Human Genetics (2022) 141:1649–1658
https://doi.org/10.1007/s00439-021-02419-4

ORIGINAL ARTICLE

Evaluating the relevance of sequence conservation in the prediction of pathogenic missense variants

Emidio Capriotti[1] · Piero Fariselli[2]

Received: 6 September 2021 / Accepted: 12 December 2021 / Published online: 31 January 2022

Abstract

Evolutionary information is the primary tool for detecting functional conservation in nucleic acid and protein. This information has been extensively used to predict structure, interactions and functions in macromolecules. Pathogenicity prediction models rely on multiple sequence alignment information at different levels. However, most accurate genome-wide variant deleteriousness ranking algorithms consider different features to assess the impact of variants. Here, we analyze three different ways of extracting evolutionary information from sequence alignments in the context of pathogenicity predictions at DNA and protein levels. We showed that protein sequence-based information is slightly more informative in the annotation of Clinvar missense variants than those obtained at the DNA level. Furthermore, to achieve the performance of state-of-the-art methods, such as CADD and REVEL, the conservation of reference and variant, encoded as frequencies of reference/alternate alleles or wild-type/mutant residues, should be included. Our results on a large set of missense variants show that a basic method based on three input features derived from the protein sequence profile performs similarly to the CADD algorithm which uses hundreds of genomic features. As expected, our method results in ~3% lower area under the receiver-operating characteristic curve (AUC). When compared with an ensemble-based algorithm (REVEL). Nevertheless, the combination of predictions of multiple methods can help to identify more reliable predictions. These observations indicate that for missense variants, evolutionary information, when properly encoded, plays the primary role in ranking pathogenicity.

Introduction

High-throughput sequencing technologies have changed our daily research by rapidly accumulating genomic data and helping to profile patient genomes (MacArthur et al. 2014; Claussnitzer et al. 2020). These studies make variant interpretation a fundamental challenge in precision medicine (Fernald et al. 2011; Capriotti et al. 2012; McInnes et al. 2021). Missense variants by changing a single amino acid in a protein sequence can be neutral or induce loss of function.

In the last 2 decades, several methods have been developed to prioritize functional missense variants relying on protein sequence/structure information (Tennessen et al. 2012; Niroula and Vihinen 2016; Ancien et al. 2018; Petrosino et al. 2021) and the protein interaction networks (Rost et al. 2016; Capriotti et al. 2019; Ozturk and Carter 2021).

It is widely accepted that evolutionary information encoded in multiple sequence alignments of DNAs and proteins is a major resource for scoring variant pathogenicity. Several methods for scoring the nucleotide and amino acid conservation have been defined (Schneider 1997; Valdar 2002). Although there is no rigorous test for judging a conservation measure, in general, quantitative conservation measures are site-specific scores calculated from a vector representing the relative frequency of the amino acids or nucleotides in a given position of a multiple sequence alignment. Among the most commonly used scores are those calculating the Euclidean distance between two sets of amino acid frequencies (Valdar 2002) and Shannon's information-theoretic entropy (Capra and Singh 2007). Such site-specific scores, which are important for identifying functionally conserved regions, do not explicitly depend on the pairs of wild-type and mutant nucleotides/amino acids observed in a specific mutation process. For scoring the pathogenicity of

Emidio Capriotti
emidio.capriotti@unibo.it

Piero Fariselli
piero.fariselli@unito.it

[1] BioFolD Unit, Department of Pharmacy and Biotechnology (FaBiT), University of Bologna, Via F. Selmi 3, 40126 Bologna, Italy

[2] Department of Medical Sciences, University of Torino, Via Santena 19, 10126 Turin, Italy

a specific variation, we considered the most basic quantitative measures calculating the frequency of the wild-type and mutant nucleotides/residues in a given site.

This paper evaluates the relevance of this information for missense variant predictions by comparing simple scores and simple predictors with the widely used and well-performing Combined Annotation-Dependent Depletion (CADD) algorithm (Rentzsch et al. 2019) and REVEL (Ioannidis et al. 2016).

We computed the conservation scores on DNA (*PhastCons100way* and *PhyloP100way*), the frequencies of the reference and alternative alleles in the genome, and frequencies of the wild-type and mutant residues in protein multiple alignments. Our analysis showed that a machine-learning method trained on a few sequence conservation features at DNA or protein levels, achieves similar performance of a state-of-the-art algorithm. In this work, we compared the performance of CADD and REVEL with those reached by three different basic gradient boosting algorithms on a set of missense variants from the *Clinvar* database. Our result indicates that the evolutionary information provides the main features for scoring the pathogenicity of missense variants.

Materials and methods

Datasets

To evaluate the performance of different machine-learning methods for predicting the pathogenicity of missense variants, we collected two datasets from the *Clinvar* database (Landrum et al. 2020). For building the two datasets, we considered two versions of *Clinvar* released in June 2020 and August 2021, respectively. The first dataset (*CommonClinvar*) consists of the missense variants annotated as *Pathogenic* and *Benign* in both versions of the database while the second dataset (*NewClinvar*) collects the new missense variants reported in the last version of *Clinvar* since June 2020 (Fig. S1). The variants reported in the older version of *Clinvar* not confirmed in the last version were discarded. Thus, the *CommonClinvar* consists of 36,751 missense variants from 7582 proteins 53.5% of which are annotated as *Benign* and the remaining ones (46.5%) as *Pathogenic*. *NewClinvar*, which includes only the newly annotated variants, is composed of 5172 from 1855 proteins 43.4% of which are reported as *Benign* and 56.6% as *Pathogenic*. The composition of the two datasets is summarized in Table S1. Both *CommonClinvar* and *NewClinvar* datasets are available as supplementary files.

Conservation features

In this work, we analyzed the performance in the prediction of pathogenic variants using three basic methods based on sequence conservation features. Each method considers only three input features, which are described as follows.

As a baseline, we implemented a method considering two site-specific conservation scores (*PPScore*) calculated on a genome level multiple sequence alignment and made available through the UCSC genome browser (Kent et al. 2002). The conservation scores used in the first method are calculated by *PhastCons* (Siepel et al. 2005) and *PhyloP* (Pollard et al. 2010) algorithms.

In the second method (*DNAProf*), the frequencies of the reference and alternative alleles in the mutated site are calculated for each variant from the *multiz100way* multiple sequence alignments. The *PhastCons100* and *PhyloP100way* scores as well as the *multiz100way* alignments for the hg38 human reference genome are available at https://hgdownload.cse.ucsc.edu/goldenpath/hg38/.

In the third method (*ProtProf*), we calculated the frequencies of the wild-type and mutant residues in the mutated sites for each mutation. These frequencies are derived from the *BLAST* (Altschul et al. 1997) search alignments against the *UniRef90* database (Suzek et al. 2007) released in June 2020. For the *BLAST* search, we used an *e*-value cutoff of 10^{-9} as suggested in previous works (Capriotti et al. 2006, 2013, 2017; Calabrese et al. 2009). By definition, the *E*-value represents the number of expected hits found by chance and depends on the number of sequences in the database. In our case, selecting an E-value threshold of 10^{-9} for a BLAST search on UniRef90, which contains ~ 10^{-8} sequences, ensures that less than one random hit can be found by chance from our search.

In summary, for each mutated loci, we considered the *PhastCons100way* (*PC*) and *PhyloP100way* (*PP*), the frequencies of the reference (f_{ref}) and alternative (f_{alt}) alleles from the *multiz100way* multiple sequence alignment, and the total number of aligned genomic sequences (N_{g}). For the protein-based method, the protein sequence profile was obtained considering the sequences returned by *BLAST*. For each mutated site, we calculated the frequencies of wild-type (f_{wt}) and mutant (f_{mut}) residues and the number of aligned proteins (N_{p}). The nucleotide and amino acid frequencies are calculated as follows:

$$f(x_i) = \frac{n(x_i)}{\sum_{i=1}^{i=k} n(x_i)},$$

where $n(x_i)$ is the number of the nucleotide or amino acid x_i in the sequence alignment and k is equal to 5 (including the generic nucleotide *N*) and 20 for DNA and protein sequences, respectively. N_{g} and N_{p} represent the denominators of the equation above for DNA and protein sequences, respectively.

Machine-learning algorithms

Using the eight features described above, we develop three binary classifiers (*PPScore, DNAProf, ProtProf*) using the following groups of three features:

PPScore PhastCons100way (PC) and *PhyloP100way* (PP) scores, and number aligned genomic sequences (N_g) in *multiz100way*
DNAProf Frequencies of the reference (f_{ref}) and alternative (f_{alt}) alleles, and number aligned genomic sequences in *multiz100way* (N_g).
ProtProf Frequencies of the wild-type (f_{wt}) and mutant (f_{mut}) residues, and number aligned protein sequences (N_p) from a *BLAST* search on *UniRef90.*

For each group of features defined above, we developed a binary classifier based on the gradient boosting algorithm as implemented in the *scikit-learn* package (Pedregosa et al. 2011). The proposed groups of features are summarized in Table 1.

Training and testing procedure

We first evaluated the performance of each method on *CommonClinvar* using a tenfold cross-validation procedure for a fair evaluation of the proposed method performance. To reduce at the minimum the possible overfitting, we mapped each missense variant on the relative protein sequence and we clustered all the sequences using the *blastclust* algorithm (https://ftp.ncbi.nih.gov/blast/documents/blastclust.html) with a sequence identity threshold of 25% and a coverage of 50%. Using the clustering based on sequence similarity, we perform a tenfold cross-validation procedure keeping all the variants belonging to the same cluster in the same subset. A second test is performed considering the *NewClinvar* dataset. In this case, the impact of the variants of a given protein is predicted excluding from the training set (*CommonClinvar*) all the variants belonging to proteins of the same cluster. We extracted a balanced set of *Pathogenic* and *Benign* variants from *CommonClinvar* and *NewClinvar* dataset for each test, randomly downscaling the most abundant class. The reported scoring measures for all the methods are averaged over ten randomly selected sets.

Table 1 Three groups of features used for the development of the binary classifiers

Group	Features		
PPScore	PC	PP	N_g
DNAProf	f_{ref}	f_{alt}	N_g
ProtProf	f_{wt}	f_{mut}	N_p

Benchmarking and performance measures

To characterize the prediction power of the main features described in this work, for each of them, we developed a single feature binary classifier based on a single threshold. For each feature, the classification threshold is optimized on the *CommonClinvar* dataset maximizing both the true-positive and true-negative rates. The optimized threshold is tested in the classification of the *NewClinvar* variant dataset.

Finally, the performances of all the binary classifiers described above are compared with those achieved by the CADD (Rentzsch et al. 2019) and REVEL (Ioannidis et al. 2016) algorithms. The optimized raw score threshold for the classification of CADD output was calculated on the *CommonClinvar* dataset as binary classifier. The performances of the methods are scored considering two subsets of the *NewClinvar* dataset including the consensus predictions (*Consensus*) and those with at least one predictor in disagreement with the remaining ones (*NotConsensus*).

All the measures considered for scoring the performance of the methods are defined in Supplementary Materials.

Results

Feature analysis and single feature classification

In the first part of our work, we analyzed the distributions of the main features used for the classification task. We focused on the six conservation features (PC, PP, f_{ref}, f_{alt}, f_{wt}, f_{mut}) comparing their distributions for the subsets of *Pathogenic* and *Benign* variants. The average, median and standard deviation of such distributions are reported in Table S2. As observed in previous works (Kircher et al. 2014; Capriotti and Fariselli 2017), the distribution of the *PhyloP100way* score (*PP*) in mutated loci associated with *Pathogenic* and *Benign* variants are significantly different (Fig. 1). Indeed, the two distributions show median values of 7.5 and 1.5, respectively, with a Kolmogorov–Smirnov distance (D_{KS}) of 0.57 (Fig. 1B and Table S2). This distance is greater than the D_{KS} observed for the *PhastCons100way* score (PC).

A higher difference between the distributions of the conservation scores for the subset of *Pathogenic* and *Benign* variants is observed when the frequencies in sequence profile from genomic and proteins are considered. The most remarkable differences are generally detected when comparing the distributions of the frequency of the alternative allele (f_{alt}) and the mutant residue (f_{mut}) for which the D_{KS} is ~ 0.60. Analyzing the frequencies of the reference allele (f_{ref}) and wild-type residue (f_{wt}) their D_{KS} is 0.58 and 0.55,

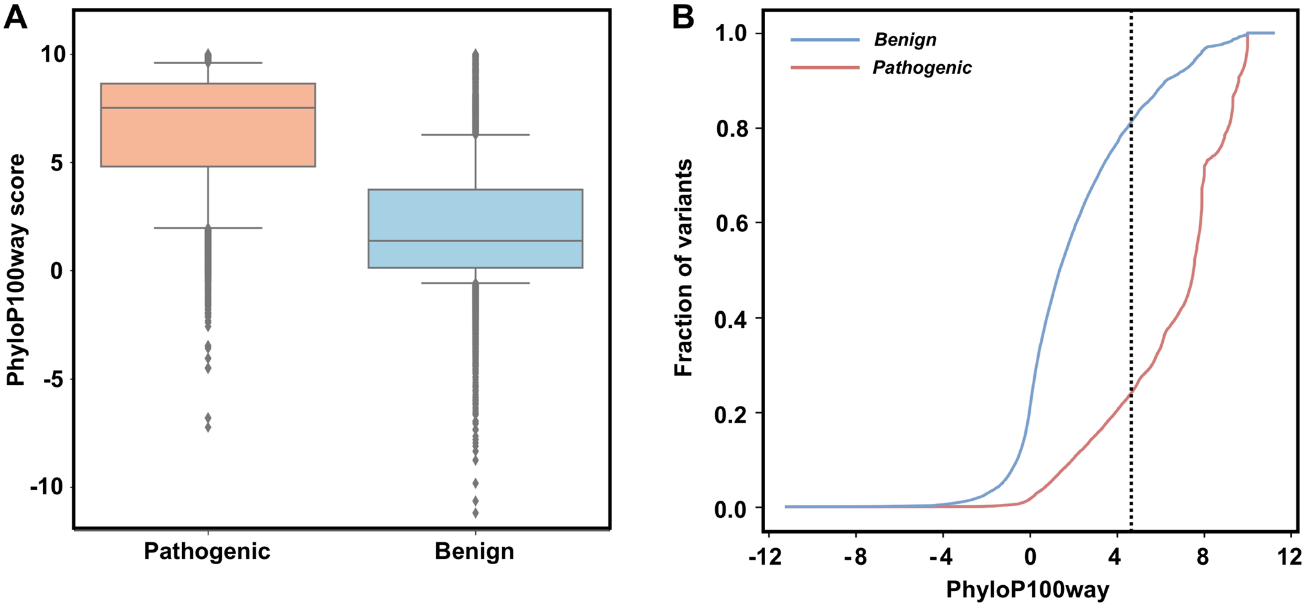


Fig. 1 (**A**) Box plot and (**B**) cumulative distributions of the *PhyloP100way* score in the variation sites for the subsets of *Pathogenic* and *Benign* variants in the *CommonClinvar* dataset. The maximum distance between the two distributions is at 4.7 that corresponds to a Kolmogorov–Smirnov distance of 0.57

respectively (Table S2). The distributions of the four types of frequencies (f_{ref}, f_{alt},, f_{wt}, and f_{mut}) for the subsets of *Pathogenic* and *Benign* variants are plotted in Fig. 2

This observation agrees with the results obtained in the prediction of *Pathogenic* variants using a classification threshold on a single feature. The classification threshold is optimized on the *CommonClinvar* dataset maximizing both the true-positive and -negative rates (Table S3). Applying the optimized thresholds on the prediction of the variants in the *NewClinvar* dataset, we found that a simple classifier

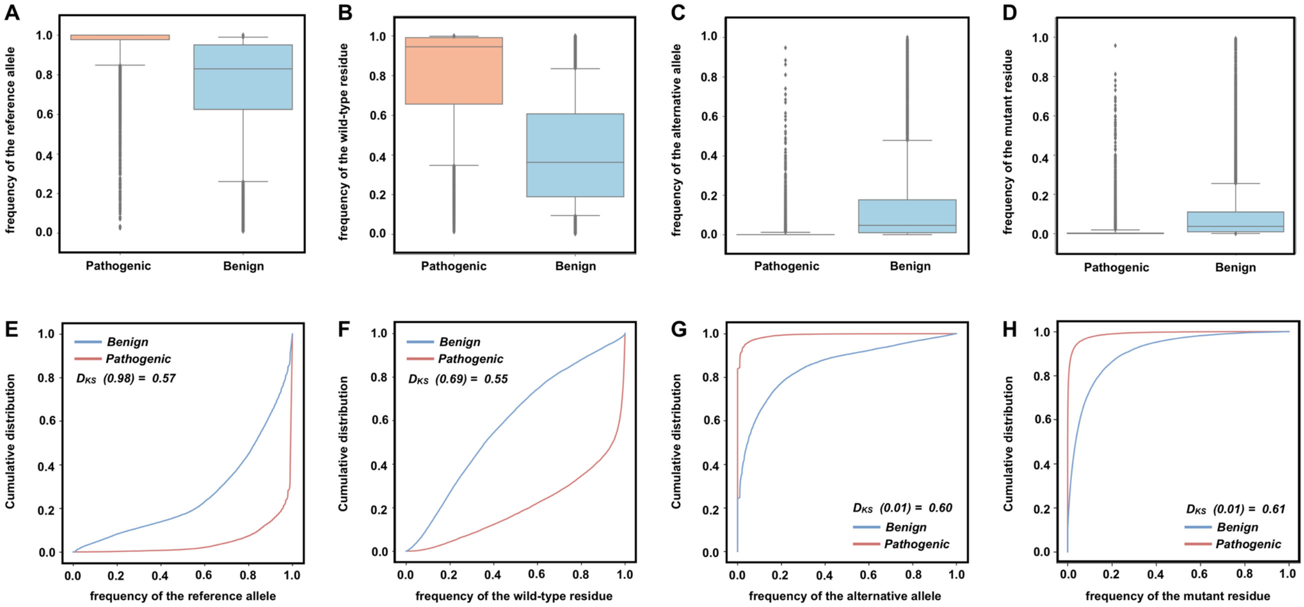


Fig. 2 Box plots and cumulative distributions of the frequencies of references (**A**, **E**)/alternative (**C**, **F**) alleles in the variation sites and the frequencies of wild-type (**B**, **G**)/mutant (**D**, **H**) residues in protein mutation sites for the subset of *Pathogenic* and *Benign* variants in the *CommonClinvar* dataset. In the cumulative distribution plot (**E**, **F**, **G**, **H**), the Kolmogorov–Smirnov distance (D_{KS}) which represents the maximum distance between the distributions of the frequencies for *Pathogenic* and *Benign* variants is reported

based on the frequency of the mutant residue extracted from a protein sequence profile achieve 81% overall accuracy (Q2), 0.63 Matthews correlation coefficient (MCC) and an Area Under the Receiver-Operating Characteristic Curve (AUC) of 0.86 (Table 2).

According to the previous observation, the *PhastCons100way* score (PC) is the least discriminating feature. When using the optimized threshold on the classification of the *NewClinvar* variants, the method based on a single *PC* threshold achieves 74% overall accuracy, 0.49 MCC and 0.75 AUC (Table 2). Slightly lower performances are obtained when the frequencies of the reference allele and the wild-type residue in the sequence profile are considered. In this case, the method based on a single f_{ref} threshold results in 78% Q2, 0.56 MCC and 0.84 AUC. These results can also be observed plotting the Receiving-Operating Characteristic (ROC) and Precision-Recall (PR) curves reported in Fig. S2.

Assessment of the machine-learning methods

Starting from the previous observations, we developed three machine-learning approaches based on the different groups of conservation features. The *PPScore* method is based on the *PhastCons100way*, *PhyloP100way* scores representing unique conservation measures not describing the type of nucleotides observed in the mutated loci. The other two methods consider the frequencies of the nucleotides or residues in the original and new sequences that correspond to f_{ref}, f_{alt} and f_{wt}, f_{mut} for *DNAProf* and *ProtProf*, respectively. To these groups of measures, we added a third feature representing the total number of sequences aligned in the mutated loci (N_{g}, N_{p}). Although these values are not related to the conservation, they are considered as features for differentiating cases of mutated loci aligned with low and high number of sequences. We implemented three machine-learning methods for predicting *Pathogenic* variants based on the gradient boosting algorithm with these groups of features. First, the performance of these methods is tested with a tenfold cross-validation procedure on the *CommonClinvar* dataset. To avoid possible overfitting, we clustered all the proteins based on the sequence identity and grouped all their variants in a unique subset. The average performance of *PPScore*, *DNAProf* and *ProtProf* on a balanced set of *Pathogenic* and *Benign* variants is reported in Table S4. The results show that among the three methods *ProtProf*, which is based on protein sequence profile, achieved the highest performance reaching 83% overall accuracy (Q2), 0.67 Matthews correlation coefficient and 0.91 Area Under the Receiver-Operating Characteristic Curve (AUC). *PPScore* which is based on *PhastCons100way*, *PhyloP100way* show the lowest performance resulting in ~4% lower AUC and ~9% lower MCC. An intermediate level of performance is achieved by *DNAProf* which results in ~2% lower AUC and ~3% lower MCC with respect to *ProtProf*. Similar results are obtained when assessing the performance of the three methods on the *NewClinvar* dataset. In addition, in this case, we predicted the impact of each variant removing from the training set all the variants in the *CommonClinvar* training set belonging to the same cluster of proteins. The performance of *PPScore*, *DNAProf* and *ProtProf* on a balanced set of variants from the *NewClinvar* dataset is summarized in Table 3. For scoring the contribution of N_{g} and N_{p} to the predictions of the three methods, we removed such features from the input and compared their performances. The results reported in Table S5 show that the methods improve their performance by ~1% for MCC and AUC indices considering *N*g and *N*p in the input features.

Comparison with CADD and REVEL algorithms

In the final part of our analysis, we compared the performance of our simple gradient boosting-based algorithms with those obtained with CADD (Rentzsch et al. 2019) and REVEL (Ioannidis et al. 2016). REVEL (Rare Exome Variant Ensemble Learner) is an ensemble method for predicting the pathogenicity of missense variants on the basis of 13 individual tools. When tested on independent test sets, REVEL shows the best overall performance as compared to any of the individual tools and 7 previously developed ensemble methods. CADD is one of the most accurate and

Table 2 Performance of basic predictors based on a single feature on the *NewClinvar* dataset Prediction threshold are optimized on the *CommonClinvar* dataset

Feature	Threshold	Q2	TNR	NPV	TPR	PPV	MCC	*F1*	AUC	AUP
PC	1.000	0.737	0.611	0.816	0.862	0.689	0.489	0.766	0.755	0.815
PP	4.704	0.769	0.796	0.756	0.743	0.784	0.539	0.763	0.841	0.828
f_{ref}	0.977	0.779	0.815	0.760	0.742	0.801	0.559	0.770	0.836	0.843
f_{alt}	0.000	0.794	0.750	0.821	0.837	0.770	0.589	0.802	0.828	0.863
f_{wt}	0.702	0.769	0.806	0.750	0.731	0.791	0.539	0.759	0.844	0.836
f_{mut}	0.005	0.815	0.819	0.812	0.810	0.817	0.629	0.814	0.857	0.856

Q2 overall accuracy, *TNR* true-negative rate, *NPV* negative predicted value, *TPR* true-positive rate, *PPV* positive predicted value, *MCC* Matthews Correlation Coefficient, *F1* harmonic mean of precision and sensitivity, *AUC* area under the receiver operator characteristic curve, *AUP* area under the precision recall curve. All the performance measures are defined in Supplementary Materials

Table 3 Testing prediction on the *NewClinvar* variant dataset

Method	Q2	TNR	NPV	TPR	PPV	MCC	*F*1	AUC	AUP
CADD	0.844	0.821	0.860	0.867	0.829	0.688	0.847	0.911	0.905
REVEL	0.871	0.918	0.843	0.825	0.909	0.747	0.865	0.945	0.942
ProtProf	0.831	0.865	0.809	0.796	0.855	0.662	0.824	0.910	0.905
DNAProf	0.812	0.780	0.834	0.845	0.794	0.626	0.818	0.881	0.873
PPScore	0.771	0.776	0.769	0.767	0.774	0.543	0.770	0.855	0.846

All the performance measures are defined in Supplementary Materials. For CADD, a raw score classification threshold of 3.1 was considered

Q2 overall accuracy, *TNR* true-negative rate, *NPV* negative predicted value, *TPR* true-positive rate, *PPV* positive predicted value, *MCC* Matthews Correlation Coefficient, *F1* harmonic mean of precision and sensitivity, *AUC* area under the receiver operator characteristic curve, *AUP* area under the precision recall curve. All the performance measures are defined in Supplementary Materials

popular methods for predicting *Pathogenic* variants in coding and non-coding regions (Benevenuta et al. 2021). This method, which is based on more than hundreds of genomic features, was trained on more than 30 million variants. To use CADD as a binary classifier, we considered the raw output of the program and we selected the threshold that maximizes the true-positive and -negative rates on the *CommonClinvar* dataset. The performance of CADD at the optimal raw score classification threshold of 3.1 is reported in Table S4. This threshold was used for the classification of the variants in the *NewClinvar* dataset. The performances of CADD and REVEL on the *NewClinvar* dataset are summarized in Table 3. This analysis shows that CADD and *ProtProf* algorithms result in a similar performance in the classification of *Pathogenic* missense variants in terms of Area Under the Receiver-Operating Characteristic (AUC) and Precision-Recall (AUP) curves on both *CommonClinvar* and *NewClinvar* datasets. As expected, REVEL outperforms *ProtProf* and CADD reaching ~ 3% higher overall accuracy (Q2) and AUC. We can also observe that *DNAProf* which is based on the sequence profile extracted from the *multiz100way* sequence alignments results only in ~ 3% lower AUC and AUP. The Receiver-Operating Characteristic and Precision-Recall curves for CADD, REVEL and the three methods presented in this manuscript are plotted in Fig. 3.

Analysis and comparison of the predictions

To better understand the difference among the presented methods, we compared the predictions of three gradient boosting-based algorithms (*PPScore, DNAProf and ProtProf)*. For performing this analysis, we defined two subsets of the *NewClinvar* dataset: the *Consensus* subset where the three predictions agree and the *NotConsensus* subset where the predictions differ. When performing such comparison, we observe that predictions overlap in ~ 73% of the cases in

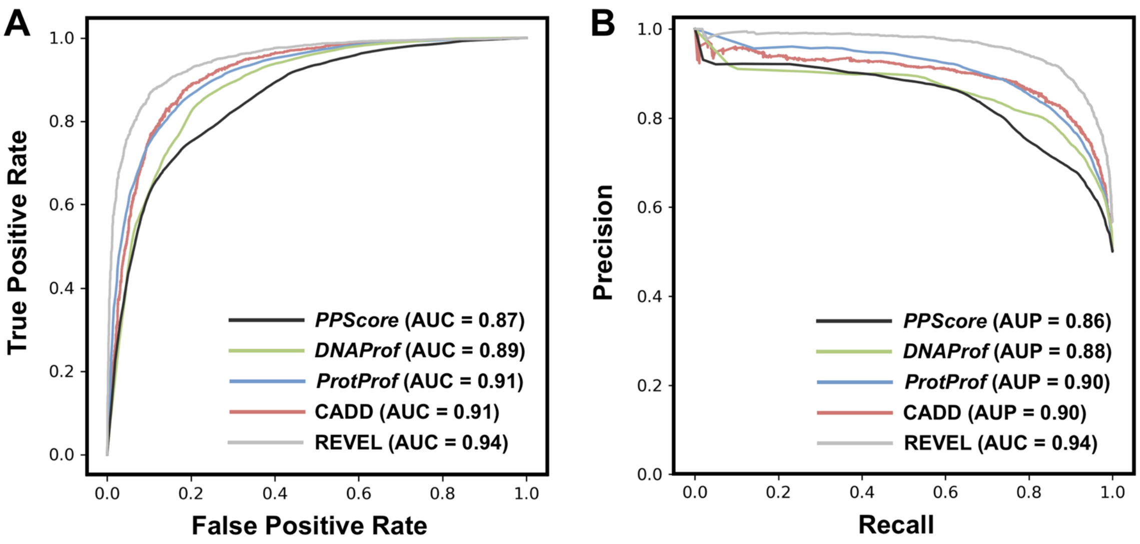
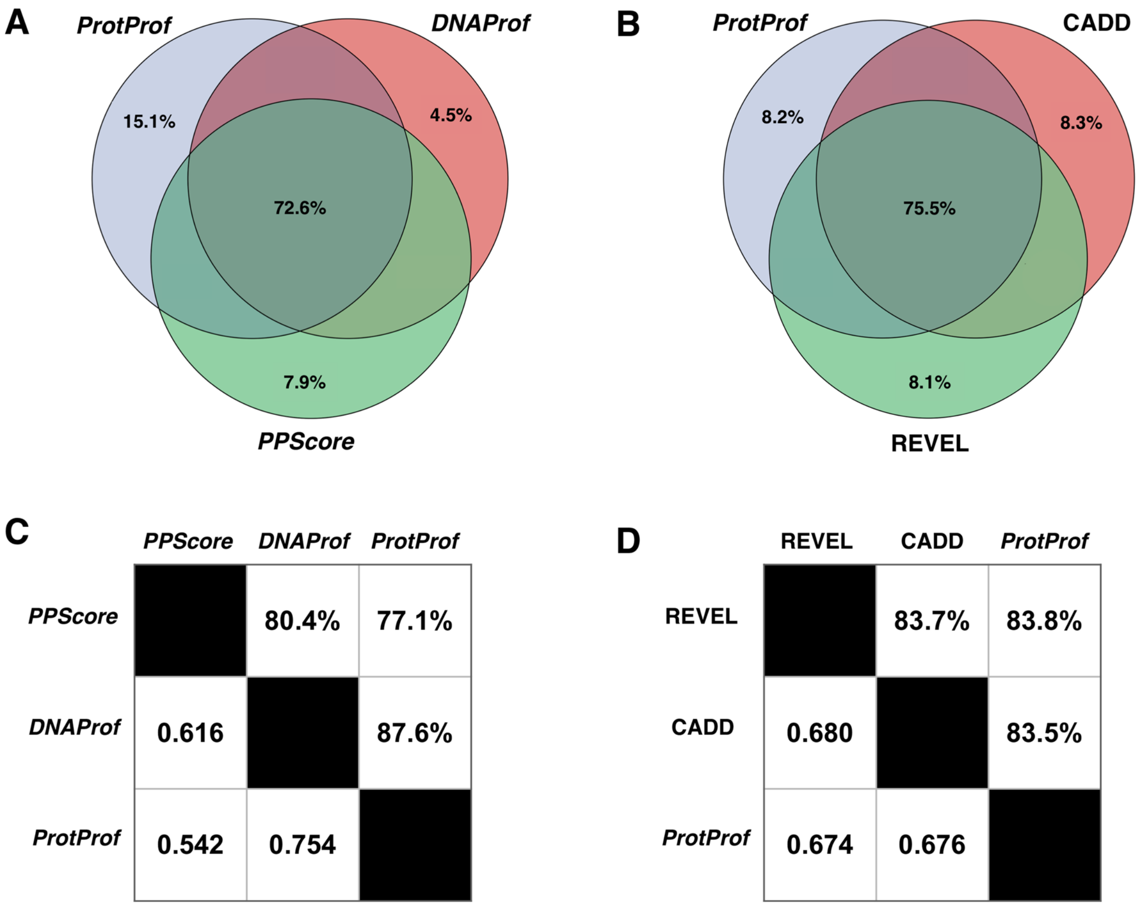


Fig. 3 Receiver-Operating Characteristic (**A**) and Precision-Recall (**B**) curves for the different gradient boosting algorithms (*PPScore, DNAProf, ProtProf*), CADD and REVEL on the *NewClinvar* dataset

the *NewClinvar* dataset (Fig. 4A), while in the remaining 23% a single predictor differs from the other.

ProtProf and *DNAProf* provide the highest prediction similarity, agreeing in 88% of the cases with a correlation of 0.75 (Fig. 4C). In terms of performance, when focusing on the *Consensus* subset, the performance of *ProtProf* reaches 0.91 AUC and 0.81 Matthews correlation, while on the remaining subset, all the methods show an AUC < 0.65 and *PPScore* results on AUC < 0.5 (Table 4). The decrease in the performance can be explained by comparing the distribution of the frequencies of the wild-type and mutant residues on the *Consensus* and *NotConsensus* subsets. Indeed, the



	PPScore	DNAProf	ProtProf
PPScore		80.4%	77.1%
DNAProf	0.616		87.6%
ProtProf	0.542	0.754	

	REVEL	CADD	ProtProf
REVEL		83.7%	83.8%
CADD	0.680		83.5%
ProtProf	0.674	0.676	

Fig. 4 Overlap of the predictions of the different methods on the *NewClinvar* dataset. **A** Venn diagram of the prediction of the 3 gradient boosting algorithms (*PPScore, DNAProf, ProtProf*) and **B** Venn diagram of the prediction of *ProtProf,* CADD and REVEL. **C** Pairwise overlap of the prediction of the 3 gradient boosting algorithms and **D** Pairwise overlap of the prediction of the *ProtProf,* CADD and REVEL. The numbers above the diagonal represent the fraction of common predictions while those below the diagonal are the correlations between the predictions

Table 4 Performance of *PPScore, DNAProf* and *ProtProf* on two subsets of the NewClinvar dataset

Method	Subset	Q2	TNR	NPV	TPR	PPV	MCC	*F*1	AUC	AUP
ProtProf	*Consensus*	0.905	0.907	0.909	0.904	0.901	0.810	0.902	0.905	0.929
ProtProf	*NotConsensus*	0.633	0.740	0.577	0.542	0.712	0.286	0.615	0.641	0.735
DNAProf	*NotConsensus*	0.567	0.402	0.536	0.706	0.583	0.113	0.639	0.554	0.665
PPScore	*NotConsensus*	0.418	0.385	0.369	0.445	0.461	− 0.170	0.453	0.415	0.561

Consensus Subset of *NewClinvar* (72.6%) for which the predictions of the three methods are in agreement, *NotConsensus* subset of *NewClinvar* (27.4%) for which one method differs from the remaining two

Q2 overall accuracy, *TNR* true-negative rate, *NPV* negative predicted value, *TPR* true-positive rate, *PPV* positive predicted value, *MCC* Matthews Correlation Coefficient, *F1* harmonic mean of precision and sensitivity, *AUC* area under the receiver operator characteristic curve, *AUP* area under the precision recall curve. All the performance measures are defined in Supplementary Materials

Kolmogorov–Smirnov distance (*DKS*) of the distributions of fwt and fmut decrease by ~ 50% when considering the *NotConsensus* subset (Fig. 5).

A similar analysis is performed comparing the predictions of *ProtProf*, CADD and REVEL. Here, the predictions of the three algorithms overlap in ~ 75% of the cases (Fig. 4B) with an average of ~ 83% of common predictions and a correlation of ~ 0.68 for pairwise comparison (Fig. 4D). The performance of the methods on the *Consensus* subsets achieves ~ 95% in terms of overall accuracy and ~ 0.95 AUC, while for the *NotConsensus* subset, the performance of CADD and *ProtProf* are similar to those of a random predictor. REVEL shows the best performance on this subset of the *NewClinvar* dataset, achieving 67% Q2 and 0.68 AUC (Table 5).

The analysis of the distributions of f_{wt} and f_{mut} on shows a $D_{\mathrm{KS}} > 0.70$ for the *Consensus* subset, while for the *NotConsensus* it drops below 0.2 with both the distributions of the fwt and fmut for *Benign* and *Pathogenic* variants strongly overlapping (Fig. 6).

Conclusion and discussion

Here, we analyzed different evolutionary information encodings for missense variant pathogenicity predictions. We compared the encoding at DNA and protein levels, where different multiple alignments techniques apply. The multiple sequence alignment includes a larger number of proteins and more remote homologs for many genes than

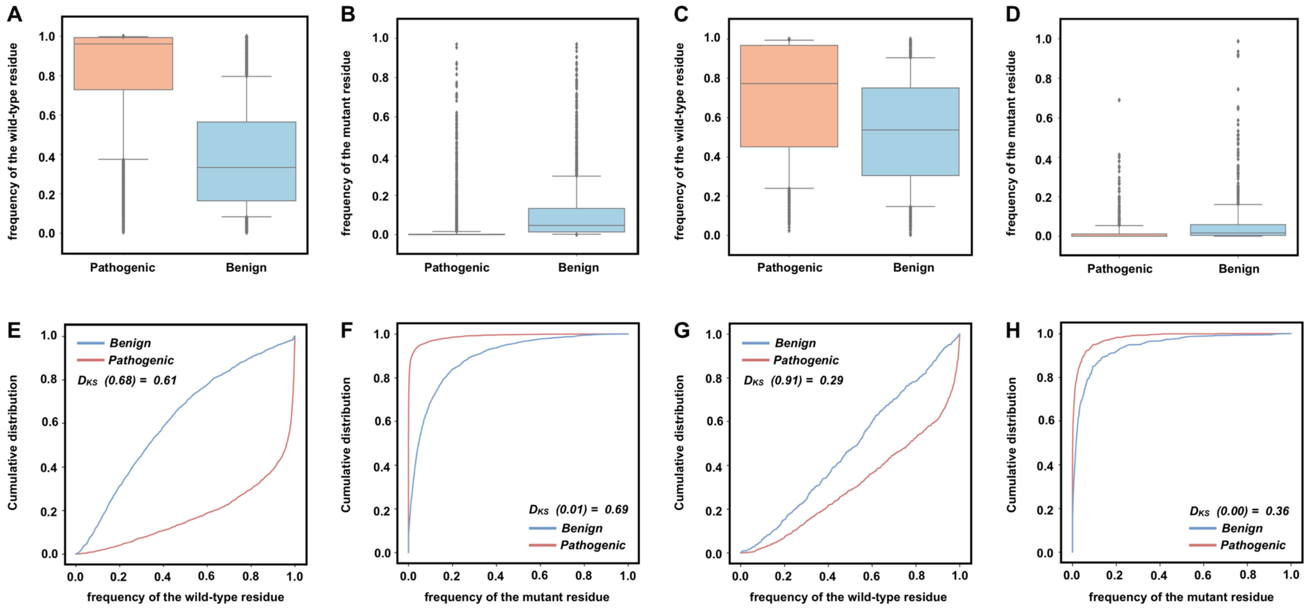


Fig. 5 Box plots and cumulative distributions of the frequencies of wild-type and mutant residues subset of *NewClinvar* for which the predictions of *PPScore, DNAProf* and *ProtProf* are in agreement (**A**, **B**, **E**, **F**) or in disagreement (**C**, **D**, **G**, **H**). In the cumulative distribution plot (**E**, **F**, **G**, **H**), the Kolmogorov–Smirnov distance (D_{KS}) which represents the maximum distance between the distributions of the frequencies for *Pathogenic* and *Benign* variants is reported. Average and standard deviations of the distributions are reported in Table S6

Table 5 Performance of REVEL, CADD and *ProtProf* on two subsets of the NewClinvar dataset

Method	Subset	Q2	TNR	NPV	TPR	PPV	MCC	*F*1	AUC	AUP
ProtProf	*Consensus*	0.945	0.954	0.941	0.935	0.950	0.890	0.943	0.945	0.959
ProtProf	*NotConsensus*	0.479	0.539	0.429	0.431	0.541	– 0.029	0.480	0.485	0.616
REVEL	*NotConsensus*	0.662	0.777	0.589	0.570	0.763	0.350	0.653	0.674	0.760
CADD	*NotConsensus*	0.531	0.337	0.458	0.684	0.565	0.022	0.619	0.510	0.628

Consensus: Subset of *NewClinvar* (75.5%) for which the predictions of the three methods are in agreement. *NotConsensus*: subset of *NewClinvar* (24.5%) for which one method differs from the remaining two

Q2 overall accuracy, *TNR* true-negative rate, *NPV* negative predicted value, *TPR* true-positive rate, *PPV* positive predicted value, *MCC* Matthews Correlation Coefficient, *F1* harmonic mean of precision and sensitivity, *AUC* area under the receiver operator characteristic curve, *AUP* area under the precision recall curve. All the performance measures are defined in Supplementary Materials

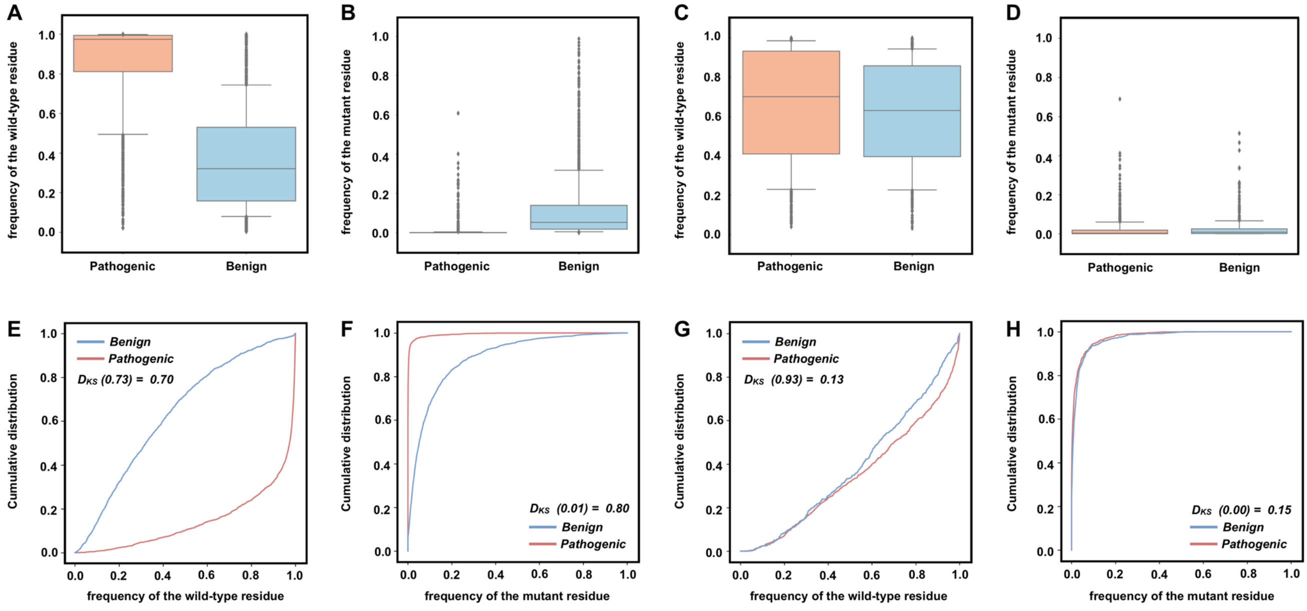


Fig. 6 Box plots and cumulative distributions of the frequencies of wild-type and mutant residues subset of *NewClinvar* for which the predictions of REVEL, CADD and *ProtProf* are in agreement (**A**, **B**, **E**, **F**) or in disagreement (**C**, **D**, **G**, **H**). In the cumulative distribution plot (**E**, **F**, **G**, **H**), the Kolmogorov–Smirnov distance (D_{KS}) which represents the maximum distance between the distributions of the frequencies for *Pathogenic* and *Benign* variants is reported. Average and standard deviations of the distributions are reported in Table S7

pre-calculated genome alignments from the UCSC genome browser. This condition can be the reason why the performance of a method trained using the protein-based information is slightly better. Thus, at least for the missense variants, an input based on evolutionary information of the wild-type and mutated residue performs better than evolutionary measures based on DNA sequence alignment in the prediction of pathogenic variants. Indeed, on the ~ 27% of the variants on which the predictions of our gradient boosting algorithms disagree, *ProtProf* reaches 6% high overall accuracy and 9% higher AUC with respect to *DNAProf* which rely on pre-calculated DNA sequence alignment from UCSC.

With these simple inputs based on evolutionary information, a machine-learning method can perform comparably to CADD, which uses more sophisticated inputs. When compared with an ensemble-based approach (RAVEL), our basic method (*ProtProf)* results in ~ 3% lower overall accuracy and AUC. Nevertheless, our analysis based on the comparison of the predictions of different methods allows the identification of a more reliable subset of predictions on which *ProtProf* reaches an overall accuracy above 90% and AUC > 0.9.

Recently, it has been suggested that protein positions have a significant role and can act as *Neutral*, *Toggle* or *Rheostat* (Miller et al. 2019). Here, we indicate an alternative view of protein positions that can be seen as a nonlinear combination of the frequencies of wild-type/mutant residues at protein level or reference/alternative allele at DNA level. The results of our analysis suggest that the performances of new and more sophisticated machine-learning algorithms should always be compared with those achieved by simple conservation-based methods. As recently proposed (Walsh et al. 2021), the design of such benchmark tests should consider the adoption of specific guidelines for avoiding bias in the training and testing sets. This procedure is important to exclude overfitting on the context-dependent features (Grimm et al. 2015) and identify new important features for improving the performance of variant scoring algorithms.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00439-021-02419-4.

Acknowledgements PF thanks the Italian Ministry for Education, University and Research for the programme “Dipartimenti di Eccellenza 20182022D15D18000410001”.

Author contributions EC developed and collected and analyzed the data. Both the authors designed the research and contributed to the writing of the manuscript.

Funding This work was supported by the PRIN project, “Integrative tools for defining the molecular basis of the diseases: Computational and Experimental methods for Protein Variant Interpretation” of the Ministero Istruzione, Università e Ricerca (PRIN201744NR8S).

Availability of data and material Additional information is available in Supplementary Materials. The *CommonClinvar* and *NewClinvar* datasets are provided as supplementary files and available at https://doi.org/10.5281/zenodo.5722031.

Declarations

Conflict of interest None.

References

Altschul SF, Madden TL, Schaffer AA et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Ancien F, Pucci F, Godfroid M, Rooman M (2018) Prediction and interpretation of deleterious coding variants in terms of protein structural stability. Sci Rep 8:4480. https://doi.org/10.1038/s41598-018-22531-2

Benevenuta S, Capriotti E, Fariselli P (2021) Calibrating variant-scoring methods for clinical decision making. Bioinformatics. https://doi.org/10.1093/bioinformatics/btaa943

Calabrese R, Capriotti E, Fariselli P et al (2009) Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum Mutat 30:1237–1244. https://doi.org/10.1002/humu.21047

Capra JA, Singh M (2007) Predicting functionally important residues from sequence conservation. Bioinformatics 23:1875–1882. https://doi.org/10.1093/bioinformatics/btm270

Capriotti E, Fariselli P (2017) PhD-SNPg: a webserver and lightweight tool for scoring single nucleotide variants. Nucleic Acids Res 45:W247–W252. https://doi.org/10.1093/nar/gkx369

Capriotti E, Calabrese R, Casadio R (2006) Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics 22:2729–2734. https://doi.org/10.1093/bioinformatics/btl423

Capriotti E, Nehrt NL, Kann MG, Bromberg Y (2012) Bioinformatics for personal genome interpretation. Brief Bioinform 13:495–512. https://doi.org/10.1093/bib/bbr070

Capriotti E, Calabrese R, Fariselli P et al (2013) WS-SNPs&GO: a web server for predicting the deleterious effect of human protein variants using functional annotation. BMC Genom 14(Suppl 3):S6. https://doi.org/10.1186/1471-2164-14-S3-S6

Capriotti E, Martelli PL, Fariselli P, Casadio R (2017) Blind prediction of deleterious amino acid variations with SNPs&GO. Hum Mutat 38:1064–1071. https://doi.org/10.1002/humu.23179

Capriotti E, Ozturk K, Carter H (2019) Integrating molecular networks with genetic variant interpretation for precision medicine. Wiley Interdiscip Rev Syst Biol Med 11:e1443. https://doi.org/10.1002/wsbm.1443

Claussnitzer M, Cho JH, Collins R et al (2020) A brief history of human disease genetics. Nature 577:179–189. https://doi.org/10.1038/s41586-019-1879-7

Fernald GH, Capriotti E, Daneshjou R et al (2011) Bioinformatics challenges for personalized medicine. Bioinformatics 27:1741–1748. https://doi.org/10.1093/bioinformatics/btr295

Grimm DG, Azencott C, Aicheler F 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. https://doi.org/10.1002/humu.22768

Ioannidis NM, Rothstein JH, Pejaver V et al (2016) REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. Am J Hum Genet 99:877–885. https://doi.org/10.1016/j.ajhg.2016.08.016

Kent WJ, Sugnet CW, Furey TS et al (2002) The human genome browser at UCSC. Genome Res 12:996–1006. https://doi.org/10.1101/gr.229102

Kircher M, Witten DM, Jain P et al (2014) A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46:310–315. https://doi.org/10.1038/ng.2892

Landrum MJ, Chitipiralla S, Brown GR et al (2020) ClinVar: improvements to accessing data. Nucleic Acids Res 48:D835–D844. https://doi.org/10.1093/nar/gkz972

MacArthur DG, Manolio TA, Dimmock DP et al (2014) Guidelines for investigating causality of sequence variants in human disease. Nature 508:469–476. https://doi.org/10.1038/nature13127

McInnes G, Sharo AG, Koleske ML et al (2021) Opportunities and challenges for the computational interpretation of rare variation in clinically important genes. Am J Hum Genet 108:535–548. https://doi.org/10.1016/j.ajhg.2021.03.003

Miller M, Vitale D, Kahn PC et al (2019) funtrp: identifying protein positions for variation driven functional tuning. Nucleic Acids Res 47:e142. https://doi.org/10.1093/nar/gkz818

Niroula A, Vihinen M (2016) Variation interpretation predictors: principles, types, performance, and choice. Hum Mutat 37:579–597. https://doi.org/10.1002/humu.22987

Ozturk K, Carter H (2021) Predicting functional consequences of mutations using molecular interaction network features. Hum Genet. https://doi.org/10.1007/s00439-021-02329-5

Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: machine learning in Python. JMLR 12:2825–2830

Petrosino M, Novak L, Pasquo A et al (2021) Analysis and interpretation of the impact of missense variants in cancer. Int J Mol Sci 22:5416. https://doi.org/10.3390/ijms22115416

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A (2010) Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 20:110–121. https://doi.org/10.1101/gr.097857.109

Rentzsch P, Witten D, Cooper GM et al (2019) CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res 47:D886–D894. https://doi.org/10.1093/nar/gky1016

Rost B, Radivojac P, Bromberg Y (2016) Protein function in precision medicine: deep understanding with machine learning. FEBS Lett 590:2327–2341. https://doi.org/10.1002/1873-3468.12307

Schneider TD (1997) Information content of individual genetic sequences. J Theoret Biol 189:427–441. https://doi.org/10.1006/jtbi.1997.0540

Siepel A, Bejerano G, Pedersen JS et al (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15:1034–1050. https://doi.org/10.1101/gr.3715005

Suzek BE, Huang H, McGarvey P et al (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 23:1282–1288. https://doi.org/10.1093/bioinformatics/btm098

Tennessen JA, Bigham AW, O'Connor TD et al (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337:64–69. https://doi.org/10.1126/science.1219240

Valdar WSJ (2002) Scoring residue conservation. Proteins 48:227–241. https://doi.org/10.1002/prot.10146

Walsh I, Fishman D, Garcia-Gasulla D et al (2021) DOME: recommendations for supervised machine learning validation in biology. Nat Methods. https://doi.org/10.1038/s41592-021-01205-4