Improving databases for human variation

To determine the true pathogenicity of genetic variants, data sharing is essential.

Single-nucleotide variants in the human genome number in the tens of millions. High-throughput sequencing has made identifying them relatively easy, but interpreting them is a far more difficult task. If we are to gain insight into the molecular underpinnings of disease and use that knowledge for better clinical decisions, it is essential to understand the impact of human genetic variation.

Progress will be made much faster if the variation data, and information on the methods used to obtain them, are shared in repositories accessible to all.

Hundreds of variant databases exist; many focus on a particular subset of genes (a list is maintained by the Human Genome Variation Society). Some seek to cover the entire human genome. They provide valuable resources for basic researchers and clinicians alike, but as Heidi Rehm, director of the Laboratory for Molecular Medicine at Partners Healthcare Personalized Medicine, puts it, "There is no true truth for most variants. Most interpretations are expert opinions based on an evaluation of evidence."

It is therefore important to know the content, curation methods and quality of interpretation for the different databases to best incorporate the data into research or clinical efforts.

Among the large, widely used databases is the Online Mendelian Inheritance in Man (OMIM), a detailed catalog of over 15,000 human genes and Mendelian disorders begun in 1960 and currently curated by researchers at Johns Hopkins University School of Medicine. OMIM contains a representative set of variants that supports the gene-disease relationship and disease mechanism.

The Human Gene Mutation Database (HGMD), established in 1996 and maintained by the Institute of Medical Genetics in Cardiff, UK, is a collection of over 174,000 germline variants published in the peer-reviewed literature as disease causing for inherited disease in humans.

Although these databases provide an important resource, they are not without shortcomings. Work by Bell *et al.* (2011) has shown that 27% of annotations for recessive disease-causing genes are incorrect. Dorschner *et al.* (2013) re-reviewed the primary literature supporting 239 unique variants entered in HGMD as disease causing and found that only 7.5% actually fit that category. Bell *et al.* concluded that OMIM and HGMD are insufficient arbiters of whether variants are pathogenic for disease.

An additional hurdle in using HGMD is that its full content can be accessed only through a subscription. The free version is of limited utility because it is not up to date, does not provide chromosomal coordinates and has few search functions.

This has led to calls for a more comprehensive and centralized resource for information on variant pathogenicity. In 2012 the US National Institutes of Health (NIH) started ClinVar, a free database that seeks to connect germline and somatic variants to their clinical interpretations. It initially drew submissions from OMIM and locus-specific databases but soon started to accept direct submissions from clinical testing labs and researchers. ClinVar requires the classification of each variant into a clinical category (benign, likely benign, likely pathogenic, pathogenic or uncertain significance).

To help facilitate submissions to ClinVar and improve the accuracy of variant interpretations, the NIH started funding the ClinGen initiative in 2013.

ClinGen helped implement a ranking system to denote the quality associated with each submission to ClinVar. For a single submitter to obtain a one-star entry, variants need to be classified into at least three clinical-significance tiers, and the methods used for the assignment and supporting evidence must be provided. Importantly, users can get in touch with a submitter to discuss the basis of their interpretations. Two-star entries describe variants from multiple submitters with no conflicting interpretation, and three- and four-star submissions come from expert panels and large consortia with ClinGen-approved methods for variant interpretation. Of the 172,870 variants in ClinVar (as of January 4, 2016), 56,742 meet at least the one-star criterion. The ~105,000 entries without a star should be treated with more caution.

With Variant Explorer, a ClinVar user can easily determine how the full set of variant interpretations by a specific submitter compares to those of others. This is useful in alerting labs to a tendency to over- or undercall pathogenic variants and may prompt some to alter their methods.

ClinGen has put together expert panels for different diseases to resolve differences in interpretation in ClinVar and provide expert interpretation of variants. But for this to be achieved, knowledge must be shared; clinical labs as well as those doing genetic testing or engaging in basic research should make their variant data freely available.