METHODS

Human Mutation

OFFICIAL JOURNAL
HGVS
HUMAN GENOME VARIATION SOCIETY
www.hgvs.org

A Standardized DNA Variant Scoring System for Pathogenicity Assessments in Mendelian Disorders

Izabela Karbassi,[1†] Glenn A. Maston,[1†] Angela Love,[1] Christina DiVincenzo,[1] Corey D. Braastad,[1] Christopher D. Elzinga,[1] Alison R. Bright,[1] Domenic Previte,[1] Ke Zhang,[2] Charles M. Rowland,[3] Michele McCarthy,[1] Jennifer L. Lapierre,[1] Felicita Dubois,[1] Katelyn A. Medeiros,[1] Sat Dev Batish,[1] Jeffrey Jones,[1] Khalida Liaquat,[1] Carol A. Hoffman,[1] Malgorzata Jaremko,[1] Zhenyuan Wang,[1] Weimin Sun,[2] Arlene Buller-Burckle,[2] Charles M. Strom,[2] Steven B. Keiles,[2] and Joseph J. Higgins[1]*

[1]Quest Diagnostics, Athena Diagnostics, Marlborough, Massachusetts; [2]Quest Diagnostics, Nichols Institute, San Juan Capistrano, California; [3]Quest Diagnostics, Celera, Alameda, California

Communicated by David E. Goldgar

Received 28 April 2015; accepted revised manuscript 7 October 2015.
Published online 15 October 2015 in Wiley Online Library (www.wiley.com/humanmutation). DOI: 10.1002/humu.22918

ABSTRACT: We developed a rules-based scoring system to classify DNA variants into five categories including pathogenic, likely pathogenic, variant of uncertain significance (VUS), likely benign, and benign. Over 16,500 pathogenicity assessments on 11,894 variants from 338 genes were analyzed for pathogenicity based on prediction tools, population frequency, co-occurrence, segregation, and functional studies collected from internal and external sources. Scores were calculated by trained scientists using a quantitative framework that assigned differential weighting to these five types of data. We performed descriptive and comparative statistics on the dataset and tested interobserver concordance among the trained scientists. Private variants defined as variants found within single families (n = 5,182), were either VUS (80.5%; n = 4,169) or likely pathogenic (19.5%; n = 1,013). The remaining variants (n = 6,712) were VUS (38.4%; n = 2,577) or likely benign/benign (34.7%; n = 2,327) or likely pathogenic/pathogenic (26.9%, n = 1,808). Exact agreement between the trained scientists on the final variant score was 98.5% [95% confidence interval (CI) (98.0, 98.9)] with an interobserver consistency of 97% [95% CI (91.5, 99.4)]. Variant scores were stable and showed increasing odds of being in agreement with new data when re-evaluated periodically. This carefully curated, standardized variant pathogenicity scoring system provides reliable pathogenicity scores for DNA variants encountered in a clinical laboratory setting.

Hum Mutat 37:127–134, 2016. Published 2015 Wiley Periodicals, Inc.*

KEY WORDS: databases; nucleic acid; polymorphism; mutation; decision support techniques; clinical laboratory techniques

Introduction

Genetic testing is fast becoming a formidable tool in the diagnostic armamentarium for common and rare diseases. Many specific genes in the human genome cause Mendelian disorders and many common diseases are associated with a constellation of genes harboring risk factors. The identification of disease genes permits research to move beyond searching for a cause to seeking a cure. As gene-specific therapies are developed, it will become increasingly important to identify which genetic variants provide diagnostic and prognostic information [Allen, 2015]. Burgeoning technologies provide the capability to rapidly sequence disease targeted multigene panels, the exome and the entire genome, but do not address the growing problem of interpreting the clinical significance of variants uncovered during the course of diagnostic testing. Several schemes for interpreting clinical variants have been proposed for cancer [Goldgar et al., 2004; Plon et al., 2008; Pastrello et al., 2011; Lindor et al., 2012; Eggington et al., 2014; Thompson et al., 2014], the mitochondrial genome [Wang et al., 2012], and for non-specific mutations [Bean et al., 2013; Duzkale et al., 2013; Kircher et al., 2014]. Recently, the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) updated guidance for the interpretation of sequence variants in clinical laboratories [Richards et al., 2015]. The report recommends the use of five specific categories for describing variants including pathogenic, likely pathogenic, uncertain significance (VUS), likely benign, and benign [Richards et al., 2015]. In the present study, we describe the application of these recommendations using a standardized, rules-based process that provides a variant pathogenicity risk score based on clinical grade information in a CLIA-certified laboratory.

Materials and Methods

Data Collection and Storage

Individual genetic diagnostic tests ordered by referring healthcare providers for Sanger and next-generation sequencing evaluations were interpreted by individuals board-certified by the American Board of Medical Genetics and Genomics over a four year period between 2010 and 2014 at a CLIA-certified clinical laboratory. Over 16,500 pathogenicity assessments on 11,894 distinct variants in 338 genes causing neurological, endocrine, and nephrotic genetic disorders were performed on the sequencing data. A list of the genes

Additional Supporting Information may be found in the online version of this article.

[†]These authors contributed equally to this work.

[*]Correspondence to: Joseph J. Higgins, Quest Diagnostics, 200 Forest Street, Suite 200, Marlborough, MA 01752. E-mail: Joseph.J.Higgins@QuestDiagnostics.com

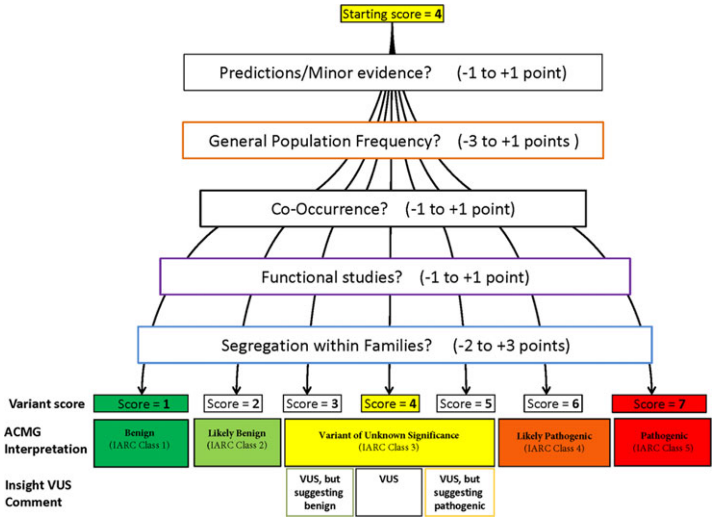

Starting score = 4
Predictions/Minor evidence? (-1 to +1 point)
General Population Frequency? (-3 to +1 points)
Co-Occurrence? (-1 to +1 point)
Functional studies? (-1 to +1 point)
Segregation within Families? (-2 to +3 points)
Variant score
Score = 1
Score = 2
Score = 3
Score = 4
Score = 5
Score = 6
Score = 7
ACMG Interpretation
Benign (IARC Class 1)
Likely Benign (IARC Class 2)
Variant of Unknown Significance (IARC Class 3)
Likely Pathogenic (IARC Class 4)
Pathogenic (IARC Class 5)
Insight VUS Comment
VUS, but suggesting benign
VUS
VUS, but suggesting pathogenic

Figure 1. Multiple lines of evidence used in the variant pathogenicity scoring system. Interpretation categories are aligned to the American College of Medical Genetics (ACMG) recommendations [Richards et al., 2015]. A midpoint score of 4 (yellow) does not favor pathogenicity or benignity. Benign scores are shown in green and pathogenic scores in red. Variants of uncertain significance (VUS) have three subclasses; score of 3 is suggestive of the variant being benign, score of 5 is suggestive of the variant being pathogenic, and score of 4 does not favor either side of the pathogenicity scale.

is provided in Supp. Table S1 with their OMIM numbers, NCBI reference transcripts, and summaries of the gene's variants. Coding exonic regions and up to 20 nucleotides of their flanking intronic sequences were tested and analyzed. Untranslated regions (UTR) were sequenced when variants were reported in the HGMD® Human Gene Mutation Database (Biobase, Waltham, MA), Online Mendelian Inheritance in Man (OMIM®), or cited in publications found in PubMed (NCBI, Bethesda, MD). Variants were identified by aligning gene sequencing results with the National Center for Bioinformatic Information human transcript reference sequence. The Alamut HT standalone (version 1.1.11) and Alamut database 2013.12.15 (version 2.2) (Interactive Biosoftware, Rouen, France) were used for basic DNA sequence annotation. Alamut Visual (version 2.2) was used for alignment, conservation, SIFT/PolyPhen dbSNP, and Exome Sequencing Project data collection to evaluate the variant in its surrounding genomic context. Alamut's standardized text string was used for literature and other variant-specific data searches. Five types of variant data were collected in a stepwise manner including data from minor evidence/prediction tools, population frequency, co-occurrence, segregation, and functional studies (Fig. 1). Variant information from external databases (e.g., ClinVar, LOVD, UMD, and specialty gene/disease databases) and publications was organized in a standardized, retrievable format within an SQL database. Novel variants were scored on their initial identification. Likely benign, VUS, or likely pathogenic variant scores were re-assessed if more than 4 months lapsed after scoring. Pathogenic and benign variant scores were evaluated every 4 years or upon the request of a patient, physician, genetic counselor, or a laboratory director. All data collection, database annotation, and pathogenicity assessments were performed by scientists trained in variant scoring using standardized training modules and annual proficiency testing.

A Rules-Based, Weighted Variant Scoring System

We used an internal 7 point scale with three subclasses in the variant of unknown significance (VUS) category that aligns with the five variant categories recommended previously [Richards et al., 2015], including pathogenic (score = 7), likely pathogenic (score = 6), VUS (score = 3–5), likely benign (score = 2), and benign (score = 1) (Fig. 1). The midpoint score of 4 was considered baseline and all variants began at this score prior to addition of data. Point values ranging from −3 to +3 were derived from five types of data, with 0.5 being the smallest change in scoring (Supp. Table S2). The sum of all point values was added to the starting score of 4 to produce a pathogenicity score ranging from 1 to 7 (Fig. 1). A special consideration was given to genes where null variants (e.g., frameshift, nonsense, canonical splice site variants at the ±1,2 positions associated with out-of-frame events) were documented in the literature to cause well-characterized disease phenotypes. These variants were assigned +2 points which raised their score to 6 (likely pathogenic) (Supp. Table S2). Exceptions to this rule were applied to null variants near the C-terminus that were not likely subject to nonsense-mediated RNA decay, those variants occurring in a non-relevant isoform, or in gene-specific cases where the disease mechanism or molecular biology was not well characterized.

Five Types of Variant Data

Minor evidence and prediction tools

Minor evidence was based on prediction tools, important functional domains, known pathogenic variants at the same residue, and the report of an affected patient with the variant (Supp. Table S2). Alamut Visual (version 2.2) was used to query prediction tools to analyze variants included SIFT (http://sift.jcvi.org) [Kumar et al., 2009], PolyPhen-2 (http://genetics.bwh.harvard.edu/pph2) [Adzhubei et al., 2010], and other prediction tools that evaluated post-translational modifications and mRNA splicing (MaxEntScan: http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html; NNSPLICE: http://www.fruitfly.org/seq_tools/splice.html; GeneSplicer: http://www.cbcb.umd.edu/software/GeneSplicer/gene_spl.shtml; Human Splice Finder: http://www.umd.be/HSF/)

[Reese et al., 1997; Zhang, 1998; Pertea et al., 2001; Yeo and Burge, 2004; Desmet et al., 2009]. These tools were used together in accordance with guidelines for using prediction methods [Vihinen, 2013] (Supp. Table S3). The predicted effect of the variant using these tools was given a lower-weight of evidence as compared with other lines of evidence (Supp. Table S2). A minimal value of +0.5 was assigned if the variant was reported in one or more patients having the clinical phenotype associated with disease. Splicing prediction scores at least 15% lower than the wild-type score [Houdayer et al., 2012] and a consensus among three or more prediction tools [Hellen, 2009] were used to assess the variant's effect on known splice sites. Exon variants predicted to cause cryptic splice sites, but not predicted to change natural splice sites, were not weighted in variant scoring. Published variants or variants in the internal database that occurred in important structural or functional domains of a protein were considered evidence of pathogenicity but did not carry the same weight as published functional evidence. New variants that changed an amino acid in a codon containing pathogenic variants were considered as minor evidence (score value of +0.5) if the amino acid differed from that of the pathogenic variant (Supp. Table S2).

Frequency data in the general population

The population frequencies of variants were estimated from internal studies, published control groups, and data reported in dbSNP, 1000 Genomes, and the Exome Sequencing Project. The variant frequencies were compared with estimated disease allele frequencies, taking into account published information on disease prevalence, varying disease penetrance, and the gene-specific attributable risk in polygenic disorders. A conservative approach was taken in calculating the disease allele frequency, to account for underestimates of disease prevalence. If a variant was found to exceed the expected disease allele frequency by 10-fold, the score was reduced by 3 points. Pathogenicity scores were reduced by 2 points if the observed frequency of the variant was threefold to 10-fold above the estimated disease allele frequency, and reduced by 1 point if the variant frequency equaled or exceeded the expected disease allele frequency by <3-fold. This rule did not apply in most cases when a founder variant was identified in the literature or if the variant was significantly enriched in a self-reported ethnic population. Alternatively, pathogenicity scores were increased by 0.5 to 1.0 points if the frequency of a variant was significantly higher in affected individuals as compared with controls matched for ethnicity using either internal datasets or well described published datasets, respectively (Supp. Table S2). In cases where the data sets were too large to allow a calculation by the Fishers exact test, we used a chi squared test with the Yates correction.

Co-occurrence

The term "co-occurrence" was defined as the presence of two or more variants that paired together in the same gene or in another gene related to the same disease. Variants that co-occurred with otherwise positive results (i.e., a known pathogenic variant in dominant disorders or two pathogenic variants in recessive disorders) were considered less likely pathogenic. Recessive variants that co-occurred less than expected with recessive pathogenic variants in trans were considered less likely to be pathogenic. If a variant in a recessive gene co-occurred frequently in trans with a single known pathogenic variant, but not with second variants in controls, then the variant was considered more likely pathogenic. The observed frequency of co-occurrences of a pathogenic variant with the variant in question was compared for significant differences to the expected frequency of co-occurrences calculated from our internal positive pathogenic rate by the binomial test [Waples, 1988; Kuk et al., 2014]. Phenotype and age were considered in the comparisons when evaluating co-occurrence especially for dominant disorders with an adult or late-life onset, or for slowly progressive degenerative diseases. Generally, missense variants that co-occurred once with an otherwise positive result received a low point reduction of –0.5 because the majority of genetic diagnostic testing is performed on post-natal samples that are not known to be associated with an embryonic lethal phenotype (Supp. Table S2).

Variant segregation analysis in families

The segregation of variants in family pedigrees was analyzed by estimating the logarithm of the odds (LOD) score or by a statistical association test if the family data were incomplete. The LOD score was estimated based on the number of meiotic events and weighted as evidence for the segregation between the disease locus and the variant in family pedigrees. For example, the LOD score at a recombination fraction of zero for one known non-recombinant meiosis was estimated at $\log(2) = 0.3$, as described previously [Ott, 1999]. A LOD score ≥ 1.0 was used as evidence to support the pathogenicity of a variant. Increasing statistical evidence for variant segregation added points to the pathogenicity score (e.g., 1 point for a LOD score between 1 and 2, 2 points for a LOD score between 2 and 3, and 3 points for a LOD score over 3). The Fisher's exact test was used to calculate the statistical significance of variant segregation in pedigrees with incomplete family data especially when the proband's siblings were tested without the parents. Two points were added for de novo variants in an affected patient when paternity and maternity were confirmed by identity testing. If identity testing was not performed, then the unconfirmed de novo event was given partial weight (Supp. Table S2).

Functional studies

The functional significance of variants was based on in vitro and in vivo published studies that showed whether or not a variant damaged the normal function of a protein. One point was either added or subtracted to the variant score based on published evidence that described the molecular, biochemical, or pathophysiological role of the variant in the clinical disorder. Publications describing experimental assays evaluating transcript splicing were generally not considered as functional evidence of pathogenicity unless the aberrant splicing led to truncation of the mRNA in a haploinsufficient disease model. The functional consequences of in-frame exon skipping events were considered on an individual basis (Supp. Table S2).

Interobserver Variation in Variant Scoring

Interobserver variation was analyzed over a period of eight months by measuring the concordance among variants scored in duplicate by independent scientists. As a separate assessment changes in scores were tracked after two trained scientists, distinct from the scientist that originally scored the variant, reviewed and discussed the variant score prior to reporting the result. Binomial 95% confidence intervals (CIs) were calculated to assess interobserver consistency and to estimate reliability.

Table 1. Summary of Variants by Type and Pathogenicity Score

	Variant score							
Variant type	1 (Benign)	2 (Likely benign)	3 (VUS suggesting benign)	4 (VUS)	5 (VUS suggesting pathogenic)	6 (Likely pathogenic)	7 (pathogenic)	Total
Missense	468	200	643	2,248	1,307	322	552	5,740
Coding synonymous	846	363	1,529	77	26	1	4	2,846
Intronic	296	108	453	81	124	22	23	1,107
Frameshift	1	0	0	3	1	642	195	842
Nonsense	0	0	0	3	1	328	286	618
Consensus splice site	0	0	0	1	0	278	127	406
In-frame insertion or deletion	8	2	5	162	30	13	26	246
UTR	33	2	8	44	0	1	1	89
Total	1,652	675	2,638	2,619	1,489	1,607	1,214	11,894

Results

Summary Statistics of Pathogenicity Assessments

Table 1 provides a summary of variant types and their assigned pathogenicity scores. Almost half of variants were classified in the likely pathogenic/pathogenic (scores 6 and 7, 23.7%; n = 2,821) and likely benign/benign (scores 1 and 2, 19.6%; n = 2,327) categories. The remaining variants, designated as VUS (scores 3, 4, and 5, 56.7%; n = 6,746), were periodically reassessed for pathogenicity. Variants limited to single families, also known as private variants (n = 5,182), were scored as VUS in 4,169 (80.5%) cases and likely pathogenic in 1,013 (19.5%). Non-private variants (n = 6,712) were in the VUS range in 2,577 (38.4%) cases, likely benign/benign in 2,327 (34.7%), and likely pathogenic/pathogenic in 1,808 (26.9%). The majority of VUS subclass scores (Fig. 1) were toward the benign scores (score = 3) (39.1%; n = 2,638) or remained at the midpoint (score = 4) (38.8%; n = 2,619). VUS with some data supporting potential pathogenicity (score = 5) (22.1%; n = 1,489) were less frequent.

Variant Classification Types

Missense variants (n = 5,740) accounted for nearly half (48.3%) of all distinct variants, followed by synonymous variants (n = 2,846) at 23.9%, and intronic variants outside the canonical splice site (n = 1,107) at 9.3%. Frame-shifting insertions and deletions (n = 842; 7.1%), nonsense (n = 618; 5.2%), and consensus splice site variants (n = 406; 3.4%) were less common. In-frame deletions and insertions (n = 246) explained 2.1% of variants. Variants in UTRs were the least common variant type (n = 89; 0.7%). The distribution of mutation types among variants classified as pathogenic or likely pathogenic shows that the majority (65.8%) were nonsense, frameshift, or consensus splice site changes, whereas missense variants accounted for 31% of pathogenic variants. Only 1% (n = 28) of pathogenic variants did not alter the amino-acid residue (synonymous, intronic but outside canonical splice sites, or UTR). The majority of these variants were predicted to affect splicing. Most benign variants (70.0%; n = 1,630) did not alter the coding sequence (coding synonymous, intronic, or in UTR), and the remaining 30.0% (n = 697) were missense or in-frame insertions or deletions. Definitive truncating variants were not classified as benign.

Variant Types by Pathogenicity Score

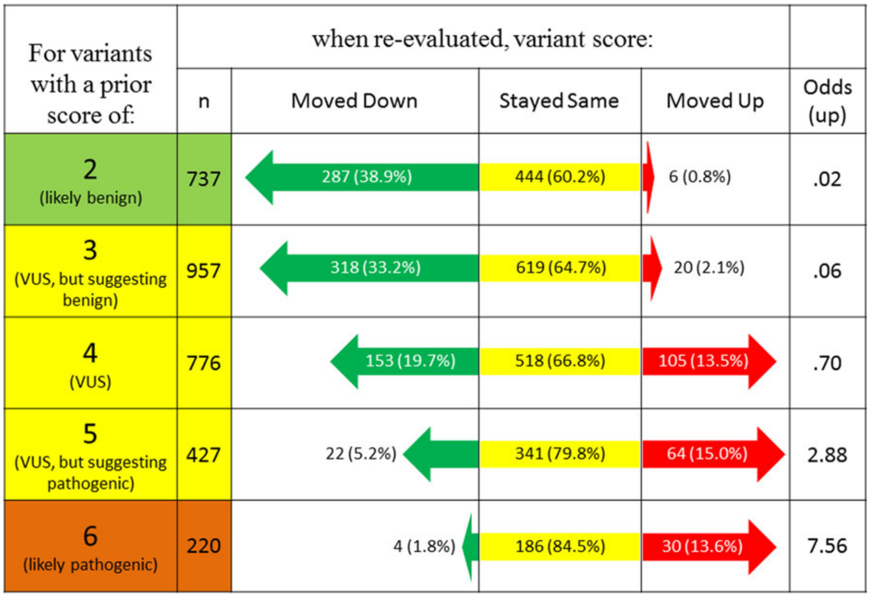
Almost all loss-of-function variants (99.5%; n = 1,856) were either pathogenic or likely pathogenic. A small number of cases with truncating or frameshift mutations were considered VUS (n = 9). These rare cases involved a few specific genes where truncating variants were not clearly known to cause disease (e.g., specific truncating mutations in *NOTCH3* and *TRPV4* in CADASIL and Charcot-Marie-Tooth disease) [Fawcett et al., 2012; Rutten et al., 2013]. Supp. Figure S1 shows the distribution of pathogenicity scores by synonymous (n = 2,846), missense (n = 5,740), intronic (n = 1,107), and in-frame insertions/deletion (n = 1,107) variant types. Thirty-eight percent of missense variants scored as VUS-suggesting pathogenic (score = 5), likely pathogenic (score = 6) or pathogenic (score = 7), whereas only 1% of synonymous changes scored within this range. Twenty-three percent of missense variants scored in the benign range (scores = 1, 2, or 3) as compared with 96% of synonymous variants. Sixty-six percent of in-frame insertions and deletions were scored as VUS (score = 4) followed by 28% in the pathogenic range (scores = 5, 6, or 7) and 6% in the benign range (scores = 1, 2, or 3).

Splicing Pathogenicity Scoring

Intronic variants, excluding canonical splice-site position, were more likely to be pathogenic because of their effect on splicing (15%) as compared with synonymous changes in coding regions (1%). Variants predicted to affect splicing outside of the consensus sites occurred in the +3 to +5 range at the beginning of introns, and from −3 to −10 at the end of introns. Variants that were predicted to affect splicing by creating new potential acceptor or donor ("cryptic") splice sites were identified in 153 cases (Supp. Table S4). The vast majority of these predictions have not been experimentally tested. Evaluations to determine the gain or loss of a binding site for an exonic splicing enhancer or silencer were not evaluated because of the difficulties in predicting the effect of a specific variant [Holste and Ohler, 2008].

Scoring Changes and Re-Analysis

Figure 2 shows the number of times a variant score changed as a function of new data. When variants with a prior score of 4 were re-evaluated (n = 776), most (66.8%) had no additional data and their scores were unchanged (n = 518). Among the remaining, new data lowered the score toward benign (19.7%, n = 153) more often than raising the score (P = 0.004). For variants initially scored as 5 (n = 427), new data increased (15% of cases; n = 64) the score more often than lowering the score (5.2%; n = 22). There were no score changes in 341 cases. Overall, a re-evaluation based on new data was more likely to increase the pathogenicity score for variants with a prior score of 5 than for those with a prior score of 4 (P < 0.0001). Variants with a prior score of 6 (n = 220) were more



For variants with a prior score of:	n	Moved Down	Stayed Same	Moved Up	Odds (up)
2 (likely benign)	737	287 (38.9%)	444 (60.2%)	6 (0.8%)	.02
3 (VUS, but suggesting benign)	957	318 (33.2%)	619 (64.7%)	20 (2.1%)	.06
4 (VUS)	776	153 (19.7%)	518 (66.8%)	105 (13.5%)	.70
5 (VUS, but suggesting pathogenic)	427	22 (5.2%)	341 (79.8%)	64 (15.0%)	2.88
6 (likely pathogenic)	220	4 (1.8%)	186 (84.5%)	30 (13.6%)	7.56

when re-evaluated, variant score:

Figure 2. Likelihood of variant score changes as a function of new data. The percentage of cases changing classification categories is depicted by green (decreases) and red (increases) arrows, cases where variant score is staying the same are depicted in yellow. The last column shows the odds of a variant score increasing to a more pathogenic score . The number of re-scoring events (n) in each scoring category is shown in the first column. Variants scored as 2 led to a lower score in 38.9% and a higher score in 0.8% on re-evaluation. Variants scored as 3 were lowered in 33.2% and raised in 2.1%. Variants scored as 2 or 3 had a significant ($P < 0.0001$) tendency to move down in scoring to classification as benign or benign/likely benign, respectively. Variants with a prior score of 5 or 6 were more likely (odds ratios of 2.88 and 7.56, respectively) to increase to more pathogenic scores ($P < 0.0001$).

likely ($P < 0.0001$) to be scored higher (13.6%, $n = 30$) than lower (1.8%, $n = 4$) (Fig. 2) on re-evaluation based on new data. None of 159 cases that scored 7 were downgraded on re-evaluation. Variants scored as 3 were re-evaluated 957 times; re-evaluation lowered the score in in 33.2% ($n = 318$) and raised it in 2.1% ($n = 20$). Only 0.8% ($n = 6$) of 737 re-evaluations of variants scored 2 led to a higher score and 38.9% ($n = 287$) were scored in the benign category. None of the variants scored 1 changed their score on re-evaluation ($n = 368$). Cases with an original score of 2 and 3 had a significant ($P < 0.0001$) tendency to continue to move further down the benign part of the scoring scale.

A review of the scoring evidence shows that general population frequency data was the most common reason for a variant score moving down toward the benign end of the scale. Co-occurrence with pathogenic positive variants was the second most common data type to move variant scores down. Family segregation data often moved variant scores up toward pathogenicity; especially the identification of de novo variants. Functional studies were less likely to change scores toward pathogenicity. This observation is expected given the number of "private variants" that we encounter.

Interobserver Scoring Consistency

Variant scores ($n = 2{,}710$) agreed in 98.5% of cases when reviewed separately and discussed by two trained scientists [95% CI (98.0, 98.9)] during routine evaluations. When nine trained scientists scored variants ($n = 104$) independently in a blinded fashion, the interobserver consistency was 97% [95% CI (91.5, 99.4)]. In both studies there was only a one-point score difference in discrepant cases with no variant scoring changes from pathogenic to benign, or benign to pathogenic. The score differences occurred in genes with unknown or ill-defined allele frequencies, disease prevalence, mixed modes of transmission, or associated with multiple phenotypes.

Table 2. The Distribution of Assigned Variant Scores Compared with the Results of Published Functional Studies ($n = 597$)

	Assigned variant score							
Published effect on protein function	1	2	3	4	5	6	7	Total
Damaging	0	10[a]	8[a]	16[a]	36	143	275	488
NOT damaging	21	22	30	3[b]	1[b]	0	2[b]	79
Conflicting	5	2	6	8	3[c]	0	6[c]	30

[a] Variant scores were lowered in 34 of 488 variants with damaging results (7.0%).
[b] Six out of 79 variants with functional study results of "not damaging" (7.6%) had a score of 4 or higher.
[c] Functional studies with conflicting information were scored in the pathogenic range in nine cases (30%).

Reliability of Functional Studies and Predictors as Indicators of Pathogenicity

Table 2 shows the relationship between the results of published functional studies and variant scores. Published functional studies that suggested that a variant damaged the protein were scored upward from a baseline score of 4 to a score of 5 in accordance with the scoring system rules (Supp. Table S2). Scores were lowered in 34 of 488 cases (7.0%) due to differences between the published functional data and other evidence. Six out of 79 variants with functional study results of "not damaging" (7.6%) had a score of 4 or higher, also suggesting that other evidence contradicted the functional study. Variant scores with functional studies containing conflicting information ($n = 30$) were distributed evenly between the

Table 3. SIFT and PolyPhen Predictions for Missense Variants Classified as Benign (n = 353) and Pathogenic (n = 363)

	SIFT tolerated	SIFT NOT tolerated	PolyPhen benign	PolyPhen damaging[a]	Both benign[b]	Both damaging[b]
Pathogenic	37	326	28	335	12	310
Benign	243	110	276	77	220	54

[a] Polyphen predictions of "probably damaging" and "possibly damaging" are combined into the damaging category.
[b] Both SIFT and PolyPhen predictions agree.

Table 4. The Performance of SIFT and PolyPhen Predictions Based on the Concordance between the Prediction Tools and Variant Classification

Parameter[a]	SIFT	PolyPhen	Both agree
Sensitivity	0.898	0.923	0.963
Specificity	0.688	0.782	0.803
PPV	0.748	0.813	0.852
NPV	0.868	0.908	0.948
FDR	0.252	0.187	0.148
Accuracy	0.795	0.853	0.889

[a] Calculations were based on the data summarized in Table 3. The number of true positives and negatives were based on the concordance between the pathogenicity scores and the SIFT and PolyPhen predictions.
Abbreviations: PPV, positive predictive value; NPV, negative predictive value; FDR, false discovery rate.

pathogenic range (30%; scores = 5, 6, or 7), at the midpoint (27%; score = 4) and in the benign range (43%; scores = 1, 2, or 3). The assigned variant scores were in agreement with damaging (n = 454), not damaging (n = 73), and conflicting (n = 8) published functional results in 90% of cases [95% CI (87.4, 92.2)]. These results suggest that published functional studies as an independent line of evidence may predict variant pathogenicity in 90% of cases.

The performance of the prediction tools, SIFT [Kumar et al., 2009] and PolyPhen 2.0 [Adzhubei et al., 2010], were compared with a set of 716 missense variants from 205 genes that were classified as benign (n = 353) or pathogenic (n = 363) by the variant scoring system (Supp. Table S2). Table 3 shows that when the two programs agreed, their predictions matched the scoring system in 95% of cases (220/232) for benign predictions and in 85% of cases (310/364) for pathogenic predictions. Table 4 shows the performance of SIFT and PolyPhen based on their agreement with the variant classification scores generated by the scoring system. The prediction tools had a combined accuracy of 89%, with a high sensitivity (96%) and lower specificity (80%) to correctly identify benign or pathogenic variants.

Case Illustrations

The Supporting Information describes the application of variant scoring system in the pathogenicity assessment of benign and pathogenic variants in two patient cases.

Discussion

We describe a standardized, rules-based system for evaluating variant pathogenicity in a diagnostic clinical laboratory. This variant scoring system uses an objective assessment by the acquisition of weighted evidence using five types of data including prediction tools, population frequency, co-occurrence, functional studies, and segregation. These variant assessments are conducted by trained scientists whose competency is regularly evaluated by standard training modules. The strength of this study is the number of pathogenicity assessments (n = 16,500), the consistent interobserver scoring of variants, and the reliability of the scores. The variant scoring system provides a standardized, rules-based iterative process for clinical grade variant scoring that tracks the history of variant classification. The database itself is valuable because it contains the cumulative knowledge from well-described disease mechanisms and disease population trends observed over years of testing patients in a clinical laboratory. The information, placed in the database by a standardized procedure is updated regularly by scientists trained in the collection of variant data to assure the quality of the variant scoring assessments. The uniformity of the process creates an awareness of the completeness and deficiencies of the data. Although the variant scoring system was operational before the recently published ACMG/AMP guidelines, it meets their recommendations [Richards et al., 2015]. We are in the process of incorporating certain supporting category elements of the ACMG guidelines that may not be well represented in our current iteration. We will then re-assess our scores generated with the new iteration of the scoring system to evaluate the impact of these changes. Updates to the scoring algorithm can be found at http://www.athenadiagnostics.com/athenainsight.

The variant evaluation process presents several challenges. The most common is determining the significance of variant frequencies in control populations. The variant frequencies can vary significantly in different geographic or ethnic populations, and the disease prevalence is not well-established for many rare disorders. The normal population frequency must be in excess of the disease prevalence before this data is used to score a variant. A recently available data set of 60,706 unrelated individual provided by the Exome Aggregation Consortium (ExAC) (http://exac.broadinstitute.org) provides a wealth of ethnically diverse variant frequency data to assist in resolving these issues. This data will be used as part of our scoring assessment (Supp. Table S2).

Many of the VUS are private, novel variants. To further clarify these private variants, we perform family segregation studies and collaborate with academia to perform functional studies. Sometimes we observe variants published as pathogenic at higher than expected frequencies compared with control populations [Norton et al., 2012; MacArthur et al., 2014]. Both published variants and variants in public databases are commonly misclassified as pathogenic [Bell et al., 2011] and are reclassified as either VUS, VUS-suggesting benign or benign after an expert review [Xue et al., 2012; Shearer et al., 2014; Tabor et al., 2014]. The variant scoring system described in this report minimizes these types of misclassifications by using multiple independent lines of evidence in a standardized, rules-based, weighted fashion.

Figure 2 shows the stability of the scoring system over time. Only 10 of 3,117 scores (0.3%) changed from likely benign to VUS or likely pathogenic to VUS. In these rare cases the scores were changed because of better control population data or because the family segregation data were weak. More importantly, no cases were rescored differently once designated as benign or pathogenic. These properties instill a high degree of confidence in each step of the variant scoring system. For example, a VUS with a score of 4 has lower odds of scoring up than does a VUS with a score of 5. It also has weaker odds of scoring lower than does a VUS with a score of 3. This shows the value of the three VUS subclasses and their corresponding data to mitigate drastic changes in pathogenicity designations. In addition, the three variant subclasses provide likelihood data to provide genetic counseling guidance and to assess the need for family segregation studies.

Studies of the effects of DNA variants on protein function are valuable in clarifying the pathogenicity of a variant. In 10% of our cases, published functional studies were either not clear or appeared to contradict other lines of evidence. Identifying the reasons for the incongruent results was challenging for several reasons. In some cases, different labs published results that contradicted each other. Typically, the experiments were performed in different cell lines or a slightly different assay was used to assess protein function. Experiments were sometimes performed on patient-derived cell lines or tissue samples. These types of experiments were difficult to interpret because of the effects of confounding genetic or cellular factors in interpreting whether the variant in question caused the aberrant protein function. The relationship between the variant's consequences and the molecular basis for disease was not always clear. The effect of a partial loss of function on the ability of molecular and metabolic pathways to tolerate such changes was difficult to determine. Functional studies that included analyses of both known damaging and known benign variants were the most valuable because the results allow comparisons between these variants and uncharacterized variants. This study, like the InSiGHT study [Thompson et al., 2014], shows that large data sets in standardized formats provide unbiased variant classification that help resolve the inherent difficulties in interpreting apparently discordant functional assays.

Prediction tools such as SIFT and PolyPhen-2 are based on scores that consider the position of amino acids in highly conserved protein domains because they are likely to be important for protein function. These programs, when they agree, have 89% accuracy for predicting damaging protein effects. However, they do not predict pathogenicity in all cases, because highly conserved sites often occur in blocks due to selection for a neighboring site and may tolerate some amino acid changes that are biochemically similar. In addition, not all functional sites are highly conserved. Therefore, the results SIFT and PolyPhen are not assigned as much weight as other data in the variant scoring system because of the degree of unreliability in their predictive value. Scores are adjusted if both agree, and then only when there is complimentary data [Stanley et al., 2014].

Conclusions

The use of multiple lines of evidence, the inclusion of three subclasses of VUS, standardized training, and a weighted rules-based scoring system are factors that are directly responsible for the stability and reliability of the variant pathogenicity scoring system described in this report. Replicating and testing this system in other clinical laboratories may enhance our ability to reliably assign variant pathogenicity scores and create standards for variant interpretation. Healthcare providers, patients and their families will benefit by the accurate interpretation of complex genetic information using a standardized variant scoring system such as we described in this report. This scoring system is currently applied to Mendelian disorders but may also identify non-Mendelian genetic risk factors or phenotypic modifiers. This rules-based scoring system may be designed to analyze somatic variants as well.

Acknowledgments

Disclosure statement: All authors are employees of Quest Diagnostics. Quest Diagnostics Laboratories offer extensive, fee-based genetic tests for molecular analyses. The variant scoring system is described in a pending patent.

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. Nat Methods 7:248–249.

Allen TC. 2015. Payment for cancer biomarker testing. Arch Pathol Lab Med 139:300–304.

Bean LJ, Tinker SW, da Silva C, Hegde MR. 2013. Free the data: one laboratory's approach to knowledge-based genomic variant classification and preparation for EMR integration of genomic data. Hum Mutat 34:1183–1188.

Bell CJ, Dinwiddie DL, Miller NA, Hateley SL, Ganusova EE, Mudge J, Langley RJ, Zhang L, Lee CC, Schilkey FD, Sheth V, Woodward JE, et al. 2011. Carrier testing for severe childhood recessive diseases by next-generation sequencing. Sci Transl Med 3:65ra64.

Desmet FO, Hamroun D, Lalande M, Collod-Beroud G, Claustres M, Beroud C. 2009. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res 37:e67.

Duzkale H, Shen J, McLaughlin H, Alfares A, Kelly MA, Pugh TJ, Funke BH, Rehm HL, Lebo MS. 2013. A systematic approach to assessing the clinical significance of genetic variants. Clin Genet 84:453–463.

Eggington JM, Bowles KR, Moyes K, Manley S, Esterling L, Sizemore S, Rosenthal E, Theisen A, Saam J, Arnell C, Pruss D, Bennett J, et al. 2014. A comprehensive laboratory-based program for classification of variants of uncertain significance in hereditary cancer genes. Clin Genet 86:229–237.

Fawcett KA, Murphy SM, Polke JM, Wray S, Burchell VS, Manji H, Quinlivan RM, Zdebik AA, Reilly MM, Houlden H. 2012. Comprehensive analysis of the TRPV4 gene in a large series of inherited neuropathies and controls. J Neurol Neurosurg Psychiatry 83:1204–1209.

Goldgar DE, Easton DF, Deffenbaugh AM, Monteiro AN, Tavtigian SV, Couch FJ, Breast Cancer Information Core Steering C. 2004. Integrated evaluation of DNA sequence variants of unknown clinical significance: application to BRCA1 and BRCA2. Am J Hum Genet 75:535–544.

Hellen CU. 2009. IRES-induced conformational changes in the ribosome and the mechanism of translation initiation by internal ribosomal entry. Biochim Biophys Acta 1789:558–570.

Holste D, Ohler U. 2008. Strategies for identifying RNA splicing regulatory motifs and predicting alternative splicing events. PLoS Comput Biol 4:e21.

Houdayer C, Caux-Moncoutier V, Krieger S, Barrois M, Bonnet F, Bourdon V, Bronner M, Buisson M, Coulet F, Gaildrat P, Lefol C, Leone M, et al. 2012. Guidelines for splicing analysis in molecular diagnosis derived from a set of 327 combined in silico/in vitro studies on BRCA1 and BRCA2 variants. Hum Mutat 33:1228–1238.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46:310–315.

Kuk AY, Nott DJ, Yang Y. 2014. A stepwise likelihood ratio test procedure for rare variant selection in case-control studies. J Hum Genet 59:198–205.

Kumar P, Henikoff S, Ng PC. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4:1073–1081.

Lindor NM, Guidugli L, Wang X, Vallee MP, Monteiro AN, Tavtigian S, Goldgar DE, Couch FJ. 2012. A review of a multifactorial probability-based model for classification of BRCA1 and BRCA2 variants of uncertain significance (VUS). Hum Mutat 33:8–21.

MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, Adams DR, Altman RB, Antonarakis SE, Ashley EA, Barrett JC, Biesecker LG, et al. 2014. Guidelines for investigating causality of sequence variants in human disease. Nature 508:469–476.

Norton N, Robertson PD, Rieder MJ, Zuchner S, Rampersaud E, Martin E, Li D, Nickerson DA, Hershberger RE, National Heart L, Blood Institute GOESP. 2012. Evaluating pathogenicity of rare variants from dilated cardiomyopathy in the exome era. Circ Cardiovasc Genet 5:167–174.

Ott J. 1999. The informativeness of family data. In: Ott J, editor. Analysis of human genetic linkage. Baltimore, MD: The Johns Hopkins University Press. p 106–113.

Pastrello C, Pin E, Marroni F, Bedin C, Fornasarig M, Tibiletti MG, Oliani C, Ponz de Leon M, Urso ED, Della Puppa L, Agostini M, Viel A. 2011. Integrated analysis of unclassified variants in mismatch repair genes. Genet Med 13:115–124.

Pertea M, Lin X, Salzberg SL. 2001. GeneSplicer: a new computational method for splice site prediction. Nucleic Acids Res 29:1185–1190.

Plon SE, Eccles DM, Easton D, Foulkes WD, Genuardi M, Greenblatt MS, Hogervorst FB, Hoogerbrugge N, Spurdle AB, Tavtigian SV, Group IUGVW. 2008. Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. Hum Mutat 29:1282–1291.

Reese MG, Eeckman FH, Kulp D, Haussler D. 1997. Improved splice site detection in Genie. J Comput Biol 4:311–323.

Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL. 2015. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 17:405–424.

Rutten JW, Boon EM, Liem MK, Dauwerse JG, Pont MJ, Vollebregt E, Maat-Kievit AJ, Ginjaar HB, Lakeman P, van Duinen SG, Terwindt GM, Lesnik Oberstein SA. 2013. Hypomorphic NOTCH3 alleles do not cause CADASIL in humans. Hum Mutat 34:1486–1489.

Shearer AE, Eppsteiner RW, Booth KT, Ephraim SS, Gurrola J, 2nd, Simpson A, Black-Ziegelbein EA, Joshi S, Ravi H, Giuffre AC, Happe S, Hildebrand MS, et al. 2014. Utilizing ethnic-specific differences in minor allele frequency to recategorize reported pathogenic deafness variants. Am J Hum Genet 95: 445–453.

Stanley CM, Sunyaev SR, Greenblatt MS, Oetting WS. 2014. Clinically relevant variants-identifying, collecting, interpreting, and disseminating: the 2013 annual scientific meeting of the Human Genome Variation Society. Hum Mutat 35:505–510.

Tabor HK, Auer PL, Jamal SM, Chong JX, Yu JH, Gordon AS, Graubert TA, O'Donnell CJ, Rich SS, Nickerson DA, Project NES, Bamshad MJ. 2014. Pathogenic variants for Mendelian and complex traits in exomes of 6,517 European and African Americans: implications for the return of incidental results. Am J Hum Genet 95:183–193.

Thompson BA, Spurdle AB, Plazzer JP, Greenblatt MS, Akagi K, Al-Mulla F, Bapat B, Bernstein I, Capella G, den Dunnen JT, du Sart D, Fabre A, et al. 2014. Application of a 5-tiered scheme for standardized classification of 2,360 unique mismatch repair gene variants in the InSiGHT locus-specific database. Nat Genet 46:107–115.

Vihinen M. 2013. Guidelines for reporting and using prediction tools for genetic variation analysis. Hum Mutat 34:275–282.

Wang J, Schmitt ES, Landsverk ML, Zhang VW, Li FY, Graham BH, Craigen WJ, Wong LJ. 2012. An integrated approach for classifying mitochondrial DNA variants: one clinical diagnostic laboratory's experience. Genet Med 14:620–626.

Waples RS. 1988. Estimation of allele frequencies at isoloci. Genetics 118:371–384.

Xue Y, Chen Y, Ayub Q, Huang N, Ball EV, Mort M, Phillips AD, Shaw K, Stenson PD, Cooper DN, Tyler-Smith C, Genomes Project C. 2012. Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. Am J Hum Genet 91:1022–1032.

Yeo G, Burge CB. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol 11:377–394.

Zhang MQ. 1998. Statistical features of human exons and their flanking regions. Hum Mol Genet 7:919–932.