Supplementary Material

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

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

[1] BioFolD Unit, Department of Biological, Geological, and Environmental Sciences (BiGeA),
University of Bologna, Via F. Selmi 3, Bologna, 40126, Italy.

[2] Department of Comparative Biomedicine and Food Science. University of Padova,
Viale dell'Università, 16, 35020 Legnaro (PD), Italy.

[*]Correspondence should be addressed to E.C. (emidio.capriotti@unibo.it).

Input features

PhD-SNP^g takes in input sequence and conservation-based features from the UCSC (University of California, Santa Cruz) repository (http://hgdownload.cse.ucsc.edu/).

PhD-SNP^g input data is a 35-element vector, 25 elements encode for the sequence of a 5-nucleotide window centered around the mutated position (2 nucleotides in each direction). Each position in the window is represented by a 5-element vector for the 5 possible nucleotides (A,C,G,T,N). The element corresponding to the nucleotide in the sequence is set to +1 and the remaining elements are 0. The 5-element vector encoding for the mutated position in the center of the window is built by settings: -1 to the element associated to reference nucleotide; +1 to the element associated to the mutant nucleotide and 0 to the remaining elements.

The 35-element vector is completed by two 5-element vectors corresponding to the PhyloP conservation scores (1) in the 5-nucleotide window. PhyloP conservation scores in each vector are derived from 7way (PhyloP7) and 100way (PhyloP100) UCSC alignments. Predictions can be performed providing in input genomics coordinates from both hg19 and hg38 human assemblies. The representation of the PhD-SNP^g input features is reported in Fig. 1B in the main manuscript.

Training and testing procedure

The performance of PhD-SNP^g has been assessed by a 10-fold cross-validation procedure in which all the variants corresponding to each chromosome were kept in the same subset to reduce possible overfittig. The variants in X and Y chromosomes have been grouped in the same subset. This procedure also allows to keep the variants belonging to the same gene in the same subset, assigning them either to the testing or training set. To balance our dataset, which is composed by a larger number of “Pathogenic”, we included in the training step variants from dbSNP database (2) (http://www.ncbi.nlm.nih.gov/SNP/). In particular we randomly selected 12,732 Single Nucleotide Variants (SNVs) with allele frequency >10% from dbSNP. These variants are only considered in training step and therefore their predictions are removed for the calculation of the evaluation measures.

Additional tests

In a recent paper (3), Grimm and colleagues described two types of possible sources of bias that can affect the variant scoring methods. These sources of bias consist in having the same variants (type-1 circularity) or different variants from the same protein (type-2 circularity) assigned to both training and testing sets. The type-1 circularity bias, which is easier to detect when variants are provided by chromosome location, was certainly excluded by the procedure defined above. The type-2 circularity bias can be more difficult to detect and to estimate the possible impact on the performance of PhD-SNP^g, we adopted the test suggested by Grimm and coworkers (3). This test measures the performance on the subsets of variants from the *“mixed”* genes, which have both pathogenic and benign SNVs in different proportions, and in *“pure”* genes with only class of variants (either pathogenic or benign).

If a predictive tool is affected by type-2 circularity the performance on the subset of variants from gene enriched for one class of SNVs (either pathogenic or benign) tends to be higher than the performance on the subset of *“mixed”* genes with a balanced fraction of pathogenic and Benign SNVs. The results of the test proposed by Grimm and coworkers are reported in Table S7 and Figure S3.

For a further evaluation of the performance of PhD-SNP^g on coding variants, we considered a set of nonsynonymous SNVs (nsSNVs) obtained merging the 5 datasets reported by Grimm and coworkers (3). The datasets were downloaded from VariBench database suite (4) at the webpage http://structure.bmc.lu.se/VariBench/GrimmDatasets.php. To avoid the over estimation of the performance, we removed from the previous sets the SNVs present in training set of PhD-SNP^g (Clinvar012016). After this filtering procedure we obtained a dataset (AllScoreTools) composed by 69,529 nsSNVs, 41% of which are pathogenic. The results of the comparison between the performance of PhD-SNP^g, CADD and FATHMM are reported in Table S8.

To estimate the ability of the variant scoring tools to predict the impact of SNVs on the transcriptional activity, we performed a test on 30 SNVs (LiverVariants). The change in transcriptional activity for these variants was experientially determined (5) and reported in a recent publication (6). For this test we measured the correlation coefficient (R^2) between the predicted probability of pathogenicity returned by PhD-SNP^g and the log2 of the ratio between the transcriptional activities in the mutated and wild-type mouse liver cells. The results of this test are reported in Table S9.

A summary of the composition of the AllScoreTools and LiverVariants datasets is reported in Table S1.

Method optimization

The Gradient Boosting algorithm from *scikit-learn* package (7) (http://scikit-learn.org/) was optimized considering different length of the sequence window, and the maximum depth and the number of the decision trees.

First we tested the predictive power of PhyloP and PhastCons conservation indexes and their combinations. Selecting PhyloP scores as the most informative conservation features (Table S2), we found that optimal performances are reached combining PhyloP7 and PhyloP100 (Table S3). The analysis of the PhyloP100 scores on the mutated sites for pathogenic and benign SNVs (Figure S1) is consistent with the highest discrimination power shown by PhyloP100.

In a second step we optimized window sizes and Gradient Boosting algorithm parameters (# of decision trees and maximum depth) to find the optimal trade-off between the number of features and performances. In the 10-fold cross-validation test on the Clinvar012016 dataset, we find that the predictive power saturates with a 5-nucleotide window input (Table S4). The optimal performance of the Gradient Boosting algorithm is obtained using 200 decision trees and maximum depth 7 (Table S5).

Comparison with state-of-the-art methods

One of the main aims for the development of PhD-SNP^g is the creation of a benchmark tool for testing new algorithms for SNVs prioritization. For this reason, we provided as Supplementary File the results of the 10-fold cross-validation test on the Clinvar012016 dataset and the validation test on the NewClinvar032016 dataset.

In this paper we compared PhD-SNP^g with CADD (8) and FATHMM-MKL (9), although it was not possible to compare them on the same bases, because the cross-validation predictions for CADD and FATHMM-MKL are not available. Moreover, some of the SNVs included in our dataset can overlap with the training set of both methods. For example, comparing the datasets used for training and testing CADD and PhD-SNP^g algorithms (Jun 16, 2012 and Jan 4, 2016 respectively), we estimated that a minimum of ~24% of the variants are in common. The results of our tests on Clinvar012016 and NewClinvar032016 are summarized in Tables 1 and 2 respectively. The standard error of the performance for the 10-fold cross-validation test on Clinvar012016 is reported in Tables S6. The relative ROC curves are shown in Figure S2 and Figure 1D.

We also compared the performance of PhD-SNP^g with CADD and FATHMM-MKL on the AllScoreTools and LiverVariants datasets. The predictions of CADD and FATHMM for the AllScoreTools dataset were downloaded from VariBench suite (4). The performance of CADD and FATHMM-MKL on the LiverVariants dataset were reported in a recent publication (6).

Evaluation measures for binary classifiers

For each prediction, the binary classification (*Pathogenic/Benign*) is made at the output threshold of 0.5. Thus, if probability of *Pathogenic* classification is >0.5 the mutation is predicted to be *Pathogenic*. For CADD a raw score threshold of 3 was used to calculate the performance.

In all the performance measures - assuming that positives indicate *Pathogenic* and negatives indicate *Benign* - TP (true positives) are correctly predicted Pathogenic Single Nucleotide Variants (SNVs), TN (true negatives) are correctly predicted *Benign* variants, FP (false positives) *Benign* SNVs annotated as *Pathogenic*, and FN (false negatives) are *Pathogenic* variants predicted to be *Benign*.

Predictor performance was evaluated using the following metrics: true positive and negative rates (*TPR, TNR*), positive and negative predicted values (*PPV, NPV*), *F1* score and overall accuracy (Q_2)

$$\begin{aligned}
Pathogenic:\ PPV &= \frac{TP}{TP+FP} \quad TPR = \frac{TP}{TP+FN} \\
Benign:\ NPV &= \frac{TN}{TN+FN} \quad TNR = \frac{TN}{TN+FP} \qquad \text{[Eq. 1]} \\
F1 = \frac{2TP}{2TP+FP+FN} &\quad Q_2 = \frac{TP+TN}{TP+FP+TN+FN}
\end{aligned}$$

We computed the Matthew’s correlation coefficient *MCC* (Eq. 2) as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \qquad \text{[Eq. 2]}$$

We also calculated the area under the receiver operating characteristic (ROC) curve (AUC), by plotting the True Positive Rate as a function of the False Positive Rate at different probability thresholds of annotating a variant as *Pathogenic* or *Benign*. PhD-SNP^{g} calculates the False Discovery Rate (FDR) as a function of the returned output (s_0).

$$Pathogenic: \; FDR(s > s_0) = \frac{FP}{FP + TP} \quad Benign: \; FDR(s < s_0) = \frac{FN}{FN + TN} \qquad \text{[Eq. 3]}$$

REFERENCES

1. 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.
2. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M. and Sirotkin, K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, **29**, 308-311.
3. 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.
4. Sasidharan Nair, P. and Vihinen, M. (2013) VariBench: a benchmark database for variations. *Hum Mutat*, **34**, 42-49.
5. Patwardhan, R.P., Hiatt, J.B., Witten, D.M., Kim, M.J., Smith, R.P., May, D., Lee, C., Andrie, J.M., Lee, S.I., Cooper, G.M. *et al.* (2012) Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol*, **30**, 265-270.
6. Nishizaki, S.S. and Boyle, A.P. (2017) Mining the Unknown: Assigning Function to Noncoding Single Nucleotide Polymorphisms. *Trends Genet*, **33**, 34-45.
7. 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. *Journal of Machine Learning Research*, **12**, 2825-2830.
8. 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.
9. 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.

Supplementary Tables

Dataset	Effect	All	Coding	Non-coding
Clinvar012016	*Pathogenic*	24,267	21,547	2,720
	Benign	11,535	7,593	3,942
	Total	35,802	29,140	7,884
NewClinvar032016	*Pathogenic*	808	525	283
	Benign	600	264	336
	Total	1,408	789	619
AllToolScores	*Pathogenic*	28,413	28,413	-
	Benign	41,116	41,116	-
	Total	69,529	69,529	-
LiverVariants	*Increase*	11	-	11
	Decrease	10	-	10
	No Effect	9	-	9

Table S1. Composition of the Clinvar012016, NewClinvar032016, AllToolScores and LiverVariants datasets.

Conservation	Q_2	TNR	NPV	TPR	PPV	MCC	F1	AUC
PhyloP7	0.808	0.684	0.709	0.867	0.852	0.556	0.859	0.806
PhyloP20	0.807	0.713	0.696	0.852	0.862	0.562	0.857	0.812
PhyloP100	0.834	0.759	0.734	0.869	0.884	0.623	0.877	0.895
PhastCons7	0.740	0.470	0.629	0.868	0.775	0.370	0.819	0.723
PhastCons20	0.743	0.486	0.630	0.865	0.780	0.379	0.820	0.736
PhastCons100	0.817	0.695	0.725	0.874	0.858	0.576	0.866	0.823

Table S2. Discriminative power of the PhyloP and PhastCons conservation scores. Average results of the 5 cross-validation tests (10-fold) performed on the Clinvar012016 dataset. Q_2, TNR, NPV, TPR, PPV, MCC, F1 and AUC are defined in the section above.

Combination	Q_2	TNR	NPV	TPR	PPV	MCC	F1	AUC
PhyloP7+100	0.843	0.743	0.763	0.891	0.879	0.639	0.885	0.903
PhyloP7+20	0.818	0.681	0.736	0.884	0.854	0.577	0.868	0.837
PhyloP20+100	0.842	0.763	0.751	0.880	0.886	0.640	0.883	0.901
PhyloP7+20+100	0.844	0.749	0.763	0.889	0.882	0.642	0.885	0.904

Table S3. Discriminative power of combined PhyloP conservation scores. Average results of the 5 cross-validation tests (10-fold) performed on the Clinvar012016 dataset. Q_2, TNR, NPV, TPR, PPV, MCC, F1 and AUC are defined in the section above.

Window Size	Q_2	TNR	NPV	TPR	PPV	MCC	F1	AUC
1	0.850	0.748	0.779	0.899	0.882	0.654	0.891	0.907
3	0.870	0.790	0.802	0.907	0.901	0.700	0.904	0.925
5	0.879	0.805	0.816	0.914	0.907	0.721	0.911	0.932
7	0.879	0.803	0.818	0.915	0.907	0.722	0.911	0.932
9	0.878	0.801	0.818	0.915	0.906	0.720	0.911	0.932
11	0.879	0.804	0.818	0.915	0.907	0.722	0.911	0.933
13	0.879	0.801	0.819	0.916	0.906	0.722	0.911	0.933

Table S4. PhD-SNP^g window sequence optimization. Average results of the 5 cross-validation tests (10-fold) performed on the Clinvar012016 dataset. Q_2, TNR, NPV, TPR, PPV, MCC, F1 and AUC are defined in the section above.

Depth	Estimators	Q_2	TNR	NPV	TPR	PPV	MCC	F1	AUC
3	100	0.855	0.759	0.783	0.900	0.887	0.665	0.894	0.914
3	200	0.862	0.774	0.794	0.904	0.894	0.683	0.899	0.920
3	300	0.867	0.782	0.801	0.908	0.898	0.694	0.903	0.923
3	400	0.870	0.787	0.805	0.909	0.900	0.700	0.905	0.925
3	500	0.871	0.789	0.806	0.910	0.901	0.702	0.905	0.926
5	100	0.870	0.790	0.805	0.909	0.901	0.702	0.905	0.926
5	200	0.875	0.797	0.811	0.912	0.904	0.712	0.908	0.929
5	300	0.876	0.798	0.813	0.913	0.905	0.715	0.909	0.931
5	400	0.877	0.800	0.814	0.913	0.906	0.716	0.909	0.931
5	500	0.878	0.802	0.815	0.914	0.907	0.719	0.910	0.931
7	100	0.877	0.803	0.814	0.913	0.907	0.718	0.910	0.931
7	200	0.879	0.805	0.816	0.914	0.907	0.721	0.911	0.932
7	300	0.878	0.806	0.815	0.913	0.908	0.721	0.911	0.933
7	400	0.878	0.805	0.815	0.913	0.908	0.721	0.911	0.933
7	500	0.880	0.809	0.816	0.913	0.909	0.724	0.911	0.933
9	100	0.878	0.803	0.815	0.913	0.907	0.719	0.910	0.931
9	200	0.879	0.804	0.817	0.914	0.908	0.721	0.911	0.933
9	300	0.880	0.806	0.819	0.915	0.909	0.725	0.912	0.932
9	400	0.880	0.807	0.818	0.915	0.909	0.724	0.912	0.932
9	500	0.880	0.807	0.819	0.915	0.909	0.725	0.912	0.933
11	100	0.875	0.801	0.809	0.910	0.906	0.714	0.908	0.931
11	200	0.880	0.806	0.818	0.915	0.908	0.723	0.911	0.933
11	300	0.881	0.808	0.819	0.915	0.909	0.726	0.912	0.933
11	400	0.880	0.806	0.819	0.915	0.908	0.724	0.912	0.933
11	500	0.881	0.807	0.821	0.917	0.909	0.727	0.913	0.933

Table S5. Optimization of the Gradient Boosting parameters tree depth (Depth) and number of classifier (Estimators). Average results of the 5 cross-validation tests (10-fold) performed on the Clinvar012016 dataset. Q_2, TNR, NPV, TPR, PPV, MCC, F1 and AUC are defined in the section above.

Method	Dataset	SE_{Q_2}	SE_{TNR}	SE_{NPV}	SE_{TPR}	SE_{PPV}	SE_{MCC}	SE_{F1}	SE_{AUC}
PhD-SNP^g	All	0.01	0.03	0.03	0.02	0.02	0.02	0.02	0.01
	Coding	0.02	0.02	0.05	0.02	0.02	0.03	0.02	0.01
	Non-Coding	0.03	0.03	0.05	0.05	0.03	0.05	0.03	0.02
FATHMM-MKL*	All	0.02	0.05	0.03	0.02	0.02	0.03	0.02	0.01
	Coding	0.03	0.04	0.04	0.03	0.03	0.04	0.02	0.01
	Non-Coding	0.03	0.05	0.03	0.03	0.05	0.04	0.03	0.02
CADD*	All	0.02	0.02	0.05	0.03	0.01	0.05	0.02	0.02
	Coding	0.01	0.01	0.03	0.01	0.01	0.02	0.01	0.01
	Non-Coding	0.07	0.01	0.09	0.09	0.01	0.09	0.07	0.03

Table S6. Standard error for the 10-fold cross-validation test on the Clinvar012016 dataset. The standard error (SE) values refer to the predictions of PhD-SNP^g, FATHMM-MKL and CADD reported in Table 1 of the main manuscript.

PathRate	Q_2	TNR	NPV	TPR	PPV	MCC	F1	AUC	DB
pure	0.885	0.758	0.474	0.900	0.969	0.539	0.933	0.907	0.217
0.0-0.1	0.792	0.782	0.993	0.927	0.244	0.410	0.386	0.926	0.022
0.1-0.2	0.802	0.782	0.979	0.910	0.438	0.537	0.591	0.925	0.067
0.2-0.3	0.823	0.799	0.954	0.890	0.607	0.621	0.721	0.918	0.035
0.3-0.4	0.842	0.798	0.948	0.921	0.715	0.691	0.805	0.931	0.056
0.4-0.5	0.866	0.875	0.874	0.857	0.858	0.732	0.858	0.936	0.111
0.5-0.6	0.875	0.792	0.911	0.939	0.852	0.747	0.894	0.945	0.057
0.6-0.7	0.885	0.803	0.853	0.928	0.901	0.742	0.914	0.942	0.078
0.7-0.8	0.904	0.786	0.823	0.943	0.929	0.741	0.936	0.941	0.093
0.8-0.9	0.919	0.794	0.697	0.940	0.964	0.696	0.952	0.940	0.127
0.9-1.0	0.920	0.767	0.403	0.929	0.985	0.520	0.956	0.940	0.137

Table S7. PhD-SNP^g performance on the subsets of Clinvar012016 variants from genes with different proportions of pathogenic variants (PathRate). The subset *pure* is composed by the variants present in the genes with only one class of SNVs (either Pathogenic or Benign). Q_2, TNR, NPV, TPR, PPV, MCC, F1 and AUC are defined in the section above. DB is the fraction of the Clinvar012016 dataset.

Method	Q_2	TNR	NPV	TPR	PPV	MC	F1	AUC	DB
PhD-SNP^g	0.736	0.774	0.778	0.680	0.676	0.454	0.678	0.797	1.000
CADD”	0.729	0.699	0.815	0.771	0.639	0.462	0.699	0.794	1.000
FATHMM*	0.699	0.781	0.733	0.576	0.639	0.364	0.606	0.741	0.941

Table S8. Performances of PhD-SNP^g, CADD and FATHMM on the AllScoreTools dataset composed by nonsynonymous SNVs. *CADD and FATHMM predictions were downloaded from VariBench (http://structure.bmc.lu.se/VariBench/GrimmDatasets.php). Q_2, TNR, NPV, TPR, PPV, MCC, F1 and AUC are defined in the section above. Optimized thresholds are considered for scoring all the methods. DB is the fraction of the AllScoreTools dataset for which the predictions were available.

	PhD-SNP^g	CADD	FATHMM-MKL
All	0.24	0.18	0.17
Negative	0.57	0.29	0.46
Positive	0.09	0.06	0.05

Table S9. Correlation coefficients (R^2) between the output of the PhD-SNP^g, CADD, FATHMM-MKL and the absolute value $\log_2$ change of the transcriptional activity of the variant compared with wild type (logES). According to a recent study (6), Negative and Positive correspond to the subsets of variants with *logES<0* and *logES>0.02* respectively.

Supplementary Figures

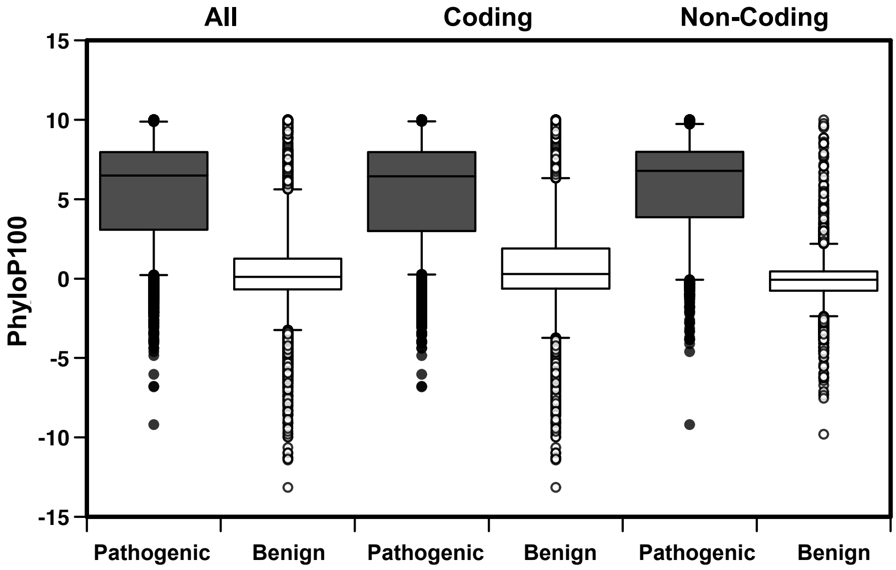

Figure S1. Distribution of the PhyloP100 conservation scores for *Pathogenic* and *Benign* Single Nucleotide Variants (SNVs) from Clinvar012016 dataset and its subsets of coding and non-coding SNVs.

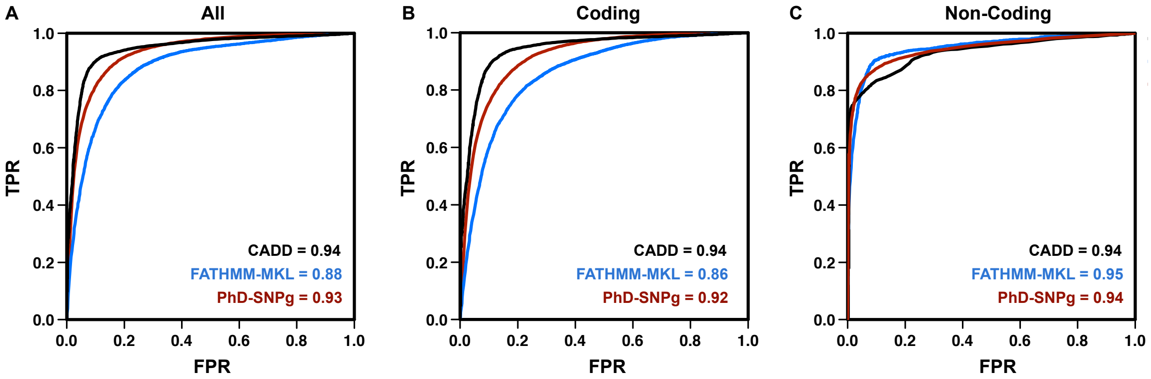

Figure S2. ROC curves of CADD, FATHMM-MKL and PhD-SNP^{g} calculated on the Clinvar012016 dataset and its subset of coding and non-coding Single Nucleotide Variants.

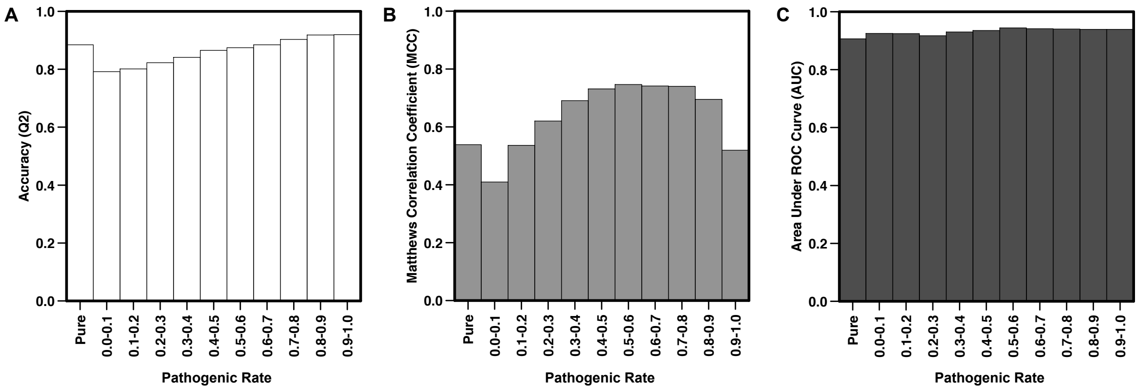

Figure S3. Performance of PhD-SNP^g on the subsets of Clinvar012016 variants in genes with different proportion of Pathogenic SNVs (Pathogenic Rate). The subset "*Pure"* is composed by the variants from genes with only on class of SNVs (either Pathogenic or Benign). Q_2, MCC and AUC are defined in the section above.

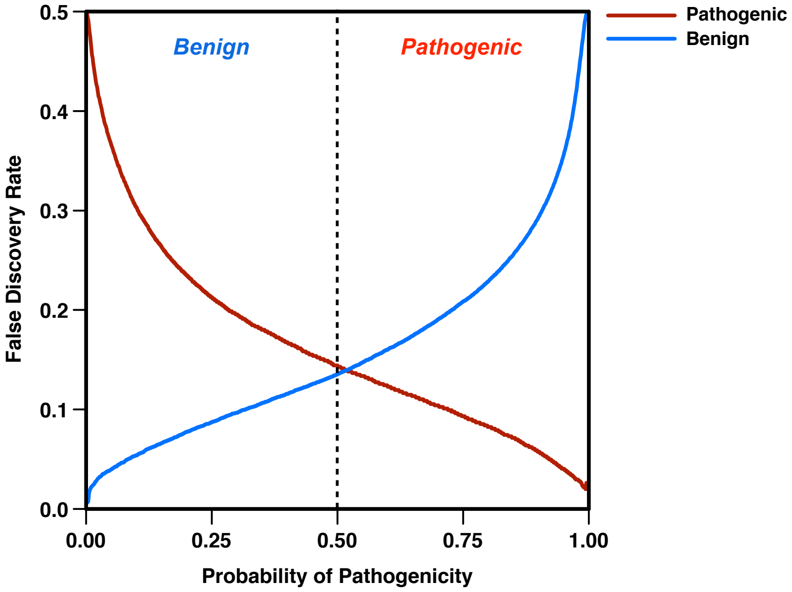

Figure S4. Empirical False Discovery Rate (FDR) as function of the PhD-SNP^g output (probability pathogenicity *s*) for the *Pathogenic* (*s*>*0.5*) and *Benign* (*s*≤*0.5*) predicted SNVs. The functions have been estimated on the Clinvar012016 dataset.