PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

How good are pathogenicity predictors in detecting benign variants?

Abhishek Niroula, Mauno Vihinen*

Protein Structure and Bioinformatics, Department of Experimental Medical Science, Lund University, Lund, Sweden

* mauno.vihinen@med.lu.se

Abstract

Computational tools are widely used for interpreting variants detected in sequencing projects. The choice of these tools is critical for reliable variant impact interpretation for precision medicine and should be based on systematic performance assessment. The performance of the methods varies widely in different performance assessments, for example due to the contents and sizes of test datasets. To address this issue, we obtained 63,160 common amino acid substitutions (allele frequency ≥1% and <25%) from the Exome Aggregation Consortium (ExAC) database, which contains variants from 60,706 genomes or exomes. We evaluated the specificity, the capability to detect benign variants, for 10 variant interpretation tools. In addition to overall specificity of the tools, we tested their performance for variants in six geographical populations. PON-P2 had the best performance (95.5%) followed by FATHMM (86.4%) and VEST (83.5%). While these tools had excellent performance, the poorest method predicted more than one third of the benign variants to be disease-causing. The results allow choosing reliable methods for benign variant interpretation, for both research and clinical purposes, as well as provide a benchmark for method developers.

Author summary

In precision/personalized medicine of many conditions it is essential to investigate individual's genome. Interpretation of the observed variation (mutation) sets is feasible only with computational approaches. We assessed the performance of variant pathogenicity/tolerance prediction programs on benign variants. Variants were obtained from high-quality ExAC database and selected to have minor allele frequency between 1 and 25%. We obtained 63,160 such cases and investigated 10 widely used predictors. Specificities of the methods showed large differences, from 64 to 96%, thus users of these methods have to be careful when choosing the one(s) they will use. We investigated further the performances on different populations, allele frequencies, separately for males and females, chromosome wise and for population unique and non-unique variants. The ranking of the tools remained the same in all these scenarios, i.e. the best methods were the best irrespective on how the data was filtered and grouped. This is to our knowledge the first large scale evaluation of method performance on benign variants.

Citation: Niroula A, Vihinen M (2019) How good are pathogenicity predictors in detecting benign variants? PLoS Comput Biol 15(2): e1006481. https://doi.org/10.1371/journal.pcbi.1006481

Editor: Anna R R Panchenko, National Institutes of Health, UNITED STATES

Received: August 31, 2018

Accepted: December 19, 2018

Published: February 11, 2019

Data Availability Statement: All variant files are available from the VariBench database http://structure.bmc.lu.se/VariBench/exac_aas.php.

Funding: MV acknowledges financial support from Swedish Research Council (Vetenskapsrådet) VR 2015-02510. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Next Generation Sequencing (NGS) is widely used in clinical diagnosis as well as in population genetics to investigate patterns of genetic variants in healthy individuals. The large numbers of variants, millions per genome in comparison to reference sequences, pose challenges for detecting disease-causing variants. There are on average about 10,000 variants per genome that cause amino acid substitutions [1]. Several databases enable annotation of disease relevance of variants and frequencies among healthy individuals. These include numerous locus specific variation databases (LSDBs) that are curated by experts in the genes and diseases. While LSDBs typically concentrate on individual genes and proteins or diseases, the general databases have much wider scope such as ClinVar [2], Online Mendelian Inheritance in Man (OMIM) [3] and the UniProt Knowledgebase (UniProtKB) [4].

The most harmful variants confer adverse impacts and reduce the fitness of the carrier, and are therefore selected against and removed from the population. On the other hand, the benign variants are tolerated and are inherited through the generations. Therefore, variants occurring at high frequencies in a population are likely benign. Information for variants and their frequencies in various populations are available e.g. in the database of short genetic variations (dbSNP) [5], the 1000 Genomes Project [6], the Exome Sequencing Project (ESP) Exome Variant Server (EVS) [7], and recently in the Exome Aggregation Consortium (ExAC) database [8]. These resources are widely used to filter out likely benign variants as well as for training and testing computational tools. Variants with allele frequencies (AFs) $\geq$1% are generally assumed to be benign, assumption widely used by e.g. predictor developers [9–12]. There are some exceptions e.g. in late onset diseases or due to incomplete penetrance. We are not aware of reliable estimates of such cases. Sickle cell anemia-causing E6V substitution in β-globin is probably the best known example. The number of such cases is so low that it does not affect results based on large scale studies, as in here. Most variants in these databases are rare, for example in the ExAC database, 99% of the variants have AF below 1% [8], and have unknown clinical relevance.

Prediction tools are instrumental for variant effect interpretation in personalized and precision medicine since experimental methods cannot deal with the amounts of variation data generated in sequencing projects. The American College of Medical Genetics and Genomics (ACMG) and the European Society of Human Genetics (ESHG) guidelines recommend using computational predictions as one of several lines of evidence for variant interpretation [13, 14]. Similarly, the joint consensus recommendation for the interpretation of variants in cancer by the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists include the use of computational predictions [15].

Numerous computational tools based on different principles have been developed to predict the tolerance and pathogenicity of genetic variants [16–19]. The performance of these tools varies widely [16, 20–23]. Even a minor difference in the performance leads to misinterpretation of large numbers of variants in genome or exome-wide scale. Hence, the choice of the tools is critical for reliable variant interpretation. The assessment of method performance requires benchmark datasets with known outcomes. In this field, such datasets are available at VariBench [24] and VariSNP [25]. Further, the assessment has to be made in a systematic way and reporting the full performance of the analyzed methods [26, 27], which unfortunately is often not the case, especially for commercial products [28]. In addition to pathogenicity/tolerance method assessment, the performance of some other predictor classes have been assessed including alternative splicing [29, 30], protein stability [31, 32], protein solubility [33], and protein localization [34].

A comprehensive predictor assessment requires a benchmark with both positive (showing the effect) and negative (not having an effect) variants. Here, we tested the predictor specificity

i.e. the capability to recognize variants not having phenotypic effect using the largest available dataset of likely benign variants. Recently, the ExAC database that has been carefully curated and contains quality-controlled data for altogether 60,706 exomes was released [8]. The database contains the overall frequencies of variations across all the individuals as well as the frequencies for several populations. We obtained the common variants from the ExAC database and identified those leading to amino acid substitutions (AASs). In total, 63,160 AASs had AF $\geq$1% and <25% in at least one of the cohorts in the dataset. These AASs are widely considered as benign and therefore were used to assess the performance of the prediction tools. We investigated the performance of 10 widely used prediction methods and found that the best tools are excellent while some others have poor performance.

Materials and methods

Variation data

The variation data were obtained from the ExAC database (release 0.3.1) [8] in a Variant Call Format (VCF) file. We identified the variants leading to amino acid substitutions (AASs) by using the annotations from the Variant Effect Predictor (VEP) [35] included in the downloaded VCF file. The amino acid substitutions were further filtered by using the AFs in the whole dataset as well as in different populations. The VCF file contained AFs for various datasets and populations. The adjusted AF (AF for all individuals with genotype quality (GQ) $\geq$20 and depth (DP) $\geq$10) as well as the AFs in all geographical populations (African, American, East Asian, Finnish, non-Finnish European, South Asian, and Other) were used in the analysis. In addition, we defined the AFs for variants in males and females. Variants having AFs $\geq$1% and <25% in any of the 9 populations were included to the study. We set an upper threshold of AF to 25%, so that the AFs represented the minor alleles. If the four nucleotides have a random distribution in a position, a minor allele cannot have a frequency >25% without becoming the major allele. In total, there are 63,197 variants that meet these criteria. The dataset is available at VariBench (http://structure.bmc.lu.se/VariBench/exac_aas.php).

Computational predictions

The predictions were obtained from the dbNSFP database (version 3.2a) [36] for several tools. The database contains annotations and predictions for all potential single nucleotide substitution-caused AASs. We obtained the predictions for Combined Annotation Dependent Depletion (CADD) [37], Functional Analysis through Hidden Markov Models (FATHMM) [38], Likelihood Ratio Test (LRT) [39], MutationAssessor [40], MetaLR [9], MetaSVM [9], MutationTaster2 [41], Polymorphism Phenotyping v2 (PolyPhen-2) [42], Protein Variation Effect Analyzer (PROVEAN) [43], Sorting Intolerant From Tolerant (SIFT) [44], and Variant Effect Scoring Tool (VEST) [45]. If there were multiple predictions for a variant from the same tool, we took the most frequent classification. If two classes were equally frequent, then the classification was considered as ambiguous. In addition, we obtained predictions for PON-P2 [22] by using the tool's Application Programming Interface (API).

Training datasets

Training datasets were obtained for FATHMM, MetaLR, MetaSVM, PolyPhen-2, VEST, and PON-P2 and cases in them were excluded from assessment of those tools. Since no variations were left for Meta-LR and Meta-SVM after excluding the training data, we could not evaluate these methods.

Common variants

Variants with AF ≥1% and <25% in a specific population are considered as common for that population. This criterion was used to obtain 10 subsets of variation data (Adj, AFR, AMR, EAS, FIN, NFE, SAS, OTH, MALE, and FEMALE). For the six geographical populations: African/African American (AFR), Latino (AMR), East Asian (EAS), Finnish (FIN), Non-Finnish European (NFE), and South Asian (SAS), the datasets were further partitioned into population-specific unique and non-unique datasets. The unique dataset contains variants with AF ≥1% and <25% in the specific population but <1% in all other populations and the non-unique dataset consists of the remaining variants. For example, the variants with AF ≥1% and <25% in AFR population are indicated as common variants for AFR population. From those, the variants with AF <1% in all the five other geographical populations are unique variants for the AFR population. The remaining common variants in the AFR population are non-unique variants.

To exclude misclassified pathogenic variants in the dataset filtered with the AF threshold, we obtained from ClinVar all the 24,232 variants that lead to AASs and were annotated as pathogenic or likely pathogenic (13 July 2018) [2]. There were 37 variants which had AF ≥1% and <25%, some of which had been used for predictor training: FATHMM (14 variants), PON-P2 (14), PolyPhen-2 (4), and VEST (6). The reason at least for some of these variants to be included into the training datasets is that more data may have accumulated to reclassify variants after the methods were trained.

Performance comparison

Except for CADD and VEST, the investigated methods classify the variants into harmful and benign. We used these classifications for the method performance assessment. For CADD, we classified the variants based on the phred-like score with a cutoff 20, below which the variants were classified as benign and otherwise harmful, as suggested by the authors. For VEST, we classified the variants based on the VEST score with a cut-off 0.5, below which the variants were classified as benign and otherwise harmful. The terms *deleterious*, *damaging*, *probably damaging*, *possibly damaging*, *disease-causing*, *functional*, and *pathogenic* were all considered to be harmful and the terms *tolerated*, *benign*, *neutral*, *non-functional*, and *polymorphism* were all considered to be benign. MutationTaster2 provides automatic annotations for harmful and benign variants based on annotations in variation databases and predicts the impacts for others. In this study, the automatic annotations of MutationTaster2 were excluded to test the actual prediction capability of the tool. PON-P2 and LRT classify variants into three classes, the third class being variants of unknown significance. The variants classified as unknown were excluded.

Several measures are needed to describe the overall performance of prediction methods [26, 27]. Since we investigated only one type of variants, the benign ones, it was possible to calculate only a single measure, the specificity. Specificity is the proportion of correctly predicted benign variants,

$$Specificity = \frac{Number\ of\ predicted\ benign\ variants}{Total\ number\ of\ predicted\ variant\ effects\ (harmful\ or\ benign)}.$$

The scores can be multiplied by 100 to show results in percentages.

Results

Specificity of tolerance predictors

To assess the quality of variant pathogenicity/tolerance prediction methods we collected from the ExAC database all variants that had AF $\geq$1% and <25%. Because of their high frequency, these variants are usually considered to be neutral and were used in here to assess the specificity of prediction methods. We tested whether 10 widely used methods having different background and design principles showed differences in their performance for benign variants. The predictions for 9 tools were collected from the dbNSFP database [36]. For PON-P2 [22], we run the predictions using the Application Programming Interface.

We could not evaluate four tools MetaLR [9], MetaSVM [9], M-CAP [46], and REVEL [11]. MetaLR and MetaSVM are meta predictors, after excluding the training datasets of their constituent tools no variants were left for evaluation. REVEL has been trained with several datasets including Exome Sequencing Project and The 1000 Genomes project that form a substantial part of the ExAC dataset that we used for testing. Thus, analysis of the performance with ExAC data would introduce circularity and not indicate true performance, instead denote how well the methods have learned the training data. M-CAP is aimed for rare variants, therefore predictions for common variants were not available and the method performance could not be assessed.

The tools are based on different principles and include those based on evolutionary information only, LRT [39], PROVEAN [43], and SIFT [44], and those combining different types of features, CADD [37], FATHMM [38], MutationAssessor [40], MutationTaster2 [47], PolyPhen2 [42], PON-P2 [48], and VEST [45]. Most of the investigated tools have been trained with known disease-causing and benign variants. The methods that use only sequence conservation information have not been trained. If variants used for training are used for assessing the methods, the obtained performance measures are likely inflated [20, 26, 49]. Hence, we excluded the training datasets for FATHMM, PON-P2, PolyPhen-2, and VEST. The remaining tools were either not trained or the training datasets were not available.

All the tested tools classify variants into pathogenic and benign classes except for CADD and VEST. CADD predicts a continuous phred-like score that ranges from 1 to 99, higher values indicating more deleterious cases. The score for VEST indicates benign when 0 and pathogenic when 1. For CADD we used the highest phred-like score cutoff recommended by the authors i.e. 20. For VEST, we classified the variants into two classes using VEST score cutoff of 0.5. To evaluate usability of the CADD and VEST cutoffs, we analyzed the sensitivities and specificities of the tools at different cutoffs which showed that the optimal VEST score cutoff is between 0.45 and 0.5 and phred-like score cutoff is between 20 and 25 (S1A and S1B Fig).

The performances of some of these tools have been assessed previously several times, however not with this kind of high-quality and large dataset for benign variants. It is important both in research and clinical practice to be able to sort out variants that have no relevance for the condition under investigation. The specificities of the methods range from 0.63 for SIFT and 0.64 for MutationTaster2 to 0.96 for PON-P2 (Table 1). FATHMM and VEST have the second and third highest performance i.e. 0.86 and 0.84, respectively. It should be noted that variants are classified into three classes by PON-P2 and two classes by FATHMM, and VEST, and CADD does not group variants into pathogenic and benign categories, instead predicts continuous probabilities. For VEST, we classified the variants into two classes using a cutoff of 0.5. The methods that use evolutionary data only are towards the end of the list (Table 1). Their specificities are 0.724 for LRT, 0.774 for PROVEAN and 0.634 for SIFT. Machine learning methods populate both ends of specificity spectrum. PON-P2, FATHMM and VEST have the highest scores while the specificities for MutationTaster2 and CADD are 0.640 and 0.643,

Table 1. Specificities of variant interpretation tools.

	All variants (n = 63,197)[a]				Variants predicted by all tools (n = 7,268)[b]		
Tools	**VUS[c]**	**Benign**	**Harmful**	**Specificity**	**Benign**	**Harmful**	**Specificity**
PON-P2[d]	21,373	34,529	1,626	0.955	6655	613	0.916
VEST[d,e]	1,168	22,614	4,480	0.835	5984	1284	0.823
FATHMM[d]	5,531	43,005	6,766	0.864	6287	981	0.865
PROVEAN	3,908	45,868	13,421	0.774	5712	1556	0.786
PPH2[d,f]	6,386	37,124	13,602	0.732	5404	1864	0.744
LRT	19,333	31,736	12,128	0.724	5465	1803	0.752
MA	8,044	39,493	15,660	0.716	5306	1962	0.730
CADD[g]	0	40,659	22,538	0.643	4539	2729	0.625
SIFT	5,099	36,808	21,290	0.634	4868	2400	0.670
MT2[h]	15,313	30,632	17,252	0.640	4764	2504	0.655

[a]All variants having AF> = 1% and <25% in at least one population and not present in the training dataset for the method. After excluding cases in the training datasets, the total number of variants was 57,528 for PON-P2, 28,262 for VEST, 55,302 for FATHMM, and 57,112 for PPH2.

[b]Variants classified as benign or harmful. Variants present in training dataset of any of the tools were excluded. All variants that were automatically annotated without making predictions were excluded.

[c]Variants for which the predictions were not available, were ambiguous, or were predicted to have unknown significance.

[d]Variants present in the training datasets were excluded.

[e]Variants were not classified into benign and harmful by the program. A cutoff of 0.5 was used so that variants with score greater than or equal to 0.5 were classified as harmful, otherwise benign.

[f]HumVar version of PolyPhen-2 was used as the performance was higher than for HumDiv version.

[g]Variants were not classified into benign and harmful by the program. A cutoff of 20 was used so that variants with score greater than or equal to 20 were grouped as harmful and otherwise benign. The authors have recommended a cutoff ranging from 10 to 20. The highest cutoff was used so that the highest possible specificity was obtained.

[h]Variants that were automatically detected to be harmful or benign were not included in the classified cases as they are not real predictions by the tool, instead annotations based on known data.

https://doi.org/10.1371/journal.pcbi.1006481.t001

respectively. It is not possible to draw definitive conclusions from the ways methods have been implemented, except saying that machine learning methods can reach much higher specificities in the best installations.

To systematically assess the performance of prediction tools, it would be important to include both pathogenic and benign variants. However, since there is no dataset of pathogenic variants that has not been used for training any of the tools, we could not perform a similar analysis for the pathogenic variants. Therefore, we used a small set of pathogenic and likely pathogenic variants from ClinVar to compare sensitivities of the tools side by side with the specificities (S1C Fig). Since we could not filter out variants used for training of all the tools, we did not do this for any of the methods. High sensitivities indicate that the tools with high specificities are not overfitted towards predicting all the variants as benign. Apart from that, we do not recommend to use the sensitivity scores presented here as reliable estimates of performance. S1D Fig shows almost identical results to those in S1C Fig when the ClinVar variants were evaluated together with the variants predicted by all the methods.

PON-P2 had the highest proportion of unclassified variants, however with far better specificity compared to the other tools (Fig 1 and Table 1). The end users have to decide what is most relevant for them—large coverage with additional false positives or lower coverage but highly reliable predictions. One percent difference in specificity means >100 false positives more or less per exome, thus the differences accumulate very fast.

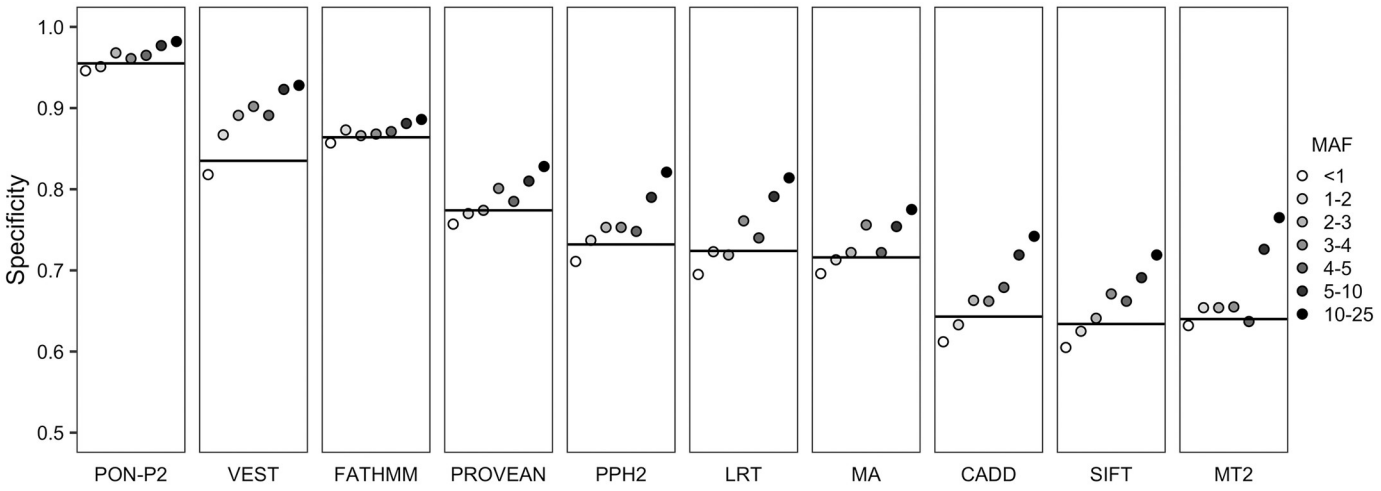

Fig 1. Performance of variant tolerance predictors. Specificities of 10 prediction tools for variants with different AFs. The black horizontal line indicates performance for all variants (AF ≥1% and <25%). The variants with AF <1% have low AF in the whole dataset but have higher AF in at least one of the populations. MA, MutationAssessor; MT2, MutationTaster2; PPH2, PolyPhen-2.

https://doi.org/10.1371/journal.pcbi.1006481.g001

To compare the performance of tools on the same set of variants, we computed the specificities of the tools on variants for which all tools made predictions (Table 1). The scores are somewhat different for all the methods and that is normal for different test datasets. The largest difference is seen for PON-P2, however, it is still the best predictor also on this dataset. The number of variants predicted by all the tools, 7,268, is only 11.5% of the total number of cases.

There are various reasons for differences in coverage, some data items may be missing, some sequences are unique for human and may therefore not be evaluated, etc. All the methods have their limitations. Comparison of both the sets in Table 1 shows that the ranking order of the methods remains practically the same. The major differences are that FATHMM has higher specificity than VEST, and CADD has the lowest specificity of all, for the variants that all the tools can predict. The other analyses are reported for all the variants that each method predicted to cover as many variants as possible.

Next we investigated whether the differences in specificities could be due to certain types of variations or whether they were due to differences in the methods. To assess the performance of tools for variants with different AFs, we divided the dataset into groups based on adjusted AF on the whole dataset. The predictor performance is higher for variants with higher AFs for all the tools (Fig 1). The specificity differences between the AF bins are the smallest for PON-P2 and FATHMM while several other methods, including CADD, LRT, PolyPhen, SIFT and VEST, had very strong correlation between specificity score and allele frequency.

As mentioned above, 1% difference in specificity means a difference of over 100 false classifications in an exome. Since the dataset is so large even a small difference in specificity is statistically significant. Results for Fisher exact tests between the pairs of tools show that the differences are significant for all variants (S2A Fig) as well as for variants predicted by all the tools (S2B Fig). The tools with similar performances have high p-value (low negative logarithm of p-value). CADD, SIFT and MT2 form one group where the results are somewhat similar, PolyPhen2, LRT and MutationAssessor form another group, The rest of the tools have significantly different performances for all variants, VEST, FATHMM and PROVEAN have similar

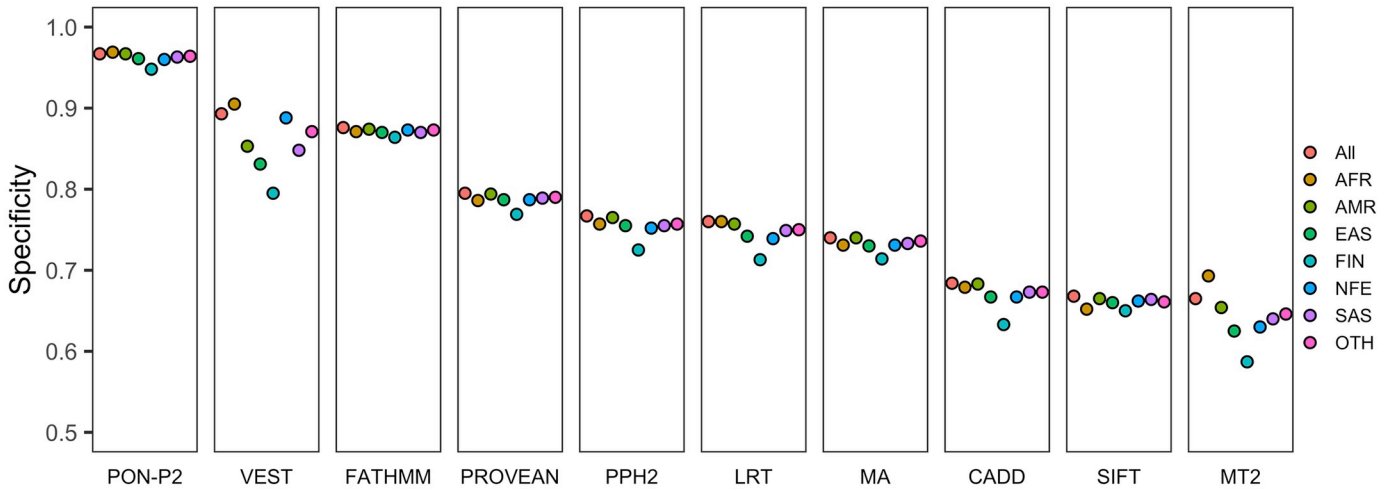

Fig 2. Performance of variant tolerance predictors for variants in ethnic groups. Specificities of prediction tools for common variants (AF ≥1% and <25%) in different populations. AFR, African; AMR, American; EAS, East Asian; FIN, Finnish; NFE, Non-Finnish European; OTH, Other; SAS, South Asian; MA, MutationAssessor; MT2, MutationTaster2; PPH2, PolyPhen-2.

https://doi.org/10.1371/journal.pcbi.1006481.g002

performances. The differences are large as the p value scale ranges from 1 to 10^{-16}. Thus, practically all the differences are statistically highly significant.

Population-specific performance

ExAC database contains information for the genetic ancestry of the individuals. Thus, in addition to the general performance, we were able to analyze also ancestry-based assessment. The same three tools, i.e. PON-P2, FATHMM, and VEST, showed the highest specificities also on the data for the ancestry groups (called for populations hereafter) (Fig 2). The methods, however, show somewhat different performances for different populations. PON-P2 and FATHMM have small performance differences between the populations, 2 and 1%, respectively, while VEST has bigger performance differences, 11% between FIN and AFR. Interestingly, all the tools have the lowest specificity for the Finnish population. This is presumably because the small, and in the past rather closed population passed through a narrow bottleneck some 300 years ago during which certain unique alleles were highly enriched.

We analyzed whether the differences in specificities in the populations were due to the differences in the percentages of variants predicted as unknown (S1 Table). The percentages of the predicted unknown variants for most tools are similar across all populations except for the Finnish population. Most tools, except for PON-P2, have the lowest percentage of variants that could not be predicted for the Finnish population. On the other hand, PON-P2 has slightly higher percentage of unknown variants in Finnish population. The difference in performance between the populations is much bigger than the difference in the percentage of unknown variants.

Next, we identified population-specific common variants which have AF ≥1% and <25% in one population but have AF <1% in all the other populations. These are referred to as population-specific unique variants and the remaining variants for the population are referred to as non-unique variants. The proportions of unique variants vary in the populations, ranging from 6.8% in European population (excluding Finnish) to 62.4% in the

African population (S2 Table). Humans have their origin in Africa and it is well known that the African population has the highest variation as most variants are recent, see e.g. [50]. The tools showed lower specificities for the unique variants than for the non-unique variants in the populations (Fig 3A). The lowest performance is seen for the unique variants in the Finnish population.

Performance differences vary largely depending on the tools and the populations. The performance differences between the unique and non-unique variants are the lowest in the African population (0.6–3.5%) and the highest in the Finnish population (3.2–12.1%) (S3 Table). With respect to the tools, the differences for unique variants are the lowest for FATHMM (ranging from 1.3 to 4.1%) and PON-P2 (ranging from 1.2 to 8.0%) and the largest for MutationTaster2 (18.4%), VEST (16.4%) and LRT (14.9%). The differences for the unique and non-unique variants in each population are visualized in Fig 3A. The differences are the smallest for FATHMM and PROVEAN, up to 3.6 and 6.6%, and the largest for LRT and CADD, up to 18.7 and 12.2%.

As the tools have lower performances for unique variants, we investigated the frequencies of unique variants and those that were not unique (i.e. non-unique). Most unique variants have low AF, between 1% and 5%, while the proportions of non-unique variants with different AFs are similar (Fig 3B). Since many predictors have been trained with variants with high allele frequencies as benign variants, the lower specificities for unique variants could be due to disparity in the frequencies. To control the bias due to frequency, we compared the performance of the tools for unique and non-unique variants with AF in the same range (i.e. 1–5%) in each population. The comparison showed that the tools indeed have poorer performance for unique variants than for non-unique variants (Fig 3C). The differences are the smallest for FATHMM, PON-P2 and PROVEAN, up to 3.7, 6.7 and 6.7%, and the largest for CADD and MutationTaster, up to 12.2% for both (S4 Table, Fig 3C). For Finnish population there are generally the largest differences (3.2 to 12.2%).

Effects of the sex and chromosomal location on prediction performance

Finally, we evaluated the performance for variants from males and females in the populations. No differences were observed in predictor performance. Most of the variants in these two datasets are overlapping. The proportions of unique variants in male (AF $\geq$1% in male but <1% in female) and female (AF $\geq$1% in female but <1% in male) populations are 5.6% and 16.9%, respectively (S5 Table). The number of unique variants in females is 3.4 times higher than the unique variants in males. This may be because of the larger numbers of females than males with African ancestry (1.75 times) in the ExAC dataset. The AFR population has the largest percentage of unique variants compared to the other groups. The performance for unique variants in male is lower than for the common variants and unique variants in female (Fig 4).

To assess the influence of variants in sex chromosomes for the lower performance of tools for unique variants in males, we examined the proportions of variants for females and males in all chromosomes. As there were only 3 variants in Y chromosome we could not investigate performance for variants in this chromosome. In the remaining chromosomes, the ratio of unique variants in males to females range from 0.17 to 0.39, with a median of 0.30. The ratio is 0.28 in the X chromosome, i.e. very close to the median (S5 Table). The tools show only minor differences in the specificities for variants in different chromosomes (Fig 5).

Discussion

Performance comparison of the computational tools enables choosing the most reliable methods. Critical Assessment of Genome Interpretation (CAGI, https://genomeinterpretation.org/)

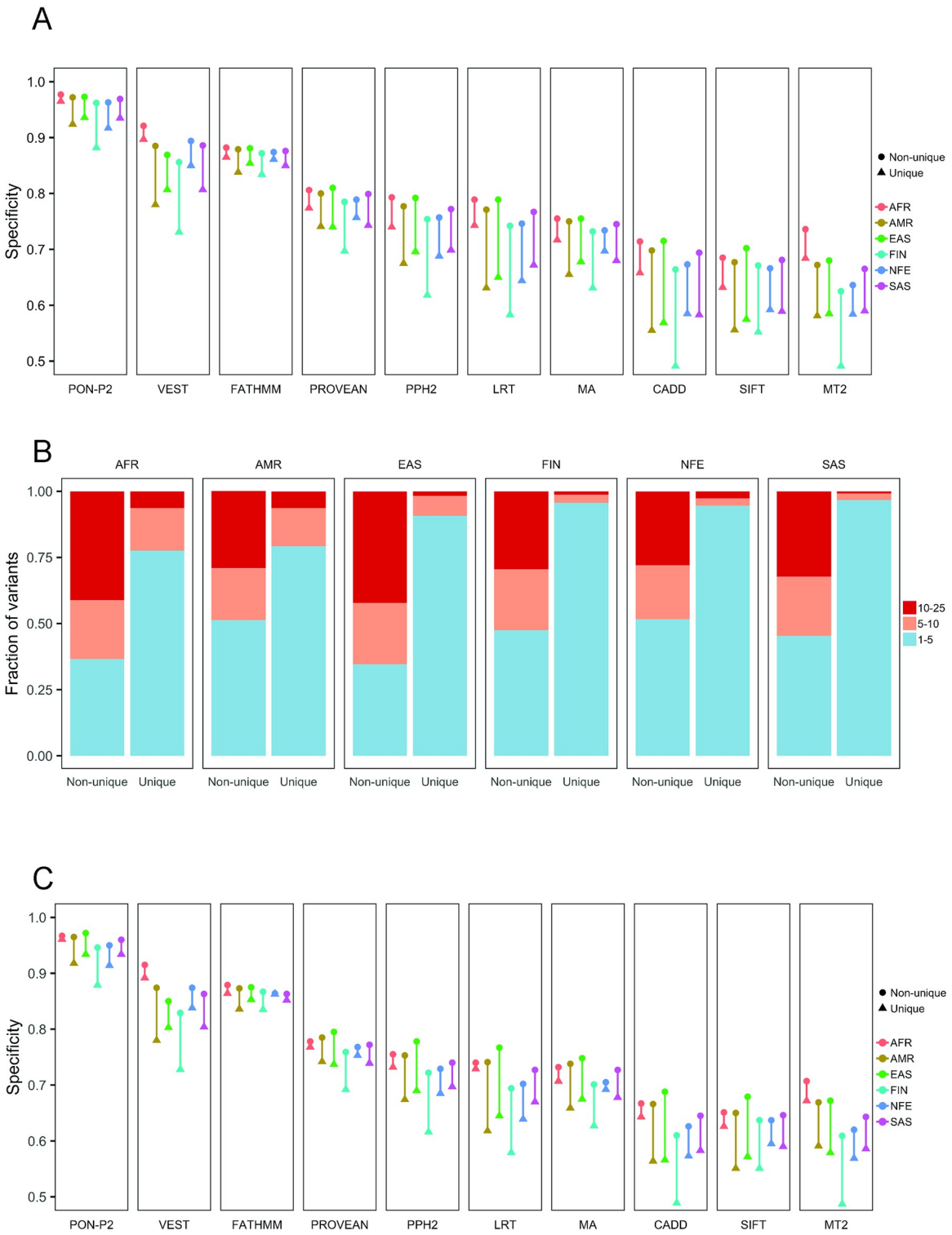

Fig 3. Analysis of unique and non-unique variants in populations. (A) Performance of tools on unique and non-unique variants with different minor allele frequencies in different populations. AFR, African; AMR, American; EAS, East Asian; FIN, Finnish; NFE, Non-Finnish European; SAS, South Asian. The unique dataset contains variants with AF ≥1% and <25% in the specific population but <1% in all other populations and the non-unique dataset consists of the remaining variants. The differences are shown by the lines containing the values for each population. (B) Fractions of unique and non-unique variants in relation to AF. The colors for AF ranges are shown to the right. (C) Specificities of prediction tools on unique and non-unique variants (AF 1–5%) for each ancestry group. Unique variants

have AF ≥1% in specific ancestry group but AF < 1% in all other ancestry groups. Non-unique variants have AF ≥1% in more than one ancestry groups.

https://doi.org/10.1371/journal.pcbi.1006481.g003

is a community wide effort to assess variant interpretation tools and approaches in the form of competitions [51]. In addition, performance of the tools has been tested by the developers as well as independent researchers. Since some predictors are frequently updated and new ones are developed, they should be assessed regularly [17]. Large datasets of both positive and negative classes are required to assess the performance comprehensively. Due to the lack of a large dataset of disease-causing variations that does not overlap with the training datasets used by the method developers, we could not assess the true positive and false negative rates for the tools. Although several performance measures are required to describe the overall performances of prediction methods [26, 27], we could only compare specificities of the tools, i.e. the capabilities of the tools to detect benign variants. We used the common variants from the ExAC database and the variants predicted to be neutral were considered as correct predictions and those predicted to be disease-related as false negatives. The large size of the ExAC database lends strength for the analysis.

Many tools have been trained with disease-causing and likely benign variants. In most cases, the benign variants have been selected based on their allele frequencies in general population(s). The common variants are considered as benign and the tools have been benchmarked against them. In some rare cases disease-related variants can have high frequency at least in some populations (e.g. sickle cell anemia HbS allele). However, such cases are very rare and are not considered to affect statistics when using large number of cases, as in here.

The analysis of burden of the harmful variants revealed that most harmful variants have extremely low AFs [52]. However, benign variants can have equally low AFs as harmful ones. Performance assessments of tools with variants with all AFs for both harmful and benign variants are desirable; however, such dataset does not exist. In this study, we defined variants with AF ≥1% and <25% as benign variants. The upper limit of 25% was set so that the variant allele

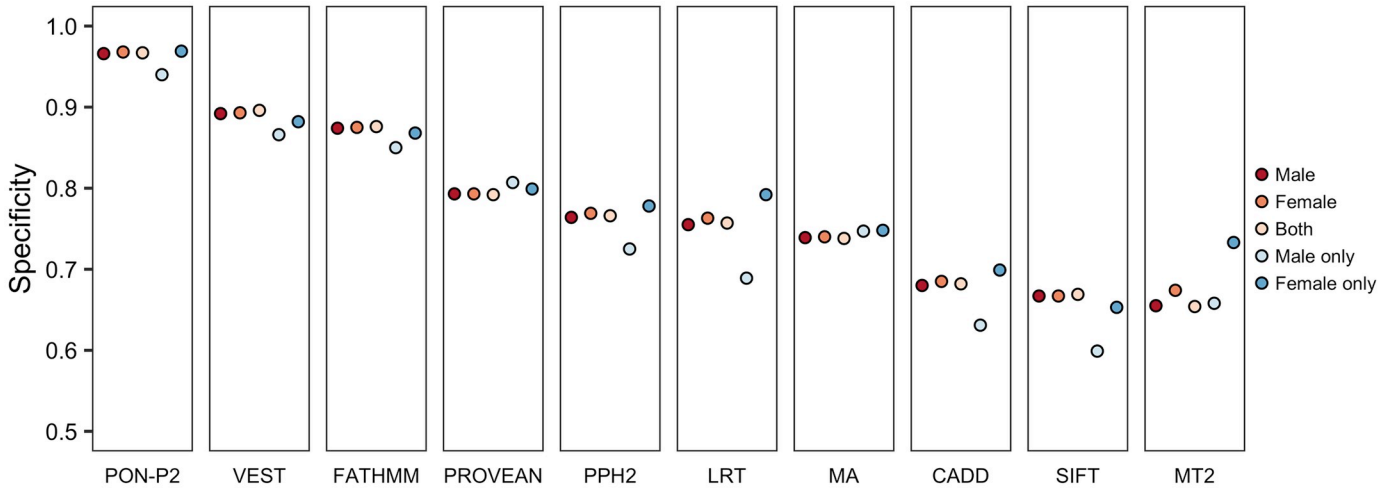

Fig 4. Performance of variant tolerance predictors for variants in males and females. Results are shown for all variants for males and females, both, as well as for unique variants in male (AF ≥1% in male but <1% in female) and female (AF ≥1% in female but <1% in male) populations.

https://doi.org/10.1371/journal.pcbi.1006481.g004

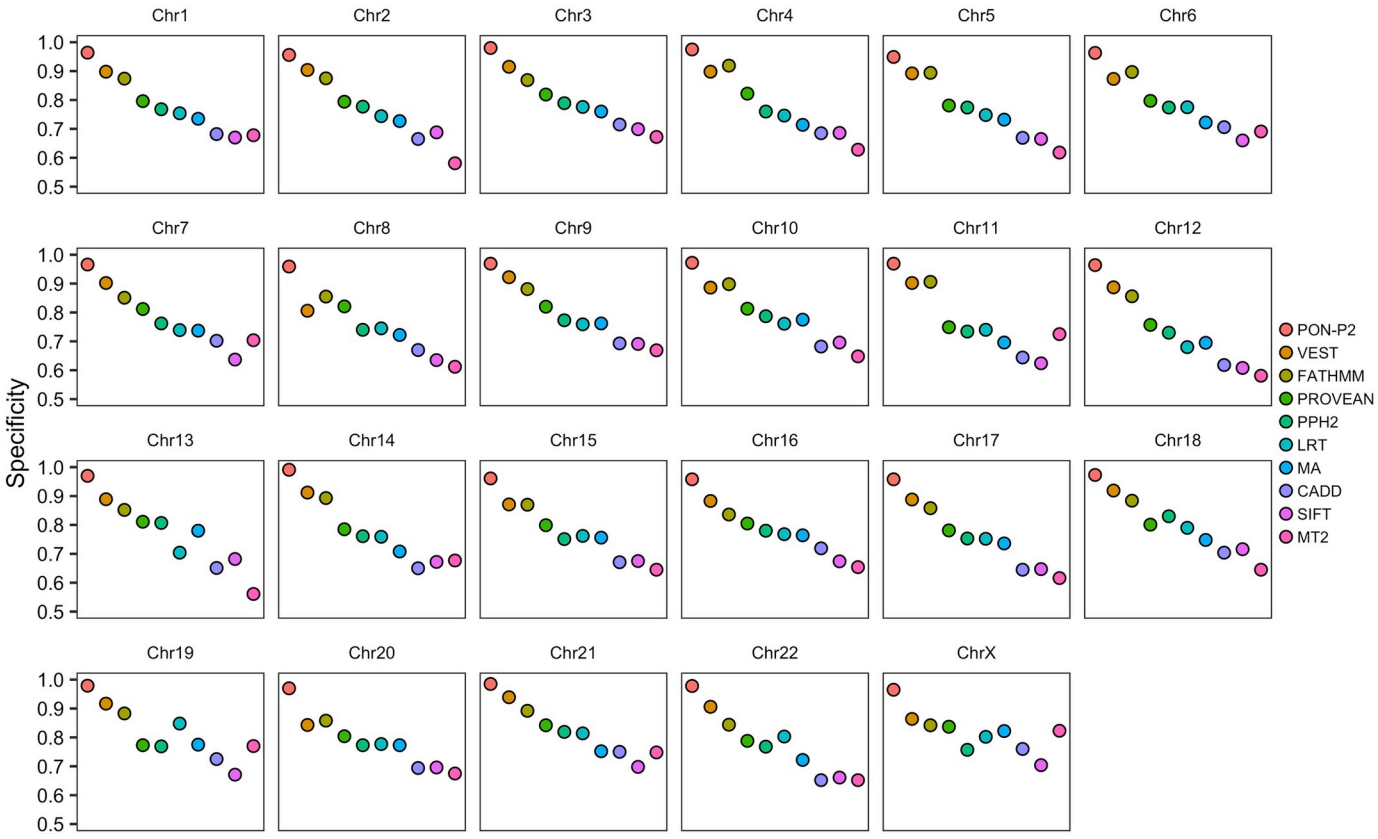

Fig 5. Chromosome-wise performance of tools. Variants in chromosome Y were excluded because there were only 3 variants. MA, Mutation Assessor; MT2, MutationTaster2; PPH2, PolyPhen-2.

https://doi.org/10.1371/journal.pcbi.1006481.g005

analyzed is a minor allele even when the variant site has a random distribution of the four nucleotide bases in the population. Although performance evaluation of prediction tools on such common variants may overestimate specificities of the tools, validated benign variants with low AF values are rare. Our results show that specificities increase with AF and have similar trend for all the tools (Fig 1). Therefore, assessments using the common variants provide useful comparison of the performance of predictors.

Our results show that the performances of tools in detecting the benign variants vary widely. The specificities of the tools ranged from 63.4% to 95.5% (Table 1). PON-P2 [22] had the best performance while MutationTaster2 [41], SIFT [44], and CADD [37] showed the poorest specificities. MutationTaster2 directly annotates the variants as disease-causing or benign based on the dbSNP [5], The 1000 Genomes Project [6], ClinVar [53], and HGMD [54] data. We excluded such automatic annotations in this study to compare the predictive performance of MutationTaster2.

In addition to the specificities of the tools, we also compared the performance on variants common in different geographical populations. All the methods showed performance differences for populations, the lowest specificity was achieved for the variants in the Finnish population (Fig 2). The variants that were unique in specific populations (AF $\geq$ 1% and $<$ 25% in the specific population but AF $<$ 1% in all other populations) were more difficult to predict. The tools showed from slightly to markedly lower performance for these variants (S3 and S4 Tables). Most of the unique variants had AFs $<$ 5% (Fig 3). To investigate the possibility of the

performance associated with low AF, we compared the performance for the unique variants and the non-unique variants (those with AF $\geq$ 1% in more than one population) with AF $<$ 5% in the same population. The comparison showed that the specificities were slightly poorer for the unique variants than for the non-unique variants. Differences in the performance on chromosome-wide analysis were very small for all the tools (Fig 5).

The methods showed very broad spectrum of performances; thus, it is important for the end-users in research as well as in precision medicine to pick a reliable one. Our results enable comparison of the tools and choosing the most reliable ones for interpretation of benign variants.

Supporting information

S1 Fig. Sensitivities and specificities of the tested predictors. (A) Sensitivities and specificities of CADD at different cutoffs of phred-like scores. The authors recommend using a phred-like score between 10 to 20 for distinguishing pathogenic and benign variants. Sensitivities (black) are calculated for 1301 pathogenic and likely pathogenic variants from ClinVar. The pathogenic variants in training datasets of tools could not be excluded. Specificities (grey) are calculated for 20602 variants with adjusted allele frequencies (AF Adj) between 1% to 25% obtained from ExAC. (B) Sensitivities and specificities of VEST at different cutoffs of VEST score. (C) Sensitivity and specificity for all the tested variant interpretation tools. (D) Sensitivity and specificity for variants that were predicted by all the methods. Variants that could not be predicted by any of the tools were excluded. The number of pathogenic variants was 480 and of neutral variants was 7268. The numbers of cases were normalized prior to calculation of sensitivity and specificity.
(DOCX)

S2 Fig. Statistical analysis of method performances. Fisher exact test was used for pairwise comparison of methods. The color coding indicates p value that ranges from 1 to 10–16, i.e. the steps indicate ten differences. (A) Comparison of all the data, and (B) variants that all the methods predicted.
(DOCX)

S1 Table. Percentages of variants that were not classified as pathogenic or benign.
(DOCX)

S2 Table. Proportion of unique variants in the populations.
(DOCX)

S3 Table. Specificity differences of tools between non-unique and unique variants in six populations (Specificity for non-unique variants—Specificity for unique variants).
(DOCX)

S4 Table. Specificity differences of tools between non-unique and unique variants with AF $\geq$1% and $<$5% in the populations.
(DOCX)

S5 Table. Chromosome-wide numbers of variants with AF $\geq$1% and $<$25% in male and female populations.
(DOCX)

Acknowledgments

We thank Rachel Karchin and Christopher Douville for providing the training dataset for VEST.

Author Contributions

Conceptualization: Mauno Vihinen.

Data curation: Abhishek Niroula.

Formal analysis: Abhishek Niroula, Mauno Vihinen.

Funding acquisition: Mauno Vihinen.

Investigation: Abhishek Niroula, Mauno Vihinen.

Methodology: Abhishek Niroula.

Project administration: Mauno Vihinen.

Resources: Mauno Vihinen.

Software: Abhishek Niroula.

Supervision: Mauno Vihinen.

Validation: Abhishek Niroula, Mauno Vihinen.

Visualization: Abhishek Niroula.

Writing – original draft: Abhishek Niroula, Mauno Vihinen.

Writing – review & editing: Abhishek Niroula, Mauno Vihinen.

References

1. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, et al. A map of human genome variation from population-scale sequencing. Nature. 2010; 467(7319):1061–1073. https://doi.org/10.1038/nature09534 PMID: 20981092
2. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016; 44(D1):D862–868. https://doi.org/10.1093/nar/gkv1222 PMID: 26582918
3. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. Nucleic Acids Res. 2015; 43(Database issue):D789–798. https://doi.org/10.1093/nar/gku1205 PMID: 25428349
4. The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017; 45 (Database issue):D158–D169. https://doi.org/10.1093/nar/gkw1099 PMID: 27899622
5. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001; 29(1):308–311. PMID: 11125122
6. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015; 526(7571):68–74. http://www.nature.com/nature/journal/v526/n7571/abs/nature15393.html#supplementary-information. PMID: 26432245
7. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature. 2013; 493(7431):216–220. https://doi.org/10.1038/nature11690 PMID: 23201682
8. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016; 536(7616):285–291. https://doi.org/10.1038/nature19057 PMID: 27535533
9. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet. 2015; 24(8):2125–2137. https://doi.org/10.1093/hmg/ddu733 PMID: 25552646
10. Korvigo I, Afanasyev A, Romashchenko N, Skoblov M. Generalising better: Applying deep learning to integrate deleteriousness prediction scores for whole-exome SNV studies. PLoS One. 2018; 13(3): e0192829. https://doi.org/10.1371/journal.pone.0192829 PMID: 29538399
11. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet. 2016; 99 (4):877–885. https://doi.org/10.1016/j.ajhg.2016.08.016 PMID: 27666373

12. Rogers MF, Shihab HA, Mort M, Cooper DN, Gaunt TR, Campbell C. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. Bioinformatics. 2018; 34(3):511–513. https://doi.org/10.1093/bioinformatics/btx536 PMID: 28968714
13. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. 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. 2015; 17(5):405–424. https://doi.org/10.1038/gim.2015.30 PMID: 25741868
14. Matthijs G, Souche E, Alders M, Corveleyn A, Eck S, Feenstra I, et al. Guidelines for diagnostic next-generation sequencing. Eur J Hum Genet. 2016; 24(1):2–5. https://doi.org/10.1038/ejhg.2015.226 PMID: 26508566
15. Li MM, Datto M, Duncavage EJ, Kulkarni S, Lindeman NI, Roy S, et al. Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer: A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. J Mol Diagn. 2017; 19(1):4–23. https://doi.org/10.1016/j.jmoldx.2016.10.002 PMID: 27993330
16. Thusberg J, Olatubosun A, Vihinen M. Performance of mutation pathogenicity prediction methods on missense variants. Hum Mutat. 2011; 32(4):358–368. https://doi.org/10.1002/humu.21445 PMID: 21412949
17. Niroula A, Vihinen M. Variation interpretation predictors: principles, types, performance and choice. Hum Mutat. 2016; 37(6):579–597. https://doi.org/10.1002/humu.22987 PMID: 26987456
18. Tang H, Thomas PD. Tools for predicting the functional impact of nonsynonymous genetic variation. Genetics. 2016; 203(2):635. https://doi.org/10.1534/genetics.116.190033 PMID: 27270698
19. Peterson TA, Doughty E, Kann MG. Towards precision medicine: advances in computational approaches for the analysis of human variants. J Mol Biol. 2013; 425(21):4047–4063. https://doi.org/10.1016/j.jmb.2013.08.008 PMID: 23962656
20. Grimm DG, Azencott CA, Aicheler F, Gieraths U, MacArthur DG, Samocha KE, et al. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. Hum Mutat. 2015; 36(5):513–523. https://doi.org/10.1002/humu.22768 PMID: 25684150
21. Bendl J, Stourac J, Salanda O, Pavelka A, Wieben ED, Zendulka J, et al. PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. PLoS Comput Biol. 2014; 10(1): e1003440. https://doi.org/10.1371/journal.pcbi.1003440 PMID: 24453961
22. Niroula A, Urolagin S, Vihinen M. PON-P2: prediction method for fast and reliable identification of harmful variants. PLoS One. 2015; 10(2):e0117380. https://doi.org/10.1371/journal.pone.0117380 PMID: 25647319
23. Riera C, Padilla N, de la Cruz X. The Complementarity Between Protein-Specific and General Pathogenicity Predictors for Amino Acid Substitutions. Hum Mutat. 2016. https://doi.org/10.1002/humu.23048 PMID: 27397615
24. Nair PS, Vihinen M. VariBench: a benchmark database for variations. Hum Mutat. 2013; 34(1):42–49. https://doi.org/10.1002/humu.22204 PMID: 22903802
25. Schaafsma GC, Vihinen M. VariSNP, A Benchmark Database for Variations From dbSNP. Hum Mutat. 2015; 36(2):161–166. https://doi.org/10.1002/humu.22727 PMID: 25385275
26. Vihinen M. How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. BMC Genomics. 2012; 13 Suppl 4:S2. https://doi.org/10.1186/1471-2164-13-s4-s2 PMID: 22759650
27. Vihinen M. Guidelines for reporting and using prediction tools for genetic variation analysis. Hum Mutat. 2013; 34(2):275–282. https://doi.org/10.1002/humu.22253 PMID: 23169447
28. Vihinen M. No more hidden solutions in bioinformatics. Nature. 2015; 521(7552):261. https://doi.org/10.1038/521261a PMID: 25993922
29. Desmet F, Hamroun G, Collod-Beroud G, Claustres M, Beroud C. Bioinformatics identification of splice site signals and prediction of mutation effects. In: Mohan RM, editor. Research Advances in Nucleic Acids Research. Kerala: Global Reseach Network; 2010. p. 1–16.
30. Jian X, Boerwinkle E, Liu X. In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Res. 2014; 42(22):13534–13544. https://doi.org/10.1093/nar/gku1206 PMID: 25416802
31. Khan S, Vihinen M. Performance of protein stability predictors. Hum Mutat. 2010; 31(6):675–684. https://doi.org/10.1002/humu.21242 PMID: 20232415
32. Potapov V, Cohen M, Schreiber G. Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. Protein Eng Des Sel. 2009; 22(9):553–560. https://doi.org/10.1093/protein/gzp030 PMID: 19561092

33. Yang Y, Niroula A, Shen B, Vihinen M. PON-Sol: prediction of effects of amino acid substitutions on protein solubility. Bioinformatics. 2016; 32(13):2032–2034. https://doi.org/10.1093/bioinformatics/btw066 PMID: 27153720

34. Laurila K, Vihinen M. Prediction of disease-related mutations affecting protein localization. BMC genomics. 2009; 10:122–122. https://doi.org/10.1186/1471-2164-10-122 PMID: 19309509

35. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016; 17(1):1–14. https://doi.org/10.1186/s13059-016-0974-4 PMID: 27268795

36. Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: A one-stop database of functional predictions and annotations for human non-synonymous and splice site SNVs. Hum Mutat. 2016; 37(3):235–241. https://doi.org/10.1002/humu.22932 PMID: 26555599

37. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014; 46(3):310–315. https://doi.org/10.1038/ng.2892 PMID: 24487276

38. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GL, Edwards KJ, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat. 2013; 34(1):57–65. https://doi.org/10.1002/humu.22225 PMID: 23033316

39. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009; 19(9):1553–1561. https://doi.org/10.1101/gr.092619.109 PMID: 19602639

40. Reva B, Antipin Y, Sander C. Determinants of protein function revealed by combinatorial entropy optimization. Genome Biol. 2007; 8(11):R232. https://doi.org/10.1186/gb-2007-8-11-r232 PMID: 17976239

41. Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods. 2014; 11(4):361–362. https://doi.org/10.1038/nmeth.2890 PMID: 24681721

42. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010; 7(4):248–249. https://doi.org/10.1038/nmeth0410-248 PMID: 20354512

43. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS One. 2012; 7(10):e46688. https://doi.org/10.1371/journal.pone.0046688 PMID: 23056405

44. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003; 31(13):3812–3814. PMID: 12824425

45. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics. 2013; 14 Suppl 3:S3. https://doi.org/10.1186/1471-2164-14-s3-s3 PMID: 23819870

46. Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nat Genet. 2016; 48(12):1581–1586. https://doi.org/10.1038/ng.3703 PMID: 27776117

47. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. 2010; 7(8):575–576. https://doi.org/10.1038/nmeth0810-575 PMID: 20676075

48. Niroula A, Urolagin S, Vihinen M. PON-P2: Prediction method for fast and reliable identification of harmful variants. PLoS ONE. 2015;(10(2):e0117380). https://doi.org/10.1371/journal.pone.0117380 PMID: 25647319

49. Walsh I, Pollastri G, Tosatto SC. Correct machine learning on protein sequences: a peer-reviewing perspective. Brief Bioinform. 2016; 17(5):831–840. https://doi.org/10.1093/bib/bbv082 PMID: 26411473

50. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491(7422):56–65. https://doi.org/10.1038/nature11632 PMID: 23128226

51. Hoskins RA, Repo S, Barsky D, Andreoletti G, Moult J, Brenner SE. Reports from CAGI: The Critical Assessment of Genome Interpretation. Hum Mutat. 2017; 38(9):1039–1041. https://doi.org/10.1002/humu.23290 PMID: 28817245

52. Kobayashi Y, Yang S, Nykamp K, Garcia J, Lincoln SE, Topper SE. Pathogenic variant burden in the ExAC database: an empirical approach to evaluating population data for clinical variant interpretation. Genome Med. 2017; 9(1):13. https://doi.org/10.1186/s13073-017-0403-7 PMID: 28166811

53. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014; 42(Database issue):D980–985. https://doi.org/10.1093/nar/gkt1113 PMID: 24234437

54. Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet. 2014; 133(1):1–9. https://doi.org/10.1007/s00439-013-1358-4 PMID: 24077912