Published online 8 May 2017 *Nucleic Acids Research, 2017, Vol. 45, Web Server issue* ***W247–W252***
doi: 10.1093/nar/gkx369

PhD-SNP^g: a webserver and lightweight tool for scoring single nucleotide variants

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

[1]Department of Biological, Geological, and Environmental Sciences (BiGeA), University of Bologna, Via F. Selmi 3, Bologna 40126, Italy and [2]Department of Comparative Biomedicine and Food Science. University of Padova, Viale dell'Università, 16, 35020 Legnaro, PD, Italy

Received January 31, 2017; Revised April 11, 2017; Editorial Decision April 21, 2017; Accepted April 24, 2017

ABSTRACT

One of the major challenges in human genetics is to identify functional effects of coding and non-coding single nucleotide variants (SNVs). In the past, several methods have been developed to identify disease-related single amino acid changes but only few tools are able to score the impact of non-coding variants. Among the most popular algorithms, CADD and FATHMM predict the effect of SNVs in non-coding regions combining sequence conservation with several functional features derived from the ENCODE project data. Thus, to run CADD or FATHMM locally, the installation process requires to download a large set of pre-calculated information. To facilitate the process of variant annotation we develop PhD-SNP^g, a new easy-to-install and lightweight machine learning method that depends only on sequence-based features. Despite this, PhD-SNP^g performs similarly or better than more complex methods. This makes PhD-SNP^g ideal for quick SNV interpretation, and as benchmark for tool development. Availability: PhD-SNP^g is accessible at http://snps.biofold.org/phd-snpg.

INTRODUCTION

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

Thus, predicting the functional effect of genetic variants is a key challenge for the interpretation of the human genome, and in turn, for the implementation of more accurate diagnostic and treatment strategies (6,7). In the last few years, several methods have been developed for predicting the impact of single nucleotide variants (SNVs) on human health, nevertheless only few of them are capable of assessing the effect of SNVs in non-coding regions (8).

In this paper, we present PhD-SNP^g, which is an extension of the PhD-SNP algorithm (9) for predicting the impact of human SNVs, both in coding and non-coding regions. PhD-SNP^g is available both as web server, and standalone software to process large datasets of variants locally. PhD-SNP^g, which is designed to be simple and lightweight, consists of a machine-learning core, trained only on comparative information in the form of the conservation score calculated from multiple sequence alignments. This information is extracted from the UCSC (University of California, Santa Cruz) repository (https://genome.ucsc.edu/). With respect to the state-of-the-art methods, such as CADD (10), FATHMM-MKL (11) and GVAWA (12), our tool requires a relatively small amount of input resources, and this makes PhD-SNP^g easier to install and run on new sets of variations, even on laptop computers. As an example, to run the full version of PhD-SNP^g <30 Gb data from UCSC are needed. This must be contrasted with the 400 Gb (or more) required by FATHMM-MKL and CADD. In addition, the lightest version of PhD-SNP^g (∼100 Mb) can run in a 'web mode' by retrieving the UCSC data directly from their URLs, without downloading the whole genome files.

Given its simple input (only nucleotide sequence and conservation are required), PhD-SNP^g can also be regarded as baseline tool for benchmarking algorithms based on more complex input features. In particular, PhD-SNP^g can be used for estimating the improvement of the performance obtained by adding new input features (such as open chromatin, histone modification, transcription factor binding sites etc.). For this reason, all the training and testing datasets created for implementing PhD-SNP^g are available online. The availability of benchmark datasets is a particularly critical point for evaluating the discriminative power of new methods with different input features, and at the same time, avoiding an overestimation of the performances (13).

*To whom correspondence should be addressed. Tel: +39 51 2094303; Fax: +39 51 2094286; Email: emidio.capriotti@unibo.it

METHOD OUTLINE

Technically, PhD-SNP^g is a binary classifier based on a Gradient Boosting algorithm, as implemented in *scikit-learn* package (14). PhD-SNP^g was trained and tested using a set of ∼36,000 *Pathogenic* and *Benign* SNVs extracted from Clinvar dataset (15) (Supplementary Table S1). In Figure 1A, the location and the type of each mutation is depicted on the corresponding human chromosome cartoon (*Pathogenic* in red and *Benign* in blue).

Dataset selection

The dataset of SNVs used for training and testing PhD-SNP^g was extracted from Clinvar (15) (http://www.ncbi.nlm.nih.gov/clinvar/). The Clinvar dataset (version January 2016) was filtered by selecting the SNVs with either *Pathogenic* or *Benign* annotation. After this filtering, we ended up with a dataset (Clinvar012016) that consists of 24,267 *Pathogenic* and 11,535 *Benign* SNVs. In the Clinvar012016 dataset, 2,720 (11%) of the *Pathogenic* and 3,942 (34%) of the *Benign* SNVs are in non-coding regions.

To evaluate the method on new incoming data, we derived a second test set based on a more recent version of Clinvar (March 2016), by selecting annotated SNVs not present in the training set (Clinvar012016). The new dataset, indicated as NewClinvar032016, is composed by 1,408 SNVs, 808 of which are annotated as *Pathogenic* and 600 as *Benign*. In the NewClinvar032016 dataset, 283 (35%) of the *Pathogenic* and 336 (56%) of the *Benign* SNVs are in non-coding regions. The files containing the Clinvar012016 and NewClinvar032016 datasets with the PhD-SNP^g predictions are included as Supplementary Files. The genomic location in those files is based on the hg38 human genome assembly.

To further evaluate the performance of PhD-SNP^g we have collected a dataset (AllToolScores) composed only by nonsynonymous SNVs (nsSNVs). This dataset was obtained by merging the five datasets by Grimm and coworkers from VarIBench website (16), and removing the nsSNVs occurring in genomic locations included in the PhD-SNP^g training set. The AllToolScores set consists of 69,529 nsSNVs, ∼41% of which have been annotated as *Pathogenic*. A final test for scoring PhD-SNP^g was performed on a set of 30 non-coding SNVs (LiverVariants) whose change in transcriptional activity was experimentally determined (17).

A summary of the composition of all datasets is reported in Supplementary Table S1.

Feature evaluation

The PhD-SNP^g input consists of 35 values, 25 encoding for the sequence and mutation and 10 for the PhyloP conservation scores (18), as pre-computed at the UCSC repository (Figure 1B).

In details they are: (i) 25 values representing the five-nucleotide window sequence centered on the mutated position (five times five possible nucleotides: A, C, G, T, N); (ii) 10 values mapping the conservation scores of the seven-species (PhyloP7) and 100-species alignments (PhyloP100) to the five window positions. Among the different input features, PhyloP100 shows the highest discriminative power (see Supplementary Table S2), as confirmed by plotting its distribution for *Pathogenic* and *Benign* SNVs (Figure 1C and Supplementary Figure S1). More details about the input features and the optimization procedure of PhD-SNP^g are reported in Supplementary Materials (see also Supplementary Tables S2–S5).

Testing prediction performance

First, PhD-SNP^g performances were assessed by performing a 10-fold cross-validation test on ∼36,000 SNVs. On this subset PhD-SNP^g achieves an Area Under the Receiver Operating Characteristics (19) Curve (AUC) of 0.93 (1 and 0.5, are the scores of a perfect and random predictors, respectively). This result is shown in Table 1 and Supplementary Figure S2. In the first test PhD-SNP^g performs as well as state-of-the-art methods (CADD and FATHMM-MKL) even though, for this test, their scores were not calculated in cross-validation. Furthermore, to evaluate the generalization capability of our predictor, and to better compare PhD-SNP^g to the state-of-the-art methods, we extracted a set of ∼1400 newly annotated SNVs from a more recent version of Clinvar (March 2016). On the NewClinvar032016 testing set, the AUC of PhD-SNP^g is 0.92, which is still high and comparable with that obtained in the cross-validation test. It is worth noticing, that PhD-SNP^g score compares well with those obtained on the same set by the state-of-the-art methods, CADD and FATHMM-MKL (Table 2 and Figure 1D). The same trend is observed on the subsets of mutations located in coding and non-coding regions. These are surprising results, considering the limited information employed by PhD-SNP^g in comparison with the other approaches.

It was pointed out that prediction tools can be hindered by two types of bias (13), such as: the same variants (type-1 circularity), or different variants from the same protein (type-2 circularity) occurring in both the training and validation sets. To exclude these sources of bias, we split our training and testing sets in a way that variants in the same chromosome (and in same gene) are kept in the same subset. To avoid that variants belonging to the same gene were assigned to different subsets, all the SNVs in the sex chromosomes (X and Y) were kept together. Nonetheless, we further checked the presence of hidden type-2 circularity by calculating the performance of our method for the subset of variants in genes with different ratio of pathogenic to benign SNVs. This test was recently introduced for checking the presence of type-2 circularly bias (13). In our analysis, we divided Clinvar012026 in subsets of variants from '*mixed*' genes, which have both pathogenic and benign SNVs in different proportions, and '*pure*' genes with only one class of variants (either pathogenic or benign). The result shows that PhD-SNP^g is not affected by type-2 circularity bias because it achieves on average similar AUC or better MCC (Matthews correlation coefficient) on the subsets of variants from the '*mixed*' genes with respect to the '*pure*' subset (Supplementary Table S7 and Supplementary Figure S3).

To provide a further comparison of the performance of PhD-SNP^g, CADD and FATHMM-MKL in predicting the impact of coding variants, we scored the three algorithms

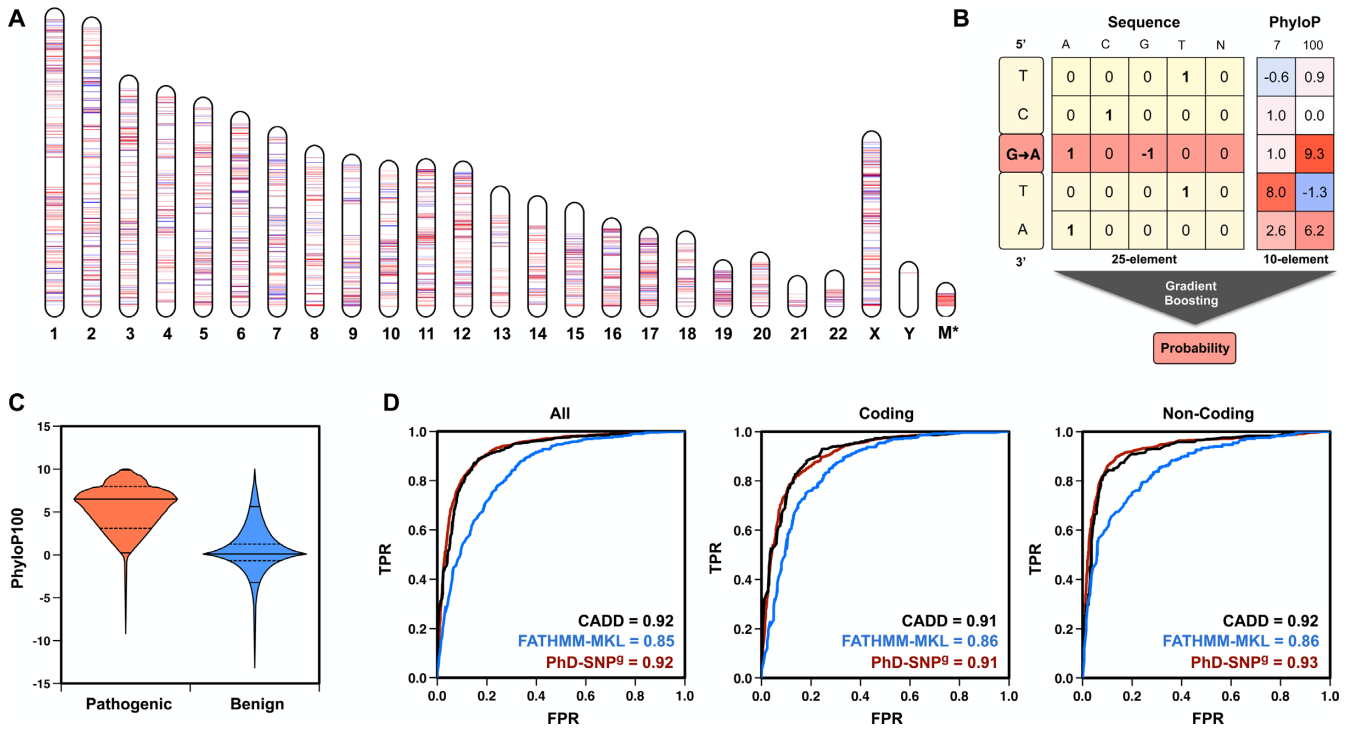

Figure 1. (**A**) Distribution of *Pathogenic* (red) and *Benign* (blue) single nucleotide variants (SNVs) along the chromosomes. *The size of the mitochondrial chromosome (M) in panel A is increased 2,500 times. (**B**) Schematic view of the PhD-SNP^g algorithm and its input features. (**C**) Distribution of PhyloP100 scores in the loci where *Pathogenic* (red) *Benign* (blue) SNVs are detected. (**D**) Performance of PhD-SNP^g (red), CADD (black) and FATHMM-MKL (blue) on the testing set (NewClinvar032016).

Table 1. Performance of PhD-SNP^g, FATHMM-MKL and CADD on the Clinvar012016 dataset

Method	Dataset	Q_2	TNR	NPV	TPR	PPV	MCC	F1	AUC
PhD-SNP^g	All	0.88	0.81	0.82	0.91	0.91	0.72	0.91	0.93
	Coding	0.88	0.74	0.77	0.92	0.91	0.67	0.92	0.92
	Non-coding	0.90	0.92	0.91	0.86	0.88	0.78	0.87	0.94
FATHMM-MKL[a]	All	0.84	0.67	0.79	0.91	0.85	0.61	0.88	0.88
	Coding	0.83	0.58	0.70	0.91	0.86	0.53	0.89	0.86
	Non-coding	0.88	0.84	0.95	0.92	0.79	0.75	0.85	0.95
CADD[a]	All	0.90	0.90	0.81	0.90	0.95	0.78	0.93	0.95
	Coding	0.91	0.85	0.80	0.93	0.95	0.77	0.94	0.94
	Non-coding	0.88	0.99	0.83	0.71	0.99	0.76	0.82	0.94

Q_2: 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. PhD-SNP^g: performance evaluation measures (defined in Supplementary Materials) are averaged over five cross-validation tests (10-fold). The standard errors for all the performance measures are reported in Supplementary Table S6.
[a]FATHMM-MKL and CADD returned predictions respectively on 99.3% and 99.9% of the total dataset.

Table 2. Performances of PhD-SNP^g, FATHMM-MKL and CADD on the NewClinvar032016 dataset.

Method	Dataset	Q_2	TNR	NPV	TPR	PPV	MCC	F1	AUC
PhD-SNP^g	All	0.86	0.77	0.88	0.93	0.85	0.72	0.88	0.92
	Coding	0.85	0.67	0.85	0.94	0.85	0.65	0.89	0.91
	Non-coding	0.88	0.86	0.91	0.90	0.84	0.75	0.87	0.93
FATHMM-MKL[a]	All	0.78	0.58	0.85	0.93	0.75	0.55	0.83	0.85
	Coding	0.81	0.58	0.82	0.94	0.81	0.57	0.87	0.86
	Non-coding	0.73	0.57	0.89	0.91	0.64	0.51	0.75	0.86
CADD[a]	All	0.86	0.82	0.85	0.89	0.87	0.72	0.88	0.92
	Coding	0.86	0.70	0.85	0.94	0.86	0.68	0.90	0.91
	Non-coding	0.87	0.92	0.85	0.81	0.90	0.74	0.85	0.92

Q_2: 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. PhD-SNP^g: performance evaluation measures (defined in Supplementary Materials) are averaged over five tests with previous Clinvar012016 models. The standard error for all the performance measures for PhD-SNP^g is below 1%.
[a]FATHMM-MKL and CADD returned predictions respectively on 99.6% and 99.8% of the total dataset.

on a dataset of nonsynonymous SNVs (AllToolScores) from VariBench (16). This test confirmed that PhD-SNP^g performs similarly to CADD and better than FATHMM-MKL (Supplementary Table S8).

Finally, we also evaluated the ability of PhD-SNP^g to predict the effect of non-coding variants on transcriptional activity. We estimated the correlation coefficient (R^2) between the output of PhD-SNP^g (probability of pathogenicity) and the logarithm of the ratio between the transcription activities in the mutated versus the wild-type mouse liver cells. This test, based on the correlation coefficients for the whole set of 30 SNVs and its subsets (17), shows that PhD-SNP^g achieved better R^2 than CADD and FATHMM-MKL (Supplementary Table S9).

More information about the procedure for comparing PhD-SNP^g with the state-of-the-art methods as well as the definition of the performance evaluation measures are provided in Supplementary Materials.

Method usage

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. Our scripts accept as input genomic coordinates from both assemblies of human genome: hg19 and hg38.

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 PhyloP100 score is available.

Prediction output

The main PhD-SNP^g output is a probabilistic score between 0 and 1. When the score is >0.5 the SNVs is predicted as *Pathogenic* otherwise *Benign*. PhD-SNP^g also returns three values that provide additional information in support of the prediction. They are: the false discovery rate (FDR), the PhyloP100 score in the mutated position and the average PhyloP100 score calculated on the five-nucleotide input window. The false discovery rate, defined in supplementary materials, can be used to filter out less reliable predictions. The empirical function for the calculation of the FDR is plot in Supplementary Figure S4.

SERVER DETAILS

Predicting the impact of single nucleotide variants

PhD-SNP^g server predicts the impact of a single nucleotide variant provided as comma-separated value (CSV) text or variant calling format (VCF). For each SNV the CSV input is composed by four elements, which indicate the chromosome, the position, the reference and alternative alleles. For example, the variation of a Thymine to Cytosine in chromosome 17, position 41 251 803 is represented by 1741251803, T,C. Multiple SNVs can be provided by copy/pasting in the input box a list of variants in separated rows. For formatting reasons, the input in VCF format should be provided by uploading a file, which contains an header starting with a hashtag (#) followed by the identifiers of at least five columns (CHROM, POS, ID, REF, ALT) separated by a tab character. After the header line, each SNV is indicated in a separated row. If the variant's ID in the third column is missing or not available a dot sign (.) must be used.

When the list of SNVs is provided, either in CSV or VCF formats, 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), which quickly extracts a portion of the human genome from a sequence file in binary format. A window sequence of five nucleotides centered around the mutated position is used to generate the 25-element vector encoding for the sequence information. If the nucleotide in input matches the reference allele, the server extracts the corresponding conservation indexes (PhyloP7 and PhyloP100) 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 (20). The PhyloP7 and PhyloP100 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 in input to the Gradient Boosting algorithm, which returns the prediction output described above. In the final step of the prediction task, the PhD-SNP^g server annotates the input variants using *TransVar* tool (21). *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

To facilitate the task of predicting the impact of single amino acid variants (SAVs), PhD-SNP^g server also takes as input a list of SAVs. Each SAV is represented by the approved HGNC (HUGO Gene Nomenclature Committee) gene symbol (22) and the amino acid change separated by a comma. The amino acid change is indicated putting together the one-letter symbol of the wild-type residue, the position along the protein sequence and the one-letter symbol of the mutant residue. For example, the change of the Methionine (M) in position 237 to Isoleucine (I) in TP53 is represented by the tuple TP53,M237I. When the PhD-SNP^g input is provided in this format (MUT) the server internally maps the protein change back to variant at the genomic level using *TransVar* algorithm. After this step, the impact of the SNVs is predicted using the procedure described above.

Input interface

The web interface of PhD-SNP^g consists of a textarea box where the SNVs, in CSV and MUT format are provided. Below a 'Browse' button allows to upload CSV and VCF files either in standard text or *gzipped* format. When the list of SNVs is provided, three '*Input Type*' buttons (CSV, VCF and MUT) allow to select the appropriate input format. A second group of buttons (Assembly) is used to indicate the human reference genome (hg19 or hg38) to which the SNVs are referred. Examples of inputs in CSV and MUT format can be provided using respectively the '*chr,pos,ref,alt*' and '*gene,mut*' links at the top of web interface. Although an example of VCF-like input is linked in the '*Help*' web page,

the usage of the textarea box for the VCF input format is discouraged.

On the bottom of the PhD-SNP^g web page, the e-mail box (optional) is available for receiving PhD-SNP^g output by e-mail.

Server output

The PhD-SNP^g output is an interactive web page where the prediction output is reported in tabular form. 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 to each SNV are reported in rows composed by nine columns. The first four columns define the SNV 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 (avgphylop100). A plus sign (+) at the beginning of each row allows to visualize the results of the annotation performed by *TransVar* algorithm. When a SNV maps on a coding region, four rows report the following information: i) RefSeq (23) code of the longest transcript (Transcript), ii) the HGNC gene symbol and the associated UniProt (24) identifiers (gene), iii) the sense of the translated strand (strand) and iv) 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 bottom part information about errors and warnings occurring during the prediction process.

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

CONCLUSIONS

The PhD-SNP^g web server is a user-friendly interface to predict the impact of SNVs in coding and non-coding regions. The standalone version of PhD-SNP^g can be easily installed and executed on standard laptop machines. It can run on an Intel Xeon 2.40 GHz machine, with 8GB of RAM and can predict the effect of 1,000 SNVs in <2 min. This time increases, depending on the network speed, when the program runs in the web mode.

Despite its simple input features, PhD-SNP^g performs similarly to the state-of-the-art methods that require more information and resources. This makes PhD-SNP^g a reliable and lightweight tool for evaluating the impact of new variants as well as a baseline benchmark tool for comparing predictors based on more complex input features.

AVAILABILITY AND REQUIREMENTS

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. PhD-SNP^g standalone tool is stored on GitHub (https://github.com/biofold/PhD-SNPg), and can be installed by running a python script that automatically downloads the programs and data from the UCSC repository, with few library dependencies.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

E.C. thanks the Geneifx—Genome Informatics Service at the University of Alabama at Birmingham, AL (USA) for providing the computational resources.

FUNDING

Department of Biological, Geological, and Environmental Sciences (BiGeA) at the University of Bologna (Italy) (to E.C.). Department of Comparative Biomedicine and Food Science, University of Padova (Italy) (to P.F.). Funding for open access charge: BCA-DOR 2016 University of Padova and RFO 2016 University of Bologna.

Conflict of interest statement. None declared.

REFERENCES

1. Capriotti,E., Nehrt,N.L., Kann,M.G. and Bromberg,Y. (2012) Bioinformatics for personal genome interpretation. *Brief. Bioinformatics*, **13**, 495–512.
2. 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.
3. Genomes Project,C., Auton,A., Brooks,L.D., Durbin,R.M., Garrison,E.P., Kang,H.M., Korbel,J.O., Marchini,J.L., McCarthy,S., McVean,0.A. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
4. Cancer Genome Atlas Research, N., Weinstein,J.N., Collisson,E.A., Mills,G.B., Shaw,K.R., Ozenberger,B.A., Ellrott,K., Shmulevich,I., Sander,C. and Stuart,J.M. (2013) The cancer genome atlas pan-cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.
5. International Cancer Genome, C., Hudson,T.J., Anderson,W., Artez,A., Barker,A.D., Bell,C., Bernabe,R.R., Bhan,M.K., Calvo,F., Eerola,I. *et al.* (2010) International network of cancer genome projects. *Nature*, **464**, 993–998.
6. Fernald,G.H., Capriotti,E., Daneshjou,R., Karczewski,K.J. and Altman,R.B. (2011) Bioinformatics challenges for personalized medicine. *Bioinformatics*, **27**, 1741–1748.
7. 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.
8. Niroula,A. and Vihinen,M. (2016) Variation interpretation predictors: principles, types, performance, and choice. *Hum. Mutat.*, **37**, 579–597.
9. Capriotti,E., Calabrese,R. and 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.
10. Kircher,M., Witten,D.M., Jain,P., O'Roak,B.J., Cooper,G.M. and Shendure,J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310–315.

11. 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.
12. Ritchie,G.R., Dunham,I., Zeggini,E. and Flicek,P. (2014) Functional annotation of noncoding sequence variants. *Nat. Methods*, **11**, 294–296.
13. Grimm,D.G., Azencott,C.A., 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.
14. 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. *J. Mach. Learn. Res.*, **12**, 2825–2830.
15. Landrum,M.J., Lee,J.M., Benson,M., Brown,G., Chao,C., Chitipiralla,S., Gu,B., Hart,J., Hoffman,D., Hoover,J. *et al.* (2016) ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.*, **44**, D862–D868.
16. Sasidharan Nair,P. and Vihinen,M. (2013) VariBench: a benchmark database for variations. *Hum. Mutat.*, **34**, 42–49.
17. Nishizaki,S.S. and Boyle,A.P. (2017) Mining the unknown: assigning function to noncoding single nucleotide polymorphisms. *Trends Genet.*, **33**, 34–45.
18. 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.
19. Baldi,P., Brunak,S., Chauvin,Y., Andersen,C.A. and Nielsen,H. (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–424.
20. 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.
21. 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.
22. Gray,K.A., Yates,B., Seal,R.L., Wright,M.W. and Bruford,E.A. (2015) Genenames.org: the HGNC resources in 2015. *Nucleic Acids Res.*, **43**, D1079–D1085.
23. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
24. UniProt Consortium (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.