nature | **methods**

A method and server for predicting damaging missense mutations

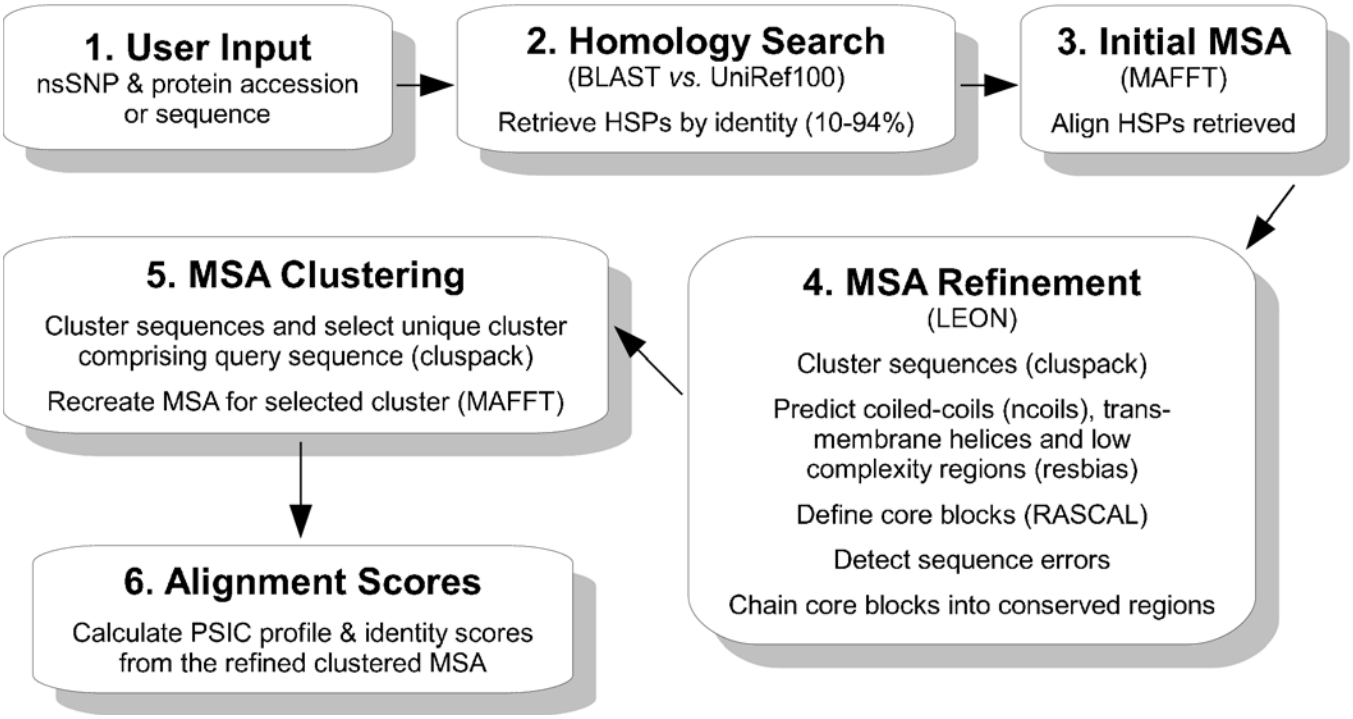
Ivan A Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov & Shamil R Sunyaev

Supplementary figures and text:

Supplementary Figure 1	PolyPhen-2 automated sequence alignment pipeline
Supplementary Figure 2	Stacked bar histograms showing distributions of difference in PSIC score
Supplementary Figure 3	Distributions of the values of all 11 features utilized by PolyPhen-2 classifier
Supplementary Figure 4	Scatter plots of score_delta, id_q_min and id_p_max features
Supplementary Table 1	Complete list of all 32 initial features considered
Supplementary Table 2	Receiver operating characteristics for PolyPhen, PolyPhen-2, SIFT, SNAP, and SNPs3D
Supplementary Methods	Description of datasets and algorithms used

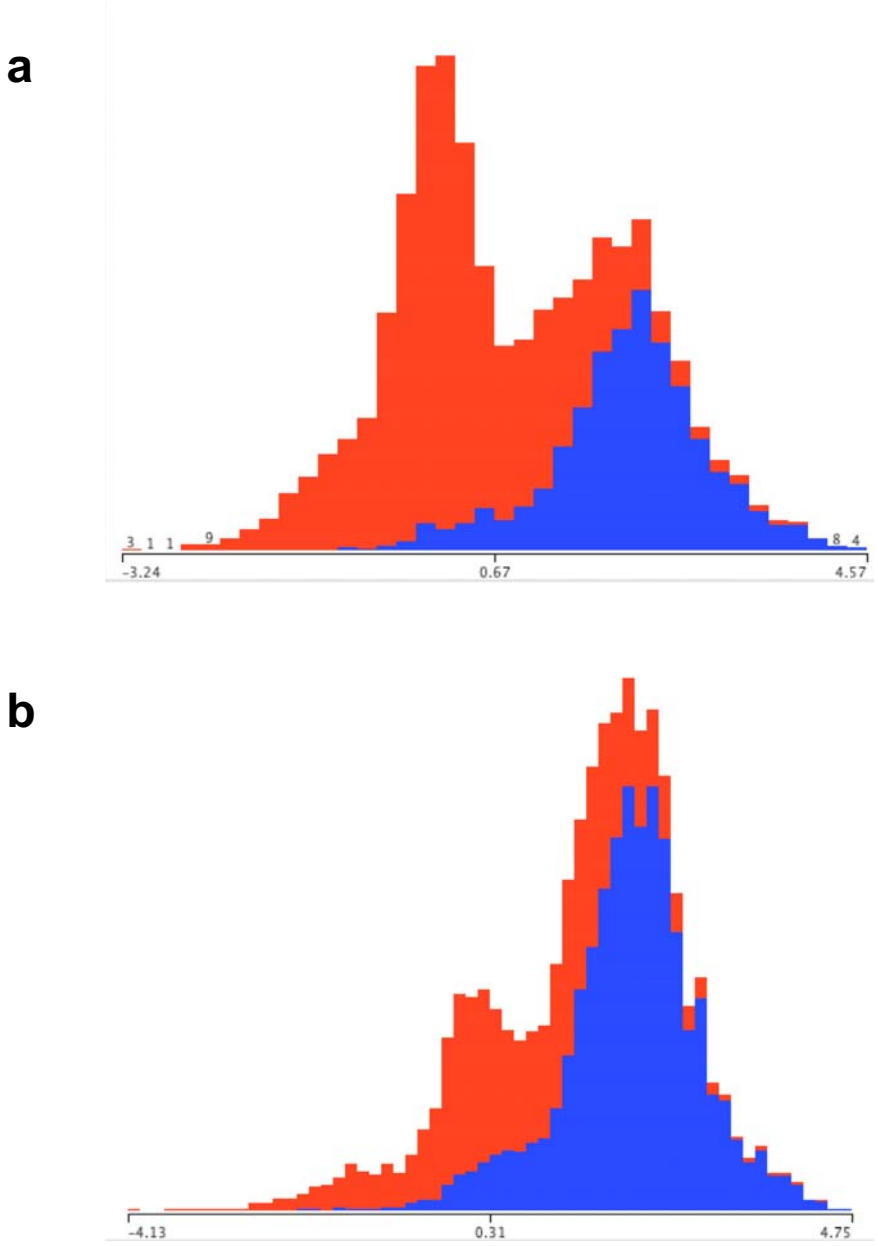
Note: Supplementary Software is available on the Nature Methods website.

Supplementary Figure 1.



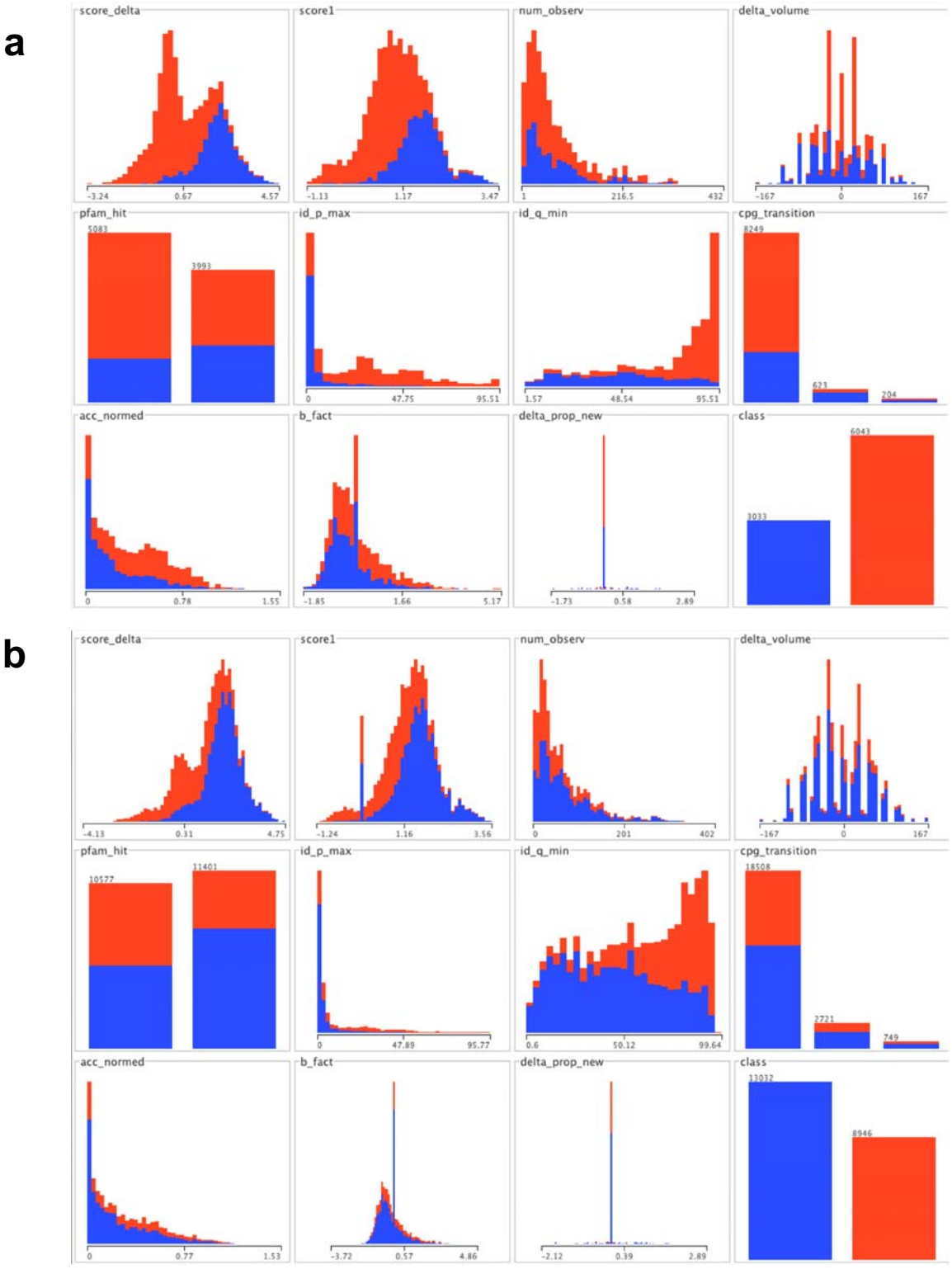
Supplementary Figure 1. PolyPhen-2 automated sequence alignment pipeline. The choice of homologs to be included and the quality of the multiple sequence alignment (**MSA**) are critical for the performance of the method. We have found that including both orthologs and paralogs of the analyzed sequence in MSA leads to more accurate predictions, perhaps because a majority of disease-causing replacements affect protein structure, rather than specific aspects of function[1]. We identify homologs of the analyzed sequence using **BLAST+**[2] and align the amino acid sequences using **MAFFT**[3]. Because the resulting alignments often contain poor-quality segments, we refine them using **Leon** software[4]. Finally, we cluster reliably aligned sequences using **Secator** algorithm[5] implemented in **ClusPack** software (http://www-bio3d-igbmc.u-strasbg.fr/~wicker/programs.html). Only the homologs that belong to a compact cluster which includes the analyzed sequence are taken into account. Using the remaining sequences decreases the quality of prediction, perhaps due to accumulation of compensatory changes[6].

Supplementary Figure 2.



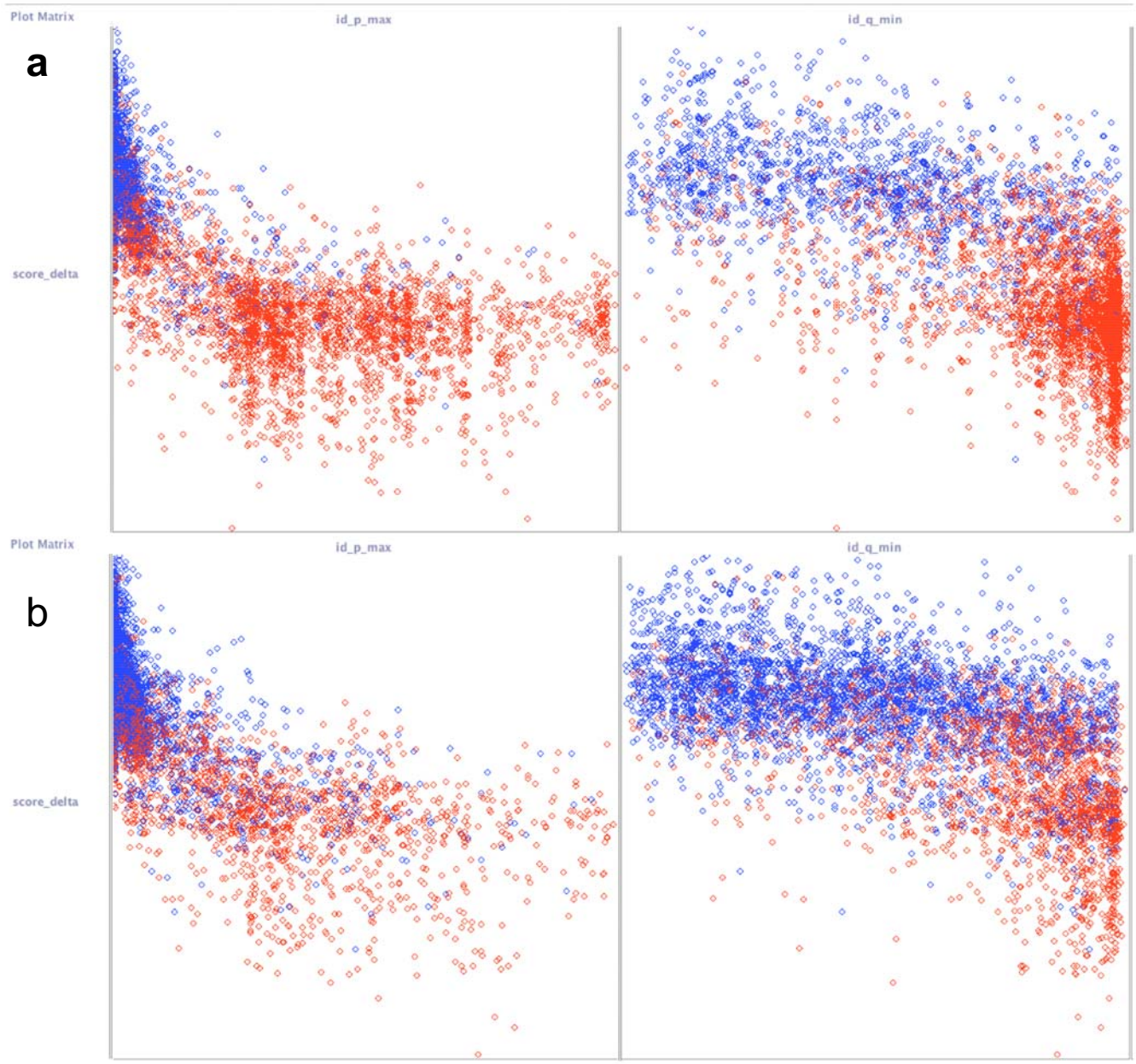
Supplementary Figure 2. Stacked bar histograms showing distributions of difference in PSIC score between a wild type and the corresponding mutant amino acid for HumDiv (**a**) and HumVar[7] (**b**). Each bar contains two parts, a red part representing the fraction of benign replacements with this value of the parameter and a blue part representing a corresponding fraction of damaging replacements. In both cases red and blue distributions are unimodal with well-separated peaks, although the benign mode (red) is quite different between HumDiv and HumVar.

Supplementary Figure 3.



Supplementary Figure 3. Distributions of the values of all 11 features utilized by PolyPhen-2 classifier and a class label for HumDiv (**a**) and HumVar[7] (**b**). Stacked bar histograms presented here follow the same conventions as in **Supplementary Fig. 2**.

Supplementary Figure 4.



Supplementary Figure 4. Scatter plots of the difference in PSIC score between a wild type and the corresponding mutant amino acids (**score_delta**) together with the sequence identity of the closest homolog carrying a non wild-type amino acid (**id_q_min**) and with the congruency of the mutant allele to multiple alignment (**id_p_max**) in HumDiv (**a**) and HumVar[7] (**b**). Together, these two features provide somewhat better separation of benign and damaging mutations than any single feature.

Supplementary Table 1.

Name	Definition	Values with ranges in HumDiv
nt1	wild type allele nucleotide	A,C,G,T
nt2	mutation allele nucleotide	A,C,G,T
site	SITE annotation from UniProt/Swiss-Prot	Yes, No
region	REGION annotation from UniProt/Swiss-Prot	NO, PROPER, SIGNAL, TRANSMEM
phat	PHAT matrix element in the TRANSMEM region	[-8.0, 4.0], mean = -0.04
score1	PSIC score for the wild type allele	[-1.1], mean = 1.07
score2	PSIC score for the mutant allele	[-1.39, 2.64], mean = .166
score_delta	difference of PSIC scores (Score1-Score2)	[-3.23, 4.57], mean = .905
num_observ	number of residues observed at the position of the multiple alignment	[1, 432], mean 69.3
delta_volume	change in residue side chain volume	[-167, 167], mean = -1,93
transv	mutation origin by transversion or transition	Yes, No
CpG	mutation origin in the CpG hypermutable context	Yes, No
pfam_hit	position of the mutation within/outside a protein domain as defined by Pfam	Yes, No
id_p_max	congruency of the mutant allele to the multiple alignment	[0, 95.5], mean = 24
id_q_min	sequence identity with the closest homologue deviating from wild type allele	[1.56, 95.5], mean 68.76
cpgVar1Var2	presence of the CpG context combined with wild type and mutant amino acid types	NO, AA1_AA2
cpg_transition	whether variant happened as transition in CpG context	No, Transition, Transversion
charge_change	change in electrostatic charge	0,1,2
hydroph_change	change in hydrophobicity	[0, 2.85], mean 0.80
ali_ide	sequence identity with the closest homolog with known 3D structure	[0, 1], mean 0.33
ali_len	alignment length with the closest homolog with known 3D structure	[0, 1213], mean 130.0
acc_normed	normalized accessible surface area of amino acid residue	[0, 1.55], mean .35
sec_str	secondary structure	HELIX, SHEET, OTHER
map_region	region of the Ramachandran map	ALPHA, BETA, OTHER
delta_prop	change in accessible surface area propensity	[-2.89, 2.89], mean -0.07
b_fact	crystallographic beta-factor	[-1.85, 5.17], mean 0.0
het_cont_ave_num	average number of contact with heteroatoms	Yes, No
het_cont_min_dist	minimal distance to a heteroatom	Yes, No
inter_cont_ave_num	average number of interchain contacts in a protein complex	Yes, No
inter_cont_min_dist	average minimal interchain distance	Yes, No
delta_volume_new	change in residue volume for buried residues	[-119, 138], mean -0.5
delta_prop_new	change in accessible surface area propensity for buried residues	[-1.83, 2,89], mean 0.0026

Supplementary Table 1. Complete list of all 32 initial features considered, with 11 final features selected for use in PolyPhen-2 classifier highlighted in blue font.

Supplementary Table 2.

Software:	PolyPhen				PolyPhen-2						SIFT	SNAP	SNPs3D
Dataset:	HumDiv		HumVar		HumDiv		HumVar		HumVar-HumDiv		HumVar		
Database:	UR	SP	UR	SP	UR	SP	UR	SP	UR	SP	SP	n/a	n/a
FPR	TPR	TPR	TPR	TPR	TPR	TPR	TPR	TPR	TPR	TPR	TPR	TPR	TPR
0.10	0.696	0.599	0.499	0.422	0.767	0.720	0.552	0.509	0.522	0.495	n/a	0.474	0.485
0.15	0.765	0.690	0.593	0.524	0.866	0.807	0.659	0.614	0.635	0.590	0.514	0.569	0.597
0.20	0.820	0.754	0.660	0.603	0.918	0.870	0.734	0.684	0.717	0.679	0.616	0.641	0.672
0.25	0.852	0.793	0.711	0.664	0.947	0.910	0.787	0.740	0.782	0.735	0.689	0.695	0.726
0.30	0.876	0.823	0.751	0.718	0.963	0.932	0.836	0.792	0.836	0.793	0.745	0.737	0.769
0.35	0.896	0.846	0.788	0.760	0.973	0.951	0.867	0.833	0.866	0.827	0.789	0.774	0.807
0.40	0.908	0.864	0.820	0.792	0.976	0.964	0.897	0.868	0.896	0.861	0.823	0.802	0.838
0.45	0.920	0.882	0.847	0.817	0.978	0.973	0.921	0.897	0.920	0.893	0.853	0.830	0.862
0.50	0.932	0.902	0.873	0.840	0.980	0.977	0.940	0.919	0.937	0.921	0.880	0.852	0.882
0.55	0.940	0.918	0.894	0.861	0.982	0.979	0.954	0.938	0.950	0.938	0.904	0.875	0.901
0.60	0.949	0.930	0.913	0.881	0.984	0.982	0.965	0.957	0.963	0.954	0.922	0.893	0.915
0.65	0.956	0.942	0.927	0.903	0.986	0.984	0.973	0.969	0.969	0.966	0.936	0.911	0.932
0.70	0.962	0.949	0.939	0.919	0.988	0.986	0.979	0.977	0.973	0.973	0.951	0.926	0.946
0.75	0.969	0.960	0.952	0.938	0.990	0.989	0.984	0.985	0.978	0.978	0.964	0.939	0.958
0.80	0.975	0.971	0.963	0.953	0.992	0.991	0.988	0.990	0.982	0.982	0.975	0.953	0.969

Supplementary Table 2. Receiver operating characteristics (ROC) for PolyPhen[8], PolyPhen-2, SIFT[9], SNAP[10], and SNPs3D[11] prediction methods. **FPR**, False Positive Rate; **TPR**, True Positive Rate. **PolyPhen** (v1.18), ROC based on the absolute value of difference between PSIC[12] profile scores of the two allelic variants; **PolyPhen-2** (v2.0.17), ROC based on the probabilistic score derived from the Naïve Bayes model with discretization (see **Supplementary Methods**); **SIFT** (v3.0), ROC based on SIFT Score; **SNAP** (http://cubic.bioc.columbia.edu/services/SNAP/), ROC based on Expected Accuracy; **SNPs3D** (http://www.snps3d.org/), ROC based on SVM Profile score. **HumDiv**, consists of 3,155 damaging alleles with known effects on the molecular function causing human Mendelian diseases annotated in the UniProt database, together with 6,321 differences between human proteins and their closely related mammalian homologs, assumed to be non-damaging; **HumVar**[7], consists of all the 13,032 human disease-causing mutations from UniProt, together with 8,946 human nsSNPs without annotated involvement in disease, which were treated as non-damaging; **HumVar-HumDiv**, consists of SNPs present in HumVar data set with all SNPs corresponding to the proteins also present in the HumDiv data set excluded to avoid bias due to overlap of training and testing data (resulting in a total of 10,583 SNPs in the final set). All predictions by **PolyPhen-2** were obtained using 5-fold cross-validation procedure as described in **Supplementary Methods**, except for **HumVar-HumDiv** data set which was tested using model trained on **HumDiv** dataset. This latter approach was utilized to improve compatibility with the results obtained from SIFT, SNAP, and SNPs3D methods. **UR**, UniRef100[13] Release 15.12 of 15-Dec-2009; **SP**, UniProtKB/Swiss-Prot[13] Release 57.12 of 15-Dec-2009.

Supplementary Methods

HumDiv dataset compilation

The set of damaging mutations was retrieved from UniProtKB[13]. Mutations were considered damaging if their annotations contained keywords implying causal mutation-phenotype relationship ("lethal", "complete loss of function", "causes", "abolishes", "no detectable activity", "impairs", etc.). Among those, we excluded a small number of ambiguous mutations:

- mutations in hemoglobins (203 cases)
- mutations annotated as "unknown" (36 cases)
- cancer-related mutations (182 cases)
- mutations in proteins whose annotations do not contain keyword "disease mutation" (32 cases).

The set of non-damaging mutations was compiled from differences in homologous protein sequences of closely related mammalian species. The data can be downloaded from:

ftp://genetics.bwh.harvard.edu/datasets/HumDiv.tar.gz

Feature selection

PolyPhen-2 uses 11 predictive features which were selected automatically by an iterative greedy algorithm out of a set of nineteen sequence-based and thirteen structure-based candidate features. The performance of classifiers is often negatively affected by the presence of irrelevant or redundant features (**Supplementary Table 1**). To select the optimal set of features, we employed machine-learning methods[14]. Both feature selection and classifier testing were performed in a 5-fold cross-validation scheme, where folds were not randomized but created to ensure that mutations in the same protein would all fall into the same fold. This was done in order to avoid selecting feature values serving as "proxies" to specific proteins, since protein identity could hold a strong clue to whether the mutation is damaging or not. After initially testing on HumDiv, we tested the best performing classifiers on the holdout pair of datasets HumVar[7] and retained features, which provided some improvement. We selected features using both forward selection and backward elimination. In the forward greedy search we continuously added features in the order of their individual performance until there was no increase in prediction accuracy in a cross-validation test. The prediction accuracy was measured by the area under the ROC curve, *i.e.*, as integral value of sensitivity over all specificity thresholds. In the reverse search, we excluded individual features until a reduction in prediction accuracy was observed. We further controlled the feature selection by measuring performance on the holdout HumVar pair of datasets. Both approaches resulted in the same set of eleven features described below.

PSIC score. This profile score reflects how likely it is for a particular amino acid to occupy a specific position in the protein sequence, given the pattern of amino acid substitutions observed in the multiple sequence alignment. This score has the form of likelihood ratio and is computed using the PSIC algorithm[12], which takes the relatedness of homologous sequences into account and uses prior probabilities derived from the amino acid substitution matrix (BLOSUM62). The PSIC score of the wild-type amino acid and the difference between the PSIC scores of the wild type and the mutant amino acids were treated as two separate features.

The sequence identity to the closest homologue carrying any amino acid that differs from the wild-type allele at the site of the mutation considered.

Congruency of the mutant allele to the multiple alignment. For each amino acid at the analyzed site, we computed the sequence identity between the analyzed protein and its closest homologue in which this amino acid is observed. We further computed a product of this sequence identity and the probability, provided by BLOSUM matrix, that this amino acid would be substituted by the mutant amino acid. The maximal value of this product over all amino acids was treated as a feature.

CpG context. CpG context of transition mutations was treated as a feature.

Structural features. Three additional features were selected for proteins with known 3D structures: 1) the accessible surface area of the wild-type amino acid residue, 2) the change in the hydrophobic propensity in the form of "knowledge-based potential", and 3) crystallographic B-factor reflecting conformational mobility of the wild-type amino acid residue[15].

The remaining three features are alignment depth (excluding gaps) at the site of the mutation, change in the amino acid volume between wild type and mutant amino acids, and whether the site of the mutation resides within an annotated Pfam[16] domain.

Classification method

Our method of classification has to deal with data consisting of a mixture of discrete and continuous-valued features and containing a substantial fraction of irregularly scattered missing values. These challenges are naturally handled by a Naïve Bayes approach coupled with entropy-based discretization[17]. This approach performed about equally well as a number of other machine-learning approaches: logistic regression, alternating decision tree, and support vector machine, and outperforms decision trees and random forests. We chose Naïve Bayes for its simplicity because, in contrast to other approaches, it does not contain any parameters, except for representing factored probabilities and smoothing, which is done by Laplace estimators. For a mutant allele, Naïve Bayes approach produces the likelihood that this allele affects protein function, phenotype, and fitness, or, in other words, is damaging, as opposed to benign.

Naïve Bayes approach, as any method of supervised classification, requires data for training and for testing. We used the same pair of datasets, containing known damaging and known benign alleles, for both purposes. Each pair of datasets was split into 5 approximately equal parts, in such a way that all mutations of a protein were assigned to the same part. Then, 4 parts were used for testing and the remaining one for validation, and the procedure was repeated 5 times with different parts used for validation (5-fold cross-validation[14]).

Qualitative appraisal of mutations

The true positive rate was calculated as the fraction of correctly predicted damaging mutations for a given threshold of Naïve Bayes probabilistic score. The false positive rate was calculated as fraction of benign mutations erroneously predicted as damaging. A mutation is classified as "probably damaging" if its probabilistic score is above 0.85, corresponding to the fraction of false positives under 10% on HumDiv and under 19% on HumVar (true positive rates are 78% and 71%, respectively). A mutation is classified as "possibly damaging" if its probabilistic score is above 0.15, corresponding to the fraction of false positives under 18% on HumDiv and 40% HumVar (true positive rates are 89% and 90%, respectively). The remaining mutations are classified as benign.

References

1. Wang, Z. & Moult, J. Hum. Mutat. **17**, 263-270 (2001).
2. Camacho, C. *et al*. BMC Bioinformatics **10**, 421 (2008).
3. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. Nucleic Acids Res. **30**, 3059-3066 (2002).
4. Thompson, J.D., Prigent, V. & Poch, O. Nucleic Acids Res. **32**, 1298-1307 (2004).
5. Wicker, N., Perrin, G.R., Thierry, J.C. & Poch, O. Mol. Biol. Evol. **18**, 1435-1441 (2001).
6. Kondrashov, A.S., Sunyaev, S. & Kondrashov, F.A. Proc. Natl. Acad. Sci. USA **99**, 14878-14883 (2002).
7. Capriotti, E., Calabrese, R. & Casadio, R. Bioinformatics **22**, 2729-2734 (2006).
8. Ramensky, V., Bork, P. & Sunyaev, S. Nucleic Acids Res. **30**, 3894-3900 (2002).

9. Ng, P.C. & Henikoff, S. Nucleic Acids Res. **31**, 3812-3814 (2003).
10. Bromberg, Y., Yachdav, G. & Rost, B. Bioinformatics **24**, 2397-2398 (2008).
11. Yue, P., Melamud, E. & Moult, J. BMC Bioinformatics **7**, 166 (2006).
12. Sunyaev, S.R. *et al*. Protein Eng. **12**, 387-394 (1999).
13. The UniProt Consortium. Nucleic Acids Res. **38**, D142-D148 (2010).
14. Witten, I.H. & Frank, E. Data Mining: Practical machine learning tools and techniques. (Morgan Kaufmann: San Francisco, CA, 2005).
15. Chasman D. & Adams R.M. J. Mol. Biol. **307**, 683-706 (2001).
16. Finn, R.D. *et al*. Nucleic Acids Res. **36**, D281-D288 (2008).
17. Fayyad, U.M. & Irani, K.B. In: Proceedings of the 13th International Joint Conference on Artificial Intelligence, pp. 1022-1027 (1993).