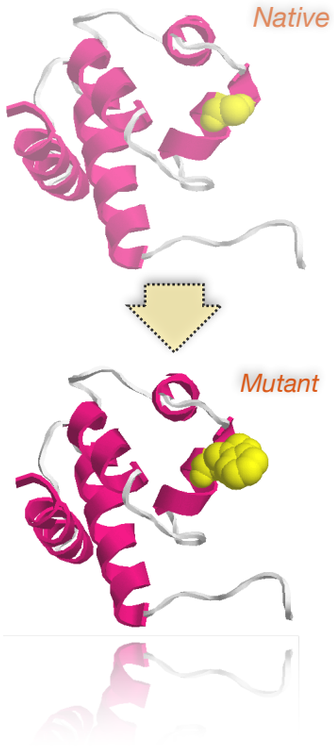
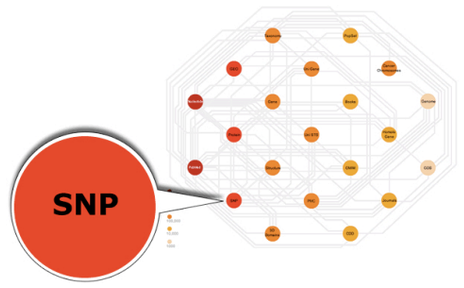
Computational methods for predicting the impact of the genetic variants

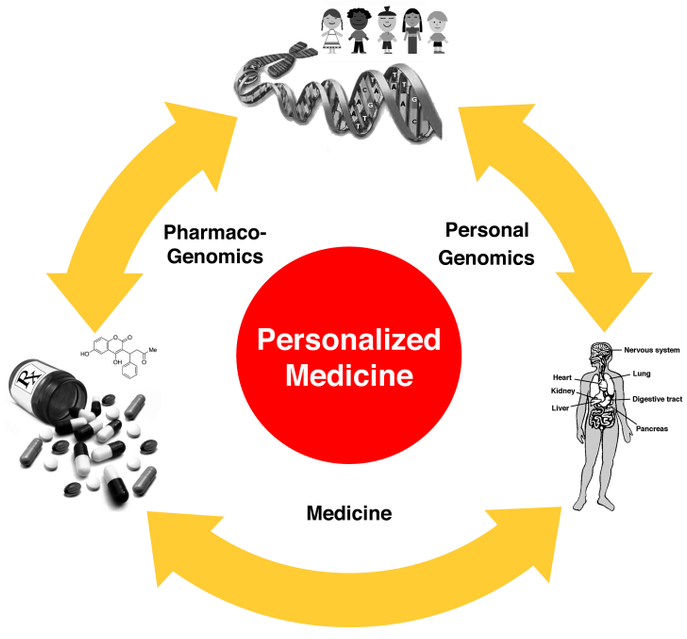


University of Lisboa (Portugal)
July 17, 2019



OMIM
Online Mendelian Inheritance in Man

Johns Hopkins University



Emidio Capriotti

http://biofold.org/

Biomolecules Folding and Disease

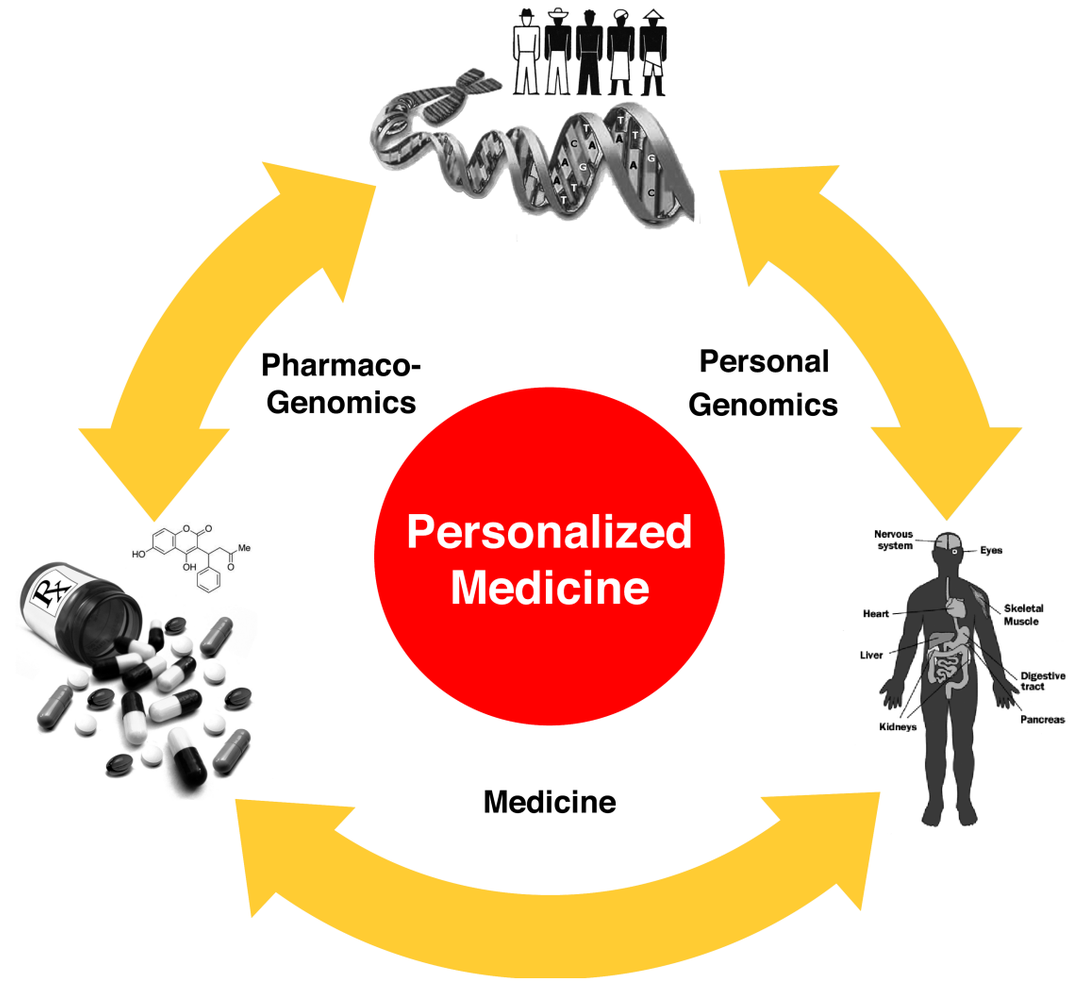
Department of Pharmacy and Biotechnology (FaBiT)
University of Bologna

Personalized medicine

Currently direct to consumers company are performing genotype test on markers associated to genetic traits, and soon full genome sequencing will cost ~$1000.

The future bioinformatics challenges for personalized medicine will be:

1. Processing Large-Scale Robust Genomic Data
2. Interpretation of the Functional Effect and the Impact of Genomic Variation
3. Integrating Systems and Data to Capture Complexity
4. Making it all clinically relevant



Fernald GH, et al (2011). Bioinformatics. 27: 1741-1748.

Single Nucleotide Variants

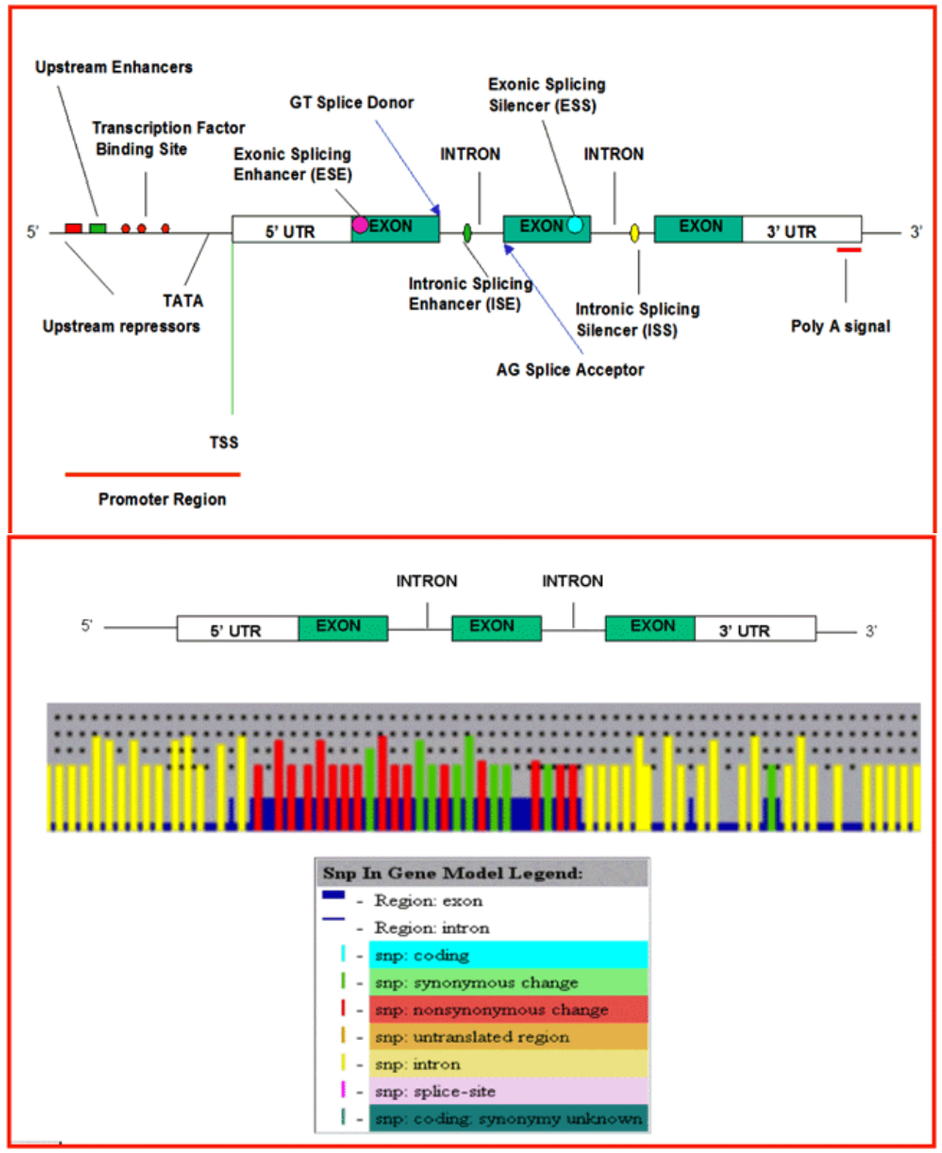
Single Nucleotide Variants (SNVs)
is a DNA sequence variation occurring when a single nucleotide A, T, C, or G in the genome differs between members of the species.
It is used to refer to Polymorphisms when the population frequency is ≥ 1%

SNVs occur at any position and can be classified on the base of their locations.

Coding SNVs can be subdivided into two groups:

Synonymous: when single base substitutions do not cause a change in the resultant amino acid

Non-synonymous or Single Amino Acid Variants (SAVs): when single base substitutions cause a change in the resultant amino acid.



http://www.ncbi.nlm.nih.gov

1000 Genomes

The 1000 Genomes Project aims to create the largest public catalogue of human variations and genotype data. Last version released the genotype of ~2,500 individuals.

Table 1 | Variants discovered by project, type, population and novelty

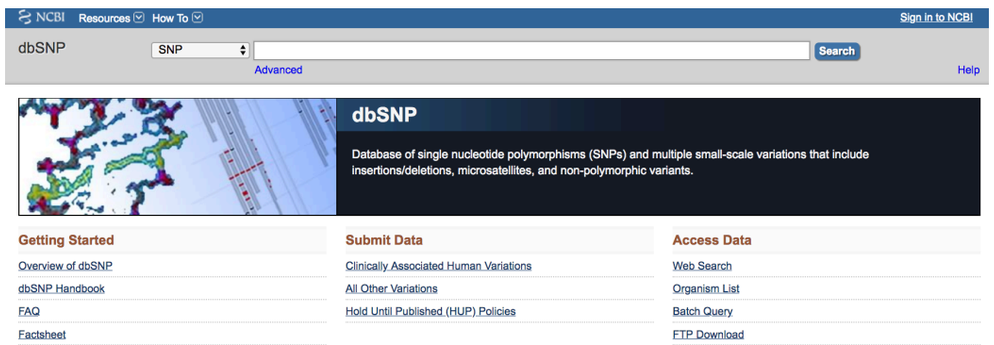
a Summary of project data including combined exon populations

	Low coverage				Trios				
Statistic	CEU	YRI	CHB+JPT	Total	CEU	YRI	Total	Exon (total)	Union across projects
Samples	60	59	60	179	3	3	6	697	742
Total raw bases (Gb)	1,402	874	596	2,872	560	615	1,175	845	4,892
Total mapped bases (Gb)	817	596	468	1,881	369	342	711	56	2,648
Mean mapped depth (×)	4.62	3.42	2.65	3.56	43.14	40.05	41.60	55.92	NA
Bases accessed (% of genome)	2.43 Gb (86%)	2.39 Gb (85%)	2.41 Gb (85%)	2.42 Gb (86.0%)	2.26 Gb (79%)	2.21 Gb (78%)	2.24 Gb (79%)	1.4 Mb	NA
No. of SNPs (% novel)	7,943,827 (33%)	10,938,130 (47%)	6,273,441 (28%)	14,894,361 (54%)	3,646,764 (11%)	4,502,439 (23%)	5,907,699 (24%)	12,758 (70%)	15,275,256 (55%)
Mean variant SNP sites per individual	2,918,623	3,335,795	2,810,573	3,019,909	2,741,276	3,261,036	3,001,156	763	NA
No. of indels (% novel)	728,075 (39%)	941,567 (52%)	666,639 (39%)	1,330,158 (57%)	411,611 (25%)	502,462 (37%)	682,148 (38%)	96 (74%)	1,480,877 (57%)
Mean variant indel sites per individual	354,767	383,200	347,400	361,669	322,078	382,869	352,474	3	NA
No. of deletions (% novel)	ND	ND	ND	15,893 (60%)	6,593 (41%)	8,129 (50%)	11,248 (51%)	ND	22,025 (61%)
No. of genotyped deletions (% novel)	ND	ND	ND	10,742 (57%)	ND	ND	6,317 (48%)	ND	13,826 (58%)
No. of duplications (% novel)	259 (90%)	320 (90%)	280 (91%)	407 (89%)	187 (93%)	192 (91%)	256 (92%)	ND	501 (89%)
No. of mobile element insertions (% novel)	3,202 (79%)	3,105 (84%)	1,952 (76%)	4,775 (86%)	1,397 (68%)	1,846 (78%)	2,531 (78%)	ND	5,370 (87%)
No. of novel sequence insertions (% novel)	ND	ND	ND	ND	111 (96%)	66 (86%)	174 (93%)	ND	174 (93%)

1000 Genomes Project Consortium (2010). Nature. 467: 1061-1073.

SNVs and SAVs databases

dbSNP (Mar 2018) @ NCBI



http://www.ncbi.nlm.nih.gov/snp

Single Nucleotide Variants

Homo sapiens	**113,862,023**
Gallus gallus	15,104,956
Zea mays	14,672,946

SwissVar (Oct 2018) @ ExPASy

SIB

swissvar

http://www.expasy.ch/swissvar/

Single Amino acid Variants

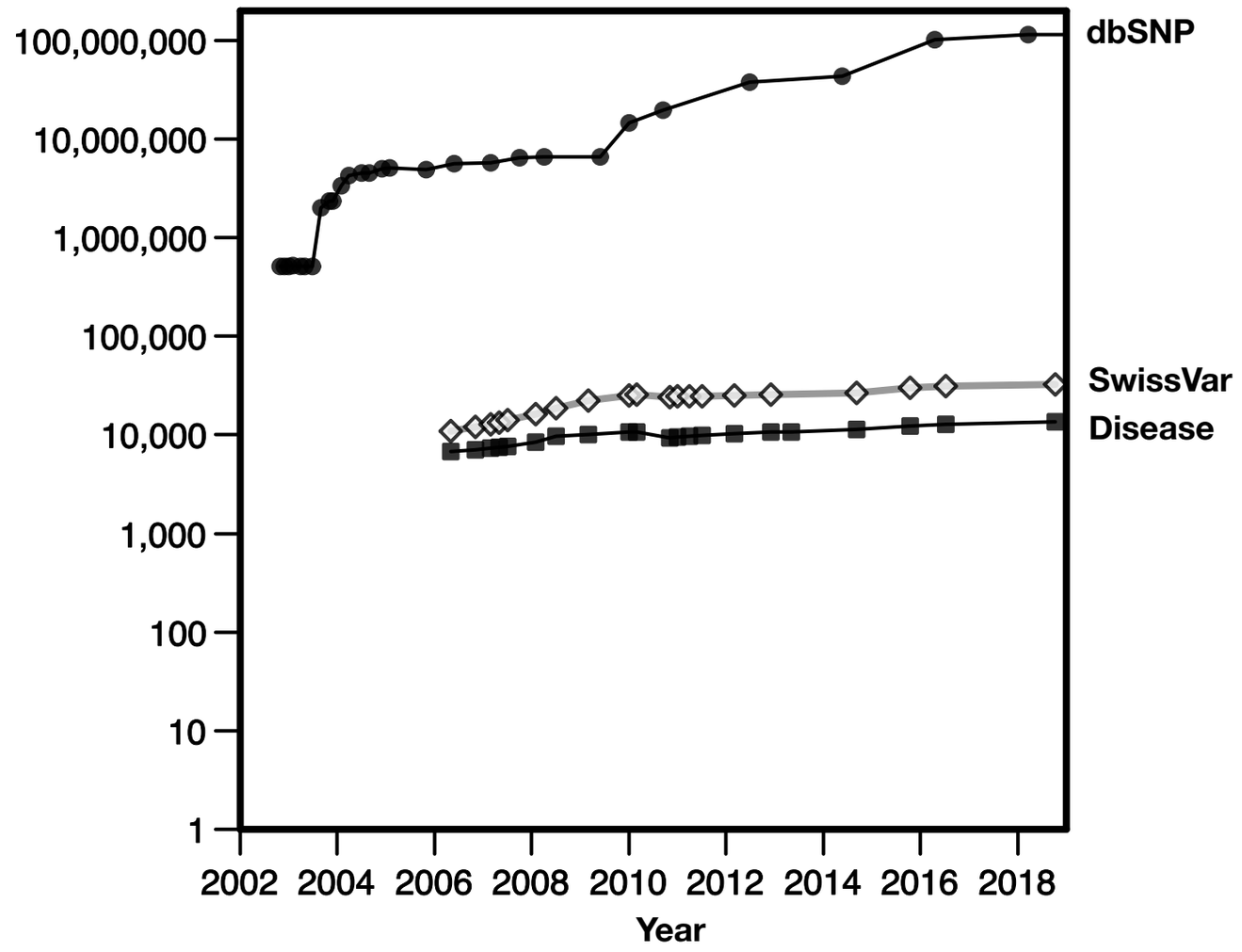
Homo sapiens	76,608
Disease	*29,529*
Polymorphisms	*39,779*

Oct 2018

SNVs and Disease

Single Nucleotide Variants (SNVs) are the most common type of genetic variations in human accounting for more than 90% of sequence differences (1000 Genome Project Consortium, 2012).

SNVs can also be responsible of genetic diseases (Ng and Henikoff, 2002; Bell, 2004).



nonsynonymous SNVs

- disease-related SNVs → the mutations are related to a Mendelian pathologies
- neutral SNVs → the mutations do not compromise the organism's health

Effects of variants

It is important to understand the functional effect of Single Nucleotide Polymorphisms (SNPs) that are very common type of variations, but also the impact rare variants which have allele frequencies below than 1%

Impact of coding variants

- Properties of amino acid residue substitution
- The evolutionary history of an amino acid position
- Sequence–function relationships
- Structure–function relationships

Impact of non-coding variants

- Transcription
- Pre-mRNA splicing
- MicroRNA binding
- Altering post-translational modification sites

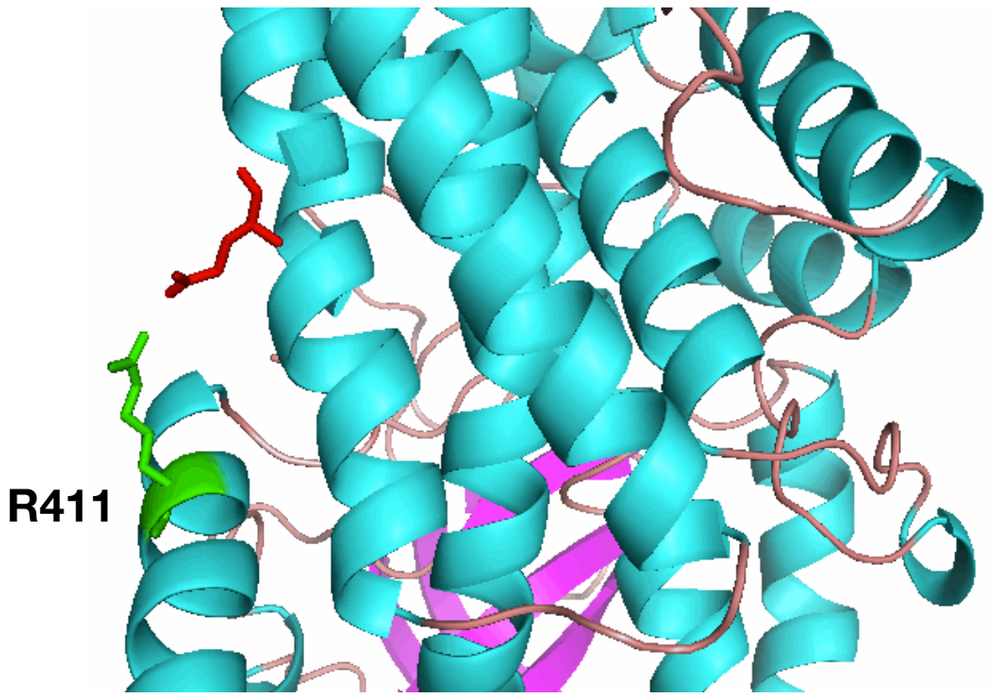
Cline and Karchin (2011) Bioinformatics, 27; 441-448.

Protein variants

Sequence, Structure & Function

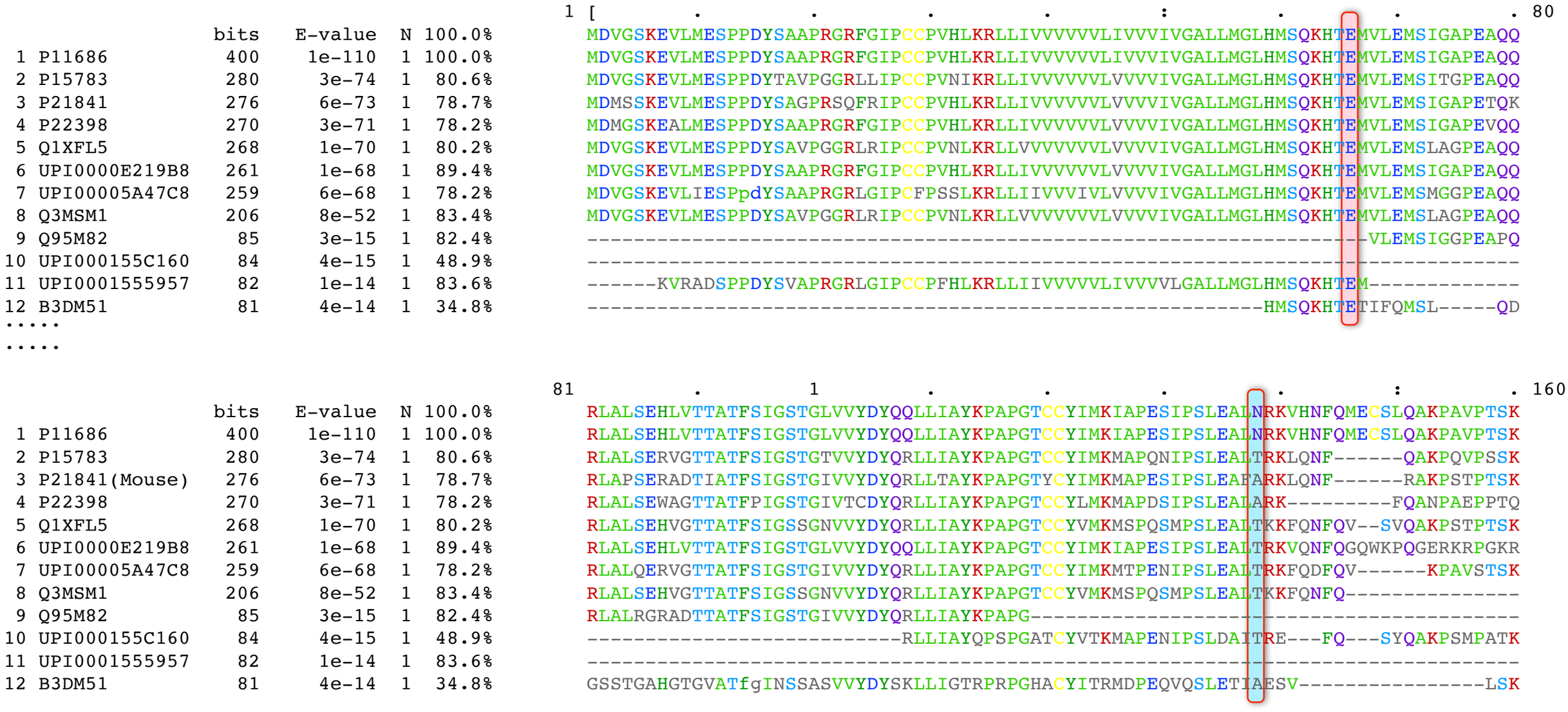
Genomic variants in sequence motifs could affect protein function.
Mutation S362A of P53 affect the interaction with hydrolase USP7 and the deubiquitination of the protein.

Nonsynonymous variants responsible for protein structural changes and cause loss of stability of the folded protein.
Mutation R411L removes the salt bridge stabilizing the structure of the IVD dehydrogenase.



Conserved or not?

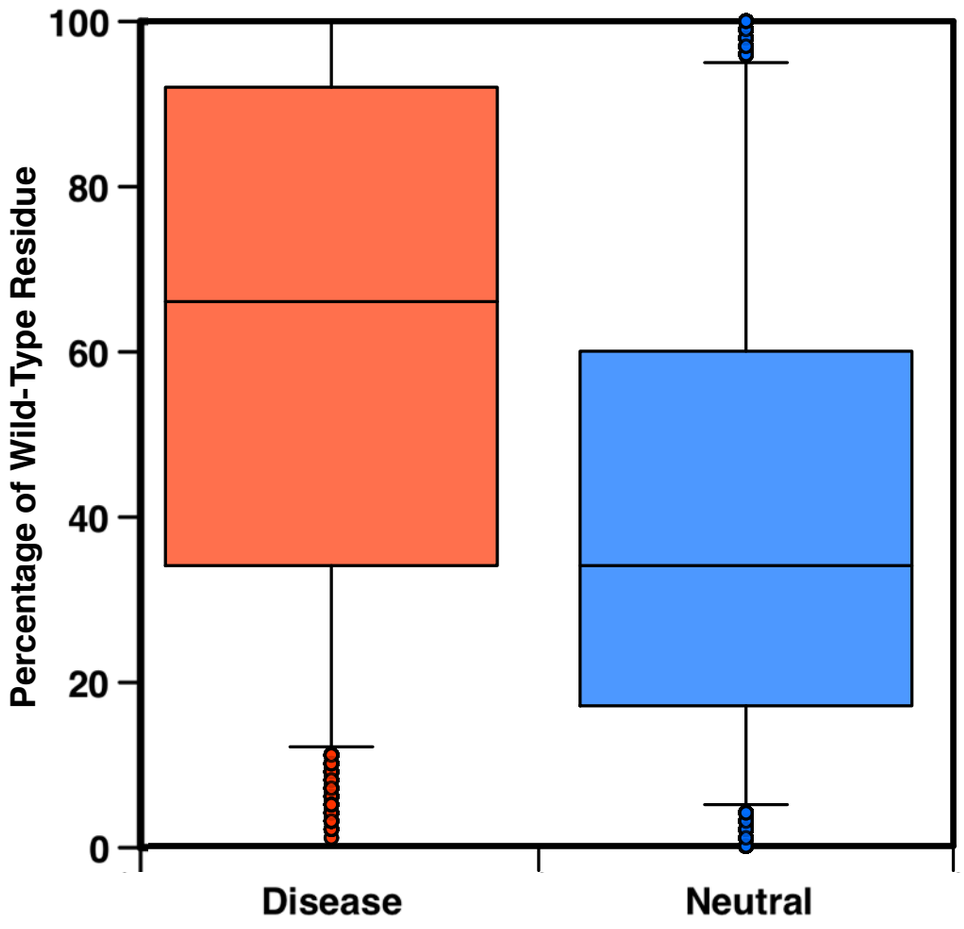
In positions 66 the Glutamic acid is highly conserved Asparagine in position 138 is mutated Threonine or Alanine



Sequence profile

The protein sequence profile is calculated running BLAST on the UniRef90 dataset and selecting only the hits with e-value < 10^{-9}.

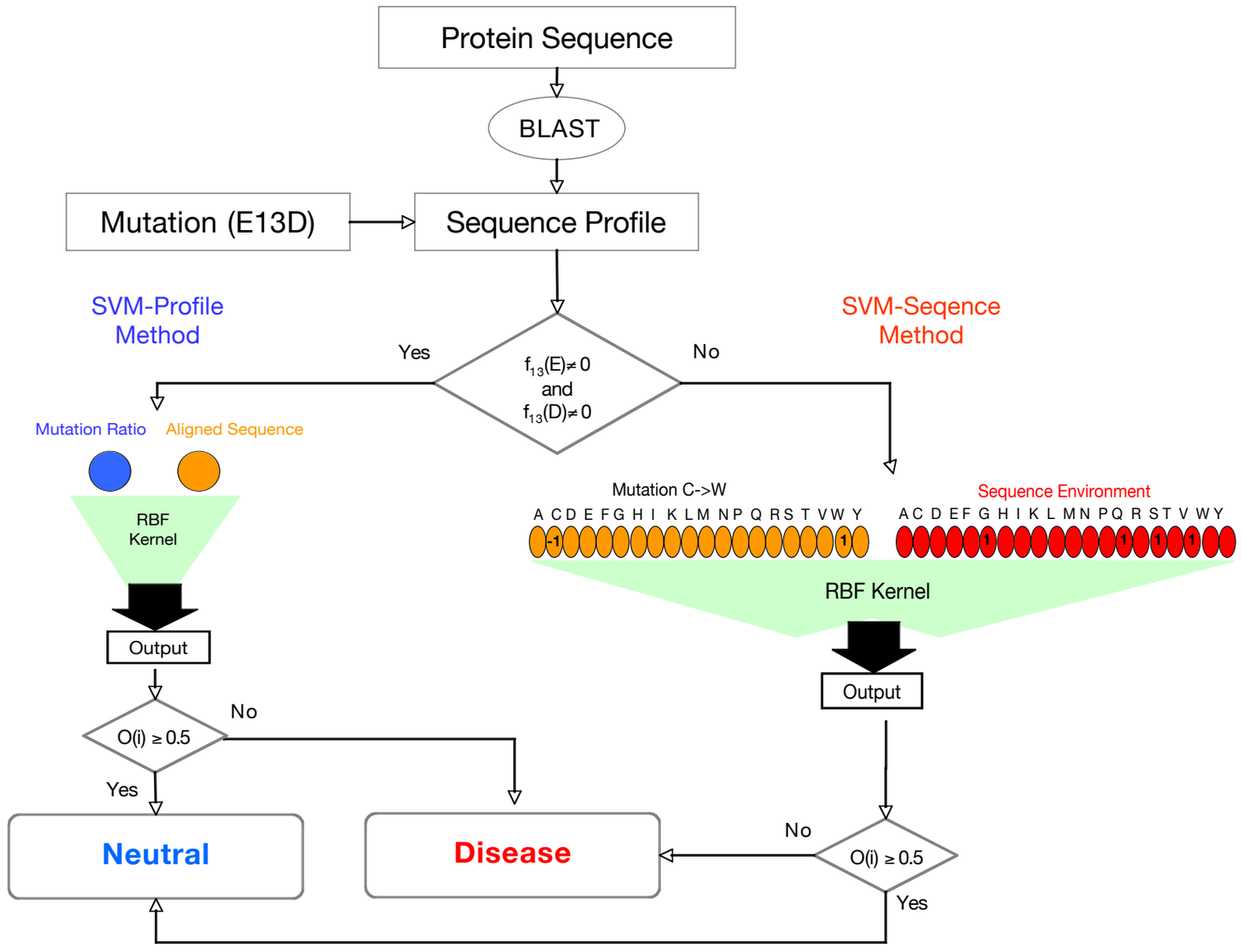
The frequency distributions of the wild-type residues for disease-related and neutral variants are significantly different (KS p-value=0).



Capriotti et al (2012). Briefings in Bioinformatics. 13; 495-512.

Hybrid method structure

Hybrid Method is based on a decision tree with SVM-Sequence coupled to SVM-Profile. Tested on more than 21,000 variants our method reaches 74% of accuracy and 0.46 correlation coefficient.



Capriotti et al. (2006) Bioinformatics, 22; 2729-2734.

Classification results

SVM–Sequence is more accurate in the prediction of disease related mutations and SVM-Profile is more accurate in the prediction of neutral polymorphism.
Both methods have the same Q2 level.

	Q2	P[D]	Q[D]	P[N]	Q[N]	C
SVM-Sequence	0.70	0.71	0.84	0.65	0.46	0.34
SVM-Profile	0.70	0.74	0.49	0.68	0.86	0.39
HybridMeth	0.74	0.80	0.76	0.65	0.70	0.46

D = Disease related N = Neutral

The Hybrid Method have higher accuracy than the previous two methods increasing the accuracy up to 74% and the correlation coefficient up to 0.46.

http://snps.biofold.org/phd-snp

Gene Ontology

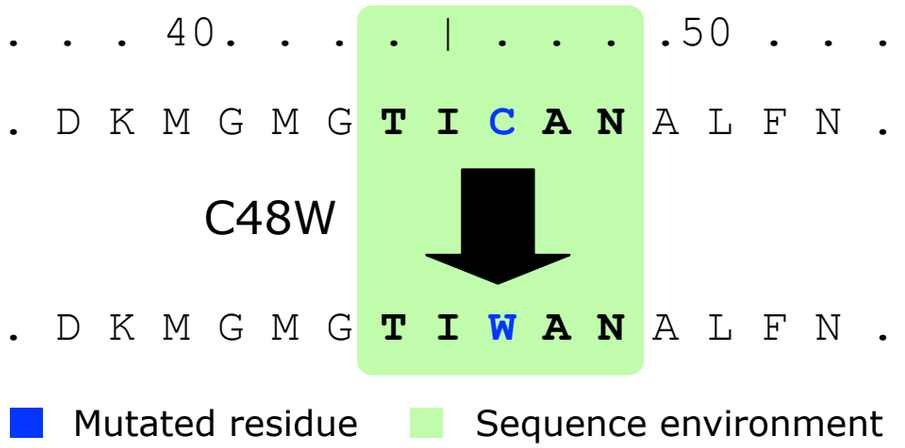
The Gene Ontology project is a major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases. The project provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data.

http://www.geneontology.org/

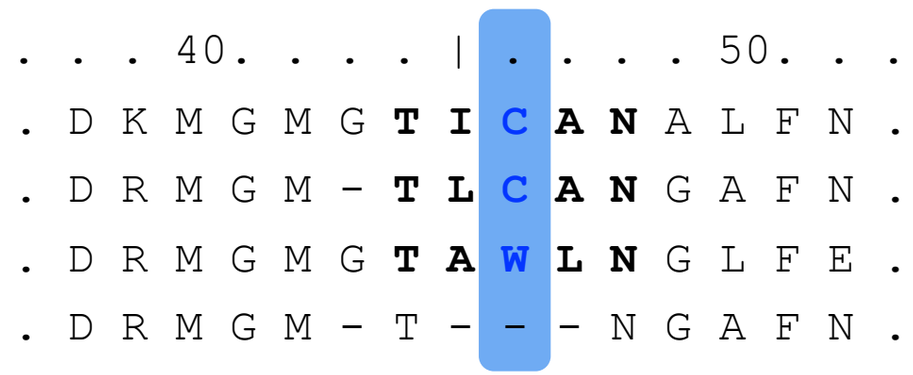
The ontology is represented by a direct acyclic graph covers three domains;

- cellular component, the parts of a cell or its extracellular environment;
- molecular function, the elemental activities of a gene product at the molecular level, such as binding or catalysis
- biological process, operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells,tissues, organs and organisms.

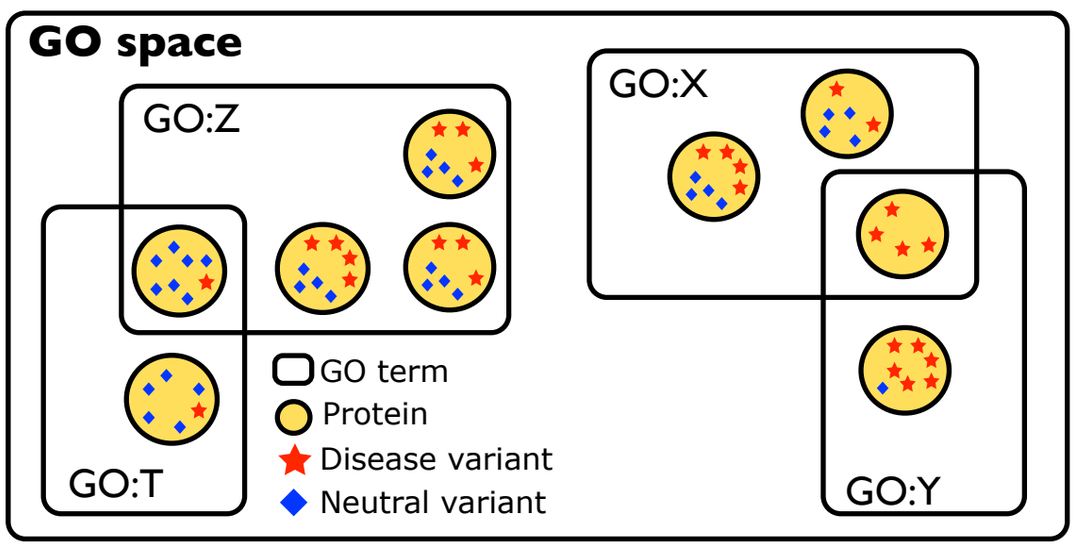
SNPs&GO input features



Sequence information is encoded in 2 vectors each one composed by 20 elements. The first vector encodes for the mutation and the second one for the sequence environment



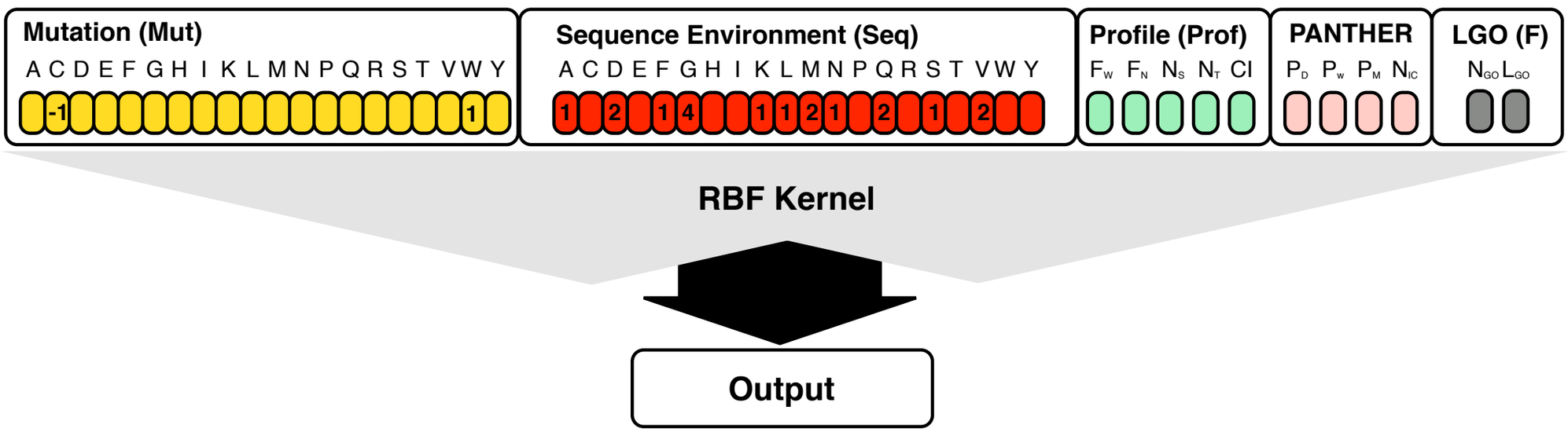
Protein sequence profile information derived from a multiple sequence alignment. It is encoded in a 5 elements vector corresponding to different features general and local features



The GO information are encoded in a 2 elements vector corresponding to the number unique of GO terms associated to the protein sequences and the sum of the logarithm of the total number of disease-related and neutral variants for each GO term.

SNPs&GO performance

SNPs&GO results in better performance with respect to previously developed methods.



Method	Q2	P[D]	Q[D]	P[N]	Q[N]	C	PM
PolyPhen	0.71	0.76	0.75	0.63	0.64	0.39	58
SIFT	0.76	0.75	0.76	0.77	0.75	0.52	93
PANTHER	0.74	0.77	0.73	0.71	0.76	0.48	76
SNPs&GO	0.82	0.83	0.78	0.80	0.85	0.63	100

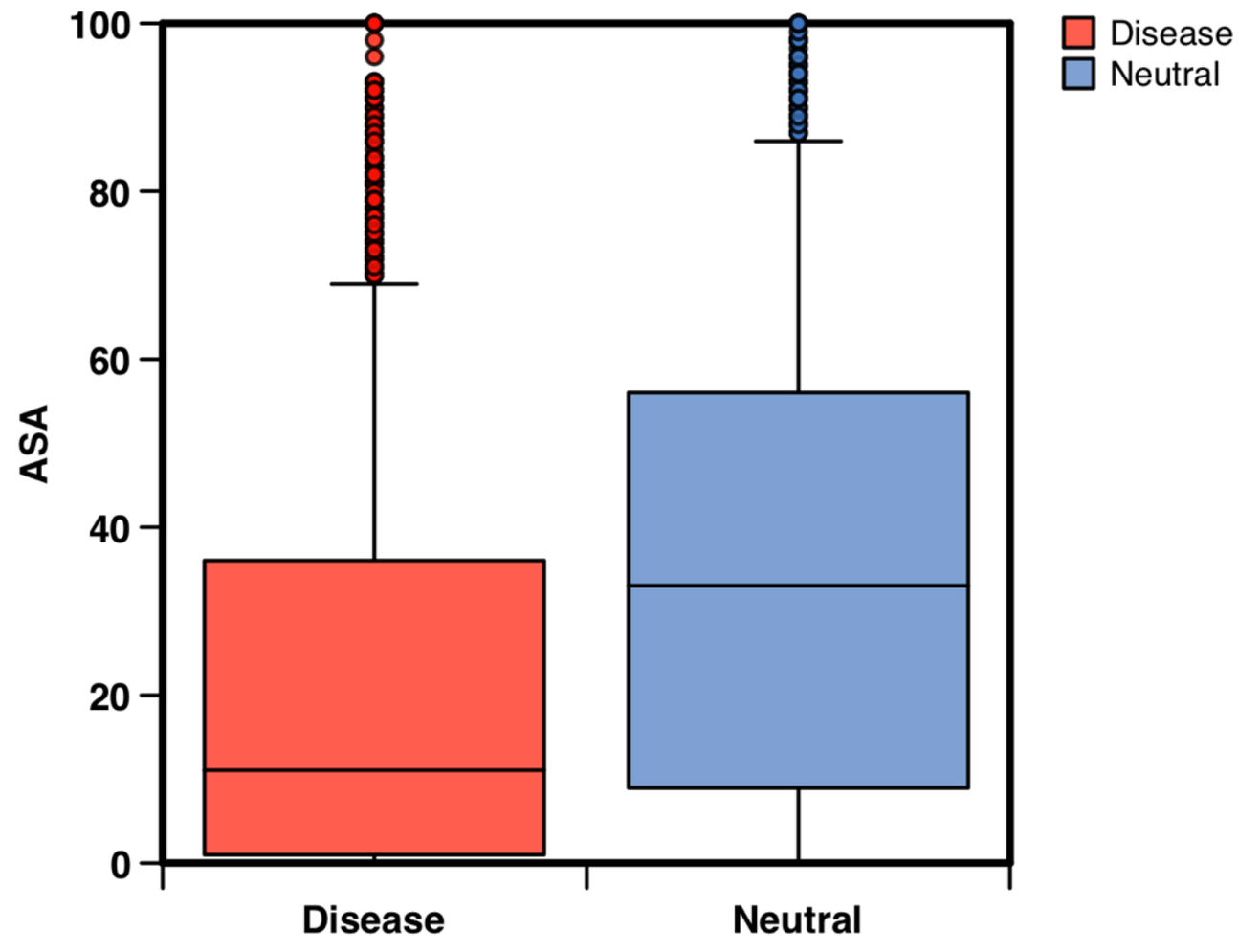
D = Disease related N = Neutral

DB= 33672 nsSNVs

Calabrese et al. (2009) Human Mutation 30, 1237-1244.

Structure environment

There is a significant difference (KS p-value = 2.8×10^{-71}) between the distributions of the Relative Accessible Solvent Area for disease-related and neutral variants. Their mean values are respectively 20.6 and 35.7.

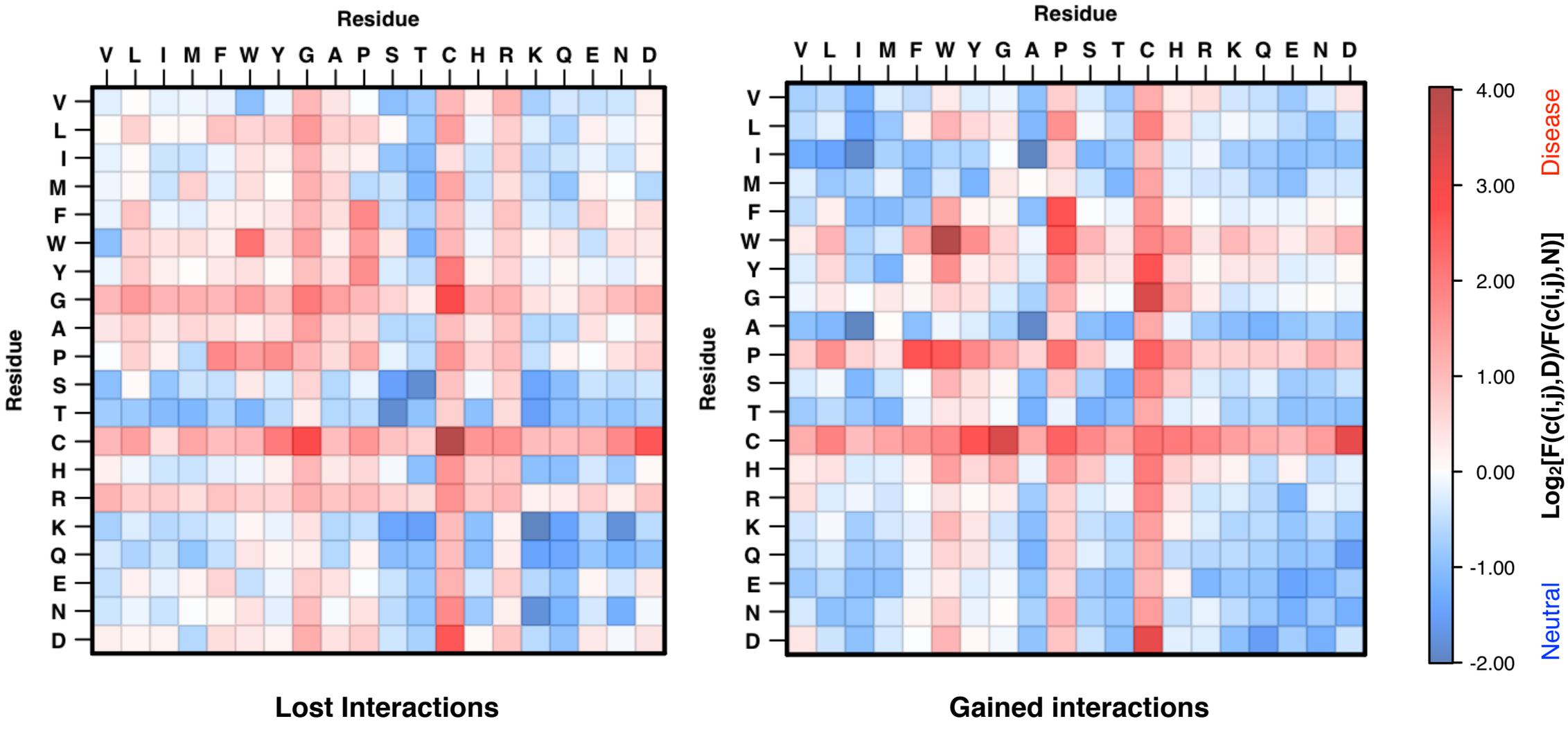


Capriotti and Altman. (2011) BMC Bioinformatics.12 (Suppl 4); S3.

Analysis of the 3D interactions

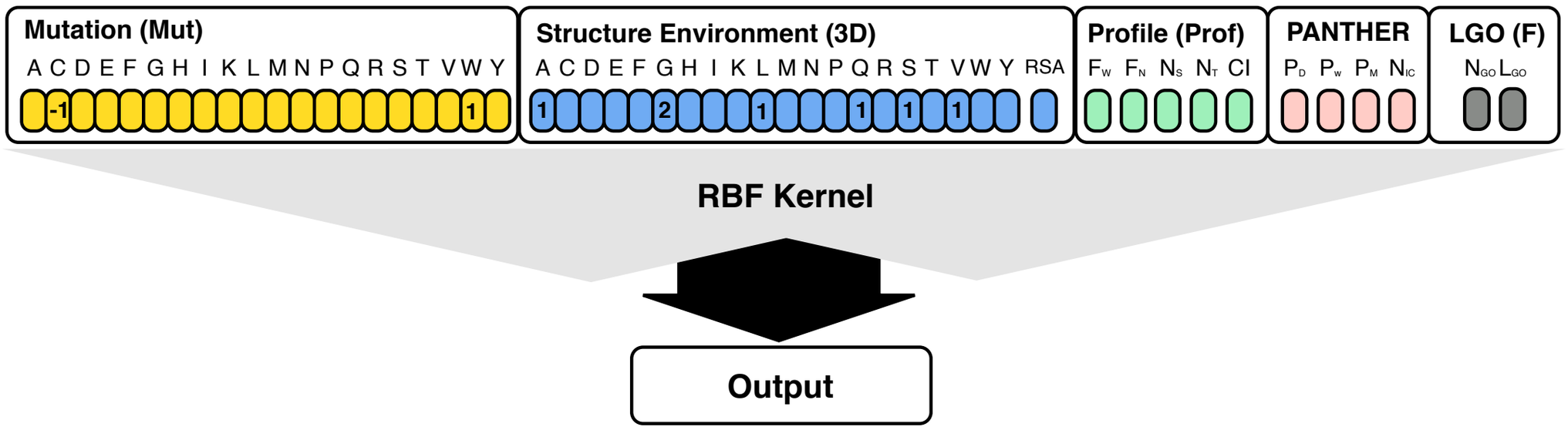
Using the whole set of SAVs with known structure, we calculate the log odd score of the ratio between the frequencies of the interaction between residue i and j for disease-related and neural variants.

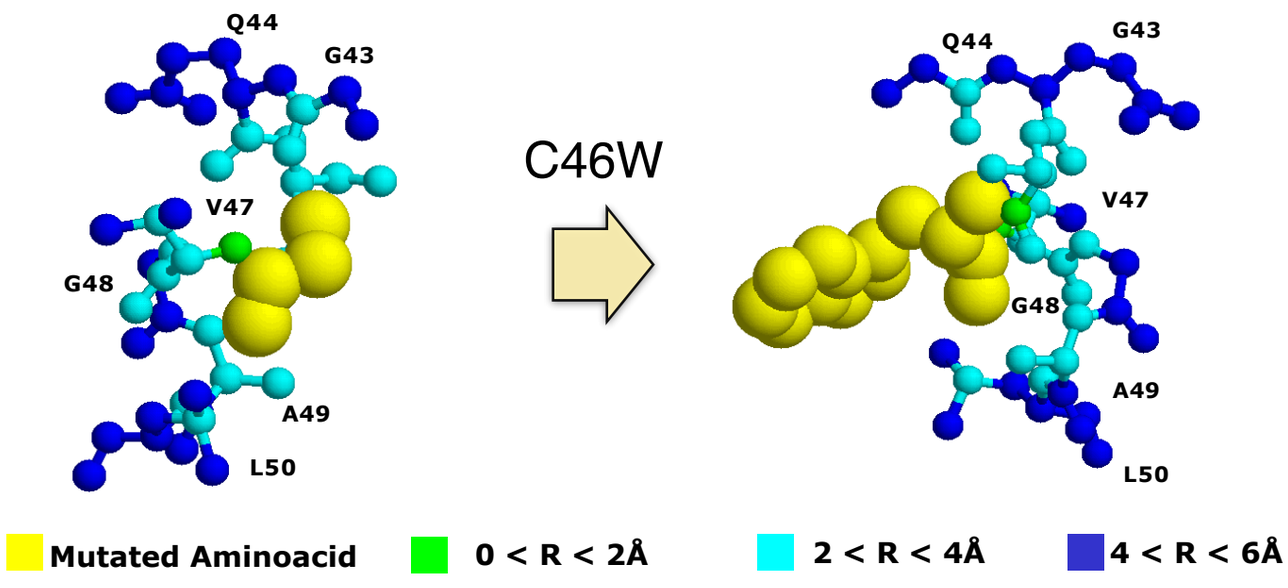
$$LC = \log_2 \left[\frac{n(i,j,Disease)/N(Disease)}{n(i,j,Neutral)/N(Neutral)} \right]$$



The structure-based method

The method takes in to input 4 types of information encoded in a 48 elements vector.
The input features are: mutation data; structure environment, sequence profile and functional score based on GO terms.

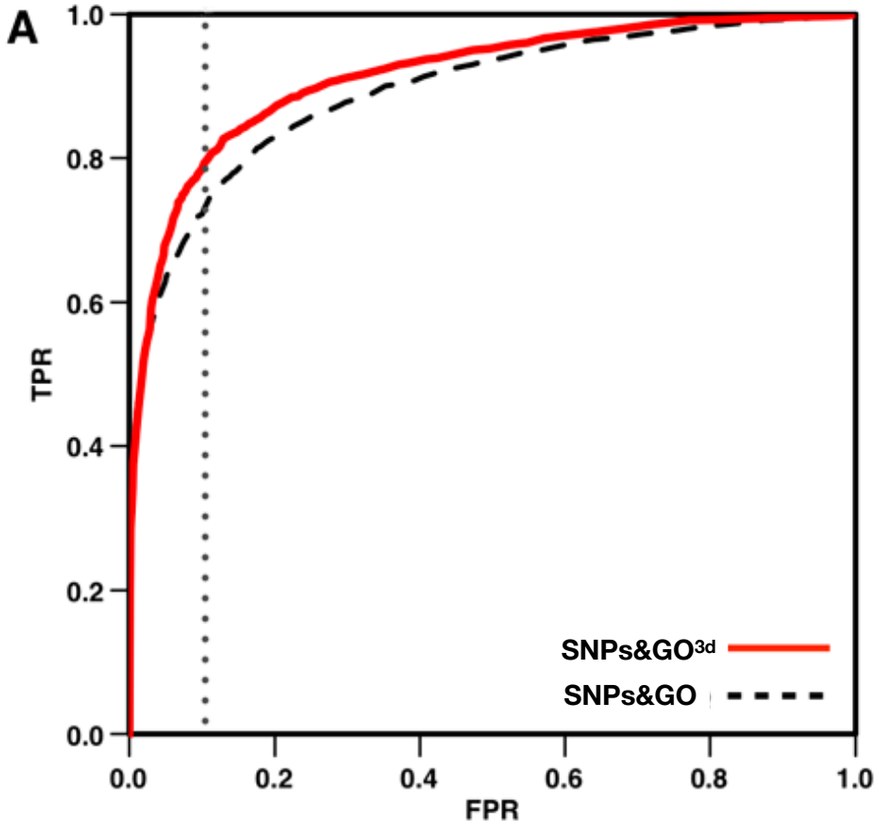


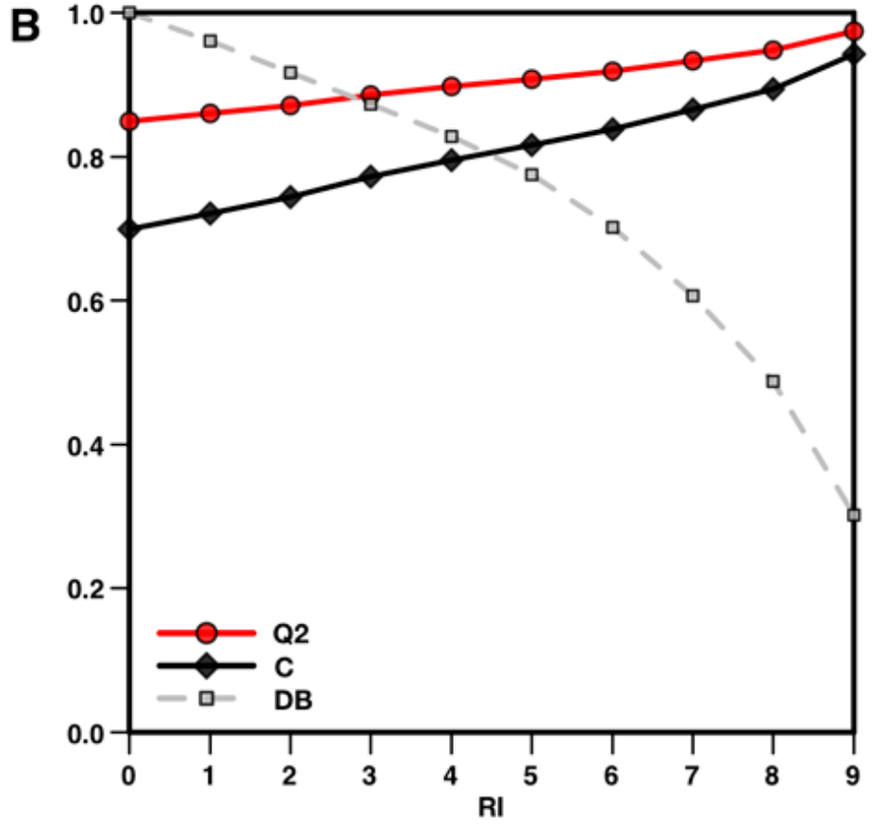


Sequence vs Structure

The structure-based method results in better accuracy with respect to the sequence-based one. Structure based prediction are 3% more accurate and correlation coefficient increases of 0.06. If 10% of FP are accepted the TPR increases of 7%.

	Q2	P[D]	S[D]	P[N]	S[N]	C	AUC
SNPs&GO	0.82	0.81	0.83	0.82	0.81	0.64	0.89
SNPs&GO^{3d}	0.85	0.84	0.87	0.86	0.83	0.70	0.92



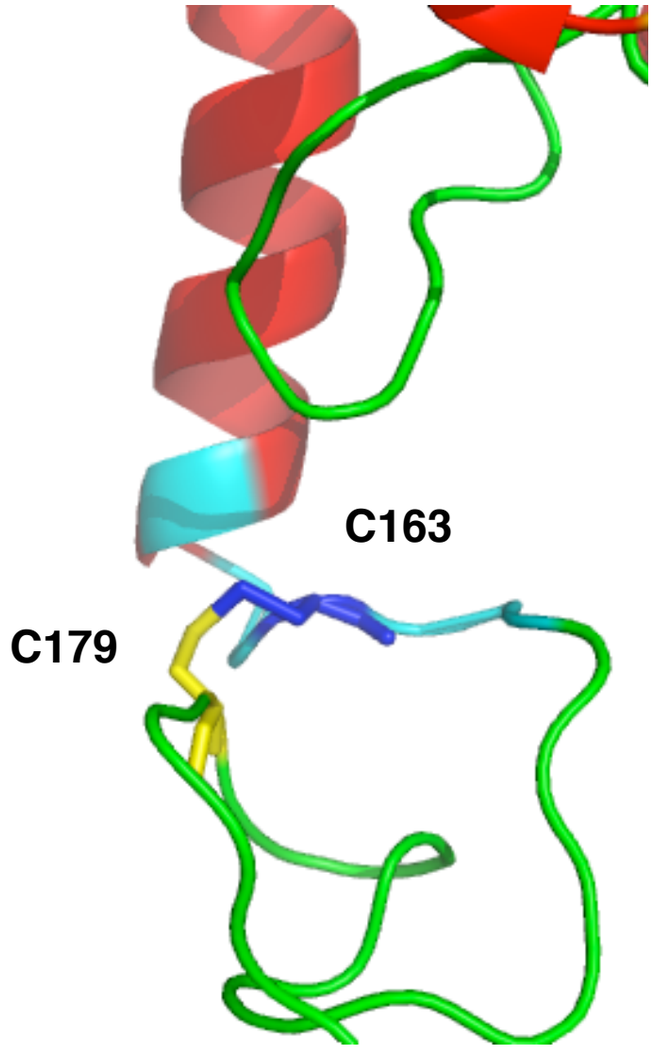


http://snps.biofold.org/snps-and-go

Prediction example

Damaging missing Cys-Cys interaction in the Glycosylasparaginase. The mutation p.Cys163Ser results in the loss of the disulfide bridge between Cys163 and Cys179. This SAP is responsible for Aspartylglucosaminuria.

1APY: Chain A, Res: 2.0 Å



Meta prediction approach

Protein variant predictors

Many predictor of the effect of Single Amino acid Variants (SAVs) are available. They mainly use information from multiple sequence alignment to predict the effect of a given mutation. In his study we consider

- PhD-SNP: Support Vector Machine-based method using sequence and profile information (Capriotti et al. 2006).
- PANTHER: Hidden Markov Model-based method using a HMM library of protein families (Thomas and Kejariwal 2004).
- SNAP: Neural network based method to predict the functional effect of single poit mutations (Bromberg et al. 2008).
- SIFT: Probabilistic method based on the analysis of multiple sequence alignments (Ng and Henikoff 2003).

Predictors accuracy

The accuracy of each predictor has been tested on a set of 35,986 mutations equally distributed between disease-related and neutral polymorphisms. PhD-SNP results in better accuracy but is the only one optimized using a cross-validation procedure. SNAP shows lowest accuracy but it has been developed for a different task.

	Q2	P[D]	S[D]	P[N]	S[N]	C	PM
PhD-SNP	0.76	0.78	0.74	0.75	0.78	0.53	100
PANTHER	0.74	0.79	0.73	0.69	0.74	0.48	74
SNAP	0.64	0.59	0.90	0.79	0.38	0.33	100
SIFT	0.70	0.74	0.64	0.68	0.76	0.41	92

DB: Neutral 17883 and Disease 17883

Prediction matching

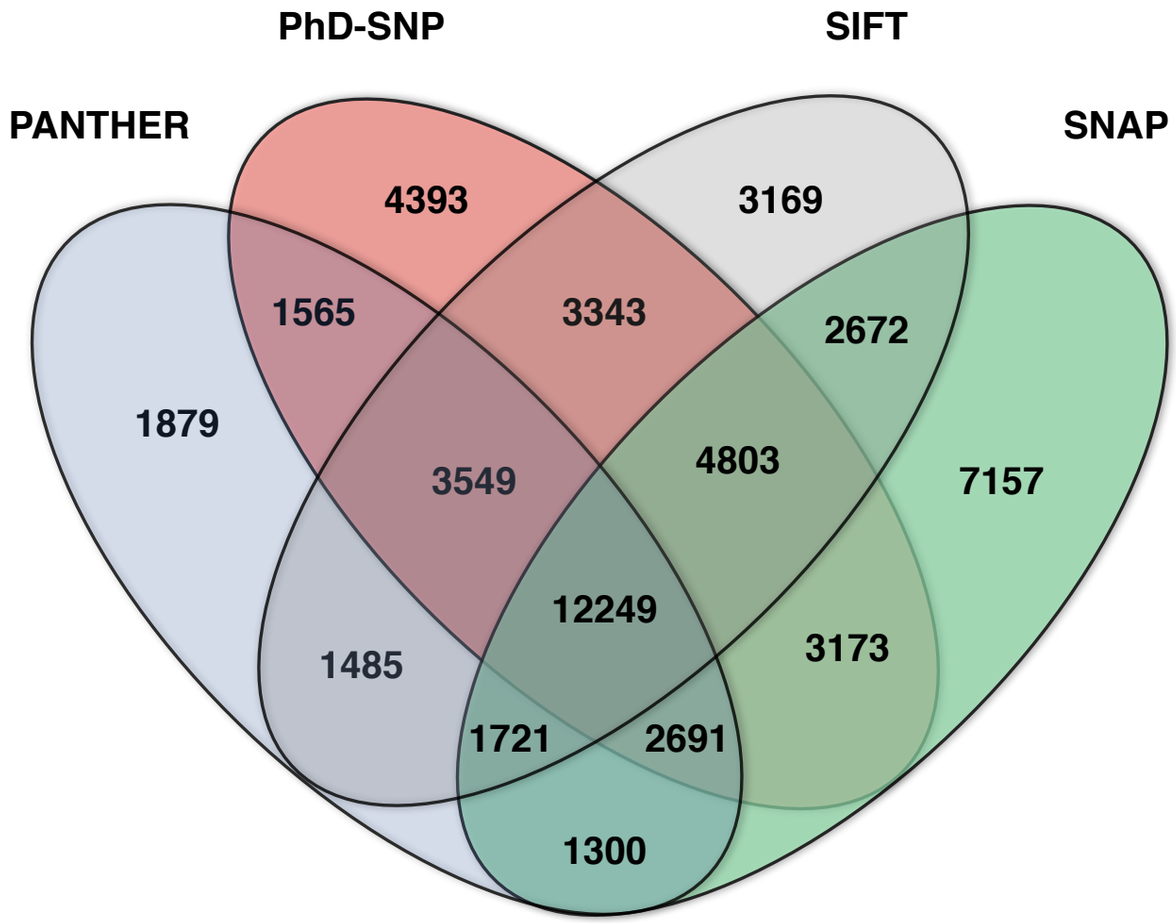
The highest correlation coefficient is between PANTHER and SIFT predictions. SNAP shows lowest correlation with PhD-SNP and PANTHER but high correlation with SIFT which input is included in SNAP

C \ O	PhD-SNP	PANTHER	SNAP	SIFT
PhD-SNP	-	0.76	0.64	0.78
PANTHER	0.51	-	0.67	0.79
SNAP	0.37	0.40	-	0.69
SIFT	0.55	0.58	0.48	-

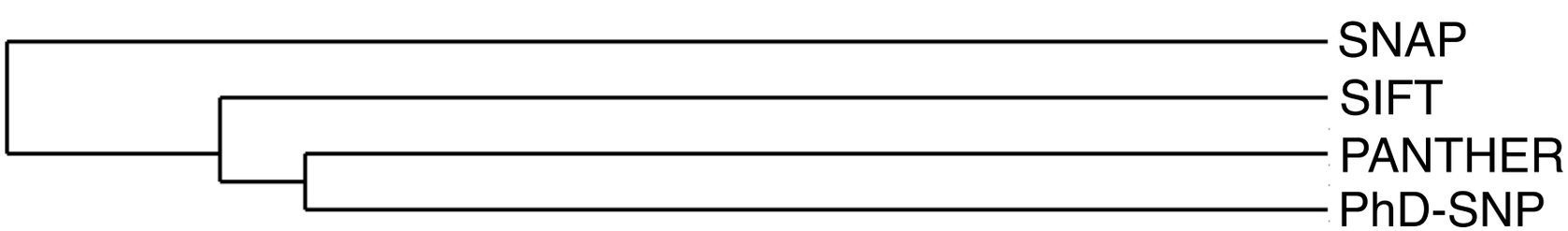
DB: Neutral 17993 and Disease 17993

Predictors tree

Using the prediction similarity we can build the predictors tree



UPGMA tree based on correlations



Prediction analysis

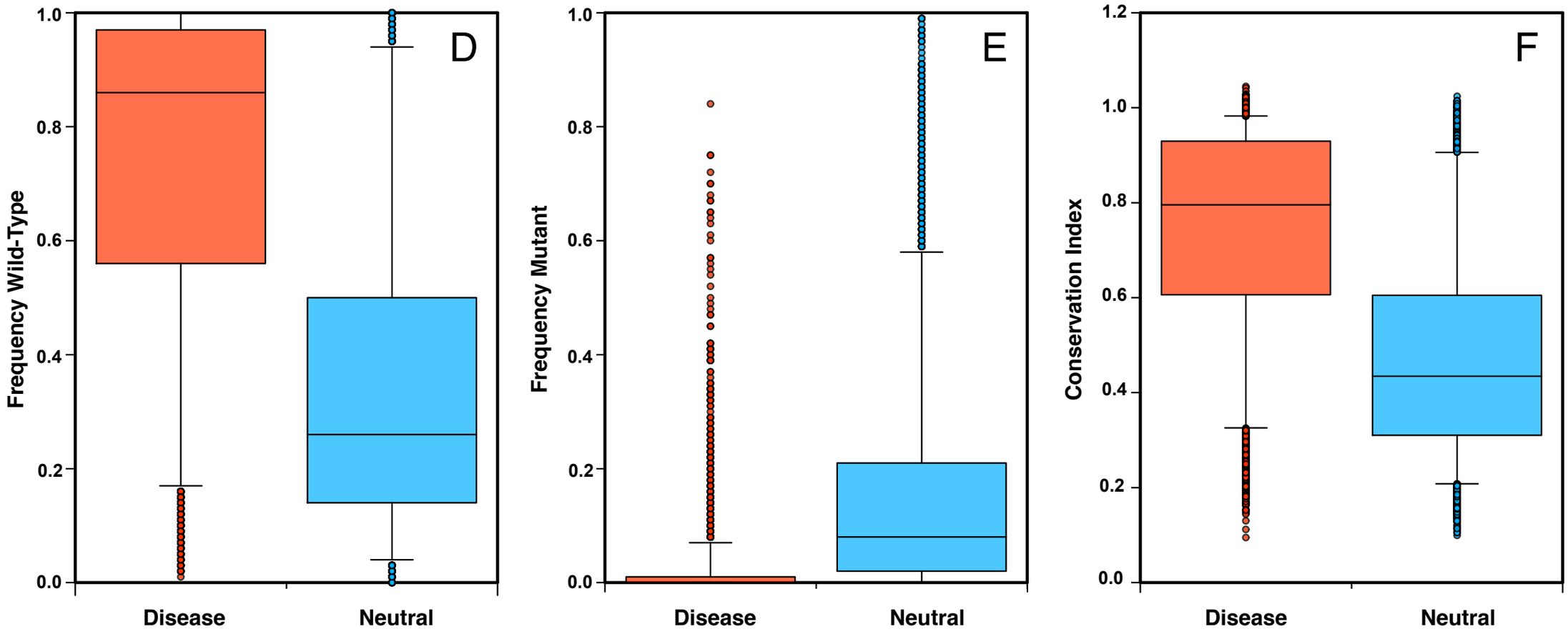
The accuracy of the predictions has been evaluated considering three different subset

- Consensus: all the predictions returned by the methods are in agreement.
- Tie: equal number of methods predicting disease and polymorphism
- Majority: One of the two possible classes is predominant

	Q2	P[D]	S[D]	P[N]	S[N]	C	AUC	%DB
PhD-SNP	0.76	0.78	0.74	0.75	0.78	0.53	0.84	100
Consensus	0.87	0.87	0.92	0.87	0.79	0.73	0.89	46
Majority	0.70	0.67	0.56	0.72	0.80	0.37	0.82	40
Tie	0.61	0.51	0.43	0.66	0.73	0.16	0.67	14

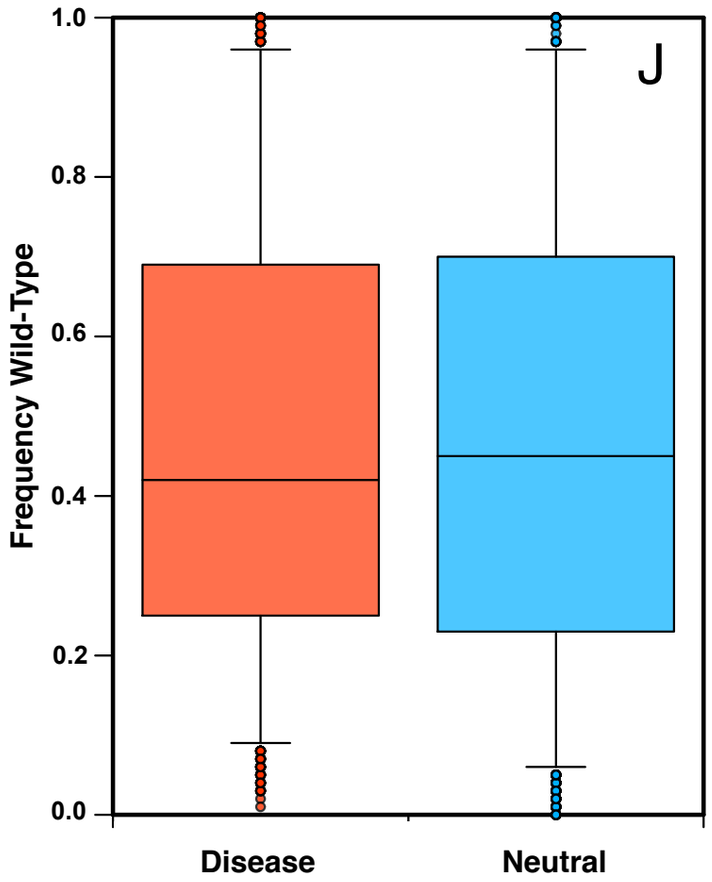
Consensus subset

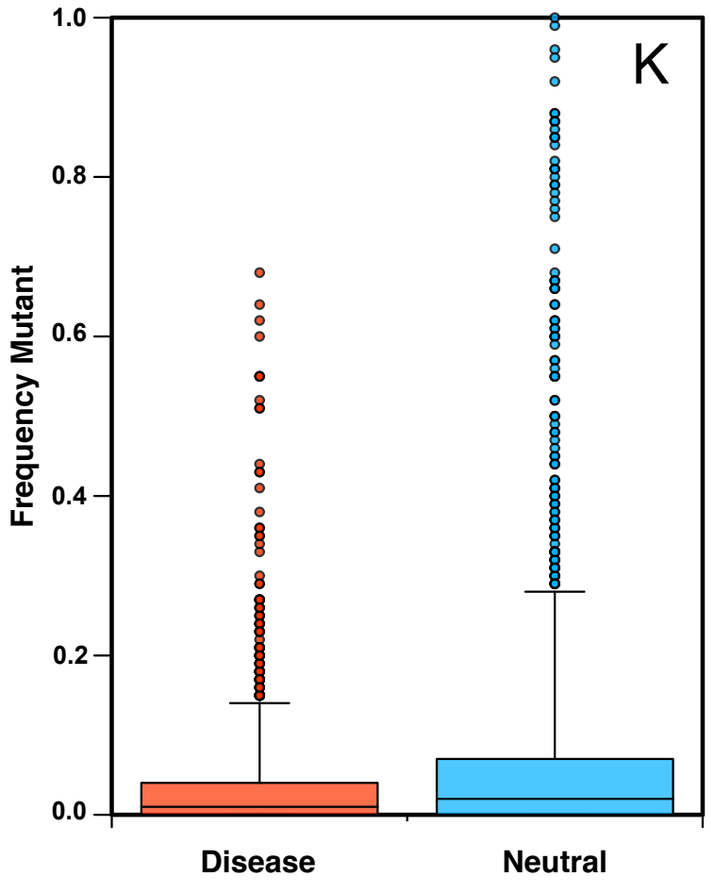
The distributions of the wild-type and new residues frequencies and CI for disease-related variants and polymorphisms on the *Consensus* subset have very little overlap.

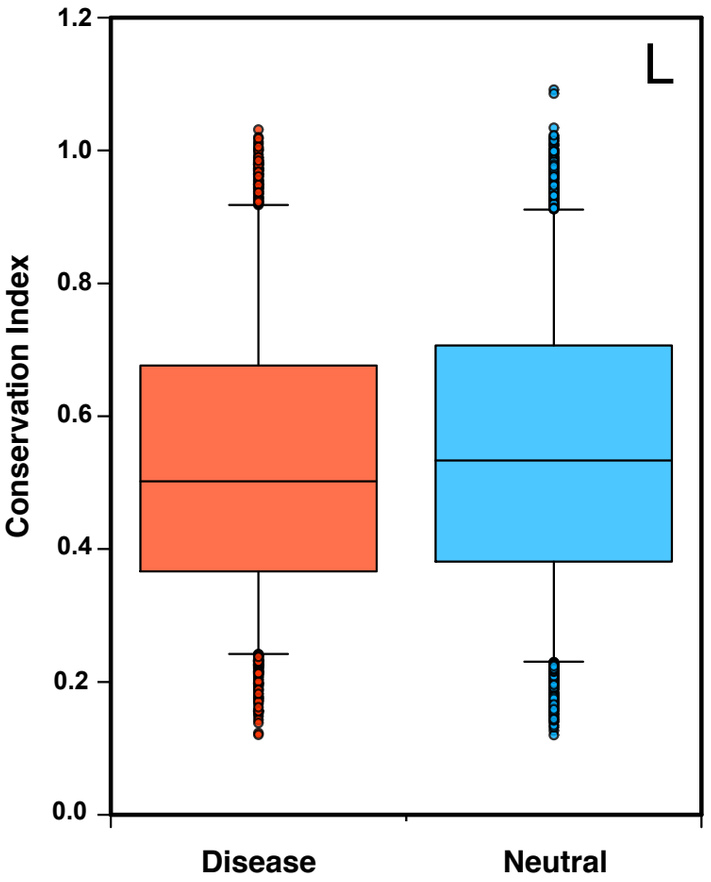


Tie subset

The distributions of the wild-type and new residues frequencies and CI for disease-related variants and polymorphisms on the *Tie* subset have almost complete overlap.

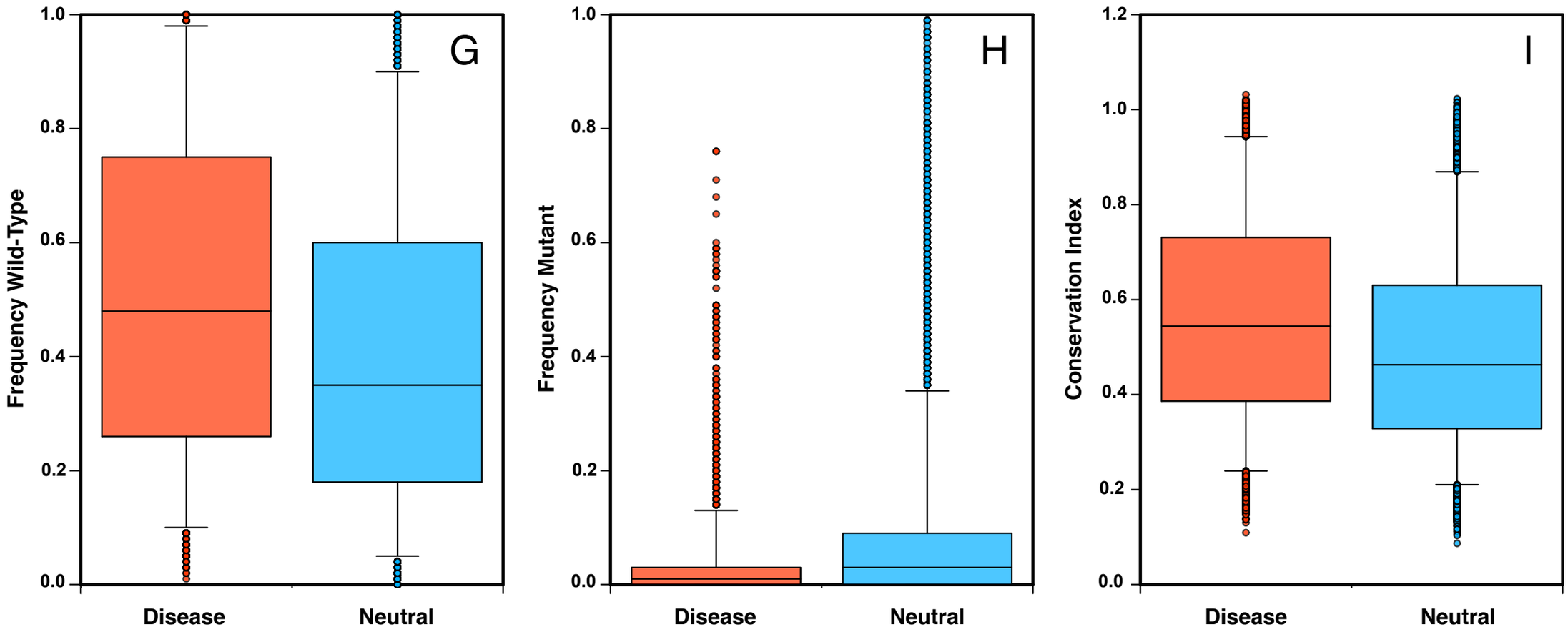






Majority subset

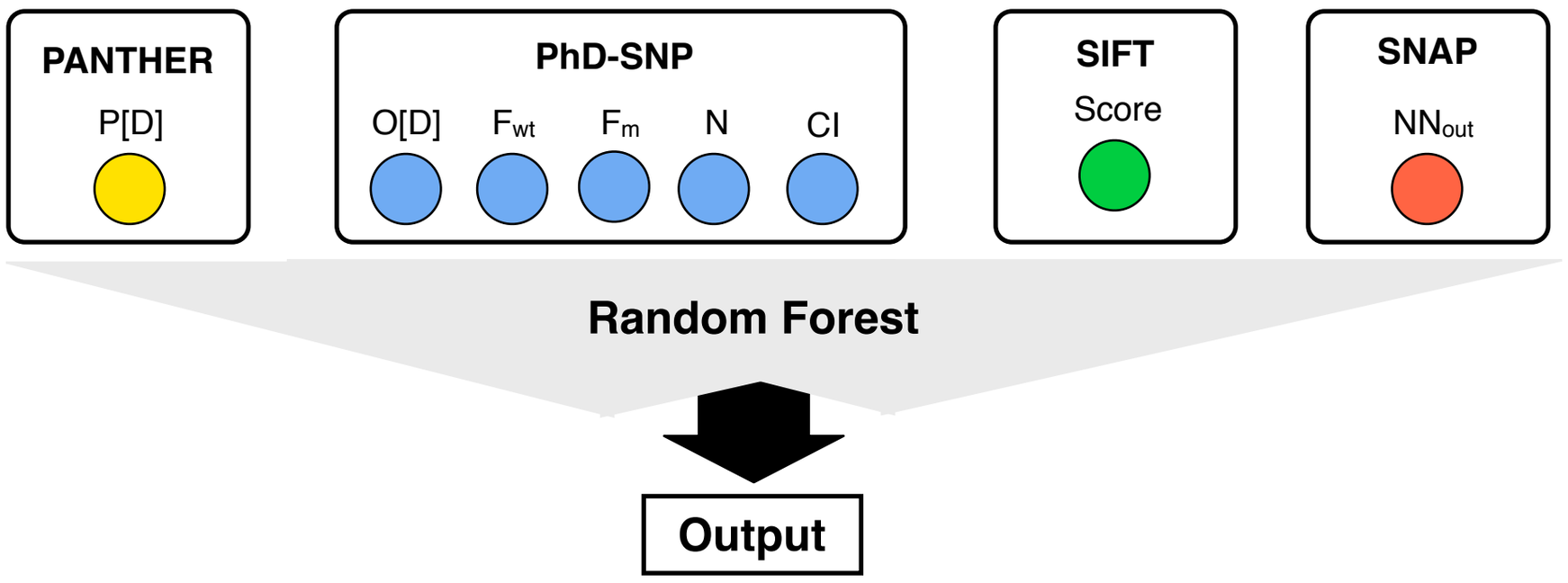
The distributions of the wild-type and new residues frequencies and CI for disease-related and polymorphism on the *Majority* subset are in an intermediate situation with respect to the previous cases.



Meta-SNP

The Meta-SNP is a RF-based meta predictor that takes in input * input features from the output of PhD-SNP, PANTHER, SNAP and SIFT.

The output of the methods can be analyzed dividing the dataset in consensus predictions (all the methods in agree), tie predictions (same number of disease and non-disease predictions) and other predictions (the remaining cases) .



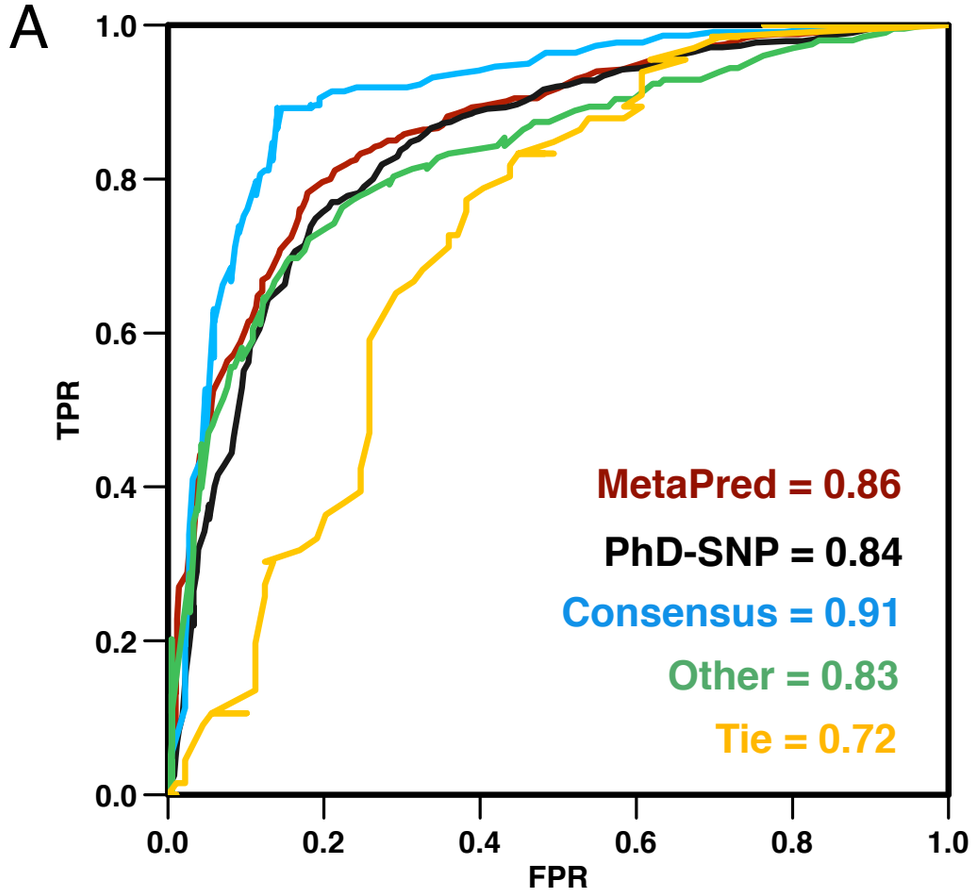
http://snps.biofold.org/meta-snp

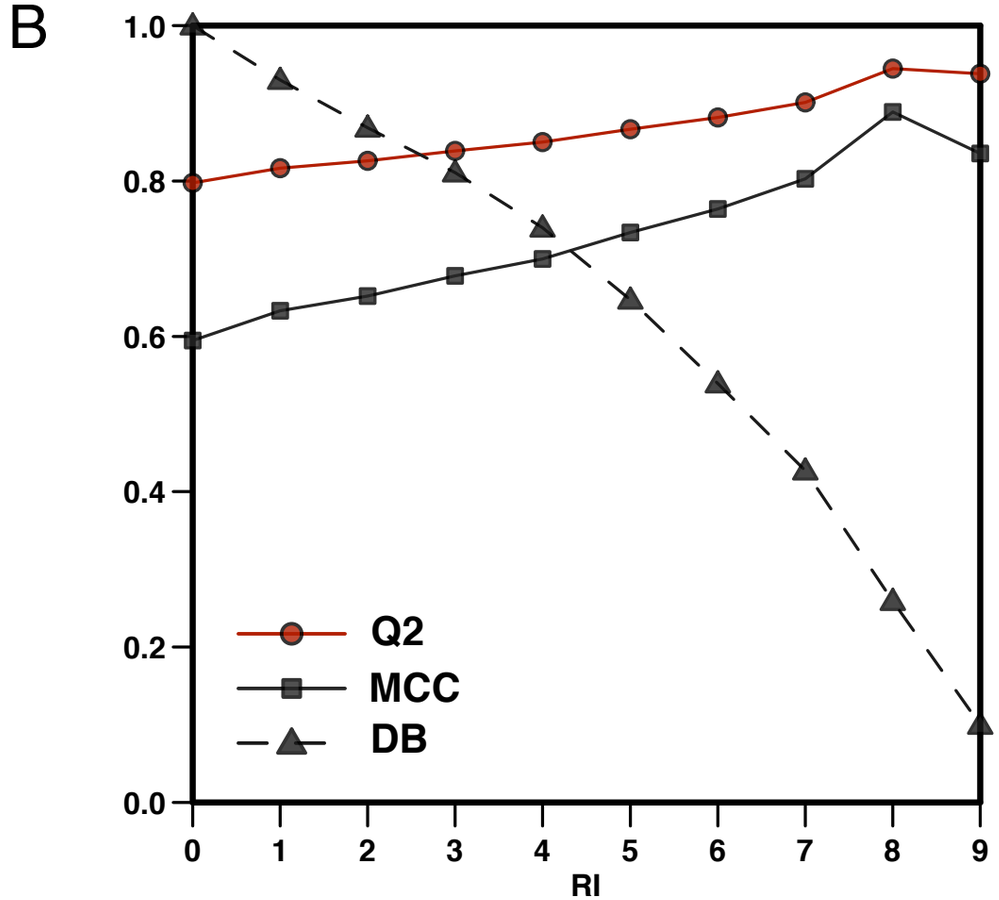
Testing Meta-SNP

Performances of Meta-Pred on the test set of 972 variants from 577 proteins

	Q2	P[D]	S[D]	P[N]	S[N]	C
Meta-SNP	0.79	0.79	0.80	0.80	0.79	0.59
PhD-SNP	0.77	0.78	0.77	0.77	0.78	0.55

DB: Neutral 486 and Disease 486



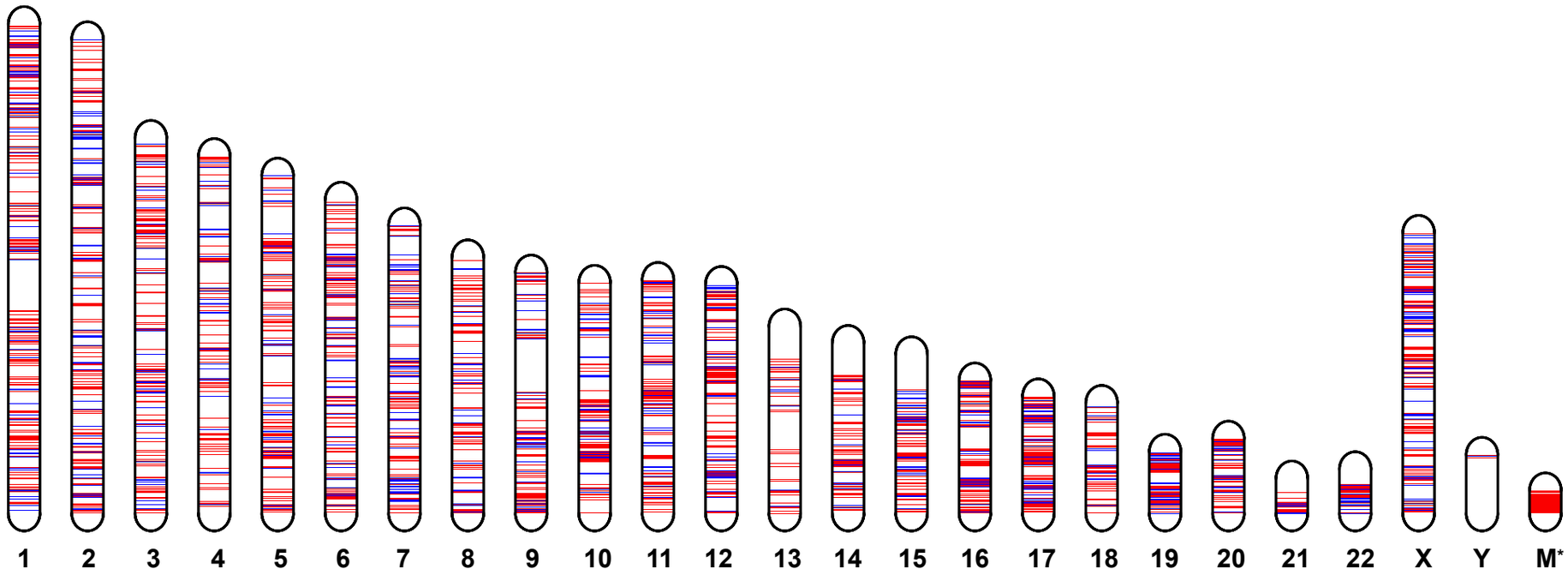


Capriotti et al. (2013). BMC Genomics. 14 (S3), In press.

From coding to non-coding

Whole-genome predictions

Most of the genetic variants occur in non-coding region that represents >98% of the whole genome.

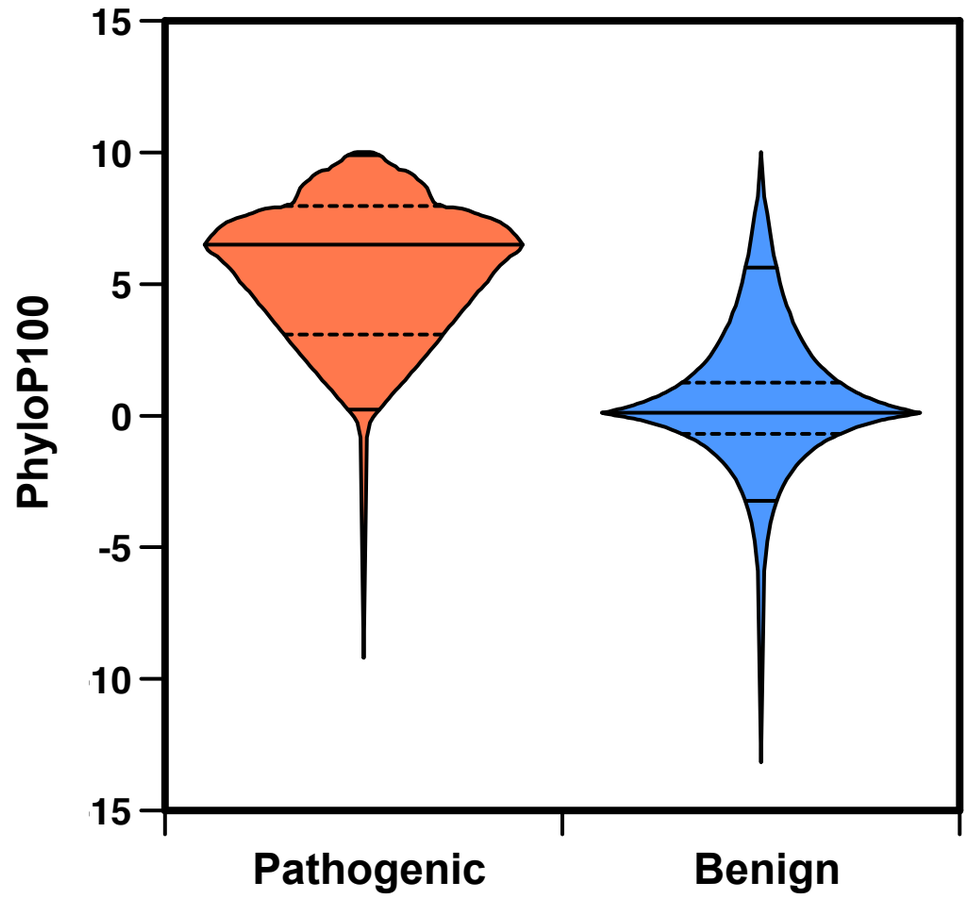


Predict the effect of SNVs in non-coding region is a challenging task because conservation is more difficult to estimate.

Sequence alignment is more complicated for sequences from non-coding regions.

PhyloP100 score

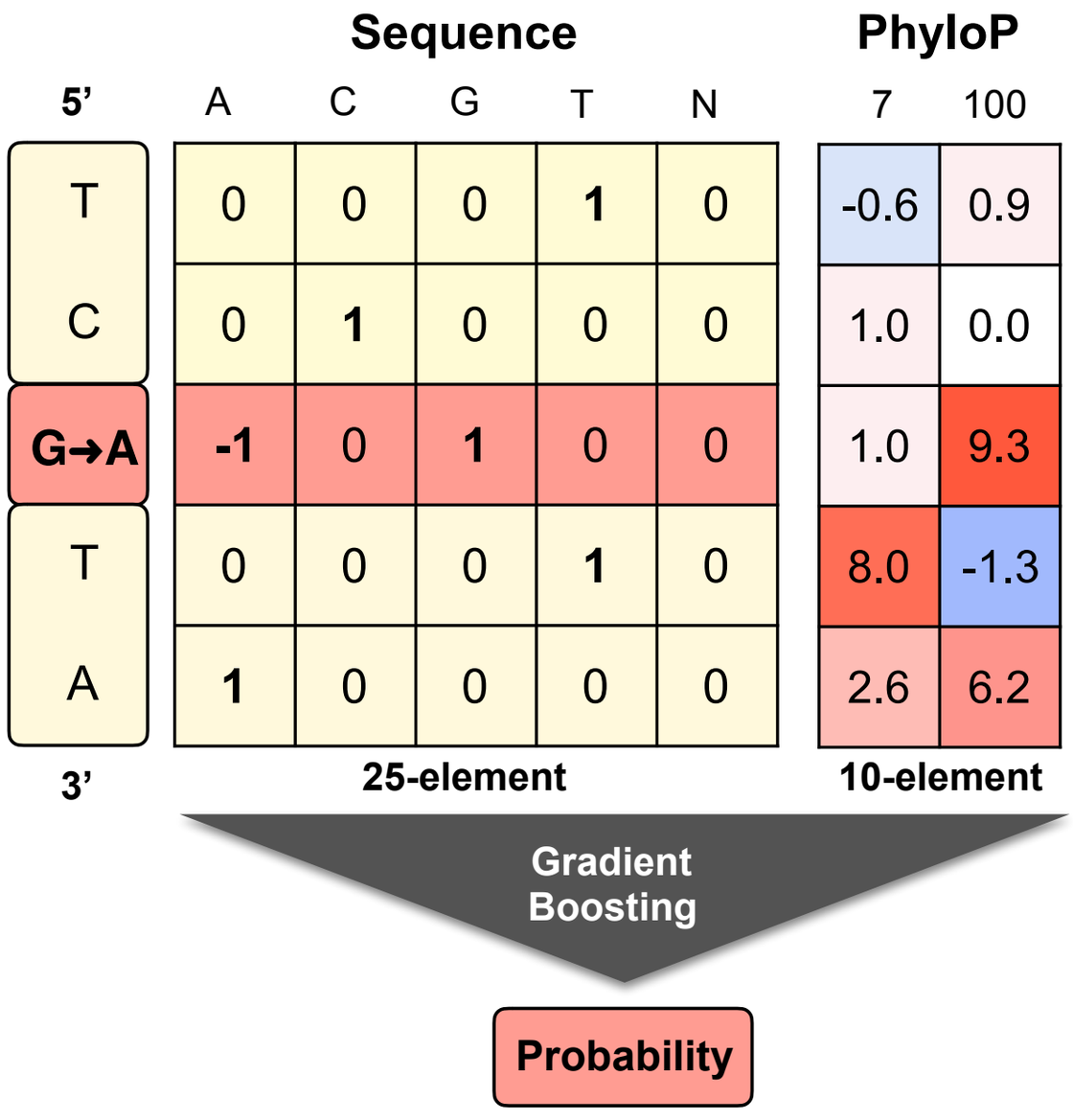
Conservation analysis based on the pre-calculated score available at the UCSC revealed a significant difference between the distribution of the PhyloP100 scores in Pathogenic and Benign SNVs.



PhyloP: Pollard et al,, Genome Research 2010

PhD-SNPg

PhD-SNPg is a simple method that takes in input 35 sequence-based features from a window of 5 nucleotides around the mutated position.

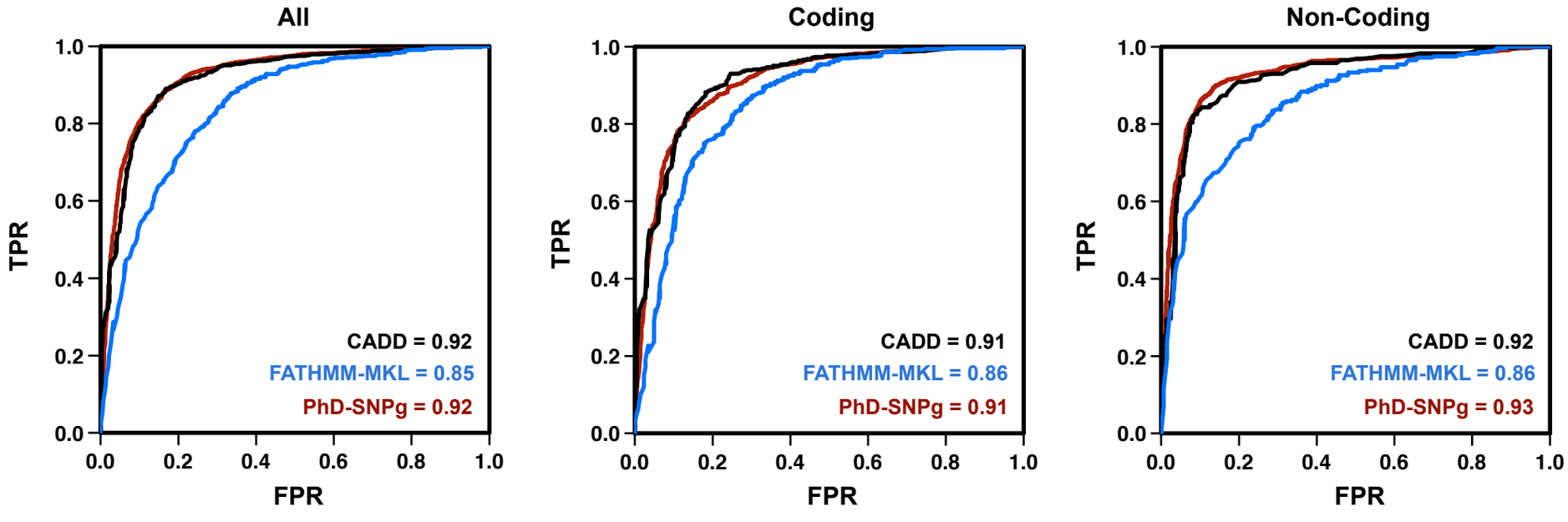


http://snps.biofold.org/phd-snpg/

Benchmarking

PhD-SNP^{g} has been tested in cross-validation on a set of 35,802 SNVs and on a blind set of 1,408 variants recently annotated.

	Q2	TNR	NPV	TPR	PPV	MCC	F1	AUC
PhD-SNP^{g}	0.861	0.774	0.884	0.925	0.847	0.715	0.884	0.924
Coding	0.849	0.671	0.845	0.938	0.850	0.651	0.892	0.908
Non-Coding	0.876	0.855	0.911	0.901	0.839	0.753	0.869	0.930

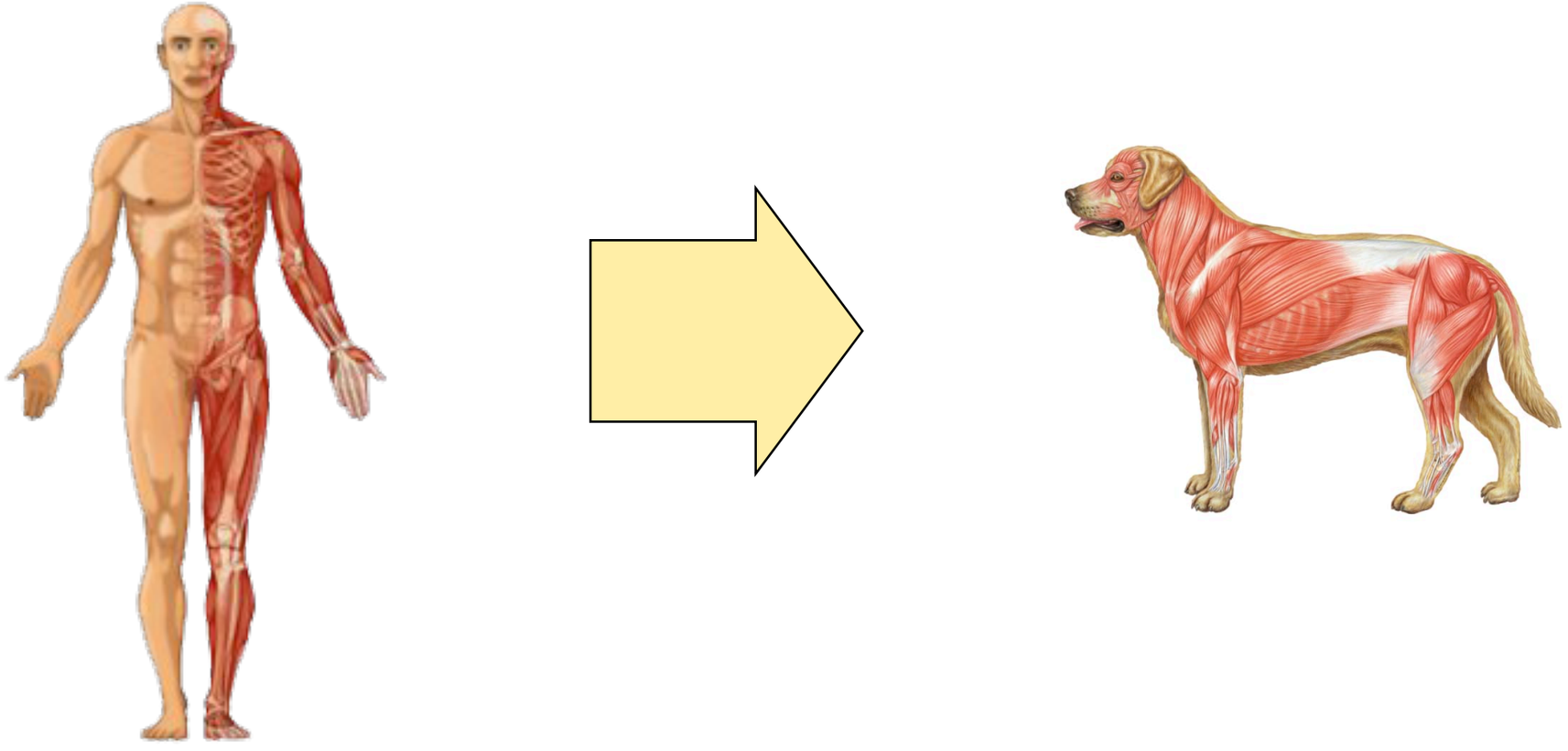


Capriotti and Fariselli. (2017) Nucleic Acids Res. PMID: 28482034.

From human to animals

Is it possible to develop similar algorithms for animals? The main limitations are the lack of

- curated databases of disease costing variants
- pre-calculated conservation scores

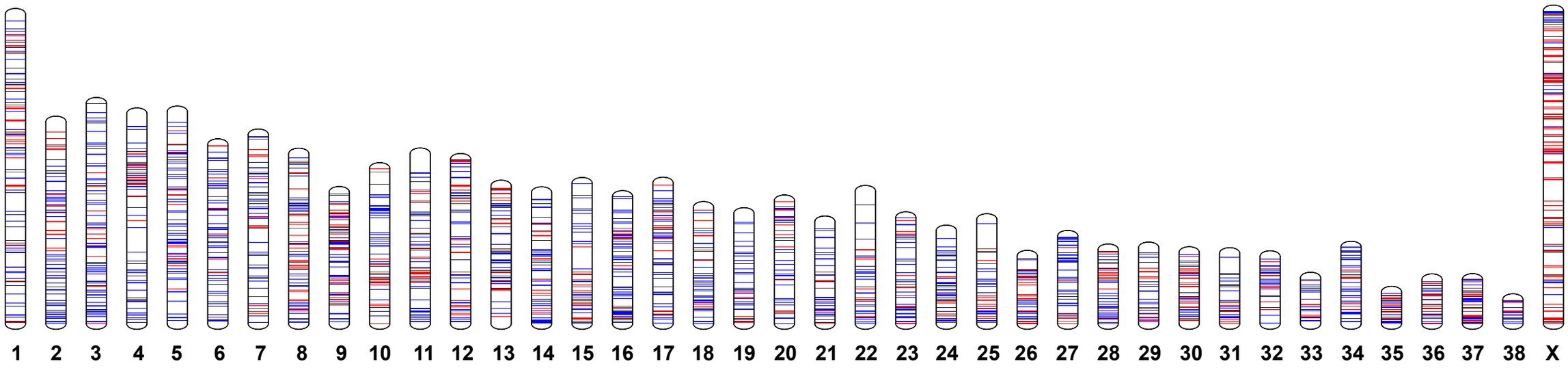


Algorithm optimization

To overcome this limitation we adapted the human predictor to the dog genome with the following steps:

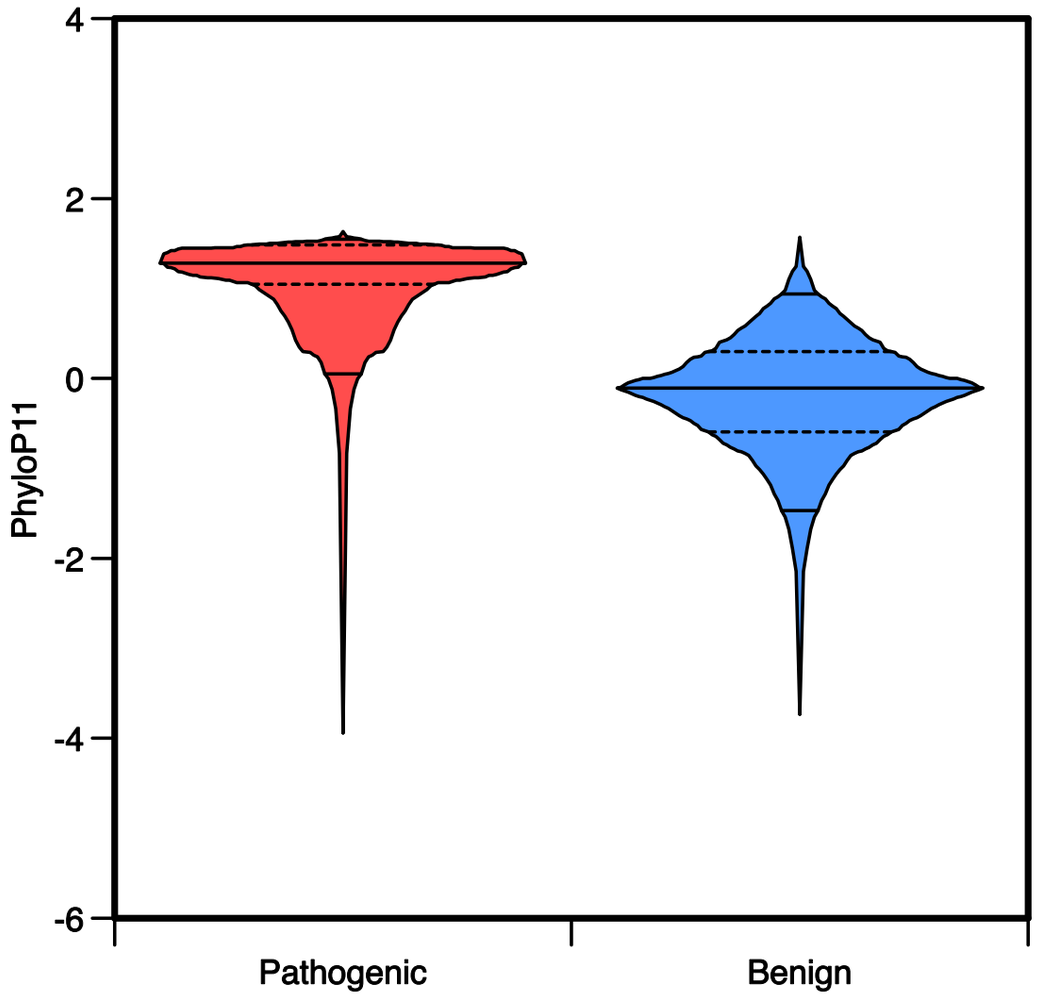
- we calculated the conservation score for the dog genome with limited number of species

- Calibrated the classification threshold using a curated set of ~1,500 highly-conserved human disease causing variants that we mapped on the dog genome



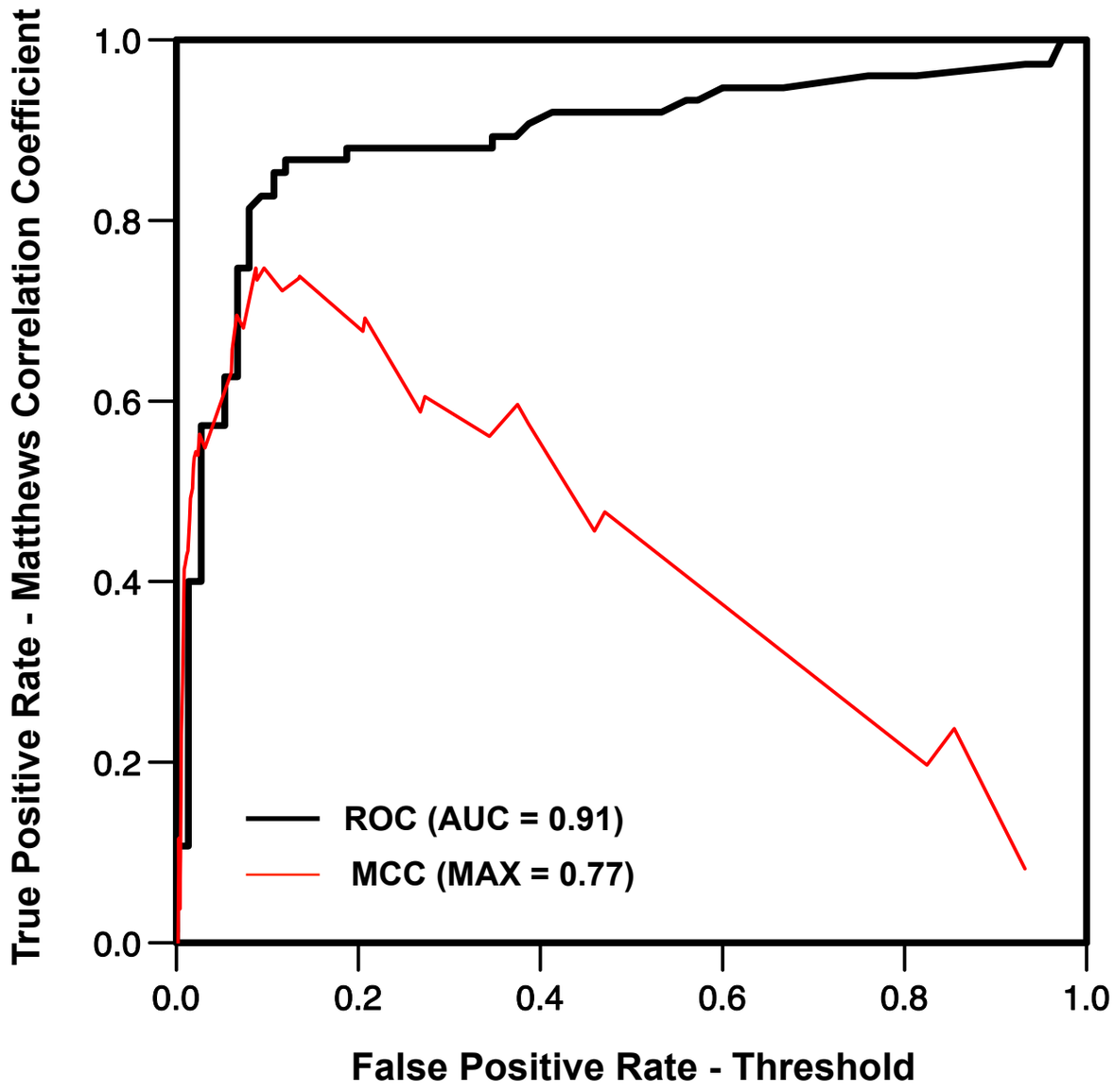
Conservation score

The distributions of the PhyloP11 scores for potentially ***pathogenic*** and ***benign*** variants in the dog genome are significantly different



Method calibration

We calculated the performance and different classification thresholds and found that for 0.1 our algorithm reaches the maximum value of Matthews Correlation Coefficient



Method validation

We selected a small set of dog variants (75) annotated in OMIA dataset and tested the performance of our method. We found the Fido-SNP reaches the same performance in the calibration and validation steps.

Dataset	TH	Q2	TNR	NPV	TPR	PPV	MCC	AUC
Human-Dog	0.09	0.87	0.91	0.86	0.86	0.85	0.77	0.91
OMIA	0.10	0.88	0.92	0.85	0.85	0.84	0.77	0.91

http://snps.biofold.org/fido-snp/

Capriotti et al.. (2019) Nucleic Acids Res. PMID: 31114899.

Blind testing

CAGI experiments

The Critical Assessment of Genome Interpretation is a community experiment to objectively assess computational methods for predicting the phenotypic impacts of genomic variation.

https://genomeinterpretation.org/

The P16 challenge

CDKN2A is the most common, high penetrance, susceptibility gene identified to date in familial malignant melanoma. $p16^{INK4A}$ is one of the two oncosuppressor which promotes cell cycle arrest by inhibiting cyclin dependent kinase (CDK4/6).

Challenge: Evaluate how different variants of p16 protein impact its ability to block cell proliferation.

Provide a number between 50% that represent the normal proliferation rate of control cells and 100% the maximum proliferation rate in case cells.

SNPs&GO prediction

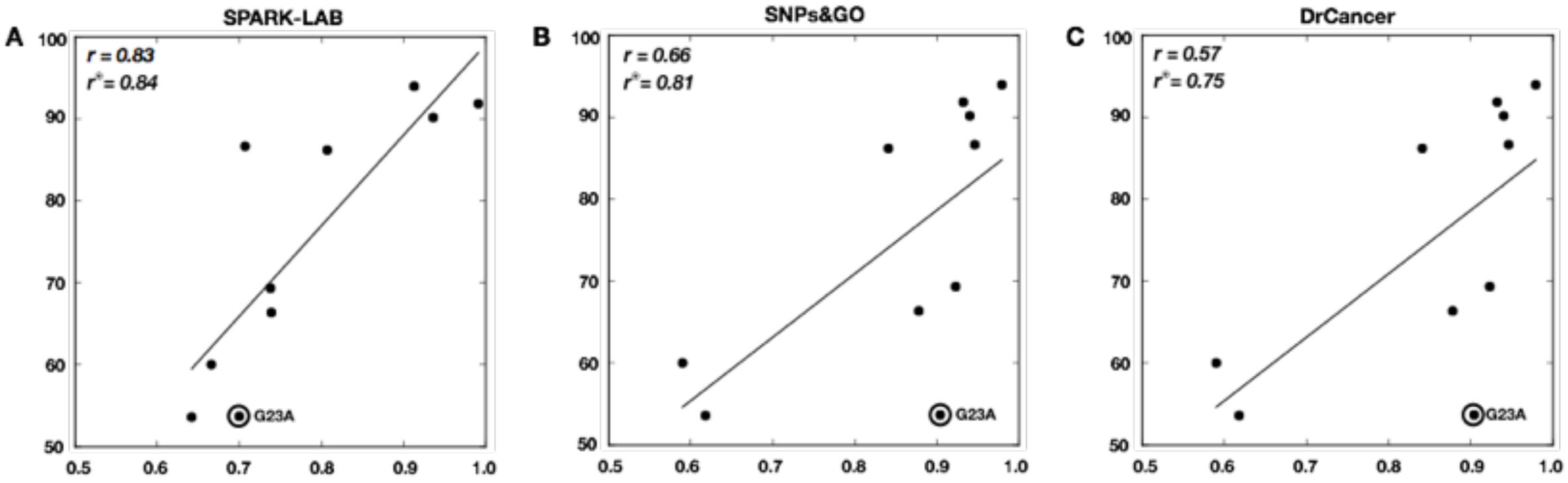
Proliferation rates predicted using the output of SNPs&GO without any optimization.

Variant	Prediction	Real	Δ	%WT	%MUT
G23R	0.932	0.918	0.014	84	0
G23S	0.923	0.693	0.230	84	1
G23V	0.940	0.901	0.039	84	0
G23A	0.904	0.537	0.367	84	2
G23C	0.946	0.866	0.080	84	0
G35E	0.590	0.600	0.010	12	14
G35W	0.841	0.862	0.021	12	0
G35R	0.618	0.537	0.081	12	4
L65P	0.878	0.664	0.214	15	1
L94P	0.979	0.939	0.040	56	0

P16 predictions

SNPs&GO resulted among the best methods for predicting the impact of P16INK4A variants on cell proliferation.

Method	Q2	AUC	MC	RMSE	$r_{Pearson}$	$r_{Spearman}$	$r_{KendallTau}$
SPARK-LAB	0.900	0.920	0.816	0.30	0.595	0.619	0.443
SNPs&GO	0.700	0.880	0.500	0.33	0.575	0.616	0.445
DrCancer	0.600	0.840	0.333	0.46	0.477	0.495	0.409



Capriotti et al. (2017) Human Mutations. PMID: 28102005.

The NAGLU challenge

NAGLU is a lysosomal glycohydrolyase which deficiency causes a rare disorder referred as Sanfilippo B disease

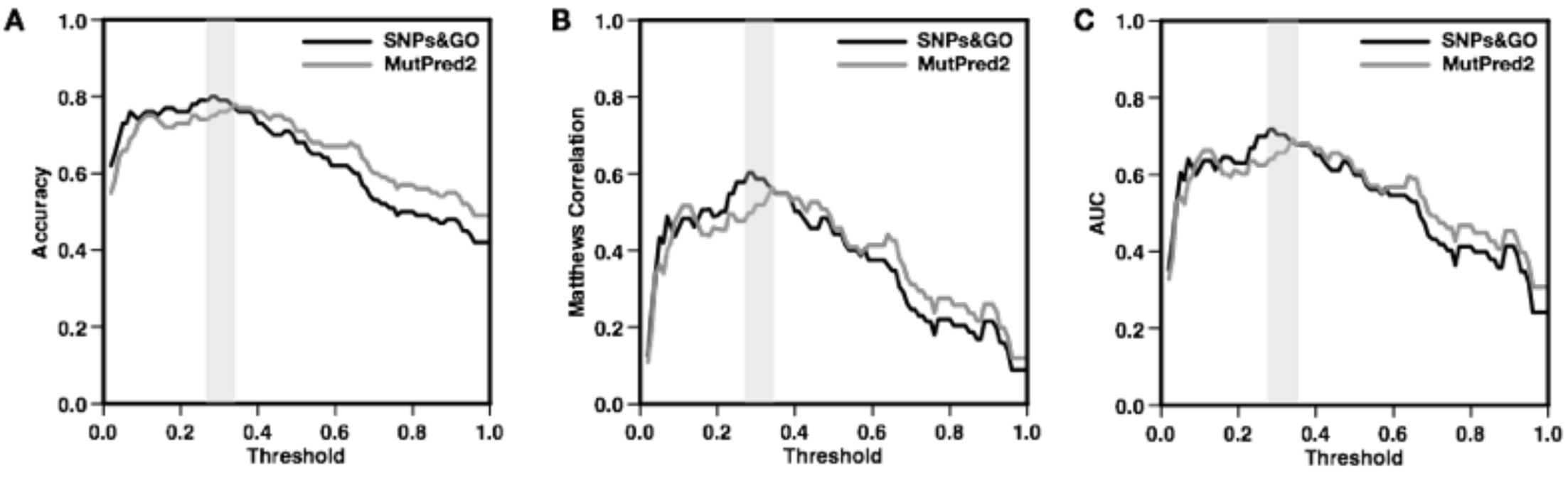
Challenge: Predict the effect of the 165 variants on NAGLU enzymatic activity.

The submitted prediction should be a numeric value ranging from 0 (no activity) to 1 (wild-type level of activity).

A posteriori evaluation

I performed a posteriori evaluation of the performance based on my version of the predictor and found that SNPs&GO reaches similar accuracy than the best method (MutPred2)

Method	Q2	AUC	MC	RMSE	$r_{Pearson}$	$r_{Spearman}$	$r_{KendallTau}$
MutPred2	0.780	0.850	0.565	0.30	0.595	0.619	0.443
SNPs&GO	0.800	0.854	0.603	0.33	0.575	0.616	0.445
SNPs&GO^{09}	0.750	0.749	0.499	0.46	0.477	0.495	0.409



Conclusions

- Evolutionary information is an important feature for the prediction of deleterious variants. The wild-type residues in disease-related variant sites are more conserved than in neutral sites.

- Among the algorithms for predicting the effect of a single amino acid variants on human health, the methods based on functional information are the most accurate ones.

- Structural information encoded through the relative solvent accessible area and the structure environment improves the predictions of disease-causing variants.

- The implementation of meta-prediction based approach allows to select highly accurate predictions.

- Nucleotide conservation is an important feature to predict the impact of SNVs in non coding regions

Acknowledgments

Structural Genomics @CNAG
Marc A. Marti-Renom
Francois Serra

Computational Biology and Bioinformatics Research Group (UIB)
Jairo Rocha

Division of Informatics at UAB
Malay Basu
Division Clinical Immunology & Rheumatology
Harry Schroeder
Mohamed Khass

Helix Group (Stanford University)
Russ B. Altman
Jennifer Lahti
Tianyun Liu
Grace Tang

Bologna Biocomputing Group
Rita Casadio
Pier Luigi Martelli
University of Torino
Piero Fariselli
University of Camerino
Mario Compiani

Mathematical Modeling of Biological Systems (University of Düsseldorf)
Markus Kollmann
Linlin Zhao

Other Collaborations
Yana Bromberg, Rutger University, NJ
Hannah Carter, UCSD, CA
Francisco Melo, Universidad Catolica, Chile
Sean Mooney, Buck Institute, Novato
Cedric Notredame, CRG Barcelona
Gustavo Parisi, Univesidad de Quilmes
Frederic Rousseau, KU Leuven
Joost Schymkowitz, KU Leuven

FUNDING
Italian MIUR: PRIN 2017
NIH: 1R21 AI134027- 01A1
Italian MIUR: FFABR 2017
UNIBO: International Cooperation
Startup funding Dept. of Pathology UAB
NIH:3R00HL111322-04S1 Co-Investigator
EMBO Short Term Fellowship
Marie Curie International Outgoing Grant
Marco Polo Research Project
BIOSAPIENS Network of Excellence
SPINNER Consortium

Biomolecules, Folding and Disease

http://biofold.org/