Human Molecular Genetics, 2013, Vol. 22, Review Issue 1 **R27–R31**
doi:10.1093/hmg/ddt384
Advance Access published on August 19, 2013

Databases of genomic variation and phenotypes: existing resources and future needs

Jennifer J. Johnston* **and Leslie G. Biesecker**

National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA

Received July 12, 2013; Revised August 2, 2013; Accepted August 4, 2013

Massively parallel sequencing (MPS) has become an important tool for identifying medically significant variants in both research and the clinic. Accurate variation and genotype–phenotype databases are critical in our ability to make sense of the vast amount of information that MPS generates. The purpose of this review is to summarize the state of the art of variation and genotype–phenotype databases, how they can be used, and opportunities to improve these resources. Our working assumption is that the objective of the clinical genomicist is to identify highly penetrant variants that could explain existing disease or predict disease risk for individual patients or research participants. We have detailed how current databases contribute to this goal providing frequency data, literature reviews and predictions of causation for individual variants. For variant annotation, databases vary greatly in their ease of use, the use of standard mutation nomenclature, the comprehensiveness of the variant cataloging and the degree of expert opinion. Ultimately, we need a dynamic and comprehensive reference database of medically important variants that is easily cross referenced to exome and genome sequence data and allows for an accumulation of expert opinion.

INTRODUCTION

Massively parallel sequencing (MPS) has become an important tool in discovering the genes and variants associated with inherited disease. The vast quantity of information that is generated from MPS demands new ways of analyzing data, which are wholly dependent on accurate genotype–phenotype databases. Understanding the relationship of variants to disease and pathophysiology is the core of human molecular genetics in both research and the clinic.

Making sense of the data

The MPS pipeline generates raw read data, which must be aligned to a reference, and variant genotype calls are made against that reference. These genotype calls are commonly provided using the variant call file (.vcf) format and typically include 100 000 variants for an exome and 3 600 000 (1) variants for a genome. Then, the medical genomicist must process this file by filtering these thousands of variants down to a manageable subset, distinguishing medically important variants from the others.

Ideally, this filtering would be a single operation—simply intersecting a .vcf file with a comprehensive reference database of medically important variants. However, the necessary genotype–phenotype resource does not yet exist, and so a filtering strategy needs to be implemented that optimizes the time and expense of analysis against the amount of useful information returned. Initial filtering often focuses on genotype quality, frequency [to remove common variants, minor allele frequency (maf) > 0.1] and variant type. Variant frequency can be extracted from databases such as 1000 Genomes, the National Heart Lung and Blood Institute's Exome Sequencing Project (ESP) and the Database of Short Genetic Variations (dbSNP). However, these databases do not include significant amounts of phenotypic data and, without that, the threshold to exclude a variant as pathogenic must be set high to minimize the error of falsely rejecting a variant. This is especially problematic because some of the datasets are enriched for samples from patients with rare disorders. When annotated mutation types are restricted to non-synonymous, nonsense, insertion/deletion and splice site alterations, using these three databases to filter out variants with maf of >0.1 can reduce the number of variants by $\sim$99%.

The hundreds of remaining variants require further filtering to distill them down to those that are highly penetrant. This is a time-consuming process that requires the use of multiple additional databases and the primary literature to assess individual variants. Available databases, in addition to the variation-centric

*To whom correspondence should be addressed at: Building 49, Room 4C60, Bethesda, MD 20892-4472, USA. Tel: +1 301 594 3981; Fax: +1 301 402 2170; Email: jjohnsto@mail.nih.gov

databases mentioned above, include fee-based curated databases focused on variant information (e.g. Human Gene Mutation Database; HGMD), expert curated databases focused on variant information (locus-specific databases; LSDB), expert curated databases focused on clinical information with some variant data (e.g. GeneReviews), curated databases providing information on inherited phenotypes and genes with selected variant information (e.g. Online Mendelian Inheritance in Man; OMIM) and open-source databases (e.g. ClinVar) (Table 1). Each of these models has strengths and weaknesses and fulfills distinct needs.

AVAILABLE DATABASES

1000 Genomes

The 1000 Genomes project (http://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/, last accessed on July 12 2013) includes sequence data from 2500 individuals from 26 different populations from Europe, Asia, Africa and the Americas (2). Currently, the database contains information on 1092 individuals from 14 populations and >39 million variants. Variants are mapped to the reference genome (GRCh37), and dbSNP identification numbers are provided. The database does not include phenotypic data on participants—other than that which can be deduced from the fact that they were adults who were sufficiently intellectually capable to read and sign a consent form. For example, one can conclude that a variant found in 1000 Genomes is unlikely to be the cause of a highly penetrant autosomal dominant intellectual disability disorder. For the purposes of MPS annotation, it serves primarily as a frequency filter. Both combined and population-specific allele frequencies are available allowing for the filtering of variants that are common to a single population. The numbers of some of the population groups are sufficiently small; however, such that the absence of a variant in 1000 Genomes does not prove it is rare.

NHLBI GO Exome Sequencing Project

The NHLBI GO ESP is focused on understanding the contribution of rare genetic variation to heart, lung and blood disorders through the sequencing of well-phenotyped populations. Variant count data are available on the Exome Variant Server (EVS; http://evs.gs.washington.edu/EVS/, last accessed on July 12 2013), which currently contains exome sequence data on 6503 individuals, and allele frequencies are provided for African-Americans and European-Americans. As ESP is an exome sequencing project, the database is not useful to assess the frequency of most variants identified in a genome shotgun sequence. The ESP variants are mapped onto the reference genome (GRCh37) and Human Genome Variation Society (HGVS) nomenclature is provided with reference accession numbers. Conservation and polyphen2 predictions are also available. While individual genotype and phenotype data are not available on EVS, these data can be accessed through dbGAP. For studies not focused on heart, lung and blood disorders, ESP functions primarily as a frequency filter much like 1000 Genomes. An important difference is the inclusion in ESP of samples from individuals with rare disorders [e.g. 418 individuals with cystic fibrosis which is 6% of ESP, or ~125 times the population frequency of this disease in individuals of Northern European ancestry (M. Bamshad, personal communication)], requiring care when using the data as a frequency filter in studies with overlapping disease phenotypes.

dbSNP

The National Center for Biotechnology Information (NCBI) Short Genetic Variations database (http://www.ncbi.nlm.nih.gov/projects/SNP/, last accessed on July 12 2013) (3), or dbSNP, currently includes 38 072 522 validated human reference single nucleotide polymorphism (RefSNP) clusters. Variants are submitted by researchers and can include frequency data and clinical significance. Variants are described by genomic position, and HGVS nomenclature is given for cDNA and protein references with accession numbers. For the purpose of the database, 'SNP' means only 'a report of variation', does not imply anything about variant type, frequency or causation and can also include variants seen only in tumors. Entries range from common, benign variants to rare, highly penetrant disease-causing variants and it is important to realize that an entry in dbSNP is not necessarily evidence against causation, regardless of disease. Frequency data in dbSNP include data from 1000 Genomes and ESP and may contain additional populations not present in the other databases. Additionally, some RefSNP clusters include a prediction of clinical significance and associated PubMed citations; and in this way, dbSNP begins to offer limited phenotype information.

Human Gene Mutation Database

The HGMD is a curated resource that focuses on published variants present in genes known or suspected to be related to human disease (4). The public version of the database (http://www.hgmd.org, last accessed on July 12 2013) has limited access for academic institutions and non-profit organizations, and full access is available through a subscription to HGMD Professional (http://www.biobase-international.com/product/hgmd, last accessed on July 12 2013). HGMD Professional currently contains entries for 5621 genes and a total of 141 161 variants. The HGMD offers a consistent tabular format with genomic coordinates and HGVS nomenclature based on a reference sequence. Additional information including SIFT (5) and MutPred (6) model predictions as well as conservation are included for missense variants. Each entry contains a prediction of pathogenicity and variants are assigned to categories including disease-associated polymorphisms (DP), functional polymorphism (FP), disease-associated functional polymorphisms (DFP), frameshift or truncating with no reported disease association (FTV), potentially disease-causing (DM?) or disease-causing (DM). The mutation category is based primarily on a curated literature review. Evidence used to determine causation includes absence in normal controls, segregation in families, multiple occurrences in affecteds, functional data and disruption of protein domains or loss of conservation. In practice, limited information is typically available in the literature for many variants, and HGMD relies on the judgments of authors, reviewers and editors. It is important for medical genomicists to confirm that the evidence for causation cited by HGMD is sufficient for their particular implementation. HGMD professional users can

Table 1. Properties of genomic variation and phenotype databases

	1000 Genomes	NHLBI Exome Variant Server	dbSNP	Human Gene Mutation Database	Locus-specific databases	OMIM	GeneReviews	ClinVar
Focus	Genome/exome variation in diverse populations, germline only	Exome variation in well-phenotyped populations, germline only	Repository for all molecular variation, both germline and somatic	Detailed information on variants responsible for inherited disease, germline only	Gene-specific variants, some with expert curation, both germline and somatic	Literature review for genes and genetic phenotypes, germline and somatic variants	Expert clinical review based on the literature for genes and the phenotypes associated with germline and somatic variants	Clinical significance of variants across all genes, both germline and somatic
Variant source	Variants from sequence data in individuals from 26 populations	Variants from sequence data in phenotyped individuals, many with rare disorders	Submitted by research/ clinical groups	Variants mined from the literature, does not include unpublished variants	Submitted by research/clinical groups, database specific	Selected variants mined from the literature	Variants selected by authors based on their phenotypic relevance	Submitted by research/ clinical groups or extracted from public databases or expert consensus reports
Phenotype	None provided	Focused phenotype information available through dbGAP	May provide clinical significance of variant	Phenotypic information limited to associated disease	May provide detailed phenotype per submission	Thorough review of the phenotype	Thorough review of the phenotype	Limited phenotypic information
Clinical resource	None	None	None	None	None	Clinical synopsis/ literature review of clinical details	Includes clinical practice guidelines	Can include variant-specific practice guidelines
Prediction of Causation	None	None	May provide submitter prediction	Yes Interpretation of the literature with prediction of causation	Yes Two part prediction, submitter/curator	Yes Associated phenotype with interpretation of the literature	Yes Interpretation of the literature with prediction of causation	Yes Submitter prediction/ expert predictions
Model-based information	No	Yes PolyPhen2 Conservation	No	Yes Sift, MutPred Conservation	No	No	Not standard, may be included	Not standard, may be included
Accessibility	Public	Public	Public	Academic and non-profit limited access/fee-based full access	Public	Public	Public	Public
Curator	1000 Genomes	University of Washington	NCBI Limited curation	HGMD Subscribers can submit feedback	Various experts	Johns Hopkins	University of Washington-based editors	NCBI Individuals can review variants and submit a reviewed record
References	None	None	If provided by submitter, may be mined from PubMed	First report of all mutations, additional reports may be included	Variant-specific references when available	Gene- and variant-specific references	Gene- and variant-specific references	Gene-specific references, variant data linked to submitter, may or may not have reference

provide feedback to the database if they have additional information about a variant allowing for ongoing annotation. Beyond its utility as a variant database, it is useful as a gene database—a single source list for all genes known or suspected to cause human disease when mutated. As HGMD is restricted to published variants and often only cites the primary reference, some LSDBs have more variants and a more complete list of references than does HGMD, but the centralization of data in HGMD makes large-scale annotation feasible.

Locus-specific databases

Numerous LSDBs exist and are typically curated by gene experts without centralized editing. HGVS maintains a list of available LSDBs at http://www.hgvs.org/dblist/glsdb.html, last accessed on July 12 2013. The quality of these databases varies widely in the number of included variants, the level of information, standards for determination of causality and overall uniformity, making their use for variant annotation challenging. To improve uniformity, the HGVS has developed the Leiden Open Variation Database (LOVD) 3.0 (http://www.lovd.nl/3.0/home, last accessed on July 12 2013), a free tool for LSDB development providing a consistent user interface (7). Each database homepage lists the curator, reference sequences and other pertinent information. Variant data can be extensive and include phenotype, functional data, family information, reported frequency and references. For each occurrence of a variant, both the submitter and the curator can make a determination of pathogenicity. These are based on a five-point scale: pathogenic (+), probably pathogenic (+?), variant of uncertain significance (?), probably no pathogenicity (−?) and no known pathogenicity (−). In practice, the submitter and curator assessments do not always agree and the medical genomicist will need to make their own assessment of the variant. The LOVD submissions are based on defined reference sequences allowing these databases to be queried using standard HGVS nomenclature. Genomic coordinates are not provided in the tabular format, but variants are mapped to the genome.

Individual databases vary in their utility. While some LSDBs are the most complete source of variant information, others do not contain a single variant. Some LSDBs do not provide a standard reference sequence, making it difficult to search for specific variants. Other LSDBs offer lists of variants without additional information, providing little annotation. That these LSDBs are separate databases makes them challenging to incorporate into a genomic analysis pipeline. For some genes, there are multiple LSDBs, increasing this complexity. Centralizing the ability to search these resources would greatly improve their utility. Finally, as noted above, they are only as good as the information provided by submitters and curators, which is variable.

Online Mendelian Inheritance of Man

The Mendelian Inheritance of Man (MIM) resource was started in the 1960s by Dr Victor A. McKusick as a physical catalog of mendelian traits and disorders with or without known molecular etiology. Curation of the database is currently performed at The Johns Hopkins University School of Medicine, and it transitioned to an online version beginning in 1985, now termed Online MIM (OMIM, http://www.omim.org, last accessed on July 12 2013). The database currently includes 21 848 entries, including 4926 phenotypes with a known molecular basis and 3003 genes with known causative mutations. For discovery, OMIM contains over 10 000 entries for genes without known mutations and ∼3500 entries for phenotypes whose molecular basis is not yet known. Each entry is a text summary of information from the literature with references. Only selected allelic variants are reviewed and criteria for inclusion include the first mutation to be discovered, high population frequency, distinctive phenotype, historic significance, unusual mutation mechanism, unusual pathogenesis or distinctive inheritance (e.g. dominant with some mutations, recessive with others in the same gene). Variant information follows the textual summary and does not typically include precise HGVS nomenclature or genomic coordinates. For that reason, searching OMIM for specific variants is challenging, and OMIM is often most useful in understanding the connection of a specific gene to disease. For extensively studied variants, OMIM can be a good information source as it can offer a thorough literature review.

GeneReviews

GeneReviews, developed and maintained at the University of Washington (http://www.ncbi.nlm.nih.gov/books/NBK1116/, last accessed on July 12 2013), is a collection of peer-reviewed inherited disease descriptions written by experts with extensive and thoughtful editing. It contains phenotypic information and information on selected variants, and its strength is in the clinical summaries it offers. The entries follow a standard format and focus on the clinical aspects of the disease including diagnosis, management and counseling. Unlike the OMIM format, GeneReviews reads like a medical textbook. However, only selected, strongly implicated genomic variants are included, consistent with its clinical focus. Selected variants are typically presented in the table format with standard HGVS nomenclature based on defined reference sequences but without genomic coordinates. The inclusion of selected variants and the absence of genomic coordinates limit the utility of GeneReviews as a database against which MPS data can be filtered. However, GeneReviews, like OMIM, can be very useful for clinical genomics by providing a source for genes strongly associated with human disease and an understanding of the causal relationship of a gene to a disease and the nature of the disease.

ClinVar

ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/, last accessed on July 12 2013) was launched in 2012 as a freely accessible public archive for reports of the relationship of genomic variations to phenotypes. An initial population of ClinVar was through the extraction of variants from OMIM, GeneReviews, some LSDBs and testing laboratories. Much like individual LSDBs, the expectation is for variant information to be submitted to ClinVar by researchers and clinical laboratories. ClinVar will serve as a central repository for predictions of causation with five standard categories ranging from benign to pathogenic and additional categories covering pharmacogenomics and complex traits. Clinical assertions are included from independent submitters and can be reviewed by others. Entries include the submitter, clinical significance and method of identification. While the

option of including supporting evidence is integral to ClinVar, much like LSDBs, the inclusion of specific items is dependent on the submitter and varies among entries. These assessments will not be consistently curated for the determination of pathogenicity. In addition to individual variant entries, ClinVar, like HGMD, provides gene information from external sources and links to a wealth of information. A companion resource, Variant Reporter (http://www.ncbi.nlm.nih.gov/variation/tools/reporter, last accessed on July 12 2013), is offered through the NCBI allowing for automated analysis of variants starting from a variety of input formats, including vcf and HGVS nomenclature. Variant Reporter returns frequency data from 1000 Genomes, a prediction of clinical significance if one exists in ClinVar, and pubmed IDs for relevant references. Currently, ClinVar includes 40 367 variants due to its recent launch and time will tell if ClinVar becomes a primary repository.

FUTURE NEEDS

Ultimately, we need a comprehensive reference database of medically important variants that is easily cross referenced to exome and whole genome sequence data. While links can be provided to additional sources of information, the necessity to query multiple sources for each variant should be minimized. The creation of a comprehensive reference database requires the submission or mining of all available data for each variant, and an ability to contribute new data and annotations over time. Such a database must be searchable using standardized nomenclature to allow for computerized annotation. Genomic coordinates are likely the most useful in this regard as they can be defined for the entire dataset without the need to further define cDNA reference sequences. Standardization of genomic nomenclature is critical for insertion/deletion variants, where using left and right flanking positions followed by the reference and variant sequence may be the most exact way of defining the data.

HGMD has been successful at collecting a large number of variants into a single database, although this has been done as a fee for service instead of an open-source mechanism. As far as mining data from other primary resources, this is already a common feature of existing databases and should be easy to incorporate into a comprehensive database. Missing from all existing databases short of ClinVar, however, is the ability to collect ongoing annotation information for each variant. A dynamic database that allows for an accumulation of expert opinion will allow the field to move toward collective conclusions about causality. Individual databases meet various needs of the medical genomicist. However, the ideal database would combine the frequency data from 1000 Genomes, ESP and dbSNP with the consistency of HGMD, the completeness and disease-specific expertise of the best LSDBs as well as the external and dynamic curation available in ClinVar, for a single comprehensive database and streamlined variant analysis.

ACKNOWLEDGEMENT

The authors thank Michael Bamshad, Lisa Brooks, David Cooper, Ada Hamosh, Melissa Landrum, Donna Maglott, David Ng, Roberta Pagon and Larry Singh for thoughtful comments on this review.

Conflict of Interest statement: L.G.B. is an uncompensated advisor to the Illumina Corp. and receives royalties from Genentech Corp.

FUNDING

The authors are supported by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health.

REFERENCES

1. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T. and McVean, G.A. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.
2. Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E. and McVean, G.A. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.
3. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M. and Sirotkin, K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
4. Stenson, P.D., Mort, M., Ball, E.V., Howells, K., Phillips, A.D., Thomas, N.S. and Cooper, D.N. (2009) The Human Gene Mutation Database: 2008 update. *Genome Med.*, **1**, 13.
5. Ng, P.C. and Henikoff, S. (2001) Predicting deleterious amino acid substitutions. *Genome Res.*, **11**, 863–874.
6. Li, B., Krishnan, V.G., Mort, M.E., Xin, F., Kamati, K.K., Cooper, D.N., Mooney, S.D. and Radivojac, P. (2009) Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics*, **25**, 2744–2750.
7. Fokkema, I.F., Taschner, P.E., Schaafsma, G.C., Celli, J., Laros, J.F. and den Dunnen, J.T. (2011) LOVD v.2.0: the next generation in gene variant databases. *Hum. Mutat.*, **32**, 557–563.