https://doi.org/10.1093/nar/gkad455

PhD-SNPg: updating a webserver and lightweight tool for scoring nucleotide variants

Emidio Capriotti [1,*] and Piero Fariselli [2,*]

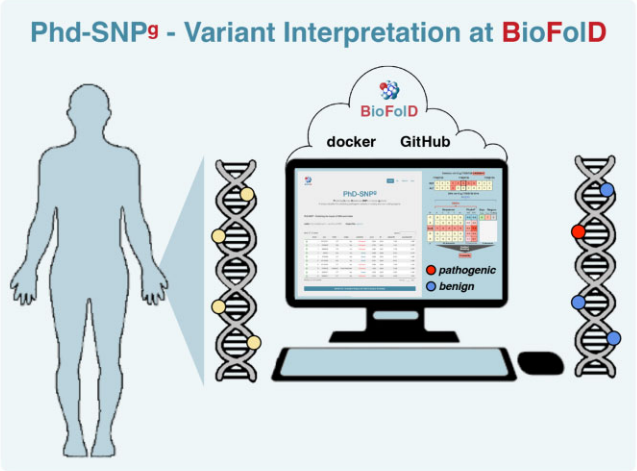
[1]BioFolD Unit, Department Pharmacy and Biotechnology (FaBiT), University of Bologna, Via F. Selmi 3, Bologna 40126, Italy and [2]Department of Medical Sciences, University of Torino, Via Santena 19, 10126, Torino, Italy

Received March 31, 2023; Revised May 08, 2023; Editorial Decision May 09, 2023; Accepted May 14, 2023

ABSTRACT

One of the primary challenges in human genetics is determining the functional impact of single nucleotide variants (SNVs) and insertion and deletions (InDels), whether coding or noncoding. In the past, methods have been created to detect disease-related single amino acid changes, but only some can assess the influence of noncoding variations. CADD is the most commonly used and advanced algorithm for predicting the diverse effects of genome variations. It employs a combination of sequence conservation and functional features derived from the ENCODE project data. To use CADD, a large set of pre-calculated information must be downloaded during the installation process. To streamline the variant annotation process, we developed PhD-SNP^g, a machine-learning tool that is easy to install and lightweight, relying solely on sequence-based features. Here we present an updated version, trained on a larger dataset, that can also predict the impact of the InDel variations. Despite its simplicity, PhD-SNP^g performs similarly to CADD, making it ideal for rapid genome interpretation and as a benchmark for tool development.

GRAPHICAL ABSTRACT



INTRODUCTION

Recent advances in sequencing technology have led to an exponential growth of the observed genetic variants in humans (1,2), whose effects are poorly understood. Most of the available data were generated by international consortiums, aiming to characterize the pattern of genetic variations across individuals (3,4) and identify mutations associated with human diseases (5,6).

Thus, predicting the functional effect of genetic variants is a key challenge for interpreting the human genome, and in turn, for implementing more accurate diagnostic and treatment strategies (7,8). In the last few years, several methods have been developed for predicting the impact of Single Nucleotide Variants (SNVs) and Insertions/Deletions (InDels) on human health. Nevertheless, few of them can assess the effect of those genome variations in noncoding regions (9).

One of the primary challenges in human genetics is determining the impact of these genetic variations. Several methods have been developed to identify disease-related single

[*]To whom correspondence should be addressed. Tel: +39 051 209 4303; Fax: +39 051 209 4286; Email: emidio.capriotti@unibo.it
Correspondence may also be addressed to Piero Fariselli. Email: piero.fariselli@unito.it

amino acid changes. However, few tools are available to score the impact of noncoding variants, which is crucial for understanding their contribution to disease (10).

Among the most popular algorithms for predicting the effect of SNVs in noncoding regions are CADD (11) and FATHMM (12,13). These tools combine sequence conservation with functional features derived from the ENCODE project data. However, the installation process for these algorithms requires downloading a large set of pre-calculated information, which can be time-consuming and resource-intensive. To address this issue, we created PhD-SNP^g, a lightweight tool that relies solely on sequence-based features (14).

This paper presents an updated PhD-SNP^g version, which introduces some novelties. First, it is trained on a more extensive set of variants (more than three times than the original method), improving the previous performance. Second, the new version of PhD-SNP^g can also deal with InDel predictions in coding and noncoding regions. Finally, PhD-SNP^g expands the evolutionary information by extending the input conservation scores from 100 to 470 aligned species (from PhyloP7/PhyloP100 to PhyloP100/PhyloP470).

PhD-SNP^g is designed to be easy to install and use, and it is available both as a web server and standalone software for processing large datasets of variants locally. The machine learning core of the tool is trained only on comparative information in the form of conservation scores calculated from multiple sequence alignments extracted from the UCSC repository. This information is obtained from the University of California, Santa Cruz (UCSC) genome browser (https://genome.ucsc.edu/), a widely-used resource in the field of genomics.

Compared to other state-of-the-art methods like CADD (11), FATHMM-MKL (14) and GVAWA (15), PhD-SNP^g requires relatively few input resources, making it easier to install and run on new sets of variations, even on laptop computers. The full version of PhD-SNP^g only needs less than 30 Gb of UCSC data, while FATHMM-MKL and CADD require 400 Gigabytes or more. Additionally, a second lightweight version of PhD-SNP^g (∼100Mb) can run in *'web mode'* by retrieving UCSC data directly from their URLs without downloading the entire genome files.

Because PhD-SNP^g has such simple input requirements (nucleotide sequence and conservation score), it can also be used as a baseline tool for benchmarking algorithms that use more complex input features, for example, to estimate the improvement in performance achieved by adding new input features such as open chromatin, histone modification, and transcription factor binding sites. All the training and testing datasets used to develop PhD-SNP^g are available online to facilitate this process.

Having benchmark datasets available is crucial for evaluating the discriminative power of new methods with different input features, while avoiding overestimating performance (16). By providing a lightweight and easy-to-use tool for predicting the functional effects of SNVs in coding and noncoding regions, the developers of PhD-SNP^g hope to make it easier for researchers to explore the complex genetic landscape of human diseases.

METHOD OUTLINE

We did not optimize the hyper-parameters for this updated version of PhD-SNP^g but kept the ones previously selected. PhD-SNP^g is a binary classifier based on a Gradient Boosting algorithm, as implemented in *scikit-learn* package (17). PhD-SNP^g was trained and trained using two sets of ∼104,000 SNVs and ∼34,000 InDels extracted from the Clinvar database (18). Both testing sets were generated from two versions of Clinvar released in December 2020 and 2022. The location and the type of the SNVs and InDels are depicted on the corresponding human chromosome cartoon (*Pathogenic* in red and *Benign* in blue) of Figure 1A and B, respectively.

Dataset selection

The datasets of SNVs and InDels used for training and testing PhD-SNP^g were extracted from Clinvar (18) (http://www.ncbi.nlm.nih.gov/clinvar/). The Clinvar database (version December 2020) was filtered by selecting the SNVs, and InDels annotated *Pathogenic, Likely pathogenic* or *Benign and Likely benign* annotation. After this filtering, we ended up with a dataset of SNVs (Clinvar122020-SNV) that consists of 51,958 *Pathogenic* and 120,826 *Benign* SNVs. In the Clinvar122020-SNV dataset, 6,261 (12%) of the *Pathogenic* and 40,878 (34%) of the *Benign* SNVs are in noncoding regions. From the same version of Clinvar, it was collected a dataset of InDels (Clinvar122020-InDel) consisting of 37,421 *Pathogenic* and 4,523 *Benign* InDels. In the Clinvar122020-InDel dataset, 1,200 (3%) of the *Pathogenic* and 3,405 (75%) of the *Benign* InDels are in noncoding regions.

To assess the performance of the method, we derived two other datasets based on a more recent version of Clinvar (December 2022), by selecting annotated SNVs and InDels not present in the Clinvar122020-SNV and Clinvar122020-InDel datasets. A new dataset of SNVs (NewClinvar122022-SNV), comprises 104,716 SNVs, 21,299 of which are annotated as *Pathogenic* and 83,417 as *Benign*. In the NewClinvar122022-SNV dataset, 2,334 (11%) of the *Pathogenic* and 62,917 (75%) of the *Benign* SNVs are in noncoding regions. From the latest version of the Clinvar database, we extracted a new dataset (NewClinvar122022-InDel) including 24,382 *Pathogenic* and 17,145 *Benign* InDels. Among them, 922 (4%) of the *Pathogenic* and 16,359 (95%) of the *Benign* InDels are in noncoding regions. The composition of all datasets is summarized in Supplementary Tables S1–S3 and Supplementary Figure S1.

After the collection of the datasets of SNVs (Clinvar122020-SNV, NewClinvar122022-SNV) and InDels (Clinvar122020-InDel, NewClinvar122022-InDel) a specific procedure was used for generating balanced training and testing sets. For scoring the performance of PhD-SNP^g in the prediction of SNVs, we generated training and testing composed of 103,916 and 42,594 SNVs, respectively. In the training and testing sets, the fraction of *Pathogenic* and *Benign* variants was balanced by downsampling randomly from the most abundant class (Benign) in the Clinvar122020-SNV, NewClinvar122022-

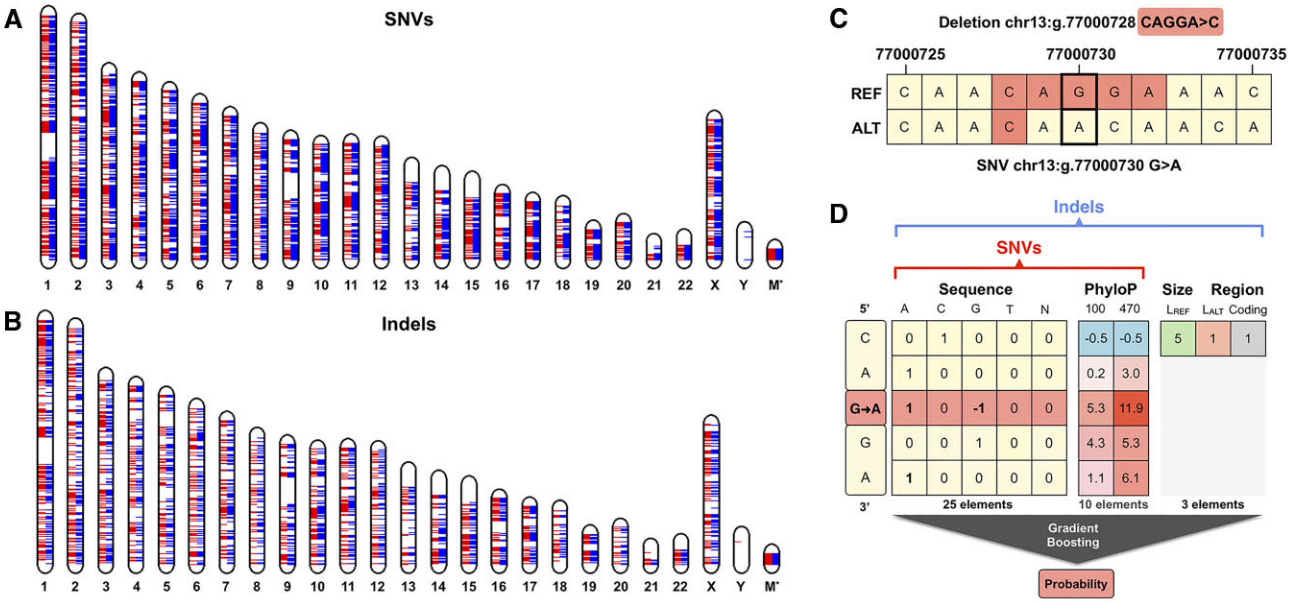


Figure 1. (**A, B**) Distribution along the chromosomes of *Pathogenic* (red) and *Benign* (blue) single nucleotide variants (SNVs) and InDels respectively. *The size of the mitochondrial chromosome (M) in panels A and B is increased 2,500 times. The impact of an InDel is predicted considering the closest SNV caused by the sequence variation. In the example in Panel **C**, the deletion of AGGA in position 77,000,728 of the chromosome 13 generates a variation from G to A in position 77,000,730. (**D**) Schematic view of the input features of PhD-SNP^g algorithm for predicting pathogenic SNVs and InDels which takes in input 35 and 38 features respectively.

SNV datasets, respectively. A similar procedure was applied for generating the testing and training sets for predicting the impact of InDels which include 34,290 and 9,046 InDels, respectively. Given the small number of benign InDels and strong unbalance between two classes of InDels in the Clinvar122020-InDel dataset, the training and testing sets of InDels were generated from NewClinvar122022-InDel and Clinvar122020-InDel respectively. In this case, the random downsampling was performed on the subset of pathogenic InDels, which represent the most abundant class. The Initial datasets of SNVs and InDels, as well as five balanced versions for each training and testing set, are provided as supplementary files. The genomic location in those files is based on the hg38 human genome assembly.

Feature evaluation

We implemented two versions of PhD-SNP^g for predicting the impact of SNVs and InDels. The input of the method predicting the impact of SNVs consists of 35 values, 25 encoding for the sequence and mutation, and 10 for the PhyloP conservation scores (19), as pre-computed at the UCSC repository. In detail, the input is composed by: (i) 25 values representing the 5-nucleotide window sequence centered on the mutated position (5 times 5 possible nucleotides: A, C, G, T, N); (ii) 10 values mapping the conservation scores of the 100-species (PhyloP100) and 470-species alignments (PhyloP470) to the five window positions. Comparing the PhyloP scores adopted in the new version of PhD-SNP^g with PhyloP7 used in the previous version, PhyloP100 shows the highest discriminative power (Supplementary Tables S4-S6). This result is confirmed by plotting its distribution for *Pathogenic* and *Benign* SNVs (Supplementary Figures S2-S3).

A similar approach was implemented for predicting the pathogenic InDels. In particular, we assume that the effect of an InDel corresponds to the effect of the closest SNV that is obtained by deleting and/or inserting a set of nucleotides in a given region of the genome. In Figure 1C, we represented the example of the deletion chr13:g.77000728 CAGGA >C which, in the closets loci, corresponds to the change of G (Guanine) to A (Adenine) in position 77,000,730 of chromosome 13. Using this assumption, we developed a second version of PhD-SNP^g for predicting the impact of the InDels which takes in input 38 values. In detail, the input is composed of 35 values used for predicting the impact of SNVs and three new features encoding for the size and location of the InDel. They represent the lengths of the reference and alternative alleles and a boolean variable corresponding to the location of the mutated loci in coding or noncoding regions (Figure 1D).

Method evaluation

We defined two datasets, one from model development, using a cross-validation procedure to evaluate the performance and a second hold-out set to evaluate the method generalization. We operated a 10-fold split for the cross-validation procedure, which was accomplished five times by employing a bootstrapping procedure to generate different cross-validation sets. The mean value of the accuracy indices of these five bootstraps is used as the final score.

In order to limit bias due to having the same genomic regions between the training and testing sets, we split the variants (SNVs and InDels) by chromosomes. Thus, when we predict variants of a given chromosome, the model used was trained using only variants from other chromosomes.

Table 1. Performance of PhD-SNP^{g}, CADD and FATHMM-MKL on the NewClinvar122022-SNV dataset. Average results of the 5 bootstrap tests (10-fold) performed on the Clinvar122020-SNV dataset. Q2: Overall Accuracy, TNR: True Negative Rate, NPV: Negative Predictive Value, TPR: True Positive Rate, PPV: Positive Predicted Value, MCC: Matthews Correlation Coefficient, AUC: Area Under the Receiver Operating Characteristic Curve, Brier: Brier Score. PhD-SNP^{g} performance evaluation measures (defined in Supplementary Materials) are averaged over 5 bootstrap tests (10-fold)

Method	Subset	Q3	TNR	NPV	TPR	PPV	MCC	F1	AUC	Brier	DB
PhD-SNP^{g}	*All*	0.898	0.883	0.909	0.912	0.887	0.796	0.899	0.959	0.076	100.0
	Coding	0.892	0.869	0.910	0.914	0.876	0.784	0.894	0.956	0.080	84.0
	Noncoding	0.930	0.958	0.909	0.900	0.954	0.861	0.926	0.970	0.055	16.0
CADD	*All*	0.911	0.845	0.973	0.976	0.863	0.828	0.916	0.982	NA	100.0
	Coding	0.901	0.817	0.981	0.984	0.844	0.813	0.909	0.983	NA	84.0
	Noncoding	0.962	0.987	0.941	0.936	0.986	0.925	0.960	0.985	NA	15.9
FATHMM-MKL	*All*	0.761	0.628	0.856	0.894	0.706	0.542	0.789	0.865	0.179	99.9
	Coding	0.728	0.567	0.833	0.887	0.673	0.480	0.766	0.836	0.204	83.9
	Noncoding	0.939	0.945	0.935	0.932	0.943	0.878	0.937	0.974	0.050	16.0

We adopted the same strategy to evaluate the variants in the hold-out set. We predict variants of a given chromosome using the model fitted during the cross-validation phase, which did not contain variants from the chromosome to test in its training set.

We tested if there are shared genes among the different chromosomes, and we found that only between chromosome X and Y there are a few shared genes and for this reason we kept the variations located into these two chromosomes in the same fold. A representation of this procedure is shown in Supplementary Figure S4.

Prediction performance

Initially, PhD-SNP^{g} performance was evaluated using a 10-fold cross-validation test on approximately 104 000 SNVs and 34 000 InDels. PhD-SNP^{g} achieved an Area Under the Receiver Operating Characteristics (19) Curve (AUC) of 0.95 and 0.99 for SNVs and InDels, respectively, on these subsets. These results are presented in Supplementary Tables S7 and S8 and Supplementary Figure S5. In the first test, the new version of PhD-SNP^{g} demonstrated comparable performance to the previous version, which obtained an AUC of ∼0.95 (Supplementary Tables S9-S10). In order to assess the generalization ability of the predictor and compare PhD-SNP^{g} performance with that of CADD and FATHMM, two sets of annotated variants composed of around 42 000 SNVs and 9000 InDels were extracted from the NewClinvar122022-SNV and Clinvar122020-InDel, respectively. On the NewClinvar122022-SNV testing set, the AUC of PhD-SNP^{g} was 0.96, which is still high and comparable to that obtained in the cross-validation test. Notably, this score is similar to that obtained on the same set by CADD (Table 1 and Figure 2A). This trend is also observed in the subsets of mutations located in coding and noncoding regions (Table 1 and Figure 2B and C). These surprising findings confirm that PhD-SNP^{g} achieves good performance despite having fewer features than other approaches. On average FATHMM-MKL achieves lower performance than PhD-SNP^{g} and CADD, nevertheless it shows comparable AUC on the subset of noncoding variants. Similar results were obtained when evaluating PhD-SNP^{g}'s performance in predicting pathogenic InDels. In this particular task, PhD-SNP^{g} achieved an AUC of approximately 0.97, which is slightly higher than that achieved by CADD. However, CADD is slightly more accurate than PhD-SNP^{g} in predicting the impact of InDels occurring in coding regions (Table 2 and Figure 2D–F). The results also show that both CADD and PhD-SNP^{g} performs better than FATHMM-Indel which reaches similar AUC on the subset of coding InDels. An examination of the predictions of PhD-SNP^{g} on the subsets of insertion and deletions, which represent approximately 32% and 65% of the entire testing set, respectively, demonstrated a comparable level of performance in terms of AUC (Supplementary Table S11). In this analysis, we also scored the calibration of the predictions. This property refers to the idea that the probability associated with the prediction of pathogenicity should match the expected value of true positives (20). The results based on the calculation of the Brier score (21) indicate that PhD-SNP^{g} predictions of pathogenic SNVs and InDel show a good level of calibration. Indeed, in the majority of the cases, the Brier score of PhD-SNP^{g} is below 0.1. In order to prevent this source of bias in all prediction tests, the training and testing sets were split in such a way that variants on the same chromosome were kept in the same subset. Supplementary Materials provide additional information on how PhD-SNP^{g} was compared to CADD and FATHMM, as well as the definition of the performance evaluation measures and the relative standard errors (Supplementary Tables S12-S13). Furthermore, we calculated the method performance on more balanced subsets with reduced variants from the most represented genes to estimate the possible effect of overrepresented variants from specific genes. The results in Supplementary Tables S14 and S15 show that the performance scores on the testing dataset obtained by selecting a maximum of 5 or 10 variants for each gene are similar to those obtained on the whole set of coding variants.

In the last part of our analysis (Table 3 and Supplementary Table S16), we evaluated the performance of PhD-SNP^{g} on different subsets of variants according to the classification reported by ANNOVAR (22). We grouped SNVs and InDels into six classes, which are: '*exonic*', '*intronic*', '*splicing*', '*noncoding RNA*' and '*other*' (the combination of all remaining noncoding elements). The distribution shows that in the ClinVar dataset, the most-representative class is the coding (*exonic*) class for both SNVs and InDels (Table 3 and Supplementary Table S16). As far as InDels are concerned, the *intronic* is well represented, leaving the other classes with few instances. For this reason, the AUC scores are generally very good for all classes (Table 3), while the measures of accuracy that require a decision threshold (such as MCC, F1, ecc.) fluctuate more, except for the most abundant *exonic* class.

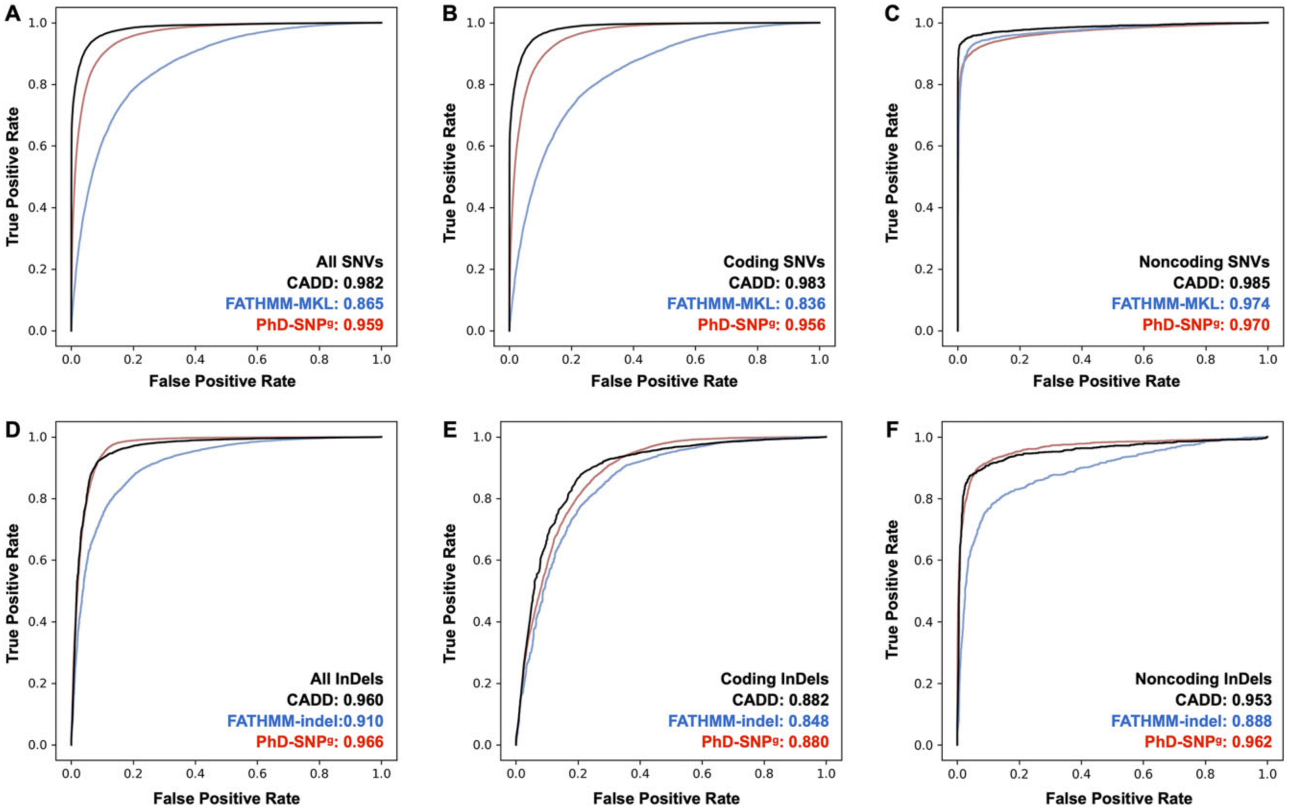


Figure 2. Comparison of the area under receiver operator characteristic curve (ROC) for CADD (red) and PhD-SNP^g (blue). Comparison of the area under receiver operator characteristic curve (ROC) for CADD (black), PhD-SNP^g (red) and FATHMM-MKL/indel (blue). The ROC curves are calculated on the datasets NewClinvar122022-SNV (**A**) and Clinvar122020-InDels (**D**) and their subset of coding and noncoding SNVs (**B**, **C**) and InDels (**E**, **F**).

Table 2. Performance of PhD-SNP^g, CADD and FATHMM-indel on the Clinvar122020-InDel dataset. Average results of the 5 bootstrap tests (10-fold) performed on the NewClinvar122022-InDel dataset. Q2: overall Accuracy, TNR: true negative rate, NPV: negative predictive value, TPR: true positive rate, PPV: positive predicted value, MCC: Matthews correlation coefficient, AUC: area under the receiver operating characteristic curve, Brier: Brier score. PhD-SNP^g performance evaluation measures (defined in Supplementary Materials) are averaged over 5 bootstrap tests (10-fold)

Method	Subset	Q3	TNR	NPV	TPR	PPV	MCC	*F*1	AUC	Brier	%DB
PhD-SNP^g	*All*	0.907	0.830	0.981	0.984	0.853	0.824	0.914	0.966	0.081	100.0
	Coding	0.877	0.357	0.933	0.994	0.873	0.532	0.930	0.880	0.107	56.6
	Noncoding	0.946	0.954	0.987	0.865	0.642	0.718	0.737	0.962	0.048	43.4
CADD	*All*	0.915	0.922	0.909	0.907	0.921	0.830	0.914	0.960	NA	99.9
	Coding	0.878	0.717	0.651	0.914	0.935	0.607	0.924	0.882	NA	56.6
	Noncoding	0.963	0.976	0.984	0.830	0.765	0.777	0.796	0.953	NA	43.3
FATHMM-indel	*All*	0.833	0.771	0.883	0.896	0.794	0.672	0.842	0.910	0.127	95.2
	Coding	0.855	0.638	0.574	0.901	0.922	0.517	0.911	0.848	0.113	53.1
	Noncoding	0.805	0.803	0.982	0.830	0.263	0.394	0.399	0.888	0.144	42.1

Method usage and output of the prediction

PhD-SNP^g can predict the effect of single and multiple SNVs from an input file. Variant Calling Format (VCF) file is also accepted as input. By default, our server, and scripts accept as input genomic coordinates from the hg38 assemblies of the human genome. Genomic coordinates based on hg19 assembly are internally converted to hg38 coordinates by the *liftOver* program.

The application of our method is limited by the availability of the conservation score. Indeed, PhD-SNP^g predictions can be performed only on genomic regions for which either the PhyloP100 or PhyloP470 score is available.

The primary output of PhD-SNP^g probabilistically assigns as '*Pathogenic*' or '*Benign*' score. Additionally, PhD-SNP^g generates three supplementary values to support the prediction, namely the False Discovery Rate (FDR), the PhyloP100 score at the mutated site, and the average PhyloP100 score calculated over the 5-nucleotide input window. FDR can be employed to eliminate less reliable predictions.

SERVER DETAILS

Predicting the impact of single nucleotide variants and InDels

The PhD-SNP^g server is able to predict the impact of a single nucleotide variant through either a CSV or VCF text

Table 3. Performance of PhD-SNP^g on the testing set of SNVs (NewClinVar122022) and InDels (ClinVar122020) classified according to their location. The variants are classified in exonic, intronic, splicing, noncoding RNA and UTR using ANNOVAR. Average performance on the 5 bootstrap tests (10-fold) performed on both datasets. Q2, TNR, NPV, TPR, PPV, MCC, F1, AUC and Brier are defined in Supplementary Materials

Variants	Subset	Q_2	TNR	NPV	TPR	PPV	MCC	F1	AUC	Brier	%DB
SNVs	*All*	0.898	0.883	0.909	0.912	0.887	0.796	0.899	0.959	0.076	100.0
(NewClinVar122022)	*Exonic*	0.892	0.871	0.911	0.914	0.875	0.785	0.894	0.956	0.080	84.9
	Intronic	0.892	0.963	0.916	0.463	0.676	0.502	0.550	0.835	0.084	5.9
	Splicing	0.975	0.705	0.058	0.976	0.999	0.198	0.988	0.960	0.020	6.6
	ncRNA	0.923	0.958	0.953	0.683	0.712	0.653	0.697	0.897	0.059	0.8
	UTR	0.920	0.951	0.964	0.416	0.342	0.335	0.374	0.744	0.062	1.5
	Other	0.815	0.963	0.821	0.391	0.783	0.461	0.517	0.747	0.151	0.2
InDels	*All*	0.907	0.830	0.981	0.984	0.853	0.824	0.914	0.966	0.081	100.0
(ClinVar122020)	*Exonic*	0.872	0.376	0.920	0.992	0.868	0.539	0.926	0.885	0.111	60.7
	Intronic	0.979	0.984	0.994	0.448	0.226	0.307	0.299	0.891	0.019	27.9
	Splicing	0.848	0.844	0.902	0.856	0.777	0.689	0.814	0.929	0.135	1.6
	ncRNA	0.947	0.971	0.974	0.422	0.376	0.367	0.390	0.864	0.046	1.4
	UTR	0.938	0.942	0.996	0.430	0.045	0.123	0.082	0.822	0.055	7.9
	Other	0.902	0.991	0.907	0.190	0.500	0.277	0.267	0.763	0.096	0.3

format (as previously described in 14). In this newer version, it is possible to provide multiple SNVs and InDels by copying and pasting a list of variants in separate rows within the input box. If the input is in VCF format, a file containing a header and at least five columns (CHROM, POS, ID, REF, ALT) separated by a tab character should be uploaded. Each SNV and InDel is then listed on separate rows after the header. If the variant's ID in the third column is missing or unavailable, a dot (.) can be used.

When the list of SNVs and InDels are provided, the server analyzes each variant and checks if the reference allele corresponds to the allele reported in the selected version of the human genome (hg19 or hg38). This task is performed using the *twoBitToFa* program (20). A window sequence of 5 nucleotides centered around the mutated position is used to generate the 25-element vector encoding for the sequence information. If the nucleotide in the input matches the reference allele, the server extracts the corresponding conservation indexes (PhyloP100 and PhyloP470) for the positions around the mutation site. The pre-calculated conservation indexes, which are available on the UCSC repository, are collected using the *bigWigToBedGraph* program (23). The PhyloP100 and PhyloP470 scores are used to generate a 10-element vector, which represents the conservation features. After this step, the 35-element vector encoding for the sequence and conservation features is given as input to the Gradient Boosting algorithm, which returns the prediction output described above. For predicting pathogenic InDels, the server identifies variants occurring at the closest loci. The 35-element vector described above as generated and complemented with three features describing the size and location of the variation. After the calculation of the prediction, in the final step of the prediction task, the PhD-SNP^g server annotates the input variants using the *TransVar* tool (24). *TransVar* finds the possible effect on the amino acid sequence of the longest matching transcript corresponding to the mutated region.

Alternative input format for single amino acid variants

The PhD-SNP^g server simplifies the prediction of single amino acid variants (SAVs) by accepting a list of SAVs as input. Each SAV in the list should be indicated by the approved HGNC (HUGO Gene Nomenclature Committee) gene symbol (25) and the amino acid change, which should be separated by a comma. The amino acid change can be described by combining the one-letter symbol of the wild-type residue, the position of the residue along the protein sequence, and the one-letter symbol of the mutant residue. For instance, the substitution of Methionine (M) at position 237 with Isoleucine (I) in TP53 is represented as TP53,M237I. By submitting the input to the PhD-SNP^g server in this MUT format, the server maps the protein change back to the corresponding genomic variant using the *TransVar* algorithm. The SNVs are then predicted for their impact using the same method described earlier.

Input interface

PhD-SNP^g web interface features a textarea box that allows users to input SNVs and InDels in CSV and MUT formats. Additionally, a 'Browse' button allows users to upload CSV and VCF files in either standard text or gzipped format. Three 'Input Type' buttons (CSV, VCF and MUT) enable users to select the appropriate input format for the list of variants. A second group of buttons (Assembly) indicates the human reference genome (hg19 or hg38) to which the SNVs refer. Users can access examples of inputs in CSV and MUT format using the 'chr,pos,ref,alt' and 'gene,mut' links at the top of the web interface. However, the usage of the textarea box for the VCF input format is discouraged, although an example of VCF-like input is linked in the 'Help' web page. Finally, an optional email box is available at the bottom of the PhD-SNP^g web page for receiving the output via email.

Server output

The output of PhD-SNP^g is a web page that displays the prediction results in a tabular format. On the top of the page, the *JobID* of the prediction process and the link to the output in text format (output.txt) are provided. In the JavaScript *d3* (https://d3js.org/) table, the predictions associated with each SNV are reported in rows composed of nine columns. The first four columns define the variants, and the remaining five provide information about the prediction. From left to right, the five prediction columns are: the result of the binary classifier (prediction), the probabilistic output (score) defined above, the associated false discovery rate (fdr), the value of the PhyloP100 score in the mutated site (phylop100) and the average value of the PhyloP100 scores for the five positions centered on the mutated site (avg-phylop100). A plus sign (+) at the beginning of each row allows the visualization of the results of the annotation performed by the *TransVar* algorithm. When a variant maps to a coding region, four rows report the RefSeq (26) code of the longest transcript (Transcript), the HGNC gene symbol and the associated UniProt (27) identifiers (Gene), the sense of the translated strand (Strand) and information about the nucleotide change at DNA, RNA and protein levels (Region). When available, the links to the RefSeq and UniProt databases are provided. The output file summarizes the prediction and annotation information in a VCF-like format. The same file includes in the bottom part information about errors and warnings occurring during the prediction process.

At the top of the page, a second web interface, accessible through the *Job* link (http://snps.biofold.org/phd-snpg/find-job.html), allows retrieving the output stored on the PhD-SNP^g server for ∼1 day. The prediction output is accessible using the *JobID* provided at the beginning of the output page.

CONCLUSIONS

The PhD-SNP^g web server provides an intuitive interface for predicting the impact of genomic variations in coding and noncoding regions. In terms of performance, although our tool is able to return predictions for all the regions of the genome for which a PhyloP score is available, the performance of the method could be affected by the amount of data used in the training step. In this regard, some classes of variants are underrepresented in the dataset derived from Clinvar. In particular, the pathogenic variants in noncoding regions represent ∼3% of the SNVs and InDels datasets and benign InDels in coding regions are ∼2%. However, PhD-SNP^g reported performance is not significantly affected by the different distributions of the gene variants (Supplementary Tables S14 and S15), indicating that the method is very robust for the coding regions.

It is also worth noting that this paper considers all noncoding sequence types a unique group. Some are more represented in our data (introns) than others (noncoding RNAs, UTR, etc.); thus, our results, and in turn, the performance, may be significantly different from those reported here if a user tests on an InDels set of different composition in noncoding regions.

From the technical standpoint, the updated PhD-SNP^g version is trained on a more extensive set of variants and can also deal with InDel variations. Furthermore, now PhD-SNP^g extends the evolutionary information of the input conservation scores from 100 to 470 species (from PhyloP10/PhyloP100 to PhyloP100/PhyloP470).

Nonetheless, the standalone version of PhD-SNP^g can be easily installed and executed on standard laptop machines. Specifically, it can run on an Intel Xeon 2.40 GHz machine with 8 GB of RAM, and can predict the effect of 1000 SNVs in less than 2 min. The runtime may increase when running the program in web mode, depending on network speed. Despite its simple input features, PhD-SNP^g performs comparably to state-of-the-art methods that require more demanding information and resources. PhD-SNP^g is a reliable and lightweight tool for assessing the effect of novel variants. It can be a fundamental benchmark for comparing predictors that rely on more intricate input features.

AVAILABILITY AND REQUIREMENTS

The PhD-SNP^g server is freely available on the Internet at http://snps.biofold.org/phd-snpg. The web interface and the PhD-SNP^g scripts are written in Python. The PhD-SNP^g standalone tool is stored on GitHub (https://github.com/biofold/PhD-SNPg), and can be installed by running a python2.7 script that automatically downloads the programs and data and programs (*twoBitToFa*, *liftOver* and *bigWigToBedGraph*) from the UCSC repository, with few library dependencies including *scikit-learn* package.

DATA AVAILABILITY

PhD-SNP^g is accessible at http://snps.biofold.org/phd-snpg.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors acknowledge financial support of the Rare Disease implementation study of the *European Network for Life Science ELIXIR* (IT-UNIBO-2022-Strategic 202201217). We thank PNRR M4C2 HPC - 1.4 'CENTRI NAZIONALI' - Spoke 8.

FUNDING

PRIN project, 'Integrative tools for defining the molecular basis of the diseases: Computational and Experimental methods for Protein Variant Interpretation' of the Italian Ministry for Education, University and Research [PRIN201744NR8S]. Funding for open access charge: Italian Ministry for Education, University and Research (PRIN201744NR8S)

Conflict of interest statement. None declared.

REFERENCES

1. Lappalainen,T., Scott,A.J., Brandt,M. and Hall,I.M. (2019) Genomic analysis in the age of Human genome sequencing. *Cell*, **177**, 70–84.
2. Capriotti,E., Nehrt,N.L., Kann,M.G. and Bromberg,Y. (2012) Bioinformatics for personal genome interpretation. *Brief. Bioinform.*, **13**, 495–512.
3. Durbin,R.M., Abecasis,G.R., Altshuler,D.L., Auton,A., Brooks,L.D., Gibbs,R.A., Hurles,M.E. and McVean,G.A. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

4. 1000 Genomes Project Consortium, Auton,A., Brooks,L.D., Durbin,R.M., Garrison,E.P., Kang,H.M., Korbel,J.O., Marchini,J.L., McCarthy,S., McVean,G.A. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
5. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020) Pan-cancer analysis of whole genomes. *Nature*, **578**, 82–93 .
6. 100,000 Genomes Project Pilot Investigators, Smedley,D., Smith,K.R., Martin,A., Thomas,E.A., McDonagh,E.M., Cipriani,V., Ellingford,J.M., Arno,G., Tucci,A. *et al.* (2021) 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. *N Engl. J. Med.*, **385**, 1868–1880.
7. Fernald,G.H., Capriotti,E., Daneshjou,R., Karczewski,K.J. and Altman,R.B. (2011) Bioinformatics challenges for personalized medicine. *Bioinformatics*, **27**, 1741–1748.
8. MacArthur,D.G., Manolio,T.A., Dimmock,D.P., Rehm,H.L., Shendure,J., Abecasis,G.R., Adams,D.R., Altman,R.B., Antonarakis,S.E., Ashley,E.A. *et al.* (2014) Guidelines for investigating causality of sequence variants in human disease. *Nature*, **508**, 469–476.
9. Niroula,A. and Vihinen,M. (2016) Variation interpretation predictors: principles, types, performance, and choice. *Hum. Mutat.*, **37**, 579–597.
10. Capriotti,E., Ozturk,K. and Carter,H. (2019) Integrating molecular networks with genetic variant interpretation for precision medicine. *Wiley Interdiscip. Rev. Syst. Biol. Med.*, **11**, e1443.
11. Rentzsch,P., Witten,D., Cooper,G.M., Shendure,J. and Kircher,M. (2019) CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.*, **47**, D886–D894.
12. Shihab,H.A., Rogers,M.F., Gough,J., Mort,M., Cooper,D.N., Day,I.N., Gaunt,T.R. and Campbell,C. (2015) An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics*, **31**, 1536–1543.
13. Ferlaino,M., Rogers,M.F., Shihab,H.A., Mort,M., Cooper,D.N., Gaunt,T.R. and Campbell,C. (2017) An integrative approach to predicting the functional effects of small indels in non-coding regions of the human genome. *BMC Bioinformatics*, **18**, 442.
14. Capriotti,E. and Fariselli,P. (2017) PhD-SNPg: a webserver and lightweight tool for scoring single nucleotide variants. *Nucleic Acids Res.*, **45**, W247–W252.
15. Ritchie,G.R.S., Dunham,I., Zeggini,E. and Flicek,P. (2014) Functional annotation of noncoding sequence variants. *Nat. Methods*, **11**, 294–296.
16. Grimm,D.G., Azencott,C., Aicheler,F., Gieraths,U., MacArthur,D.G., Samocha,K.E., Cooper,D.N., Stenson,P.D., Daly,M.J., Smoller,J.W. *et al.* (2015) The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum. Mutat.*, **36**, 513–523.
17. Pedregosa,F., Varoquaux,G., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Prettenhofer,P., Weiss,R., Dubourg,V. *et al.* (2011) Scikit-learn: machine learning in Python. *JMLR*, **12**, 2825–2830.
18. Landrum,M.J., Chitipiralla,S., Brown,G.R., Chen,C., Gu,B., Hart,J., Hoffman,D., Jang,W., Kaur,K., Liu,C. *et al.* (2020) ClinVar: improvements to accessing data. *Nucleic Acids Res.*, **48**, D835–D844.
19. Pollard,K.S., Hubisz,M.J., Rosenbloom,K.R. and Siepel,A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, **20**, 110–121.
20. Benevenuta,S., Capriotti,E. and Fariselli,P. (2021) Calibrating variant-scoring methods for clinical decision making. *Bioinformatics*, **36**, 5709–5711.
21. Brier,G.W. (1950) Verification of forecasts expressed in terms of probability. *Monthly Weather Rev.*, **78**, 1–3.
22. Yang,H. and Wang,K. (2015) Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat. Protoc.*, **10**, 1556–1566.
23. Kent,W.J., Zweig,A.S., Barber,G., Hinrichs,A.S. and Karolchik,D. (2010) BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, **26**, 2204–2207.
24. Zhou,W., Chen,T., Chong,Z., Rohrdanz,M.A., Melott,J.M., Wakefield,C., Zeng,J., Weinstein,J.N., Meric-Bernstam,F., Mills,G.B. *et al.* (2015) TransVar: a multilevel variant annotator for precision genomics. *Nat. Methods*, **12**, 1002–1003.
25. Tweedie,S., Braschi,B., Gray,K., Jones,T.E.M., Seal,R.L., Yates,B. and Bruford,E.A. (2021) Genenames.Org: the HGNC and VGNC resources in 2021. *Nucleic Acids Res.*, **49**, D939–D946.
26. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
27. UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.