INFORMATICS

Human Mutation

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

Assessing the Pathogenicity of Insertion and Deletion Variants with the Variant Effect Scoring Tool (VEST-Indel)

Christopher Douville,[1†] David L. Masica,[1†] Peter D. Stenson,[2] David N. Cooper,[2] Derek M. Gygax,[3] Rick Kim,[3] Michael Ryan,[3] and Rachel Karchin[1,4*]

[1]Department of Biomedical Engineering and Institute for Computational Medicine, The Johns Hopkins University, Baltimore, Maryland; [2]Institute of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff, UK; [3]In Silico Solutions, Fairfax, Virginia; [4]Department of Oncology, Johns Hopkins University School of Medicine, Baltimore, Maryland

Communicated by William S. Oetting

Received 9 July 2015; accepted revised manuscript 14 September 2015.
Published online 7 October 2015 in Wiley Online Library (www.wiley.com/humanmutation). DOI: 10.1002/humu.22911

ABSTRACT: Insertion/deletion variants (indels) alter protein sequence and length, yet are highly prevalent in healthy populations, presenting a challenge to bioinformatics classifiers. Commonly used features—DNA and protein sequence conservation, indel length, and occurrence in repeat regions—are useful for inference of protein damage. However, these features can cause false positives when predicting the impact of indels on disease. Existing methods for indel classification suffer from low specificities, severely limiting clinical utility. Here, we further develop our variant effect scoring tool (VEST) to include the classification of in-frame and frameshift indels (VEST-indel) as pathogenic or benign. We apply 24 features, including a new "PubMed" feature, to estimate a gene's importance in human disease. When compared with four existing indel classifiers, our method achieves a drastically reduced false-positive rate, improving specificity by as much as 90%. This approach of estimating gene importance might be generally applicable to missense and other bioinformatics pathogenicity predictors, which often fail to achieve high specificity. Finally, we tested all possible meta-predictors that can be obtained from combining the four different indel classifiers using Boolean conjunctions and disjunctions, and derived a meta-predictor with improved performance over any individual method.

Hum Mutat 37:28–35, 2016. Published 2015 Wiley Periodicals, Inc.*

KEY WORDS: insertion deletion variant; indel; in-frame frameshift; bioinformatics pathogenicity predictor; meta-predictor

Introduction

The average human exome contains over four hundred naturally occurring microinsertion/deletion variants (referred to as indels throughout) [Consortium, 2010]. In-frame indels account for ~50% of these variants, and result from the insertion/deletion of an integer number of codons, and ultimately amino acids. Frameshifts account for the other ~50% of indels, and result from contiguous nucleotide insertions/deletions of a length not divisible by three. This fractional change in the number of codons shifts the translational reading frame, resulting in an entirely new downstream sequence thereby shifting the position at which the first stop codon is encountered. Thus, frameshift indels translate to protein that is very distinct from the native protein, particularly if the indel occurs early in the transcript sequence. Both in-frame and frameshift indels alter protein sequence and length.

Because they drastically alter protein primary structure, but are also highly prevalent in healthy populations, indels present a unique classification challenge. Clearly, the principles governing pathogenicity are not identical to those governing changes in protein function and stability; otherwise, most indels would be pathogenic. The challenge then arises because protein sequence, structure, function, and stability are typically considered when assessing a variant of unknown impact on disease liability [Steward et al., 2003]. These protein-based criteria could lead to a high false positive rate, and therefore low specificity, because the fraction of indels that appreciably impact health might be overestimated.

The increased utilization of high-throughput genomic sequencing technologies and hopes for their clinical application, coupled with the high prevalence of indels, has led to a demand for bioinformatic predictors of indel pathogenicity. Most computational methods for assessing genetic variation initially focused on missense variants; more recently, several groups have extended these methods to handle indels [Choi et al., 2012; Hu and Ng, 2012; Hu and Ng, 2013; Zhao et al., 2013; Folkman et al., 2015]. Most of these methods utilize supervised machine learning classifiers and are trained on two classes of indel: pathogenic from disease mutation databases and benign from either population variation databases or tolerated interspecies variations derived from genomic alignments. DDIG-in is based on a support vector machine, and the authors of this method reported a sensitivity of 0.86 and specificity of 0.72 for frameshift indels [Zhao et al., 2013], and a sensitivity of 0.89 for in-frame indels [Folkman et al., 2015]; the authors did not report prediction specificity for in-frame indels. PROVEAN uses an unsupervised approach that compares the reference protein sequence with a sequence that incorporates a variant of interest [Choi et al., 2012]. The authors of PROVEAN reported high sensitivities of 0.93 and 0.96 for in-frame insertions and deletions, respectively, and a specificity of 0.80 for in-frame insertions and 0.68 for in-frame deletions; PROVEAN does not assess frameshift indels. SIFT-indel, based on a J48 Decision Tree [Hall et al., 2009], achieved good balanced accuracies for

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

[†]These authors contributed equally to this work.

[*]Correspondence to: Rachel Karchin, Department of Oncology, Johns Hopkins University School of Medicine, Baltimore, MD. E-mail: karchin@jhu.edu

Contract grant sponsor: National Institutes of Health, National Cancer Institute (NIH NCI U01CA180956).

in-frame (sensitivity = 0.81; specificity = 0.82) [Hu and Ng, 2013] and frameshift indels (sensitivity = 0.90; specificity = 0.78) [Hu and Ng, 2012]. However, the neutral dataset used in those studies comprised indels derived from cross-species comparisons. As the authors state, SIFT-indel was trained to predict impact on gene function, irrespective of impact on disease. Indeed, when the method was applied to variants from human variation databases, the majority of the indels were predicted to be deleterious; thus, specificities would be below 50% for predicting indel pathogenicity. The CADD classifier utilized a unique approach, in which a support vector machine was trained to discriminate fixed (or nearly fixed) derived alleles in humans from a set of simulated variants [Kircher et al., 2014]. The CADD classifier was developed to predict deleterious variants rather than variant pathogenicity or impact on protein function, but with the stated assumption that these quantities are all related. The authors of CADD reported classifier performance on missense variants and indels together, but not on indels separately [Kircher et al., 2014].

We present a novel method for assessing the impact of indels with an emphasis on discriminating between benign and disease-causing indels. This new approach expands the functionality of our Variant Effect Scoring Tool (VEST), which was initially designed to assess the impact of missense variants; we refer to the new functionality as VEST-indel, throughout. In addition to more conventional protein sequence and functional considerations, VEST-indel includes a feature based on PubMed search results for the gene of interest, which is a measure of known relevance to human health. Utilizing one dataset for cross-validation and a second, non-overlapping dataset for testing, we report high sensitivity and specificity for both in-frame and frameshift indels. We provide direct comparison between VEST-indel and existing indel prediction methods by constructing an additional benchmark dataset of neutral and disease-causing indels on which none of the methods had been trained, and show that VEST-indel achieves improved performance. Next, we exhaustively test Boolean combinations of all tested classifiers and identify a meta-predictor that further improves performance relative to any of the methods individually. Finally, we used Tajima's D statistic [Tajima, 1989] to detect signatures of positive, balanced, or relaxed selection for variants that resulted in false positive predictions. The presence of these selective pressures could suggest a context dependence in which indels defined as neutral could also be considered pathogenic, in turn justifying false positive predictions.

Materials and Methods

Data Collection

A curated set of in-frame and frameshift indels (microdeletions and microinsertions) of ≤20 base-pairs in length, annotated as being pathogenic from publications in the biomedical literature, was downloaded from Human Gene Mutation Database [Stenson et al., 2014] (2014v.3). Only high-confidence annotations with the "DM" designation were included. A second curated set of in-frame and frameshift indels was downloaded from the NCBI ClinVar database on August 7, 2014. Only entries annotated as "likely pathogenic" (Clinical Significance 4) or "pathogenic" (Clinical Significance 5) and not annotated as a somatic mutation were included. Any entry from ClinVar that was also present in HGMD was removed from the ClinVar set. Annotated in-frame and frameshift variants were downloaded from the Exome Variant Server using (ESP6500SI-V2-SSA137) [Fu et al., 2013] and from the 1000 Genomes Project Phase 3 (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/) [Clarke et al., 2012]. To increase the likelihood that variants from the Exome Variant Server and 1000 Genomes Project were benign common polymorphisms, and to retain sufficient variants for our training set, we only used variants with a minor allele frequency (MAF) ≥0.01 and occurring in either African individuals or those of African ancestry. In ESP600, these were identified as "African-American" and in 1000G as the AFR superpopulation comprising YRI (Yoruba in Ibadan, Nigeria), LWK (luhya in Webuye, Kenya), GWD (Gambian in Western Divisions in the Gambia), MSL (Mende in Sierra Leone), ESN (Esan in Nigeria), ASW (Americans of African ancestry in SW USA), and ACG (African Caribbeans in Barbados). The other populations represented in ESP6500 and 1000G are believed to have experienced severe bottlenecks in recent history, and hence individuals from these populations may harbor potentially pathogenic variants at higher MAF than individuals of African ancestry [Lohmueller et al., 2008; MacArthur and Tyler-Smith, 2010; Ng et al., 2008; Hu and Ng, 2012]. A curated set of putatively benign in-frame and frameshift indels, derived from pairwise genome alignments of human and cow, dog, horse, chimpanzee, rhesus macaque, and rat, was generously provided to us by Pauline Ng and Jing Hu. This set had been previously used to train their SIFT-indel classifier [Hu and Ng, 2012; Hu and Ng, 2013]. Additional background information about these data sets, including probability densities for indel length and MAF, are shown in Supp. Figure S1.

The number of variants used for this study, grouped by source and ontology, were: 2,523 in-frame deletions, 565 in-frame insertions, 17,606 frameshift deletions, and 8,265 frameshift insertions from HGMD 2014.3; 43 in-frame deletions, 14 in-frame insertions, 344 frameshift deletions, and 134 frameshift insertions from HGMD 2014.4; 1,991 in-frame deletions, 404 in-frame insertions, 774 frameshift deletions, and 618 frameshift insertions from ESP6500; 86 in-frame deletions, 70 in-frame insertions, 37 frameshift deletions, and 23 frameshift insertions from 1000 Genomes, Phase 1; 304 in-frame deletions, 261 in-frame insertions, 229 frameshift deletions, and 134 frameshift insertions from 1000 Genomes, Phase 3; 16 in-frame deletions, five in-frame insertions, 32 frameshift deletions, and 74 frameshift insertions from ClinVar; 4,686 in-frame deletions, 3,406 in-frame insertions, 706 frameshift deletions, and 628 frameshift insertions from the above-mentioned genome alignments.

Feature Selection

The Random Forest Feature Importance Z-score [Breiman, 2001] was used to rank a set of 49 candidate features from [Wong et al., 2011] and five additional features (Supp. Table S1), using PARF software (http://code.google.com/p/parf), with 100 trees and default parameters. To avoid overfitting, an independent feature-selection set was used (500 pathogenic and 500 benign examples for each of the in-frame and the frameshift classifiers). We used a greedy algorithm to identify a good, minimum set of features. Briefly, beginning with the top-ranked feature, a Random Forest was trained using only that feature and 10-fold cross-validation was used to estimate the classifier's area under the receiving operator characteristic (ROC) curve (AUC). We successively added the next top-ranked feature until all candidate features were included. For the in-frame classifier, the maximum AUC was achieved with 23 features and for the out-of-frame classifier, the maximum AUC was achieved with 16 features (Supp. Table S2). These features were used for the remainder of the work described here. The selected features include measures of gene importance, the damaging effect of the variant on protein activity, evolutionary conservation and protein local environment (Supp. Table S3).

Table 1. Datasets Used in Development of the VEST-Indel Method

	Feature selection	Training	Testing	Null distribution	Multi-method benchmark
In-frame					
Pathogenic	500[a]	2,475[a]	39[b]	N/A	57[f]
Benign	500[c]	1,877[c]	8,105[d]	346[e]	156[e]
Frameshift					
Pathogenic	500[a]	24,478[a]	184[b]	N/A	478[f]
Benign	500[c]	1,350[c]	1,340[d]	537[e]	60[e]

Superscript letters indicate the source of the examples for each type of insertion/deletion variant and each stage of VEST-indel development (feature selection, classifier training, classifier validation, empirical null).

[a] HGMD.

[b] ClinVar.

[c] ESP6500.

[d] Inter-species benigns from SIFT-indel.

[e] 1000G Phase 3.

[f] HGMD2014.

There is no overlap between examples in any of the columns. N/A, not applicable because only benign examples were used to develop the empirical null distribution.

Classifier Training Protocol

Random Forest classifiers were trained to classify in-frame and frameshift variants (using PARF software with 100 trees and default parameters). For in-frame classifier training, 2,475 pathogenic and 1,877 benign examples were available, whereas 24,478 pathogenic and 1,350 benign examples were available for frameshift classifier training (Table 1). To handle class imbalance, the in-frame classifier was trained on a randomly selected set of 1,877 pathogenic examples and all 1,877 benign examples. Ten frameshift classifiers were trained on a randomly selected set of 1,350 pathogenic examples and all 1,350 benign examples (repeated 10 times, sampling without replacement). All ten classifiers were used to score frameshift variants, by computing 10 scores for each variant and averaging them.

Modeling the Analytical Null Distribution

We developed an analytical null score distribution based on Random Forest classifier scores of putative benign variants (1000 Genomes Project Phase 3, MAF ≥ 0.01, African ancestry). The scored variants (537 for in-frame insertion/deletions, 346 for frameshift insertion/deletions) did not overlap the examples used for Random Forest feature selection, training or the independent test set. An empirical cumulative distribution of scores was calculated and modeled as a Generalized Pareto Distribution (GPD) [Pickands III, 1975; Knijnenburg et al., 2009].

$$F(z) = \left\{ \frac{1 - \left(\frac{kz}{a}\right)^{\frac{1}{k}},\ \ k! = 0}{1 - e^{-\frac{z}{a}},\ \ k = 0} \right. \qquad (1)$$

where k and a are the GPD shape and scale parameters, respectively (in-frame $k = 0$, $a = -8.48$, frameshift $k = 0$, $a = -9.04$) and z is the score ($0 \leq z \leq 1$).

Analytical Null Distribution for Missense Variants

We developed an analytical null distribution by modeling a GPD based on an empirical null distribution of VEST missense Random Forest classifier scores (Eq (1), $k = 0$, $a = -5.26$). The missense Random Forest [Carter et al., 2013] was trained on 47,724 pathogenic missense mutations from HGMD v2012.2 and 45,818 putatively benign missense variants from ESP6500 accessed 07/2012. The empirical null scores were generated from 28,509 variants from the 1000 Genomes Project with AF ≥ 1%, none of which overlapped with training data.

Combined Prioritization of Indel and Missense Variants

For each In-frame, frameshift, or missense variant, a VEST score was computed using homology-restricted 10-fold cross-validation with the appropriate Random Forest (Performance Assessment). Then, a *P* value was calculated using the analytical null (Eq. (1)) for its respective type. To assess whether VEST could correctly prioritize a pathogenic variant over a benign variant, irrespective of whether the variant was in-frame, frameshift or missense, we ranked the combined set of variants according to p-value, and computed area under the ROC curve [Fawcett, 2004].

To compare VEST results with CADD, which also provides combined prioritization, we scored the same variants using the CADD Webserver (*Materials and Methods*). Variants were ranked according to their scaled *C*-scores and area under the ROC curve was computed.

Performance Assessment

Homology-restricted cross-validation

The in-frame indel Random Forest and each of the 10 frameshift Random Forests were assessed for sensitivity, specificity and balanced accuracy, using a rigorous ten-fold cross-validation protocol. The same protocol was applied to the missense Random Forest to assess combined prioritization of all three mutation types. To avoid overestimating our performance, we ensured that any examples from genes whose protein products had ≥35% sequence identity were included in the same fold. BlastP with default parameters [Altschul et al., 1997] was used for pairwise alignment of protein sequences and sequence identity calculations [Altschul et al., 1997]. Evidence suggests that homology-restricted cross-validation is important to avoid overly optimistic estimates of pathogenicity classifier performance [Capriotti and Altman, 2011].

VEST-indel independent test set

An independent set of examples, having no overlap with feature selection, training, or empirical null sets, was constructed. We removed any test set examples whose protein products had ≥35% sequence identity with any training examples [Altschul et al., 1997]. Pathogenic in-frame and frameshift mutations were taken from ClinVar [Landrum et al., 2013] and benign in-frame and frameshift variants were taken from the interspecies set (Data Collection). All data were "cleaned" so as to ensure that there was no overlap between these examples and the other three data sets. The VEST-indel independent test set is included in Additional File 1 (http://karchinlab.org/vest_indel_additional_files/Additional_File_1.xlsx).

Comparison of Insertion/Deletion Variant Pathogenicity Predictors

Four previously published methods were selected for comparison with VEST-indel. Three of the methods (SIFT-indel, DDIG-in, CADD) [Hu and Ng, 2012; Hu and Ng, 2013; Zhao et al., 2013; Folkman et al., 2015] handle both in-frame and frameshift insertion/deletions and the fourth method PROVEAN [Choi et al., 2012]

Table 2. Training and Validation Sets Used by Current Prediction Methods

	Training set (as published)		Test set (as published)		Non-overlapping multi-method benchmark set	
	Pathogenic	Benign	Pathogenic	Benign	Pathogenic	Benign
In-frame						
PROVEAN	Uniprot	Uniprot	HGMD 2011	1000G P1	HGMD2014.4	1000G P3 AA
DDIG-in	HGMD 2012	1000G P1	Uniprot	Uniprot	HGMD2014.4	1000G P3 AA
SIFT-indel	HGMD 2010	Interspecies	Uniprot	Uniprot	HGMD2014.4	1000G P3 AA
CADD	Simulated	Fixed Polymorphisms	ClinVar	ESP6500	HGMD2014.4	1000G P3 AA
VEST-indel	HGMD 2014.3	ESP6500 AA	ClinVar	Interspecies	HGMD2014.4	1000G P3 AA
Frameshift						
PROVEAN	N/A	N/A	N/A	N/A	N/A	N/A
DDIG-in	HGMD 2012	1000G P1	HGMD 2012	Interspecies	HGMD2014.4	1000G P3 AA
SIFT-indel	HGMD 2010	Interspecies	N/A	N/A	HGMD2014.4	1000G P3 AA
CADD	Simulated	Fixed Polymorphisms	ClinVar	ESP6500	HGMD2014.4	1000G P3 AA
VEST-indel	HGMD 2014.3	ESP6500 AA	ClinVar	Interspecies	HGMD2014.4	1000G P3 AA

1000G P1 and *1000G P3* are variants from 1000 Genomes Phase 1 and 3, respectively. *Interspecies* benign variants derived from pairwise genome alignments of human and cow, dog, horse, chimp, rhesus macaque, and rat. *Uniprot* variants were obtained from the UniProtKB/Swiss-Prot "Human Polymorphisms and Disease Mutations" dataset (Release 2011_09), annotated as deleterious, neutral, or unknown based on keywords from the provided Uniprot descriptions. AA, African or African American Ancestry and N/A, not applicable.

handles only in-frame variants. According to their publications, each of these methods used a customized dataset for classifier training and validation. Data sources included the UniProtKB, HGMD versions from 2010, 2011, 2012 and 2014, the ESP6500, the 1000G project, an interspecies collection of putatively benign variants developed by the SIFT-indel team, a set of variants fixed in recent evolution and a simulated set of variants developed by the CADD team [Boutet et al., 2007; Consortium, 2010; Hu and Ng, 2012; Fu et al., 2013; Hu and Ng, 2013; Stenson et al., 2014] (Table 2). To our knowledge, SIFT-indel, DDIG-in, PROVEAN and CADD are the four most widely used methods in the field. All provide an easy-to-use Web interface that can handle batch submissions: PROVEAN URL is http://provean.jcvi.org/genome_submit_2.php?species=human, SIFT-indel URL is http://sift.bii.a-star.edu.sg/www/SIFT_indels2.html, DDig-In URL is http://sparks-lab.org/ddig/, and CADD URL is http://cadd.gs.washington.edu/. The Web interfaces ensured that we applied each of these methods in the manner intended by their respective authors.

Multi-method benchmark set

To perform an unbiased comparison of VEST-indel and the four other methods, we identified a set of 553 pathogenic and 357 benign examples, which did not overlap with any examples used to train, fit parameters, select features, or validate performance by any of the four methods. This *multi-method benchmark set* comprised pathogenic examples (61 in-frame and 491 frameshift) from the most recent version of HGMD (2014v.4), excluding any examples present in earlier versions of HGMD that had been used to train DDIG-in, SIFT-indel, or VEST-indel (PROVEAN and CADD were not trained on HGMD). Benign examples (224 in-frame and 118 frameshift) were taken from 1000G Phase 3 (MAF ≥ 0.10, African Ancestry). Any examples present in 1000G Phase 1 or ESP6500, which were used to train or validate any of PROVEAN, DDIG-in, CADD or VEST-indel were omitted. Only examples for which every method returned a prediction result were included. The final multi-method benchmark set comprised 59 benign frameshift insertion/deletion and 163 benign in-frame insertion/deletion variants (MAF ≥ 0.1 from 1000G AFR superpopulation), as well as 474 pathogenic frameshift and 53 pathogenic in-frame variants from HGMD v.2014.4 (Additional File 2 http://karchinlab.org/vest_indel_additional_files/Additional_File_2.xlsx).

Performance assessment

SIFT-indel and DDIG-in provide a categorical classification for each example (damaging/neutral or disease/neutral) and a confidence measure. PROVEAN, CADD, and VEST-indel provide a numerical score for each example (–40 to 12.5, 1 to 99, and 0 to 1). To compare methods, PROVEAN scores were assigned to categories of damaging (<–2.5) or neutral (≥–2.5) (as recommended by the authors) and VEST-indel scores were assigned to categories of pathogenic (≥0.5) or benign (<0.5), which represents a majority vote of decision trees in the Random Forest classifier. CADD scaled "*C*-scores" were assigned to categories of deleterious (<15) or not deleterious (≥15) (as recommended on their Webserver). Sensitivity (TP/(TP + FN)), specificity (TN/(TN + FP)) and balanced accuracy ((sensitivity + specificity)/2) were calculated for each method, where TP = the number correctly classified as pathogenic (or damaging or disease) examples, FN = the number of incorrectly classified pathogenic examples, TN = the number of correctly classified benign (or neutral) examples, and FP = the number of incorrectly classified benign examples.

Meta-Predictors that Combine Classifications of Multiple Methods

In these Boolean expressions, each method is represented by a variable X_i, which is set to TRUE when the method classifies an example as pathogenic and FALSE when the method classifies an example as benign. For combinations of two methods, candidate meta-predictors were (X_1 and X_2) and (X_1 or X_2). For combinations of three methods, candidate meta-predictors (X_1 and X_2 and X_3), (X_1 or X_2 or X_3),(X_1 or X_2 or X_3), ((X_1 and X_2) or X_3), ((X_1 or X_2) and X_3), ((X_1 and X_3) or X_2),((X_1 or X_3) and X_2), ((X_2 and X_3) or X_1), ((X_2 or X_3) and X_1). For combinations of four methods, there are 64 possible combinations (Supp. Table S4). We used a brute-force approach and limited the number of methods in the meta-predictor to a maximum of four to avoid a combinatorial explosion. All possible four-way combinations of the five methods were explored.

Selective Pressures on Genes and False Positive Classifications

A standard method for identifying genes under selection is Tajima's D statistic [Tajima, 1989], and for each gene harboring, a variant in the multi-method benchmark set, we computed this statistic based on its longest annotated RefSeq transcript [Pruitt et al., 2007]. If RefSeq transcripts were not available for the gene, the longest annotated Ensembl transcript [Cunningham et al., 2015] was used. These calculations were performed using SNPs in 1000 Genomes Phase 3 AFR samples and the PopGenome package in R [Pfeifer et al., 2014]. Each gene was assessed for the presence of statistically significant positive or balancing selection ($P < 0.05$). The PopGenome package estimates P values by simulation using Hudson's coalescent model [Hudson, 2002].

Availability

VEST-indel is freely available for non-commercial use via a batch Web service at http://cravat.us. Choose the analysis program "VEST". Results for small-size submissions (<100) are returned on average in less than five minutes. Very large submissions are supported, and results for a submission of ∼1,000,000 mutations are returned on average in less than 24 hr.

Results

In protein-coding exons, in-frame indels generally have a less severe impact than frameshifts [Ng et al., 2008]. Since the biological effect of in-frame and frameshift indels is different, we chose to develop two distinct Random Forest classifiers to handle these two distinct variant types. We assessed the performance of the classifiers in three phases: (1) we estimated their sensitivity and specificity with stringent, homology-restricted 10-fold cross-validation (pathogenic class from HGMDv2014.3 [Stenson et al., 2014], benign from ESP6500 African Ancestry [Fu et al., 2013]) (Table 2); (2) we re-estimated sensitivity and specificity on an independent test set of variants (pathogenic class from ClinVar [Landrum et al., 2013], benign Interspecies alignments [Hu and Ng, 2012; Hu and Ng, 2013]) (Table 2) that had not been used in classifier training and had been filtered for homology overlap with the cross-validation set; (3) we re-estimated sensitivity and specificity on a second independent test set of variants (pathogenic class from new entries in HGMDv2014.4, benign from 1000 Genomes Phase III African Ancestry [Consortium, 2010] that did not overlap with any training data used by previously published methods (*multi-method benchmark set*). These experiments are detailed in *Materials and Methods.*

In our cross-validation experiments, VEST-indel achieved a sensitivity and specificity of 0.90 for in-frame indels (Table 3). Cross-validation performance for frameshift indels was slightly lower, with a sensitivity of 0.83, specificity of 0.88, and balanced accuracy of 0.85. During the testing phase, VEST-indel maintained good performance for classifying in-frame indels, with a balanced accuracy of 0.82, and sensitivity and specificity of 0.80 and 0.85, respectively (Table 3). Frameshift prediction improved slightly from the cross-validation to testing phase. For testing, frameshift classification resulted in a balanced accuracy of 0.87, with a sensitivity of 0.89 and specificity of 0.86.

Table 4 compares VEST-indel performance with that previously reported for other methods. The VEST-indel specificity of 0.90 for classifying in-frame indels was higher than that achieved by SIFT-indel (0.82) or PROVEAN (0.80 for insertions and 0.68 for deletions); the authors of the DDIG-in method did not provide a specificity for in-frame classification. The specificity of 0.88 achieved by VEST-indel for classifying frameshift indels was 0.10 higher than that achieved by SIFT-indel and 0.16 higher than that reported for DDIG-in (Table 3); PROVEAN does not classify frameshift indels. The higher specificity achieved by VEST-indel results from a relatively low false positive rate, indicating improved ability to discriminate neutral indels from those that are pathogenic.

Four of the compared methods attained reasonably high sensitivities, demonstrating the ability to identify truly pathogenic variants (Table 4). PROVEAN had the highest sensitivity for classifying in-frame variants, which was higher that realized by VEST-indel (0.90), DDIG-in (0.89), or SIFT-indel (0.81). Conversely, the SIFT-indel sensitivity of 0.90 was highest among compared methods for classifying frameshift indels, with DDIG-in at 0.86 and VEST-indel at 0.83. The authors of the CADD method do not report sensitivity or specificity statistics for either in-frame or frameshift indels [Kircher et al., 2014].

Although the above-cited comparison of previously published results provides some indication of relative performance, the five methods compared in Table 4 utilized different datasets for cross-validation and testing (see Table 2). The datasets being different, to some extent, limits the insights that can be gleaned from

Table 3. VEST-Indel Performance Metrics

	Sensitivity	Specificity	Balanced accuracy
In-frame cross validation	0.90	0.90	0.90
In-frame testing	0.80	0.85	0.82
Frameshift cross validation	0.83	0.88	0.85
Frameshift testing	0.89	0.86	0.87

Training utilized 10-fold cross validation and pathogenic variants from Human Gene Mutation Database 2014.3 and benign examples from Exome Sequencing Project (minor allele frequency in African Ancestry ≥ 0.01). The test set consisted of pathogenic examples from ClinVar and benign examples derived from pairwise genome alignments of human and cow, dog, horse, chimp, rhesus macaque, and rat.

Table 4. Comparing Performance with Previously Published Results and Testing all Methods with the New Multi-Method Benchmark Dataset

	Previously published		Multi-method benchmark		
	Sensitivity	Specificity	Sensitivity	Specificity	Balanced Accuracy
In-frame					
VEST-indel	0.90[a]	0.90[a]	0.81	0.96	0.88
SIFT-indel	0.81	0.82	0.86	0.76	0.81
DDIG-in	0.89	N/A	0.78	0.91	0.84
PROVEAN	0.93/0.96	0.80/0.68	0.95	0.80	0.88
CADD	N/A	N/A	0.74	0.88	0.81
Frameshift					
VEST-indel	0.83[a]	0.88[a]	0.85	0.95	0.90
SIFT-indel	0.90	0.78	0.94	0.25	0.59
DDIG-in	0.86	0.72	0.75	0.80	0.77
CADD	N/A	N/A	0.98	0.05	0.52

Previously published sensitivity and specificity based on author's cross-validation experiments. PROVEAN does not use cross validation so the reported numbers are from validation set experiments done separately for insertion and deletion variants. N/A, not applicable. Published results for the DDIG-in in-frame classifier do not include specificity; their self-reporting consists of an accuracy (not balanced accuracy) of 0.84 and precision of 0.81. The authors of CADD did not report the performance achieved with indels separately.

[a] Results from Table 1 included here for comparison. Multi-method benchmark set consisted of pathogenic examples from Human Gene Mutation Database 2014.4 and benign examples 1000 Genomes Phase 3 (minor allele frequency in African Ancestry ≥ 0.1).

comparison. In addition to using the same dataset to benchmark all methods, direct and fair comparison requires that no training examples from any of the methods be present in the benchmark dataset. We constructed a *multi-method benchmark set*, comprising indels not present in any method's training set, by using mutations recently added to HGMD and variants from the latest release of the 1000 Genomes project (Phase 3; Table 2). These indels were new to all of the methods, and were intended to reduce overly optimistic estimates of performance.

Table 4 compares performance achieved by the five methods for classifying neutral and disease-causing indels from the multi-method benchmark set. VEST-indel shows superior specificity for classifying both in-frame (0.96) and frameshift (0.95) indels. These high specificities further validate the ability of VEST-indel to accurately reject neutral variants as disease causing. All methods had reasonably high balanced accuracies for in-frame indel classification, with VEST-indel and PROVEAN yielding the highest balanced accuracy of 0.88. Of note, VEST-indel and PROVEAN achieved nearly identical balanced accuracies with approximately equal trade-offs in sensitivity and specificity (Table 3). For frameshift variants, VEST-indel outperformed the other methods, having a balanced accuracy of 0.90, compared with 0.77 for DDIG-in, 0.59 for SIFT-indel, and 0.52 for CADD. In the case of DDIG-in, VEST-indel showed substantially improved sensitivity and specificity (Table 4). The dramatic gain in performance achieved by VEST-indel, relative to SIFT-indel and CADD, resulted from a marked gain in specificity (0.95 vs. 0.25 for SIFT, and 0.05 for CADD); this is consistent with previous reports for SIFT-indel, which maintains good specificity when predicting protein-damaging indels, but suffers low specificity when predicting pathogenicity [Hu and Ng, 2012; Hu and Ng, 2013].

Although they are often confounded, bioinformatics prediction of protein-damaging variation is not necessarily the same as prediction of pathogenic/disease-causing variation [Capriotti et al., 2012]. A variant that reduces protein stability, function, or even results in complete loss of protein production is certainly protein damaging, but implications for health and disease will also depend on the importance of that particular protein in complex networks of interacting molecules [Capriotti et al., 2012]. Thus, although all disease-causing variants are likely to be protein damaging, all protein-damaging variants are not invariably disease causing. A good set of predictive features should capture both protein damage and gene importance. VEST-indel uses 23 features for in-frame classification and 16 features for frameshift classification; these features were top-ranked by a Random Forest Z-score feature selection method and a greedy algorithm that maximized Random Forest ROC AUC (see *Materials and Methods*). Features include measures of DNA sequence conservation, DNA natural variation, gene-level annotations, transcript-level annotations, computational predictions of protein local structure, protein local regional sequence composition and protein-level annotations from UniProtKB (Supp. Table S1). Some of these features are similar to those adopted by previously published methods, whereas others have not been previously applied to indel classification. In the first category, we include DNA sequence conservation scores, indel length, indel location within the transcript, predicted highly flexible or disordered regions in protein structure, predicted solvent accessibility in protein structure, and occurrence of the indel within a repetitive (low-information) sequence. In the second category, we include a hidden-Markov model-based score [Karchin et al., 2005] of an alanine substitution, several properties of amino-acid residue sequence in a 15-residue window around the position where the indel begins, the density of single nucleotide polymorphisms in the coding exon where the indel begins, and the number of results returned from a PubMed search of the HUGO-approved gene name harboring the indel. A feature matrix that compares features used in VEST-indel to those used by SIFT-indel, DDIG-in, PROVEAN, and CADD is provided (Supp. Table S4).

Among all VEST-indel features, the PubMed feature was consistently the most informative, for both frameshift and in-frame variants. For this feature, the algorithm searches the title and abstract of every publication indexed by PubMed, and returns the number of publications mentioning the HUGO gene name (no aliases) in which the mutation occurs [Schuler et al., 1996; Wheeler et al., 2007]. Thus, genes that have been subject to relatively greater attention in the biomedical literature will be scored as more relevant to human health. This result is consistent with previous reports that literature mining is a useful proxy for the importance of a given gene to human health [Schuler et al., 1996; Karchin et al., 2005; Wheeler et al., 2007; Masica and Karchin, 2011; Capriotti et al., 2012; Clarke et al., 2012]. Insertion/deletion variants in less important genes may knock-down or knock-out protein activity but may not necessarily cause disease [Ng et al., 2008].

Not all variants in genes important for human health are necessarily pathogenic, and over reliance on gene-level disease relevance might ultimately decrease performance. Therefore, we compared the five tested methods on a difficult set of variants, limited to genes that contained at least one pathogenic and one benign variant (141 pathogenic and 78 benign in-frame variants in 57 genes, and 561 pathogenic and 88 benign frameshift variants in 86 genes). Using the scores from the cross-validation experiments (Table 3), we recomputed sensitivity and specificity for VEST-indel for these variants only, and also scored them using the DDIG, SIFT, PROVEAN, and CADD Webservers (Supp. Table S5). VEST-indel had the highest balanced accuracy of the five methods on this difficult set (0.78 in-frame, 0.67 frameshift).

Although VEST-indel, SIFT-indel, DDIG-in, PROVEAN, and CADD share some similarities with respect to training sets and features, we considered that they might be different enough to provide independent information about an indel of interest. Therefore, they could be combined into a meta-predictor to yield improved performance. This approach has had some success in predicting the pathogenicity of missense variants [Gonzalez-Perez and Lopez-Bigas, 2011; Frousios et al., 2013; Martelotto et al., 2014]. Using the multi-method benchmark set, we assessed the classification performance resulting from each pair, trio, or quartet of methods combined using Boolean conjunctions and disjunctions. See Supp. Tables S6 and S7 for a complete list of the tested combinations.

For in-frame classification, the combination of ((VEST-indel AND PROVEAN) OR (CADD AND DDIG-in)) yielded a substantially improved sensitivity (0.93) while retaining good specificity (0.97), when compared to VEST-indel alone (sensitivity = 0.81, specificity = 0.96), and indeed any of the methods alone (Supp. Table S6). This result indicates that these methods are highly complementary when combined in the described fashion. Conversely, for frameshift classification, the combination of ((VEST-indel AND (SIFT-indel OR DDIG-in)) had roughly equivalent sensitivity (0.83) and specificity (0.97) to VEST-indel alone (sensitivity = 0.85, specificity = 0.95). This results because the most specific method (VEST-indel) is combined using the AND operation (i.e., sensitivity could not possibly increase, nor could specificity decrease).

The strategy of classifying a variant as pathogenic if any of the classifiers predicted it to be pathogenic (i.e., combining classifiers with a Boolean OR) did not yield good results. For the in-frame classifier, the combination (VEST-indel OR SIFT-indel OR PROVEAN OR CADD) had a sensitivity of 1 but a specificity of 0.56, with balanced accuracy of 0.78. Combining four classifiers or three classifiers with the OR operator consistently yielded good sensitivity but a substantial decrease in specificity. This result is, to some extent, expected because combining classifiers with the OR operation increases the possibility of accepting a variant as pathogenic. Conversely, requiring that all classifiers agree (i.e., combining classifiers with a Boolean AND) reduces the probability of a pathogenic classification. Indeed, all meta-predictors that used only AND operators had high specificity, but low sensitivity. For example, the (VEST-indel AND SIFT-indel AND CADD AND DDIG-in) meta-predictor had a specificity of 1.00 and sensitivity of 0.46. Taken together, these results highlight the benefit of developing meta-predictors that combine Boolean conjunctions and disjunctions, rather than considering only a single type of Boolean operation.

The benign examples in the multi-method benchmark set were taken from the 1000 Genomes Phase 3 samples, limited to individuals in the AFR (African) super-population [Consortium, 2010] and having MAF ≥ 0.1. Whereas common variants are generally considered to be non-pathogenic [Tennessen et al., 2010], datasets of common variants may be contaminated by pathogenic variants if they occur in genes that are not under purifying selection [Hu and Ng, 2012]. We assessed the possibility that the multi-method benchmark set might include common pathogenic variants. If this were the case, a false positive call from one of the methods might represent the correct identification of a truly pathogenic variant. Genes not subject to purifying selection might alternatively be under positive, balancing, or relaxed (neutral) selection [Hu and Ng, 2012]. For each of VEST-indel, SIFT-indel, DDIG-in, PROVEAN, and CADD we assessed the relationship between variants under selective pressure and those called as false positives, using Fisher's exact test (two-tailed, $\alpha = 0.05$). None of the benign variants were under balancing selection, defined as a statistically significant (nominal $P < 0.05$) positive Tajima's D statistic [Tajima, 1989]. Thirteen frameshift and 56 in-frame variants were under positive selection, defined as a statistically significant (nominal $P < 0.05$) negative Tajima's D statistic. With the exception of a borderline P value for DDIG-in in-frame variant classification ($P = 0.051$), there were no statistically significant relationships between positively selected variants and variants that were called as false positives. For frameshift variants, $P = 1.0$ for VEST-indel, $P = 0.26$ for SIFT-indel, $P = 0.18$ for DDIG-in, and $P = 0.40$ for CADD. For in-frame variants, $P = 0.25$ for VEST-indel, $P = 0.56$ for SIFT-indel, $P = 0.13$ for PROVEAN, and $P = 0.79$ for CADD.

VEST-indel P values for in-frame and frameshift indels are comparable to VEST P values for missense variants and as a result, multiple variant types can be jointly prioritized. We assessed joint prioritization performance by combining variants from the VEST-indel in-frame and frameshift training sets (Table 1) and variants from the VEST missense training set [Carter et al., 2013] (2,475 pathogenic and 1,877 benign in-frame indels; 24,478 pathogenic and 1,350 benign frameshift indels; 38,221 pathogenic and 38,221 benign missense variants). We also assessed performance in a balanced set, in which we randomly selected 1,350 pathogenic and 1,350 benign variants of each type for the combined set. VEST P values and scaled CADD scores were used to compute ROC area under the curve (AUC) as described in *Materials and Methods.* For the combined set, VEST and CADD achieved a similar ROC area under the curve (AUC) of 0.90 and 0.88, respectively. For the balanced set, VEST classification resulted in an AUC of 0.91 and CADD classification resulted in an AUC of 0.74.

Discussion

In this study, we sought to develop a method for predicting indel pathogenicity. This functionality is distinct from existing classifiers that were developed to predict indel impact on protein structure or function. Although clinical utility appears to be a common goal for much of bioinformatics methods development, indel pathogenicity prediction presents the challenge of distinguishing variants that affect protein structure and function from those that adversely affect health [Consortium, 2010]. Given the enrichment for protein sequence and annotation features available for algorithmic development [Boutet et al., 2007; Pollard et al., 2010], the difficulty discriminating neutral and disease-causing indels might be unsurprising. Our newly developed classifier, VEST-indel, partially addresses previous methodological limitations, and achieves high balanced accuracy even when tasked with sorting disease-associated indels from those present in the general population. In particular, VEST-indel realized substantial gains in specificity relative to existing methods, highlighting reductions in falsely classifying neutral variants as pathogenic. To realize these performance gains, VEST-indel heavily utilized a new feature that captures the known relevance of a gene to human health. This new "PubMed" feature leverages decades of community-wide biomedical research. Thus, the algorithm uses features that ultimately estimate indel impact on protein, and the PubMed feature additionally estimates the biological context of the protein. Given that poor specificity also limits the utility of methods aimed at assessing the pathogenicity of missense variants [Chan et al., 2007; Hicks et al., 2011; Thusberg et al., 2011; Shihab et al., 2013], the approach presented here might prove beneficial for variant classification in general.

The in-frame meta-predictor ((VEST-indel OR DDIG-in) AND (PROVEAN)) achieved excellent sensitivity (0.93) and specificity (0.94) when applied to our multi-method benchmark dataset. This complementarity results because the two high-specificity classifiers are combined using the OR operation, which is then combined with the high-sensitivity classifier PROVEAN, using the AND operation (see individual classifier performance, Table 4). The Boolean OR operation increases the possibility of pathogenic classification; importantly, pathogenic classification from VEST-indel and DDIG-in is complementary rather than entirely overlapping, hence the increased sensitivity relative to either method alone. As expected, however, the specificity of the (VEST-indel OR DDIG-in) classifier decreased (see Table S6). Next, even though the highly sensitive PROVEAN is slightly more prone to false positives, the specificity of the meta-predictor cannot decrease owing to the unanimity required by the AND operation; on the contrary, the complementarity of true-negative calls among these three classifiers restores a high specificity. We are deliberate in this explanation because meta-predictor derivation relying on a single Boolean operation type is limiting and can result in significant trade-offs in sensitivity and specificity. As our results show, taking advantage of the complementarity that can result from combining Boolean conjunctions and disjunctions can be beneficial when maximizing balanced accuracy is desired.

Our new VEST-indel method can be used in combination with VEST scoring of missense variants to yield a jointly prioritized list of both variant types. This analysis requires a single batch submission to the CRAVAT server [Douville et al., 2013]. To our knowledge, the only other automated method available for such joint ranking is CADD. For data sets in which the number of missense variants far exceeds the number of indels, VEST and CADD have similar performance. However, when variant types (indels and missense) and classes (pathogenic vs. benign) are evenly distributed, VEST significantly outperforms CADD.

Author Contributions

C.D., D.L.M., and R.K. conceived of the study and designed the experiments. C.D. performed the experiments and wrote the software prototype. D.M.G., R. Kim, and M.R. wrote the production version of the software. P.D.S. and D.N.C. obtained and compiled the HGMD data and provided critical feedback about the manuscript. C.D., D.L.M., and R.K. wrote the manuscript.

Acknowledgment

D.N.C. and P.D.S. acknowledge the financial support of Qiagen Inc. through a License Agreement with Cardiff University.

Disclosure statement: The authors declare no conflict of interest.

References

Altschul SF, Madden TL, Schaeffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402.

Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A. 2007. *Uniprotkb/Swiss-Prot*. Plant Bioinformatics: Springer. p 89–112.

Breiman L. 2001. Random forests. Mach Learn 45:5–32.

Capriotti E, Altman RB. 2011. A new disease-specific machine learning approach for the prediction of cancer-causing missense variants. Genomics 98:310.

Capriotti E, Nehrt NL, Kann MG, Bromberg Y. 2012. Bioinformatics for personal genome interpretation. Brief Bioinform 13:495–512.

Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. 2013. Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics 14(Suppl 3):S3.

Chan PA, Duraisamy S, Miller PJ, Newell JA, McBride C, Bond JP, Raevaara T, Ollila S, Nyström M, Grimm AJ, Christodoulou J, Oetting WS, Greenblatt MS. 2007. Interpreting missense variants: comparing computational methods in human disease genes CDKN2A, MLH1, MSH2, MECP2, and tyrosinase (TYR). Hum Mutat 28:683–693.

Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. 2012. Predicting the functional effect of amino acid substitutions and indels. PloS ONE 7:e46688.

Clarke L, Zheng-Bradley X, Smith R, Kulesha E, Xiao C, Toneva I, Vaughan B, Preuss D, Leinonen R, Shumway M. 2012. The 1000 Genomes Project: data management and community access. Nat Methods 9:459–462.

Consortium GP. 2010. A map of human genome variation from population-scale sequencing. Nature 467:1061–1073.

Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S. 2015. Ensembl 2015. Nucleic Acids Res 43:D662–D669.

Douville C, Carter H, Kim R, Niknafs N, Diekhans M, Stenson PD, Cooper DN, Ryan M, Karchin R. 2013. CRAVAT: cancer-related analysis of variants toolkit. Bioinformatics 29:647–648.

Fawcett T. 2004. ROC graphs: notes and practical considerations for researchers. Mach Learn 31:1–38.

Folkman L, Yang Y, Li Z, Stantic B, Sattar A, Mort M, Cooper DN, Liu Y, Zhou Y. 2015. DDIG-in: detecting disease-causing genetic variations due to frameshifting indels and nonsense mutations employing sequence and structural properties at nucleotide and protein levels. Bioinformatics 31:1599–1606.

Frousios K, Iliopoulos CS, Schlitt T, Simpson MA. 2013. Predicting the functional consequences of non-synonymous DNA sequence variants—Evaluation of bioinformatics tools and development of a consensus strategy. Genomics 102:223–228.

Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Rieder MJ, Altshuler D, Shendure J. 2013. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493:216–220.

Gonzalez-Perez A, Lopez-Bigas N. 2011. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genetics 88:440–449.

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. 2009. The WEKA data mining software: an update. ACM SIGKDD Explor Newslett 11:10–18.

Hicks S, Wheeler DA, Plon SE, Kimmel M. 2011. Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. Hum Mutat 32:661–668.

Hu J, Ng PC. 2012. Predicting the effects of frameshifting indels. Genome Biol 13:R9.

Hu J, Ng PC. 2013. SIFT Indel: predictions for the functional effects of amino acid insertions/deletions in proteins. PloS ONE 8:e77940.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337–338.

Karchin R, Kelly L, Sali A. Improving functional annotation of non-synonomous SNPs with information theory; 2005.

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.

Knijnenburg TA, Wessels LF, Reinders MJ, Shmulevich I. 2009. Fewer permutations, more accurate P-values. Bioinformatics 25:i161–i168.

Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. 2013. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 42(Database Issue):D980–985.

Lohmueller KE, Indap AR, Schmidt S, Boyko AR, Hernandez RD, Hubisz MJ, Sninsky JJ, White TJ, Sunyaev SR, Nielsen R. 2008. Proportionally more deleterious genetic variation in European than in African populations. Nature 451:994–997.

MacArthur DG, Tyler-Smith C. 2010. Loss-of-function variants in the genomes of healthy humans. Hum Mol Genet 19(R2):R125–R130.

Martelotto LG, Ng CK, De Filippo MR, Zhang Y, Piscuoglio S, Lim R, Shen R, Norton L, Reis-Filho JS, Weigelt B. 2014. Benchmarking mutation effect prediction algorithms using functionally validated cancer-related missense mutations. Genome Biol 15:484.

Masica DL, Karchin R. 2011. Correlation of somatic mutation and expression identifies genes important in human glioblastoma progression and survival. Cancer Res 71:4550–4561.

Ng PC, Levy S, Huang J, Stockwell TB, Walenz BP, Li K, Axelrod N, Busam DA, Strausberg RL, Venter JC. 2008. Genetic variation in an individual human exome. PLoS Genet 4:e1000160.

Pfeifer B, Wittelsburger U, Onsins SER, Lercher MJ. 2014. PopGenome: an efficient Swiss army knife for population genomic analyses in R. Mol Biol Evol 31:1929–1936.

Pickands III J. 1975. Statistical inference using extreme order statistics. Ann Stat 3:119–131.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 20:110–121.

Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35(Suppl 1):D61–D65.

Schuler GD, Epstein JA, Ohkawa H, Kans JA. 1996. Entrez: molecular biology database and retrieval system. Methods Enzymol 266:141.

Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GL, Edwards KJ, Day IN, Gaunt TR. 2013. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat 34:57–65.

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

Steward RE, MacArthur MW, Laskowski RA, Thornton JM. 2003. Molecular basis of inherited diseases: a structural perspective. Trends Genet 19:505–513.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595.

Tennessen JA, Madeoy J, Akey JM. 2010. Signatures of positive selection apparent in a small sample of human exomes. Genome Res 20:1327–1334.

Thusberg J, Olatubosun A, Vihinen M. 2011. Performance of mutation pathogenicity prediction methods on missense variants. Hum Mutat 32:358–368.

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S. 2007. Database resources of the national center for biotechnology information. Nucleic Acids Res 35(Suppl 1):D5–D12.

Wong WC, Kim D, Carter H, Diekhans M, Ryan MC, Karchin R. 2011. CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. Bioinformatics 27:2147–2148.

Zhao H, Yang Y, Lin H, Zhang X, Mort M, Cooper DN, Liu Y, Zhou Y. 2013. DDIG-in: discriminating between disease-associated and neutral non-frameshifting micro-indels. Genome Biol 14:R23.