Standards for Clinical Grade Genomic Databases

Sophia L. Yohe, MD; Alexis B. Carter, MD; John D. Pfeifer, MD, PhD; James M. Crawford, MD, PhD; Allison Cushman-Vokoun, MD, PhD; Samuel Caughron, MD; Debra G. B. Leonard, MD, PhD

• ***Context.***—**Next-generation sequencing performed in a clinical environment must meet clinical standards, which requires reproducibility of all aspects of the testing. Clinical-grade genomic databases (CGGDs) are required to classify a variant and to assist in the professional interpretation of clinical next-generation sequencing. Applying quality laboratory standards to the reference databases used for sequence-variant interpretation presents a new challenge for validation and curation.**

Objectives.—**To define CGGD and the categories of information contained in CGGDs and to frame recommendations for the structure and use of these databases in clinical patient care.**

Design.—**Members of the College of American Pathologists Personalized Health Care Committee reviewed the literature and existing state of genomic databases and developed a framework for guiding CGGD development in the future.**

Results.—**Clinical-grade genomic databases may provide different types of information. This work group defined 3 layers of information in CGGDs: clinical genomic variant repositories, genomic medical data repositories, and genomic medicine evidence databases. The layers are differentiated by the types of genomic and medical information contained and the utility in assisting with clinical interpretation of genomic variants. Clinical-grade genomic databases must meet specific standards regarding submission, curation, and retrieval of data, as well as the maintenance of privacy and security.**

Conclusion.—**These organizing principles for CGGDs should serve as a foundation for future development of specific standards that support the use of such databases for patient care.**

(*Arch Pathol Lab Med.* 2015;139:1400–1412; doi: 10.5858/arpa.2014-0568-CP)

Next-generation sequencing (NGS) technology is now affordable for clinical laboratories, and many are implementing NGS tests for patient care. Because of major differences in how NGS data are produced and analyzed, compared with other laboratory tests, pathologists and other laboratory professionals are faced with a new set of challenges in analyzing, interpreting, and reporting NGS test results.[1,2] A fundamental requirement of clinical laboratory testing is reproducibility and accuracy of results within and among laboratories. Next-generation sequencing performed in a clinical environment must meet this same reproducibility standard for all aspects of testing, including generation of raw NGS sequence data, data analysis using multiple bioinformatics software packages to align sequence reads and to detect sequence variants (ie, bioinformatics pipelines), and final clinical interpretation. The final interpretation of the clinical relevance of a patient's NGS test results should be based on the highest possible levels of medical evidence according to standards for evaluating such evidence.[3] Similarly, use of standard reporting elements, including use of standard nomenclature to describe and categorize variants, is critical to avoiding clinically significant omissions, to eliminating confusion with other variants, and to ensuring data integrity.[4–6]

Unlike data produced in other areas of the laboratory, NGS data go through iterations of analysis using multiple software packages to transition from raw-sequence data to a final report. This so-called *bioinformatics pipeline* uses algorithms to align multiple copies of overlapping raw sequences to a human reference sequence and then uses other algorithms to detect where the patient's DNA differs from the reference sequence. These tools, if improperly designed or used, can introduce errors into the analysis.[1,7–10] One major problem with the use of multiple bioinformatics software packages is the differences in sensitivities and specificities for detection of different types of DNA sequence variants.[1,11,12] Types of sequence variants that can be detected by NGS are listed in Table 1.

Accepted for publication March 2, 2015.

Supplemental digital content is available for this article at www.archivesofpathology.org in the November 2015 table of contents.

From the Department of Laboratory Medicine and Pathology, University of Minnesota Medical Center, Minneapolis (Dr Yohe); the Department of Pathology and Laboratory Medicine and the Department of Biomedical Informatics, Emory University, Atlanta, Georgia (Dr Carter); the Department of Pathology, Washington University School of Medicine, St. Louis, Missouri (Dr Pfeifer); the Department of Pathology and Laboratory Medicine, Hofstra North Shore–Long Island Jewish School of Medicine, Hempstead, New York (Dr Crawford); the Department of Pathology and Microbiology, University of Nebraska Medical Center, Omaha (Dr Cushman-Vokoun); the MAWD Pathology Group, North Kansas City, Missouri (Dr Caughron); and the Department of Pathology and Laboratory Medicine, University of Vermont College of Medicine, Burlington (Dr Leonard).

The authors have no relevant financial interest in the products or companies described in this article.

Reprints: Sophia L. Yohe, MD, Department of Laboratory Medicine and Pathology, University of Minnesota Medical Center, MMC 609 Mayo Bldg, 420 Delaware St SE, Minneapolis, MN 55455 (e-mail: yohe0001@umn.edu).

Downloaded from http://meridian.allenpress.com/doi/pdf/10.5858/arpa.2014-0568-CP by guest on 03 December 2022

Table 1. Types of Sequence Variants That Can Be Detected by Next-Generation Sequencing	
Single-nucleotide variants	Substitution of a single base pair
Small insertions and deletions (indels)	Insertion or deletion of base pairs
Copy-number variants	More or fewer copies of a large region of DNA
Structural variants	Rearrangements, inversions, copy-number variants

Beyond the variability of the bioinformatics pipeline, the interpretation of the significance of a specific variant can be unique to the laboratory that performed the testing (ie, nonreproducible among laboratories). Variability in interpretation for sequence variants is due, in part, to the lack of professionally curated information to support clinical decision making, combined with the amount of information typically generated by such analyses. A single pathologist or other laboratory professional cannot understand the significance of all possible variants that could be generated without database support. Currently, investigation of multiple databases is required to assess the potential significance of even one sequence variant, and that is a cumbersome, time-consuming, and increasingly unfeasible process because the scope of NGS testing continues to increase in the clinical environment.[13–17] Adding to that complexity, not all databases contain accurate information, and a single database may have variability in the quality of its information for different variants. For example, laboratories that fail to follow strict quality control and quality assurance practices may submit inaccurate sequences to the databases and, therefore, may confound genotype-phenotype correlations and the interpretation of the clinical significance of specific variants.[18] "Clinical grade" databases—that is, the NGS results generated under clinical quality standards, which can be used to identify which variants infer risk of disease, to guide diagnosis, to predict prognosis, and/or to indicate a potential therapeutic target—are needed for broad and effective clinical use of NGS in clinical laboratories. Lack of clinical grade, evidence-based databases poses risks to patient care because use of less-than-sufficient or inaccurate evidence may lead to interpretation error and patient harm.

Several groups have looked at standards for genomic variation databases. However, those groups have not focused on the issues of reproducibility, quality, and clinical laboratory standards for the data being submitted, issues that are required for clinical patient care.[19–22] Standards for data submission, data curation, and data retrieval have been proposed, but those standards have generally focused on research use. Quality control by the laboratory generating and analyzing the data has received little attention in the literature to date, despite laboratories being required to ensure that the data are correct throughout the process, from specimen collection to data submission to later retrieval.[1,11]

To help address these issues, the College of American Pathologists (Northfield, Illinois) introduced an NGS testing section to the Molecular Pathology Checklist in the College of American Pathologists Laboratory Accreditation Program.[7] The College of American Pathologists developed NGS clinical laboratory accreditation requirements, and several groups are working on reference standards and proficiency testing materials. The US National Institute of Standards and Technology (Gaithersburg, Maryland)[23] is developing NGS reference standards. Horizon Discovery (Cambridge, United Kingdom), AcroMetrix (Life Technologies, Benicia, California), the College of American Pathologists, and others have developed or are developing proficiency testing modules or controls that will assess the reproducibility in variant detection among laboratories performing clinical NGS testing. These proficiency test materials will assess both the data-generation components of NGS tests and the bioinformatics pipelines that are used to align the sequence reads and to detect the sequence variants. To achieve a clinical grade database, data quality from the laboratory that initially reports a variant must be high and must follow standards for data submission, retrieval, and curation. Phenotype data also require standardization in terminology and completeness of the observations submitted.[24] Additionally, data security and individual privacy must be addressed for all aspects of the data submitted to the database (see the special section on "Data Security and Privacy" below).

Therefore, a work group of the College of American Pathologists Personalized Health Care Committee examined challenges specifically related to evidence-based resources available to assist with the interpretation of hereditary and acquired sequence variants in the clinical setting. Early in the process, the work group realized that the definition of a *clinical grade genomic database* (CGGD) needed to be standardized and categorized by function. This is because CGGDs may provide different types of information about and surrounding the variants they describe. The clinical utility of the database is driven, in part, by the type of information contained therein. Therefore, the purpose of this article is to describe the definition of a CGGD as well as the various categories of information that may be included within it. In addition, this article frames recommendations for the structure and use of such databases in the clinical patient-care setting. These organizing principles for CGGDs should serve as a foundation for future development of specific standards that support the use of such databases for patient care.

CLINICAL GRADE GENOMIC DATABASES

A CGGD is a clinical decision-support tool that can be used in the interpretation of human sequence variants for clinical use. Clinical decision-support tools provide evidence and support for decision making, but they do not mandate or require decisions. The final interpretation is dependent on the specific patient for whom testing was performed and the pathologist or other laboratory professionals examining the case, as well as clinical discussions with other health care providers. For a database to be used as a CGGD, the database *must* contain sequences and/or variants that have been produced from human samples in a laboratory that meets clinical quality standards for the analysis that generates the sequence and/or the variant (the so-called *high-quality human sequence/variant* [HQHSV]). In the United States, a laboratory certified under the Clinical Laboratory Improvement Amendments of 1988 (CLIA) and accredited by CLIA or a CLIA-deemed organization for high-complexity testing meets high clinical quality standards.[25] Similar standards in other countries may include the International Organization for Standardization (Geneva, Switzerland)[26] or the UK National External Quality Assess-

Downloaded from http://meridian.allenpress.com/doi/pdf/10.5858/arpa.2014-0568-CP by guest on 03 December 2022

Table 2. Layers of a Clinical Grade Genomic Database (CGGD)				
Layer No.	Term	Contains HQHSVs	Contains Clinical Information About HQHSVs	Contains Evidence-Based Clinical Significance for Variants
1	CGVR	Required	. . .	. . .
2	GMDR	Required	Required	. . .
3	GMED	Optional[a]	Optional[a]	Required

Abbreviations: CGVR, clinical genomic variant repository; GMDR, genomic medical data repository; GMED, genomic medical evidence database; HQHSVs, high-quality human sequence/variants.

[a] Although the presence of HQHSVs and their associated clinical information is not required for layer-3 GMEDs, sound evidence for the clinical significance of a variant can only be derived from the information contained in layer-1 CGVR and layer-2 GMDR databases. If the original data used to derive clinical significance are not present in a layer-3 database, the layer-3 GMED database must cite the source of the HQHSVs and the clinical information from which clinical significance was derived.

ment Service (Sheffield, United Kingdom).[27] A noncertified laboratory can meet the same quality standards but would need a mechanism to document and define their quality standards to ensure that their data are generated under the same quality standards as certified and accredited clinical laboratories. Regardless, applying those standards to reference databases for sequence variants presents a new challenge for validation and curation. If sequence data from laboratories not meeting these standards are included in a CGGD, then a mechanism for identifying the data generated under clinical standards must be available in the search criteria, as well as the determination of whether recommendations are made based on clinical quality data only or on all data.

To achieve practical utility, CGGDs should be easy to access. Given the cost to develop and maintain such databases, a fee may be required for access and use. The CGGDs may be developed by an institution or by a group of institutions, a commercial enterprise, the government, or through a public-private partnership. Currently, multiple databases are cross referenced for interpretation of NGS data. Use of multiple databases for interpretation of clinical case material presents certain challenges as well as opportunities. Even if a database contains HQHSVs and medical evidence of the highest quality, it is possible for different databases to have disparate information for the same variant. In addition, some databases may have sequences and variants from selected populations, which may not be representative of the reference population for the clinical case being interpreted. Understandably, international efforts at development and integration of CGGDs may yield the most-complete variant database from a "population" perspective.[28] However, although different databases present challenges, competition among databases may ultimately lead to better products. Some databases may focus on sequence and variants in particular areas of the genome or in reference to a particular disease or group of diseases. These smaller, less-comprehensive efforts may provide in-depth knowledge for the area of interest.

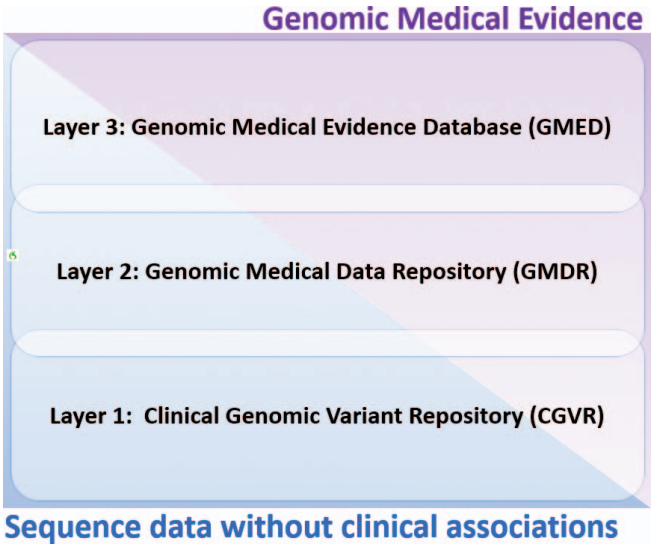

Figure 1. *Layers of a clinical grade genomic database (CGGD). Layer-1 information consists of sequence data only without associated clinical information or medical evidence; layers 2 and 3 contain progressively more clinical information and/or medical evidence.*

Therefore, in a CGGD, the breadth of the human genome covered (eg, genome, exome, gene, or single variant) may vary, provided that the database contains HQHSVs. In addition, the types of information available in a CGGD may also vary. Specifically, some databases are simply repositories of human sequence and variants, whereas other databases contain large amounts of sound medical evidence related to the clinical significance of variants. In discussing the latter, the work group identified 3 different classes of information for CGGDs. We chose to call these different classes "layers," because their value content is applied differently in interpretation of clinical case material.

Layers of CGGDs

The work group defined 3 layers of information for CGGDs: clinical genomic variant repositories (CGVRs), genomic medical data repositories (GMDRs), and genomic medicine evidence databases (GMEDs). The layers are differentiated by the types of genomic and medical information contained and the utility in assisting with clinical interpretation of genomic variants.

To be designated a CGGD, the database must contain information from at least one of the layers that meets the criteria described below. However, a single database may also include information from 2 or all 3 layers as well. In most database constructs, the types of information are additive in sequential order, but that may not always be the case because layer-3 GMED databases may contain little if any genomic sequence data. The names and the general proportion of sequence data are given in Table 2 and are shown graphically in Figure 1. The types of information needed for clinical interpretation are heavily dependent on the type of tissue being analyzed, the population being studied, and the clinical questions being asked. As such, CGGDs are intended to be high-quality clinical tools for assisting with interpretation of genomic test results in the context of these factors and the clinical judgment of the pathologist or other laboratory professional.

Downloaded from http://meridian.allenpress.com/doi/pdf/10.5858/arpa.2014-0568-CP by guest on 03 December 2022

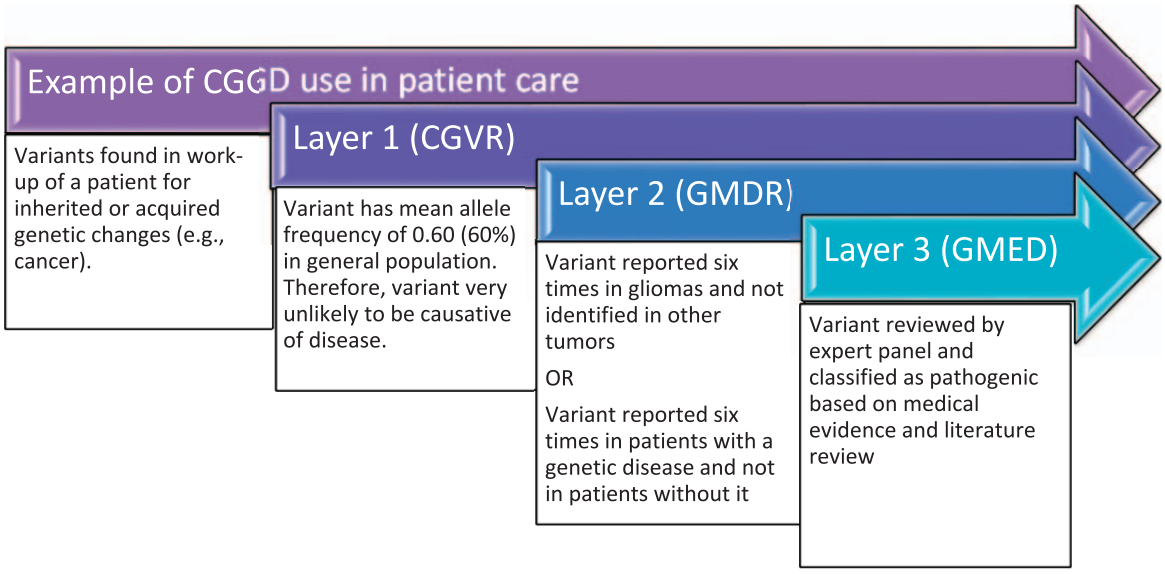

Figure 2. *Examples of how databases containing different layers of genomic information are used when evaluating individual variants. Abbreviations: CGGD, clinical grade genomic database; CGVR, clinical genomic variant repository; GMED, genomic medicine evidence database; GMDR, genomic medical data repository.*

Layer 1: CGVRs.—Clinical genomic variant repository databases are aggregations of sequence data, which may range from raw sequence reads to a list of identified sequence variants (compared with a reference sequence). The CGVRs must contain HQHSVs. The work group could not agree on whether CGVRs could contain nonhuman sequences or sequences not generated using clinical quality standards, so our compromise position was as follows: Some CGVRs may contain sequences from nonhumans or human sequences from noncertified laboratories, and there is no doubt that sequence information from nonhuman model systems (eg, yeast and mice) and from basic science investigations of human disease (eg, cell lines) can provide extremely useful information regarding the potential effect of a sequence change. Nonetheless, a CGVR *must* have functionality that allows end users to restrict their examination of sequences to HQHSVs only.

Although CGVRs are generally of limited use in clinical interpretation, they may be used to identify the frequency of a given variant in the human population (mean allele frequency) or to provide insight into the amount of data supporting identification of a novel and/or pathogenic sequence variant. The CGVRs that are restricted to sequence information do not have associated clinical information about the patients from whom the HQHSVs were derived, so the mean allele frequency may not be representative of the population of interest or as a whole. In addition, they do not have information on the clinical significance of the HQHSV. Similarly, CGVRs can be helpful in determining whether a variant has been previously described, but the clinical implications of the variant (benign versus pathogenic, hereditary versus acquired) can only be determined in the context of other information (as illustrated in Figure 2). Examples of existing databases that contain the types of information found in a layer-1 database include the 1000 Genomes Project and the Exome Sequencing Project (US National Heart, Lung, and Blood Institute, Bethesda, Maryland).[17,29,30] However, these databases lack assurance regarding the quality under which sequences were generated and, therefore, are not currently consistent with a CGGD.

Layer 2: GMDRs.—Genomic medical data repository databases are CGVRs plus associated clinical information of significance to genomic test interpretation. These include, but are not limited to, the race, ethnicity, phenotype, and diagnoses for the patient from whom the specimen was collected. In addition, information regarding the specimen source and associated findings (presence of tumor, percentage of tumor, tumor type, histologic grade, confirmed somatic, among others) is present. Clinical information may include treatment and outcome data, if available and submitted. Outcome data for an individual patient's sequence may be included in GMDRs, but that requires the information to be updated as the patient's condition and treatment changes for it to be useful. The GMDRs are useful for generating allele frequencies in specific populations (such as patients with breast cancer or hereditary ataxia) and across populations. However, GMDRs remain largely descriptive databases because synthesis of variant occurrence and outcomes across patients with the same variant or variants is not included (see Figure 2). An example database that includes the type of information contained in both layers 1 and 2 is the Catalogue of Somatic Mutations in Cancer (COSMIC; Wellcome Trust Sanger Institute, Hinxton, Cambridge, England) database, but this database currently lacks assurances regarding the quality under which the sequences were generated and the certainty of the associated findings and is, therefore, not consistent with a GMDR.[31,32]

Layer 3: GMEDs.—Genomic medicine evidence databases build on the variant and clinical information from the first 2 layers but are distinct because they only house information related to variants whose clinical significance has been proven in a sound and reproducible, scientific manner. These databases describe the classification, disease-association, or therapy association for a given variant, with citations of the evidence supporting that information. The content of layers 1

Downloaded from http://meridian.allenpress.com/doi/pdf/10.5858/arpa.2014-0568-CP by guest on 03 December 2022

Table 3. Submission Standards for Each Layer of a Clinical-Grade Genomic Database			
Submission Standards	**Layer 1**	**Layer 2**	**Layer 3**
Authorization	All required for layers 1, 2, 3		
Database grants user permission to submit data			
User must be authorized to disclose the data (see "Attestation")			
Selection	Required for layers 1, 2, 3		
Submitter selects data to upload			
Transmission	All required for layers 1, 2, 3		
Upload mechanism for large files			
Encrypt data during transmission			
Verification	All required for layers 1, 2, 3		
Check integrity of data after transmission to ensure that it is unaltered			
Check that data transmitted meets required formatting standards			
Check that all required annotation fields are completed (if annotation data are being uploaded rather than manually entered)			
Annotation	For details, see Supplemental Table[a]		
Patient metadata			
Specimen metadata			
Test performance metadata			
Sequence data			
Variant data			
Peer-reviewed medical evidence with references			
Added/corrected/removed information by the submitter			
Attestation	All required for layers 1, 2, 3		
Data accurately represents information gained from patient's sample			
Equipment and procedures used to generate the data adhere to the quality practices required of clinically certified laboratories			
Data were generated using the same process used for clinical patient care in a clinically certified laboratory			
Submitter has appropriate and documented patient consent where required for the data submitted			
Submitter has notified the database of the restrictions required by the patient on who can view the patient's data			
Notification	All required for layers 1, 2, 3		
Submitter is notified when changes to submitter's database entries are made			
Curators are notified when new or revised database entries are submitted			
Publication	Required for layers 1, 2, 3		
Database entries submitted by noncurators should be reviewed as part of the curation process before publication to end users (see "Curation")			

[a] See supplemental digital content file at www.archivesofpathology.org in the November 2015 table of contents.

and 2 serve as a foundation upon which the medical evidence for layer 3 is either derived or built. Such medical evidence may include functional genomic studies and clinical trials reports. Layer-3 databases may also include data on successful treatment of lesions associated with specific variants. Of all the layers of databases, a GMED is the most useful clinical decision-support tool for laboratory professionals and pathologists in the interpretation of NGS results (Figure 2). ClinVar (US National Center for Biotechnology Information, Bethesda, Maryland) is an example of a database that includes information from all 3 layers. However, ClinVar currently lacks assurances regarding the quality under which sequences were generated, and the standards used for the medical evidence of clinical usefulness are not yet defined.[13,33,34] Although it remains a useful tool, it cannot currently be classified as a GMED under the requirements listed above.

To different degrees, each database layer is useful for the interpretation of a sequence variant for a specific patient. However, all layers, even GMEDs, must be used in the context of the patient being examined, along with other evidence, professional standards, practice guidelines, and clinical judgment for accurate interpretation.

RECOMMENDED STANDARDS FOR EACH CGGD LAYER

Each layer of a CGGD must meet specific standards. Most standards will apply to all 3 layers of a CGGD, whereas other standards are unique to 1 or 2 layers. Standards are categorized by activity, and activities include submission, curation, and retrieval of data, as well as implementation and maintenance of privacy and security. Within the submission, curation, and retrieval activities, there are shared subcategories, which include authorization, selection, transmission, verification, annotation, attestation, notification, and publication. Implementation and maintenance of privacy and security are activities performed by database programmers and administrators rather than end users, so those subcategories do not apply.

Standards for submission must describe who is allowed to submit data to the CGGD, what data to submit, the format of the data, and the transfer of data, and they must ensure the accuracy of the data after transfer (Table 3). Standards for curating the database include the selection, organization, and maintenance of the data within the database over time, including how frequently data are reviewed and updated, and the tracking of that information (ie, versions of the database) (Table 4). Submissions should be held to a

Downloaded from http://meridian.allenpress.com/doi/pdf/10.5858/arpa.2014-0568-CP by guest on 03 December 2022

Table 4. Curation Standards for Each Layer of a Clinical-Grade Genomic Database			
Curation Standards	**Layer 1**	**Layer 2**	**Layer 3**
Authorization	Required for layers 1, 2, 3		
Database grants user permission to curate data			
Selection	All required for layers 1, 2, 3		
Database administration determines what data should be curated			
Schedules for curation are defined			
Transmission	N/A		
Not applicable for curation			
Verification	All required for layers 1, 2, 3		
Procedures exist for determining the accuracy of the data			
Procedures exist for determining the currency of the data			
Annotation	All required for layers 1, 2, 3 except as noted below		
For all records curated (with and without changes)			
Database records and displays date of review			
Database records and displays the curator who reviewed the information			
Database maintains an audit of all changes			
Database entries are versioned (each change or correction gets a new version number)			
Old versions of the data are maintained (but may not be viewable to standard end users)			
Database cites applicable peer-reviewed literature references for medical evidence regarding clinical significance of a variant	Optional	Optional	Required
For any record that is changed by a curator	All required for layers 1, 2, 3		
Database records a description of what changes were made and why			
Database cites applicable peer-reviewed literature references supporting the changes			
Attestation	All required for layers 1, 2, 3		
Submitter attests to the same items under "Submission" (Table 3)			
Curator attests that the changes made to the entry are accurate and current to the best of his or her knowledge			
Curator attests that changes are made according to the database's procedures and policies			
Notification	All required for layers 1, 2, 3		
Submitters are notified when curators update their database entries			
Curators are notified when a submitter changes his or her submitted data			
End users may wish to be notified when some or all database entries are modified			
Publication	All required for layers 1, 2, 3		
For curators: Publication takes place after changes are made			
For submitters: Publication takes place only after a curator has reviewed and approved the changes			

Abbreviation: N/A, not applicable.

scheduled review, with e-mail reminders or other forms of communication. Standards for data retrieval and use include determining who can access the database, the process for database queries, the tools for filtering query results, the format of the data for viewing, and the transfer of data (Table 5). Data security and privacy issues include patient consent, controls regarding access to data, and the security of the data at rest and during transmission.

The standards that are layer specific include how and what clinical or phenotypic data to submit (GMDRs and GMEDs), the definition of variant classification (GMEDs), and the levels of evidence used to determine classification, disease-association, or therapy association (GMEDs).

Recommended Standards Common to All CGGDs

The standards in this section are discussed in the context of a CGVR (layer 1) but are applicable to GMDRs (layer 2) and GMEDs (layer 3) as well. No standards are unique to CGVRs (layer 1).

Submission.—Standards for data submission include what data to submit, who is allowed to submit data, the quality of the data, the format of the data, and the transfer of data. Submitted data include the sequencing data and metadata. Sequencing data could include raw sequencing reads (eg, FASTQ-formatted files), sequence data after alignment, and quality metrics (eg, binary alignment map [BAM] files), or identified variants compared with a specified reference sequence of genome (eg, variant call format [VCF] files) (Table 6). If only identified variants are submitted, all variants identified (including benign variants, variants of unknown significance, and presumed pathogenic variants) should be included in the database so that population-based information is more readily available. The genome reference sequence must also be specified.

There are pros and cons to the different options above. For example, raw-sequencing reads require large amounts of storage but allow reanalysis by others using different bioinformatics pipelines. Because clinical laboratories do not have a gold standard for actual sequence, reanalysis allows research on the comparison of different bioinformatics pipelines for assessment of different types of variants.

Metadata provides information about one or more aspects of the data, such as who, what, when, where, and how data were generated and the quality standards used by the

Downloaded from http://meridian.allenpress.com/doi/pdf/10.5858/arpa.2014-0568-CP by guest on 03 December 2022

Table 5. Retrieval Standards for Each Layer of a Clinical-Grade Genomic Database			
Retrieval Standards	**Layer 1**	**Layer 2**	**Layer 3**
Authorization	All required for layers 1, 2, 3		
Database grants permission to viewer (end user) to see/retrieve data			
End user sees only data permitted by law (no consent required, consented without restrictions or consented for this individual/institution to see)			
Selection			
Queries allow end user to filter data to HQHSVs only	Required for layers 1, 2, 3		
Queries allow examination of allele frequencies for certain types of submitted data only (eg, sequences versus single variants)	Required for layers 1, 2, 3		
Queries filter on any required annotation element	Optional for layers 1, 2, 3		
Transmission	All optional for layers 1, 2, 3[a]		
Databases can transmit clinically actionable information on variants in a standard format back to a LIS or EHR (Note: National/international standards for such transmission have not yet been developed)			
Transmissions should include the database source, the version date of the information, the applicable references, among other information			
Verification	Optional for layers 1, 2, 3[a]		
The LIS or EHR receiving the transmission must have a mechanism to verify that the data downloaded from the database are accurately represented			
Annotation	All required for layers 1, 2, 3		
End users (viewers) must have a mechanism to notify the curators of a database of any potential inaccuracies in the data			
Database curators incorporate notifications regarding potential inaccuracies into their curation procedures			
Attestation	All required for layers 1, 2, 3		
The database has a statement that the officially published data are curated and accurate to the best of the curator's knowledge			
Each database entry must publish the date of its last curated review, the date of its last update, and the version number			
Notification			
End users can be notified when one, multiple, or all database entries are updated	Optional for layers 1, 2, 3		
Curators are notified when users submit potential inaccuracies in database entries (may be immediately notified or queued and batched)	Required for layers 1, 2, 3		
Publication	N/A		
Does not apply to this section			

Abbreviations: EHR, electronic health record; HQHSVs, high-quality human sequence and/or variants; LIS, laboratory information system, N/A, not applicable; Opt, optional.

[a] Currently optional but should be required once a transmission standard is developed.

laboratory generating the data. Metadata regarding the specimen, the laboratory testing process, and bioinformatics process should be included in all CGGDs. Metadata about the specimen may include specimen type (blood, prostate, kidney, lung, and other types), specimen preparation (fresh; frozen; formalin-fixed, paraffin-embedded, among others), the species from which the specimen was obtained, the time between collection and fixation or processing, and the percentage of reads with a variant. Metadata about the testing process may include extraction methods, library preparation methods, target-enrichment methodology (if used), instrument platform, instrument software version, quality scores, depth of coverage (total and variant), date of sequencing and analysis, and the submitting institution or laboratory identifier with contact information. Bioinformatics metadata includes the algorithm(s) used, the filters applied, and the reference genome. Each database will have to make decisions on who is allowed to submit data and exactly what data are submitted; that information should be transparent. Contact information for the submitter is useful if the users of the database need to contact the submitter with questions regarding specific data. Inclusion of data lacking one or more required annotations is not recommended because that would compromise the value of the database for clinical use.

Table 6. Type of Sequencing Data Files	
File Type	**Description**
FASTQ	Raw sequencing reads
BAM	Sequence data after alignment and quality metrics
VCF	Identified variants compared with a reference

Abbreviations: ASCII, American Standard Code for Information Interchange; BAM, binary alignment map; FASTQ, the Wellcome Trust Sanger Institute text-based format for storing biological sequences, particularly nucleotide sequences, with both the sequence letter and the quality score encoded with a single ASCII character for brevity; VCF, variant call format.

Different aspects of data quality and integrity to be considered in CGVRs include the technical quality, the variant call quality, and minimizing data input/transfer errors. Standardized data submission is important to ensure the quality of the data in the database and to avoid errors during the transfer of data. Standards for data quality should include not only the quality data for the sequencing itself but also some assurance that there has been quality control

Downloaded from http://meridian.allenpress.com/doi/pdf/10.5858/arpa.2014-0568-CP by guest on 03 December 2022

during the entire process from patient identification and specimen collection to DNA extraction to bioinformatics analysis. Performance of the testing in a high-complexity, clinically certified laboratory is one mechanism for ensuring minimal quality standards for the entire testing process. Laboratories without clinical accreditation/certification would need a process to demonstrate equivalent validation, quality management, and test-performance standards, with a mechanism for noting that in the database. Alternately, a database might include that information as a field that can be filtered (see section on "Data Retrieval").

Data should be in a standard format for submission to a database. Variant calls should use standard nomenclature, including the official Hugo Gene Nomenclature Committee (HGNC, European Bioinformatics Institute, Hinxton, Cambridge, England) gene names, standardized variant nomenclature (eg, http://www.hgvs.org/mutnomen/),[35] and unambiguous location coordinates (either chromosomal coordinates and/or complementary DNA [cDNA] coordinates with a reference transcript).[36,37] Such metadata should have a defined format.[38]

Use of automated data transfer is preferred to eliminate errors form manual entry. Automated data transfer will require standard messaging formats and interoperable communication systems. These standard messaging formats are still in their infancy and are not completely developed. However, some metadata may still need to be entered manually. Who submits or annotates those data and ensures the accuracy of the manually entered data fields is important. Trained annotators generally provide higher quality and more-uniform data with fewer mistakes,[39] and clinical laboratories that use manual entry most often have a second check of manually entered information before release to ensure accuracy and to reduce errors. Quick, easy, and automated data transfer will improve submission rates, although incentives for data submission may be required.[40]

Curation.—Curation of information for all CGGD databases ensures that information is complete and accurate at the time of submission. Curators also monitor changes to data and correct errors. Mechanisms for feedback to the database managers regarding errors or other problems identified by users accessing the database should be established. Given the size of these databases, some automation of this process will likely be required.[37] Lastly, updates need to be tracked, and each version should have a publication or certification date.

Data Retrieval.—Data retrieval should use standard formats. If databases contain data from nonhuman species, those data should be able to be filtered for human entries. Likewise, if a database contains data both from laboratories that have used stringent validation and quality practices (certification to perform high-complexity testing in a clinical patient-care environment) and from laboratories that have not, the database should allow an end user to filter the data to only HQHSVs. The ability to query by either gene or variant will be necessary. A Web-based interface would be optimal. Once relevant data are retrieved, the display and manipulation of the data will need to be user friendly and relatively intuitive. Tools for data analysis or manipulation may be part of the database, or the format of the retrieved data may be compatible with third-party tools. Having a standardized format that allows interaction with other databases or tools is also desirable.[20,39] The data should also be displayed in a format that is easy to interpret and to visualize. The database should track database queries, and access to that information by database users may be useful.

Data Security and Privacy.—Data security and privacy are applicable to all 3 layers of CGGDs. There are no security or privacy standards that are unique to any of the layers; therefore, security and privacy will only be discussed in this section.

By definition, CGGDs are repositories of human genetic information. During the past 10 years, the US Federal Government has enacted increasingly protective laws for the privacy and security of health information, especially genetic information. A list of the major US federal statutes regarding privacy and security of health information is provided in Table 7, along with a brief description of the pertinent information related to genetic data. Many current databases containing genetic/genomic information were developed without realizing the requirements for compliance with federal law in the United States or other countries. The collection, manipulation, management, storage, and retrieval of sensitive and individually identifiable health information must comply with applicable privacy laws at the national, regional, and local levels. Given the cloud-based nature of many of these databases, both users and managers of the database must be aware of the country in which the data are housed and ensure that its location is compliant with applicable regulations. For example, if protected health information from US patients is stored in non-US databases, those databases have no requirement to comply with US privacy and security laws. In that situation, the individual or entity who submits the data to a noncompliant database may be legally liable for security breaches under the US Health Information Technology for Economic and Clinical Health Act of 2009 (Title XIII of Pub L 111-5) and the US Health Insurance Portability and Accountability Act of 1996 (HIPAA, Pub L 104–191). Conversely, if sensitive and protected information on a patient from a non-US country is stored in a database located within the United States, those data are subject to the US Patriot Act of 2001 (Pub L 107–56).[41] Access to those data by representatives of the US federal government may violate the privacy laws of the country in which the specimen was collected and analyzed.

Human genetic information as defined by the Genetic Information Nondiscrimination Act of 2008 (GINA, Pub L 110–233; see Table 7) should only be submitted to a CGGD if the laboratory has informed consent from the patient. Although it is certainly true that short sequences containing commonly identified variants are not individually identifiable, those sequences and variants are included in the definition of genetic information under GINA. The HIPAA Omnibus Rule (effective date: September 23, 2013) specifically references the definition of genetic information under GINA when it updated the definition of protected health information to include genetic information. No clarification on this significant conundrum for both patient care and research has been provided by the federal government, and laboratories should consult with an attorney before submitting any sequence or other data from a human patient to a database without informed consent. To support the requirement for informed consent before submission of genetic information, a CGGD must have a place for a laboratory to attest that the laboratory has obtained informed consent from the patient. This work group does not recommend the actual consent be uploaded to the CGGD because that would release unnecessary additional patient identifiers

Downloaded from http://meridian.allenpress.com/doi/pdf/10.5858/arpa.2014-0568-CP by guest on 03 December 2022

Table 7. Summary of US Laws Relating to Genetic Health Information

Source, y	Definitions
HIPAA,[58] 1996	**Individually identifiable health information** (from HIPAA) is: Information, including demographic data, that relates to • The individual's past, present, or future physical or mental health or condition, • The provision of health care to the individual, or • The past, present, or future payment for the provision of health care to the individual, and that identifies the individual or for which there is a reasonable basis to believe can be used to identify the individual. Individually identifiable health information includes many common identifiers (eg, name, address, birth date, Social Security No.).
	PHI The PHI is individually identifiable health information in any form or media, whether electronic, paper, or oral, with certain exclusions, including employment records, which a covered entity maintains in its capacity as an employer, and education and certain other records subject to, or defined in, the Family Educational Rights and Privacy Act (20 USC §1232g).
	Deidentified Health Information There are no restrictions on the use or disclosure of deidentified health information. Deidentified health information neither identifies nor provides a reasonable basis to identify an individual. There are 2 ways to deidentify information, either (1) a formal determination by a qualified statistician; or (2) the removal of specified identifiers of the individual and of the individual's relatives, household members, and employers is required and is adequate only if the covered entity has no actual knowledge that the remaining information could be used to identify the individual.
HIPAA final security rule,[42] 2003	**ePHI** Information that comes within paragraphs (1)(i) or (1)(ii) of the definition of protected health information as specified in this section.
	Electronic Media 1. Electronic storage media, including memory devices in computers (hard drives) and any removable/transportable digital memory medium, such as magnetic tape or disk, optical disk, or digital memory card; or 2. Transmission media used to exchange information already in electronic storage media. Transmission media include, for example, the Internet (wide open), an extranet (using Internet technology to link a business with information accessible only to collaborating parties), leased lines, dial-up lines, private networks, and the physical movement of removable/transportable electronic storage media. Certain transmissions, including paper via facsimile and voice via telephone, are not considered transmissions via electronic media because the information being exchanged did not exist in electronic form before the transmission.
GINA,[59] 2008	**Genetic information** includes: a. An individual's genetic tests; b. The genetic tests of that individual's family members; c. The manifestation of disease or disorder in family members of the individual (family medical history); d. An individual's request for, or receipt of, genetic services, or the participation in clinical research that includes genetic services, by the individual or a family member of the individual; or e. The genetic information of a fetus carried by an individual or by a pregnant woman who is a family member of the individual and the genetic information of any embryo legally held by the individual or family member using an assisted reproductive technology. Genetic information does **not** include information about the sex or age of the individual, the sex or age of family members, or information about the race or ethnicity of the individual or family members that is not derived from a genetic test.
HIPAA omnibus rule,[60] 2013	**Revised Definition of Health Information Under HIPAA to:** Health information means information, **including genetic information**, whether oral or recorded in any form or medium.
	Genetic information means: 1. Subject to paragraphs 2 and 3 of this definition, with respect to an individual, information about a. The individual's genetic tests; b. The genetic tests of family members of the individual; c. The manifestation of a disease or disorder in family members of such individual; or d. Any request for, or receipt of, genetic services or participation in clinical research, which includes genetic services, by the individual or any family member of the individual. 2. Any reference in this subchapter to genetic information concerning an individual or family member of an individual shall include the genetic information of a. A fetus carried by the individual or family member, who is a pregnant woman; and b. Any embryo legally held by an individual or family member using an assisted reproductive technology. 3. Genetic information **excludes** information about the sex or age of any individual.
	Genetic test means An analysis of human DNA, RNA, chromosomes, proteins, or metabolites, if the analysis detects genotypes, mutations, or chromosomal changes. **Genetic test does not include** an analysis of proteins or metabolites that is directly related to a manifested disease, disorder, or pathologic condition.

Downloaded from http://meridian.allenpress.com/doi/pdf/10.5858/arpa.2014-0568-CP by guest on 03 December 2022

Table 7. Continued	
Source, y	**Definitions**
HIPAA omnibus rule,[60] 2013	**Unsecured PHI** Unsecured PHI means PHI that is not rendered unusable, unreadable, or indecipherable to unauthorized persons through the use of a technology or methodology specified by the Secretary in the guidance issued under section 13402(h)(2) of Public Law 111–5.

Abbreviations: ePHI, electronic protected health information; GINA, Genetic Information Nondiscrimination Act; HIPAA, Health Insurance Portability and Accountability Act; PHI, protected health information.

(name, date of birth, and other such information) and violate the minimum necessary rule under HIPAA.

In addition to informed consent, many other privacy and security measures exist and require compliance by the managers of a CGGD. In the United States, the CGGD must comply with the Final Security Rule of HIPAA. For an extensive discussion of the requirements of the Final Security Rule of HIPAA, the reader is referred to several references.[42,43] This rule requires administrative (assessments, policies, and procedures), physical (hardware), and technical (software) safeguards for protected health information. Databases in the cloud that contain protected health information (including genetic information as defined by GINA) must satisfy the same requirements as data that are not in the cloud, specifically regarding the HIPAA Final Security Rule. Many cloud services, for example, state that they are compliant with HIPAA, but that compliance typically only extends to some of the administrative and most, if not all, of the physical safeguards. Those cloud services cannot comply with the technical safeguards if they do not support or maintain the actual software being used by the end user. Software used to deploy CGGDs must enable controls for person authentication, access, auditing, data integrity, and transmission security. Person authentication means that the CGGD has confirmed the identity of the user who desires access to the database before granting logon. Access controls require that the individual user have a unique user identifier and password as well as automatic log offs. Authorization of the individual to view the data has been previously described in the submission, curation, and retrieval activities but is also included among the various privacy and security measures that must be implemented. Audit controls must be enabled so that events in which software users who view or alter an individual patient's protected health information (in the case of a CGGD, genetic or medical information) are recorded. A quality program must ensure that the integrity of the data is not compromised. Data encryption is recommended, although not mandated, during transmission to and from the database, as well as within the database itself (at rest). Entities in the United States that meet or exceed the encryption standards set forth by federal guidelines both at rest[44] and during transmission[45–47] are obviated from complying with the Security Breach Notification Rule[48,49] in the event of breach, theft, or loss of data. Before submission of genetic information to any database, documented privacy and security practices should be reviewed by a knowledgeable individual to determine whether the database is qualified to house such information.

Recommended Standards for Layer-2 GMDRs

Submission.—The submission of clinical and/or phenotypic data for a GMDR (layer 2) creates opportunity for understanding clinical findings based on molecular information. Required information depends on whether the database is to be used primarily for cancer, infectious disease, or constitutional changes, including inherited conditions and polymorphisms that may affect drug therapy (ie, pharmacogenetics). Important information for a cancer database includes certain patient demographics (eg, age, race, ethnicity), tumor type and stage, previous therapeutic intervention, and response to therapy (either before or after performance of cancer genomics). Important information for an inherited-disease database includes patient demographics, clinical phenotype, family history and pedigree, segregation studies (if available), and, in some cases, clinical laboratory and/or radiologic studies. Demographic information such as race, age, and gender is essential for both cancer and inherited-disease databases. Phenotypic terminology and formatting should be standardized for databases.[36] In the case of cancer, the submitter should be required to state whether the corresponding healthy and tumor tissue were tested. Metadata should include the percentage of tumor in the analyzed specimen (taking into account any microdissection that may have taken place). Some tools and standardized formats have been developed by the International Standards for Cytogenomic Arrays (National Center for Biotechnology Information) Consortium for their chromosome microarray database.[24] Systematized Nomenclature of Medicine–Clinical Terms (SNOMED-CT) codes or International Classification of Disease (ICD)-9/ICD-10 codes could be considered for standardizing disease information because they are widely used, usually available, and transfer of data could potentially be automated; however, they have their limitations mostly related to a lack of completeness of the clinical description and accuracy.[50]

Curation.—Curation of GMDRs (layer 2) includes points discussed for CGVRs (layer 1) above, plus mechanisms to ensure the integrity of the clinical information that has been submitted. Some of the clinical information may not be available at initial sequence-data submission, so mechanisms that allow additional information, such as evolution of symptoms or clinical outcomes, to be added at a later time without compromising patient confidentiality are critically important to the utility of a CGVR. In the case of constitutional (heritable disorder) testing, follow-up information from biological relatives is also essential. Furthermore, incentives for submitters to provide follow-up data may increase submissions and the completeness of the clinical data, and therefore, the clinical usefulness of the CGGD.[40]

Retrieval.—The ability to query by phenotype (eg, tissue of origin, tumor type and subtype, clinical signs or symptoms, among others) and gene or mutation will be necessary. Standardization of entry (eg, cancer type) is imperative to facilitate data retrieval. The same data retrieval

Downloaded from http://meridian.allenpress.com/doi/pdf/10.5858/arpa.2014-0568-CP by guest on 03 December 2022

considerations for CGVRs (layer 1) also apply to GMDRs (layer 2).

Recommended Standards Unique to Layer-3 GMEDs

Submission.—Submission of data to layer-3 databases includes submitting data that specifically supports the biological classification of a particular variant.[51] Although layer 1 and layer 2 databases must contain sequences and/or variants derived from human samples, layer 3 databases may contain medical evidence pertaining only to previously identified variants. Such medical evidence may include the source of peer-review information, the quality and extent of the evidence supporting a clinical association or causal clinical linkage, reference to causal information (eg, pharmacogenomic studies), or follow-up from an author of a published study regarding RNA or protein studies, segregation analysis, or outcome data after treatment. Evidence could also be based on the analysis of data within a CGGD, with comments and edits from users of the database, as long as the lack of peer review is a searchable field. Standard classifications of the clinical significance of variants with definitions of evidence levels for each must be used for a GMED. Incentives for submitters to provide additional data may be necessary.[40]

Curation.—Although curation is important for all 3 layers, it is most critical for a GMED (layer 3). Curation of a GMED has unique issues, including the definition of clinical relevance, the levels of evidence required for classification of a variant, the frequency at which data are reviewed and updated, and the tracking of database modifications. The definition of clinical relevance of a variant and levels of evidence may differ for inherited diseases and cancer. For inherited diseases, a variant that is demonstrated to be disease-causing in some patients, may not have complete penetrance or may not cause disease in other patients or family members, with the hypothesis that the effect of the variant is modified by other variants in other genes/proteins. For cancer, a clinically relevant variant may not be disease causing but may determine diagnosis, classification, prognosis, or therapeutic effectiveness. Standards for determining the reliability and significance of information in the literature are important. Evidence synthesis and review must adhere to evidentiary standards for determining the quality of a specific article, such as those developed the Evaluation of Genomic Applications in Practice and Prevention (EGAPP) group.[52] Such evidence syntheses must account for the different significances of a variant in different disease processes or tissues. For example, a *BRAF* c.1799T>A, p.Val600Glu (V600E) mutation in melanoma has therapeutic implications, but the therapeutic implications for the same mutation in colorectal carcinoma or hairy cell leukemia is not known at this time.

Standards for interpreting and reporting variants as deleterious, benign, or uncertain in inherited diseases have been recommended by the American College of Medical Genetics (Bethesda, Maryland); however, the amount of evidence required to determine whether a variant is benign or pathogenic is less clear.[5] *Levels of evidence* refers to a ranking system describing the strength of results measured in a research study or clinical trial. Examples of levels of evidence are provided by the University of Oxford Centre for Evidence-Based Medicine (Oxford, England).[53] However, those levels of evidence may not translate well for rare diseases where the only studies are case reports or small case series. Inherited diseases may require their own levels of evidence that include functional studies of an affected protein. Possible sources of evidence include patterns emerging from GMDRs (layer 2), functional studies, clinical trials, publications, guidelines, or in silico predictions.

Levels of evidence for reporting a variant in cancer are more complex. A variant in cancer can have different clinical uses (diagnosis, prognosis, therapy selection, or genetic predisposition for cancer). Each of those would have its own level of evidence for the variant classification. Furthermore, evidence for the significance of a particular variant may be excellent in one tissue or tumor type but not in others. It is challenging to have adequate studies that cover every possible tissue/tumor type, especially for less-common tumors. The National Comprehensive Cancer Network (Fort Washington, Pennsylvania), for example, has a categorization based on the degree of evidence, as well as expert consensus.[54]

As our understanding of the clinical significance of human genetic variation in health care increases, the classification of a variant may change over time. A variant that has unknown significance on submission to a CGGD may be proven later to be benign or pathogenic. In addition, the development of new, targeted therapies may make a variant important for therapeutic reasons. As medical knowledge evolves, GMEDs (layer 3) will need to be updated. The date of the update, the version of the data, and the data content of that version need to be tracked and be referenced within each query or use of the database. In addition, notification of updates or changes in classification for a given variant could be considered but would be logistically challenging.

Retrieval.—Two unique issues should be noted regarding data retrieval from GMEDs (layer 3). The first is to automate correlation of the sequence information in the database with data from the laboratory information system or electronic health record. The second issue is to provide a mechanism for reviewing and assessing the relevance of the evidence supporting a given variant classification. If a database will be used to provide variant classification of sequencing results from an individual patient, automated retrieval of information for all variants detected in a patient is required. Ideally, a system will be in place so that the patient file automatically queries the database for all variants and receives the classification results. These results should easily link to the evidence supporting the interpretation, such as a link to references or specific databases.

Although the primary use of a CGGD is to guide care of individual patients, a further consideration in data retrieval from GMEDs (layer 3) is whether the patient's clinical phenotype should be layered across the full database of patient phenotypes. In essence, a GMED (layer 3) should provide functionality for data retrieval based on (1) individual patient demographics (race, sex, age), (2) genotypic findings, and/or (3) clinical phenotypic findings. Such analytics may help individual health care providers gain insight into the potential relevance of their patient's identified sequence variants related to specific phenotypic characteristics and therapeutic responses and may further the process of discovery as well as provide new differential diagnoses for specific patients. Whether nonhuman data should be included in GMEDs (layer 3) is problematic. We recognize that some novel variants will lack any human data regarding classification and that, in those cases, data from animal models, conservation across species, and/or cell-line data may be used to interpret whether a novel variant might be responsible for disease and warrant further consider-

Downloaded from http://meridian.allenpress.com/doi/pdf/10.5858/arpa.2014-0568-CP by guest on 03 December 2022

ation. However, those data cannot be considered clinical grade and, therefore, require both a notation in the database as to the limitation of the sources used for classification and a notation in the patient report about the limits of the data used for interpretation. This issue is related to the levels of evidence supporting variant classification, as discussed above.

SUMMARY

Discussions about clinical-grade genomic databases will have different emphases depending on what layers of information are included in a given database. We have identified 3 layers of information that may be contained in a database: the sequence data (CGVR, layer 1), the clinical and phenotypic information (GMDR, layer 2), and the classification or association information (GMED, layer 3). The composition of a given database will affect the structure, function, and clinical usefulness of the database. Information from CGVRs and GMDRs (layers 1 and 2, respectively) can be used to expand our knowledge of the significance of different variants and may identify needed scientific functional studies to determine the clinical relevance of specific variants or genes in a disease, as well as allow compilation of the rarer occurrences of variants in a specific disease across multiple testing sites to begin to identify new variant-disease associations requiring additional study. The GMEDs (layer 3) will be useful for the interpretation of patient NGS test results.

Regardless of the composition of a database, if it will be used for clinical care, all data must be high quality. Quality is important at all levels, including how the data are produced, transmitted, organized, retrieved, filtered, and interpreted. In addition, a database must be useable, be kept up to date, allow data to be added later, and still provide data-security measures to protect patient confidentiality. Documentation of all procedures and versions of the database as updates and submissions occur are essential. Databases composed of quality clinical-grade genomic data will be important for the advancement of our understanding of the clinical significance of genetic variants and for more reliable and reproducible interpretation of patient NGS test results.

The goal of this article is to provide an overview of the issues and standards to be considered when creating a CGGD. National and international efforts are underway to create comprehensive genomic databases, such as ClinGen and the Human Variome Project International (Victoria Australia).[55,56] As these databases will contain information obtained from clinical laboratories and/or be used for patient care, we encourage consideration of these standards when making decisions about the structure and function of the databases as CGGDs. We have not defined the specific requirements for these standards. Those specific requirements, such as the level of evidence required to classify a variant, will require input from multiple stakeholders. Although creation of a GCCD is a daunting task, it is achievable. Clinical-grade databases exist for other types of genetic data, for example, the Database of Genomic Variation and Phenotype in Humans Using Ensembl Resources (DECIPHER, Wellcome Trust Sanger Institute) for copy number variation, translocations, and inversions associated with inherited syndromes.[57] That project has tackled some of the issues discussed in this article, including patient privacy, data security, and curation issues, such as updating clinical information. A CGGD for sequence variants can be built using the lessons learned by the databases that have come before.

References

1. Guo Y, Ye F, Sheng Q, Clark T, Samuels DC. Three-stage quality control strategies for DNA re-sequencing data. *Brief Bioinform*. 2014;15(6):879–889.
2. Mardis ER. The $1,000 genome, the $100,000 analysis? *Genome Med*. 2010;2(11):84.
3. Stanley CM, Sunyaev SR, Greenblatt MS, Oetting WS. Clinically relevant variants—identifying, collecting, interpreting, and disseminating: the 2013 annual scientific meeting of the Human Genome Variation Society. *Hum Mutat*. 2014;35(4):505–510.
4. Ogino S, Gulley ML, den Dunnen JT, Wilson RB. Standard mutation nomenclature in molecular diagnostics: practical and educational challenges. *J Mol Diagn*. 2007;9(1):1–6.
5. Richards CS, Bale S, Bellissimo DB, et al; Molecular Subcommittee of the ACMG Laboratory Quality Assurance Committee. ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. *Genet Med*. 2008;10(4):294–300.
6. Gulley ML, Braziel RM, Halling KC, et al; Molecular Pathology Resource Committee, College of American Pathologists. Clinical laboratory reports in molecular pathology. *Arch Pathol Lab Med*. 2007;131(6):852–863.
7. Commission on Laboratory Accreditation. Laboratory Accreditation Program: Molecular Pathology Checklist. Northfield, IL: College of American Pathologists; 2013:72.
8. Cottrell CE, Al-Kateb H, Bredemeyer AJ, et al. Validation of a next-generation sequencing assay for clinical molecular oncology. *J Mol Diagn*. 2014; 16(1):89–105.
9. Gargis AS, Kalman L, Berry MW, et al. Assuring the quality of next-generation sequencing in clinical laboratory practice. *Nat Biotechnol*. 2012; 30(11):1033–1036.
10. Pritchard CC, Salipante SJ, Koehler K, et al. Validation and implementation of targeted capture and sequencing for the detection of actionable mutation, copy number variation, and gene rearrangement in clinical cancer specimens. *J Mol Diagn*. 2014;16(1):56–67.
11. Spencer DH, Tyagi M, Vallania F, et al. Performance of common analysis methods for detecting low-frequency single nucleotide variants in targeted next-generation sequence data. *J Mol Diagn*. 2014;16(1):75–88.
12. Liu X, Han S, Wang Z, Gelernter J, Yang BZ. Variant callers for next-generation sequencing data: a comparison study. *PloS one*. 2013;8(9):e75619. doi:10.1371/journal.pone.0075619.
13. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2014;42(database issue):D7–D17. doi:10.1093/nar/gkt1146.
14. Fernández-Suárez XM, Galperin MY. The 2013 Nucleic Acids Research Database Issue and the online molecular biology database collection. *Nucleic Acids Res*. 2013;41(database issue):D1–D7. doi:10.1093/nar/gks1297.
15. Fernández-Suárez XM, Rigden DJ, Galperin MY. The 2014 Nucleic Acids Research Database Issue and an updated NAR online Molecular Biology Database Collection. *Nucleic Acids Res*. 2014;42(database issue):D1–D6. doi: 10.1093/nar/gkt1282.
16. Küntzer J, Eggle D, Klostermann S, Burtscher H. Human variation databases. *Database (Oxford)*. 2010;2010:baq015. doi:10.1093/database/baq015.
17. Abecasis GR, Auton A, Brooks LD, et al; 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491(7422):56–65.
18. Soussi T. Locus-specific databases in cancer—what future in a post-genomic era?: the TP53 LSDB paradigm. *Hum Mutat*. 2014;35(6):643–653.
19. AlAama J, Smith TD, Lo A, et al. Initiating a human variome project country node. *Hum Mutat*. 2011;32(5):501–506.
20. Chervitz SA, Deutsch EW, Field D, et al. Data standards for Omics data: the basis of data sharing and reuse. *Methods Mol Biol*. 2011;719:31–69.
21. Oetting WS. Impact of next generation sequencing: the 2009 Human Genome Variation Society scientific meeting. *Hum Mutat*. 2010;31(4):500–503.
22. Ramos EM, Din-Lovinescu C, Berg JS, et al. Characterizing genetic variants for clinical action. *Am J Med Genet C Semin Med Genet*. 2014;166C(1):93–104. doi:10.1002/ajmg.c.31386.
23. National Institute for Standards and Technology. The NIST Web site: 2014. http://www.nist.gov/. Accessed August 5, 2014.
24. Riggs ER, Jackson L, Miller DT, Van Vooren S. Phenotypic information in genomic variant databases enhances clinical care and research: the International Standards for Cytogenomic Arrays Consortium experience. *Hum Mutat*. 2012; 33(5):787–796.
25. Centers for Medicare & Medicaid Services, HHS. Clinical laboratory improvement amendments (CLIA) of 1988—Laboratory Requirements, *Fed Regist*. 1992;57(40):7146. Codified at 42 CFR §493. http://www.gpo.gov/fdsys/pkg/CFR-2012-title42-vol5/pdf/CFR-2012-title42-vol5-part493.pdf. Accessed February 13, 2015.
26. International Organization for Standardization. ISO Web site: 2014. http://www.iso.org/iso/home.htm. Accessed August 5, 2014.
27. UK National External Quality Assessment Service. UK NEQAS Web site: 2014; http://www.ukneqas.org.uk/content/Pageserver.asp. Accessed August 5, 2014.

Downloaded from http://meridian.allenpress.com/doi/pdf/10.5858/arpa.2014-0568-CP by guest on 03 December 2022

28. Auerbach AD, Burn J, Cassiman JJ, et al. Mutation (variation) databases and registries: a rationale for coordination of efforts. *Nat Rev Genet*. 2011;12(12):881; discussion 881.

29. 1000 Genomes. 100 genomes Web site: 2014. http://www.1000genomes.org/. Accessed August 11, 2014.

30. National Heart, Lung, and Blood Institute. NHLBI Grand Opportunity Exome Sequencing Project (ESP) Web site. Accessed August 11, 2014. https://esp.gs.washington.edu/drupal/. Accessed August 11, 2014.

31. Wellcome Trust Sanger Institute. Catalogue of Somatic Mutations in Cancer (COSMIC): 2014. http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/. Accessed August 5, 2014.

32. Bamford S, Dawson E, Forbes S, et al. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br J Cancer*. 2004;91(2):355–358.

33. National Center for Biotechnology Information. ClinVar: 2014. http://www.ncbi.nlm.nih.gov/clinvar/. Accessed August 5, 2014.

34. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res*. 2014;42(database issue):D980–D985. doi:10.1093/nar/gkt1113.

35. Human Genome Variation Society. HGVS Web page: 2010. http://www.hgvs.org/. Accessed September 19, 2010.

36. Byrne M, Fokkema IF, Lancaster O, et al. VarioML framework for comprehensive variation data representation and exchange. *BMC Bioinformatics*. 2012;13:254. doi:10.1186/1471-2105-13-254.

37. Tong MY, Cassa CA, Kohane IS. Automated validation of genetic variants from large databases: ensuring that variant references refer to the same genomic locations. *Bioinformatics*. 2011;27(6):891–893.

38. Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet*. 2014;133(1):1–9.

39. Pool R, Esnayra J. Bioinformatics: Converting Data to Knowledge: Workshop Summary Washington, DC: National Academy Press; 2000: http://www.ncbi.nlm.nih.gov/books/NBK44940/. Accessed February 13, 2015.

40. Patrinos GP, Cooper DN, van Mulligen E, et al. Microattribution and nanopublication as means to incentivize the placement of human genome variation data into the public domain. *Hum Mutat*. 2012;33(11):1503–1512.

41. Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism (USA PATRIOT Act) Act of 2001, Pub L No. 107–156. 115 Stat. 272 (2001). http://www.gpo.gov/fdsys/pkg/PLAW-107publ56/pdf/PLAW-107publ56.pdf. Accessed February 13, 2015

42. Health Insurance Reform: Security Standards (Final Security Rule), 45 CFR §§ 160, 162, 164, http://www.hhs.gov/ocr/privacy/hipaa/administrative/securityrule/securityrulepdf.pdf. Accessed February 23, 2015.

43. Summary of the HIPAA Security Rule. 2014; http://www.hhs.gov/ocr/privacy/hipaa/understanding/srsummary.html. Accessed June 15, 2014.

44. Scarfone K, Souppaya M, Sexton M. *Guide to Storage Encryption Technologies for End User Devices: Recommendations of the National Institute of Standards and Technology*. In: Technology NIoSa, ed. Vol Special Publication 800-111. Gaithersburg, MD: US Department of Commerce; 2007:1–40.

45. Frankel S, Hoffman P, Orebaugh AD, Park R. *Guide to SSL VPNs: Recommendations of the National Institute of Standards and Technology*. In: Technology NIoSa, ed. Vol Special Publication 800-113. Gaithersburg, MD: U.S. Department of Commerce; 2008:1–87.

46. Frankel S, Kent K, Lewkowski R, Orebaugh AD, Ritchey RW, Sharma SR. *Guide to IPsec VPNs: Recommendations of the National Institute of Standards and Technology*. Gaithersburg, MD: US Department of Commerce; 2005:1–126. *Vol Special Publication 800-77*.

47. Polk T, McKay K, Chokhani S. *Guidelines for the Selection, Configuration, and Use of Transport Layer Security (TLS) Implementations*. Gaithersburg, MD: US Department of Commerce; 2014:1–67. *Vol Special Publication 800-52 Revision 1*.

48. Health Information Technology for Economic and Clinical Health (HITECH) Act, Title XIII of Division A and Title IV of Division B of the American Recovery and Reinvestment Act of 2009 (ARRA), Pub. L. No. 111-5, 123 Stat. 226 (Feb. 17, 2009), codified at 42 U.S.C. §§300jj et seq.; §§17901 et seq., http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/hitechact.pdf. Accessed February 13, 2015.

49. Health and Human Services. HHS strengthens HIPAA enforcement. http://www.hipaa.com/hhs-strengthens-hipaa-enforcement/. Posted November 3, 2009. Accessed April 15, 2015.

50. Pathak J, Wang J, Kashyap S, et al. Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE Network experience. *J Am Med Inform Assoc*. 2011;18(4):376–386.

51. MacArthur DG, Manolio TA, Dimmock DP, et al. Guidelines for investigating causality of sequence variants in human disease. *Nature*. 2014;508(7497):469–476.

52. Teutsch SM, Bradley LA, Palomaki GE, et al; EGAPP Working Group. The Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Initiative: methods of the EGAPP Working Group. *Genet Med*. 2009;11(1):3–14.

53. University of Oxford. Centre for Evidence-Based Medicine—levels of evidence (March 2009). http://www.cebm.net/oxford-centre-evidence-based-medicine-levels-evidence-march-2009/. Accessed August 15, 2014.

54. National Comprehensive Cancer Network. NCCN categories of evidence and consensus. http://www.nccn.org/professionals/physician_gls/categories_of_consensus.asp. Accessed August 15, 2014.

55. Clinical Genome Resources. ClinGen Resource. http://www.clinicalgenome.org/. Accessed April 15, 2015.

56. Human Variome Project International. The Human Variome Project Web site. http://www.humanvariomeproject.org/. Accessed August 15, 2014.

57. Firth HV, Richards SM, Bevan AP, et al. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am J Hum Genet*. 2009;84(4):524–533.

58. US Department of Health and Human Services. Health Insurance Portability and Accountability Act of 1996 (HIPAA). Pub L No. 104–191, 110 Stat 1936 (1996), codified at 42 USC §300gg and 29 USC §1181 et seq. and 42 USC §1320d et seq.

59. Equal Employment Opportunity Commission. Regulations under the Genetic Information Nondiscrimination Act (GINA) of 2008. *Fed Regist*. 2010;75(216):68912–68939. To be codified at 29 CFR §1635. http://www.gpo.gov/fdsys/pkg/FR-2010-11-09/pdf/2010–28011.pdf. Accessed February 13, 2015.

60. Department of Health and Human Services. Modifications to the HIPAA privacy, security, enforcement, and breach notification rules under the health information technology for economic and clinical health act and the genetic information nondiscrimination act; other modifications to the HIPAA rules; final rule (HIPAA omnibus rule). *Fed Regist*. 2013;78(17):5566–5702. To be codified at 45 CFR §160, §164. http://www.gpo.gov/fdsys/pkg/FR-2013-01-25/pdf/2013-01073.pdf. Accessed February 13, 2015.

Downloaded from http://meridian.allenpress.com/doi/pdf/10.5858/arpa.2014-0568-CP by guest on 03 December 2022