When a “Disease-Causing Mutation” Is Not a Pathogenic Variant

Jian Wang1,2 and Yiping Shen1,2*

The clinical utility of molecular genetic testing relies on our accurate and comprehensive knowledge about the relationships between genes/variants and diseases/symptoms. Correctly interpreting the clinical significance of detected variants continues to be a constant challenge for molecular diagnostic practice. This challenge has become substantially magnified as testing based on next generation sequencing (NGS)3 becomes rapidly integrated in routine clinical practice.

Whole exome sequencing for clinical diagnostics is a new practice. There are many challenges that lie ahead. One key issue out of many remains our ability to correctly interpret the clinical significance of gene variants, which lags far behind our ability to discover them. Consequently, many variants are currently categorized as variants of “unknown significance” in a clinical report. This certainly poses a challenge to physicians who receive the report, but we would argue that an even more pressing issue is the possibility of a false diagnosis: a non–disease-causing variant being interpreted as a pathogenic variant owing to our incorrect knowledge about gene/variant and disease/symptom associations. This possibility may be small for tests based on single genes or gene panels but markedly higher for tests based on whole exome sequencing/whole genome sequencing. Assigning a non–disease-causing variant as causative to a patient will deny the patient the future opportunity for identifying the real disease causation, and as a result the patient may never receive a truly correct disease diagnosis. Any such false diagnosis has the potential to have a huge impact on the care of the individual patient, on the diagnosis and management of family members who test negative or positive for the false-positive variant, and potentially on future family planning.

DNA sequence variants can largely be grouped into 2 categories: truncating variants and nontruncating variants. Truncating variants consist of nonsense variants, out-of-frame indels, most splicing variants, and partial gene deletions. Nontruncating variants consist of single or multiple nucleotide substitutions and in-frame indels. Truncating variants, which have a more deleterious impact on gene products, are often pathogenic to diseases that are caused by loss-of-function gene products. Nontruncating variants can result in loss of function or gain of function, and it is often more difficult to predict their pathogenicity.

Based upon the American College of Medical Genetics and Genomics guideline for variant clinical interpretation (1), pathogenic variants are determined on the basis of 1 of 2 types of evidence: (a) the sequence variant has been reported previously and is a recognized cause of disease, or (b) the sequence variant has not been reported previously and is of a type that is expected to cause the disease. These 2 guidelines are based upon 2 assumptions: (i) the previously reported causal relationship between a variant and a disease is correct, and (ii) certain types of variants can be predicted with high confidence to cause disease. New evidence has challenged both of these assumptions, although only the former are discussed here.

Early on, Bell et al. discovered that an unexpected proportion (27%) of literature-annotated disease mutations in recessive disease-causing genes were incorrect (2). When large population exome data became available, this issue was further evaluated both at the gene and variant levels. Piton et al. (3) used the population exome data set generated by the National Heart, Lung, and Blood Institute Exome Sequencing Project (http://evs.gs.washington.edu/EVS/), and raised concerns about some X-linked intellectual disability (XLID) genes based on 3 indications: (a) truncating variants are observed in the Exome Variant Project (EVP) data set, (b) previously published disease-causing mutations are detected at a higher than expected frequency in the EVP data set, and (c) the original implication of the gene’s association with XLID was based upon insufficient evidence. After gathering variant frequency data and reviewing the evidence in the original research papers, they found that 10 genes (approximately 10% of total known XLID genes) were unlikely to be true XLID genes. For example, the angio-

1 Department of Laboratory Medicine, Shanghai Children’s Medical Center, Shanghai Jiaotong University School of Medicine, Shanghai, P. R. China; 2 Department of Laboratory Medicine, Boston Children’s Hospital, Boston, MA.

* Address correspondence to this author at: Department of Laboratory Medicine, Boston Children’s Hospital, 300 Longwood Ave., Boston, MA 02115. Fax 617-730-0338; e-mail yiping.shen@childrens.harvard.edu.

Received October 27, 2013; accepted December 9, 2013.

Previously published online at DOI: 10.1373/clinchem.2013.215947

1 Nonstandard abbreviations: NGS, next generation sequencing; XLID, X-linked intellectual disability; EVP, Exome Variant Project; HGMD, Human Gene Mutation Database.
Tensin II receptor, type 2 (AGTR2) gene (OMIM 300852) was originally identified as an XLID gene because a missense variant p.Gly21Val was detected in 2 brothers with profound mental retardation and was absent from 510 control male chromosomes. The glycine residue at position 21 of this gene was poorly moderately conserved and there was only a small physiological difference between glycine and valine residues. Now this variant is observed at a 0.46% frequency in European Americans in the EVP data set. If this were a true highly penetrant disease-causing variant, 0.46% of the population would have been identified with intellectual disability due to this variant. This is inconsistent with a population disease incidence (2% of incidence in the general population for all types of ID). In light of this type of evidence, Piton et al. suggested that 10 genes not be considered as proven XLID genes and another 15 genes required further supporting evidence to establish a causal relationship with XLID.

By no means is this classification dilemma specific to XLID genes. Recently, Dorschner et al. reported an even more alarming result while evaluating the frequency of actionable incidental variants in exomes (4). Among 239 unique Human Gene Mutation Database (HGMD) variants identified as disease causing, they found that only 7.5% of these variants were pathogenic or likely pathogenic by rigorous evaluation criteria. Again, the availability of variant frequencies from population exome databases such as EVP and the 1000 Genomes Project enabled the reclassification of these variants according to the rationale that many of them were found to be too common to be highly penetrant disease-causing mutations. Other similar studies examining a defined set of genes associated with particular disorders such as sudden infant death syndrome (5) also uncovered that a variable, but significant, percentage of variants previously reported as disease causing are present at high frequency in population exome data sets. Collectively, these findings point to a common problem: a fraction of currently known disease genes in databases such as HGMD and locus-specific databases are not true disease-causing genes.

Currently there is no single database that is rigorously curated for the purpose of clinical diagnostic sequencing. This is a widely recognized problem. NIH is funding a project (the Clinical Genome Resource, or ClinGen) that is aimed to provide our community with authoritative information on genomic variants that are relevant to human disease and useful in clinical practice. We anticipate that an ideal human mutation database will apply standard and rigorous criteria to curate each gene and corresponding variant. This database will also contain sufficient details of supporting evidence that allow for ongoing reassessment when new evidence becomes available. A curation process should integrate evidence from genetics, statistics, bioinformatics, and functional biology and allow for new knowledge to be rapidly integrated. We have come to understand that during this process it is often difficult to reach a sufficient level of confidence from stand-alone evidence in a single category, and each category of evidence has its own pitfalls that have led to false claims of disease–gene (variant) association in the first place. We believe more stringent criteria and an integrated approach are necessary to reclassifying genes and reevaluating variants using the tools and resources that were not previously available.

Many disease-causing variants are very rare or are only present in a patient or the patient’s family, so statistical analysis of relative frequency lacks sufficient power to definitively prove a causal relationship. In such a scenario, interpretation needs to be dependent on evidence from multiple sources such as (a) the type of variant (e.g., truncating variants are more likely to be causal for diseases caused by loss-of-function mechanisms), (b) the location of the variant in the gene (e.g., variants located at critical residues or functional domains are more likely to cause diseases), (c) family study (i.e., does the variant segregate with phenotype in family?), (d) functional study, and (e) informatic prediction. Current informatic predictive tools based on conservation score and structural impact score are unreliable for use in clinical interpretation. Using a curated true pathogenic variant set to train predictive tools should help to improve prediction sensitivity and specificity. A better understanding of disease mechanisms will help to establish relationships between types and locations of variants and pathogenicity. Functional studies using in vitro heterologous cell models and surrogate readouts often do not directly demonstrate disease causality. It is desirable to perform functional analysis on relevant cells and validate the effect of variants by restoring the phenotype after complementing the genetic deficiency. Family study is often limited due to a small number of available family members. Variant frequency data from large population exome sequencing are extremely helpful for statistical analysis, as demonstrated by studies mentioned here, but some variants generated by NGS, such as indels, may have high false-positive and false-negative rates. Some variants show significantly different frequency in different ethnic groups. These caveats require us to be cautious in using these data. In addition, evaluating variants of incomplete penetrance requires even larger data sets for statistical significance. Reevaluating variants will be an ongoing effort that requires the involvement of the whole clinical diagnostic and research communities and a corresponding forum for reinforming patients of this new knowledge as might be needed.
The scale of the problem and significance of the challenge call for the birth of a new discipline in this medical genomics era: interpretive medical genomics. Its practitioners will require comprehensive expertise in medical genetics, genomics, bioinformatics, statistics, and in vitro and in vivo functional analysis, which, combined with the importance of the clinical variant in medicine, make clinical variant interpretation a heroic venture for our time.

**Author Contributions:** All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.

**References**