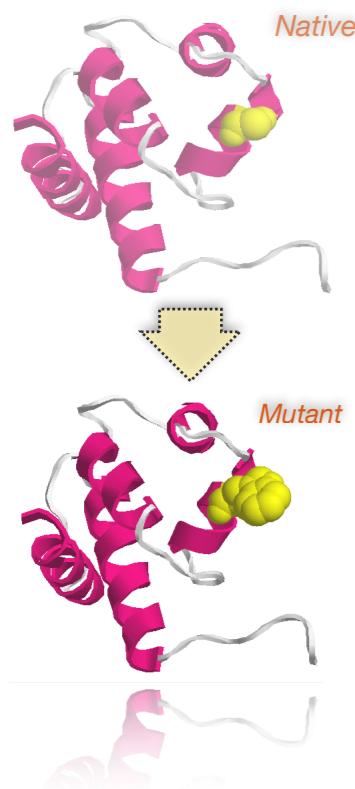
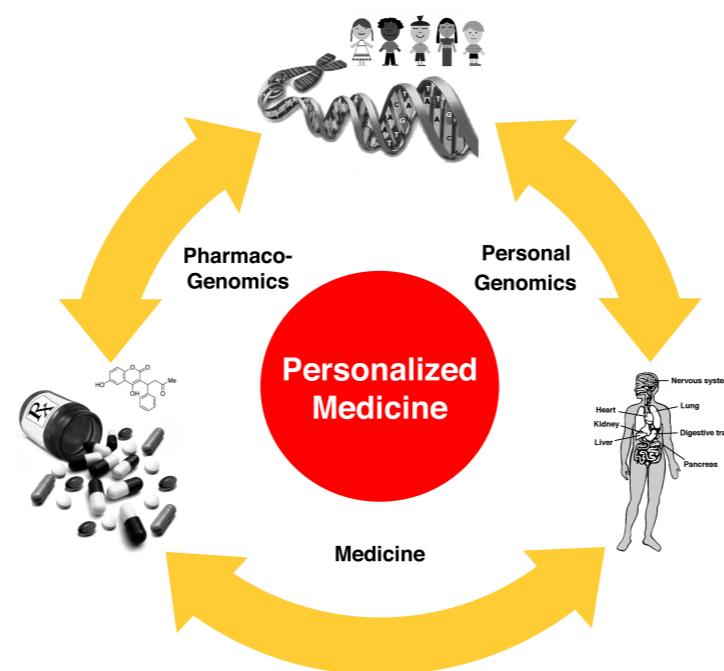


Predicting the impact of variants in the whole genome



Düsseldorf (Germany)
September 2, 2016

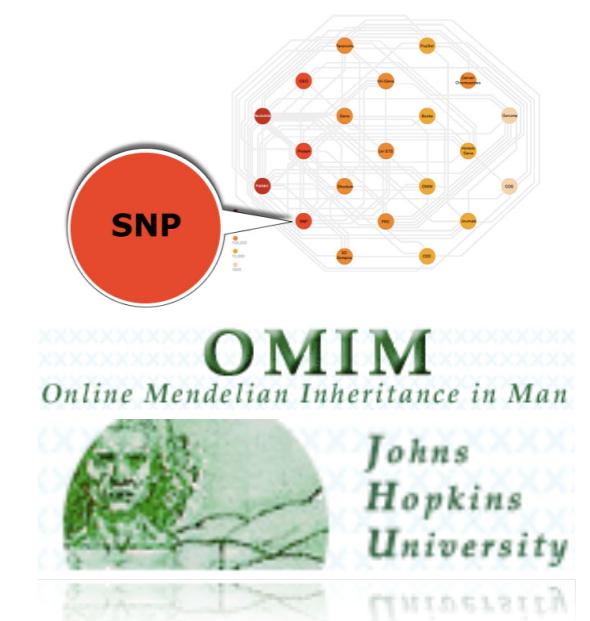


Emidio Capriotti

<http://biifold.org/>



Biomolecules
Folding and
Disease



Institute for Mathematical Modeling
of Biological Systems
Department of Biology

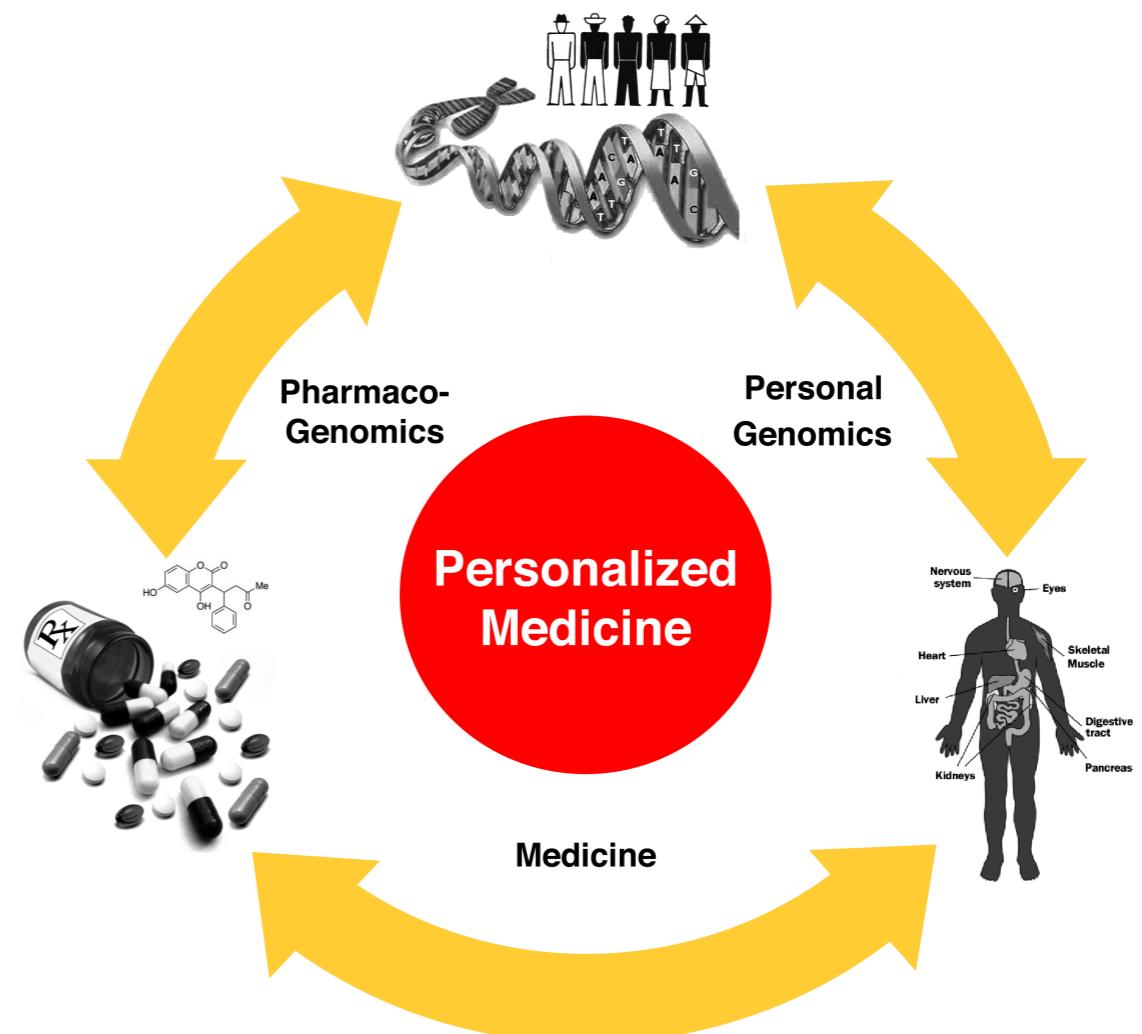
Emidio Capriotti
HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Personalized medicine

Currently direct to consumers company are performing **genotype test** on **markers associated to genetic traits**, and soon **full genome sequencing** will cost about \$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**



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.

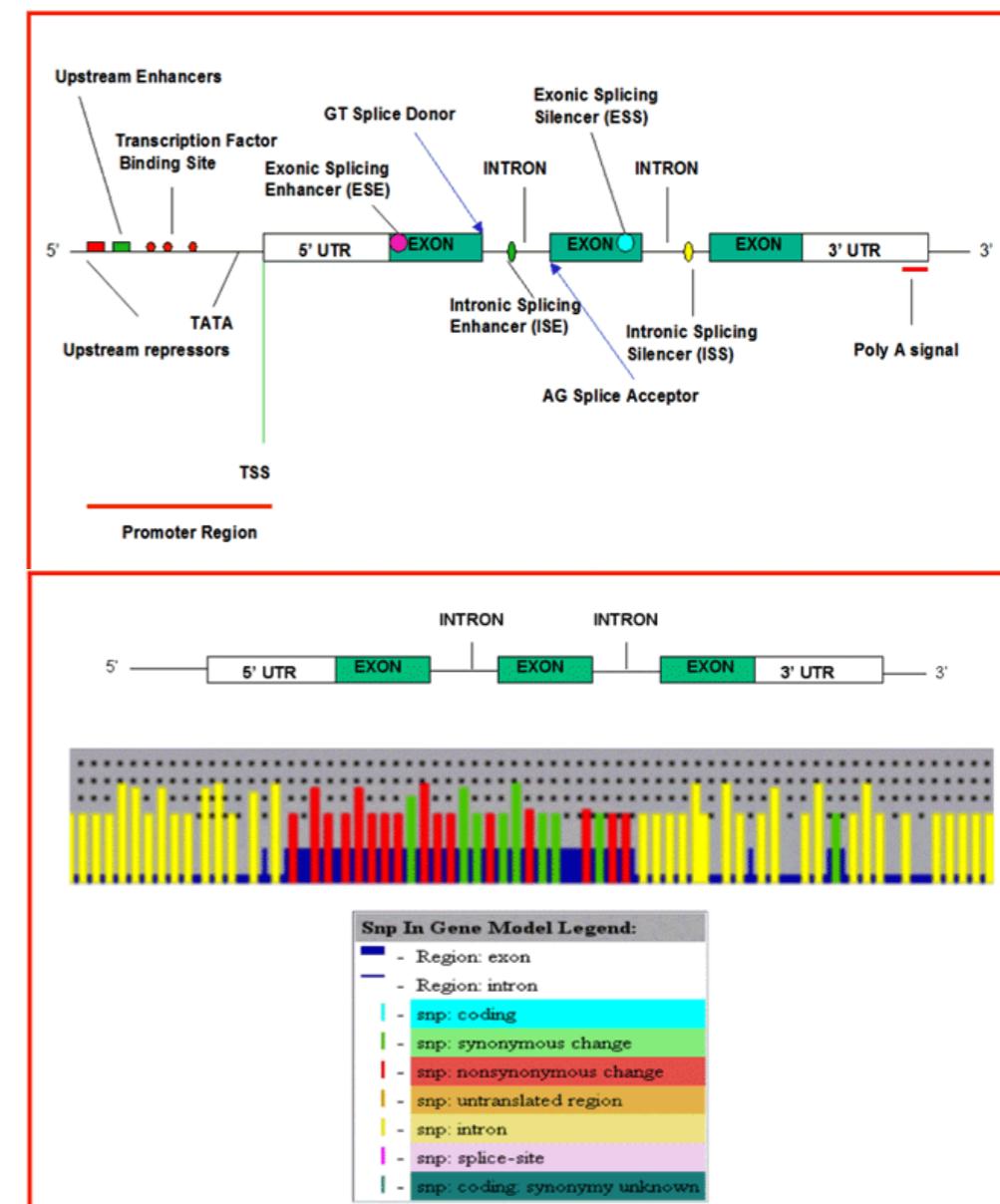
It is used to refer to Polymorphisms when the population frequency is $\geq 1\%$

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.



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**

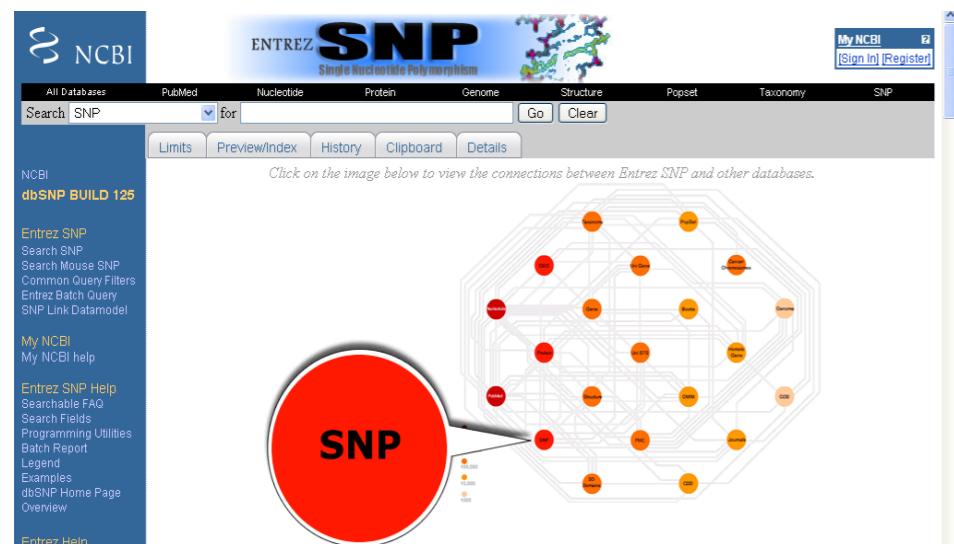
- Properties of amino acid residue substitution
- The evolutionary history of an amino acid position
- Sequence–function relationships
- Structure–function relationships

Impact of **non-coding variants**

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

SNVs and SAVs databases

dbSNP 147 (Apr 2016) @ NCBI



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

Single Nucleotide Variants

<i>Homo sapiens</i>	100,815,862*
<i>Bos taurus</i>	100,170,652
<i>Mus musculus</i>	80,443,437

* Validated

SwissVar (Jul 2016) @ ExPASy



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

Single Amino acid Variants

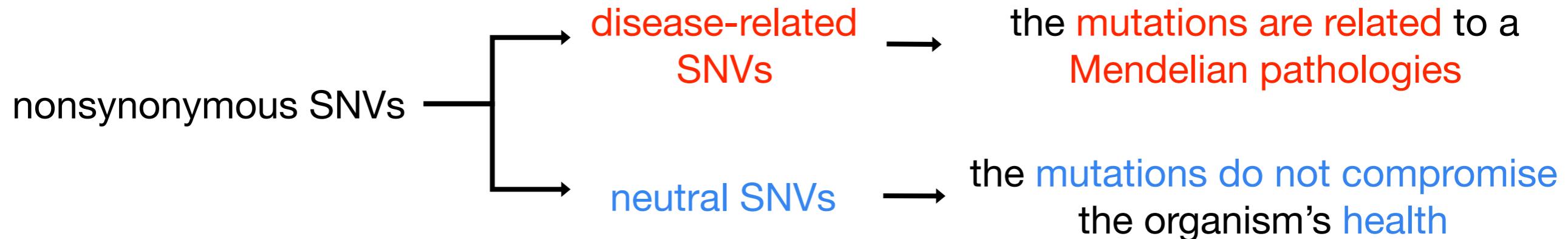
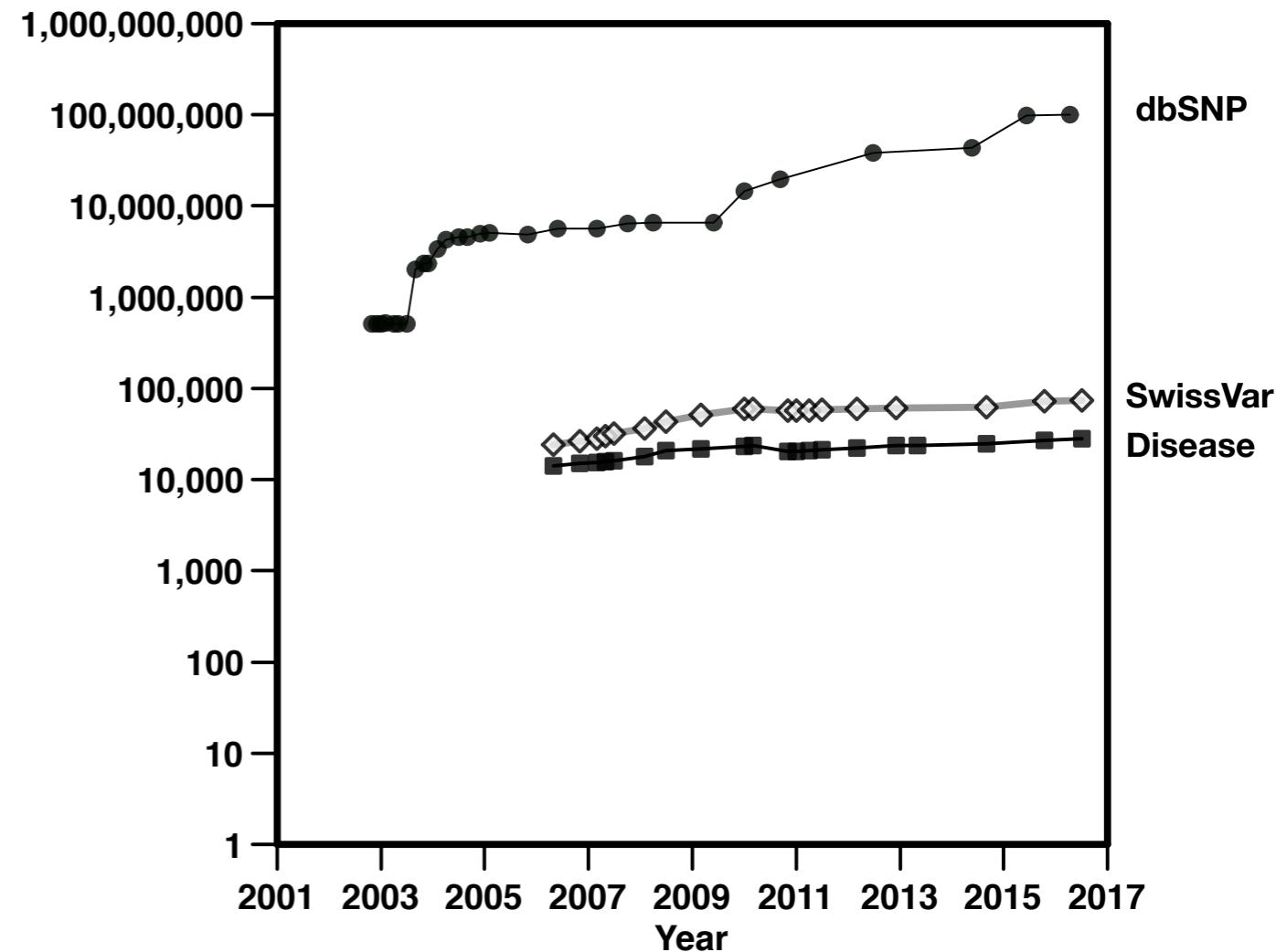
<i>Homo sapiens</i>	74,177
<i>Disease</i>	28,239
<i>Polymorphisms</i>	38,934

Sep 2016

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).



Sequence, Structure & Function

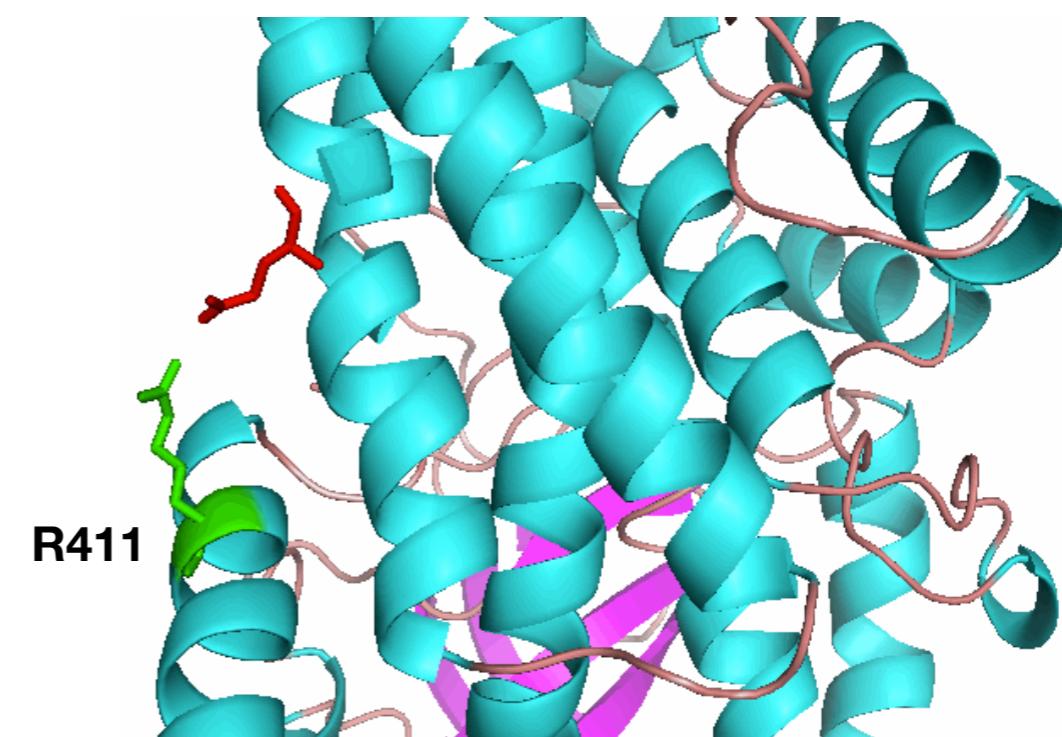
Genomic variants in sequence motifs could affect protein function.

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.



What predictions?

Given the large amount of available mutations **what can we predict?**

Develop binary classifiers to predict the impact of mutations on:

- Protein Structure
- Protein Function
- Human Health

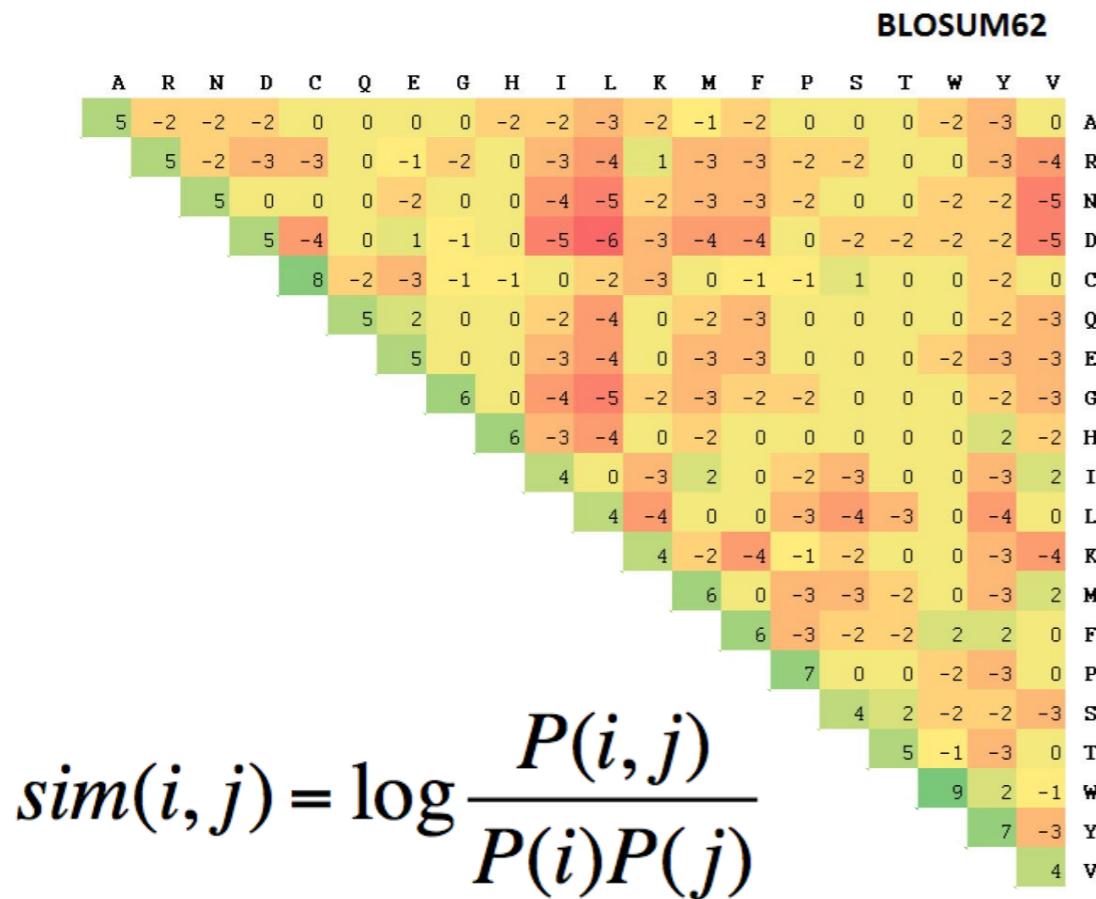
Structural changes upon mutation can be predicted using comparative modeling approaches.

Functional changes can be predicted from experimental data collected in PMD database (at <http://www.genome.jp/dbget/>)

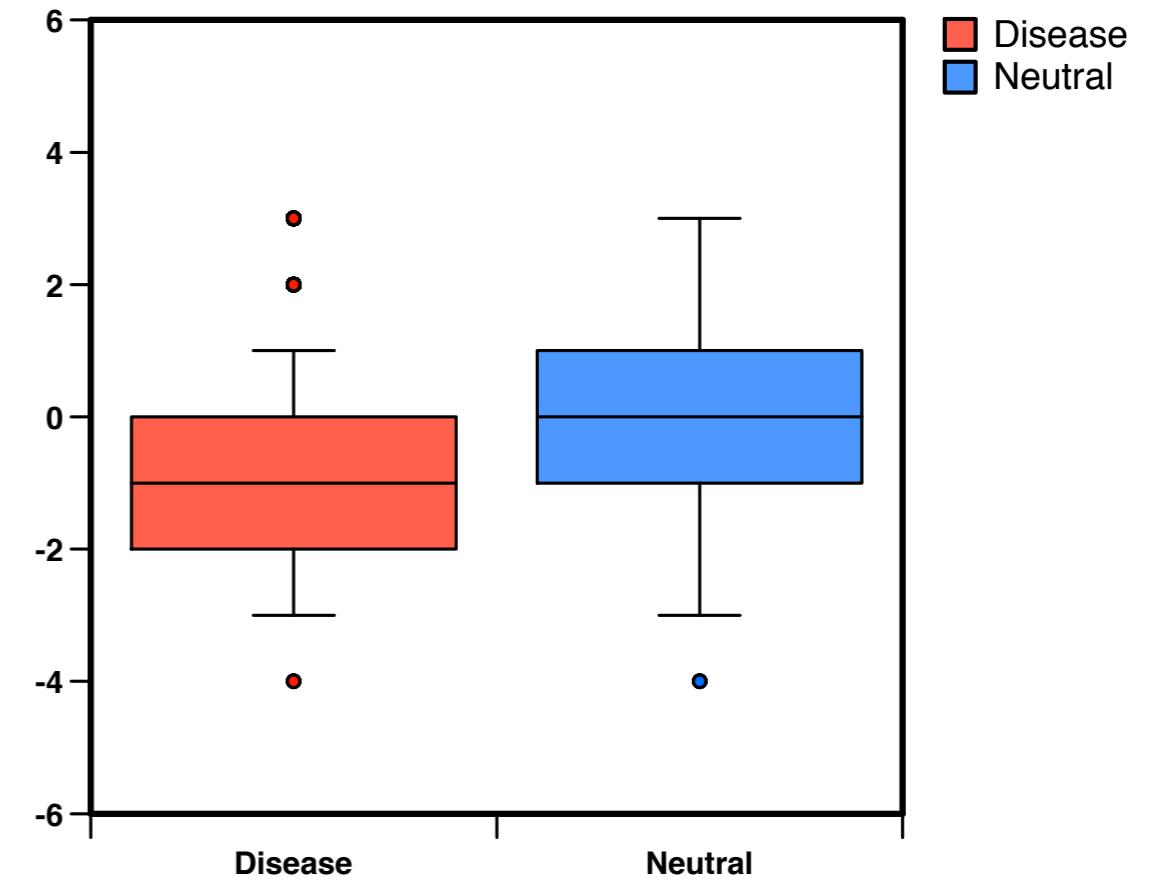
Predicting the impact of mutation on human health is a more complex task that requires the integration of several source of information.

Simple Predictor

A simple method can be developed predicting the impact of mutations using BLOSUM62 substitution matrix.



$$sim(i, j) = \log \frac{P(i, j)}{P(i)P(j)}$$

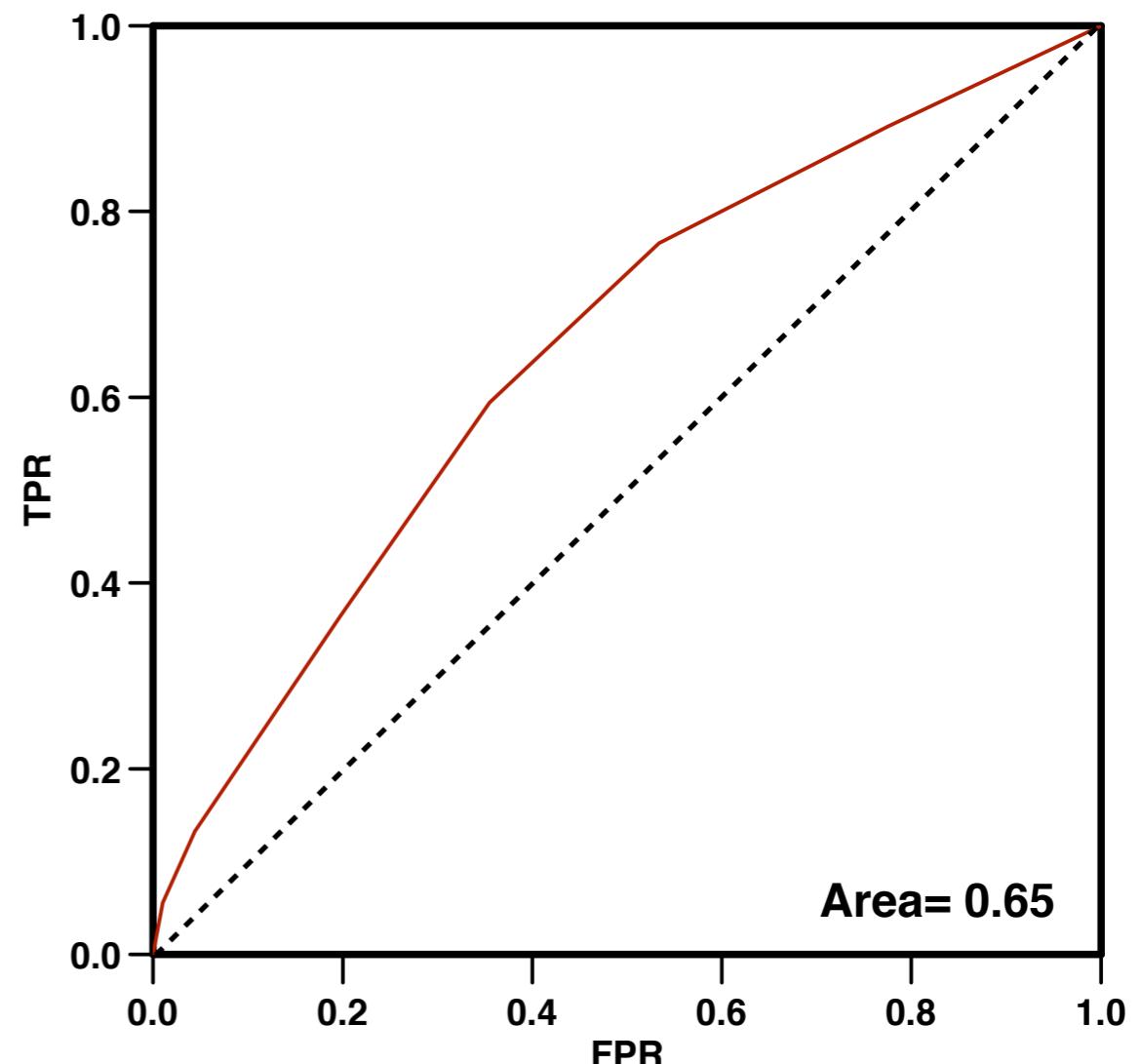


BLOSUM62 Predictions

It is possible to plot the ROC curve of the predictions moving BLOSUM62 threshold from -4 to 3.

We can calculate the Area Under the Curve and optimize the prediction threshold.

If we use a threshold equal to -1 the method result in 64% overall accuracy and 0.24 Matthews' correlation coefficient



	Q2	P[D]	S[D]	P[N]	S[N]	C
BLOSUM62	0.64	0.67	0.77	0.59	0.47	0.24

Accuracy measures

Overall Accuracy

$$Q2 = \frac{TP + TN}{TP + FN + TN + FP}$$

Sensitivity

$$S = \frac{TP}{TP + FN}$$

Precision

$$P = \frac{TP}{TP + FP}$$

Correlation

$$C = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

		Actual values	
		Positive	Negative
Predicted values	Positive	TP	FP
	Negative	FN	TN

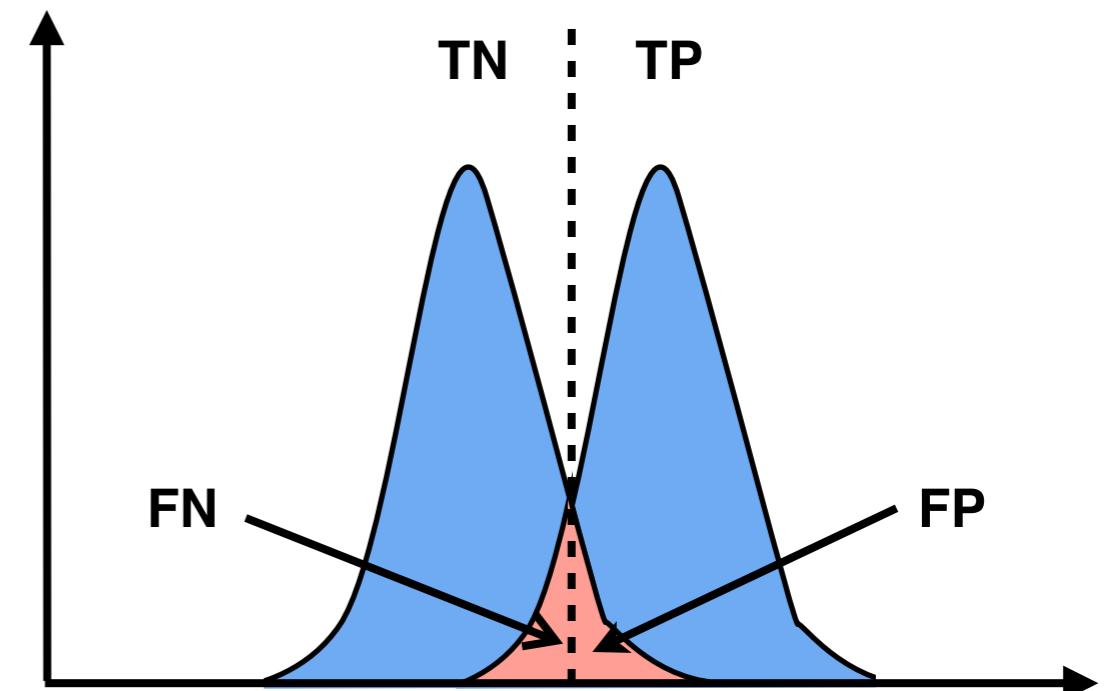
Receiving Operator Curve

True Positive Rate

$$TPR = \frac{TP}{TP + FN}$$

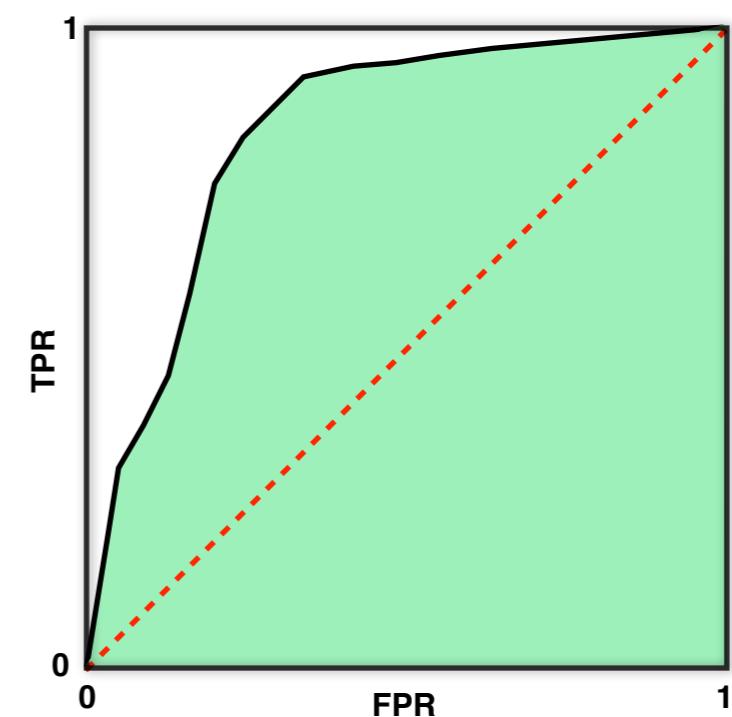
False Positive Rate

$$FPR = \frac{FP}{FP + TN}$$



The **Area Under the ROC Curve (AUC)** is an accuracy measure that is 0.5 for completely random predictors and close to 1.0 for highly accurate predictors.

Baldi et al. (2000) Bioinformatics, 16:412-424



Conserved or not?

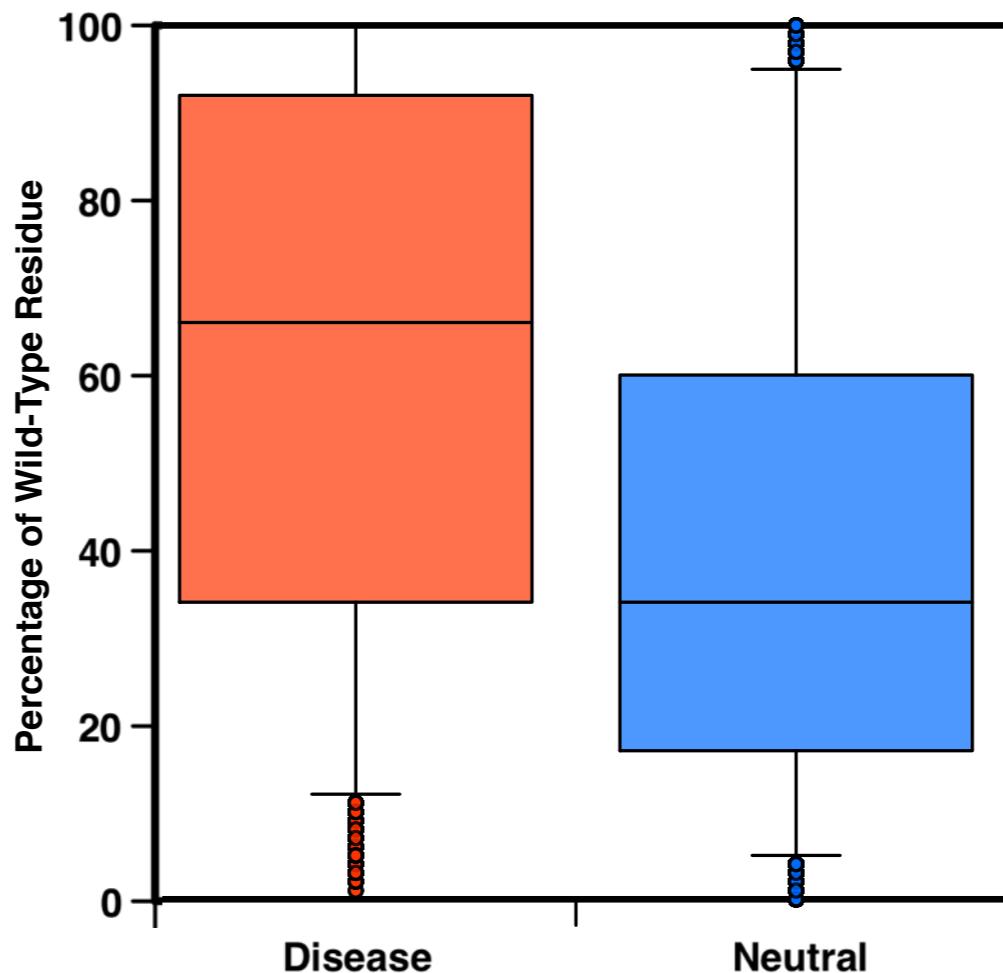
In positions 66 the Glutamic acid is highly conserved Asparagine in position 138 is mutated Threonine or Alanine

Sequence alignment and phylogenetic tree showing the evolution of the SKEVLM motif across various species. The top panel shows a sequence logo for the SKEVLM motif with a red box highlighting the invariant 'K' residue. The bottom panel shows a phylogenetic tree with colored nodes corresponding to the sequence motifs shown above. The tree root is at the top, and the sequence positions 81 and 160 are indicated.

Sequence profile

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

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



Machine learning

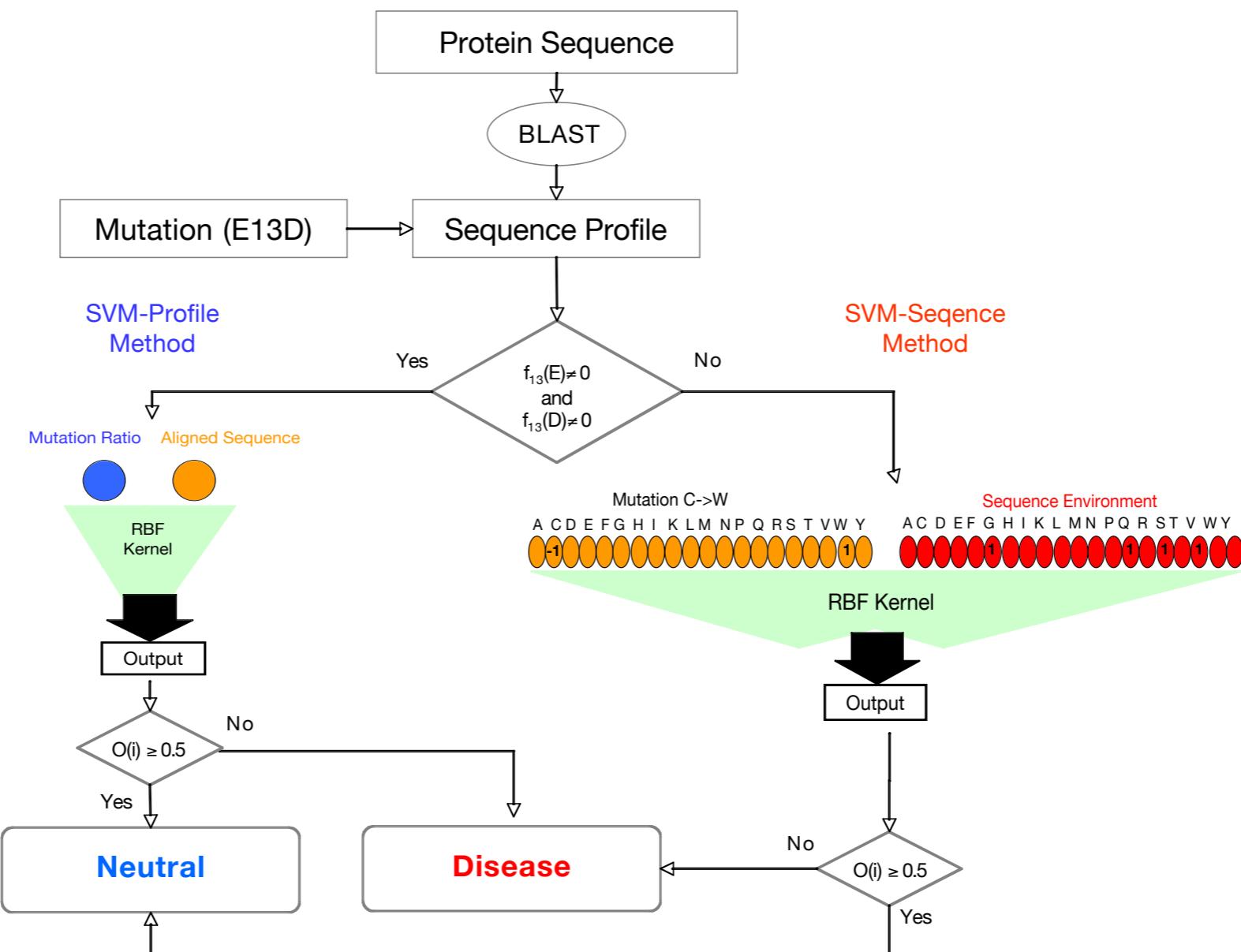
- Computational approach to build models based on the analysis on empirical data.
- Machine learning algorithms are suitable to address problems for which analytic solution does not exists and large amount of data are available.
- They are implemented selecting a representative set of data that are used in a training step and then validated on a test set with data “*not seen*” during the training.
- Most popular machine learning approaches are in computational biology are Neural Networks, Support Vector Machines and Random Forest.

Binary classifiers

- Support Vector Machine (SVM): Maps positive and negative training examples to a high-dimensional space in which they can be distinguished from each other.
- Artificial Neural Network (ANN): multi-layer network of nodes, including input features, outputs, and one or more hidden layers. Weights of input and output edges connecting nodes are adjusted to maximize prediction accuracy.
- Random Forest (RF): Trains an “ensemble” of decision trees to distinguish positive from negative training examples, utilizing a random set of input features.
- Naïve Bayes Classifiers: Probabilistic classifier that treats each feature as independent of the others; parameters are adjusted to maximize the probability of impact for positive examples and minimize probability for negative examples.

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.



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.

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.

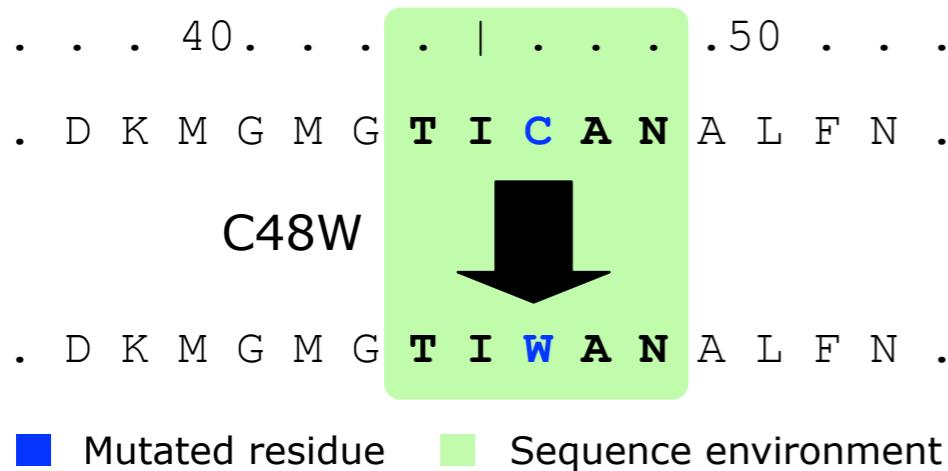


<http://www.geneontology.org/>

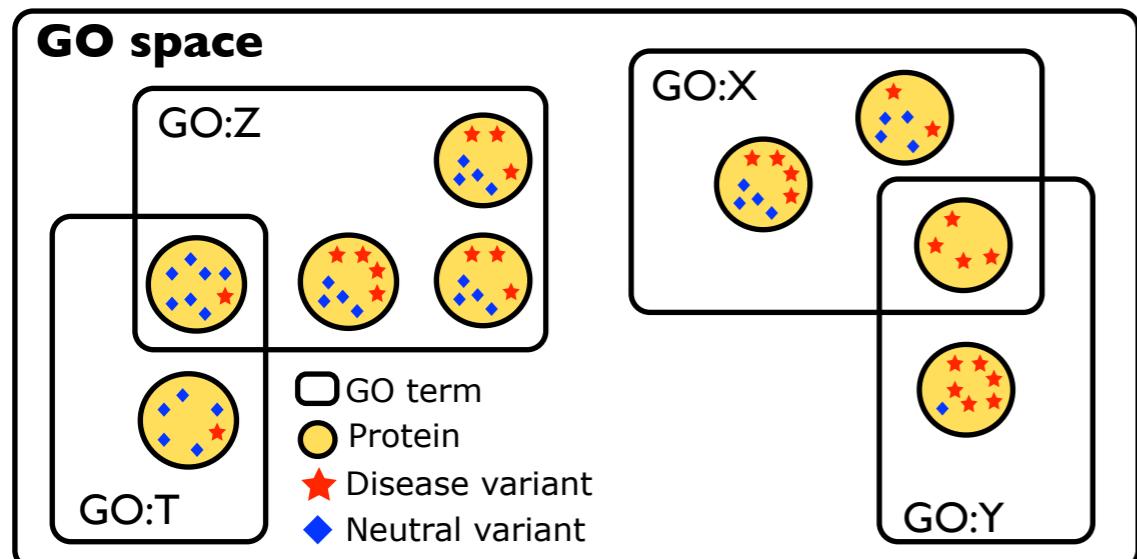
The ontology is represented by a direct acyclic graph covers three domains;

- 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.

Prediction features



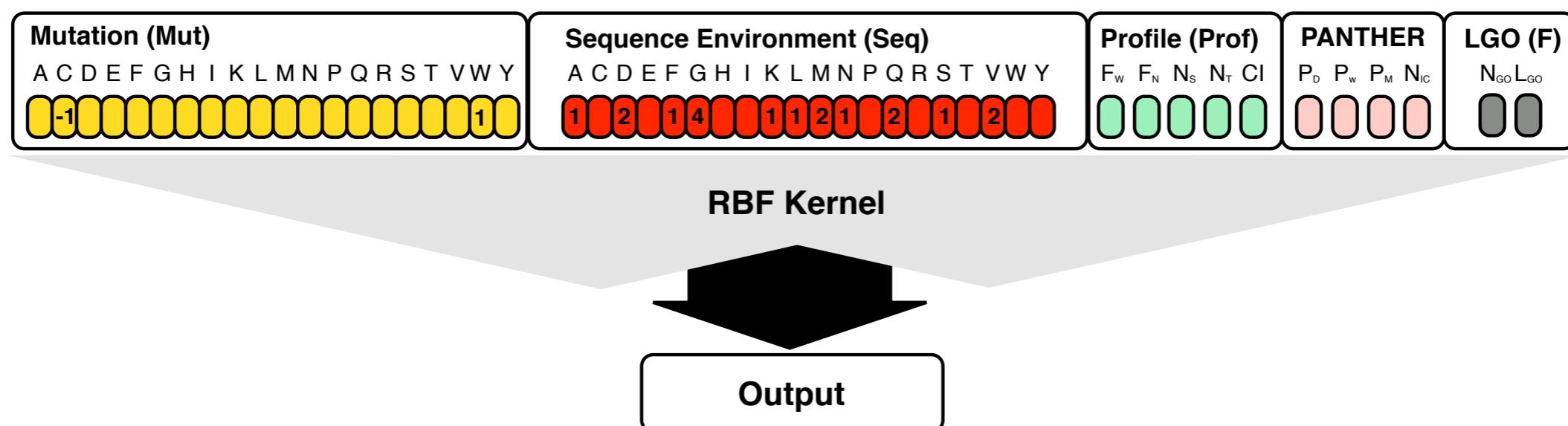
Protein sequence profile information derived from a multiple sequence alignment. It is encoded in a 5 elements vector corresponding to different features general and local features



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

SwissVar data

SwissVar (October 2009)

- Disease variants: 22,771
 - Neutral variants: 34,258
 - Unclassified variants: 2,269
 - **Total: 59,298**
-
- Disease-related mutations not clearly annotated are removed.
 - Mutations related to more than one disease are considered only once.

Training set

After this filter we collected 17,993 Disease mutations from 1,424 proteins that are balanced with the same number of neutral polymorphisms.

Protein structure data

The mapping of SwissVar mutations data on the structures available on the PDB is a difficult task. The main problems for this task are:

- incomplete PDB structures
- differences between Swiss-Prot protein sequence and PDB sequence
- different residue numeration

The mapping procedure is performed using a pre-filtered list of correspondences between Swiss-Prot and PDB.

All Swiss-Prot/PDB pairs in the list are aligned using BLAST. To have a good overlap between sequence and structure I filtered the list of alignments removing those:

- with ≥ 1 gaps
- sequence identity $< 100\%$
- shorter than 40 residues

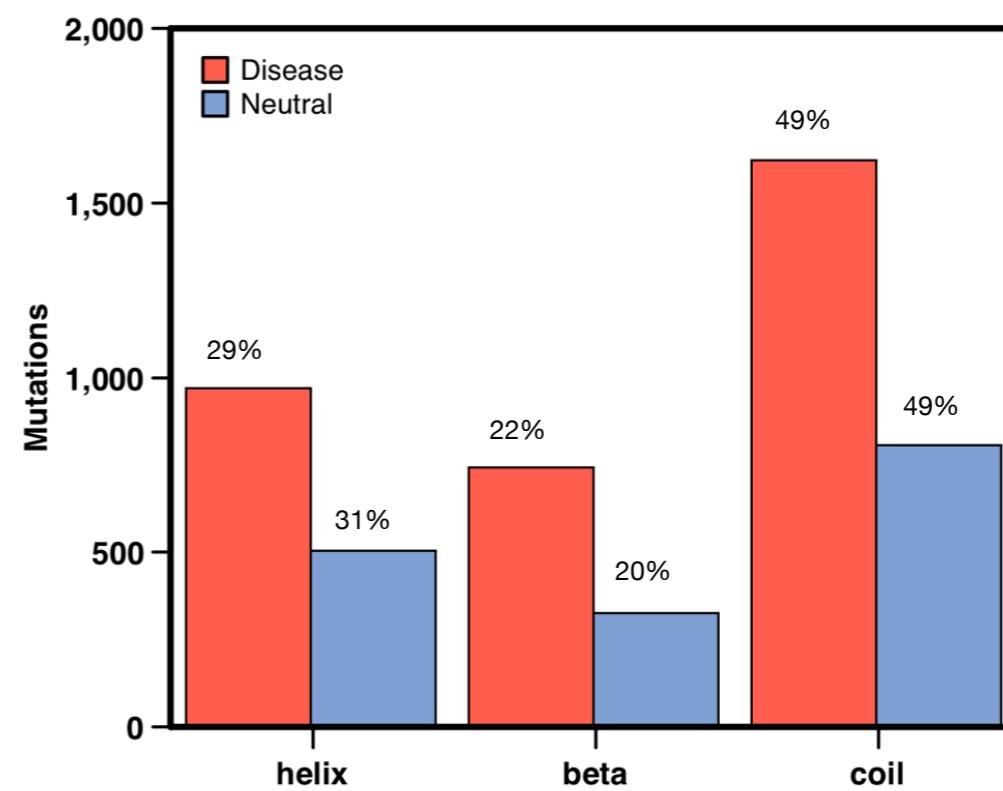
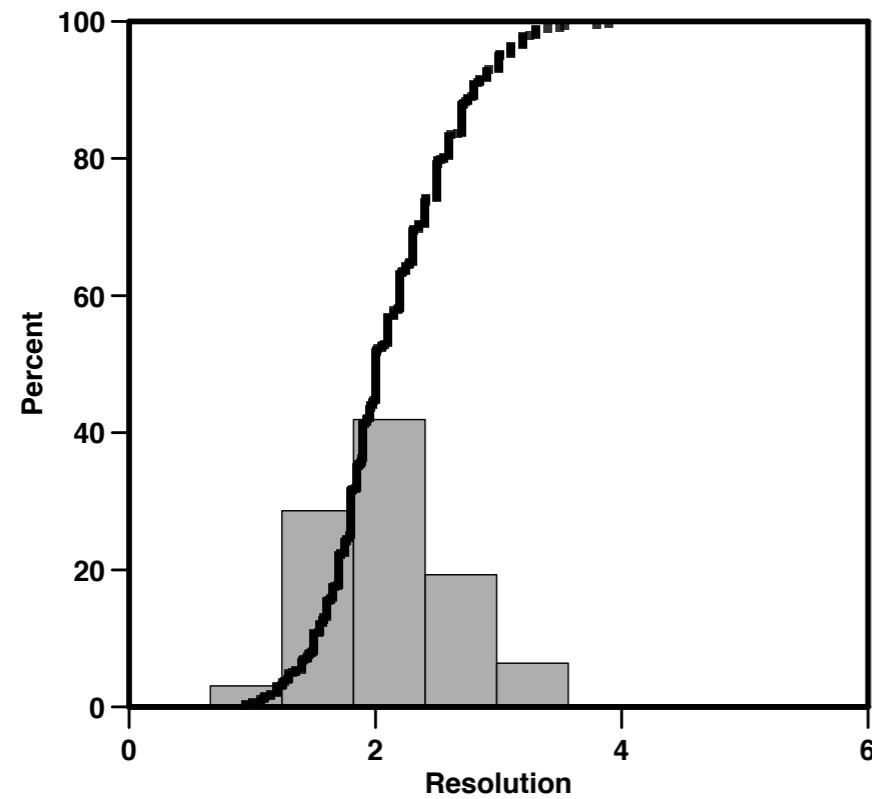
If one mutation maps on more than one PDB the one with lower resolution is selected

3D Structure Dataset

After the mapping procedure the final dataset of mutations with known 3D structure is composed by

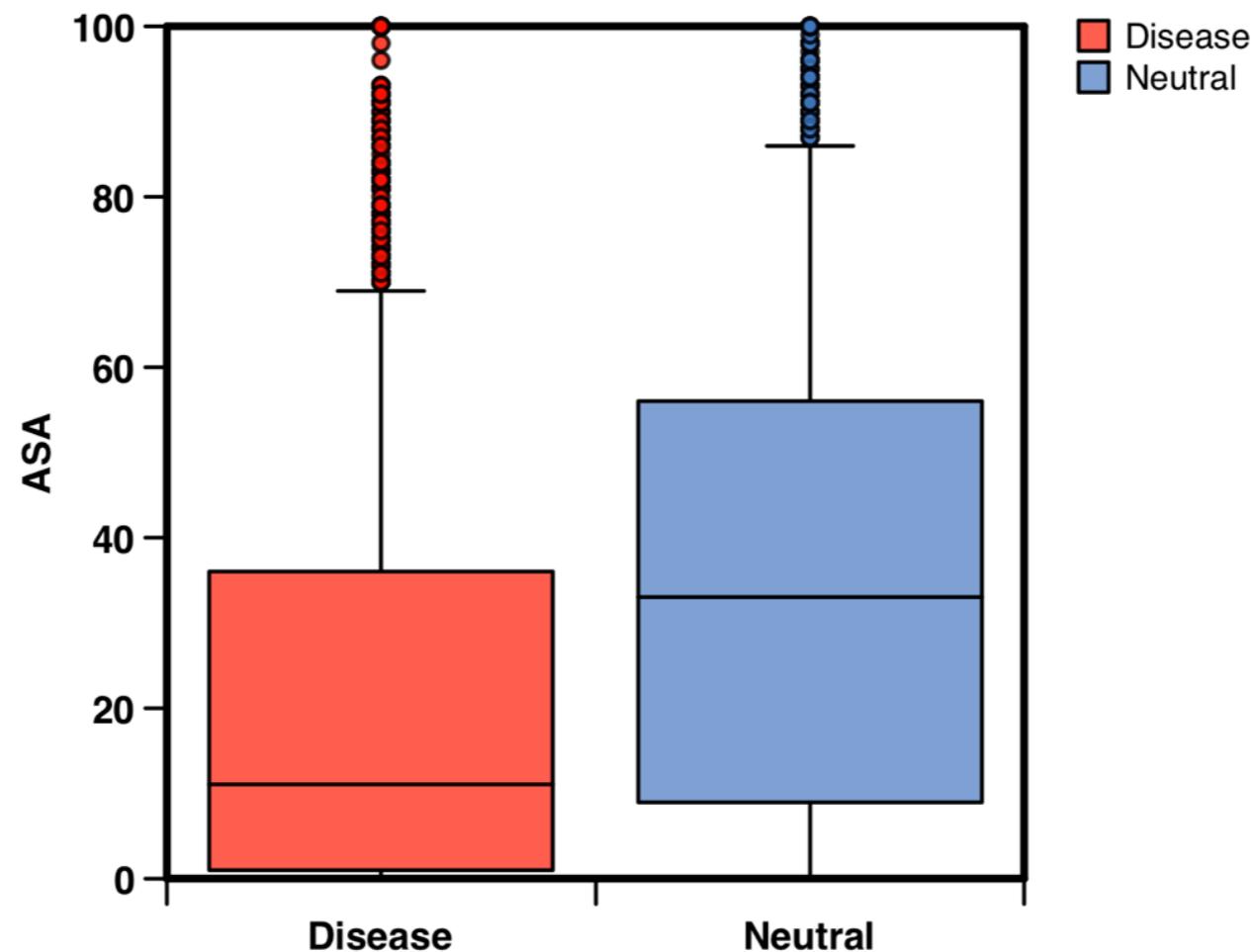
- Disease variants: 3,342
- Neutral variants: 1,644
- Total: 4,986

from 784 chains from 770 structures (584 X-ray, 92 NMR and 94 models).



Structure environment

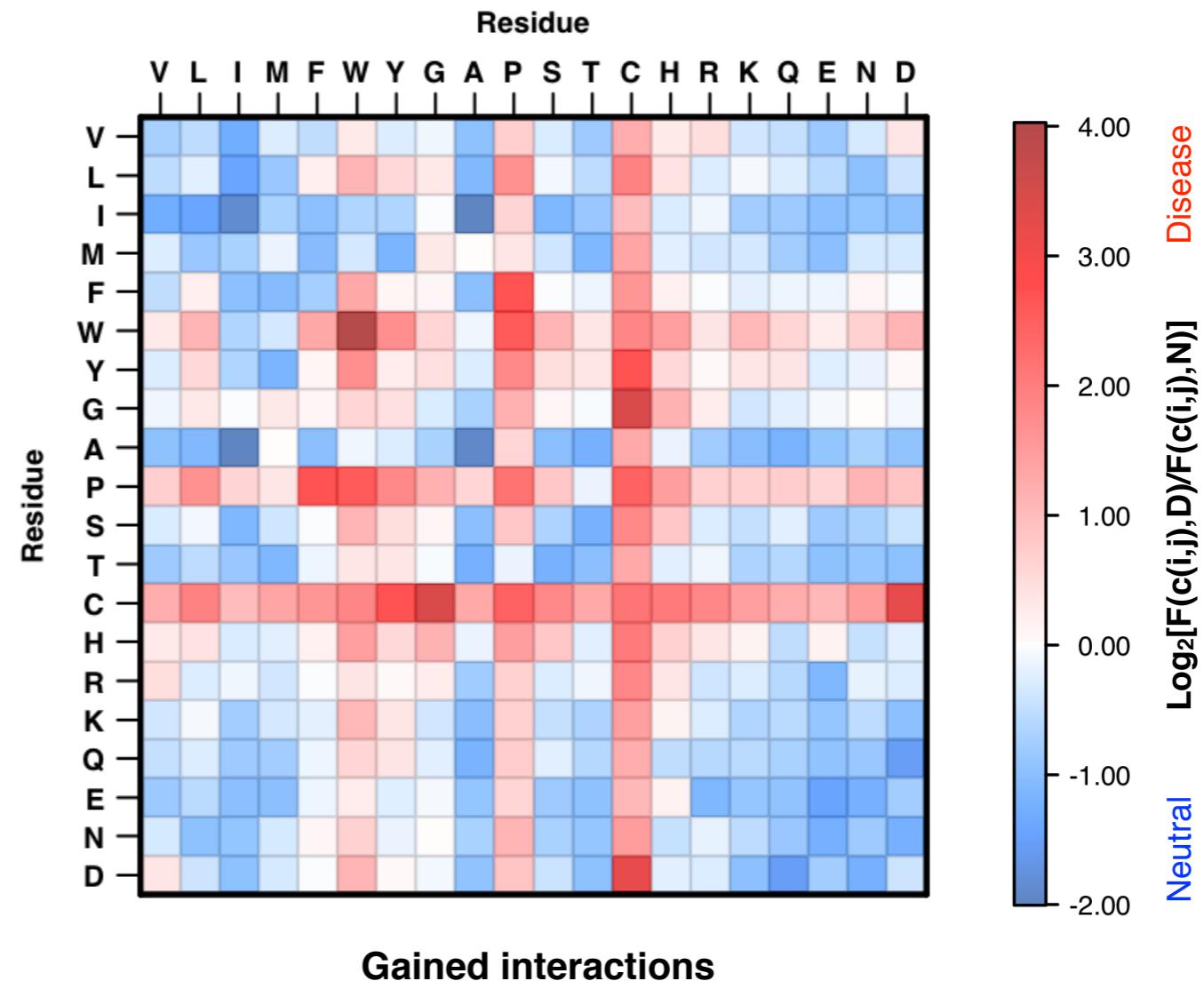
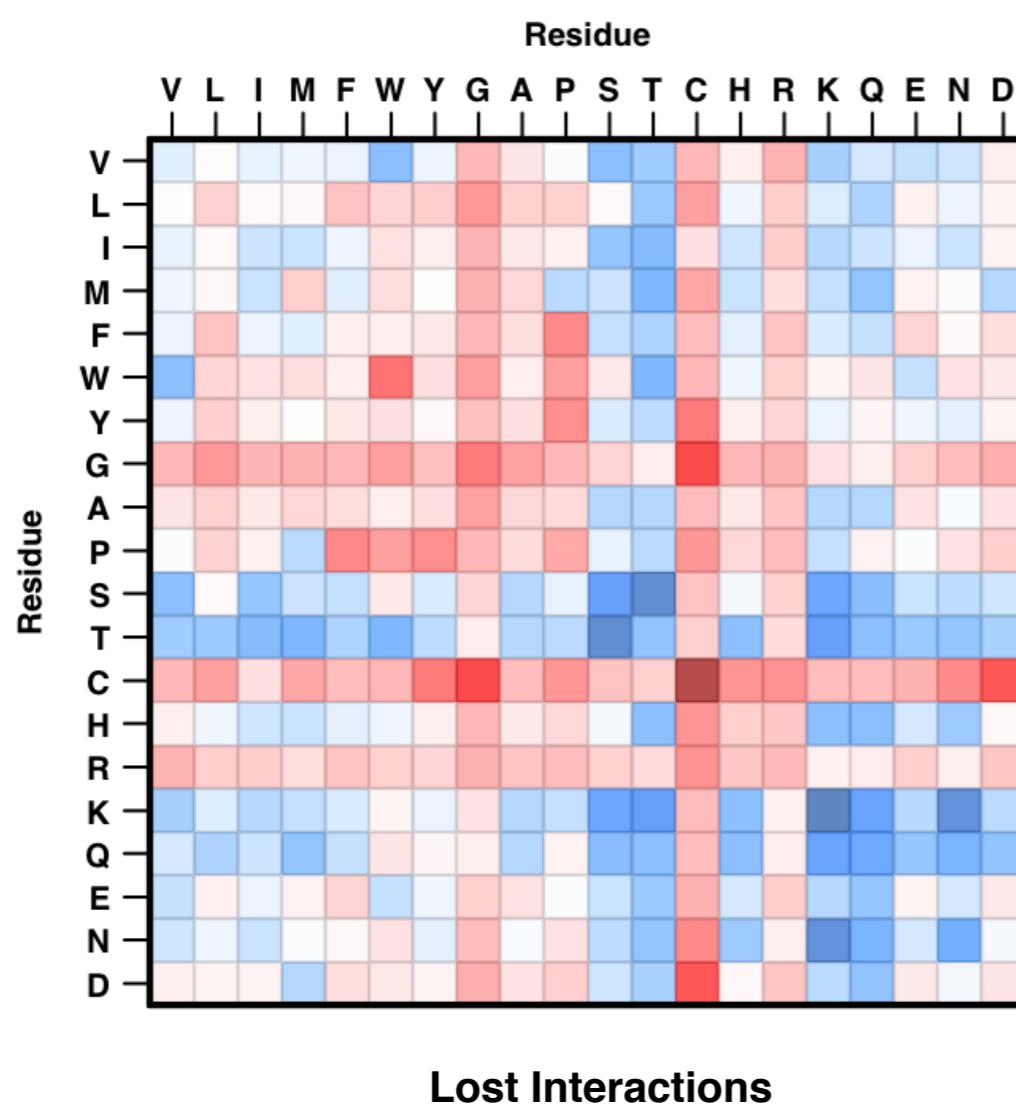
There is a significant difference (p-value $KS < 0.001$) between the distributions of the relative Accessible Solvent Area for disease-related and neutral variants. Their mean values are respectively 20.6 and 35.7.



Analysis of the 3D interactions

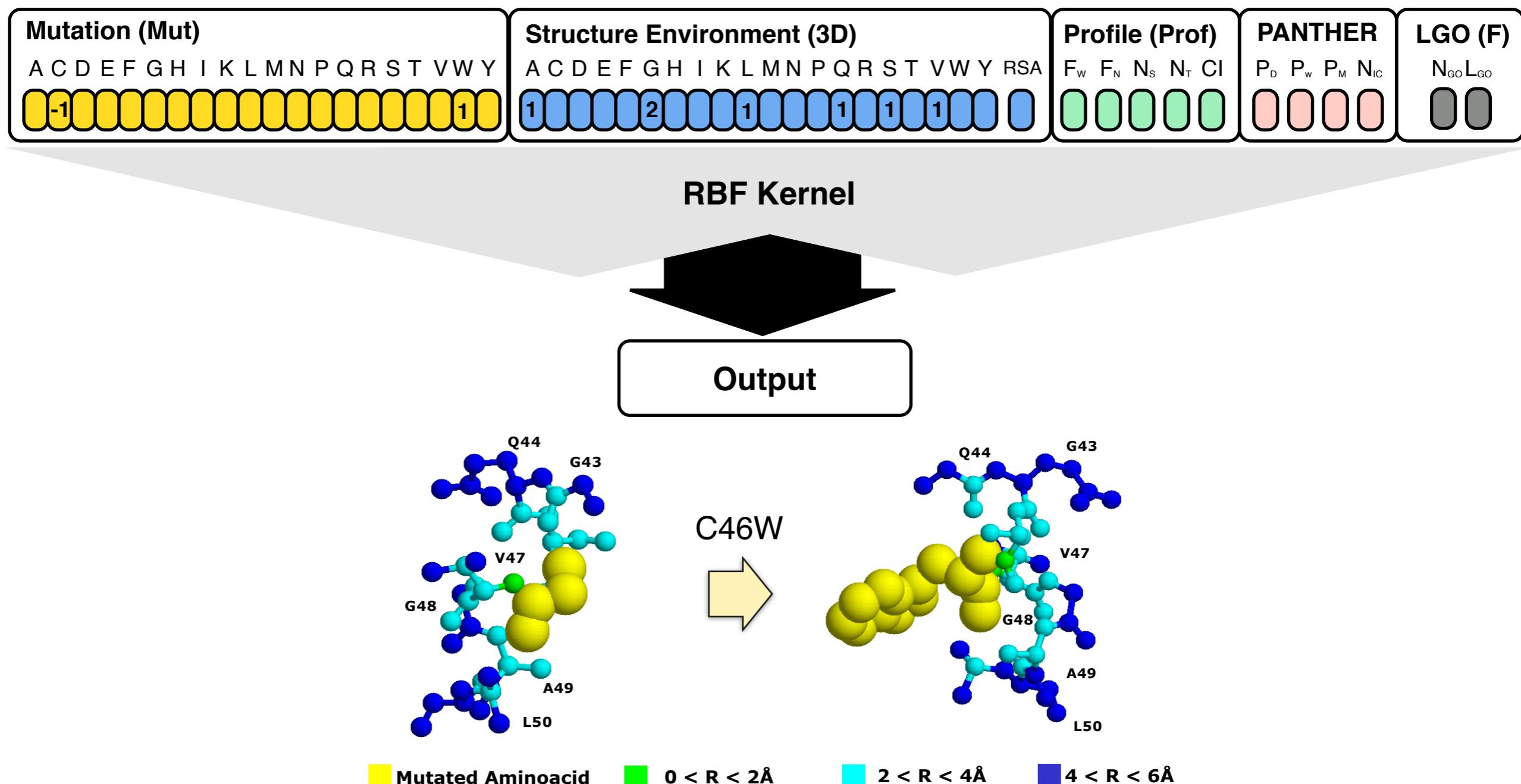
Using the whole set of SAVs with known structure, we calculate the log odd score of the ratio between the frequencies of the interaction between residue i and j for disease-related and neutral variants.

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



The structure-based method

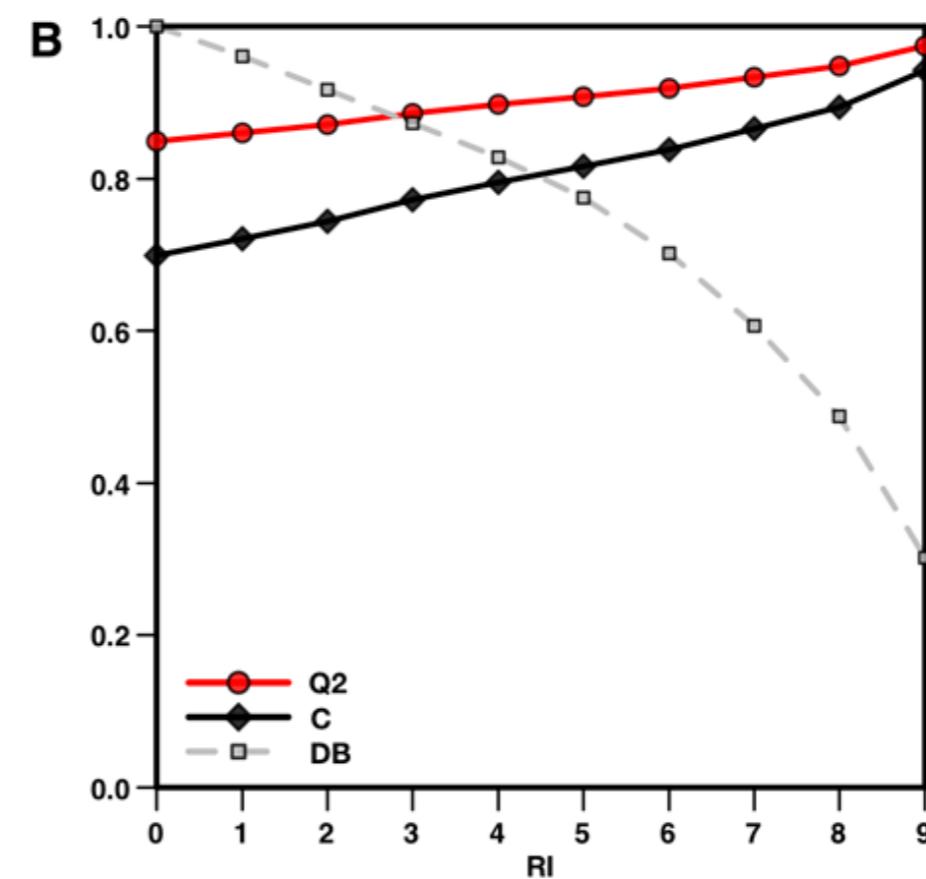
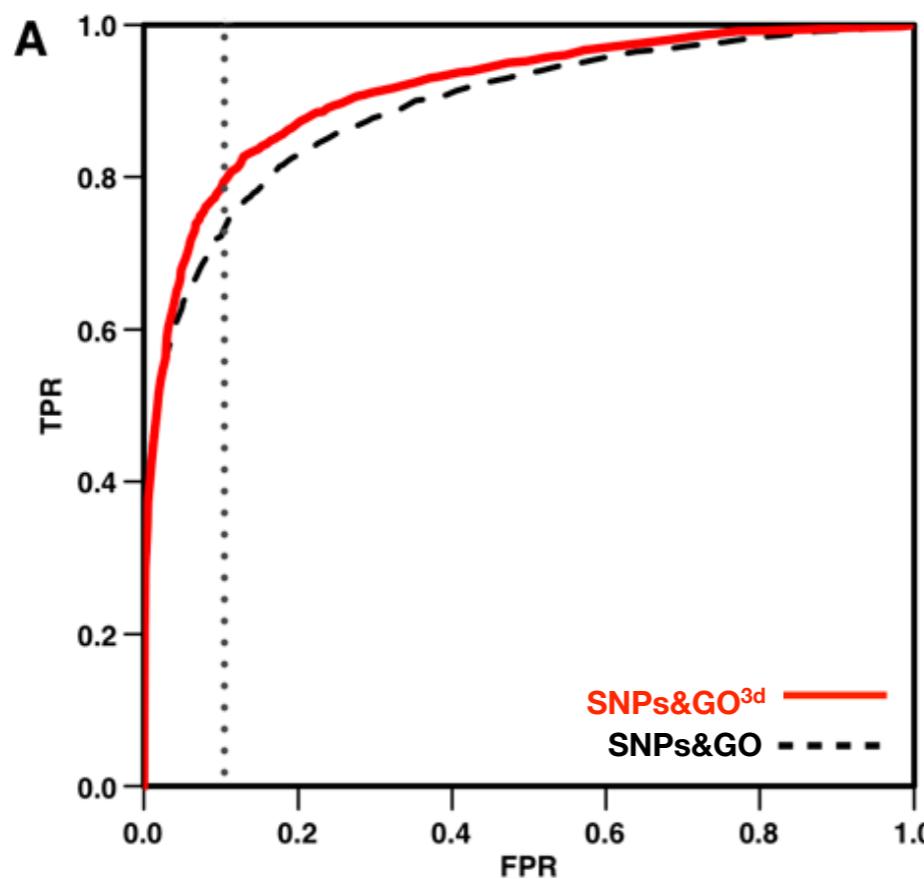
The method takes in to account 5 different types of information encoded in a **52 elements vector**. The **input features** are: mutation data; structure environment, sequence profile and functional score based on GO terms.



Sequence vs structure

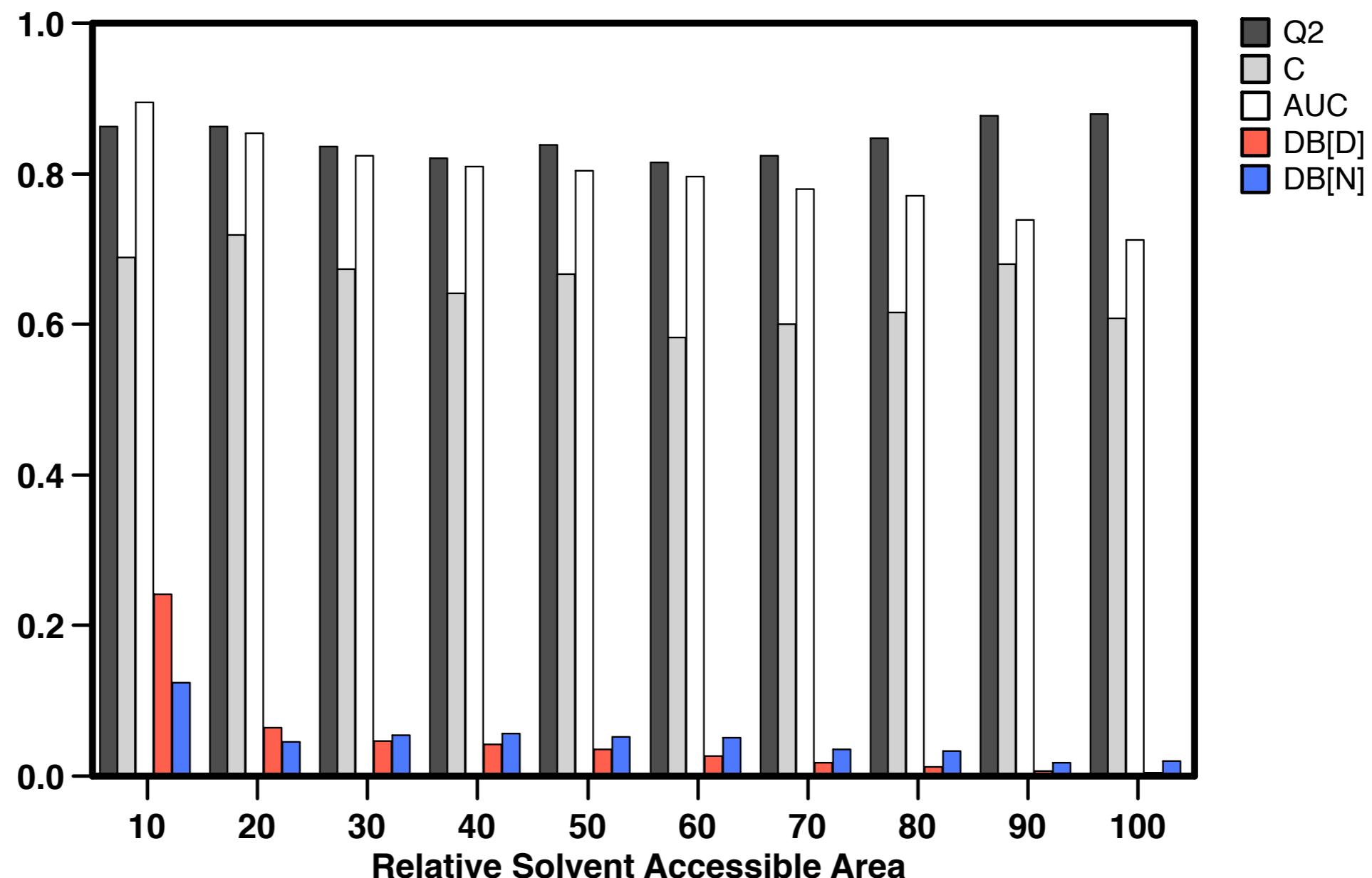
The structure-based method results in better accuracy with respect to the sequence-based one. Structure based prediction are 3% more accurate and correlation coefficient increases of 0.06. If 10% of FPR are accepted the TPR increases of 7%.

	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



Accuracy vs Accessibility

The predictions are more accurate for mutations occurring in buried region (0-30%). Mutations of exposed residues results in lower accuracy.

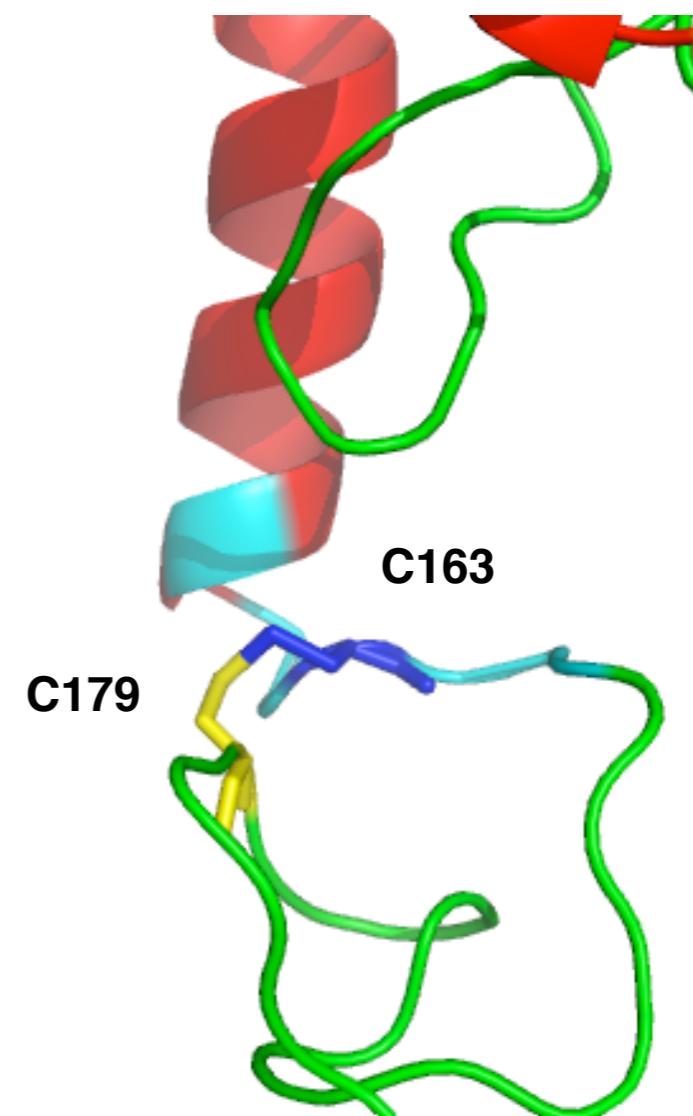


Prediction example

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.

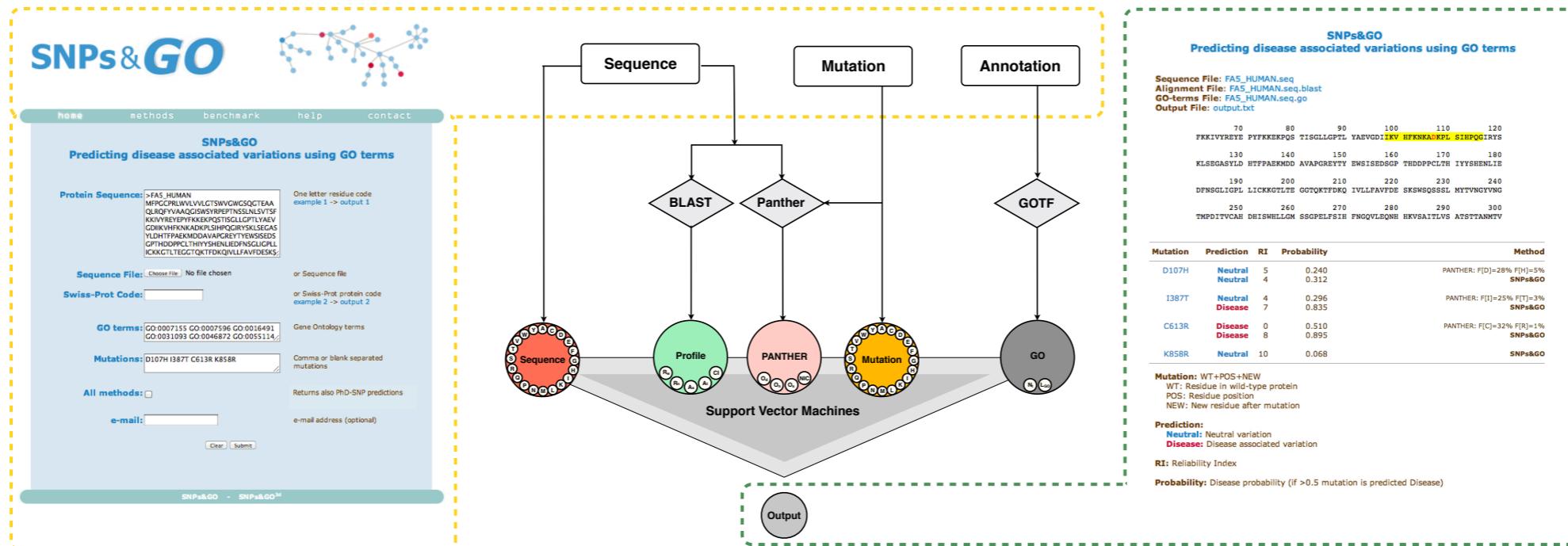


1APY: Chain A, Res: 2.0 Å

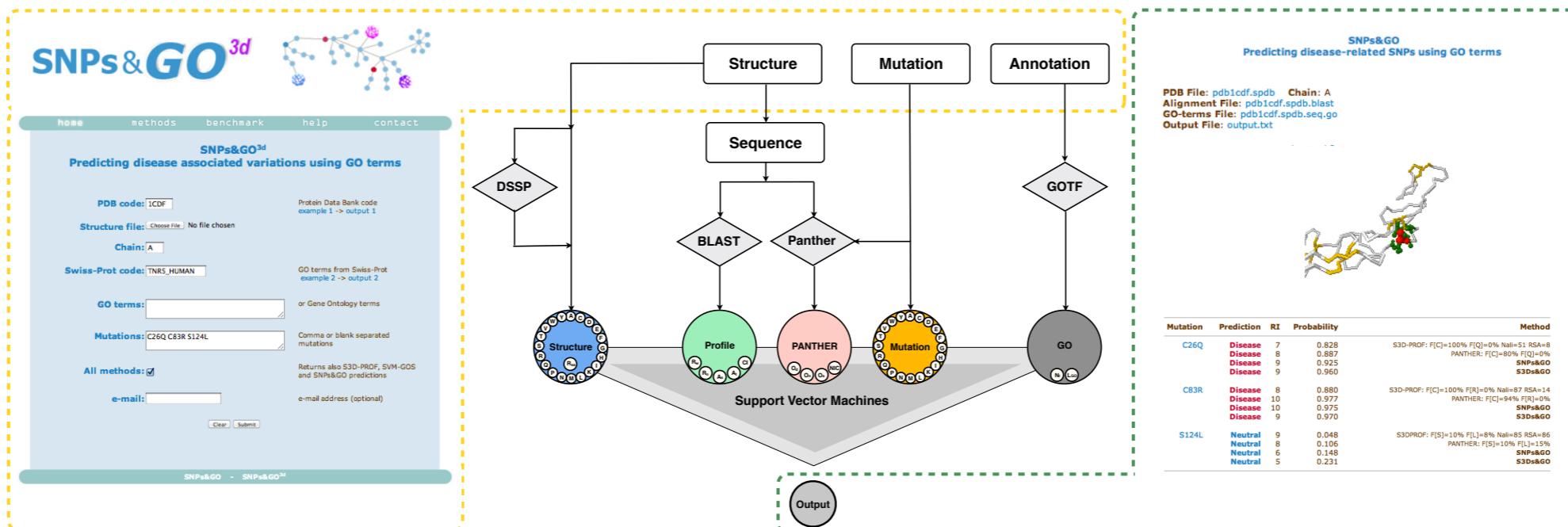


SNPs&GO web server

A



B



<http://snps.biifold.org/snps-and-go>

Capriotti et al. (2013). BMC Genomics. 14 (S3), S6.

SAVs Predictors

Many predictor of the effect of 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 point 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

SAVs Predictors

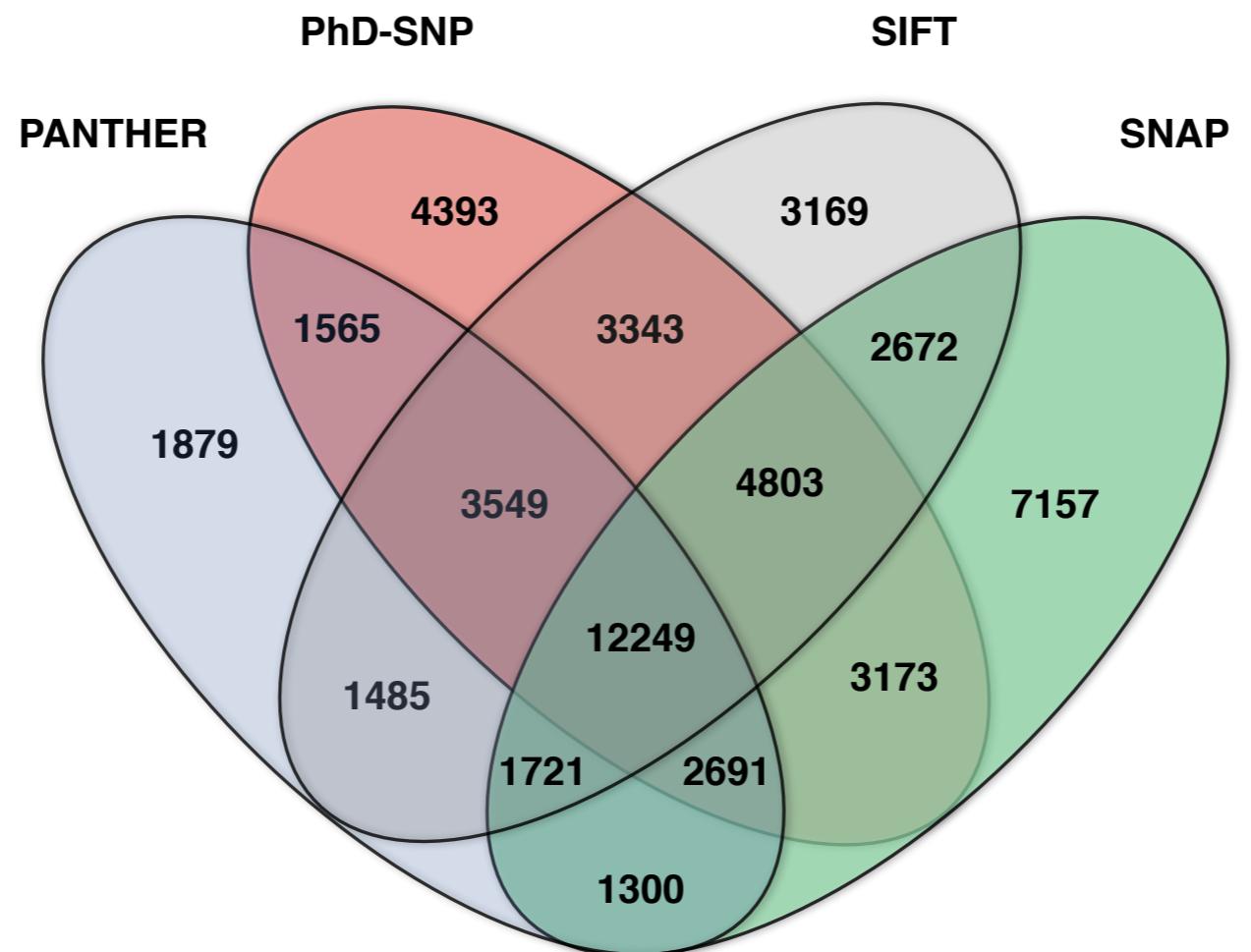
The higher correlation coefficient is between PANTHER and SIFT predictions. SNAP shows low correlation with PhD-SNP and PANTHER but higher correlation with SIFT which input is included in SNAP

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	-

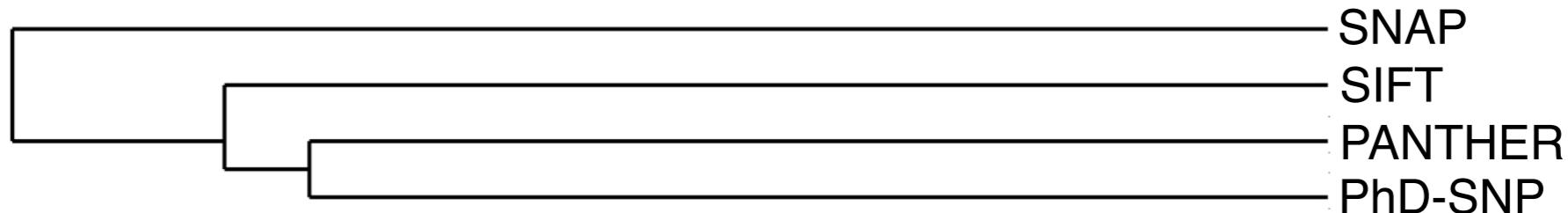
DB: Neutral 17993 and Disease 17993

Predictors tree

Using the prediction similarity we can build the predictors tree



UPGMA tree based on correlations



Prediction Analysis

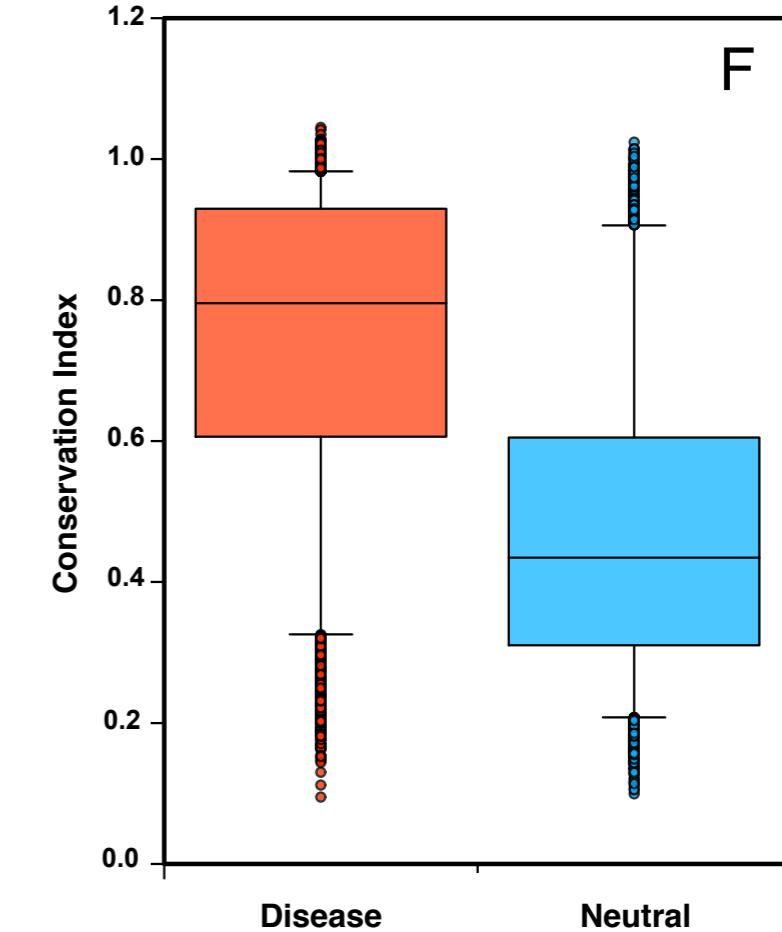
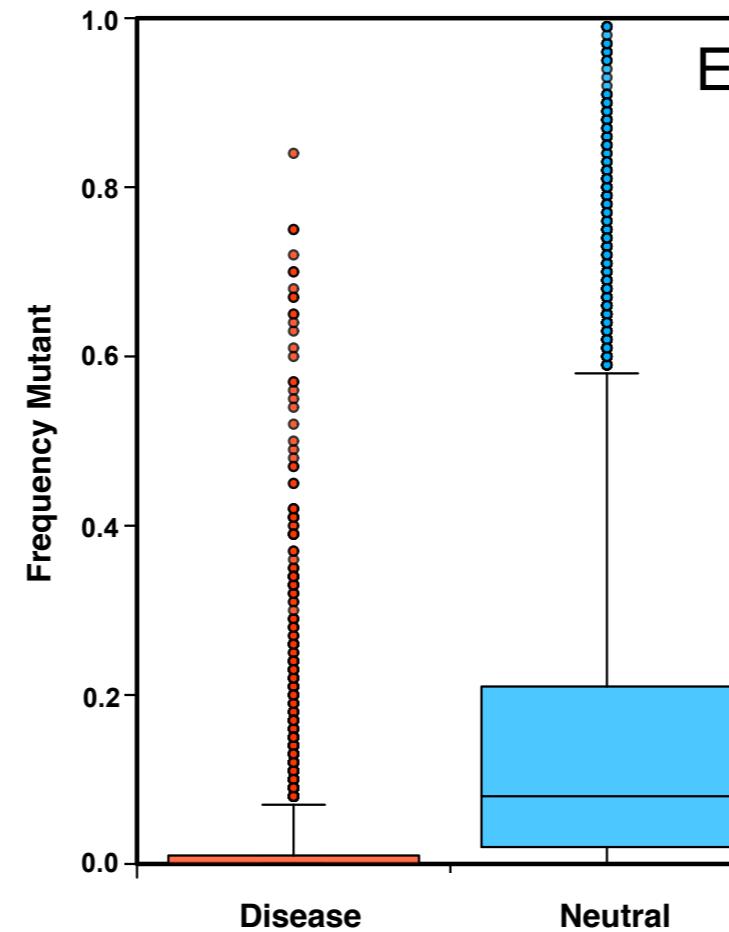
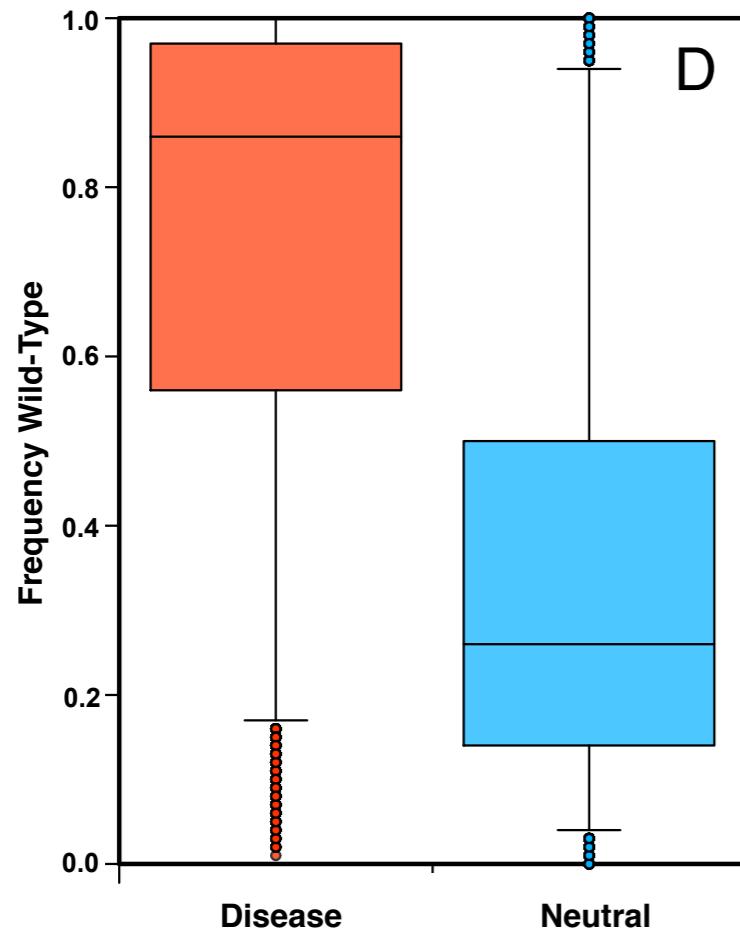
The accuracy of the predictions has been evaluated considering three different subset

- **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

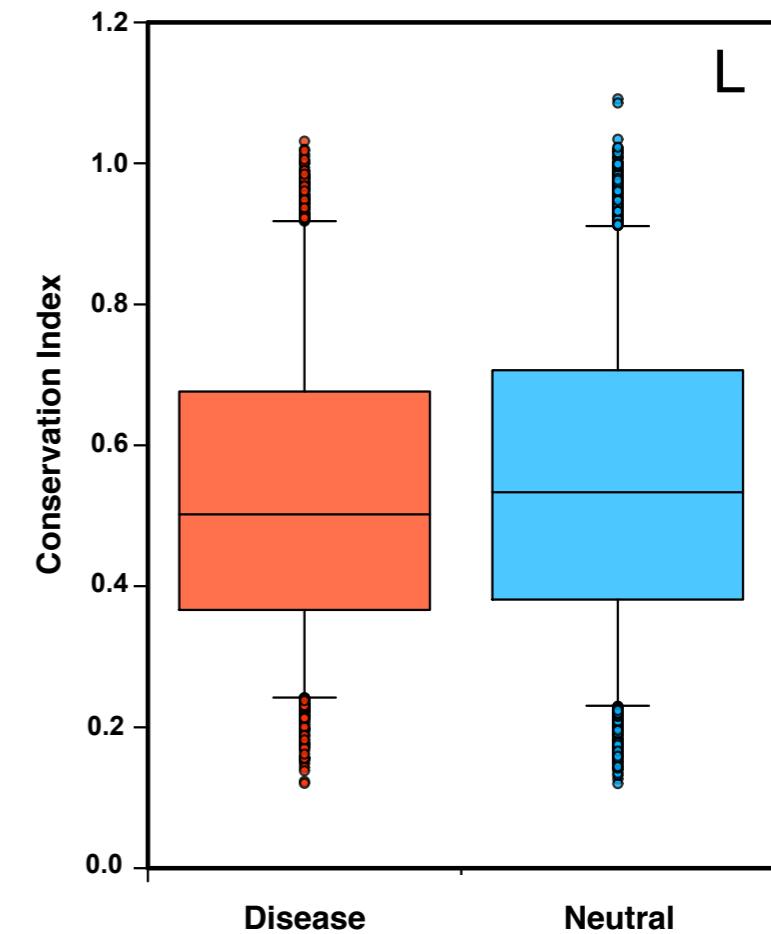
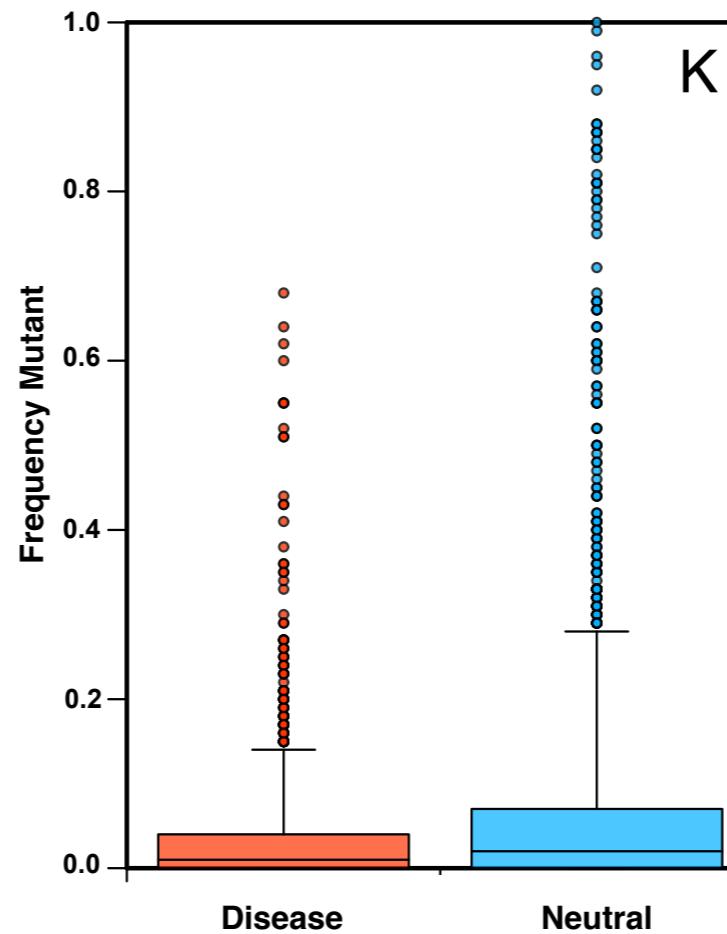
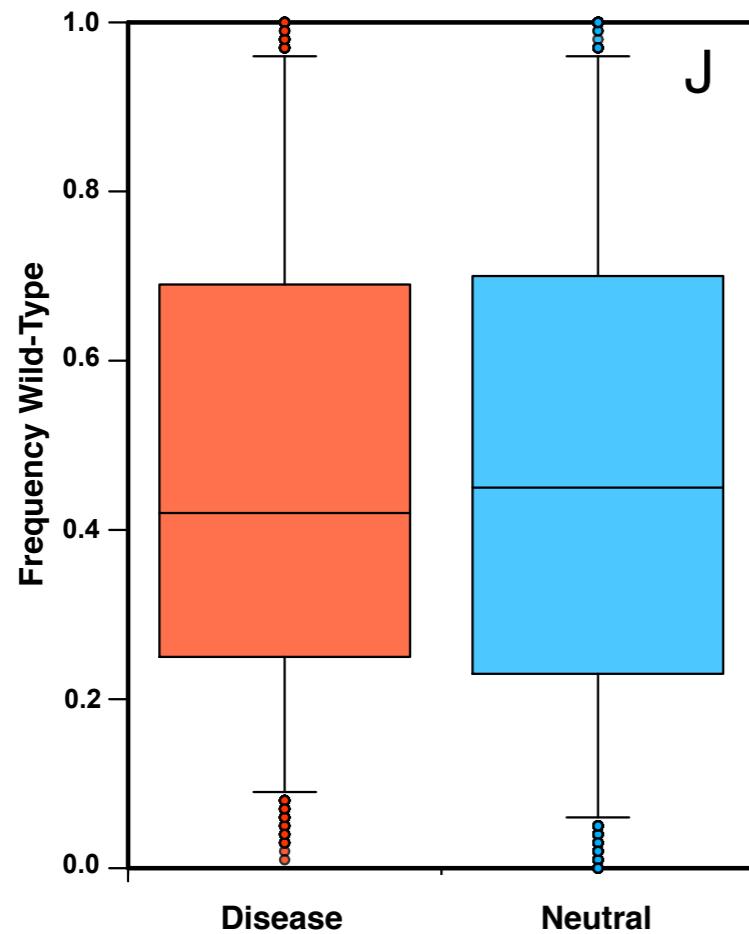
Consensus subset

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.



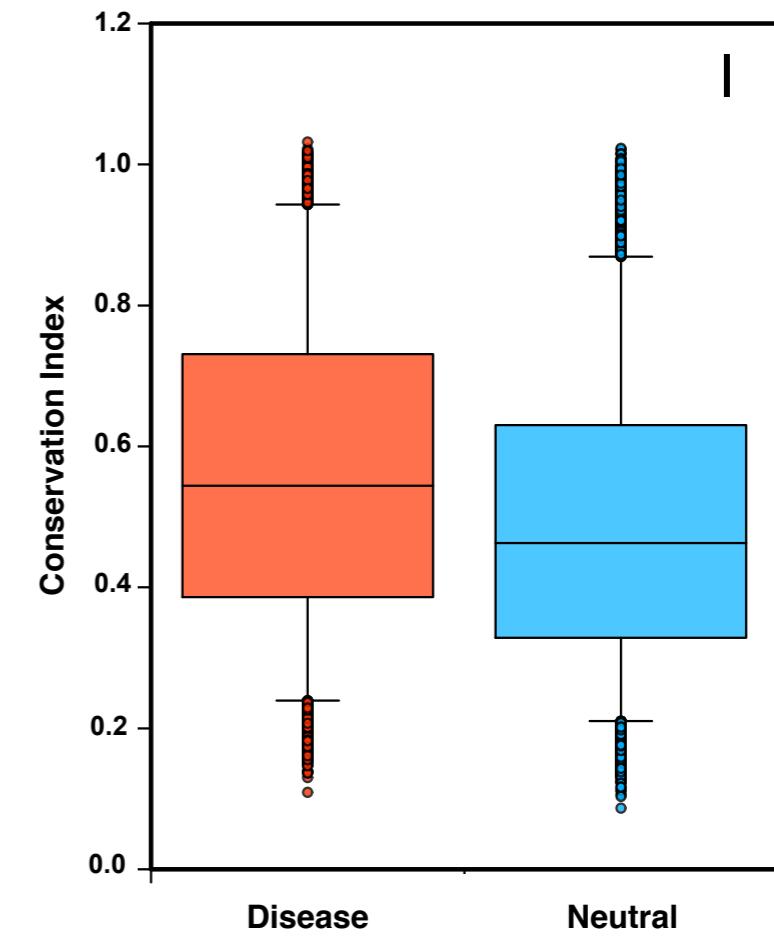
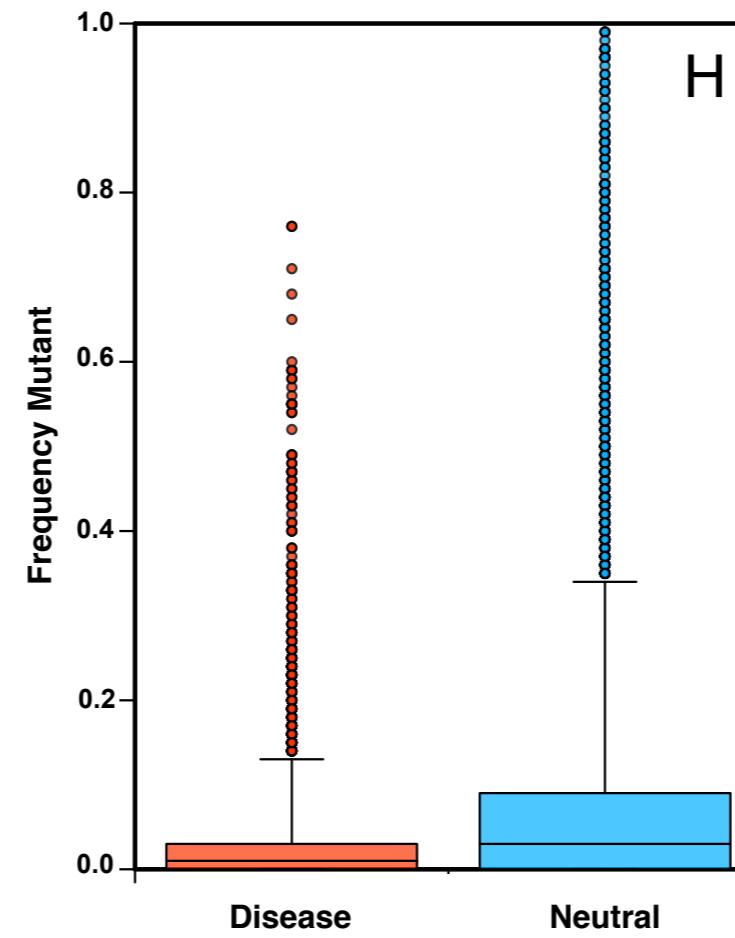
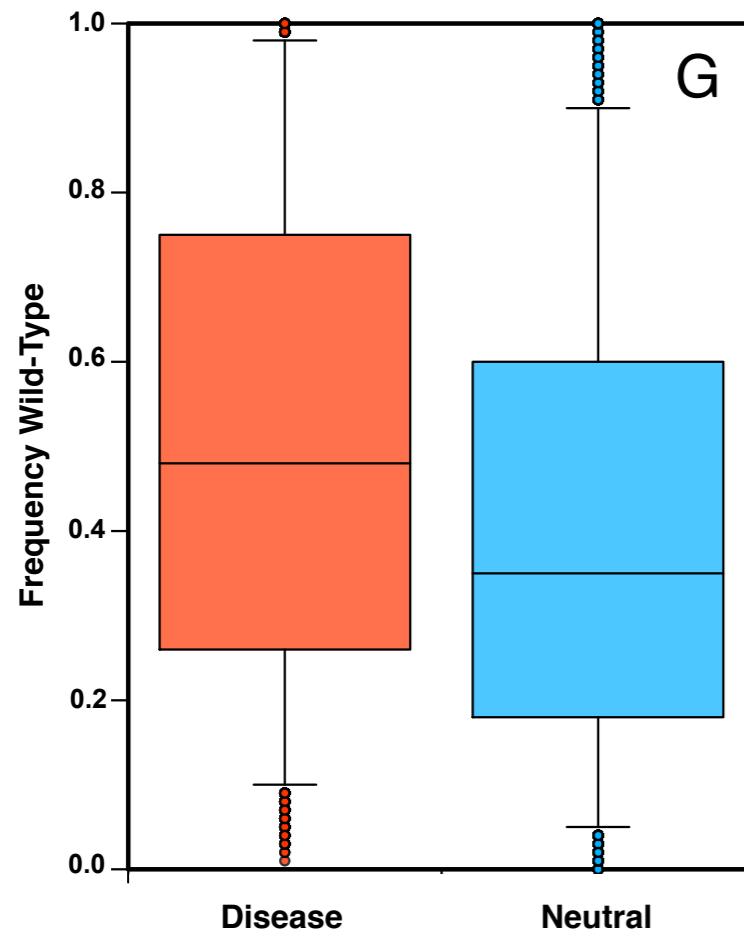
Tie subset

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

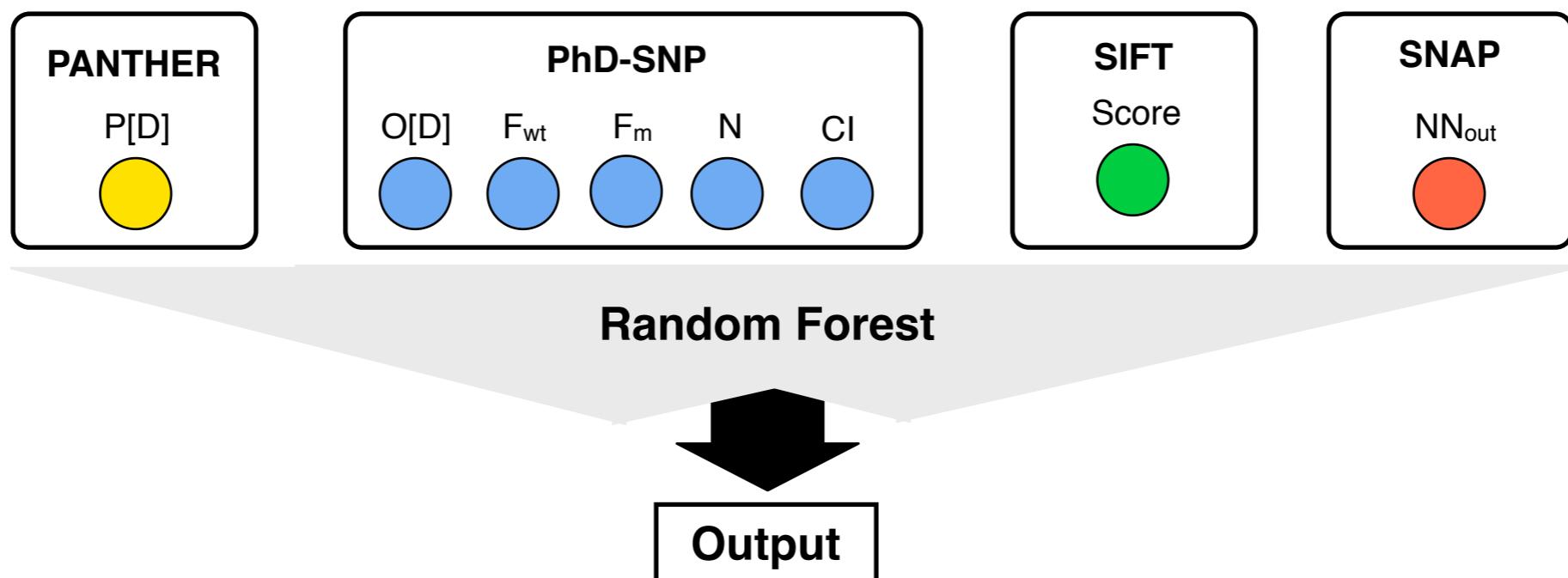
The distributions of the wild-type and new residues frequencies and CI for disease-related and polymorphism on the *Majority subset* are in an intermediate situation with respect to the previous cases.



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) .

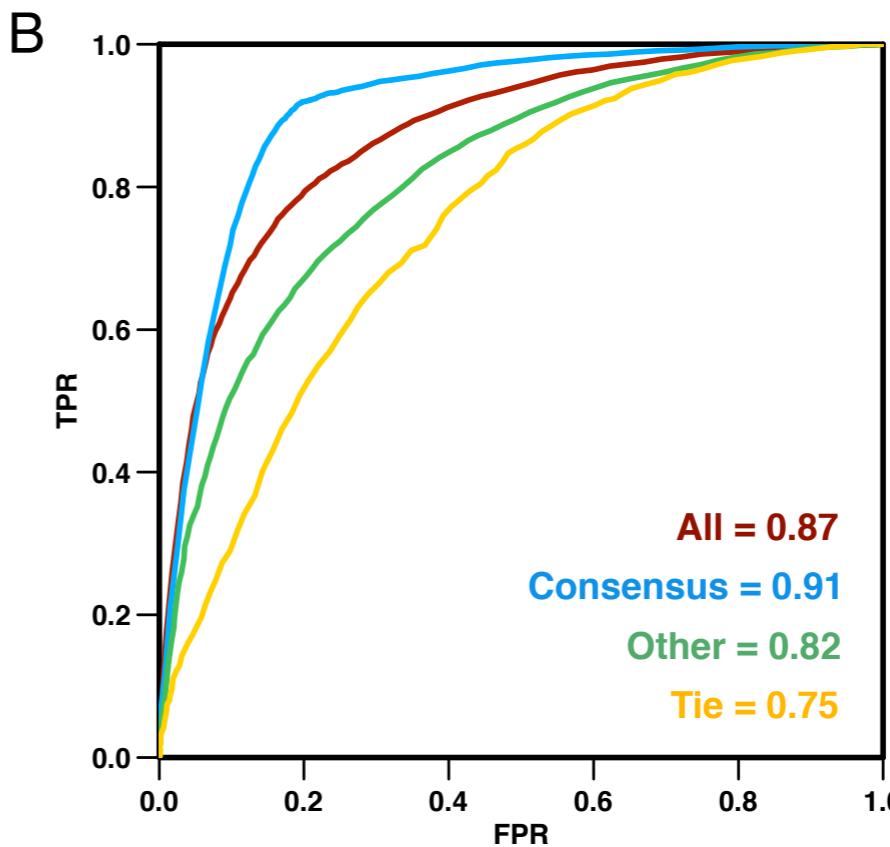
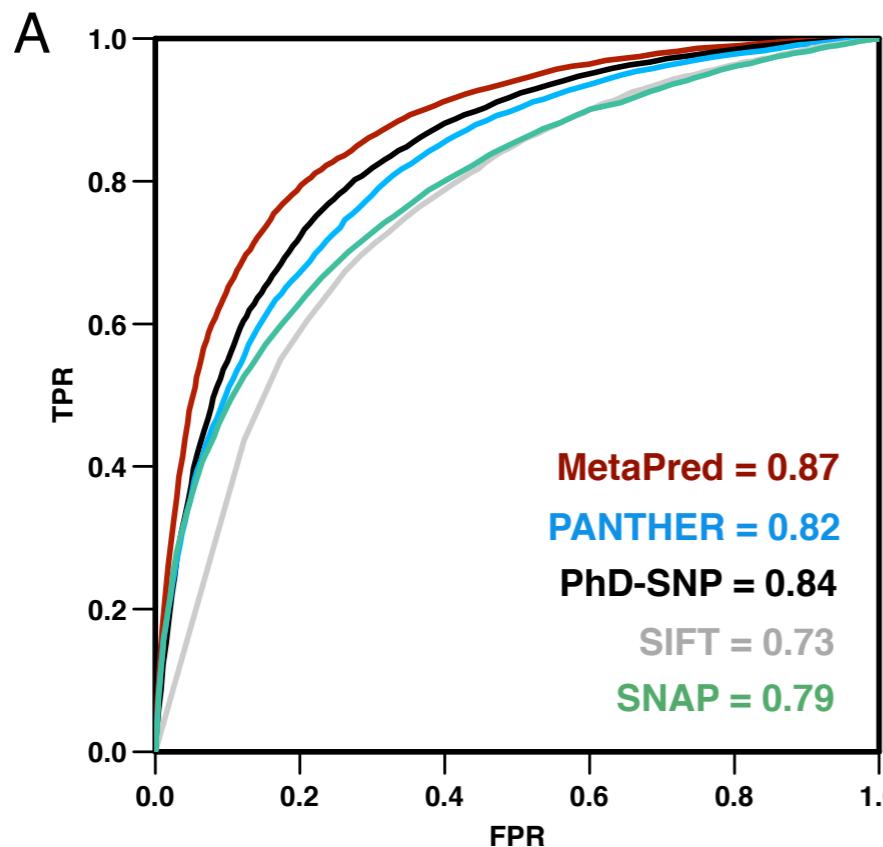


Meta-SNP accuracy

The Meta-Pred method results in better accuracy with respect to the PhD-SNP.

	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
Meta-SNP	0.79	0.80	0.79	0.79	0.80	0.59	0.87	100
Consensus	0.87	0.88	0.92	0.87	0.80	0.73	0.91	46
Majority	0.75	0.72	0.64	0.76	0.82	0.47	0.82	40
Tie	0.69	0.62	0.57	0.73	0.76	0.34	0.75	14

DB: Neutral 17993 and Disease 17993

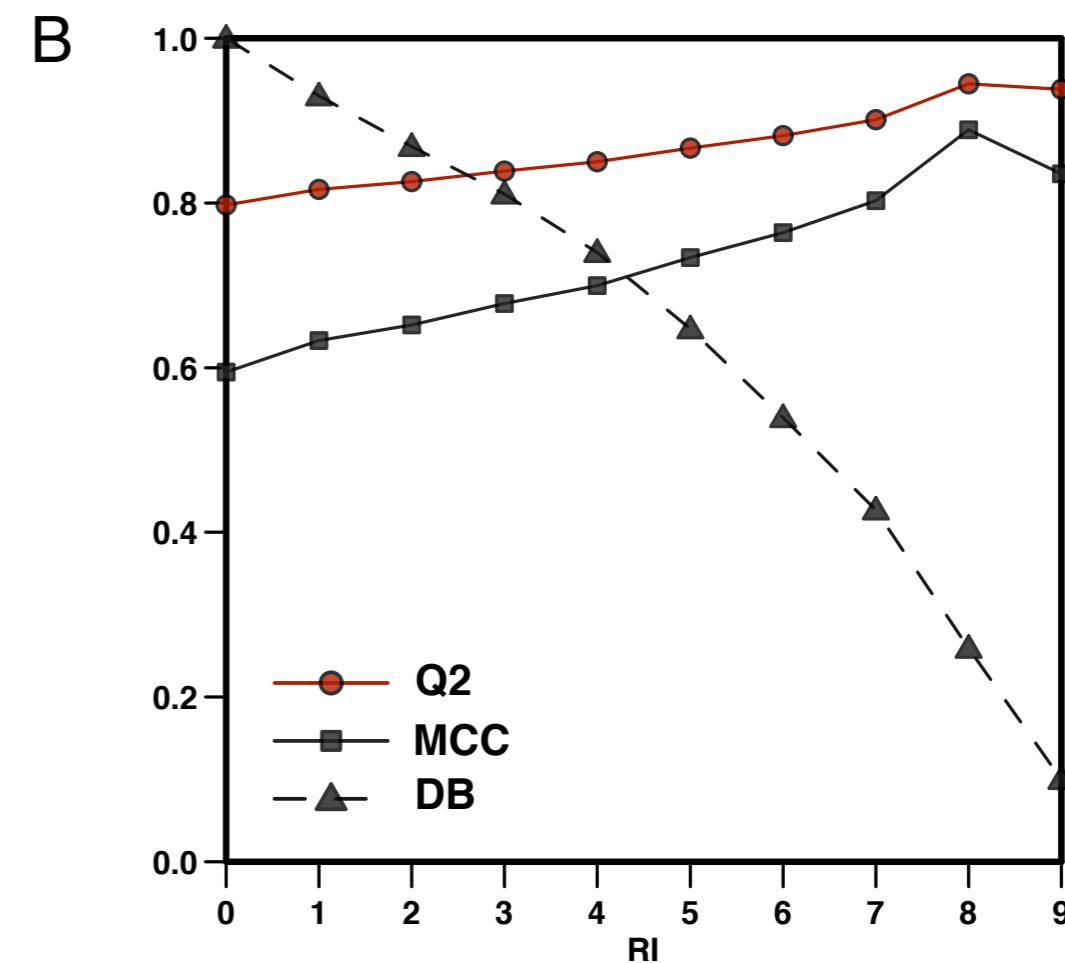
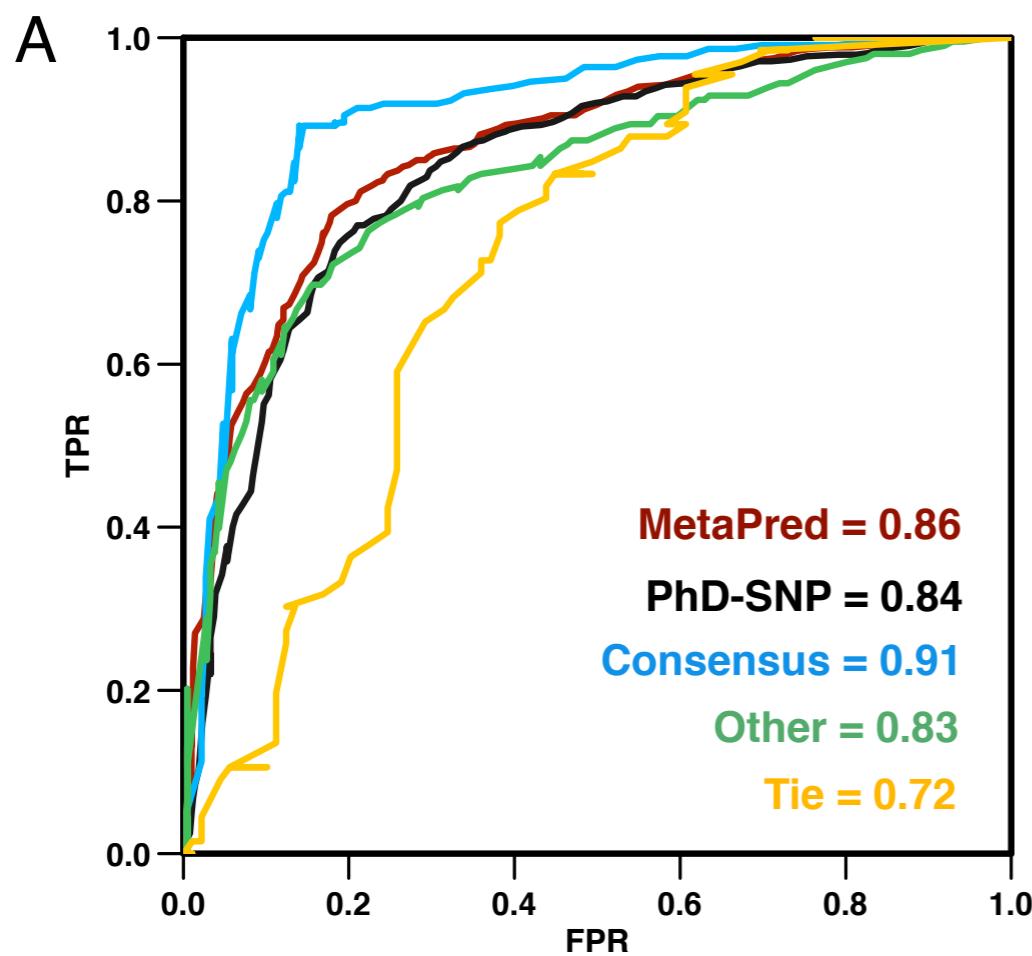


Testing Meta-SNP

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

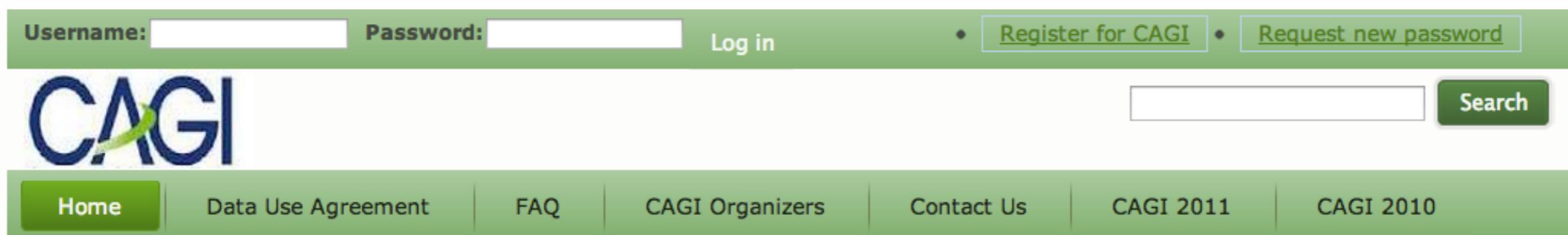
	Q2	P[D]	S[D]	P[N]	S[N]	C
Meta-SNP	0.79	0.79	0.80	0.80	0.79	0.59
PhD-SNP	0.77	0.78	0.77	0.77	0.78	0.55

DB: Neutral 486 and Disease 486



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.



The image shows the top navigation bar of the CAGI website. It features a green header bar with the following elements: a 'Username:' input field, a 'Password:' input field, a 'Log in' button, a 'Register for CAGI' link, a 'Request new password' link, the CAGI logo, a search input field, and a 'Search' button. Below this is a green footer bar with links for 'Home', 'Data Use Agreement', 'FAQ', 'CAGI Organizers', 'Contact Us', 'CAGI 2011', and 'CAGI 2010'.

CAGI 2012

- [Overview](#)
- [Key Dates](#)
- [Conference](#)
- [Challenges](#)
 - [Crohn's Disease](#)
 - [BRCA](#)
 - [Splicing](#)
 - [MRN](#)
 - [FCH](#)
 - [HA](#)
 - [riskSNPs](#)
 - [MDA](#)

Welcome to the CAGI experiment!

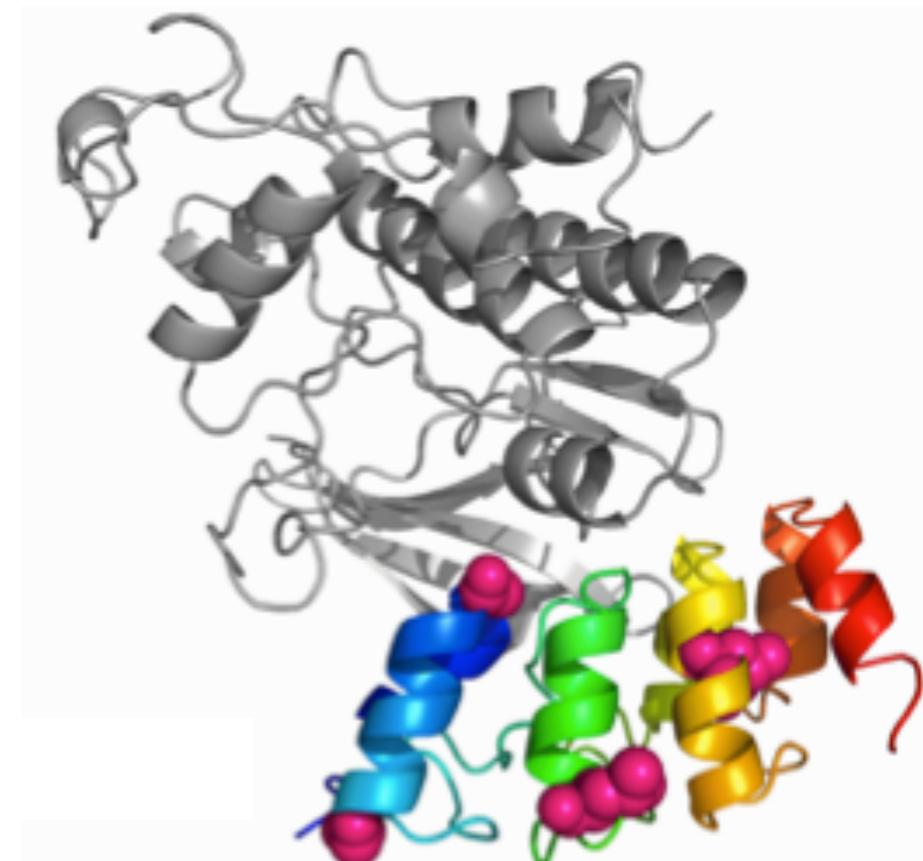
The Critical Assessment of Genome Interpretation (CAGI, \kā-jē\) is a community experiment to objectively assess computational methods for predicting the phenotypic impacts of genomic variation. In this experiment, modeled on the Critical Assessment of Structure Prediction (CASP), participants will be provided genetic variants and will make predictions of resulting molecular, cellular, or organismal phenotype. These predictions will be evaluated against experimental characterizations, and independent assessors will perform the evaluations. Community workshops will be held to disseminate results, assess our collective ability to make accurate and meaningful phenotypic predictions, and better understand progress in the field. From this experiment, we expect to identify bottlenecks in genome interpretation, inform critical areas of future research, and connect researchers from diverse disciplines whose expertise is essential to methods for genome interpretation. We want to emphasize that CAGI is a community experiment to understand and improve the interpretation of genome variation. It is not a contest and all predictors are awarded recognition for their participation in the meeting.

The CAGI P16^{INK} challenge

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

Challenge: Predict how protein variants in p16 protein impact its ability to block cell proliferation.

SNPs&GO among the best methods to blindly **predict the change in cell proliferation** associated to mutations on P16^{INK} (~70% accurate predictions).



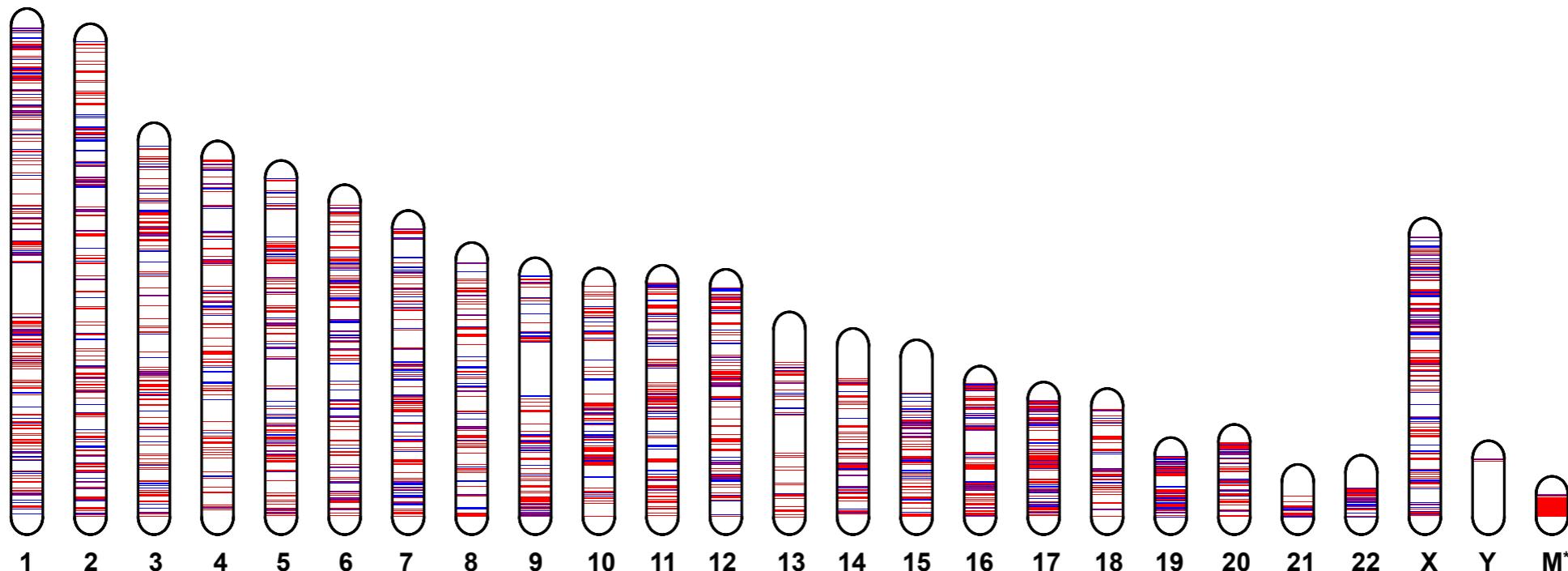
SNPs&GO prediction

Proliferation rates have been predicted using the **raw output** of SNPs&GO without any fitting

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

Whole-genome predictions

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

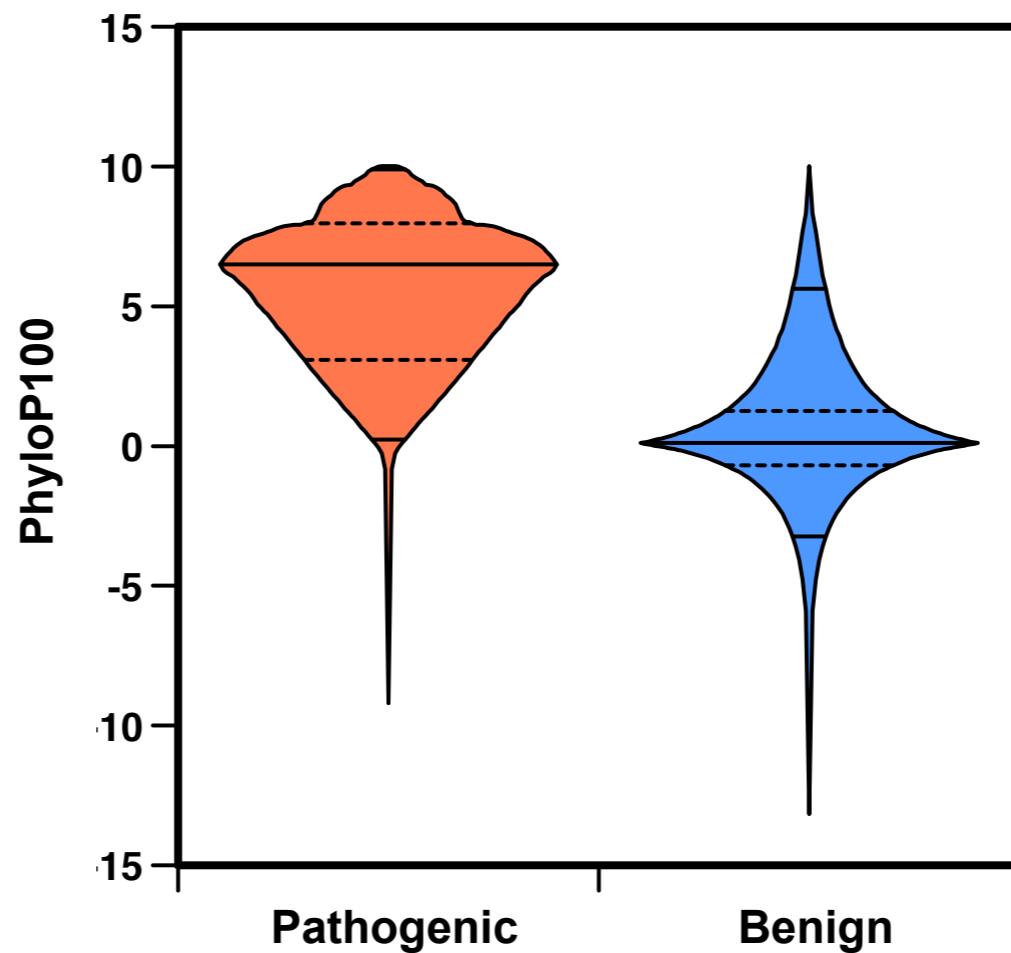


Predict the effect of SNVs in non-coding region is a challenging task because conservation is more difficult to estimate.

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

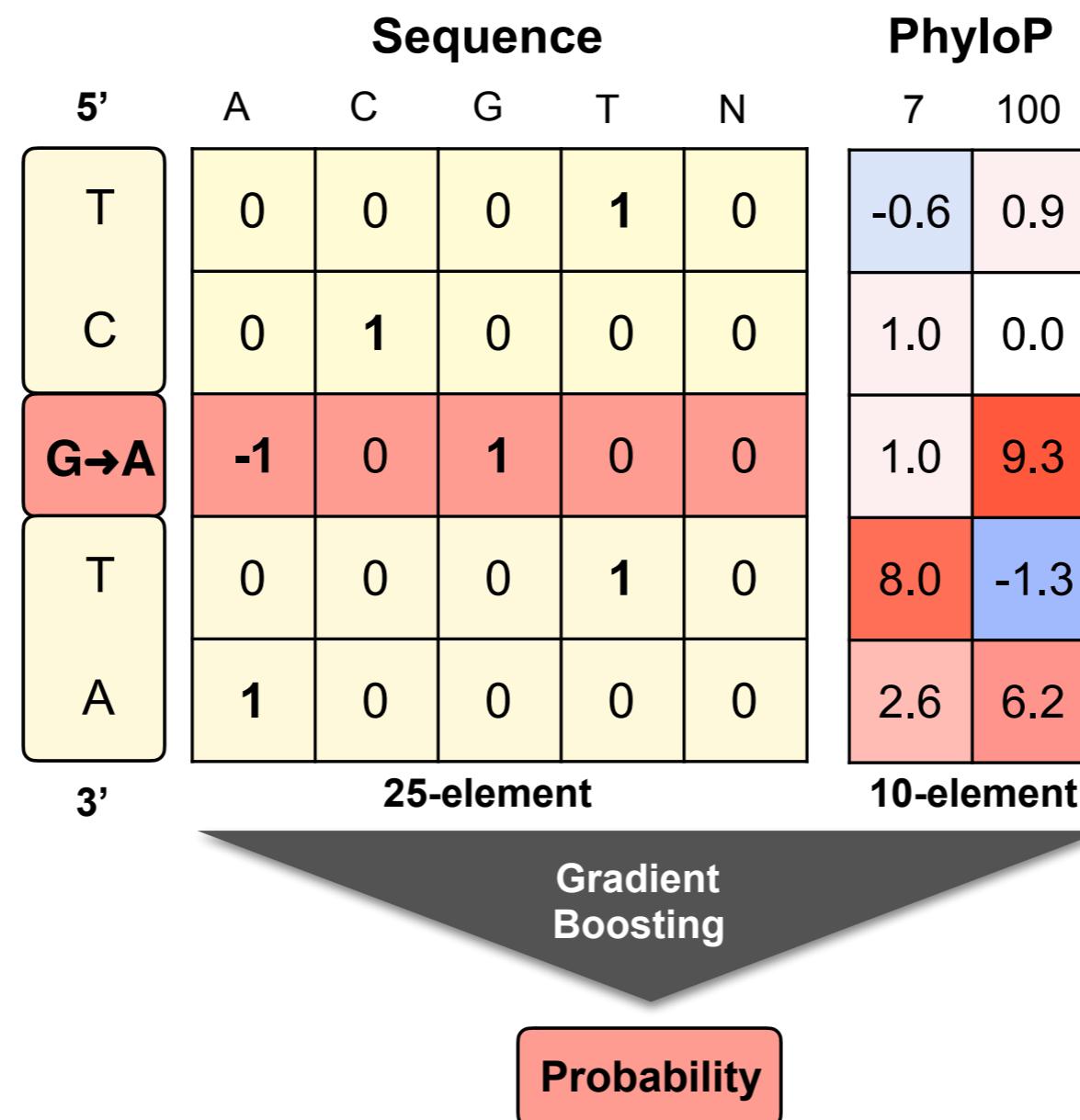
PhyloP100 score

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.**



PhD-SNPG

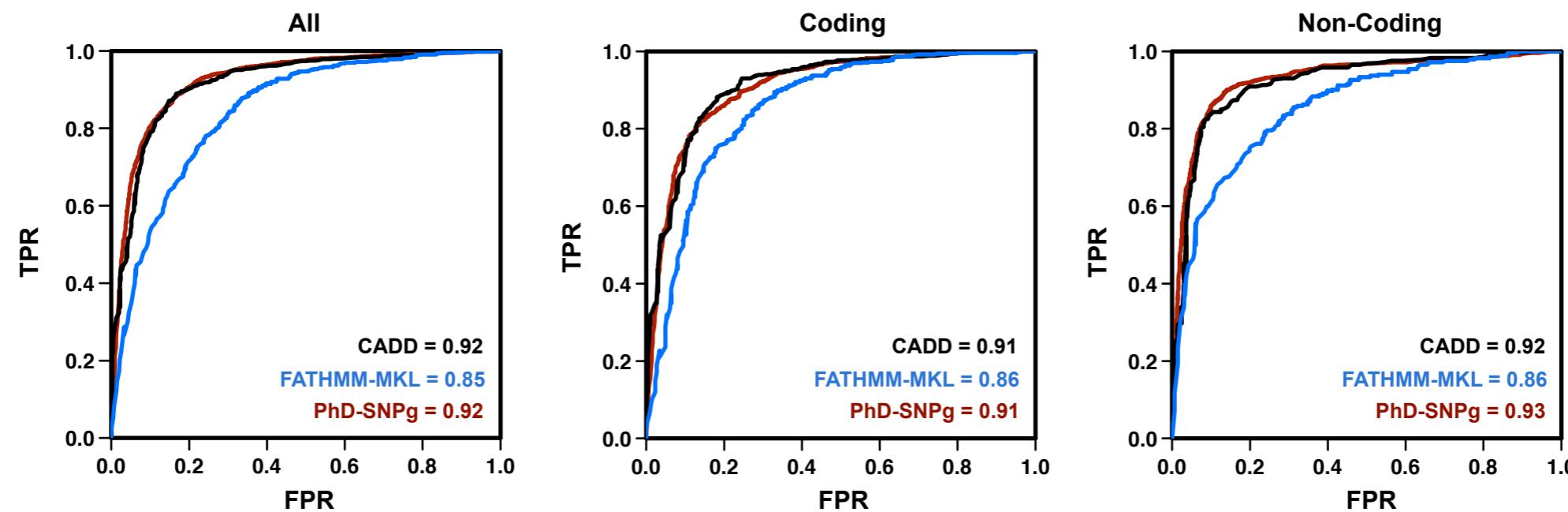
PhD-SNPG is a simple method that takes in input **35 sequence-based features** from a window of 5 nucleotides around the mutated position.



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



Web server and scripts

PhD-SNPg has been implemented ad web server and the scripts are freely available on GitHub

The screenshot shows the PhD-SNPg web application interface. At the top left is the BioFold logo, which consists of a stylized molecular structure of red and blue spheres. To the right of the logo is a navigation bar with three buttons: "Home" (which is highlighted in blue), "Method", and "Install". Below the navigation bar, the title "PhD-SNPg" is displayed in a large, bold, blue font. Underneath the title, the subtitle "Predicting **human** **Deleterious** **SNPs** in **human** **genome**" is shown in a smaller, dark font. Below the subtitle, a descriptive text reads "A binary classifier for predicting pathogenic variants in coding and non-coding regions." A large text input area is present, with the placeholder text "Insert the list of SNVs using the comma separated values format `chr, pos, ref, alt` (maximum 10,000 variants)." Below this input area, there is a note: "or upload an input file in csv or vcf format (zipped files are allowed)." Underneath the input area, there is a "Browse" button. At the bottom of the page, there are two sections: "Input type" and "Human genome". The "Input type" section includes a "DNA:" dropdown menu with "CSV" and "VCF" options, where "CSV" is currently selected. The "Human genome" section includes an "Assembly:" dropdown menu with "hg19" and "hg38" options, where "hg19" is currently selected.

Conclusions

- Evolutionary information are important for the prediction of deleterious variants. The wild-type residues in disease-related mutation sites are more conserved than in neutral sites.
- Different algorithms for genome interpretation predicting the effect of a single amino acid polymorphism on human health have been tested. The methods based on functional information are among the most accurate.
- Relative solvent accessible area and structure environment are better features than sequence-based ones to discriminate between disease-causing and neutral mutants.
- The implementation of meta-prediction based approach allows to select highly accurate predictions.
- Whole-genome SNVs prediction is a more difficult task but predictions based on pre-calculated conservation score can perform as-well-as more sophisticated methods including annotation from ENCODE project.

Acknowledgments

Structural Genomics @CNAG

Marc A. Marti-Renom

Davide Bau

David Dufour

Francois Serra

Department of Bioinformatics (CIPF)

Joaquín Dopazo

Bologna Biocomputing Group

Rita Casadio

Remo Calabrese

Piero Fariselli

Pier Luigi Martelli

University of Camerino

Mario Compiani

Computational Biology and Bioinformatics Research Group (UIB)

Jairo Rocha

Mathematical Modeling of Biological Systems (University of Düsseldorf)

Markus Kollmann

Division of Informatics at UAB

Rui Tian

Shivani Viradia

Malay Basu

Diego E. Penha

Helix Group (Stanford University)

Russ B. Altman

Jennifer Lahti

Tianyun Liu

Grace Tang

Other Collaborations

Yana Bromberg, Rutger University, NJ

Francisco Melo, Universidad Católica, Chile

Sean Mooney, Buck Institute, Novato

Cedric Notredame, CRG Barcelona

Gustavo Parisi, Universidad de Quilmes

Frederic Rousseau, KU Leuven

Joost Schymkowitz, KU Leuven

FUNDING

Startup funding Dept. of Pathology UAB

NIH:3R00HL111322-04S1 Co-Investigator

EMBO Short Term Fellowship

Marie Curie International Outgoing Grant

Marie Curie Reintegration Grant

Marco Polo Research Project

BIOSAPIENS Network of Excellence

SPINNER Consortium

Biomolecules, Folding and Disease

<http://biofold.org/>

