase the community's access to data.

This project is also collaborating with the MutaDATABASE project (www.mutadatabase.org), which is supporting clinical laboratory data sharing and centralization of variant data. Approximately 15,000 variants have already been submitted into MutaDATABASE. As such, we will work with Dr. Willems and his team to ensure that data collected within the MutaDATABASE project are also deposited into ClinVar. We have allocated IT resources in Year 2 to support an interface between the two systems, if MutaDATABASE develops as a preferred site of data submission for certain laboratories, particularly those abroad. Likewise, early discussions have also begun with other international efforts to ensure efficient data exchange and collaboration towards common goals (see letters of support from Richard Cotton and Johan den Dunnen).

ClinVar quality control and versioning

The data elements of ClinVar are stored using reliable, standard technologies, including a relational database in order to ensure transactional integrity, backup, transaction history and rollback, and custodianship. Although the technology base will be adjusted to meet future requirements as they evolve, relational technology is sufficient to support the needs of ClinVar today. Structural and sequence variants submitted to ClinVar are mapped to reference sequences and annotated according to Human Genome Variation Society (HGVS) standards. ClinVar assigns a version number to all data submissions, allowing submitters to update their records and retaining the previous version 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.

Drawing from our experience submitting data to NCBI through the ISCA Consortium, several quality control checks have been built into the system: a check that the variant is defined within valid regions of genomic sequence; a check that the variant description conforms to HGVS nomenclature; identification of conflicts among submitters and the published literature or prior submitted data; and 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. Both of these quality control checks of the variant classification system have been in place for several years in support of the ISCA project and have been highly successful in improving the quality of structural variant interpretations.

User support for software applications

NCBI and the BITW, which includes staff from NCBI, will be responsible for providing user support for ClinVar and other tools and applications used to support the activities of the grant. NCBI 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).

In addition to user support provided by NCBI, we will also provide general support for all applications and other activities developed through this grant (contact links will be provided via the applications and a unified project website that will evolve from the ISCA website once the sequencing and structural variant efforts are fully merged). User support for additional applications created specifically for the grant (data submission, curation, etc.), not including ClinVar, will be provided by the BITW. Several members of the BITW have extensive experience in providing user support for genetic software applications. We will also work closely with genetic software vendors that are supporting labs to ensure proper support is being provided for data submission and we will supplement the support as needed to ensure capture of data. One software vendor, Cartagenia, funded in part by this grant, has been working directly with individual laboratories to support the submission of structural variant calls and full raw datasets and will continue to provide this support as well as expand to supporting sequence variant submission. 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. In addition to the six vendors supporting structural variant data as listed above, three other groups are actively working to improve their system to aid in the data submission process. On the sequencing variant side, several members of the BITW also support the GeneInsight software (http://pcpgm.partners.org/it-solutions/geneinsight) which today provides variant database infrastructure to support two of the 6 clinical laboratories funded by this grant (ARUP and Partners Healthcare) enabling highly efficient data sharing and data submission.

2b. Submission of Phenotypic Information into ClinVar

Sources of phenotypic data

There will be three main sources of phenotypic data: laboratories, clinicians, and patients. For phenotypic data submitted from any source, the standard format will consist of general variables such as age at time of testing, gender, race/ethnicity, type of testing (symptomatic vs. asymptomatic), and identification of the origin of clinical data. More disease-specific and/or gene-specific data will follow the format of SNOMED-CT or HPO terms. Data collection formats, such as general clinical data collection forms and disease-specific data collection forms, are described in Aim 1. Examples of such forms can be found in the Appendix. We will also integrate clinician-driven electronic data capture methods such as the use of PhenoDB (developed by Ada Hamosh's group for the Baylor-Hopkins Mendelian Sequencing Center). Finally, we will work to develop an online patient registry of individuals who have undergone clinical genetic testing. The creation of this registry will allow both patients and clinicians to enter additional phenotypic information directly. It will also allow patients to indicate their willingness to be contacted for future studies, facilitating contact between them and interested researchers for both exploring unsolved genetic cases as well as supporting genotype-phenotype studies. The combination of these approaches capitalizes on many different potential sources of phenotypic data, increasing the scope of our phenotypic data collection process.

Collection of phenotypic data on vast numbers of patients faces two main challenges. First, there are consent issues, as addressed in Aim 1. 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 continue to be submitted on a regular basis from laboratories also submitting genotype information. Clinical laboratories will provide this type of phenotypic data at a minimum as proof-of-principle for this mechanism of data capture. 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 have been more successful in collecting detailed clinical data through disease-specific check-box forms, and, as such, we will create additional disease-specific forms and share them with all laboratories to aid in data collection. Curated data from locus-specific databases (LSDBs) 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.

Although this level of data appears to be minimal, the provision of at least some phenotypic data is more valuable than none; further, by not *requiring* in-depth phenotypic data, often necessitating full consent of the patient, many more cases can be collected. Our efforts at data collection for the ISCA Consortium have illustrated this point when compared to the data collection model of the DECIPHER database, another database of structural variation (http://decipher.sanger.ac.uk/). The DECIPHER database has attempted to collect very detailed phenotypic data on all patients, requiring a full consent process in order to make such data publically available; as a result, this database has achieved many fewer publically available database entries. The ISCA Consortium has historically collected a minimal amount of phenotypic data, such as a test indication from a clinical laboratory, and this information is displayed in a standardized, de-identified format (HPO terms); we have been able to do this utilizing an opt-out process. Because of this, the ISCA database already has at least five times the number of publically available data entries compared to DECIPHER, yet 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. The addition of our optional patient registry is unique in that it provides an optional mechanism by which registered patients and/or their clinicians can add additional detailed information, but this is not a requirement for data submission. As long as the patient has not opted-out of the database in its entirety, cases can still be submitted with the basic phenotype information provided to the laboratory.

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 during the test reporting process. In ISCA, laboratory genetic counselors working in the EEAW were able to increase the submission of phenotypic information by directly contacting providers who frequently submitted samples to their laboratories and requesting that they use the phenotype form. Other laboratory genetic counselors at independent laboratories mirrored these efforts and were also encouraged to request more clinical information (if not provided when the sample was submitted) when calling out abnormal array results. 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, and/or providing access to secure online portals in which to enter such information, such as PhenoDB. 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 clinical data collection forms that can be downloaded from our websites, including the ISCA and ClinVar websites.

Although clinicians will be the best source of clinical phenotypic data, we anticipate that participation by busy clinicians will be variable and should therefore not be our sole source for the collection of detailed phenotypic data. Because of this, 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 rare chromosome disorder support group), the Hypertrophic Cardiomyopathy (HCM) Association, and CureCMD (Congenital Muscular Dystrophies), 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.Patient Crossroads.com/), has a demonstration project with dbVar for Congenital Muscular Dystrophy where they are linking mutation results from clinical laboratories with phenotypic information from their 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.

Creation of a new, online patient registry to facilitate enhanced phenotype data collection and research initiatives

Not all patient groups will have an existing patient registry in which they can register, such as those with rare disorders, or those that have not yet received specific diagnoses. Further, it is cumbersome to establish multiple

distinct disease-specific registries, and then attempt to integrate each one into a unified environment. To address these issues, we will work with Patient Crossroads, currently supporting the Global Rare Diseases Registry project (http://www.grdr.info), to create a new patient registry that will support registration of any patient who has undergone clinical genetic testing. One advantage of this approach is that it will allow both patients with any disorder and their clinicians to enter phenotype information through a structured electronic data capture system. Patients will complete phenotypic questions that have been validated and shown to have a high level of accuracy when completed by patients. This will complement clinician-entered data and alleviate some of the burden referring clinicians face regarding providing such information. The interface will also allow clinicians to continue to access phenotypic data on any of their patients that have registered and consented to this access, motivating them to use the system to support their own research and clinical interests. 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.

Contacting patients to collect additional phenotypic information or to recruit them for research is often difficult for clinical laboratories. The laboratory must send the request to the ordering clinician who must then reach out to the patient. Online patient registries have successfully solved this roadblock by not requiring additional work from busy clinicians. DuchenneConnect (https://www.duchenneconnect.org/), CureCMD (http://curecmd.org/), and Simons VIP Connect (http://www.simonsvipconnect.org/) are successful examples of patient based registries that contain both phenotypic and genotypic information. On 2/10/2012 the NIH Office of Rare Disease Research (ORDR) and Patient Crossroads announced the Global Rare Diseases Patient Registry and Data Repository (GRDR) (http://rarediseases.info.nih.gov/GRDR) to develop 15 new patient registries and link with 20 existing registries, representing over 40 genetic conditions. This registry model was chosen due to the ability to collect large volumes of phenotypic information with a high level of accuracy. Each registry chose a subset of the information including the genotype information to review and validate as part of the registration process. Our proposed registry will go beyond this important effort, bringing the benefits of an organized registry program not only to other patient groups with rare diseases that may not otherwise be able to fund/organize an independent registry effort, but currently undiagnosed patients as well. Patient Crossroads is providing the software platform for the GRDR, and we will leverage the same platform for this project. Co-Investigators in this proposal have worked with Patient Crossroads previously on the development of DuchenneConnect and Simons VIP Connect.

Online registration of patients after genetic testing

Participating laboratories will include information regarding the availability of this registry as part of the clinical report, encouraging patients to register online. One-page handouts will also be developed and distributed to the laboratories to include with each report: one handout will focus on patients with negative or inconclusive results, explaining why participation is important, while a second handout will focus on patients who received a positive result, explaining how registering online can enhance research in their disease.

At registration patients will complete a series of questions regarding: molecular genetic and cytogenetic testing history and results; clinical status by review of organ systems; and their interest in research. As part of the GRDR, Patient Crossroads has developed a question bank that includes over 400 questions to collect phenotypic information that have been tested and found to have high reliability when completed by patients. Having similar questions in multiple registries allows researchers to search across registries and diseases for common phenotypic results. The GRDR, working with Children’s Hospital of Philadelphia, plans to expand this question bank to over 2,000 validated questions to cover most medical conditions. In addition, responses to select questions will be validated by the Project Coordinator (to be named) for the registry. This process will include reviewing laboratory reports (mailed or uploaded by the patients) to confirm the genotype and reviewing other reports that are felt to be critical for a given diagnosis, e.g.., cardiac and lung function in Duchenne Muscular Dystrophy. Information on the website will explain how to request a copy of their report. This process has been successfully employed by DuchenneConnect, Simons VIP Connect and CureCMD to insure the accuracy of the genotype data. Based on the patient’s diagnosis, other clinical reports have been requested and curated by the Project Coordinator for each online registry. Patient Crossroads and CureCMD have successfully uploaded genetic test result data and phenotypic data to ClinVar and dbGaP. In previous work with multiple groups developing online patient portals, we have determined that too many questions on the registration survey results in a lower registration rate. In general, the patient should be able to complete the survey in thirty minutes or less. Registries developed by Patient Crossroads are secure. Patient Crossroads is hosted in a SAS 70 Type II HIPAA compliant infrastructure with dedicated firewalls and advanced intrusion detection to secure patient data. All registry network transmissions are encrypted for an added level of protection. The technology is built on RedHat Linux, MySQL and other opensource

standards. In addition, only the Project Coordinator will have access to any particular patient's genetic testing data, as needed in order to curate the information.

Benefits of online registration for patients

In previous online registry development we have found it to be important to provide immediate rewards (information) to patients to encourage completion of their online profile. After completing their profile, registered patients will be able to compare their results to the aggregated results of the registry community. They will also be able to refine their comparison to a portion of the community. For example, they could compare to people with the same genetic test result, same clinical symptoms, same gender, and/or same age. In most cases visual representation will also be presented automatically. The website will include a method for families to search for other families with a similar diagnosis, location and age range and send an invite to begin an offline conversation. The system does not share email addresses but allows the registrant to send an introductory email to another member. If that registered member is interested in communicating, they can respond to the email and then the two patients/families are linked. Registrants can also receive information about research opportunities based on their genotype or phenotype. Through the registry new studies will be announced by targeted emails based on the registrant's genotype, genetic testing status, clinical symptoms or other criteria.

Online registration of researchers and clinicians:

We will work with professional organizations, dbGaP, and ClinVar to publicize the registry and the ability to perform phenotype and genotype searches to the research and clinical communities. Researchers and clinicians will be able to register by providing an institutional email address and contact information. Once verified, they will be allowed to search de-identified information within the registry by phenotype, genetic test results, demographic information, or a combination.

Benefits to researchers

Once registered, researchers will be able to perform searches as needed by selecting multiple parameters from the survey questionnaire. Limits will be in place to not return queries that might be used to identify a registrant. The Project Coordinator will be able to search on any combination of genotype and phenotype information and can answer researcher questions that fall below this safeguard. The researcher search capability has been used by researchers on *DuchenneConnect* to determine numbers of individuals available for studies and the feasibility of recruiting individuals quickly. There will be a tool for researchers to submit queries to subsets of registrants after review by the Project Coordinator. Researchers may promote an open study by submitting a request to the Project Coordinator with proof of IRB approval. Targeted emails can then be sent to registrants that appear to qualify for the study. Historically the response rate has been very high because individuals registered on the website have stated an interest in research. Once interested participants respond to the targeted email and are consented, the researcher would be able to contact the laboratory that performed the patient's molecular or cytogenetic test and request additional data files not stored on the registry website. The researcher may also be able to request a link to the data files stored in ClinVar or dbGaP.

Connecting the patient registry with ClinVar and dbGaP

To protect patient privacy, dbGaP and ClinVar do not collect any demographic or contact information. Through dbGaP it is possible to request the laboratory that submitted a data file and then ask the laboratory to request additional information from the clinician or ask the clinician to share information about a research project with the patient. This system is cumbersome and dependent on clinician cooperation and time to make the connection. In this proposal we plan to develop systems that would reduce the burden on the clinician. As part of registration, participants will provide a copy of their clinical laboratory report. The test results and testing laboratory will be entered into the patient registry database. This will allow the researcher to determine if a participant's full data file is likely stored in dbGaP. We will develop a system where the participant will provide consent and the researcher can request data from the laboratory or request a link to the data file stored at dbGaP. Patient Crossroads has developed a system of GUIDs (Globally Unique Identifiers) with the NIH to link registry participants with biological samples stored in biobanks at Coriell and Rutgers. We will work to develop a similar system to link registrant's phenotypic information with their data files in dbGaP and ClinVar.

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 developing and demonstrating efforts to

curate the data, including improving the classification of variants with respect to their role in human health and disease. The curation process is integral to the success of this and any effort attempting to determine the functional significance of genomic variation. Unfortunately, there is not a consensus amongst the field regarding the most efficient way to curate this type of information. As the task is monumental in scope, there is a push towards automated curation; this method would result in easy scalability, but would inevitably lose the nuanced reasoning of curation by skilled experts. Manual curation has the benefit of the experience and knowledge of the individuals evaluating each variant, but is ultimately resource-heavy in terms of time and financial support. One goal of this project is to attempt to come to consensus on the most efficient methods for data curation, particularly in terms of large-scale datasets, such as those from whole exome or whole genome sequencing. We will use demonstration projects on several different genetic disorders, taking the lessons learned and integrating them into a more streamlined process for these large datasets. The different levels of curation (e.g. strength of evidence and degree of consensus) will be described and clearly noted in ClinVar, enabling users to filter the data based upon the level of curation or the types of clinical assertions (benign to pathogenic) on the variants. 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. As shown in Figure 5 and defined below, data curation will occur at one of five different levels.

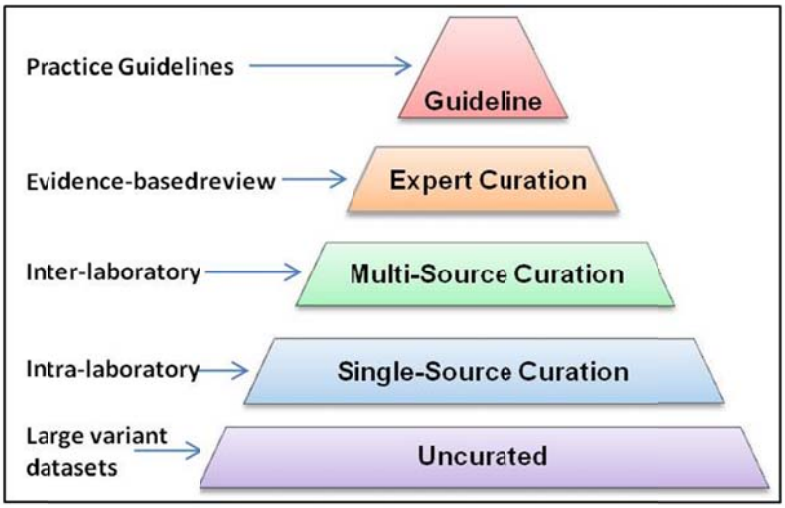


Figure 5. Levels of curation for clinical assertions

1. Uncurated

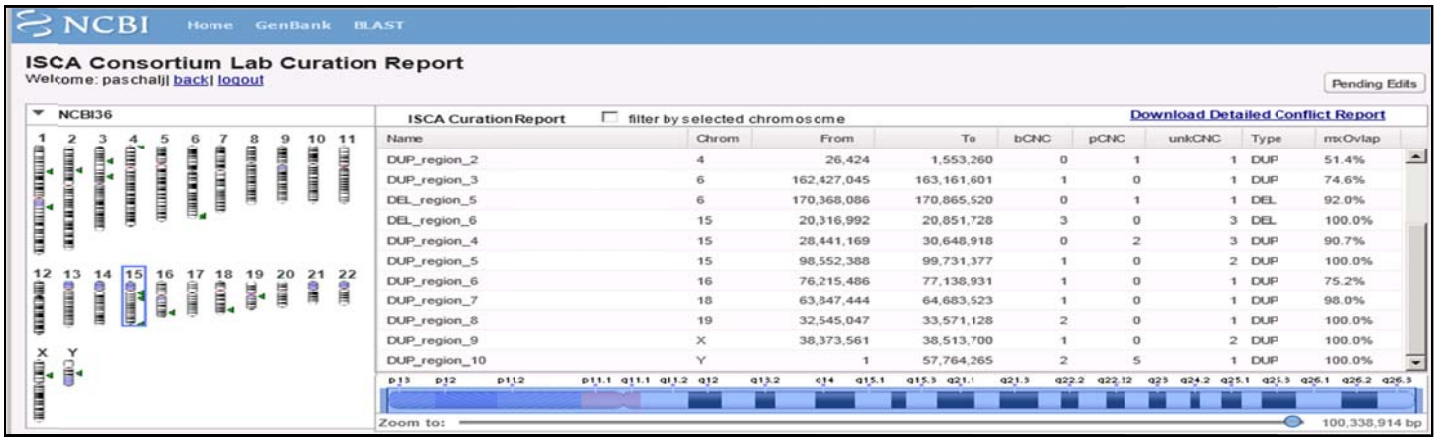
Data viewable within ClinVar without a clinical classification, such as population data from the 1000 Genomes project [2], NHLBI Exome Sequencing Project (eva.gs.washington.edu/EVS), or other large scale datasets such as non-classified variants from the increasing number of exomes and genomes being sequenced in clinical laboratories, would be considered “uncurated.” Much of this data will have been deposited directly into dbGaP with variants accessioned in dbSNP and dbVar, but this uncurated data will also be accessible within ClinVar, particularly to enable viewing of population frequency data.

2. 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 curated at this level. In an effort to improve the accuracy and standardization of single-source curation, we have developed, in partnership with NCBI, several algorithm-based curation tools and integrated them into the data submission process for the ISCA Consortium. The goal of this effort is to improve the quality of data entering the database while providing immediate value to the submitter through validation and consistency checks. Data curation is carried out at multiple levels: 1) quality checks, described in section 2a, are performed, 2) regions where the same laboratory has reported different clinical calls for similar observations are identified and flagged, and 3) the submitted calls are compared against an expert curated list of regions with established clinical significance (http://www.ncbi.nlm.nih.gov/dbvar/studies/nstd45/). 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 (Figure 6).

Each identified "conflict" region is presented to the submitter who has the option to click on the variant and change the call, or choose 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 prior to integration with the full dataset. To date, approximately 5-10% of variants submitted to the ISCA database have been flagged to the submitting laboratory

due to a discrepancy in variant classification within the laboratory's own dataset. This could be due to laboratory effort, or the changing understanding of certain variants over time. Regardless of the reason(s) for discrepancies, this type of curation process has already demonstrated significantly enhanced quality control for clinical reporting.



:n arrows

Similar laboratory-level curation tools will be developed for sequence-level data. Those tools will be available within ClinVar as described above for structural variant submission. In addition, as noted in Aims 1 and 2, we will also work with genetic software vendors that support variant evidence assessment to ensure the incorporation of consistent terminologies and rules for the clinical classification of variants to enable improved single-source data curation.

3. Multi-source Curation

Given that many laboratories will be submitting variants with clinical assertions to the database, ClinVar will automatically generate discrepancy accessions when submitted assertions are in conflict. These discrepancy accessions will then be available in the system to warn users of the discrepancies and to enable expert-level curation as described below. For sequence level variants, the tool to develop the discrepancy accessions will first ensure that the variant submissions match at the genomic reference level and then it will compare the clinical assertions made. Given that laboratories use similar but not always identical terminologies, we will develop a term mapping process to map each laboratory's terms to a common set of terminologies. For example, if the system supports "Likely Pathogenic" as the term of choice but another lab reports these variants with "Likely Deleterious", the term-mapping system will ensure that these assertions are not viewed in conflict. Over time, as the standards for variant classification are incorporated broadly, term mapping will not be needed; however, given the large amount of retrospective data that we intend to capture, these mappings will be critical.

Additionally, submitted structural variants will also be curated against other submissions. 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.

4. Expert-level Curation

To improve our understanding of disease relevance for individual variants 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 [27]. Furthermore, as evidence available regarding particular variants changes over time, clinical classifications made at the time of data submission may require re-evaluation [28]. 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, we established an Evidence-Based Review Committee (EBRC) within the ISCA Consortium that will become part of the role 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 [17]. 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 [17]. 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 through an automated notification system.

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 “likely benign” or “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/).

The clinical assertions of structural variants submitted to our database are then curated against this set of known pathogenic and benign regions. Currently, there are approximately 2,000 variants in the ISCA database that overlap with an ISCA known pathogenic region. Of these, approximately 4.6% were found to be in conflict, demonstrating the utility of this process. These discrepancies are subsequently reviewed with the individual submitting laboratories so that they can re-review their data and possibly change their original interpretation

This model of expert-level curation will be modified for use for sequence-level variants. Realizing that this effort will be a substantial undertaking, demonstration projects will be necessary in order to build the infrastructure needed to efficiently curate sequence-level data. Initial efforts will focus on expert-level curation of variants in several hundred genes included in eight model projects, each focusing on a different clinical phenotype. To date, most of the successful projects in evidence-based curation of data have focused on gene or disease-specific databases to organize communities for expert curation [InSiGHT (http://www.insight-group.org), ENIGMA (http://enigmaconsortium.org), CFTR 2 (http://www.cftr2.org), etc.]. However, developing separate infrastructure for each disease-specific activity is time-consuming, resulting in a bottleneck in this process. Therefore we will provide a basic environment for expert groups to organize around their disease domains. In some cases, the infrastructure provided by ClinVar will be sufficient to provide the necessary tools and data fields for evaluation of variants. In other cases, there may be a need to interface to an external system, particularly if sharing patient identifiable data. We intend to explore both models in several demonstration projects. For example in one project, we will work with the InSiGHT group, which has developed a database for genes involved in gastrointestinal hereditary tumors such as colon cancer (http://www.insight-group.org). As noted in the letter of support provided by Finlay Macrae, we will work closely with this group to exchange data. Matthew Ferber, lab director at Mayo Clinic, one of the largest providers of colorectal cancer testing in the US, will organize data submission from US clinical laboratories into the InSiGHT database. In exchange, the InSiGHT database will deposit their data into ClinVar, much of which has already gone through expert-curation.

In other disease areas where no community organized databases exist, we will create the environment for curation within ClinVar. This includes activities for rasopathies led by Sherri Bale, developmental delay led by Soma Das, inborn errors of metabolism led by Elaine Lyon and Rong Mao, *PTEN* related disorders led by Madhuri Hegde, and cardiomyopathies led by Heidi Rehm. For each of these efforts, we have already engaged between 4-20 laboratories for each disorder who have agreed to submit data and organized a broad group of experts for each disease area from research, clinical practice and laboratory genetics to work together to use evidence-based approaches to improve the classifications of variants. We will also support a curation project on congenital muscular dystrophies led by Madhuri Hegde, to demonstrate mechanisms for engaging with patient support organizations and allowing data collected through patient registries to be passed to ClinVar. For example, the CureCMD registry combines a patient registry and coordinated collection of genetic test reports from clinical laboratories to enable high quality genotype-phenotype data collection (see letter of support from Anne Rutkowski, president of CureCMD). Dr. Rehm will also work closely with the Hypertrophic Cardiomyopathy Association (HCMA) which maintains a patient repository and work to enable the mutual exchange of phenotype and genotype data from consented patients to achieve the same goals (see letter of support from Lisa Salberg, president of the HCMA). We also have a project to gather data for breast cancer genes. Given patent protection for testing in the US and

the lack of submission of data into the public domain from the sole provider of BRCA1/2 testing in the US, Robert Nussbaum has organized an effort to work with breast cancer clinics to collect data from patient test reports. To date, 43 clinics have agreed to participate in the project and information on approximately 2,000 variants in BRCA1/2 has been collected.

Recognizing the benefit of collaboration when undertaking such a substantial project as curation of genome-wide variation, we intend to foster working relationships with other groups working to curate human variation. In addition to the demonstration projects supported by our own group, we will work closely with groups such as PharmGKB and the Cancer Cytogenomics Array Consortium, to learn from their efforts as well as bring additional data to their efforts. Further, we will work with these groups to encourage the deposition of their curated variants into the ClinVar database. For groups at the early stages of organizing clinically relevant database efforts, we will work closely with those groups to enable them to use the methods and infrastructure developed to support this project. For example, we are developing a collaborative relationship with the mitochondrial disease community, led in part by Dr. Marni Falk (see letter of support), to share infrastructure and methodologies to support an effort that has united clinical laboratories, researchers, clinicians and the United Mitochondrial Disease Foundation to coordinate the collection and sharing of DNA variant data from targeted mitochondrial and nuclear gene and whole exome testing in patients with mitochondrial diseases.

The goal of all of the expert-curation demonstration projects will be analogous to those initial goals of the ISCA evidence-based review committee, including: A) establishing the *type and amount* of evidence needed to determine the pathogenicity of a given variant, B) establishing, in conjunction with the BIT working group, an efficient process by which to procure this evidence, and C) establishing an efficient process by which to evaluate this data. Once these models are developed and tested in the demonstration projects, we will examine rules of evidence across specific disease areas to allow us to elucidate patterns and approaches that will apply to other disease areas. This way, lessons learned from the seed domains enable curation rules to be determined for additional diseases. We will create a set of recommendations for supporting expert curation of human genetic variation for rare diseases allowing these activities to scale across many other clinical domains.

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 (either by targeted testing or whole exome/genome sequencing) and expert curation activities are needed. This workgroup will initiate new projects, giving higher priority to those diseases and gene sets for which multiple laboratories offer testing and for which the largest number of patients stand to benefit. 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. NCBI has also committed to directly supporting these efforts through their ongoing work on ClinVar and the Genetic Testing Registry projects.

5. 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 [29]. Very few variants will exist in this curation level, but those that are published within professional guidelines will be noted as such. We will also work with NHGRI's future Clinically Relevant Variant Resource group to ensure that the activities of this funded effort are best enabled by our project.

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 (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 evolution 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 (if submitting individual CMA, WES or WGS datasets) 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 3,000 unique visitors in the last month (April-May 2012), 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. NCBI is also maintaining a website for the ClinVar project and have created a page to directly link with community projects interacting with the effort (http://www.ncbi.nlm.nih.gov/clinvar/community/) enabling further awareness and education about ongoing efforts.

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. (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 those 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 over a year ago, this toolkit has had over 3,000 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 2011, 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. A second conference was held in 2012, representing the first joined effort between the structural variant and sequence variant communities.

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 [30] and an article was published in the *Journal of Genetic Counseling* describing the role of genetic counselors in ISCA and the importance of clinicians working with laboratories to increase submissions to databases [31]. Additional articles discussing the output of the ISCA database [15] and recommendations for CMA use [12] have been published. Additional articles on the curation process [17] and the phenotype data collection process [26] have also been published. 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. Many of the existing patient advocacy organizations already have 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. In addition to continuing to work with these existing groups, such as Genetic Alliance, UNIQUE (the rare chromosome disorder support group), Duchenne Connect, and the HCM Association, we will engage the larger patient community of individuals referred for genetic testing to participate in our patient registry, which will be integrated with the Global Rare Disease Registry network. We will work with the patient advocacy community as a whole to provide input on both the clinical data collection forms discussed in Aim 1 and the registry development as discussed in Aim 2b. 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 previously 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, and will use this experience to guide us as we create the patient registry described in this proposal.

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.

Sustainability of ClinVar and systems for improving our understanding of human variation

We have had discussions with several groups about how to sustain the long-term submission of data by clinical laboratories. Laboratory directors and laboratory genetic counselors are now recognizing 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 informed us that the NIH Genetic Testing Registry (GTR) plans to add a field where laboratories may indicate whether they participate in data exchange programs such as ClinVar, the ISCA Consortium, and LSDBs, which may influence customer choice of testing laboratories and allow market forces to stimulate data submission.

Most importantly, through our work with the ISCA Consortium and in starting to develop the same model for the sequencing community, we have found that laboratories are most eager 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.

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.

BIOINFORMATICS AND IT COORDINATION

The Bioinformatics and IT Workgroup (BITW) will be made up of staff at NCBI and staff at the participating clinical laboratory sites, and overseen jointly by Joyce Mitchell, one of the principal investigators, and Sandy Aronson, the Director of IT at Partners Healthcare Center for Personalized Genetic Medicine. The BITW will serve as the coordinating body for our information technology efforts. It will not directly build or maintain IT infrastructure. Instead, it will define and constantly refine documentation, protocols and processes that will facilitate the grant related IT activities performed by laboratories, vendors and NCBI.

The **BITW** is tasked with:

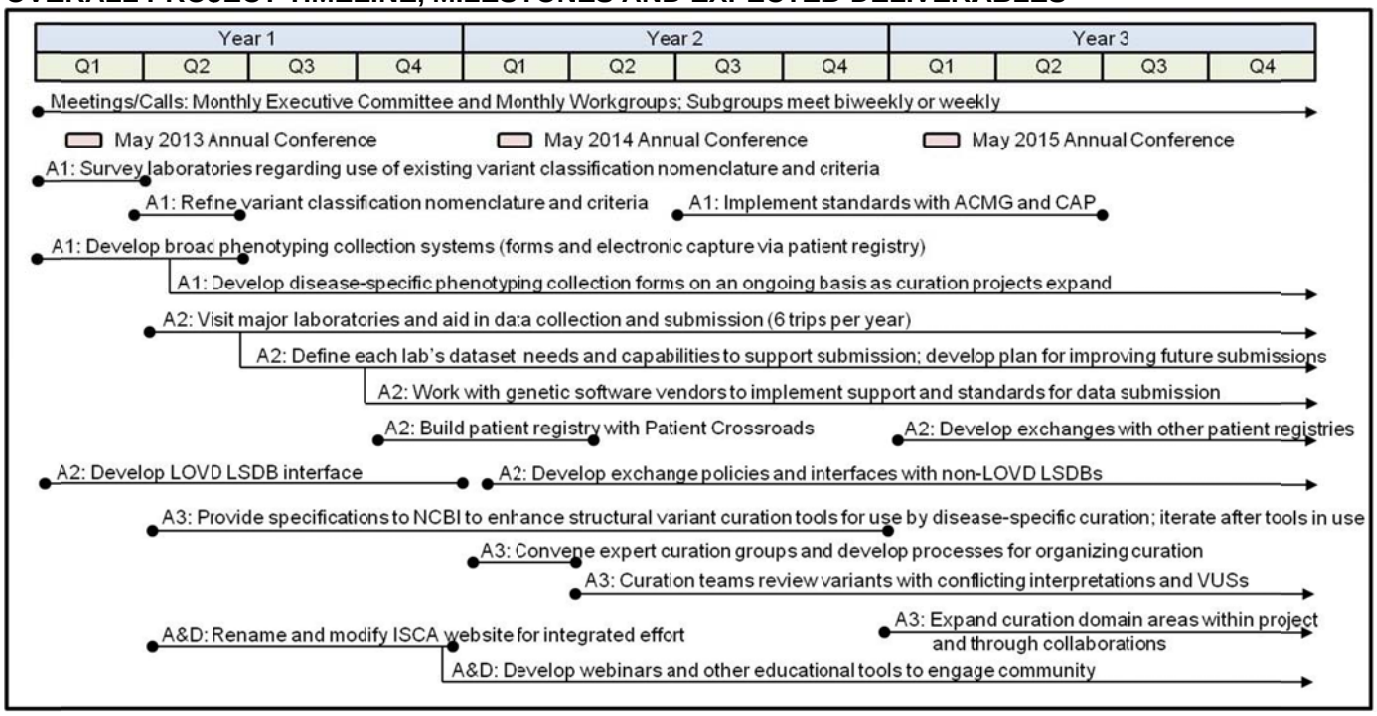
1. Cataloging existing laboratory capabilities and infrastructure: The BITW will work closely with laboratories and their supporting vendors to understand in detail both what data laboratories are currently tracking and what infrastructure is in place to support the transmission of these data to NCBI.
2. Data content and structure: The BITW will work closely with NCBI and the PSSW to develop and review the data dictionaries, ontologies, file formats, and basic content to be collected on variants and cases. This will include data formats for raw variant and case data submission as well as content and structure for variant annotations. The BITW will also work with both genetic software vendors and the laboratories themselves to ensure that data dictionaries and related ontologies are consistently implemented whenever possible. When consistency is impossible, differences will be noted so term mapping and reformatting can occur as needed. This will be an iterative process. Updated data dictionaries and ontologies will be released as the data collected by laboratory tracking capabilities evolve from existing tests on single genes or panels into full sequence analysis over the whole genome.
3. Data submission methods: The BITW will work with each laboratory as well as genetic software vendors to develop robust methods for data submission. As laboratories, vendors and NCBI implement processes required to support data submission, the BITW will track inter-organizational dependencies and facilitate the resolution to problems as they arise. Data conversion efforts to extract retrospective data from past cases will be supported while processes are established to more efficiently submit new cases.
4. Data curation systems: The BITW will specify enhancements for NCBI that will be required to evolve the curation tools developed for the ISCA Consortium to enable curation of all types of genomic variation.
5. Data security and access: The BITW will be responsible for working with NCBI, contributing laboratories, and genetic software vendors to ensure the security of the data with respect to access and adherence to HIPAA standards as well as other regulations imposed by IRBs and local, state and federal guidelines.
6. Staff and system training: The BITW will assist in training staff at submitting laboratories as well as aiding NCBI and the Engagement, Education, and Access Workgroup in developing educational resources for users of the database.

Table 2. Chart of IT responsibilities

	BITW Meetings	Analysis & BioFx Teams	Labs & Expert Curators	NCBI	Vendors
Aim 1					
Catalog each laboratory's / vendor's capabilities					
Propose initial submission process/standards based on lab capabilities					
Serve as laisons between laboratories and vendors in explaining needs / capabilities					
Respond to feedback to create version 1.0 submission process standard					
Aim 2					
Implement infrastructure required to perform submissions					
Facilitate issue characterization and resolution strategies throughout implementation process					
Implement infrastructure and process changes required to address issues					
Implement de-identification infrastructure					
Assist in reformatting initial data loads and define procedures for ongoing submissions					
Term mapping process and infrastructure					
Implement additional infrastructure for enhanced phenotype data collection					
Maintain interfaces to and exchanges between dbGaP, dbVAR, dbSNP and ClinVar					
Define dataflows, exchanges and interfaces between NCBI and external systems					
LOVD - ClinVar bidirectional interface (collaboration with Johan den Dunnen)					
PatientCrossroads - Lab data integration - keeping GUID's consistant					
Aim 3					
Data enrichment and conflict resolution infrastructure for expert review					
Extend current ISCA curation software to support sequence variant curation					
Support exchange models for external expert curation groups (InSiGHT)					
Intake of annotations from expert review					
Resourcing of Additional Activities					
Implementation of website changes: Webmaster will be assigned this task					
Evaluate HIPAA/regulatory/ethical aspects of IT processes: Consultants (Joan Scott and Jennings Aske) will assist in this area					

Legend: Black/Blue indicates primary driver of the activities. Grey/Green indicates facilitation roles.

OVERALL PROJECT TIMELINE, MILESTONES AND EXPECTED DELIVERABLES



ADMINISTRATION AND MANAGEMENT

Executive Committee:
Robert Nussbaum (PI and chair), David Ledbetter (PI), Christa Martin (PI), Joyce Mitchell (PI), and Heidi Rehm (PI)

A. Organizational structure and staff responsibilities

The project will involve six workgroups overseen by the executive committee members as diagrammed in Figure 1. Drs. Rehm and Martin will oversee the sequence and structural variant projects, respectively. The Executive Committee will meet by conference call on a biweekly 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 project is progressing appropriately and 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 six workgroups, the Policies, Standards, and Sustainability Workgroup will meet monthly and 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 and the chairs of the five other workgroups (Aradhya, Aronson, Faucett, Hegde, Miller, Thorland, Willems), 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 Bioinformatics and Information Technology Workgroup will also have deliverables that impact all activities of the grant. This group will be co-directed by Joyce Mitchell (PI) and Sandy Aronson, Director of IT at Partners Healthcare Center for Personalized Genetic Medicine (PCPGM). The group membership will include Karen Eilbeck (Utah), 3 bioinformatics/IT staff from PCPGM, ClinVar project staff from NCBI (Donna Maglott and Deanna Church) as well as the bioinformatics/IT staff from each of the contributing laboratories for both sequencing variants and structural variants. There will be monthly meetings of the entire group with attendance by Jennings Aske (Chief Information Security Office at Partners Healthcare) and ad hoc participation by Joan Scott (ELSI representative). Subgroups of the BITW will meet weekly. The workgroup leaders will review the standards and structure of all of the component data sets that will be submitted to ClinVar and put in place the processes needed to ensure high quality data is submitted and efficient methods of data submission are developed. The leaders will be part of the PSSW to assist with the discussions of the standards and the abilities of the individual data contributors to comply with those standards.

The Sequence Variant Workgroup will be chaired by Drs. Hegde and Willems and meet monthly. Both will be responsible for all activities of the working group including overseeing the development of evidence-based criteria for the classification of sequence variants and overseeing all sequence level model curation projects. 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.

The Structural Variant Workgroup will be chaired by Drs. Aradhya and Thorland and meet monthly. They 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. Church, Kearney, Martin, and South, 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 and meet monthly. 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 Executive Committee.

The Engagement, Education, and Access Workgroup will be chaired by Mr. Faucett and meet monthly. 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, in addition to guiding the development of the patient registry as described in Aim 2b. Mr. 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 (EC), 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 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. Drs. Rehm and Martin will supervise the collection of progress reports from workgroups and summarize progress for presentation at SAB and annual meetings. In addition, they 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. Consortium, G.P., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.
3. Tennessen, J.A., A.W. Bigham, T.D. O'Connor, W. Fu, E.E. Kenny, S. Gravel, S. McGee, R. Do, X. Liu, G. Jun, H.M. Kang, D. Jordan, S.M. Leal, S. Gabriel, M.J. Rieder, G. Abecasis, D. Altshuler, D.A. Nickerson, E. Boerwinkle, S. Sunyaev, C.D. Bustamante, M.J. Bamshad, and J.M. Akey, *Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes.* Science, 2012.
4. 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.
5. 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.
6. 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.
7. 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.
8. 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.
9. Cook, E.H., Jr. and S.W. Scherer, *Copy-number variations associated with neuropsychiatric conditions.* Nature, 2008. **455**(7215): p. 919-23.
10. 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.
11. Kearney, H.M., S.T. South, D.J. Wolff, A. Lamb, A. Hamosh, and K.W. Rao, *American College of Medical Genetics recommendations for the design and performance expectations for clinical genomic copy number microarrays intended for use in the postnatal setting for detection of constitutional abnormalities.* Genet Med, 2011. **13**(7): p. 676-9.
12. 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.
13. 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.
14. 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.

15. 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.
16. 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.
17. Riggs, E.R., D.M. Church, K. Hanson, V.L. Horner, E.B. Kaminsky, R.M. Kuhn, K.E. Wain, E.S. Williams, S. Aradhya, H.M. Kearney, D.H. Ledbetter, S.T. South, E.C. Thorland, and C.L. Martin, *Towards an evidence-based process for the clinical interpretation of copy number variation.* Clin Genet, 2012. **81**(5): p. 403-12.
18. 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.
19. Services, D.o.H.a.H., *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.
20. Reese, M.G., B. Moore, C. Batchelor, F. Salas, F. Cunningham, G.T. Marth, L. Stein, P. Flicek, M. Yandell, and K. Eilbeck, *A standard variation file format for human genome sequences.* Genome Biol, 2010. **11**(8): p. R88.
21. Danecek, P., A. Auton, G. Abecasis, C.A. Albers, E. Banks, M.A. DePristo, R.E. Handsaker, G. Lunter, G.T. Marth, S.T. Sherry, G. McVean, and R. Durbin, *The variant call format and VCFtools.* Bioinformatics, 2011. **27**(15): p. 2156-8.
22. 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.
23. 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.
24. 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.
25. 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.
26. Riggs, E.R., L. Jackson, D.T. Miller, and S. Van Vooren, *Phenotypic information in genomic variant databases enhances clinical care and research: the International Standards for Cytogenomic Arrays Consortium experience.* Hum Mutat, 2012. **33**(5): p. 787-96.
27. 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.
28. 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.
29. 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.

30. 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.
31. Wain, K.E., E. Riggs, K. Hanson, M. Savage, D. Riethmaier, A. Muirhead, E. Mitchell, B.S. Packard, and W.A. Faucett, *The Laboratory-Clinician Team: A Professional Call to Action to Improve Communication and Collaboration for Optimal Patient Care in Chromosomal Microarray Testing.* J Genet Couns, 2012.

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 primarily clinical testing results to a common public database to improve understanding of genomic variation and the interpretation of results (benign, pathogenic, etc.). This "digital biobank" of data 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.

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 "inconclusives")
   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. A 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 process 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. (See online process below).

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.

Online Collection of Data from Patients

The online registry will use the Informed Consent process and model consent form recommended by the NIH Office of Rare Diseases Global Rare Diseases Registry (GRDR) (http://www.grdr.info/index.php?option=com_content&view=article&id=28&Itemid=44). As part of the registration process all registrants will be shown the online consent form and required to accept the terms before registering. The project will seek IRB approval, but it is likely that the registry will receive an IRB exemption because the primary purpose of the registry is not research. It is designed to facilitate recruitment for other IRB approved research projects. The Emory University IRB determined that the online patient registry DuchenneConnect was exempt. Both the Emory University IRB and the Geisinger IRB determined that the online patient registry Simons VIP Connect was exempt.

The registry will collect patient submitted information and will not participate in clinical care. As a result, HIPAA does not apply to the registry. Only the registered participant and the Registry Coordinator will have access to identifiable information. Contact information will never be shared. Researchers or clinicians interested in contacting a participant will need to provide evidence of an IRB approved protocol. The Registry Coordinator will then forward contact requests from the approved researcher/clinician to the registrant without revealing their contact information. If the registrant is interested in responding to the request for information or wishes to participate in the study, they will be able to reply to the email and will then be linked with the researcher/clinician. Participants can then assist the researcher in gaining access to their data files from the laboratory or from dbGaP.

Registered researchers and clinicians with valid institutional email address and contact information will be able to search de-identified aggregate data in the registry. The search criteria will be restricted to insure that no query returns less than 5 responses to protect participants and to limit the possibility of identification.

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 a robust evidence-based assessment of human genomic variation to improve patient care.

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 Partners Healthcare 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 the 6 funded sequencing laboratories who have agreed to participate in data submission and the eight model curation projects as well as 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: A Unified Clinical Genomics Database

Total Planned Enrollment: 300,000

TARGETED/PLANNED ENROLLMENT: Number of Subjects			
Ethnic Category	**Females**	**Males**	**Total**
Hispanic or Latino	9,159	9,159	18,318
Not Hispanic or Latino	140,841	140,841	281,682
Ethnic Category: Total of All Subjects *	150,000	150,000	300,000
Racial Categories			
American Indian/Alaska Native	778	778	1,556
Asian	10,063	10,063	20,126
Native Hawaiian or Other Pacific Islander	265	265	530
Black or African American	13,825	13,825	27,650
White	125,069	125,069	250,138
Racial Categories: Total of All Subjects *	150,000	150,000	300,000

* 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 the project requires a broad interdisciplinary approach that brings together expertise and experience in molecular genetic and cytogenomic testing, laboratory testing policies and procedures, bioinformatics and medical informatics, human genetics research, and clinical genetics practice. Because molecular genetic testing and cytogenomic testing are traditionally carried out in laboratories that fall into two separate communities, it is critical to have experts in both these areas represented in the multiple PI leadership group. We accomplish this goal by including Drs. Ledbetter, Rehm and Martin, who have strong national profiles as laboratory directors with expertise in both testing areas. These three PIs are joined by Drs. Joyce Mitchell, and Robert Nussbaum, who are nationally recognized experts who together bring nearly 60 years of experience in human genetics research, laboratory testing, and medical and bioinformatics research to the project.

Multiple PI Leadership (in alphabetical order)

David H. Ledbetter, Ph.D., FACMG. Dr. Ledbetter is the Executive Vice-President and Chief Scientific Officer of Geisinger Health System. Dr. Ledbetter is an ABMG Board-certified clinical cytogeneticist and an expert on genomics technology and translation to clinical genetic testing. Dr. Ledbetter is formally trained in human and medical genetics, with American Board of Medical Genetics certification in Clinical Cytogenetics, and over 30 years' experience in genetic testing in a CLIA-certified laboratory environment. He has extensive experience in the translation of new genetic and genomic technology into clinical practice primarily in the area of genetic diagnostics in pediatrics and prenatal diagnosis. He has participated in and had leadership roles in many multicenter collaborative research projects, including one of the first Human Genome Centers at Baylor College of Medicine, an Intellectual and Developmental Disabilities Research Center (Baylor), and currently participates in an NIH multicenter study of whole-genome chromosomal microarrays in prenatal diagnosis (R. Wapner, PI) and an Autism Center of Excellence (ACE) Network (D. Geschwind, PI). Dr. Ledbetter has held a number of leadership positions in genetics and genomics, including as a Founding Branch Chief for the National Human Genome Research Institute's Intramural Program (1993-1996), the Founding Chair of the Department of Human Genetics at the University of Chicago (1996-2003), and as Director of the Division of Medical Genetics at Emory University School of Medicine (2003-2010). In this latter role, Dr. Ledbetter built one of the largest academic translational and diagnostic genomics programs in the country, the Emory Genetics Laboratories.

Along with Dr. Christa Martin, Dr. Ledbetter founded the International Standards for Cytogenomic Arrays (ISCA) Consortium and serves as Chair of the Steering Committee. The ISCA consortium has developed data standards for CNV and phenotypic data associated with whole genome cytogenomic array, evidence-based processes for determining functional and clinical significance of CNVs, and a publicly available database representing over 25,000 patient samples at dbVar and dbGaP at NCBI/NIH. His CNV Atlas of Human Development, funded by a Grand Opportunity (GO) grant from NICHD, is a model for the routine submission of whole exome or whole genome sequence data from routine diagnostic laboratories into central, public databases for knowledge generation. Dr. Ledbetter's interest is to apply the lessons learned from whole genome CNV analysis to developing methods for determining the functional and clinical significance of all sequence and structural variation in the human genome for improving patient care.

As joint principal investigator, Dr. Ledbetter will jointly oversee the entire project with Drs. Martin, Mitchell, Nussbaum and Rehm, and will co-chair the Policies, Standards, and Sustainability Workgroup. He will work with all of the workgroups to ensure consistency of efforts across all projects. He will meet regularly with the other members of the Executive Committee to monitor progress, set goals and organize the work of this project.

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 meet regularly with the other members of the Executive Committee to monitor progress, set goals and organize the work of this project.

Joyce A. Mitchell, PhD, FACMI, FACMG, is Professor and Department Chair of Biomedical Informatics at the University of Utah School of Medicine. She obtained her PhD in Population Genetics from the University of Wisconsin with postdoctoral training in clinical genetics at UCSF. She is certified as a PhD Medical Geneticist by the American Board of Medical Genetics and is a Founding Fellow of the ACMG. Dr. Mitchell[1]s postdoctoral training was also in Medical Informatics Sciences and she was elected to be a Fellow in the American College of Medical Informatics (ACMI). Dr. Mitchell spent 25 years on the faculty of the University of Missouri School of Medicine in two departments: Child Health (Section on Medical Genetics) and Health Management and Informatics (Division Leader of Health Informatics). Administratively, she served as the Director of the Medical Informatics Group, the Associate Dean for Information Technology for the School of Medicine, and the Chief Information Officer for University of Missouri Health Care, which included ushering the system through the Year 2000 and installing the beginnings of an Electronic Medical Record. She spent a sabbatical year at the National Library of Medicine in 2001-02 and developed the Genetics Home Reference <http://ghr.nlm.nih.gov/> to bridge the genomics research results with consumer health interests in genetic diseases. She served as the Senior Scientific Advisor for the Genetics Home Reference until 2009.

In 2005, Dr. Mitchell was recruited by the University of Utah School of Medicine to serve as Department Chair for Biomedical Informatics. In 2007 she was appointed as Associate Vice President for Health Sciences IT, where she coordinates and directs the information technology resources for the academic mission across four schools and multiple research institutes, coordinating with the IT unit of the University of Utah Health System for clinical data. She is also the director of the Biomedical Informatics Core for the Center for Clinical and Translational Sciences (CCTS).

Dr. Mitchell was elected to serve as President of the American College of Medical Informatics from 2008-2010, and will serve as immediate past-president until 2012. She is currently serving a four-year term on the Board of Regents of the National Library of Medicine (NLM), and will serve as Chair of the NLM's Board of Regents in 2012-2013. She also serves on the Council of Councils for the National Institute of Health (NIH) from 2012-2015. Dr. Mitchell serves as co-director of the NLM-sponsored course in Biomedical Informatics held annually at the Marine Biological Laboratory in Woods Hole, Massachusetts.

As joint principal investigator, Dr. Mitchell will co-Chair the Bioinformatics and IT Coordinating Workgroup with Sandy Aronson, Executive Director of IT of Partners Healthcare Center for Personalized Medicine. As Co-Chair, she will set project priorities, supervise personnel, and monitor the overall progress of the IT components of the project and how IT interacts with all other workgroups to ensure consistency with respect to bioinformatics approaches and IT best practices. She will meet regularly with the other members of the Executive Committee to monitor progress, set goals and organize the work of this project.

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 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 to patient care for the past six years.

Dr. Nussbaum is a Past President of the American Society of Human Genetics, and has served as a Director of the American College of Medical Genetics and Genomics and the American Board of Medical Genetics. He is a member of the Institute of Medicine, where he serves on the Genomic Medicine Working Group.

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. Dr. Nussbaum will Chair the Executive Committee made up of the five PIs on this project. He will organize and set the agenda for regular meetings and teleconferences.

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 joint principal investigator for this project, Dr. Rehm will oversee all aspects of the sequence-level variation projects. She 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. She will meet regularly with the other members of the Executive Committee to monitor progress, set goals and organize the work of this project.

Resolution of Disputes

Drs. Ledbetter, Martin, Mitchell, Nussbaum and Rehm, are in complete agreement as to the overall project goals, how the project should be organized and administered, and the methods to achieve these goals. They have been meeting weekly over the past few months to plan the revised proposal and have had no difficulties reaching unanimity on all topics under discussion. 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 five 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 five PIs, the members of the Executive Committee will vote and abide by a majority of 3 votes or more. 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 with Dr. Nussbaum taking the lead in that engagement process.

Responsibility for regulatory compliance and human subjects protection

All five 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 obtaining 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**, the **University of Chicago** and the **University of Utah** 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, 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 **Dr. David Ledbetter**, joint principal investigator of this proposal and have as a co-investigator **Mr. W. Andrew Faucett**. Geisinger Clinic will direct and coordinate the activities of the Education, Engagement and Ethics Workgroup. 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. He will also oversee the development of the patient registry with Patient Crossroads.

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 co-chair the Structural Variant Workgroup.

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 co-direct the Structural Variant workgroup.

MutaDATABASE Foundation
Dr. Patrick Willems will serve as the consortium PI for the MutaDATABASE Foundation. He and his team will organize data collection from laboratories who wish to submit data into MutaDATABASE.

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 developmental delay.

University of Utah
Dr. Joyce A. Mitchell, joint principal investigator, will serve as the consortium PI at the University of Utah. Mitchell will jointly oversee the entire project with Drs. Rehm, Martin, Nussbaum and Ledbetter. She will jointly oversee an IT coordinating team that will work closely with NCBI staff as well as the bioinformatics and IT staff across all of the funded sites and labs submitting data. She will also interact with all of the workgroups to ensure consistency across all projects with respect to bioinformatics approaches and IT standards and will meet regularly with the executive committee. Dr. Mitchell will work with **Dr. Karen Eilbeck**, co-investigator, who will manage the ontology development at the boundary of the SO and IDPO to ensure that the new terms and relationships generated are compliant to both ontologies.

Foreign Work
MutaDATABASE is a resource that is a foreign component to this domestic U41 application. The MutaDATABASE Foundation has supported the centralization of variant data, a mission consistent with the core mission of this proposal. The MutaDATABASE Foundation has recruited the participation of many laboratories and gene curators and therefore we have committed resources to enable Dr. Willems to facilitate interaction with outside groups who may preferentially choose to deposit data into MutaDATABASE. If this database becomes a common and preferred site for data deposition, we will support the development of a robust interface between ClinVar and MutaDATABASE in Year 2.

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), ClinVar 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 contribute specifications for software development 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 two e-mails 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' letters confirm 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 and permission for a July 3rd receipt date for the May 25th submission cycle.

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
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

Organizations, projects and consortia who will collaborate on the project

Russ Altman, M.D., Ph.D.	Principal Investigator, PharmGKB President, American Society for Clinical Pharmacology
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
Mary Claire King, Ph.D.	President, American Society of Human Genetics Professor,
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
Dan Roden, M.D.	Chair, Pharmacogenetics Research Network Professor of Medicine and Pharmacology, Vanderbilt School of Medicine
Patrick J. Willems, M.D., PhD.	Principal Investigator, The MutaDATABASE Foundation

Consultants

Jennings Aske, J.D.	Chief Information Security Officer Partners Healthcare, Inc.
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
Professor John Christodoulou, AM, M.B.B.S., Ph.D.	Head, Centre for Rett Syndrome Research, Professor, University of Sydney, AUSTRALIA
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
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
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
Professor Finlay Macrae, M.D.	Honorary Secretary, International Society for Gastrointestinal Tumours
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.
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)
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

APPENDIX

ISCA Data Submission Template - p.1

METHODS:

	For submitter to complete	Description
Submitting_Lab		Name of lab submitting this data
Submitting_person		Name of person submitting
Submitter contact (email)		(For NCBI internal use only)
Release_submitted_lab		(Yes or no) Should the submitting lab origin of these variants be exposed to users
Array_Manufactuer		Company manufacturing the array
Array_Name		Name of the array
Description of the array		Text description the design of the array
Resolution_backbone		What is the resolution, defining lower size threshold and breakpoint resolution, across the genome
Resolution_targeted		If there are targeted regions, what is the resolution in those regions
Method_description		Brief text description of aCGH method
Reference_DNA		Description of the how the reference DNA was selected (e.g. pooled, how many samples, ect)
GEO_array_ID		If the array design has been submitted to GEO, the GEO ID
Analysis Software		What analysis software was used to call variants
Analysis_methods_and_parameters		Description of algorithm and parameters used when running the variant calling software.

NOTE: If the submitter has data from more than 1 platform, thus requiring two or more "methods", please submit each as a different excel template submissions.

ISCA Data Submission Template - p.2

VARIANTS:

Please use one entry for each observed variant, a sample with more than one reported variants will thus include multiple entries with the same sample ID.

Submission_batch_ID
Local_unique_anonymous_sample_ID
Array_name
Array_version
Genome Build
Chr
Start (maximum coordinates)
Start (minimum coordinates)
Stop (minimum coordinates)
Stop (maximum coordinates)
Type (loss/gain)
Copy_number_dosage_count
Interpretation
Inheritance
Validation
Comment
Phenotype_HPO
Phenotype_Other
Gender
Age_in_months_at_time_of test referral
If_inherited_pheno_of_parent_of_origin
Family_History
Consent
Allow_recontact

ISCA Data Submission Template - p.3

SAMPLES WITHOUT VARIANTS:

This table lists IDs for additional samples for which probe level data is available, but for which no "Variant" calls were reported.

Local_unique_anonymous_sample_ID
Array_name
Array_version
Comment
Phenotype_HPO
Phenotype_Other
Gender
Age_in_months_at_time_of test referral
Family_History
Consent
Allow_recontact

ISCA Data Submission Template - p.4

Data dictionary defining fields in the "Variants" tab

Field	Description	Example:
Submission_batch_ID	The is a unique ID provided by the user to track the submission within this file a unit. This id should include the lab 3-letter prefix, and mayb include the date.	ABCD_batch5_4_26_2010
Local_unique_anonymous_sample_ID	Sample ID which will be unique in terms of all the variants submitted from a given lab in this and future submission batches. This will include a submitting lab code prefix such as "EGL" which may be an abbreviation for the lab, or may be an random ID assigned to the lab such as "ABC". Please contact paschalj@ncbi.nlm.nih.gov for details and to be assigned a prefix. NOTE: this ID must already be anonymized such that it is linked to the origianl test request only by a key held by the submitter.	ABCD000232
Array_name	Name of Array on which this daa was generated	ISCA 44K
Array_version	Version of the Array	2.3
Genome Build	(e.g. NCBI build 36) note: NCBI build 36= UCSC HG18	36
Chr	Chromosome	1
Start (maximum coordinates)	Start position (this should be the "maximal" span start location when interpolating between probes)	199925242

ISCA Data Submission Template - p.5

Data dictionary (cont.)		
Start (minimum coordinates)	Start position (this should be the "minimal" span start location when interpolating between probes)	200000242
Stop (minimum coordinates)	Stop position (this should be the "minimal" span stop location when interpolating between probes)	210234245
Stop (maximum coordinates)	Stop position (this should be the "maximal" span stop location when interpolating between probes)	210309245
Type	"loss" or "gain"	loss
Copy_number_dosage_count	Count of copy number (0 - homozygous loss, 1 - heterozygous loss, 3,4 - one or two additional copies/gain)	0
Interpretation	"pCNC", "bCNC","uncertain", "likely bCNC", "likely pCNC" (pCNC=interpreted pathogenic,, bCNC=interpreted benign, uncertain=uncertain interpretation; likely bCNC=uncertain but most likely benign; likely pCNC=uncertain but most likely pathogenic)	pCNC Note: Losses and gains as part of unbalanced translocations should be interpreted individually. To indicate that the loss and gain are part of an unbalanced
Inheritance	"Biparental", "De novo", "Maternal", "Paternal", "not_tested"	Paternal
Validation	Validation evidence, ("unknown", "RT-PCR", "FISH","Oligo_aCGH","SNP_aCGH","MLPA","Karyotype", "BAC_aCGH"), multiple methods separate by semi-colon	RT-PCR;FISH

ISCA Data Submission Template - p.6

Data dictionary (cont.)		
Comment	Free-text comment on sample or mechanism of variant	unbalanced translocation
Phenotype_HPO	Referring Diagnosis using Human Phenotype Ontology terms. In this field, use only terms which terms which appear on the ISCA phenotype collection form (available at https://www.iscaconsortium.org/) or additional HPO terms at http://www.human-phenotype-ontology.org/Phenomizer/Phenomizer.html. You may include the English terms, the HPO numbers or both as indicated in the example.	Gross motor delay (HPO:0002194);Cleft Palate(HPO:0002744) OR Gross motor delay;Cleft Palate; OR HPO:0002194;HPO:0002744
Phenotype_Other	Referring Diagnosis in free text form. Please contact paschalj@ncbi.nlm.nih.gov for details regarding collection forms and standard vocabulary for this information.	Developmental Delay;Cleft Palate
Gender	M/F	M

ISCA Data Submission Template - p.7

Data dictionary (cont.)		
Age_in_months_at_time_of test referral	This value can be calculated by the submitter using the test submission date and the birth data of subject using the Excel function DATEDIF with the "M" parameter as shown in Column F of this row. Note: this formula can be used be the submitter to do the age_in_months calculations, and then provide ONLY Age_in_months in the submission, the submission should NOT contain the actual birth date or test date.	4/11/2009
IS_parent_of_origin_affected	If inherited, does the parent from which variant is inherited have an affected phenotype (YES, NO)	UNK
Family_History	Free-text comment about Family History	Sister had similar phenotype
Consent	"legacy" or "opt-out"	opt-out
Allow_recontact	Can re-contact of patient for follow up allowed through proper ISCA process? (Yes/No)	Yes

ClinVar Data Submission Form for Variants - p.1

Variant ID	REQUIRED	Assertion ID	1
	REQUIRED	Variant Name	NM_000053.2:c2350A>G
Location and Variation	REQUIRED (either genomic or transcript location)	Genome Build	NCBI36
		chromosome accession	NC_0000013
		genomic chromosome position start	51430498
		genomic chromosome position stop	51430498
		HGVS-genomic	
		HGVS-transcript	NM_000053.2:c.2350A>G
	RECOMMENDED	HGVS-protein	NP_000044:p.Met769Val
	OPTIONAL	Type of Variation	SNV (substitution); insertion; deletion; duplication; indel; multi-nucleotide
	OPTIONAL	dbSNP rs	rs#
	OPTIONAL	Exon number	
	OPTIONAL	Intron Number	
	OPTIONAL	Exon/Intron coordinate type	RefSeq/LRG; Transcript; Historical
	OPTIONAL	Variant comment	
Variant Consequence and Context	OPTIONAL	Molecular Consequence (direct coding consequence) - comma delimited list	frameshift; missense; nonsense; synonymous; in frame; splicing
	OPTIONAL	Molecular Consequence Comment	
	OPTIONAL	Functional Consequence (comma delimited list)	Loss of function; gain of function; overexpression; under expression; splice_site_lost; splice_site_gained; NMD
	OPTIONAL	Means of determining functional consequence	Predicted; Experimental
	OPTIONAL	Functional Consequence Comment	
	OPTIONAL	Molecular Context relative to gene	upstream; downstream; 5'UTR; 3'UTR; cds; intron
	OPTIONAL	is in duplicated region / is there a pseudogene	
	OPTIONAL	Comment on adequacy of method to resolve duplicated regions	
	OPTIONAL	distance from exon/ intron junction	
	OPTIONAL	protein domain(s)	
Gene	OPTIONAL	+/- strand	
	RECOMMENDED	Gene sym. HUGO	
	OPTIONAL	Entrez gene ID	
	OPTIONAL	OMIM gene ID	
	OPTIONAL	Gene Comment	
	OPTIONAL	Total number exons	
Clinical Assertion	RECOMMENDED	Mode of Inheritance	AR; AD; XLD; XLR, YL
	APPLIED AUTOMATICALLY	Review Status of clinical significance (only authorized submitters can set reviewed categories)	uncurated (when not provided in next column); curated (by single submitter); expert curation (panel); practice guidelines
	REQUIRED	5-tier clinical significance (relative to the phenotype), + values for pharma tests	not provided; pathogenic; likely pathogenic; uncertain; likely benign; benign; responsive; not responsive, etc.
	RECOMMENDED	Date of last evaluation of clinical significance (default is submission date)	
	OPTIONAL	Submitter local clinical metric (where submitter has alternate system/nomenclature for clinical significance	
	OPTIONAL	Comment on Clinical Significance	
Phenotype	REQUIRED	Disease/Phenotype for which clinical significance was evaluated	(Diagnosis)
	OPTIONAL	Is asserted to be benign for all highly penetrant Mendelian diseases	yes/no
	REQUIRED	Type of phenotype (Disease; Disease Risk; Drug Response; Phenotype Trait; Finding)	
	REQUIRED IF	Disease scope of test (required if no Disease provided)	Test name/ test indication
	OPTIONAL	Severity (attribute of this allele relative to the disorder)	
	OPTIONAL	Age of onset (attribute of condition itself)	
	OPTIONAL	Penetrance (attribute of condition itself)	
	OPTIONAL	Alt. Phenotype Name(s)	
	OPTIONAL	Phenotype Abbrev.	

ClinVar Data Submission Form for Cases - p.2

Identifiers	REQUIRED	Anony. Subject ID	
	OPTIONAL	Anony. Pedigree ID	
	OPTIONAL	Anony. Mother ID	
	OPTIONAL	Anony. Father ID	
Sample Information	OPTIONAL	Is Proband?	yes/no
	OPTIONAL	Sex	M / F
	OPTIONAL	Age	
	OPTIONAL	Ethnicity/race	
	OPTIONAL	Country of origin	
	OPTIONAL	DNA source tissue	
Disease/Phenotype Information	REQUIRED	Disease/Phenotype or Healthy/Control	
	OPTIONAL	Type of phenotype	Disease; Disease Risk; Drug Response; Phenotype Trait; Finding;
	REQUIRED IF	Disease Scope of test (required if no Disease/Phenotype	Testname/test indication
	OPTIONAL	Reason for testing	Presymptomatic, carrier testing, diagnosis, population screening, risk assessment
	OPTIONAL	Other phenotypes present in subject	
	OPTIONAL	Tested indiv. has positive family history	yes; no;
	OPTIONAL	Reason for Testing (Free Text, indication)	
	REQUIRED	Affected (yes/no)	
	OPTIONAL	Considered Healthy (yes/no)	
Result	OPTIONAL	Overall Test Result	positive; negative
Method	OPTIONAL	Method ID(s)	external reference to more detailed method info
	REQUIRED	Primary method type	Next-gen; Sanger; genotyping-sequenom;
	OPTIONAL	Quality score for variant call from primary method (defined in method tab)	
	REQUIRED	Validation status(s)	sanger_confirmed; confirmed_by_genotyping;
	OPTIONAL	Quality score for variant call from validation method (defined in method tab)	Phred score
	OPTIONAL	Comment	
	OPTIONAL	publication	
Variant Genotype Data	REQUIRED (may be same as genomic or transcript)	Var 1 name	NM_000053.2:c2350A>G
	REQUIRED (either genomic	Var 1 genomic name	
		Var 1 transcript name	
	Recommended	Var 1 protein var name	
	REQUIRED	Var 1 geno state	heterozygous; homozygous; hemizygous; homoplasmic; heteroplasmic
	OPTIONAL	Var 1 molecular consequence	frameshift; missense; nonsense; synonymous; in frame; splicing
	OPTIONAL	Var 1 functional consequence	Loss of function; gain of function; overexpression; under expression; splice_site_lost; splice_site_gained;
	Recommended	Var1 5-tier clinical significance (relative to the phenotype)	not provided; pathogenic; likely pathogenic; uncertain; likely benign; benign;
	Recommended	Date of last evaluation of clinical significance (default is submission date)	
	OPTIONAL	Var1 suspected mode of inheritance	AR; AD; XLD; XLR, YL
	OPTIONAL	Var1 allele origin	de novo / paternal /maternal
	OPTIONAL	Var 1 dbSNP rs	rs#
	OPTIONAL	Var1 dbVar sv	
	Recommended	Var 1 Gene Symbol	
	OPTIONAL	Entrez gene ID	
	OPTIONAL	OMIM gene ID	
	OPTIONAL	Var 1 is_compund_het_with_other_var	list name of other variant
	OPTIONAL	Var 1 is_epistatic_with_other_var	list name of other variant
	OPTIONAL	var 1 count_geno+pheno+ concordant family members	list name of other variant
	OPTIONAL	Count of other family members with var1 who are affected (geno+pheno+)	
	OPTIONAL	Count of other family members with var1 who are not affected (geno+pheno-) [exclude carriers]	
		var 2 … repeating co-occurrence sets	

ClinVar Data Submission Form for Aggregated Data from Affected Probands - p.3

		Variant name
SampleSet Information (Cases)		Observation Set Type
		Pubmed_ID
		Ethnicity
		Are genotyped subjects known to be Independent?[Probands only]
		Age range (X-Y)
		Proband/Pedigree public ID(s)
		Number affected subjects tested
		Number of affected males genotyped
		Number of affected females genotyped
		Number chromosomes tested from affected subjects
Same as total subjects and chromosomes if only probands are being reported		Number Independent Affected subjects/families tested
		Number Chromosomes Tested in Independent Affected Subjects
Genotype and Allele Counts		Number Variant Alleles observed in Affected
		Number affected subjects with variant
		Number affected subjects who are homozygous variant [includes hemi-zygous]
		Total Number Affected subjects who are heterozygotes (total)
		Within the hets, for recessive MOI, the number of confirmed compound heterozygotes (compound with a different known or suspected pathogenic variant)
		Within the hets, for recessive MOI, the number of affected hets where another suspected pathogenic other variant in this gene could **not** be identified
		Number de novo variants observed in affected subjects
Same as total subjects and alleles if only probands are being reported		Number of Independent Families with one or more affected subject with the variant genotype
		Number of observed alleles from indendent affected subjects
Other Variant		number affected subjects with this variant who have another variant thought to be responsible for phenotype
Transmission and Segregation of Variant and Phenotype	geno+pheno+	Number instances observed of heterozygous parent transmitting the variant allele to affected child
	geno-pheno+	Number Instances of heterozygous parent transmitting the normal allele to affected child, rather than the variant allele (non-segregants)
	geno+pheno-	Number unaffected subjects with variant allele(s), thus (in)consistent with Mode of Inheritance, or non-penetrant
		Number independent families seen to cosegregate the variant allele and affected phenotype among two or more family members
		# Informative Meioses

ClinVar Data Submission Form for Aggregated Data from Controls - p.4

	Assertion ID
	Var. Name
Methods	Study Type
	Study Source
	Method ID
Sample Set Information (Controls)	Descr.
	Are control samples explicitly assessed as unaffected for target phenotype?
	Are genotyped subjects known to be Independent?
	max age
	min age
	tissue
	ethnicity
	country
	Number of Chromosomes tested
	Total number genotyped subjects
	Number genotyped males
	Number genotyped females
	Number of families
	public pedigree ID
	Citation
	URL
	Comment
Genotype and Allele Counts in Controls	# ref homozy.
	# het
	#var. homozy.
	flag if homoz.
	# variant alleles observed
	variant allele frequency

ClinVar Supplemental Data for Methods and Evidence - p.5

Detailed Methods	Method ID
	Method Description
	Platform Type
	Sequencer Platform Name
	Sequence Capture / Multiplexing Method description
	Coverage threshold after removal of duplicates (next gen)
	% variant reads threshold needed to call variant (next gen)
	Mapping/Alignement software
	Mapping/Alignment software version
	Variant Calling software
	Variant Calling software version
	Validation Platform Type
	Validation Platform Name
	is confirmed by independent technologies?

Computed Functional Prediction	Assertion ID
	Var. Name
	Software Name
	Software Version
	Citation
	Parameter Description
	Value Name
	Value

Experimental Functional Evidence	Assertion ID
	Var. Name
	Method Description
	Findings
	Number of Observations
	Final Result
	Citation
	Comments

ClinVar Data Dictionary

Overview

This document defines the data elements represented in the ClinVar database. The document includes descriptions of how data are managed, the XML used to represent each concept (see ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/), the field name in the spreadsheet version of the submission document, the table and column in which the data are stored in the relational database, and allowed values.

Status of this document

Working Draft. Please direct any comments to clinvar@ncbi.nlm.nih.gov

General processing

Data added computationally

Not all values included in this document are expected to be supplied from a submitter; some will be added based on information in NCBI's databases. These values are marked explicitly as '**from NCBI**'

Optional and required values

Some elements are hierarchical in representation. If a major category topic is optional, all data elements in that category are optional. But if an optional category is selected, then the data elements listed as required are required for that category.

Validation of submissions

Processing of submissions to enforce the rules presented in this document may not all be managed via the xsd provided from our ftp site (ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/). Some validation will be provided at the database level, by comparison to standard terminologies or known relationships among variants, genes, and phenotypes, or by validating reported alleles against the stated reference sequence. Documentation of the authorities used by ClinVar are provided from the Nomenclature/Authorities link on ClinVar's home page. (http://www.ncbi.nlm.nih.gov/clinvar/nomenclature/)

Data representations used in multiple contexts

Source/Status

Most elements in the database are characterized with respect to the submitter, identifiers used by the submitter, date submitted, date modified, status of the record (*e.g.* current/to be

deleted/secondary to another record), review status, and whether the data should be public or private. Rather than repeating these elements for each data category defined below, the word **Source/Status** will be used as a pointer to the Data source/Status section, where the source and status elements are defined.

AttributeSet

Many concepts in the database are represented by what ClinVar terms an attributeSet, which is an open-ended structure providing the equivalent of a type of information, the value(s) for that data type, submitter(s), free text comment(s) describing that attribute, identifier(s) for that attribute, and citation(s) related to that attribute. Rather than repeating this description per attribute, the word **AttributeSet** will be used to indicate that the data are stored using this data structure, with the attribute types expected for that database concept. Thus, by definition, an **AttributeSet** includes **Source/Status**.

Method

Note: Methods used by a submitter to capture clinical data shall be shared between ClinVar and the Genetic Testing Registry (GTR).

This section summarizes the elements that describe an assay method. We recognize that not all elements apply to the each of the approaches enumerated in the **Evidence** section. More than one method may be submitted per **Observation.** An example would be one method for primary data collection and another for validation. A set of methods should be submitted as it applies to any set of observations.

Note: When data are submitted based on curatorial review, please provide the following:

- purpose = 'Curation'
- SourceType='data review'

Note: The Methods tab of the spreadsheet requests a Method ID. This is used only to connect observations (Obs-Affected, Obs-controls, PerTestData) with the method(s) used, and will not be retained as a data element. It must be unique for any one submission.

Description

- optional
- free text describing the method. Submitters are encouraged to use specific data elements whenever possible.
  SPREADSHEET: Methods.Method Description

SPREADSHEET: Methods.Sequence Capture / Multiplexing Method description
XML: Method.Description
db: GTR.clinvar.method.description

Type of platform

- optional
- type of platform used for data capture. Examples include high throughput sequencing, microarray, *etc.*
- SPREADSHEET: Methods.Platform Type
  XML: Method.TypePlatform
  db: GTR.clinvar.method.platform_type

Platform name

- optional
- name of the platform used for data capture. Examples include HiSeq, MiSeq
  SPREADSHEET: Methods.Platform Name
  XML: Method.NamePlatform
  db: GTR.clinvar.method.platform

Confidence measures

- optional
- A structure to support representation of name/value pairs describing how the confidence in the method was assessed. These measures may describe the confidence in the technology, the confidence in identifying/calling a variation, and/or the confidence in any conclusion drawn from data analysis. Operationally, the methods used to assess confidence are themselves methods. The following types of information may be in scope:
  - coverage after removal of duplicates
  - unaligned reads
  - quality score of base call or variation type (*e.g.* are there co-occurring independent single substitutions or one multi-nucleotide variation?)
  - quality score of mapping
  - was the variant detected on both strands?

  SPREADSHEET: Methods. – Confidence measures can be user defined
  SPREADSHEET: Methods.Coverage threshold after removal of duplicates

SPREADSHEET:	Methods.% variant reads threshold needed to call variant
SPREADSHEET:	Methods.Quality score
XML:	Method.ConfidenceMeasure
db:	GTR.clinvar.obs_method_attr (attributeSet)

Purpose

- required
- A selection from a controlled list to category the primary intent of the method used to generate the data provided in the submission. Some combinations of method type and purpose will not be allowed, e.g. case-control and review.
- current options:

  review: to review primary or review data from other submitters

  primary assay: for experimental determination of results

  validation: a special type of primary assay used to confirm results

SPREADSHEET:	Methods.Purpose
XML:	Method.Purpose
db:	GTR.clinvar.method.method_purpose

Type of result

- optional
- options: variation call, number of occurrences, p value, odd ratio

SPREADSHEET:	Method.Result type
XML:	Method.ResultType
db:	GTR.clinvar.method.result_type

Minimum value reported

- optional
- For this method, what is the minimum value that can be reported?

SPREADSHEET:	Methods.Minimum value reported
XML:	Method.MinReported

db: GTR.clinvar.method.min_reported

Maximum value reported

- optional
- For this method, what is the maximum value that can be reported?

  SPREADSHEET: Methods.Maximum value reported
  XML: Method.MaxReported
  db: GTR.clinvar.method.max_reported

Software

- optional
- Description of key software, with explicit representation of name and version.

Name

- required
- Name of the software

  SPREADSHEET: Methods.Mapping/Alignment software
  SPREADSHEET: Methods.Variant Calling software
  XML: Method.Software.name
  db: GTR.clinvar.method_attr.attr_char (AttributeSet)

Version

- optional
- Version of the software

  SPREADSHEET: Methods.Mapping/Alignment software version
  SPREADSHEET: Methods.Variant Calling software version
  XML: Method.Software.version
  db: GTR.clinvar.method_attr.attr_char2

Purpose

- optional
- purpose of this software. Examples include variant calling, prediction functional consequence

  SPREADSHEET: Methods.Purpose
  XML: Method.Software.purpose

db: GTR.clinvar.method_attr.attr_char where attr_type is 'software_purpose'

Citations

- optional
- Citation(s) documenting this method.
  SPREADSHEET: Methods.Citations
  XML: Method.Citation
  db: GTR.dbo.citation; GTR.dbo.citation_many

Reference standard

- optional
- For sequence-based tests, what sequence or assembly was used as the reference standard (*e.g.* GRCh37).
- NOTE: this is an example of data that will be validated at the database level and not explicitly in the xsd.
  SPREADSHEET: Methods.Reference Standard
  XML: Method.ReferenceStandard
  db: GTR.clinvar.ref_std

Type

- required
- Type of approach in this method. Examples include clinical testing, evaluation of a reference population, case-control, curation, *in vivo*, *in vitro,* functional assay*, in silico*
  SPREADSHEET: Methods.Method type
  XML: Method.Type
  db: GTR.clinvar.method.method_type

Result

- optional
- The conclusion reached by this method for this observation. An example is a method used to validate the variant call; the MethodResult would be pass/fail/inconclusive

- SPREADSHEET:
  XML: Method.MethodResut
  db: GTR.clinvar.obs.method_attr where attr_type is method_result

GTR_test_id

- optional
  Used as a unique identifier for a test registered in the Genetic Testing Registry
  SPREADSHEET: GTR Test ID
  XML: Method.XRef
  db: GTR.clinvar.method.extrn_id/ GTR.clinvar.method.extrn_src

Sample (required)

This section is used to describe what was studied to generate the submission.

species

- required (will default to human if not supplied)
  SPREADSHEET: ExpFuncEvidence.Species
  XML: ObservedIn.Sample.Species
  db: clinvar.sample.txid

cell line

- optional
- Name of the cell line
  SPREADSHEET: ExpFuncEvidence.cell line
  XML: ObservedIn.Sample.CellLine
  db: clinvar.sample.cell_line

Strain/breed

- optional
- Name of the strain or breed that was analyzed in this observation

  SPREADSHEET: ExpFuncEvidence.strain/breed

  XML: ObservedIn.Sample.Strain

  db: clinvar.sample.strain

origin

- required
- The genetic origin of the variation being submitted.
- possible values: germline, somatic, uncertain, not determined, de novo

  SPREADSHEET: Obs-affected(SampleSet).DNA Origin
  XML: ObservedIn.Sample.origin
  db: clinvar.sample.origin

age ranges

- optional
- The range of ages included in this sample. If age range is an important variable in your submission, with different observations based on the age, please submit each observation separately, rather than lumping into one summary observation with one sample description.

  SPREADSHEET: Obs-affected(SampleSet).Age range (min-max in years)
  SPREADSHEET: Obs-controls(SampleSet).Age range (min-max in years)
  XML: ObservedIn.Sample.Age
  db: clinvar.sample.min_age
  db: clinvar.sample.max_age
  db: clinvar.sample.age_units

Geographic origin

- optional. If multiple, provided as semi-colon delimited. Can be used to indicate country, continent, region

  SPREADSHEET: Obs-affected(SampleSet)Geographic origin
  SPREADSHEET: Obs-controls(SampleSet)Geographic origin
  XML: ObservedIn.Sample.GeographicOrigin
  db: GTR.clinvar.sample.geographic_origin

ethnicity

- optional
- Name or description of the ethnicities included in this sample.

  SPREADSHEET: Obs-affected(SampleSet).Ethnicity
  SPREADSHEET: Obs-controls(SampleSet).Ethnicity

XML: ObservedIn.Sample.Ethnicity
db: clinvar.sample.ethnicity

Study name

- optional
- Public name of a study submitting these data and providing the sample. Can be used to indicate the name of a study population or cohort. Example. Framingham
  SPREADSHEET: Obs-affected(SampleSet).Study name
  SPREADSHEET: Obs-controls(SampleSet).Study or cohort name
  XML: ObservedIn.Sample.StudyName
  db: GTR.clinvar.obs_method_attr where attr_type = 17

tissue

- optional
- Name or description of the tissue that was assayed. Highly recommended if the origin is somatic or if an experimental analysis
  SPREADSHEET: Obs-Affected(SampleSet).DNA source tissue
  SPREADSHEET: PerTestData(Sample).DNA source tissue
  XML: ObservedIn.Sample.Tissue
  db: clinvar.sample.tissue

fraction of sample which is tumor-containing

- optional
- Applicable only if origin is somatic
- Free text description of the fraction of the sample
  SPREADSHEET: Obs-Affected(SampleSet).Fraction of sample which is not tumor
  XML: ObservedIn.Sample.FractionTumor
  db: clinvar.sample.fraction_tumor

affected status

- required
- Indicate whether this sample had the condition/observed phenotypes
  SPREADSHEET: *DEFINED BY TAB* (Obs-Affected vs Obs-controls)
  XML: ObservedIn.Sample.AffectedStatus

db: clinvar.sample.affected_status

Number of chromosomes tested

- optional, but highly recommended

  SPREADSHEET: Obs-affected(SampleSet).Number chromosomes tested
  XML: ObservedIn.Sample.NumberChrTested
  db: clinvar.sample.chr_tested

Number of subjects of unspecified sex.

- optional
- NOTE: this is not represented explicitly in the XML because calculation of subjects with sex unspecified will be determined by the difference between the number of individuals tested, and the sum of number of males and females.
- Number affected subjects genotyped of unspecified sex

  SPREADSHEET: Obs-affected(SampleSet).
  XML: ObservedIn.Sample.NumberTested
  ObservedIn.Sample.NumberMales
  ObservedIn.Sample.NumberFemales
  ObservedIn.Sample.AffectedStatus
  db: clinvar.sample.individuals_tested
  db: clinvar.sample.males
  db: clinvar.sample.females
  db: clinvar.sample.affected_status

Number of affected males genotyped

- optional
- Number of affected males genotyped in the observation set. This concept is represented in the database by a combination discrete data elements.

  SPREADSHEET: Obs-affected(SampleSet).Number of affected males genotyped
  XML: ObservedIn.Sample.NumberMales
  db: clinvar.sample.males
  db: clinvar.sample.affected_status

Number of affected females genotyped

- optional
- Number of affected males genotyped in the observation set. This concept is represented in the database by a combination discrete data elements
  SPREADSHEET: Obs-affected(SampleSet).Number of affected females genotyped
  XML: ObservedIn.Sample.NumberFemales
  db: clinvar.sample.females
  db: clinvar.sample.affected_status

Number of unaffected subjects

- optional
- The number of unaffected subjects on which this submission is a based
  SPREADSHEET: Obs-controls(SampleSet).Total number genotyped subjects
  XML: ObservedIn.Sample.NumberTested
  ObservedIn.Sample.AffectedStatus
  db: clinvar.sample.individuals_tested
  db: clinvar.sample.affected_status

Number of unaffected males genotyped

- optional
- The number of unaffected males on which this submission is based
  SPREADSHEET:SampleSet(obs-controls)/ Number of males genotyped
  XML: ObservedIn.Sample.NumberMales
  db: clinvar.sample.males
  db: clinvar.sample.affected_status

Number of unaffected females genotyped

- optional
  SPREADSHEET:SampleSet(obs-controls)/ Number genotyped females
  XML: ObservedIn.NumberFemales
  db: clinvar.sample.females
  db: clinvar.sample.affected_status

Number unaffected subjects genotyped of unspecified sex

- optional
- The number of unaffected males on which this submission is a based
  SPREADSHEET:SampleSet(obs-controls)/ NOT REPRESENTED
  XML: ObservedIn.Sample.NumberTested -
  ObservedIn.Sample.NumberMales –
  ObservedIn.Sample.NumberFemales
  ObservedIn.Sample.AffectedStatus
  db: clinvar.sample.individuals_tested -
  db: clinvar.sample.males -
  db: clinvar.sample.females
  db: clinvar.sample.affected_status

- Number of males genotyped with unspecified affectedstatus
  SPREADSHEET: SampleSet(obs-controls)/ NOT REPRESENTED
  XML: ObservedIn.Sample.NumberMales
  ObservedIn.Sample.AffectedStatus
  db: clinvar.sample.males
  db: clinvar.sample.affected_status

- Number females genotyped with unspecified affected status
  SPREADSHEET:SampleSet/NOT REPRESENTED
  XML: ObservedIn.Sample.NumberFemales
  ObservedIn.Sample.AffectedStatus
  db: clinvar.sample.females
  db: clinvar.sample.affected_status

- Number subjects of unspecified sex genotyped with unspecified affected status
  SPREADSHEET:SampleSet/NOT REPRESENTED
  XML: ObservedIn.Sample.NumberTested - ObservedIn.Sample.NumberMales –
  ObservedIn.Sample.NumberFemales
  ObservedIn.Sample.AffectedStatus
  db: clinvar.sample.individuals_tested –
  db: clinvar.sample.males –
  db: clinvar.sample.females
  db: clinvar.sample.affected_status

- Number of Affected families (optional)
  SPREADSHEET: SampleSet(Obs-Affected).Number of Independent Affected subjects tested
  XML: ObservedIn.FamilyData.NumFamilies
  ObservedIn.Sample.AffectedStatus
  db:clinvar.sample.families_tested
  db: clinvar.sample.affected_status

- Positive family history (optional): yes/no. Reporting that another member of a family has the reported phenotype. Does not require that other family members were included in the observation set.
  XML: ObservedIn.FamilyData.FamilyHistory
  db: clinvar.sample.positive_family_history

- Family ID (optional) Details of a family will have to be managed in BioSample, the set of data describing the family can be captured in the biosample_id
  SPREADSHEET:
  XML: ObservedIn.FamilyInfo.PedigreeID
  db: clinvar.sample.biosample_id

- Comment (optional): free text describing the sample
  SPREADSHEET:
  XML: ObservedIn.Comment
  db: GTR.dbo.comment.comment

Data source

ClinVar maintains attribution for each data element based on the description of the person and organization providing the information. In the database, these are maintained by identifiers for the organization and identifiers for the individual.

XML:
db: GTR.clinvar.attr_source.extrn_src (organization)
db: GTR.clinvar.attr_source.entered_by (individual)

Identifiers in public database records (optional)

The database cross-reference structure (XRef) is provided to represent pointers to identifiers in other databases for the same concept. For example, if a gene is being described, then XRefs can be provided to NCBI's Gene database, Ensembl, HGNC, *etc.*, including the database, the identifier, and URL.

SPREADSHEET:
XML: XRef.db, XRef.id, XRef.url
db: GTR.clinvar.attr_source

Citations (optional)

Citations include published articles and URLs. If a database name and identifier are supplied, the full text is not required.

Source

- the name of the data service providing an identifier for a citation. This value should not be completed if the citation is a URL or a free text.

SPREADSHEET: Various Tabs (Categories)
SPREADSHEET: --.Citation (Source automatically =PubMed for this col.)
SPREADSHEET: OTHER SOURCES NOT REPRESENTED IN SPREADSHEET
XML: Citation.ID@Source
db: GTR.dbo.citation.extrn.src

- ID: the identifier provided by that data source for a citation. This value should not be completed if the citation is a URL or a free text.

SPREADSHEET: Various Tabs (Categories)
SPREADSHEET: --.Citation
XML: Citation.ID
db: GTR.dbo.citation.extrn_id

- URL: complete URL

SPREADSHEET: Various Tabs (Categories)
SPREADSHEET: --. Text or URL Citation

XML: Citation.URL
db: GTR.dbo.citation.url

- CitationText: when there is no database ID for the publication

  SPREADSHEET: Various Tabs (Categories)
  SPREADSHEET: --. Text or URL Citation
  XML: Citation.CitationText
  db: GTR.dbo.citation.citation

Comments (optional)

A free text comment can be provided to describe submitted data. This is not formatted. In the database, these are connected to the content about which a comment is made, based on the name of the database table and the unique identifier in that table. This database (db) implementation will not be documented in each section where comments are supported.

- Text: (required) the content of the comment
  SPREADSHEET: Various Tabs (Categories)
  SPREADSHEET: Comment
  XML: Comment.CommentText
  db: GTR:dbo.comment.comment

- Type: (required) public (will be rendered on the web) or private (to explain a submission and be stored in the database but not rendered on the web.)
  XML: Comment.Type
  db: GTR:dbo.comment.comment_type

Information describing the submitter and the submission

Identification of the submitter

Name of the individual responsible for the submission
SPREADSHEET:
XML: Submitter.Personnel.Person.Name.First

XML: Submitter.Personnel.Person.Name.Middle
XML: Submitter.Personnel.Person.Name.Last
db: GTR.dbo.person

Private Contact information (required, not publically displayed, for contact regarding submissions)

SPREADSHEET: (SubmitterInfo)/ Submitter phone/ Submitter fax/
Submitter email
XML:Submitter.Personnel.PrivateContact.email
XML:Submitter.Personnel.PrivateContact.phone
XML:Submitter.Personnel.PrivateContact.fax
db: GTR.dbo.contact (identified as private in GTR.dbo.org_person)

Public Contact information (optional, publically displayed)

SPREADSHEET: (SubmitterInfo)/Submitter phone/Submitter fax
Submitter email
XML:Submitter.Personnel.PublicContact.email
XML:Submitter.Personnel.PublicContact.phone
XML:Submitter.Personnel.PublicContact.fax
db: GTR.dbo.contact (identified as public in GTR.dbo.org_person)

Submitter Identifier

NCBI maintains several identifier systems for submitters (*e.g.* the dbSNP submitter handle), and there may be public identifier systems as well. Thus the identifier for a submitter is managed as a database cross-reference. The submitter handle is treated explicitly in the database.

SPREADSHEET:

XML: Personnel.Organization.SubmitterHandle
db: GTR.dbo.organization.submitter_handle
XML: Submitter.Personnel.PersonRef/@id and @db where db=snp
db: GTR.dbo.person.submitter_handle
db: GTR.dbo.person.submitter_id

Descriptors of the submission

Date submitted

If provided in XML, the date in the file as indicated below. Otherwise, the date a submission was received. If the submission is an update of an existing record, this submission date will be the date of record of a new version of the submission.
SPREADSHEET: (Submitter info)/ Submission date
XML: ClinvarSubmission.ClinvarSubmissionID@submitterDate
db: GTR.clinvar.measure_target.subdate

Release status

This allows a temporary hold on data being presented publically. Options include public, hold until published, private. If not supplied, public will be the default.
SPREADSHEET: Not represented

XML: ClinvarSubmission.ReleaseStatus
db: GTR.clinvar.measure_target.pubstat (record level submission)
db: GTR.clinvar.mset.pubstat

Submitter's identifier for the record submitted (required)

If the submitter does not provide a unique key for the record, ClinVar constructs one based on the description of the variations and the phenotypes
SPREADSHEET:
XML: ClinvarSubmission.ClinvarSubmissionID.localKey
db:GTR.clinvar.measure_target.local_key

URL to submitter's record

SPREADSHEET:
XML: Xrefs/@url (The Xref structure can be provided at many levels in the submission, to indicate what URL the submitter has for that object).
db: GTR.clinvar.attr_source

Phenotype

ClinVar requires categorization of phenotypes and sets of phenotypes. Current options include:

Controlled term	Usage
Disease	Use for diagnostic name
Drug response	Usually written as drug name + response This includes pharmacodynamic and pharmacokinetic differences
Subphenotype	Use to submit a disease hierarchy
Blood group	For the name of a blood group system. If an allele of a blood group is manifest as an additional phenotype, include in the trait set.
Finding	Use for clinical features or phenotypic measures
Infection resistance	Corresponds to **genetic resistance to infectious agent, IDO_0000587**

Sets of phenotypes

Data may be acquired from individuals or families with a set of clinical findings in addition to a diagnosis. Each finding is captured explicitly, and the co-occurring findings are represented as a set.

SPREADSHEET: Not represented

XML: TraitSet.type

db: GTR.clinvar.tset.type

Relationships among members of a set of phenotypes

Phenotype-phenotype relationships are represented as parent-child, sibling, or manifestations. An example of manifestation is a trait which is a clinical finding, and a

diagnostic name. The ClinVar/GTR staff curates some hierarchical relationships, but usually uses those provided by external authorities. If you wish to suggest a revision of current hierarchies, or suggest new ones, please contact us.

If a variation results in a condition with increased risk of an additional condition, each condition should be categorized as in the table above. An attribute of type 'risk description', with a description of the relationship, can be submitted to clarify which condition is primary and which has the increased risk.

- **SPREADSHEET**: Not represented
- **XML**: Trait.TraitRelationship.type
- **db**: GTR.clinvar.tsubset.relat_type

Description of one phenotype (trait)

Representing that a variation is not pathogenic

When making a clinical assertion of "benign"/"not pathogenic", the phenotype can either be listed as not pathogenic relative to a specific condition, or to the concept representing "all highly penetrant genetic disorders". If the latter, please submit the name of the phenotype as "AllHighlyPenetrant"

Names

Preferred name

The name of the phenotype used for reporting from ClinVar by default.

When available, this will be a preferred term from SNOMED CT. Other sources may include Office of Rare Diseases Research (ORDR), Human Phenotype Ontology (HPO), OMIM®, and MeSH. The name for the phenotype that the submitter provides will be retained, but will be mapped to controlled vocabularies when possible. Because testing laboratories may know only the name of the ordered test, the test name will suffice for the phenotype name recorded as 'test disease scope/indication'. A list of disorder names used by ClinVar/GTR is provided from ClinVar's ftp site in the file named gene_condition_source_id.

AttributeSet: preferred name

SPREADSHEET: Variant (Phenotype).Disease/Phenotype name for which clinical significance was evaluated
XML: trait.name.type=preferred
db: GTR.clinvar.target_attr where attr_type = 17 (AttributeSet)

Alternate name(s)

Other names used for this phenotype. These will be added to the set of search terms used in ClinVar and GTR.

Optional, multiple allowed

AttributeSet: alternate name

SPREADSHEET: Variant (Phenotype).Alt. Phenotype Name(s) (if multiple, pipe (|) separated)
XML: trait.name.type=alternate
db: GTR.clinvar.target_attr where attr_type = 18 (AttributeSet)

Preferred acronym

The acronym of the phenotype used for reporting from ClinVar/GTR by default.

Optional, only one allowed.

AttributeSet: preferred symbol

SPREADSHEET: Variant (Phenotype).Phenotype Abbrev. (preferred Acronym)
XML: trait.symbol.type=preferred
db: GTR.clinvar.target_attr where attr_type = 19

Alternate acronym(s)

Alternate acronyms for the phenotype.

Optional, multiple allowed.

AttributeSet: alternate symbol

SPREADSHEET: Variant (Phenotype).Alt. Phenotype Abbrev (if multiple, pipe (|) separated)

XML: trait.symbol.type=alternate
db: GTR.clinvar.target_attr where attr_type = 20

Attributes (optional)

These are based on the AttributeSet structure, and thus can are used to capture values assigned to defined information categories, along with supporting documentation.. The values can be words, integers, decimals, and/or dates. Types are restricted by an enumerated list of allowed values per major information set. These restrictions may be applied in the XSD, or only in the underlying relational database. If you wish to suggest a new attribute, please contact us at clinvar@ncbi.nlm.nih.gov.

Optional, multiple allowed

Concept	attr_type	XML	column in spreadsheet
Usual age of onset	257	AgeOfOnset	Variant(Phenotype).Age of onset
Reported penetrance	353	Penetrance	Variant (Phenotype).Penetrance
Mode of inheritance*	162	ModeOfInheritance	

* ***NOTE****: mode of inheritance is stored as an attribute of the allele/phenotype relationship. If consistent for all alleles, Mode of inheritance will also be stored as an attribute of the phenotype itself.*

Category/Type

Classification of the phenotype name.

NOTE: if the submission provides the name of a trait, and a type different from ClinVar's categorization, ClinVar will retain what is submitted but continue to report the ClinVar categorization until resolved with the submitter. Only one classification is allowed per phenotype identifier. The current choices are:

- **Disease**: usually a diagnostic term
- **Drug response**: usually constructed as name of drug + response
- **Blood group**: names of blood groups
- **Finding**: for measures/clinical features
- **Test name/indication:** for clinical testing when the suspected diagnosis is not provided

SPREADSHEET: Variant(Phenotype).Type of phenotype
XML: TraitSet.Trait@Type
db: GTR.clinvar.target.id_type

Variant allele(s)

Sets of variants

If a phenotype has been observed in a complex heterozygote, the combination of alleles is presented as a set. This representation is to be distinguished from co-occurrence, which is used to report rare alleles in genes thought to contribute to a phenotype, but for which the alleles are not thought to be pathogenic in the reported context. In the XML and the database, all variations are submitted as sets, even if the size of the set is one.

SPREADSHEET:
XML: GTR.MeasureSet
db: GTR.clinvar.mset + GTR.clinvar.msubset

OMIM allelic variant ID

An OMIM allelic variant ID will reported for a set of variations as appropriate.

SPREADSHEET:
XML: GTR.MeasureSet
db: GTR.clinvar.mset_attr (AttributeSet)

Variant allele.

Each allele needs to be described unambiguously as the location of the variation and the sequence at that location. This requirement may be achieved in any of several ways.

Location

There are multiple options to specify the location of a variation. To permit unambiguous mapping to the genome, a submission in nucleotide coordinates, as accession.version+location is highly preferred. If a LRG sequence is used, the version is not applicable. If the description of the variation is provided via an HGVS expression which includes the explicit reference sequence, then location need not be reported. For chromosome locations, the build and chromosome names (or accession+version) must be supplied for the sequence to be identifiable.

Cytogenetic

For variations defined by sequence, cytogenetic location is optional and can be provided by NCBI. For large structural variations defined only cytogenetically, this is required.

SPREADSHEET: TOSTORE
XML: Measure.CytogeneticLocation
db: GTR.clinvar.seq_loc.cytogenetic + GTR.clinvar.seq_loc.chr

Nucleotide Location

Where is this variation on a defined sequence representing a genome? Locations may be a point or a range, with or without defined end points. If a point, only start needs to be provided.

SPREADSHEET: Variant(Location and variation).Genome Assembly (*e.g.* NCBI36, GRCh37.p3)
SPREADSHEET: Variant(Location and variation).Chromosome accession.version
SPREADSHEET: Variant(Location and variation).Genomic chromosome position start
SPREADSHEET: Variant(Location and variation).Genomic chromosome position stop
XML: Measure.SequenceLocation with multiple attributes to define the assembly, sequence, and position/boundaries of the variation's location
db: GTR.clinvar.seq_loc (multiple columns)

Protein Location

Note, although submitting the definition of a variation *only* in protein coordinates will be accepted, this format is not recommended. It is our goal to map sequence variation

to the genome, and protein coordinates are not always sufficient. That said, submission of a variation as both the nucleotide change and protein change is desirable, to support confirmation of location.

SPREADSHEET: Variant(Location and variation).HGVS-protein
XML: Measure.SequenceLocation @Accession=accession.version
db: GTR.clinvar.measure_attr where attr_type in (133, 134, 194, 236), depending on how the protein change is submitted.

Variant name

The name in common use for an allele. This may be an official allele name.

SPREADSHEET: Obs-Affected(Sample Set Information).Variant Name
XML: Measure.Name@type=preferred
db: GTR.clinvar.measure_attr.attr_char where attr_type = 17 (AttributeSet)

Variation Description (*e.g.* HGVS)

SPREADSHEET: Variant(Location and variation).HGVS-genomic
SPREADSHEET: Variant(Location and variation).HGVS-transcript
XML: Measure.attribute.type=HGVS *expression*
db: GTR.clinvar.measure_attr.attr_char where attr_type is of of the HGVS class (AttributeSet)

Strand

Optional representation of strand on which the allele is found. Likely of concern only when explicitly representing a haplotype.

SPREADSHEET: Variant(Gene). +/- strand
XML: Measure.SequenceLocation@Strand
db: GTR.clinvar.seq_loc.strand

Has paralogs

Optional flag that variation calls in this region may be confounded by paralogs in the genome. In other words, this field is not intended to report locations of all paralogs for this location; but a warning, projected to be computed by NCBI, that paralogs exist.

SPREADSHEET: Variant (Variant Consequence and Context).Is in duplicated region / is there a pseudogene
XML: Measure.AttributeSet.MeasureAttributeType=ParalogInfo
db: GTR.clinvar.measure_attr

Has pseudogenes

Optional flag which can be provided by the submitter, but usually computed by NCBI, that a gene has pseudogenes. In other words, this field is not intended to report locations of all pseudogenes for a gene; but a warning that pseudogenes exist which may affect confidence in variation calls in this region.

SPREADSHEET: Variant (Variant Consequence and Context).Is in duplicated region / is there a psuedogene
XML:
Measure.AttributeSet.MeasureAttributeType=PseudogeneInfo
db: GTR.clinvar.measure_attr

Identifiers in public databases

- Identifier in dbSNP/dbVar/OMIM/locus-specific databases, *etc.* Special handling is provided for identifiers generated by NCBI, namely rs#, nsv, nssv, in that they have dedicated attribute types. At times, a submission may include information that the location of a variation can be identified by an rs# or an nsv# or some other public identifier. In that case, the fact that this submitter made this statement is captured via attr_source.

SPREADSHEET: Variant (Location and variation).dbSNP rs
XML: Measure.AttributeSet.Attribute.rsNumber
XML: Measure.AttributeSet.Attribute.nsv
db: GTR.clinvar.attr_source
db: GTR.clinvar.measure_attr where attr_type in () for dbSNP, dbVar, OMIM respectively

Location relative to a gene, protein, or other genomic locations(optional)

Some of these values are based on sequence ontology terms and computed per transcript. Content can be computed by NCBI and/or provided by submitter. This category includes exon and intron numbers, position relative to splicing or regulatory regions, position in conserved protein domains, *etc*. The sequence ontology terms used by NCBI include:

- UTR (SO:0000203)
    - 5_prime_UTR (SO:0000204)
    - 3_prime_UTR (SO:0000205)
- Upstream location
    - Upstream variant (SO:0001631)
        - Within 5kb (SO:0001635)
        - Within 2kb (SO:0001636)
- Downstream location
    - downstream_gene_variant (SO:0001632)
        - 5KB_downstream_variant (SO:0001633)
        - 500B_downstream_variant (SO:0001634)
- Splice site
    - splice_site (SO:0000162)
- Distance from nearer splice junction
  (can be calculated if not provided)
- Regulatory site (yes/no or name of promoter/locus control region)
    - Promoter: SO:0000167

Intron or exon number (optional)

Submitters may provide an Intron or exon designation and Arabic numeral (*e.g.*, exon 4, intron 3, not IVS 3 or Exon IV). The sequence used to define the numbering system should also be included.

SPREADSHEET: Variant(Variant Consequence and Context).Intron or exon number
SPREADSHEET: Variant(Variant Consequence and Context).Numbering system for intron / exon
XML: Measure.AttributeSet.Attribute Type='Location'
db: GTR.clinvar.measure_attr where attr_type = 472

Region name (active site, conserved domain, unspecified, etc.)(optional, free text)

Submitters may provide a domain name in which the variation is found. NCBI will also report when the variation lies within a known domain.

SPREADSHEET: TOBEDEFINED
XML: Measure.AttributeSet.Attribute@type='Domain'

db: GTR.clinvar.mset_attr where attr_type = 473

Total exons in transcript

This optional concept is included in the dictionary because the value may be included in our public displays. The data are not to be submitted however, and will be provided based on the sequence used to define any gene annotation.

Other regions with similar sequence which may confound interpretation

Submitters may describe other regions in the genome with sequence highly similar to the context of the reported variant, and which may affect variation calls. This attribute may describe a gene or a variant.

SPREADSHEET: TOBEDEFINED
XML: MeasureSet.AttributeSet.Attribute@type='RelatedSequence'
db: GTR.clinvar.mt_attr where attr_type =

Molecular consequence: (optional)

Molecular consequence is reported from sequence ontology terms when available, and, when possible, shall be computed per transcript by NCBI. These terms are in this group because they can be calculated explicitly from the type and location of the variation, unlike the functional consequence which must be established experimentally (or predicted). (AttributeSet).

Frameshift (SO:0000865)
Missense (SO:0001783)

Nonsense (SO:0001587)

Synonymous (SO:0001588)

In frame (SO:0001650)

Stop lost (SO:0001578)

SPREADSHEET: Variant(Variant Consequence and Context).Molecular consequence
XML: Measure.AttributeSet.Attribute type='MolecularConsequence'
db:

Comment about molecular consequence

As with most other data elements, a free text comment may be submitted about the functional consequence. The comment structure should be used if the consequence being submitted/reported is not defined by the Sequence Ontology group. We strongly recommend, however, that an SO term be requested if current terms are insufficient.

SPREADSHEET: Variant(Variant Consequence and Context).Molecular consequence comment
XML: Measure.AttributeSet.Attribute Type='MolecularConsequence'
Measure.AttributeSet.Comment.CommentText

Functional consequence: (Optional)

These attributes will be provided by the submitter since they require determination of the consequences of the molecular change. Each shall be qualified by whether the submitter predicted the consequence or established it experimentally. This choice list includes options used by the LOVD databases (AttributeSet).

- Loss of function
- Gain of function
- Overexpression
- Underexpression
- Splice_site_lost
- Splice_site_gained
- Nonsense mediated decay
- affects function
- probably affects function
- probably does not affect function
- does not affect function
- unknown

SPREADSHEET: Variant(Variant Consequence and Context).Functional Consequence
XML: Measure.AttributeSet.Attribute@Type='FunctionalConsequence'

db: GTR.clinvar.measure_attr where attr_type = 474

Method for determining functional consequence

SPREADSHEET: Variant(Variant Consequence and Context).Method for determining functional consequence
XML: Method.Type=functional consequence
db: GTR.clinvar.method.method_type

Functional consequence comment

SPREADSHEET: Variant(Variant Consequence and Context).Functional consequence comment
XML: Measure.AttributeSet.Attribute Type='FunctionalConsequence'
Measure.AttributeSet.Comment.CommentText

Type of variation

Description of the type of variation, using terms from the Sequence Ontology as appropriate. Note that the option *undefined* exists as a default value. NCBI will reassign the type when necessary (NCBI_DB). We will not maintain all the allowed values in this document; please refer to the xsd for the complete list.

- single nucleotide variant (SO:0001483) [aka substitution]
- multiple nucleotide polymorphism (SO:0001013)
- Insertion (SO:0000667)
- Deletion (SO:0000159)
- Indel (SO:1000032)
- Duplication (SO:1000035)
- Microsatellite (SO:0000289)
- Repeat expansion

  SPREADSHEET: Variant(Location and variation).Type of Variation
  XML: Measure.MeasureType
  db: GTR.clinvar.measure.id_type

Description of the asserted relationship between a set of phenotypes and a set of variations

Mode of inheritance

The mode of inheritance can be reported both as an attribute of a disorder and as an attribute of the relationship between a [set of] variations and the disorder. In both contexts, the ontology being used is Human Phenotype Ontology (HPO).

SPREADSHEET: Variant(Clinical Assertion).Mode of Inheritance
XML: MeasureType.AttributeSet.Attribute Type="ModeOfInheritance"
db: GTR.clinvar.mt_attr where attr_type = (AttributeSet)

Clinical significance (required)

This entity is listed as required, only because an option must be selected. One option, however, is 'not provided' so that submitters are not required to calculate significance to submit their data. Clinical significance shall include the tiered values of

- pathogenic
- probably pathogenic
- variant of unknown significance
- probably not pathogenic
- no known pathogenicity

as well as other qualifiers such as:

- risk factor
- association
- confers resistance
- confers sensitivity

and pharmacologic descriptors:

- Ultrarapid metabolizer
- Extensive metabolizer
- Intermediate metabolizer
- Poor metabolizer

To support submissions from LOVD databases, the following options, reported as pathogenicity properties , are supported based on the mode of inheritance. The phenotype that is caused is defined as the trait.

when X-linked;
-this variant, in a female, is causative
-this variant, in a male, is causative

autosomal dominant;
-this variant is causative

autosomal recessive;
-this variant, when combined with a variant affecting function on the other allele, is causative

when imprinted;
-this variant, when inherited from the father, causes [phenotype]
-this variant, when inherited from the mother, causes [phenotype]

when mitochondrial;
-this variant, when inherited from the mother, causes [phenotype]

XML: MeasureTrait.ClinicalSignificance
db: GTR.clinvar.mt_attr where attr_type = 151

ClinicalSignificance is also reported explicitly for co-occurring variations which may contribute to the phenotype. See the co-occurrence section.

Pathogenicity property (optional)

SPREADSHEET:
XML: MeasureTrait.AttributeSet.Attribute attr_type = pathogenic property
db: GTR.clinvar.mt_attr where attr_type = 527

Last assessed

Date of the last evaluation of clinical significance. If not supplied, the submission date is used.

SPREADSHEET: Variant (Clinical Assertion).Date of last evaluation of clinical significance

XML: MeasureTrait.ClinicalSignificance.DateLastEvaluated
db: GTR.clinvar.mt_attr.attr_date where atttr_type =

Citation (optional)

SPREADSHEET:
XML: MeasureTrait.ClinicalSignificance.Citation
db: GTR.dbo.citation; GTR.dbo.citation_many

Identifiers in public databases

Identifier for this assertion in other databases

SPREADSHEET:
XML: MeasureTrait.ClinicalSignificance.XRef
db: GTR.clinvar.attr_source

Comment

Opportunity for free text comment

SPREADSHEET: Variant (Clinical Assertion).Comment on Clinical Significance
XML: MeasureTrait.ClinicalSignificance.Comment

Clinical Significance Review Status

Review status: indicates the level of confidence in any assertion (Required for Reviewed or Practice guideline status). One type per submission by spreadsheet.

- selected from among
  - Not classified by submitter (calculated)
  - Classified by single submitter (calculated)
  - Reviewed by expert panel (submitted only by authoritative panels)
  - Reviewed by professional society (submitted only by professional society)
  - Practice guideline
  - (calculated by NCBI if there are multiple submissions for the same phenotype/allele relationship)

SPREADSHEET: Variant(Clinical Assertion).Review Status of clinical significance
XML: MeasureTrait.ClinicalSignificance.ReviewStatus
db: GTR:clinvar.mt_attr

Custom Assertion Score (optional)

Submitter-specific scoring method names and the values obtained for each, (where submitter has alternate system/nomenclature for clinical significance). These will not be standardized, not stored in a normalized fashion in relational columns, but are being retained for the submitter's use.

SPREADSHEET: Variant(Clinical Assertion).Submitter local clinical metric, custom assertion score
XML: MeasureTrait.CustomAssertionScore
db: GTR:clinvar.version.xml_object

Evidence

The evidence section maintains the details necessary to review the medical importance of a set of variations with respect to a diagnosis or medical outcome. This evidence may be computational, based on experimental testing, or observations in human subjects. A submission may contain multiple observations, as defined by the description of the sample and method.

XML: ObservedIn
db: GTR.clinvar.observations

Sample

Citations (Optional)

Comments (Optional)

Observations

ClinVar uses 'observations' to store evidence generated from a combination of methods applied to a sample. Observations is also used to represent some conclusions about results from a combination of methods, such as 'confirmed by independent methods'.

XML path = /ClinvarSubmissonSet/ClinvarSubmisson/MeasureTrait/

Note: The field "Independent Observations" in the Sample section indicates if these counts are independent (probands or singletons) or if counts may include related subjects.

Number of Independent Affected Subjects tested

- optional
- count of the probands from families and singleton subjects. This may match the value provided in the sample description, or may be smaller if some of the sample included multiple family members
  SPREADSHEET
  XML: ObservedIn/ObservedData/ObsAttributeType[@val_type="name"] = "IndependentObservations"
  XML: ObservedIn/ObservedData/Attribute/@integerValue = *number*
  XML: ObservedIn.Sample.AffectedStatus = "yes"
  db: GTR.clinvar.obs_attr where attr_type =IndependentObservations
  db: GTR.clinvar.sample.affected_status='yes'

Number of Variant Alleles observed in Affected

- optional
- number of times the variant allele was observed in individuals with the phenotype
  SPREADSHEET
  XML: ObservedIn/ObservedData/ObsAttributeType[@val_type="name"] = "VariantAlleles"
  XML: ObservedIn/ObservedData/Attribute/@integerValue = *number*
  XML: ObservedIn.Sample.AffectedStatus = "yes"
  db: GTR.clinvar.obs_attr where attr_type ='VariantAlleles'
  db: GTR.clinvar.sample.affected_status='yes'

Number of affected subjects who are homozygous variant [including hemizygous]

- Optional
- number of affected individuals who are homozygous for only this variant

  SPREADSHEET

  XML: ObservedIn/ObservedData/ObsAttributeType[@val_type="name"] = "SubjectsOnlyVariant"

  XML: ObservedIn/ObservedData/Attribute/@integerValue = *number*

  XML: ObservedIn.Sample.AffectedStatus = "yes"

  db: GTR.clinvar.obs_attr where attr_type ='SubjectsOnlyVariant'

  db: GTR.clinvar.sample.affected_status='yes'

Number of affected subjects who are observed as single heterozygotes

- optional
- This count includes single heterozygotes reported in an affected subject, in the context of dominant mode of inheritance, and single heterozygotes observed (in a recessive context) but where no other pathogenic variant was identified to classify as a compound heterozygote

  SPREADSHEET

  XML: ObservedIn/ObservedData/ObsAttributeType[@val_type="name"] = "SingleHeterozygote"

  XML: ObservedIn/ObservedData/Attribute/@integerValue = *number*

  XML: ObservedIn.Sample.AffectedStatus = "yes"

  db: GTR.clinvar.obs_attr where attr_type ='SingleHeterozygote'

  db: GTR.clinvar.sample.affected_status='yes'

Number of affected subjects who are observed as compound heterozygotes

- optional
- This is a count of heterozygotes observed in affected subjects where another heterozygous pathogenic variant partner WAS identified. Both variant alleles must be submitted.

  SPREADSHEET

  XML: ObservedIn/ObservedData/ObsAttributeType[@val_type="name"] = "CompoundHeterozygote"

  XML: ObservedIn/ObservedData/Attribute/@integerValue = *number*

  XML: ObservedIn.Sample.AffectedStatus = "yes"

db: GTR.clinvar.obs_attr where attr_type =CompoundHeterozygote
db: GTR.clinvar.sample.affected_status='yes'

Number of affected subjects with genotype consistent with mode of inheritance

- optional
- The sum of single heterozygotes, compound heterozygotes, and homozygotes for the reported allele with a phenotype consistent with the asserted mode of inheritance
  SPREADSHEET
  XML: ObservedIn/ObservedData/ObsAttributeType[@val_type="name"] = "GenotypeAndMOIConsistent"
  XML: ObservedIn/ObservedData/Attribute/@integerValue = *number*
  XML: ObservedIn.Sample.AffectedStatus = "yes"
  db: GTR.clinvar.obs_attr where attr_type =GenotypeAndMOIConsistent
  db: GTR.clinvar.sample.affected_status='yes'

Number of unaffected subjects who are homozygous variant [including hemizygous]

- optional
  SPREADSHEET
  XML: ObservedIn/ObservedData/ObsAttributeType[@val_type="name"] = "SubjectsOnlyVariant"
  XML: ObservedIn/ObservedData/Attribute/@integerValue = *number*
  XML: ObservedIn.Sample.AffectedStatus = "no"
  db: GTR.clinvar.obs_attr where attr_type ='SubjectsOnlyVariant'
  db: GTR.clinvar.sample.affected_status='no'

Number of unaffected subjects who are heterozygotes.

- optional
- In a dominant context this is evidence against pathogenicity or for non-penetrance. In a recessive context these subjects are carriers.
  SPREADSHEET
  XML: ObservedIn/ObservedData/ObsAttributeType[@val_type="name"] = "SingleHeterozygote"

XML: ObservedIn/ObservedData/Attribute/@integerValue = number
XML: ObservedIn.Sample.AffectedStatus = “no”
db: GTR.clinvar.obs_attr where attr_type =’SingleHeterozygote’
db: GTR.clinvar.sample.affected_status=’no’

Number unaffected subjects with relevant variant genotype.

- optional
- This includes Heterozygous, Compound Heterozygous, and Homozygous variant-containing genotypes, as determined by the asserted mode of inheritance.
  This count is either evidence against pathogenicity, or evidence for non-penetrance
  SPREADSHEET
  XML: ObservedIn/ObservedData/ObsAttributeType[@val_type=“name”] = “UnaffectedSubjectsWithVariantGenotype”
  XML: ObservedIn/ObservedData/Attribute/@integerValue = *number*
  XML: ObservedIn.Sample.AffectedStatus = “no”
  db: GTR.clinvar.obs_attr where attr_type =’UnaffectedSubjectsWithVariantGenotypeThusInconsistantWithPathogenicAssertion’
  db: GTR.clinvar.sample.affected_status=’no’

Number de novo variants observed in affected subjects

- optional
  SPREADSHEET
  XML: ObservedIn/ObservedData/ObsAttributeType[@val_type=“name”] = “VariantAlleles”
  XML: ObservedIn/ObservedData/Attribute/@integerValue = *number*
  XML: ObservedIn.Sample.AffectedStatus = “yes”, Origin=”de novo”
  db: GTR.clinvar.obs_attr where attr_type =’VariantAlleles’
  db: GTR.clinvar.sample.affected_status=’yes’, origin = ‘de novo’

Number of affected subjects with this variant who also have another variant thought to be responsible for phenotype

This information is captured to evaluate pathogenicity, based on the logic that if another allele may account for the observed phenotype, this one has unknown pathogenicity.

- optional
  SPREADSHEET
  XML: ObservedIn/ObservedData/ObsAttributeType[@val_type="name"] = "SubjectsWithDifferentCausativeAllele"
  XML: ObservedIn/ObservedData/Attribute/@integerValue = *number*
  XML: ObservedIn.Sample.AffectedStatus = "yes"
  db: GTR.clinvar.obs_attr where attr_type='SubjectsWithDifferentCausativeAllele'
  db: GTR.clinvar.sample.affected_status='yes'

Number instances observed of heterozygous parent transmitting this variant to an affected child

- optional
- Discussion: What is the best way to capture the concepts of heritability and mode of inheritance?
  SPREADSHEET
  XML: ObservedIn/ObservedData/ObsAttributeType[@val_type="name"] = "ObservedHetParentTransmitNormalAlleleVariantToAffected "
  XML: ObservedIn/ObservedData/Attribute/@integerValue = *number*
  XML: ObservedIn.Sample.AffectedStatus = "yes"
  db: GTR.clinvar.obs_attr where attr_type ='ObservedHetParentTransmitNormalAlleleVariantToAffected '
  db: GTR.clinvar.sample.affected_status='yes'

Number instances of heterozygous parent transmitting the normal allele to affected child, rather than the variant allele

- optional
  SPREADSHEET
  XML: ObservedIn/ObservedData/ObsAttributeType[@val_type="name"] = "ObservedHetParentTransmitNormalAlleleToAffected"
  XML: ObservedIn/ObservedData/Attribute/@integerValue = *number*
  XML: ObservedIn.Sample.AffectedStatus = "yes"

db: GTR.clinvar.obs_attr where attr_type ='ObservedHetParentTransmitNormalAlleleToAffected'
db: GTR.clinvar.sample.affected_status='yes'

Number of affected families with the variant genotype

- optional

  SPREADSHEET

  XML: ObservedIn/ObservedData/ObsAttributeType[@val_type="name"] = "AffectedFamiliesWithVariantGenotype"
  XML: ObservedIn/ObservedData/Attribute/@integerValue = *number*
  XML: ObservedIn.Sample.AffectedStatus = "yes"
  db: GTR.clinvar.obs_attr where attr_type ='AffectedFamiliesWithVariantGenotype'
  db: GTR.clinvar.sample.affected_status='yes'

Number of independent families seen to co-segregate the variant allele and affected phenotype among two or more family members

- optional
- SPREADSHEET

  XML: ObservedIn/ObservedData/ObsAttributeType[@val_type="name"] = "CosegregatingFamilies"
  XML: ObservedIn/ObservedData/Attribute/@integerValue = *number*
  XML: ObservedIn.Sample.AffectedStatus = "yes"
  db: GTR.clinvar.obs_attr where attr_type ='CosegregatingFamilies'
  db: GTR.clinvar.sample.affected_status='yes'

Number of informative meioses

- optional
- SPREADSHEET

  XML: ObservedIn/ObservedData/ObsAttributeType[@val_type="name"] = "InformativeMeioses"
  XML: ObservedIn/ObservedData/Attribute/@integerValue = *number*
  XML: ObservedIn.Sample.AffectedStatus = "yes"
  db: GTR.clinvar.obs_attr where attr_type ='InformativeMeioses'
  db: GTR.clinvar.sample.affected_status='yes'

Co-occurrence with other pathogenic

This section represents multi-locus genotypes where other variant(s) suspected of affecting this phenotype are observed co-occurring in subjects with this variant.

Note: Co-occurrences affect the evaluation of the current variant in several circumstances including: 1) Where the other variant forms a compound het with this variant 2) Where another pathogenic variant is observed which is believed to be independently responsible for the phenotype 3) When the subject is unaffected but is trans with a known knock out allele in the same gene, and double knockout of this gene is lethal.

- Zygosity of the asserted variant, the target of this ClinVar assertion record. (*Value*=HomozygousVariant, SingleHeterozygote, or CompoundHeterozygote)
  SPREADSHEET:
  XML: ObservedIn/Co-occurrenceSet/Zygosity
  db: GTR.clinvar.obs_attr where att_type=*'value'*

- Affected Status (see note above)
  SPREADSHEET:
  XML: ObservedIn.Sample.AffectedStatus = "yes" *or* "no"
  db: GTR.clinvar.sample.affected_status='yes' *or* 'no'

- Count of subjects with this co-occurrence

  SPREADSHEET:
  XML: ObservedIn/Co-occurrenceSet/Count
  db: GTR

- In Co-occurence
  XML: MeasureTrait.ObservedIn.Co-occurrenceSet.AlleleDescSet.ClinicalSignificance
  db:

Last assessed

Note: The fields below are a repeating set for each unique combination of co-occurring genotypes. Currently planned to allow co-occurrence sets of up to three co-occurring variant genotypes.

- Definition of co-occurring genotype
  - Variation identifier (name)
  - Allele
  - Gene
  - Zygosity the other variation
  - If heterozygous, is it known to be in trans with the target variant.
  - Original submitted clinical interpretation of the other variant.

  MeasureTrait.ObservedIn.Co-occurrenceSet.Zygosity
  MeasureTrait.ObservedIn.Co-occurrenceSet.AlleleDescSet.Name
  MeasureTrait.ObservedIn.Co-occurrenceSet.Zygosity.Count

Controls: Unaffected for phenotype being asserted.

These data may include general population surveys such as 1000 genomes, or sample-sets such as population based carrier screening

- Number of reference homozygotes
- Number of heterozygotes
- Number of variant homozygotes
- Number of variant alleles observed

Subject-based, observation specific

NOTE: This section is included for completeness, but the data will not be stored in ClinVar. Instead, individual-level data submitted to dbGaP/genotype archive/BioSamples will be aggregated for submission to ClinVar

- Anonymous case ID (Lab ID)
- Anonymous pedigree ID
- Patient Age
- Patient Gender
- Patient sub-phenotype

- Patient Ethnicity
- Country of Origin
- DNA source tissue
- Disease/Phenotype Name
- Type of Phenotype (Disease/Drug Response, etc.)
- Disease scope/target of test, to infer likely phenotype if none provided
- Reason for testing (diagnosis, carrier screening, etc.)
- Proband status (yes|no)
- Affected (yes|no) [requires assertion of a phenotype/scope of test]
- Considered healthy (yes|no)

Description of a gene (optional)

Names

A gene, if provided, must be unambiguously defined. That definition may be supplied either by a unique name or symbol, or an identifier in a public database such as a GeneID or an HGNC id. ClinVar will coordinate the representation of names of genes with NCBI's Gene database. In other words, the name will be defined primarily by the nomenclature established by the HUGO Gene Nomenclature Committee (HGNC)

Preferred name

The preferred full name as reported by NCBI's Gene database.

Optional, only one allowed.
AttributeSet:

SPREADSHEET:	NOT REPRESENTED
XML:	measure.name.@type=preferred
db:	GTR.clinvar.measure_attr where attr_type = 17

Alternate name(s)

Other names used for this gene. Provided via the Gene database.

Optional, multiple allowed
AttributeSet:

SPREADSHEET:	NOT REPRESENTED

XML: measure.name.@type=alternate
db: GTR.clinvar.measure_attr where attr_type = 18

Preferred symbol

The official symbol from HGNC. We need some unambiguous identifier for a gene. Either an official symbol, GeneID, or HGNC_id is allowed, and one is required.

AttributeSet:

SPREADSHEET: Variant(Gene).Gene symbol
XML: measure.symbol.@type=preferred
db: GTR.clinvar.measure_attr where attr_type = 19

Alternate symbols(s)

Alternate gene symbols from Gene. Optional, multiple allowed.

AttributeSet:

SPREADSHEET:
XML: measure.symbol.@type=alternate
db: GTR.clinvar.measure_attr where attr_type = 20

Attributes

Examples would be HGNC ids, GeneID, MIM number, chromosome, cytogenetic band, chromosome sequence location, related pseudogenes/paralogs. The set of optional attributes is designed to capture information necessary to set the framework for interpretation of variation. It thus is highly desirable to report the LRG or the RefSeqGene on which interpretation of gene structure is based.

NOTE: many of these are not duplicated in the ClinVar database but are provided by NCBI as imports from the Gene database or defined by the sequence used to define the gene structure.

XML: measure.attribute@type as enumerated

db: GTR.clinvar.measure_attr where attr_type defines the content

Concept	attr_type	XML	column in spreadsheet
Location of the gene as defined by GRC assembly	Location		
Type of gene (e.g. protein-coding/non-coding/pseudogene)	NA/managed in Gene		

Instructions: The accurate interpretation and reporting of genetic test results is contingent upon the reason for referral, clinical information provided, and family history. To help provide the best possible service, please check the applicable clinical information below.

Patient Identification

Patient Name: _____ (Last) _____ (First) **Gender:** **[ ]** Male **[ ]** Female

Date of Birth: _____ (mm/dd/yyyy)

Clinical Information – Check all that apply. Use additional space at the bottom of the form if needed.

Perinatal History
[ ] Prematurity
[ ] Intrauterine growth restriction
[ ] Oligohydramnios
[ ] Polyhydramnios
[ ] Non-immune hydrops fetalis
[ ] Other: _____

Growth
[ ] Failure to thrive
[ ] Overgrowth
[ ] Short stature
[ ] Other: _____

Cognitive/Developmental
[ ] Learning disability
[ ] Developmental delay
  [ ] Gross motor delay
  [ ] Fine motor delay
  [ ] Speech delay
[ ] Intellectual disability/MR
[ ] Other: _____

Behavioral/Psychiatric
[ ] Autism
[ ] Pervasive developmental delay
[ ] Attention deficit hyperactivity disorder
[ ] Anxiety
[ ] Behavioral/psychiatric abnormality
  Specify: _____
[ ] Other: _____

Cutaneous
[ ] Hyperpigmentation
[ ] Hypopigmentation
[ ] Other: _____

Neurological
[ ] Seizures
[ ] Hypotonia
[ ] Hypertonia
[ ] Cerebral palsy
[ ] Encephalopathy
[ ] Structural brain anomaly
  Specify: _____
[ ] Other: _____

Cardiac
[ ] Atrial septal defect
[ ] Ventricular septal defect
[ ] Coarctation of the aorta
[ ] Tetralogy of Fallot
[ ] Other structural heart defect
  Specify: _____
[ ] Other cardiac abnormality
  Specify: _____

Craniofacial
[ ] Dysmorphic facial features
  Specify: _____
[ ] Ear malformation
  Specify: _____
[ ] Cleft lip
[ ] Cleft palate
[ ] Macrocephaly
[ ] Microcephaly
[ ] Other: _____

Hearing/Vision
[ ] Hearing loss
  Specify: _____
[ ] Abnormality of Vision
  Specify: _____
[ ] Abnormality of Eye Movement
  Specify: _____
[ ] Other: _____

Musculoskeletal
[ ] Contractures
[ ] Club foot
[ ] Diaphragmatic hernia
[ ] Limb anomaly
  Specify: _____
[ ] Polydactyly
  Specify: _____
[ ] Syndactyly
  Specify: _____
[ ] Vertebral anomaly
  Specify: _____
[ ] Other: _____

Gastrointestinal
[ ] Gastroschisis
[ ] Omphalocele
[ ] Anal atresia
[ ] Tracheoesophageal fistula
[ ] Pyloric stenosis
[ ] Other: _____

Genitourinary
[ ] Ambiguous genitalia
[ ] Hydronephrosis
[ ] Kidney malformation
  Specify: _____
[ ] Cryptorchidism
[ ] Hypospadias
[ ] Other: _____

Family History
[ ] Parents with ≥ 2 miscarriages
[ ] Other relatives with similar clinical history
  Explain: _____
  _____
  _____

Please include any additional relevant information not provided above (list karyotype if known).

As a participant in the ISCA (International Standards for Cytogenomic Arrays) Consortium, this clinical cytogenetics laboratory contributes submitted clinical information and test results to a HIPAA compliant, de-identified public database as part of the NIH's effort to improve understanding of the relationship between genetic changes and clinical symptoms. Confidentiality is maintained. Patients may request to opt-out of this scientific effort by: 1) checking the box below, 2) calling the laboratory at XXX-XXX-XXXX and asking to speak with a laboratory genetic counselor. Please call with any questions.
[ ] Indicate refusal for inclusion in these efforts by checking this box. If the box is not marked, data will be anonymized and used.

International Standards for Cytogenomic Arrays Consortium

Prenatal Chromosome Microarray Testing

Patient Clinical Information Form

[your logo here]

Instructions: The accurate interpretation and reporting of genetic test results is contingent upon the reason for referral, clinical information provided, and family history. To help provide the best possible service, please check applicable clinical information below. **Send this page with the specimen or return by fax to the Cytogenetics Laboratory at the contact number below. If a karyotype has been performed, please record the results at the bottom of the form.**

Lab Name: **Contact Information:**

Test Request Date: _____ **Referring Physician:** _____ **Physician Specialty:** _____

Patient Last Name:	Patient First Name:	Date of Birth:	Fetal Gender: □ Male □ Female	LMP:

Clinical Information: Please check all that apply:

Primary Indication for Testing:
- □ Abnormal serum screen
- □ Advanced maternal age
- □ Fetal abnormality as indicated
- □ None specified

Perinatal History:
- □ IUGR
- □ Oligohydramnios
- □ Polyhydramnios
- □ Increased nuchal translucency (includes cystic hygroma)
- □ Hydrops (unknown or infection)
- □ 2 vessel cord
- □ Other:

_____

Family History:
- □ Parents with ≥ 2 miscarriages
- □ Other relatives with similar clinical history (please explain):

_____
_____
_____
_____
_____

Neurological:
- □ NTD (myelomeningocele)
- □ Agenesis of the corpus callosum
- □ Dandy Walker (posterior fossa abnormality)
- □ Ventriculomegaly/ hydrocephaly
- □ Holoprosencephaly
- □ Decreased fetal movement
- □ Abnormal gyri (lissencephaly)
- □ Structural brain anomaly
- □ Cerebellar hypoplasia
- □ Other:

_____

Craniofacial:
- □ Cleft lip +/- cleft palate
- □ Cleft palate alone
- □ Hyper/Hypotelorism
- □ Macrocephaly
- □ Microcephaly
  List HC, if known:

_____
- □ Other:

_____

Pulmonary:
- □ CCAM/small thoracic cavity
- □ Diaphragmatic hernia
- □ Eventration of diaphragm
- □ Pulmonary sequestration
- □ Pleural effusion
- □ Other:

_____

Cardiac:
- □ ASD
- □ AV canal defect
- □ Coarctation of the aorta
- □ Hypoplastic left heart
- □ Tetralogy of Fallot
- □ VSD
- □ Echogenic intracardiac focus
- □ Dextrocardia or situs inversus
- □ Hypoplastic right heart
- □ Double outlet right ventricle
- □ Transposition of the great Vessels
- □ Truncus arteriosus
- □ Pulmonary valve atresia
- □ Aortic atresia
- □ Ebstein anomaly
- □ Other:

_____

Gastrointestinal:
- □ Gastroschisis
- □ Omphalocele
- □ Absent stomach
- □ Tracheoesophageal fistula
- □ Echogenic focus
- □ Meconium ileus/anal atresia
- □ Other:

_____

Musculoskeletal:
- □ Contractures (arthrogryposis)
- □ Club foot (bilateral)
- □ Limb anomaly
- □ Polydactyly
- □ Clenched hands
- □ Scoliosis
- □ Syndactyly
- □ Vertebral anomaly
- □ Micromelia
- □ Mesomelia
- □ Acromelia
- □ Skeletal dysplasia
- □ Other:

_____

Genitourinary:
- □ Ambiguous genitalia
- □ Hydronephrosis (pelvic AP diameter > 7mm)
- □ Kidney malformation
- □ Megacystis (including posterior valves)
- □ Polycystic kidneys
- □ Renal agenesis
- □ Urethra/ureter obstruction
- □ Other:

_____

Clinical description: Please include any additional relevant clinical information not provided above (list karyotype if known).

As a participant in the ISCA (International Standards for Cytogenomic Arrays) Consortium, this clinical cytogenetics laboratory contributes submitted clinical information and test results to a HIPAA compliant, de-identified public database as part of the NIH's effort to improve understanding of the relationship between genetic changes and clinical symptoms. Confidentiality is maintained. Patients may request to opt-out of this scientific effort by: 1) checking the box below, 2) calling the laboratory at 1-800-366-1502 and asking to speak with a laboratory genetic counselor. Please call with any questions. [ ] Indicate refusal for inclusion in these efforts by checking this box. If the box is not marked, data will be anonymized and used.

NOONAN SPECTRUM REQUISITION FORM

Patient Name: _____ Date of Birth: _____/_____/_____ (MM/DD/YYYY)

TEST TO BE PERFORMED

Please check box(es) to order.

Noonan Spectrum Disorders (Noonan, LEOPARD, Cardio-Facio-Cutaneous, and Costello Syndromes)

_____ ❑ Noonan Spectrum Gene Chip (*PTPN11, SOS1, RAF1, KRAS, NRAS, SHOC2, BRAF, MEK1, MEK2, HRAS*) $1,500

Please contact the laboratory for individual gene sequencing tests.

Familial Variant Testing

_____ ❑ Familial Variant(s)** OR ❑ Research Confirmation** $400

If proband testing was performed elsewhere, please attach a copy of the original result and send positive control sample, if available

Gene_____ Variant_____

Proband Name_____

Relationship to Patient_____ LMM Accession #: PM-_____

CLINICAL INFORMATION

Clinical status: ❑ Affected ❑ Unknown ❑ Unaffected

Purpose of study: ❑ Diagnostic ❑ Family history ❑ Prenatal ❑ Other: _____

Clinical diagnosis: ❑ Noonan ❑ LEOPARD ❑ CFC ❑ Costello **ICD-9 Codes:** ❑ 759.89 (Noonan/LEOPARD)
(check all known/suspected clinical diagnoses) ❑ 759.8 (Other specified abnormalities)
Age at diagnosis: _____ ❑ Other _____

Ultrasound Finding: ❑ Cystic hygroma ❑ Increased NT - Size: _____ ❑ None
❑ Heart defect - Type: _____ ❑ Other _____

Congenital heart defect: ❑ Pulmonic valve stenosis ❑ Hypertrophic cardiomyopathy ❑ None
❑ Septal defect ❑ Other _____

Facial dysmorphism: ❑ Epicanthal folds ❑ Ptosis of the eyelids ❑ Low nasal bridge
❑ Hypertelorism ❑ Downward eye slant ❑ Low set ears and posteriorly rotated
❑ Papillomatas ❑ Coarseness ❑ None

Short stature: ❑ Yes - Height(%): _____ Parental Heights: _____ ❑ No

Cognitive development: ❑ Learning disabilities ❑ Developmental delay ❑ Mental retardation ❑ Normal

Skeletal: ❑ Pectus excavatum ❑ Pectus carinatum ❑ Scoliosis ❑ Normal

Genitourinary: ❑ Cryptorchidism (undescended testes) ❑ Normal
❑ Kidney malformation If yes, please describe:_____

Hair/Skin findings: ❑ Loose anagen hair ❑ Lentigines ❑ Café-au-lait spots ❑ Other _____ ❑ No

Bleeding diathesis: ❑ Yes ❑ No If yes, please describe:_____

Malignancy: ❑ Yes ❑ No If yes, please describe:_____

Other:_____

Previous Genetic Testing: ❑ Yes ❑ No Gene(s)/Tests: _____
Result (if variants detected, please elaborate):_____

Has another family member already had genetic testing for this disease? ❑ Yes ❑ No
If yes, please describe in the comments section and attach a copy of the genetic test lab report and pedigree.

FAMILY HISTORY

Family History: ❑ Yes ❑ No (Sketch below or attach pedigree, if appropriate)

Paternal Side:_____
Maternal Side:_____
Consanguinity: ❑ Yes ❑ No

○ = Female □ = Male ◇ = Gender Unspecified
● ■ ◆ = Affected Individual ⊙ = Carrier

Laboratory for Molecular Medicine • 65 Landsdowne Street • Cambridge, MA, 02139
Phone: 617-768-8500 • Fax: 617-768-8513 • Website: http://pcpgm.partners.org/lmm • Email: lmm@partners.org

Last Revised: 14 Jul 2011

HEARING LOSS REQUISITION FORM

Patient Name: _____ Date of Birth: _____/_____/_____ (MM/DD/YYYY)

CLINICAL INFORMATION

Clinical status: ❑ Affected ❑ Unknown (no screening/evaluation(s)) ❑ Unaffected (all screening/evaluation(s) normal)
Purpose of study: ❑ Diagnostic ❑ Carrier testing ❑ Other _____

Age of onset of hearing loss:_____ **ICD9 codes:** ❑ 389.1 (sensorineural hearing loss) ❑ Other

Type of hearing loss: ❑ Sensorineural ❑ Conductive ❑ Auditory neuropathy/dys-synchrony ❑ Mixed

Laterality: ❑ Bilateral ❑ Unilateral

Severity (PTA): **Please attach audiogram if available**
Left Ear: ❑ Mild (15-30dB) ❑ Moderate (31-50dB) ❑ Moderately-severe (51-70dB) ❑ Severe (71-90dB) ❑ Profound (>90db)
Right Ear: ❑ Mild (15-30dB) ❑ Moderate (31-50dB) ❑ Moderately-severe (51-70dB) ❑ Severe (71-90dB) ❑ Profound (>90db)

Audiogram Shape / Frequencies:
Left Ear: ❑ Flat (all frequencies) ❑ Sloping (high frequency) ❑ Saucer-shaped (mid frequency) ❑ Rising (low frequency)
Right Ear: ❑ Flat (all frequencies) ❑ Sloping (high frequency) ❑ Saucer-shaped (mid frequency) ❑ Rising (low frequency)

Progression: ❑ Stable ❑ Progressive ❑ Fluctuating ❑ Unknown

Auditory neuropathy/dys-synchrony: ❑ No ❑ Present OAEs ❑ Absent ABR w/ cochlear microphonic ❑ Unknown

Vestibular problems: ❑ None ❑ Delayed walking ❑ Dizziness ❑ Vertigo ❑ Balance problems ❑ Unknown

Temporal bone abnormalities on CT/MRI: ❑ None ❑ EVA ❑ Mondini dysplasia ❑ Unknown
❑ Other (explain):_____

Stapes fixation: ❑ Yes ❑ No **Perilymphatic gusher with stapedectomy:** ❑ Yes ❑ No

Exposure to aminoglycoside antibiotics (e.g gentamicin, neomycin, tobramycin, amikacin): ❑ Yes ❑ No ❑ Unknown

Eye Findings: ❑ None ❑ Unknown ❑ Retinitis pigmentosa - age of onset:_____
❑ Other (explain):_____

BOR Features: ❑ None ❑ Ear tags ❑ Ear pits ❑ Ear abnormalities ❑ Branchial arch abnormality ❑ Renal abnormality
❑ Explain:_____

Previous Genetic Testing: ❑ No ❑ Yes - Gene(s)_____
If variants detected, please elaborate:_____

Other relevant medical problems:_____
_____

Has another family member already had genetic testing for this disease? ❑ Yes ❑ No
If yes, please describe and attach a copy of the genetic test lab report and pedigree.

FAMILY HISTORY

Sibling with or other family history of similar hearing loss? ❑ Yes ❑ No

List affected individuals and the nature of their hearing loss (Sketch below or attach pedigree if appropriate):_____
_____
_____

Paternal Side:_____
Maternal Side:_____
Consanguinity: ❑ Yes ❑ No

○ = Female □ = Male ◇ = Gender Unspecified
● ■ ◆ = Affected Individual ⊙ = Carrier

Laboratory for Molecular Medicine • 65 Landsdowne Street • Cambridge, MA, 02139
Phone: 617-768-8500 • Fax: 617-768-8513 • Website: http://pcpgm.partners.org/lmm • Email: lmm@partners.org

Last Revised: 22 Jul 2011

ISCA Member Institutions

ACL Laboratories, Rosemont, United States
Affymetrix, Santa Clara, United States
Aga Khan Univ Hospital, Karachi, Pakistan
Agilent Technologies, Santa Clara, United States
Alberta Children's Hospital, Calgary, Alberta, Cananda
Alberta Health Services Genetics Labs, Edmonton, Canada
All Childrens Hospital Cytogenetics and Molecular Cytogenetics, St Petersburg, United States
ARUP Laboratories, University of Utah, Salt lake City, United States
Athena Diagnostics, Worcester, United States
Baylor College of Medicine, Houston, United States
Beaumont Hospitals, Royal Oak Michigan, United States
Beth Israel Deaconess Medical Center, Boston, MA, United States
Bioarray SL, Crevillente, Spain
Biochemistry and Molecular Genetics, Barcelona, Spain
BioDiscovery, El Segundo, United States
BlueGnome Limited, Cambridge, United Kingdom
Boston University School of Medicine, Boston, United States
Bristol Genetics Laboratory, Bristol, United Kingdom
Cartagenia, Leuven, Belgium
Center for Medical Genetics Ghent University Hospital, Ghent, Belgium
Central Regional Genetic Services, Wellington Hospital, Wellington, New Zealand
Centre for Human Genetics and Laboratory Medicine, 82152 Martinsried, Germany
Changhua Christian Hospital, Changhua, Taiwan, Province of China
Children's Hospital at Westmead, Westmead, Australia
Children's Hospital of Philadelphia, Philadelphia, United States
Children's Medical Center of Dayton Dayton, United States
Children's Memorial Hospital, Sydney, United States
Childrens Hospital and Research Center Oakland, Oakland, United States
Chlidrens Mercy Hospital Cytogenetics Microarray Lab, Kansas City, United States
CHU Nantes, Nantes, France
Cincinnati Children's Hospital, Cinncinati, OH, United States
Columbia University Medical Center, New York, United States
CombiMatrix Diagnostics, Irivine, United States
Comprehensive Genetic Services of Richmond County, New York, United States
Comprehensive Genetics, Milwaukee, United States
Credit Valley Hospital, Canada
Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus
Cytogenetics, Vancouver, Canada
Cytogenetics, Lecce, Italy
Cytogenetics Diagnostic Genetics LabPlus Auckland City Hospital, Auckland, New Zealand
Cytogenetics The Royal Chidren's Hospital, Melbourne, Australia
Cytogenetics, London Health Sciences Centre and University of Western Ontario, Canada
Dayton Childrens Medical Center, Dayton, United States
DNA Laboratories, Bangi, Malaysia
Duke University Health System Clinical Labs, Durham, NC, United States
Emory University, Emory Genetics Laboratory, Atlanta, United States

GeneDx, Gaithersburg, United States
General Lab LabCo, Barcelona, Spain
Genetics Center, Orange, United States
Genetikum, Neu-Ulm, Germany
GENOMA Molecular Genetics Laboratory, Rome, Italy
Genomax Technologies Pte ltd, Singapore, Singapore
Genomic Software, Sunnyvale, United States
Genycell Biotech España S.L., Granada, Spain
Great Ormond Street Hospital North East Thames Regional Genetics Service, London, United Kingdom
Greenwood Genetic Center, Greenwood, United States
Guys and St Thomas Hospital, London, United Kingdom
Hamad Medical Corporation, Qatar
Hayward Genetics, New Orleans, United States
Henry Ford Hospital, United States
HMC Genetics laboratory, doha, Qatar
Hospital for Sick Children, United States
Hunter Area Pathology Service, Newcastle, Australia
Illumina, San Diego, United States
Institut fuer medizinische Genetik, Dresden, Germany
Institute of Human Genetics, Newcastle upon tyne, United Kingdom
Institute of Human Genetics, Graz, Austria
Istanbul University Cerrahpasa Medical School Department of Medical Genetics, Istanbul, Turkey
Karolinska Institute, Stockholm, Sweden
Keio University School of Medicine, Tokyo, Japan
King Faisal Hospital and Research Centre, Riyadh, Saudi Arabia
KK Hospital, Singapore, Singapore
Labor OvensRaeder, Muenchen, Germany
Laboratory of Medical Genetics, Maribor, Slovenia
Laboratory of Medical Genetics National Taiwan University Hospital, Taipei, Taiwan, Province of China
Mater Pathology, Brisbane, Australia
Mayo Clinic, Rochester, United States
Medical Genetics, Basel, Switzerland
Medical Genetics Center MGZ Munich, Munich, Germany
Medical Genetics Laboratory, Vandoeuvre les Nancy, France
Memorial University of NL, St Johns, Canada
Mission Health and Hospitals Fullerton Genetics Laboratory, United States
Montefiore Hospital, United States
Mount Sinai Cytogenetics and Cytogenomics Laboratory, New York, United States
Mount Sinai School of Medicine, United States
National Center for Biotechnology Information, Bethesda, MD
National Center for Child Health and Development, Tokyo, Japan
National University Health System, Singapore, Singapore
National Yang Ming University, Taipei, Taiwan, Province of China
NIMGenetics, Tres Cantos, Spain
Northwestern Reproductive Genetics, United States
OUHSC Genetica, Ok City, United States
Oxford Gene Technology, Oxford, United Kingdom

Pacific Laboratory Products, Melbourne, Australia
PathWest Laboratory Medicine, Perth, Australia
Phalanx Biotech Group, Inc., Hsinchu, Taiwan
Pittsburgh Cytogenetics Laboratory, Pittsburgh, United States
Prenatal Genetics, Barcelona, Spain
Prenatal medicine and genetics, Duesseldorf, Germany
Prince of Wales Hospital, Randwick Sydney, Australia
QML Pathology, Brisbane, Australia
Quantitative Genomic Medicine Laboratories, Barcelona, Spain
Roche NimbleGen, Inc., Madison, WI, United States
Royal North Shore Hospital, Sydney, Australia
SA Pathology, Adelaide, Australia
Sanford Clinic USD Genetics Laboratory, United States
SciGene, Sunnyvale, United States
Scott and White Memorial Hospital Cytogenetics, Temple, United States
Sengenics Sdn Bhd, Kuala Lumpur, Malaysia
Seoul National University, Seoul, Korea, Republic of
Servicio de Genetica Complejo Hospitalario de Toledo, toledo, Spain
South East Scotland Cytogenetics Service, Edinburgh, United Kingdom
St. Christopher's Hospital for Children, Phildadelphia, PA, United States
Stanford Hospital and Clinics, United States
Stony Brook University Medical Center, Stony Brook, United States
Sudbury Regional Hospital, Canada
Sullivan Nicolaides Pathology, Indooroopilly, Australia
Sydney Genetics, Sydney, Australia
Taipei Veterans General Hospital, Taipei, Taiwan, Province of China
Telemark Hospital, SKIEN, Norway
Texas Tech University, United States
The Chinese University of Hong Kong, Shatin, Hong Kong
The Kennedy Center, Copenhagen, Denmark
The Wellcome Trust Sanger Institute, United Kingdom
The Wessex Regional Genetics Laboratory, Salisbury United Kingdom
Thomas Jefferson University Hospital, Philadelphia, United States
Tokyo Womens Medical University, Tokyo, Japan
TOMA Advanced Biomedical Assays SpA, Busto Arsizio, Italy
TPMG Regional Cytogenetics Laboratory, Kaiser Permanente Northern California, United States
UCLA, Los Angeles, United States
UMass Memorial Memorial Center, United States
UMCG Groningen, Netherlands
UMDNJ NJ Medical School, Newark, United States
University Medical Center Ljubljana, Slovenia
University of Alabama , Birmingham, United States
University of British Columbia, Vancouver, Canada
University of Colorado Genetics Laboratory, Denver Colorado, United States
University of Florida, Gainesville, United States
University of Illinois, Urbana, IL, United States
University of Miami Pathology Cytogenetics Laboratory, Miami Florida, United States

University of Michigan, Michigan Medical Genetics Laboratories, United States
University of Minnesota, United States
University of Nebraska Medical Center, Human Genetics Laboratory, United States
University of Oxford, United Kingdom
University of Rochester, United States
University of Sao Paolo, Brazil
University of Sherbrooke, Children's Hospital, Canada
University of Wisconsin, United States
US Labs, Calabasas, United States
UT M D Anderson Cancer Center, Houston, United States
Vancouver Prostate Center, Vancouver, United States
Vejle Hospital, Vejle, Denmark
Washington University School of Medicine, St Louis, United States
Welgene Biotech, Taipei, Taiwan, Province of China
Wessex Regional Genetics Laboratory, Salisbury, United Kingdom
West Midlands Regional Genetics Laboratory, Birmingham, United Kingdom
Western Diagnostic Pathology, Perth, Australia
Yale Schoole of Medicine, New Haven, United States
Zentrum fuer Praenataldiagnostik, Berlin, Germany

Sequencing Laboratories Which Have Agreed to Submit Data

Ackerman Lab, Mayo
Alfred I Dupont Hospital for Children
All Children's Hospital St. Petersburg
Ambry Laboratories
ARUP
Athena Diagnostics
Baylor Medical Genetic Laboratories
Boston Children's Hospital
Boston University
Children's Hospital of Philadelphia
Children's Mercy Hospital, Kansas City
Cincinnati Children's Hospital
City of Hope Molecular Diagnostic Lab
CureCMD
Denver Genetic Laboratories
Detroit Medical Center
Emory University
Fullerton Genetics Laboratory
GeneDx
Cleveland Clinic
Greenwood Genetics
Harvard-Partners Lab for Molec. Medicine
Henry Ford Hospital
Huntington Medical Research Institutes
Indiana University/Perdue University
InSiGHT
Integrated Genetics
LabCorp/ Correlagen
Masonic Medical Research Laboratory
Mayo Clinic
Mt. Sinai School of Medicine
Nationwide Children's Hospital
Nemours Biomolecular Core, Jefferson Medical
Oregon Health Sciences University
Providence Sacred Heart Medical Center
Quest Diagnostics
SickKids Molecular Genetic Laboratory
Transgenomics
University of Chicago
University of Michigan
University of Nebraska Medical Center
University of Oklahoma
University of Penn
University of Sydney
University of Washington
Women and Children's Hospital
Wayne State University School of Medicine
Yale University