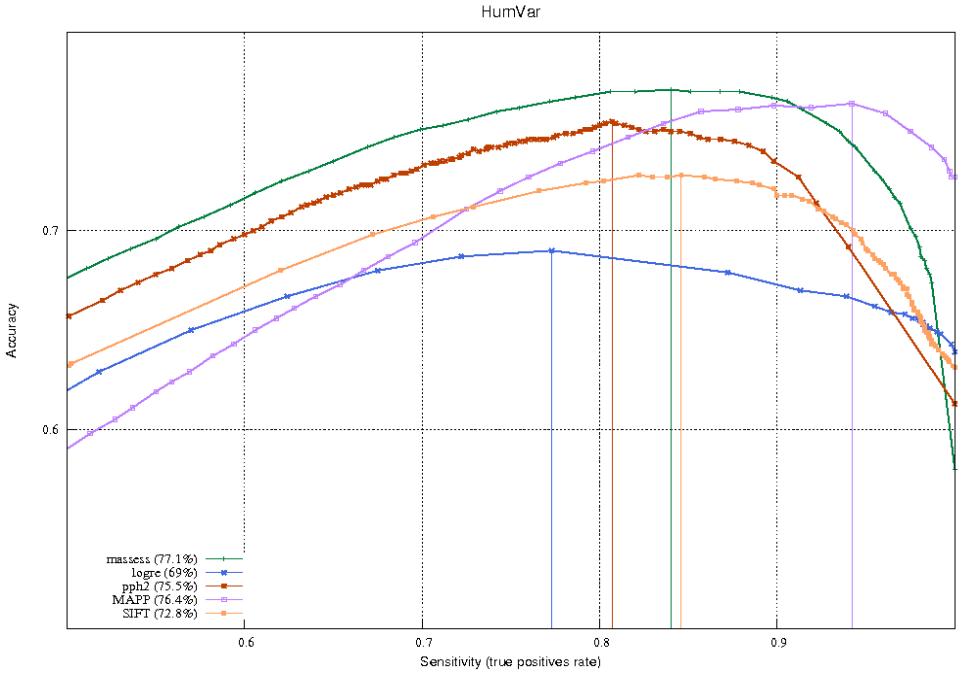
The American Journal of Human Genetics Volume *88*

Supplemental Data

Improving the Assessment of the Outcome of Nonsynonymous SNPs with a Consensus Deleteriousness Score, Condel

Abel González-Pérez and Nuria López-Bigas

Figure S1. Selecting the optimal cutoff for each tool according to their performance on the HumVar (first panel) and HumDiv (second panel) datasets. The abscissa of each graph contains the fraction of deleterious variants from each dataset correctly classified by the five tools, whereas the ordinate represents the accuracy attained at each sensitivity mark. For all tools, the accuracy increases with the sensitivity, up to a point at which the recovery of further deleterious variants is overshadowed by the missclassification of neutral variants, and hence the accuracy starts decreasing. The optimal sensitivity (producing the highest accuracy) is marked for each tool.



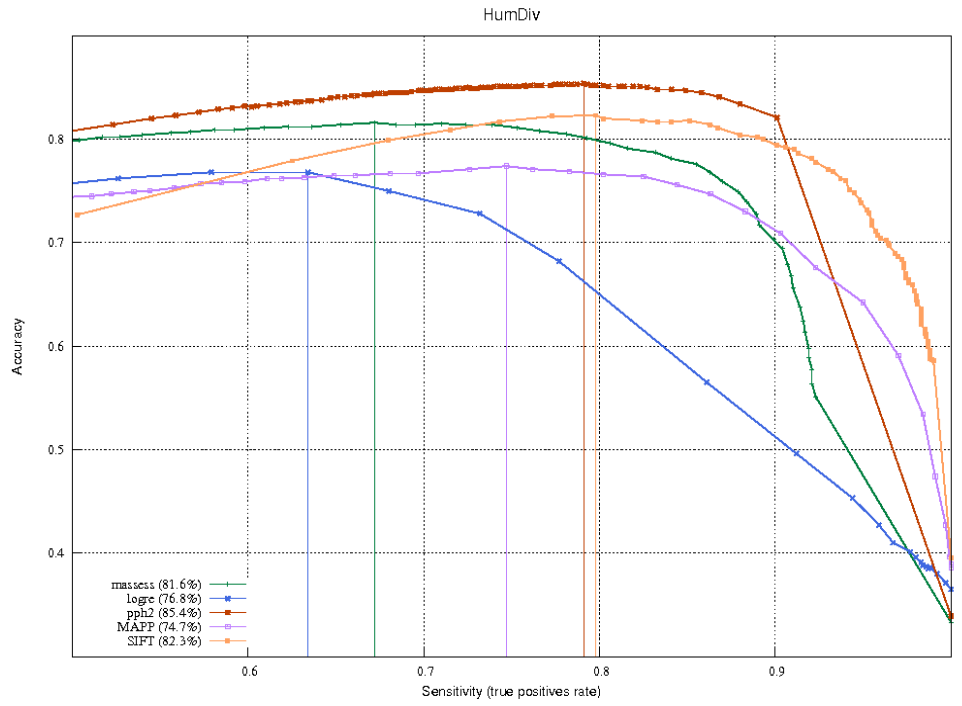


Figure S2. Demonstration of the calculation of weights employed to compute the WVS and the WAS (see Methods section) for positively and negatively classified variants.

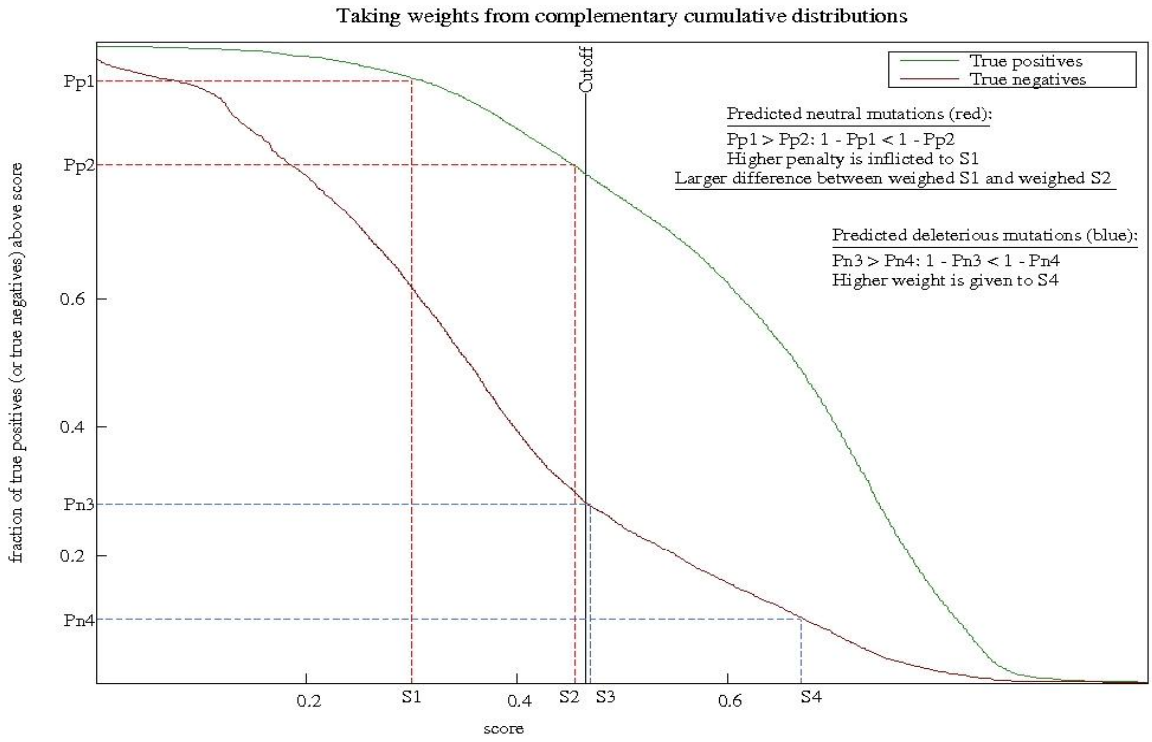


Figure S3. Fraction of variants in HumVar and HumDiv successfully classified as deleterious or neutral by at least a given number of methods. Disease: deleterious variants; polymorphisms: neutral variants.

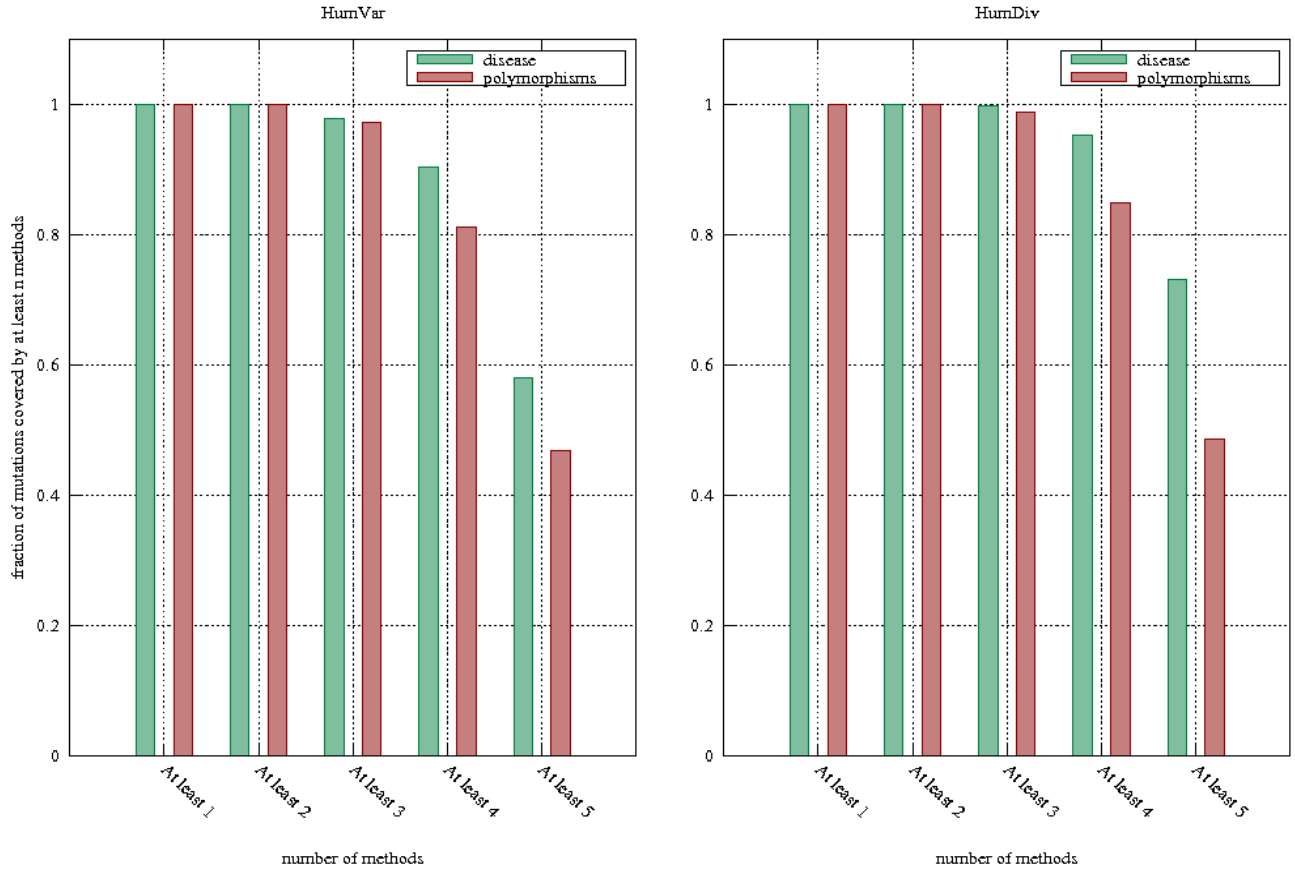
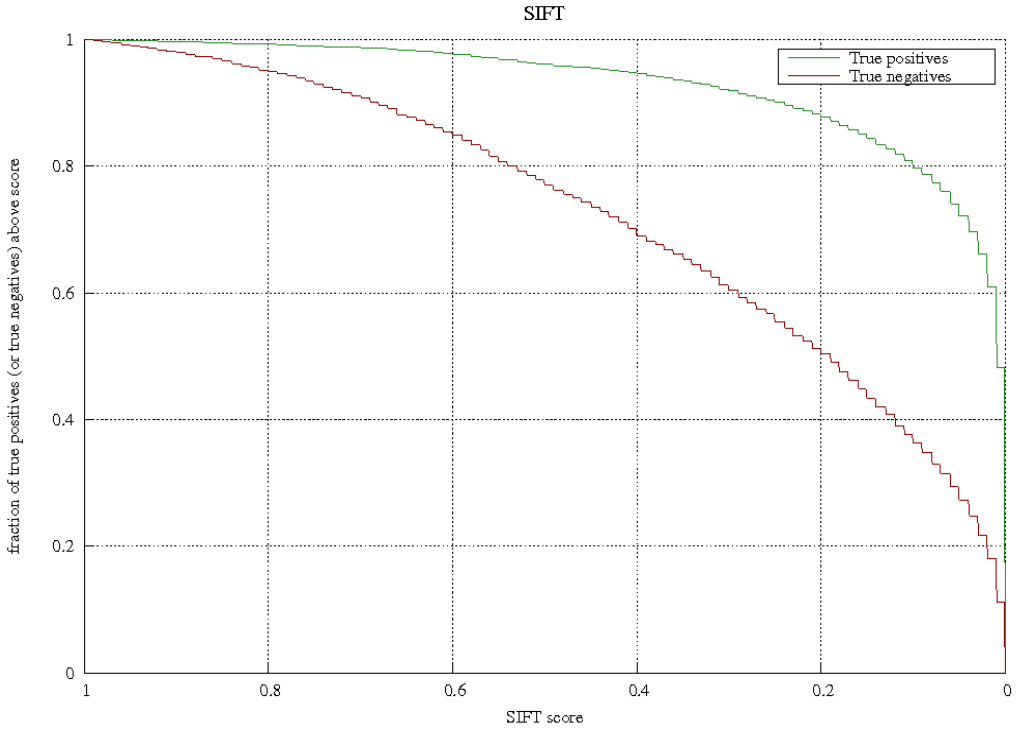
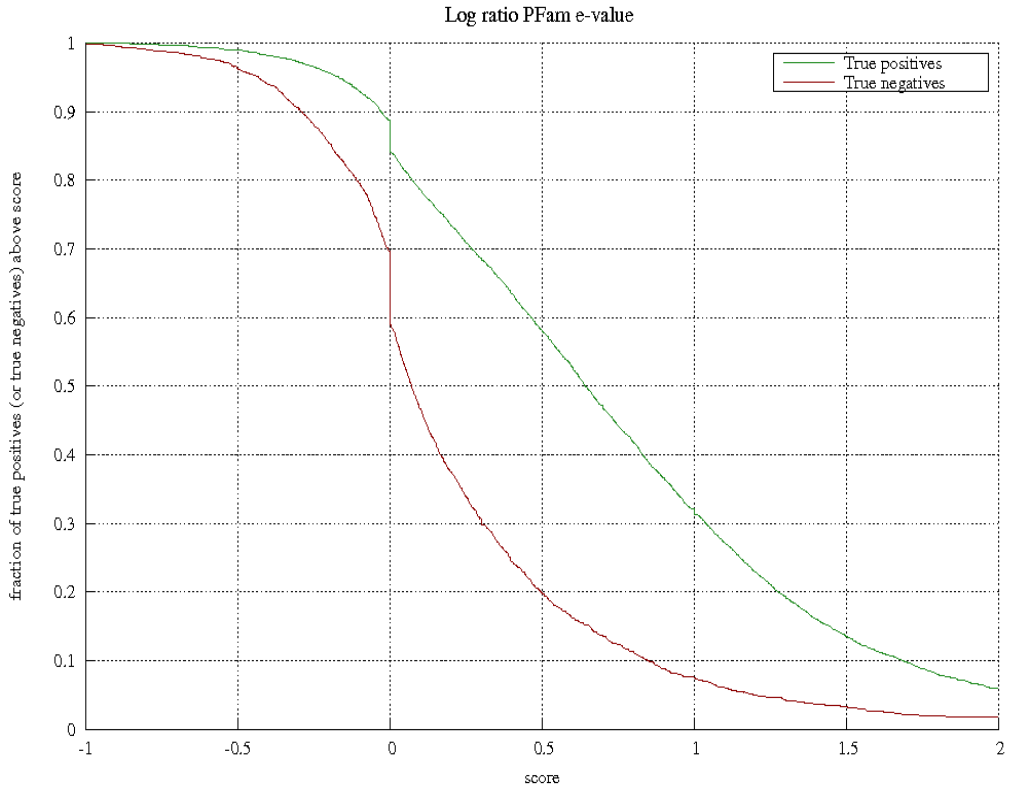
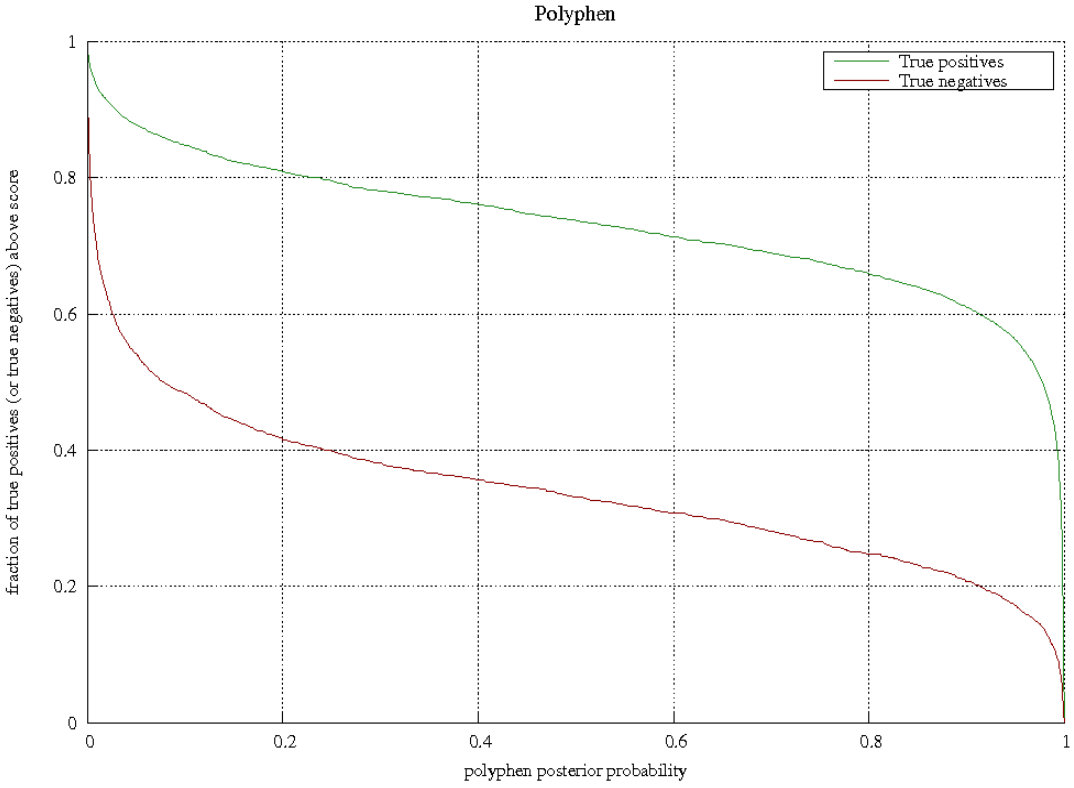
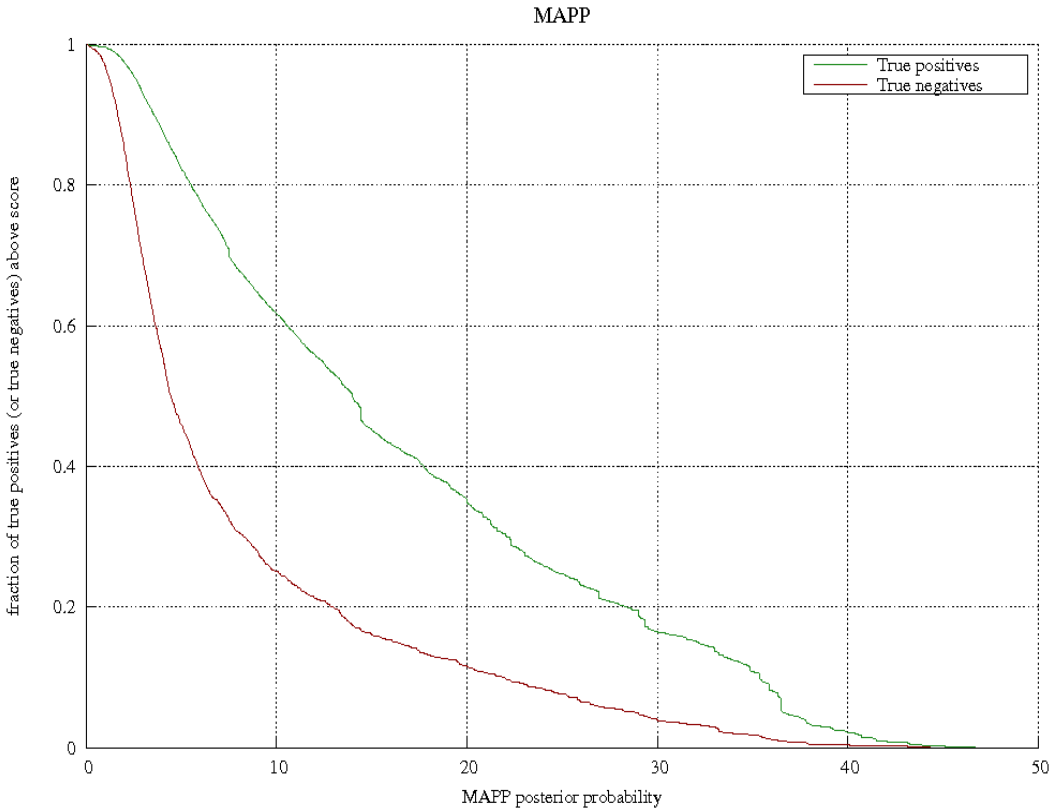


Figure S4. Complementary cumulative distributions of the scores produced by the five methods on the deleterious and neutral sets of variants of HumVar. Note that the original scores of the methods were used to compute these distributions, rather than their normalized scores.




Log ratio PFam e-value
True positives
True negatives
fraction of true positives (or true negatives) above score
score
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-1
-0.5
0
0.5
1
1.5
2


Polyphen
True positives
True negatives
fraction of true positives (or true negatives) above score
polyphen posterior probability
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1


MAPP
True positives
True negatives
fraction of true positives (or true negatives) above score
MAPP posterior probability
0
0.2
0.4
0.6
0.8
1
0
10
20
30
40
50

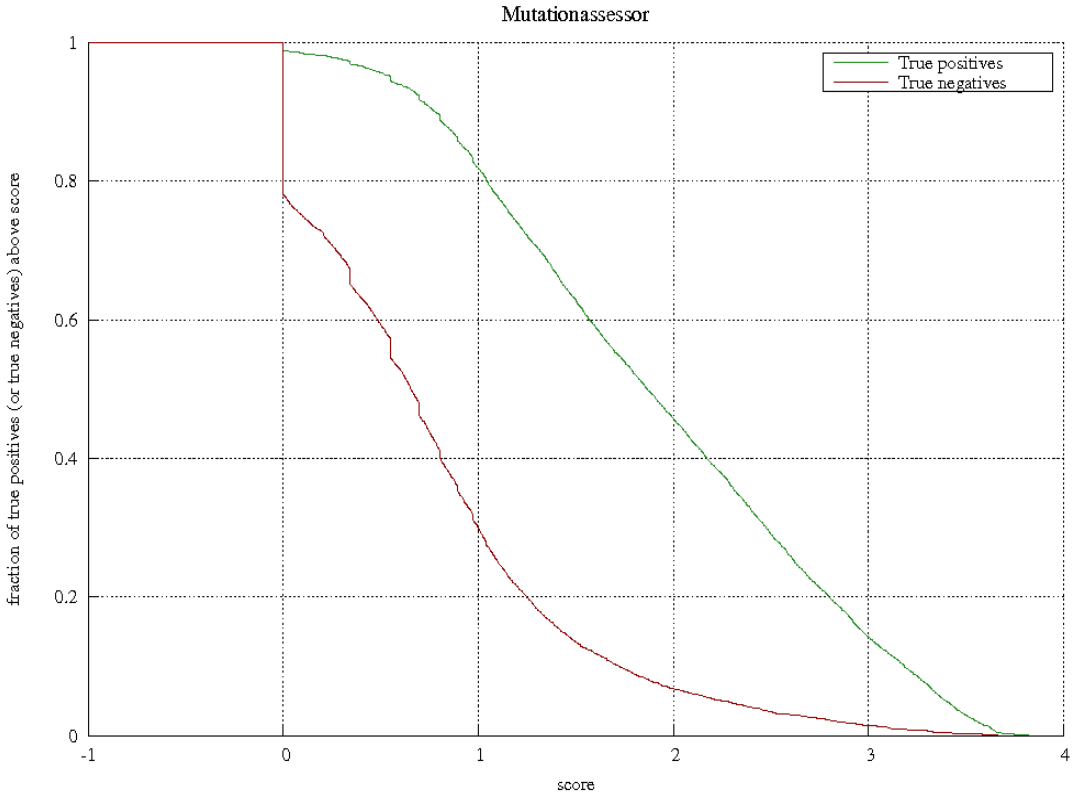

Mutationassessor
True positives
True negatives
fraction of true positives (or true negatives) above score
score
0
0.2
0.4
0.6
0.8
1
-1
0
1
2
3
4

Figure S5. ROC curve produced by the five tools and the four integrated scores with the HumDiv dataset.

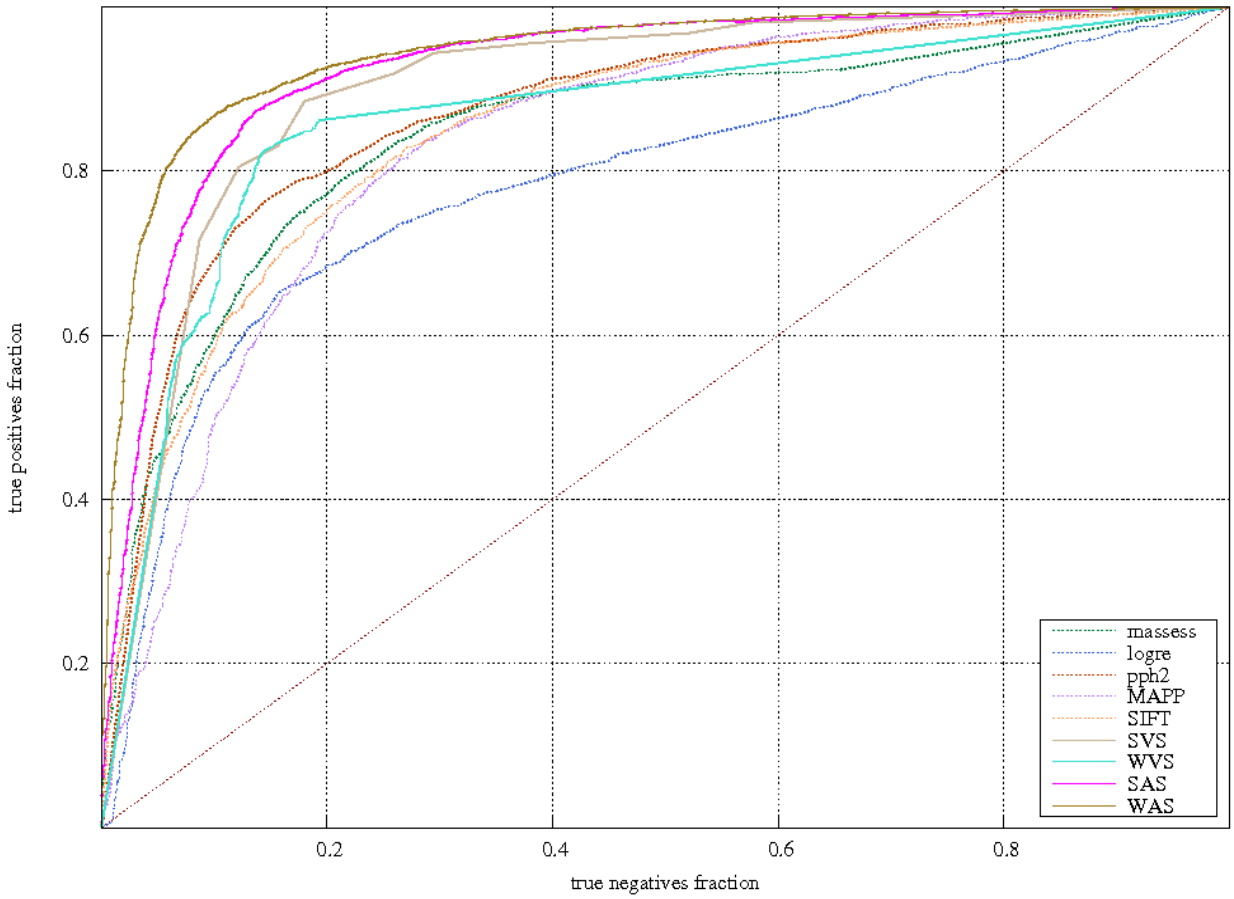
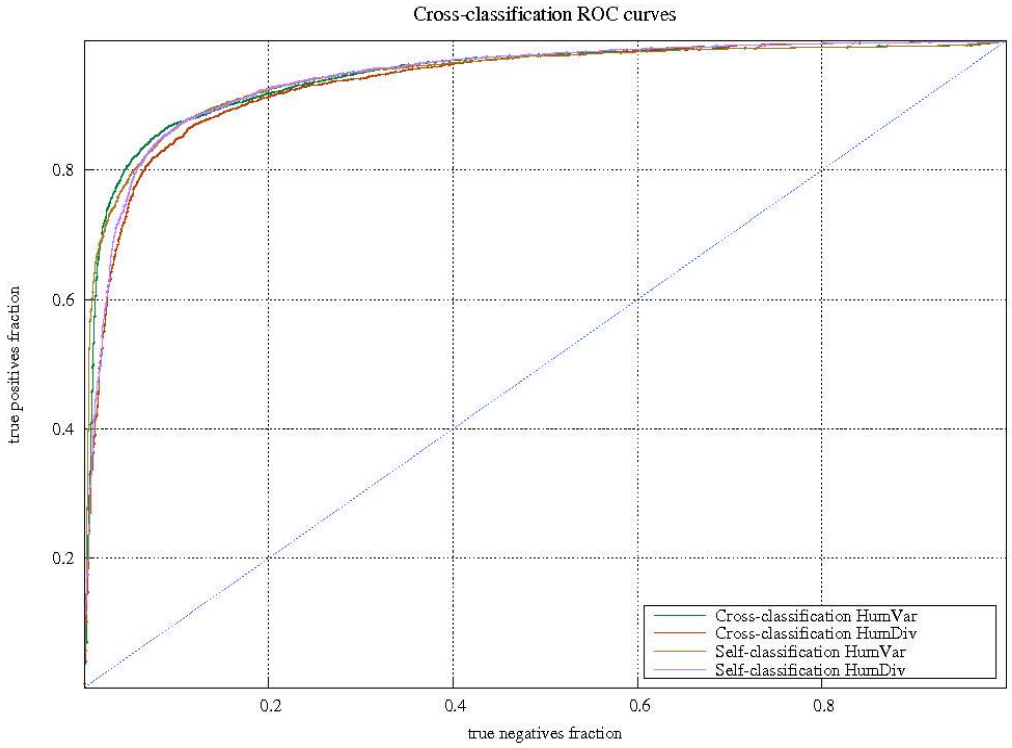


Figure S6. Accuracy with which the five tools and the four integrated scores classify the HumVar dataset.

Figure S7. Cross-validation of the WAS. Top panel presents the ROC curves resulting from classifying each dataset using the weights calculated from the tools' classification of the other dataset (cross-classification). The original self-classification ROC curves are also shown for comparison. Bottom panel presents the ROC curve resulting from performing a ten-fold cross-validation on HumVar.


Cross-classification ROC curves
true positives fraction
true negatives fraction
0.2
0.4
0.6
0.8
Cross-classification HumVar
Cross-classification HumDiv
Self-classification HumVar
Self-classification HumDiv

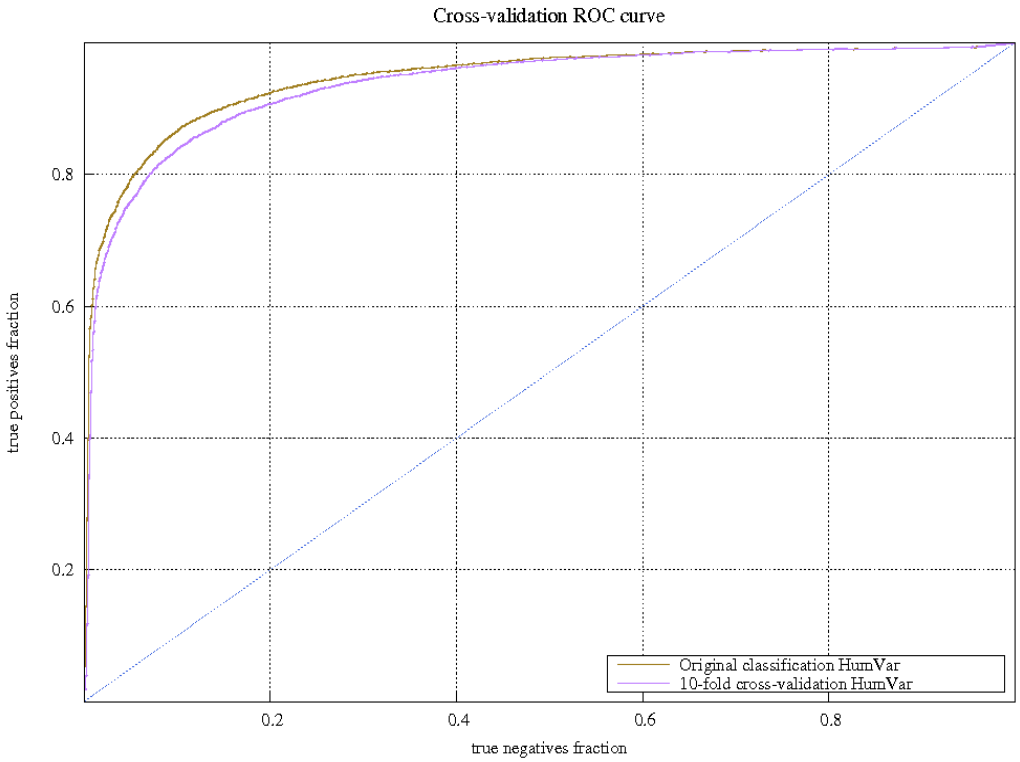

Cross-validation ROC curve
true positives fraction
true negatives fraction
0.2
0.4
0.6
0.8
Original classification HumVar
10-fold cross-validation HumVar

Figure S8. WAS of four disjoint sets of mutations from the Cosmic database compared to HumDiv neutral mutations. The five sets consist respectively, of the neutral mutations in HumDiv, the mutations appearing in only one sample (1), in two to four samples (2-4), in five to nine samples (5-9), and in ten or more samples (10+) in the Cosmic database. The points represent the mean WAS; the error bars represent the standard error of the mean. The weights were computed from the HumDiv dataset. The p-values resulting from the Wilcoxon-Mann-Whitney test of each group-group comparison are shown in the graphs. (All comparisons including neutral polymorphisms yielded p-values smaller than 10^{-318}.)

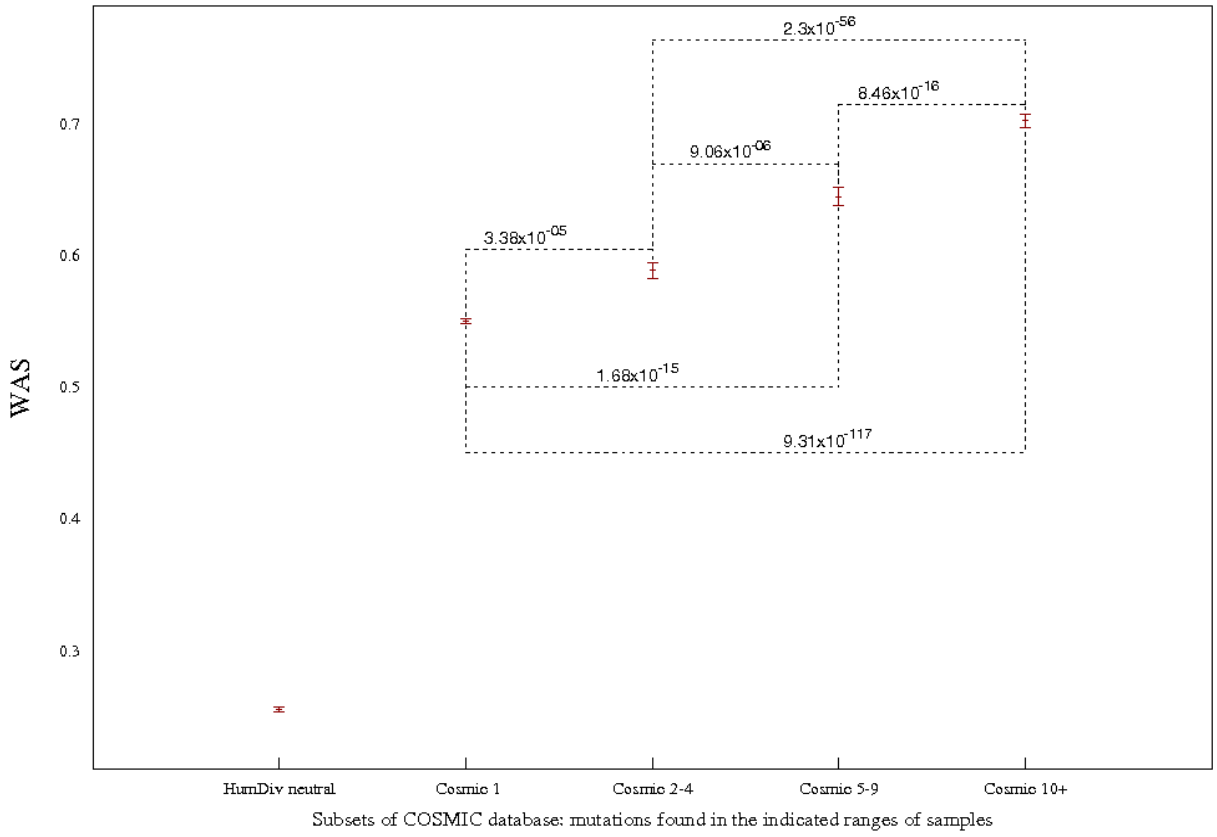


Figure S9. Venn-like diagram representing the fraction of HumVar neutral variants that are incorrectly classified by SIFT, Massessor and PPH2, and by combinations thereof.

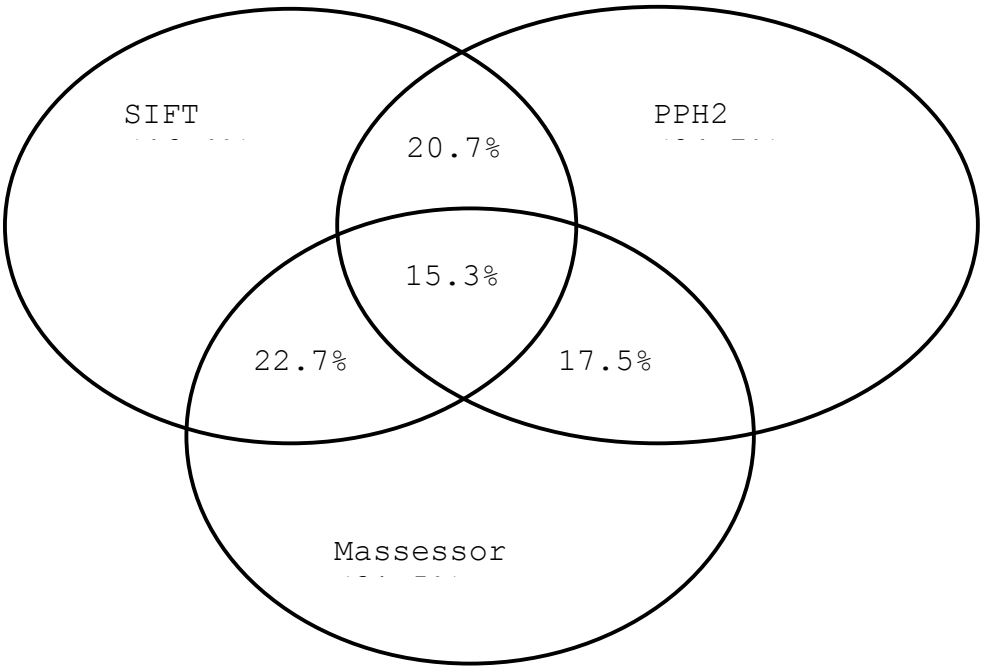


Figure S10. ROC curves produced by the WAS calculated for all variants in HumVar and HumDiv, and for the subsets that are classified by exactly 5 tools. The legend within the figure reflects the fraction of such subsets in both datasets.

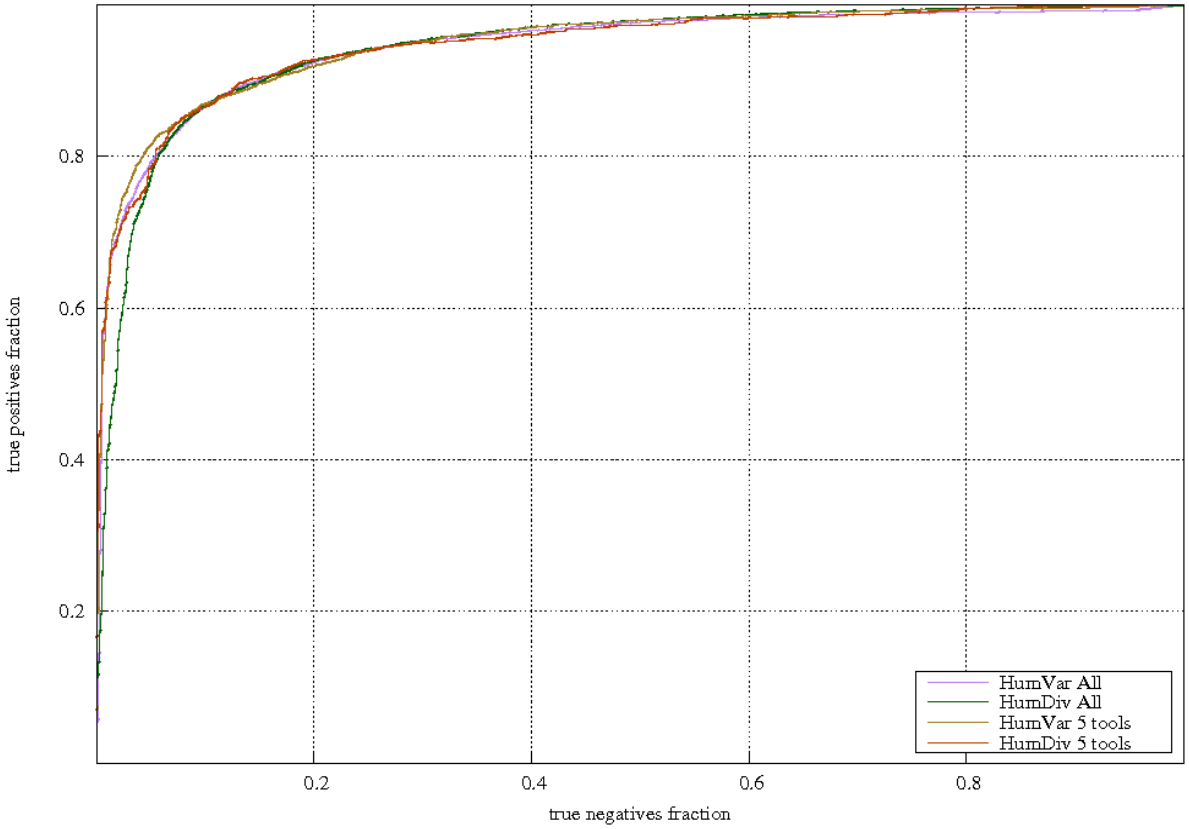


Table S1. Description of tools, and list of parameters and databases used to run them.

Parameters and versions	**SIFT**	**Logre**	**PPH2**	**MAPP**	**Massessor**
Tool version	4.0.3	N/A	polyphen-2.0.23	MAPP.jar updated 6/28/05	version 0.75 beta
Obtained from	sift.jcvi.org/www/sift4.0.3.tar.gz	N/A	genetics.bwh.harvard.edu/pph2/dokuwiki/_media/polyphen-2.0.23r349.tar.gz	mendel.stanford.edu/SidowLab/downloads/MAPP/MAPP.jar	Queried through webAPI at mutationassessor.org/
Implementation	By developers	See *Logre below	By developers	By developers Not including MSA: see *MAPP below	By developers
Input	a) Fasta file with wildtype protein sequence b) Location of the protein database to search for orthologs/paralogs c) Substitution file in the format wtaaPOSmtaa	a) Fasta file with wildtype protein sequence b) Fasta file with mutant protein sequence	a) Substitution file with 5 columns: 1, ID of the mutation; 2, Swissprot ID of the protein that bears it; 3, wtaa; 4, POS; 5, mtaa.	a) file with MSA of the protein with the mutation and a set of orthologs/paralogs b) file with phylogenetic tree of the sequences within the MSA	a) Substitution file with two columns: 1, Swissprot ID of the protein that bears the mutation; 2, mutation file with the format wtaaPOSmtaa.
Other command line arguments	median conservation observed at the mutated position – 2.75 (recommended by developers)	N/A	PSIC computation and features extraction is done first; the classifier is run after this first step is done (as recommended by developers for different datasets)	-	N/A
Protein database searched for orthologs/paralogs	Uniprotkb_swissprot downloaded on Oct. 2010 from ftp://ftp.ebi.ac.uk/pub/databases/fastafiles/uniprot/	N/A	Uniref100 downloaded on Oct. 2010 from ftp://ftp.uniprot.org/pub/databases/uniprot/uniref/ uniref100	Ensembl-compara (through its API), release 58	Internal to the tool
Program used to build MSA	Internal to the tool	N/A	Internal to the tool	Probcons v. 1.12 (recommended by	

				MAPP developers)	
Other programs/databases used		a) HMMER3 (v. 3.0) downloaded from ftp://selab.janelia.org/pub/software/hmmer3/ b) Pfam A HMMs (release 24) downloaded from ftp://ftp.sanger.ac.uk/pub/databases/Pfam/current_release/		ClustalW v. 2.0.12, used to build phylogenetic tree from sequences in MSA	
Classifier used	N/A	N/A	HumVar.UniRef100.NBd.f11.model (HumVar) HumDiv.UniRef100.NBd.f11.model (HumDiv)	N/A	N/A
Location of output	Defined through command line	Defined through command line	Defined through command line	Defined through command line	Obtained through webAPI building URL of the type http://mutationassessor.org/?cm=msa&p=SW_ID&var=aawtPOSaamt&frm=txt as defined by developers
Accuracy (HumVar/HumDiv)	72.8%/82%	69%/76.7%	74.9%/85.2%	76.4%/77.4%	77.1%/81.6%

Legend
MSA: multiple sequence alignment
wtaa: wildtype aminoacid
POS: position of the mutation in the protein sequence
mtaa: mutant aminoacid

***Logre**
Briefly, the sequences of the wildtype and mutant proteins are aligned to the HMM representing the domain where the mutation is located. Then, the Logre score is calculated as $\log_{10}$ ($E\text{-}value_{mutant}$ /$E\text{-}value_{wildtype}$) following the description of the algorithm from the paper by Clifford *et al.*, 2004.

***MAPP**

The MAPP does not build a multiple alignment on its own. Instead, it receives a multiple sequence alignment of the protein that contains the mutation and its orthologs and paralogs, along with a phylogentic tree. We automated the search for orthologs and paralogs of mutation bearing proteins through the Ensembl-compara API and the building of MSAs and phylogenetic trees using Probcons and ClustalW.