Predicting the effect of protein variants

Laboratory of Bioinformatics I
Module 2

Emidio Capriotti
http://biofold.org/

Biomolecules Folding and Disease

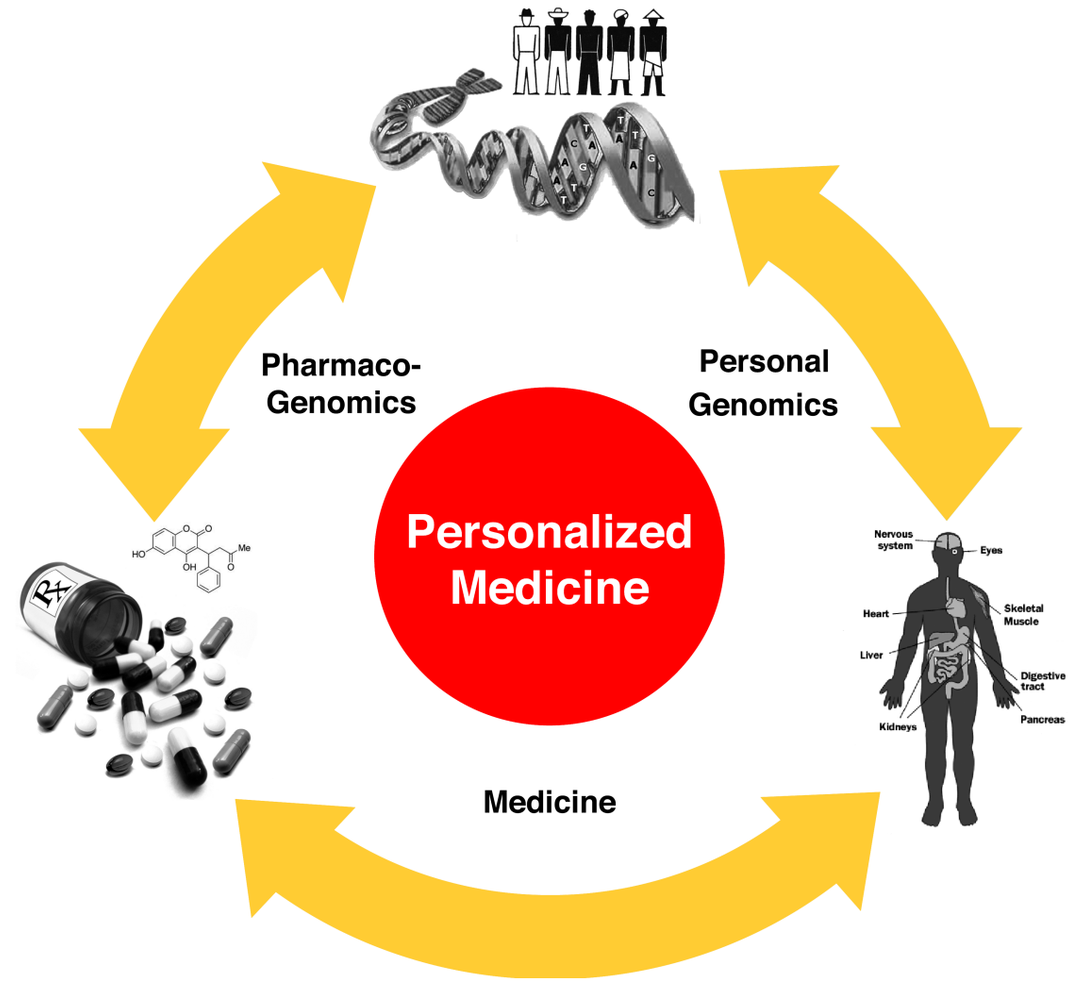
Department of Pharmacy and Biotechnology (FaBiT)
University of Bologna

Personalized medicine

Currently direct to consumers company are performing genotype test on markers associated to genetic traits, and 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



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

Single Nucleotide Variants

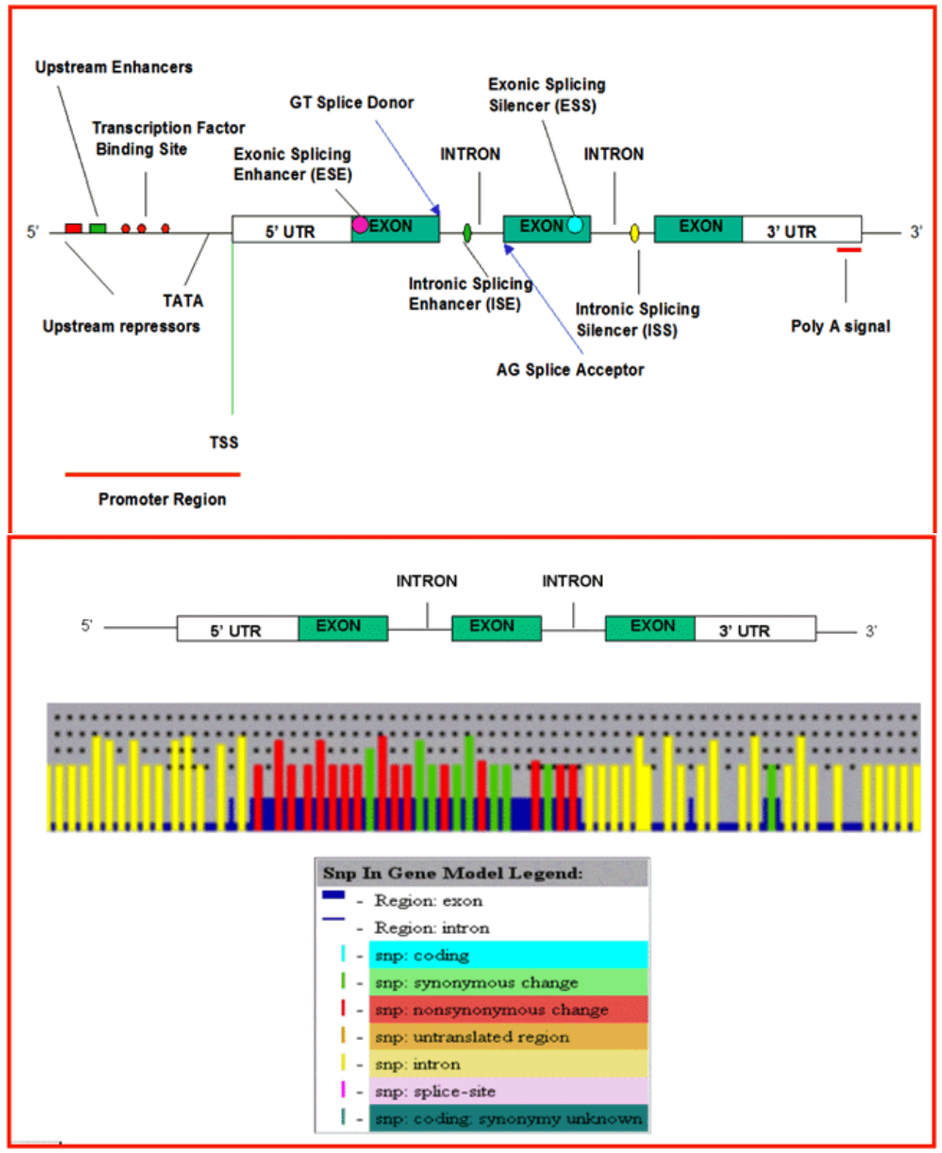
Single Nucleotide Variants (SNVs)
is a DNA sequence variation occurring when a single nucleotide A, T, C, or G in the genome differs between members of the species.
It is used to refer to Polymorphisms when the population frequency is ≥ 1%

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

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.

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

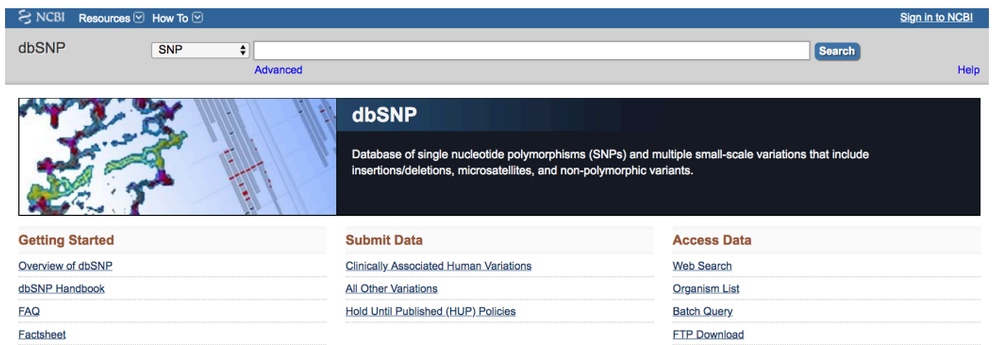
a Summary of project data including combined exon populations

	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

Single Amino acid Variants

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

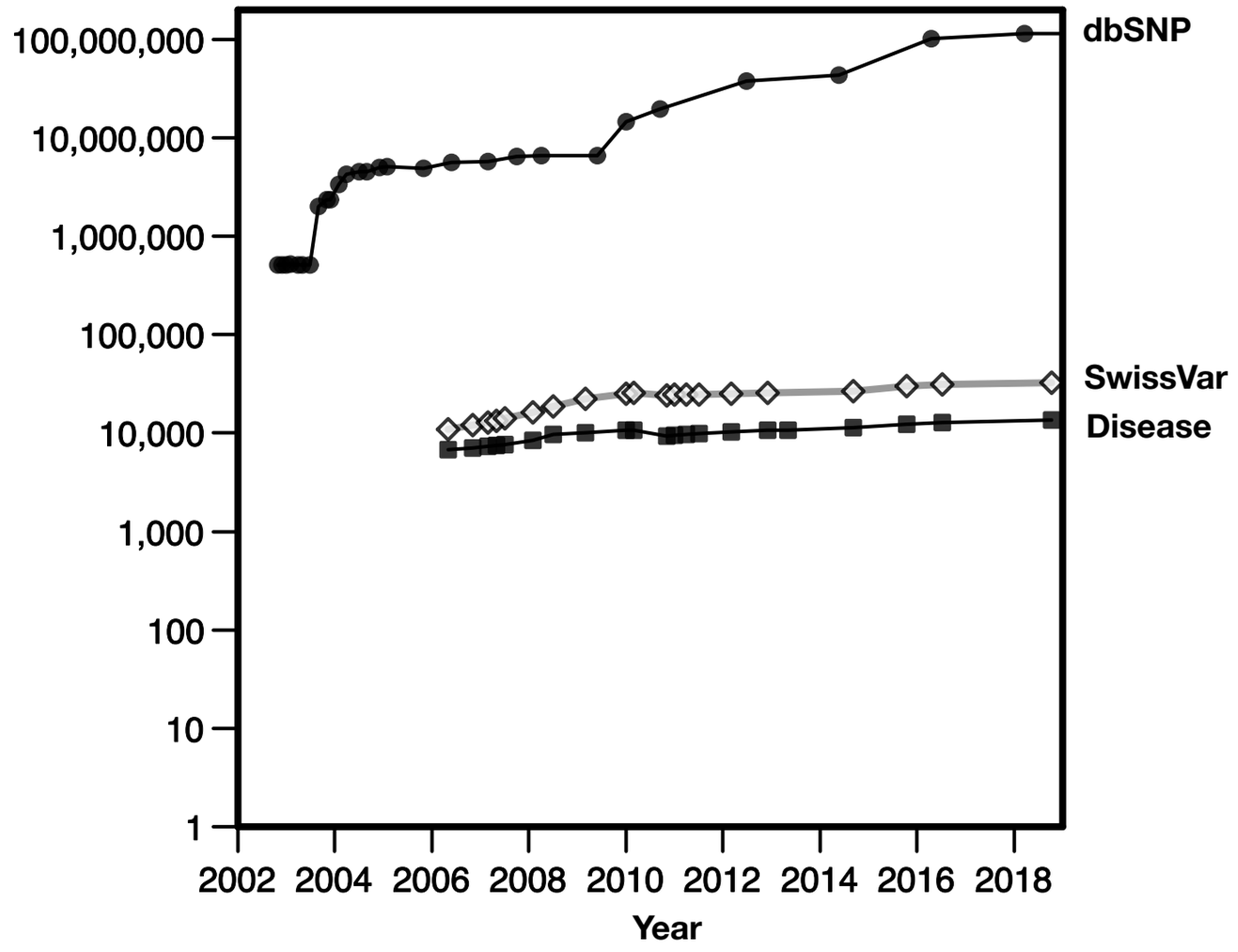
http://www.expasy.ch/swissvar/

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).



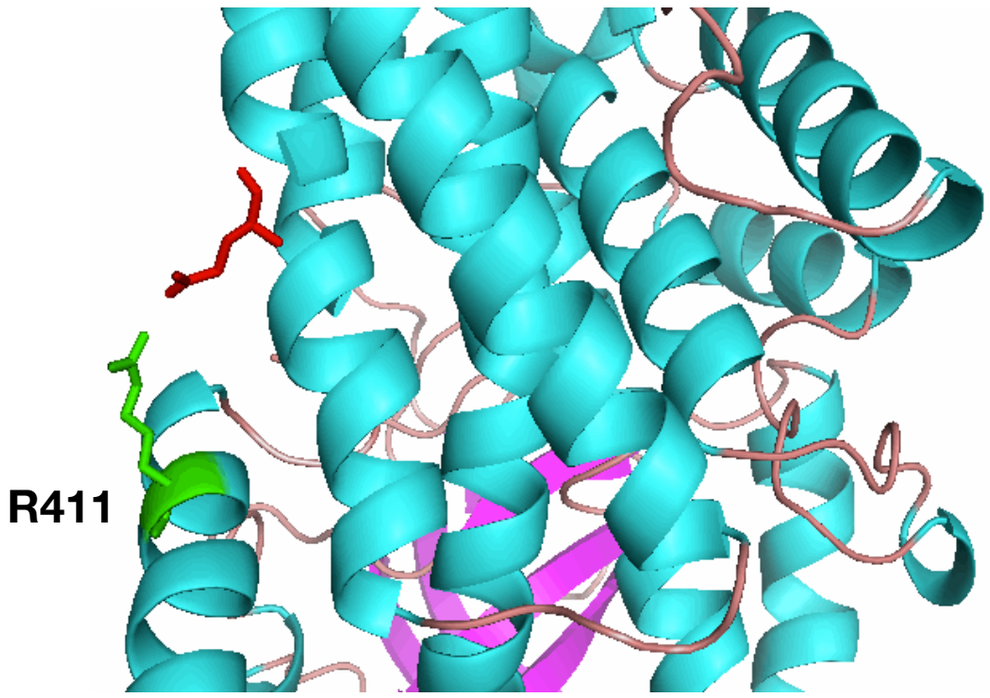
nonsynonymous SNVs

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

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.

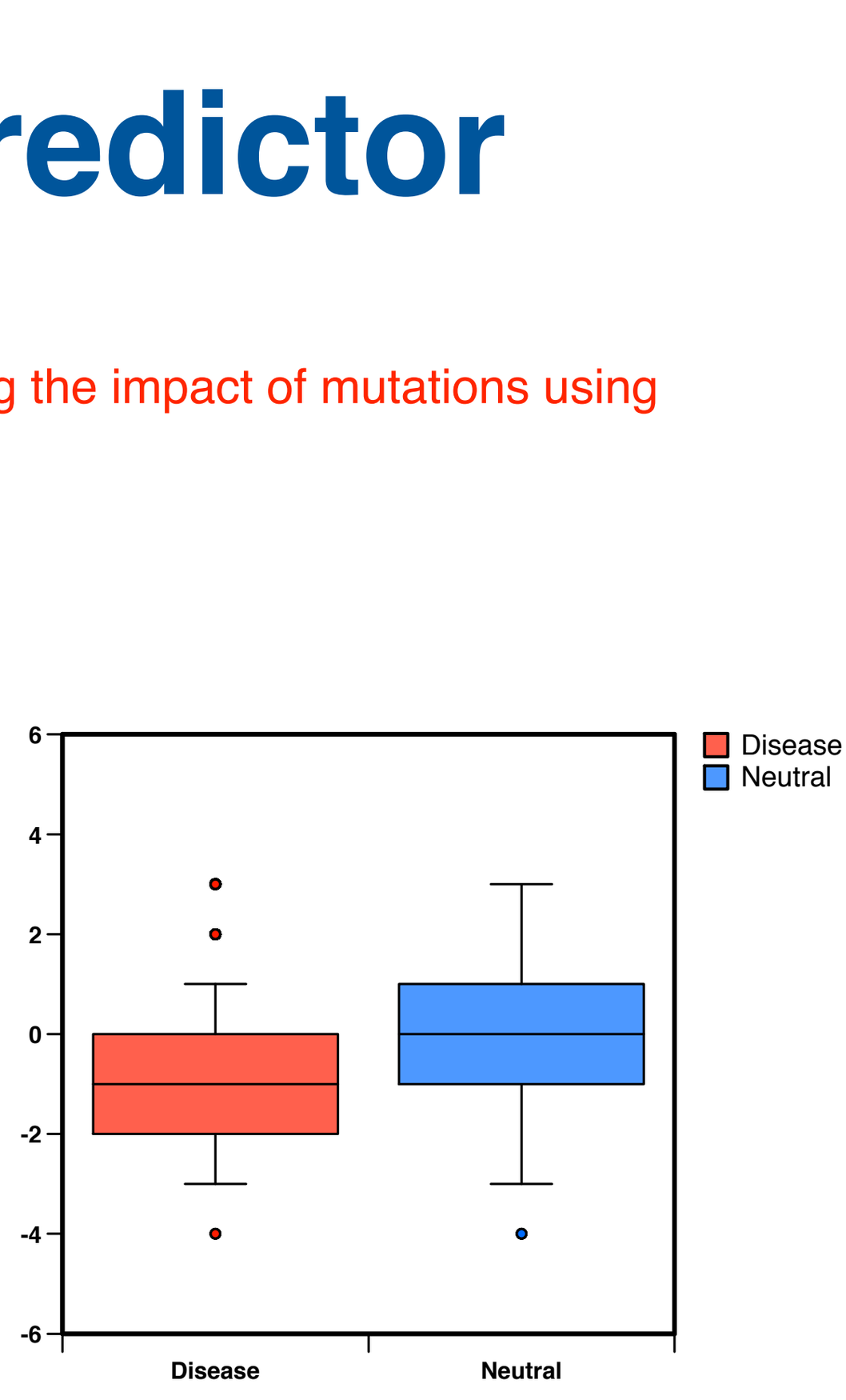
Simple Predictor

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

BLOSUM62

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	
A	5	-2	-2	-2	0	0	0	0	-2	-2	-3	-2	-1	-2	0	0	0	-2	-3	0	A
R		5	-2	-3	-3	0	-1	-2	0	-3	-4	1	-3	-3	-2	-2	0	0	-3	-4	R
N			5	0	0	0	-2	0	0	-4	-5	-2	-3	-3	-2	0	0	-2	-2	-5	N
D				5	-4	0	1	-1	0	-5	-6	-3	-4	-4	0	-2	-2	-2	-2	-5	D
C					8	-2	-3	-1	-1	0	-2	-3	0	-1	-1	1	0	0	-2	0	C
Q						5	2	0	0	-2	-4	0	-2	-3	0	0	0	0	-2	-3	Q
E							5	0	0	-3	-4	0	-3	-3	0	0	0	-2	-3	-3	E
G								6	0	-4	-5	-2	-3	-2	-2	0	0	0	-2	-3	G
H									6	-3	-4	0	-2	0	0	0	0	0	2	-2	H
I										4	0	-3	2	0	-2	-3	0	0	-3	2	I
L											4	-4	0	0	-3	-4	-3	0	-4	0	L
K												4	-2	-4	-1	-2	0	0	-3	-4	K
M													6	0	-3	-3	-2	0	-3	2	M
F														6	-3	-2	-2	2	2	0	F
P															7	0	0	-2	-3	0	P
S																4	2	-2	-2	-3	S
T																	5	-1	-3	0	T
W																		9	2	-1	W
Y																			7	-3	Y
V																				4	V

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

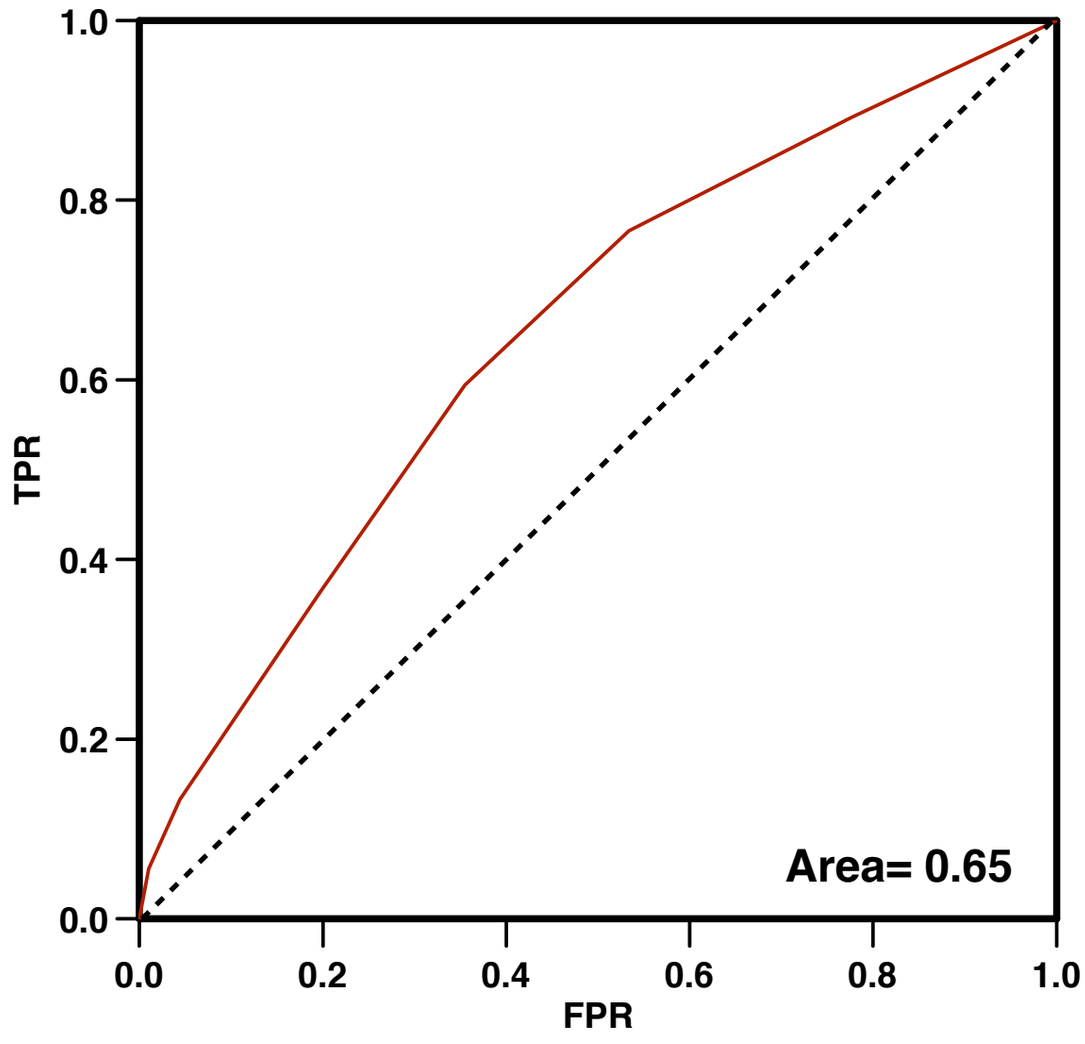


BLOSUM62 Predictions

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

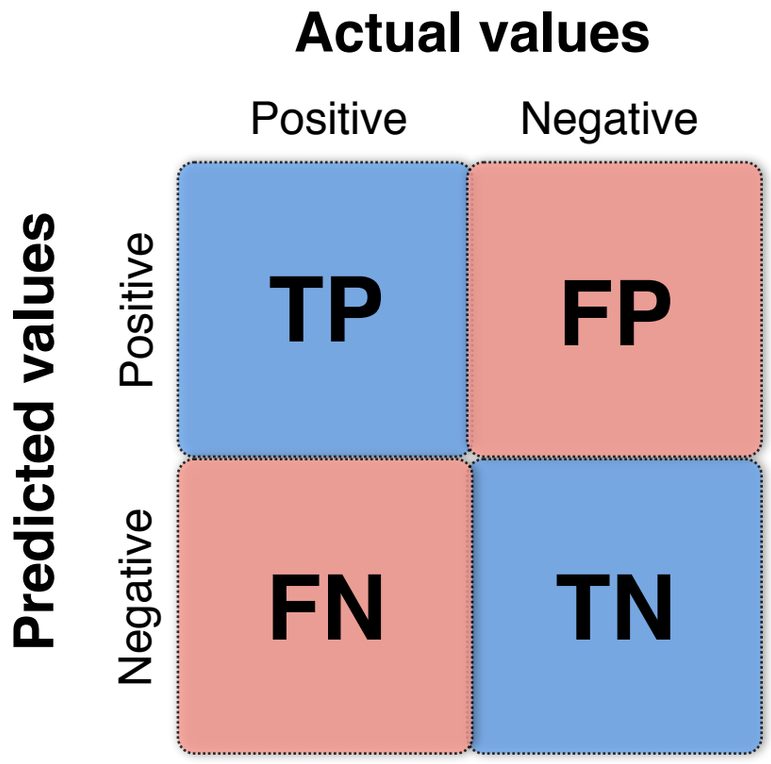
$$S = \frac{TP}{TP + FN}$$

Precision

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

Correlation

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



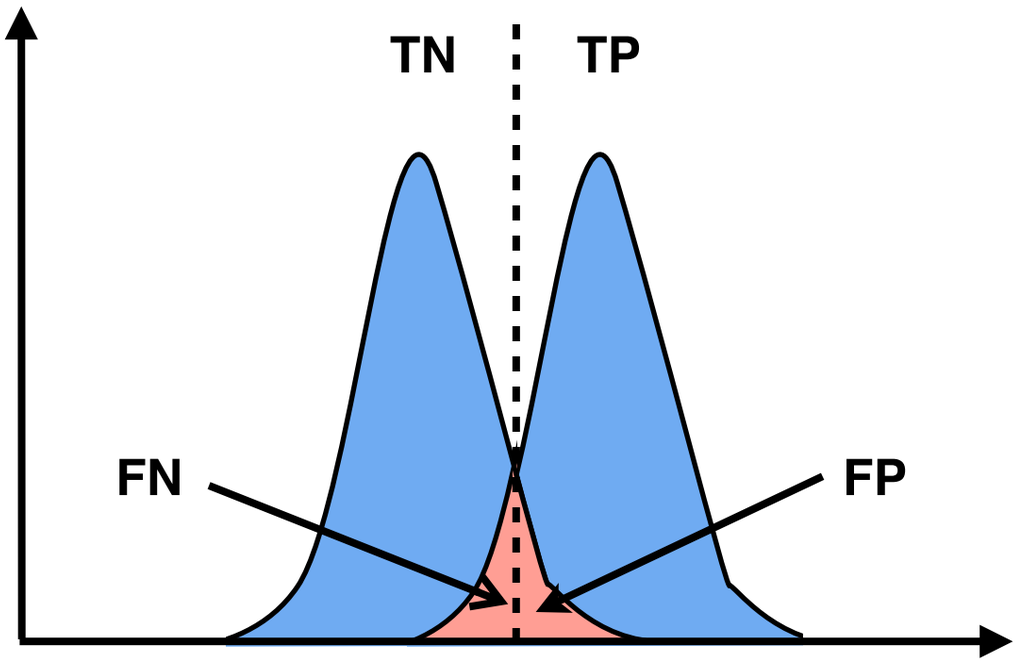
Receiving Operator Curve

True Positive Rate

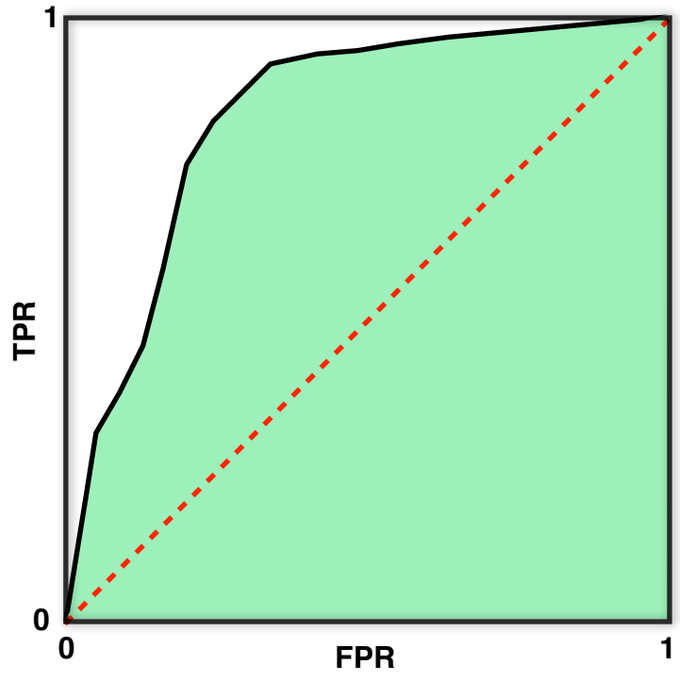
$$TPR = \frac{TP}{TP + FN}$$

False Positive Rate

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



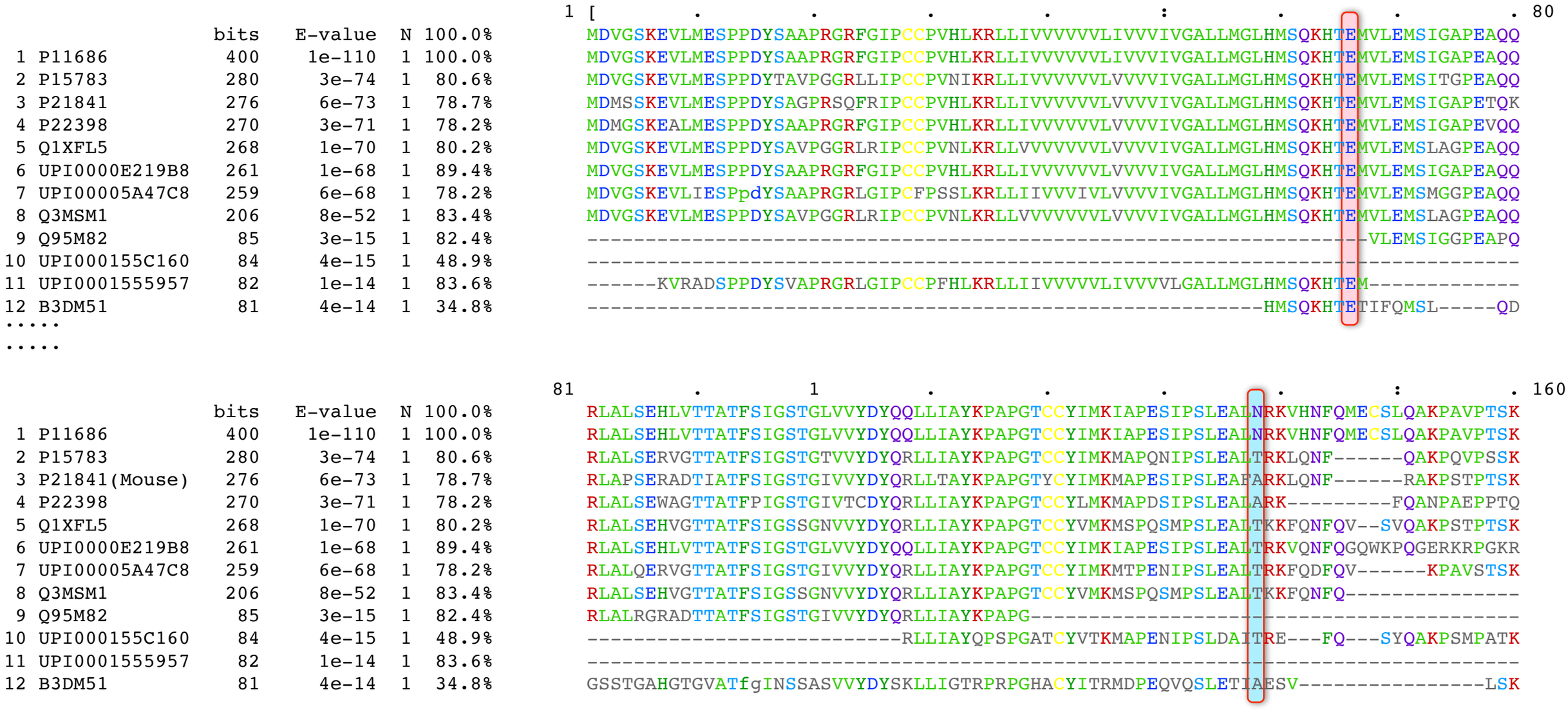
The Area Under the ROC Curve (AUC) is an accuracy measure that is 0.5 for completely random predictors and close to 1.0 for highly accurate predictors.



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

Conserved or not?

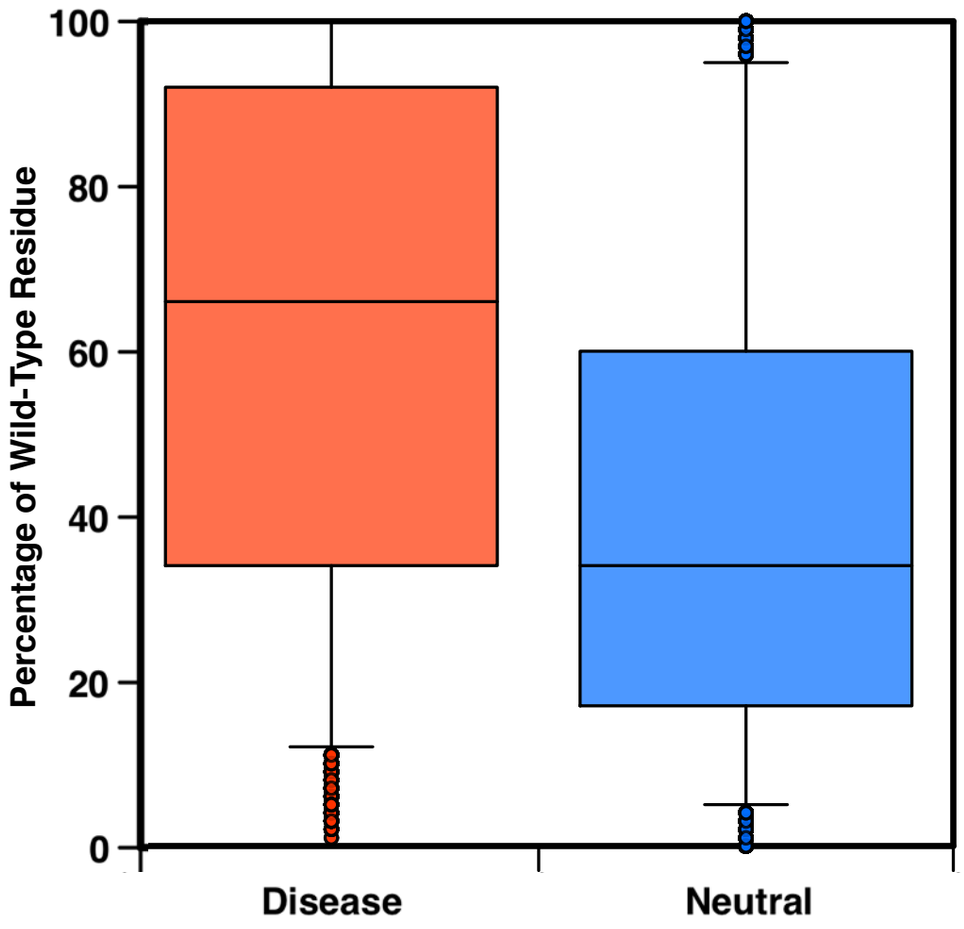
In positions 66 the Glutamic acid is highly conserved Asparagine in position 138 is mutated Threonine or Alanine



Sequence 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.

Machine learning

- Computational approach to build models based on the analysis of 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.

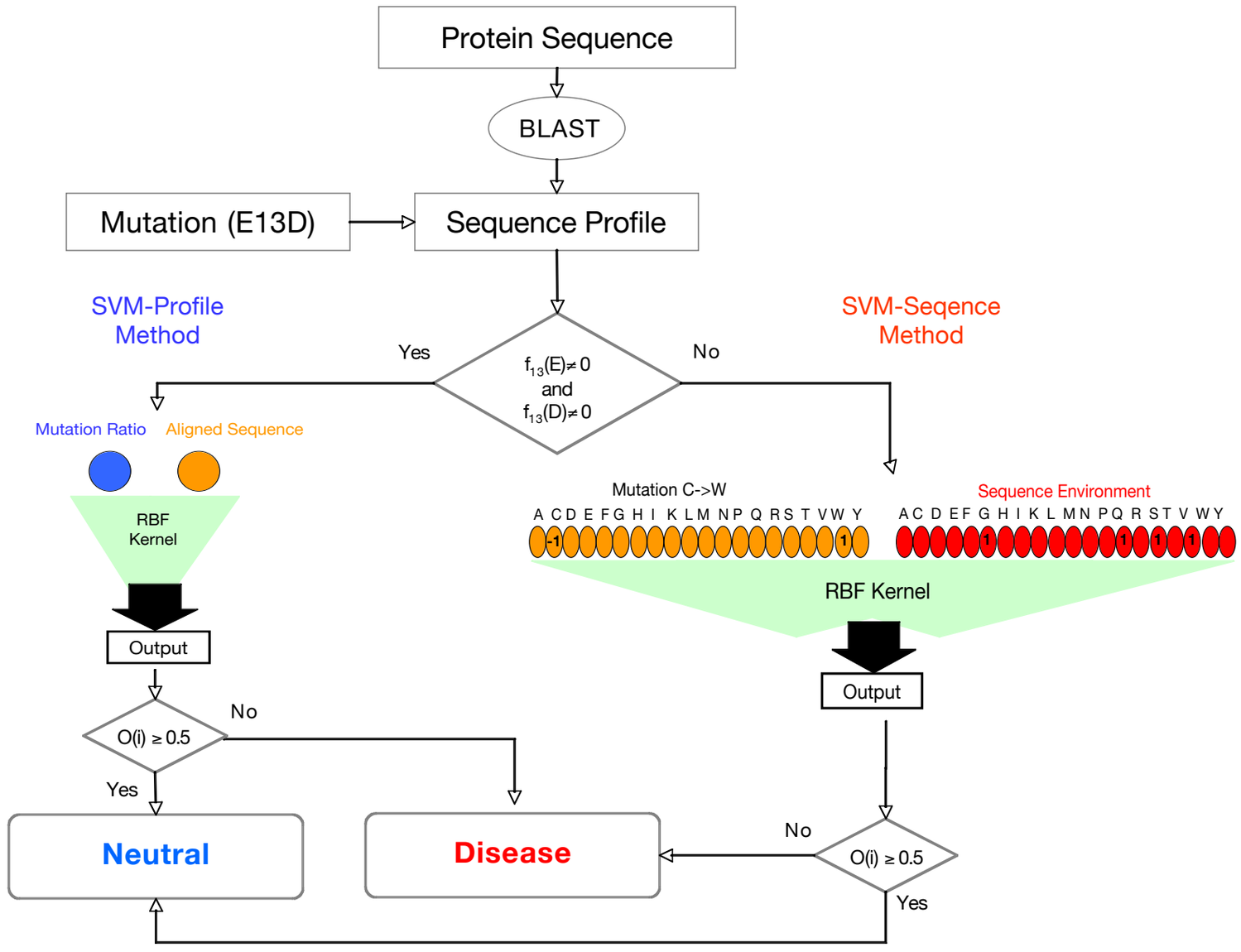
- Artificial Neural Network (ANN): multi-layer network of nodes, including input features, outputs, and one or more hidden layers. Weights of input and output edges connecting nodes are adjusted to maximize prediction accuracy.

- 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.



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

Selective pressure

In genetics, the Ka/Ks ratio is an indicator of selective pressure acting on a protein-coding gene.

It is calculated as the ratio of the number of nonsynonymous substitutions per non-synonymous site (Ka), to the number of synonymous substitutions per synonymous site (Ks), in a given period of time.

Homologous genes with:

- Ka/Ks ratio >> 1 (positive selection): mutations must be advantageous.
- Ka/Ks ratio ~ 1 (neutral selection): advantageous ~ disadvantageous
- Ka/Ks ratio << 0 (negative selection): mutations are disadvantageous

The ratio, also known as ω or dN/dS, can be calculated at gene and site levels.

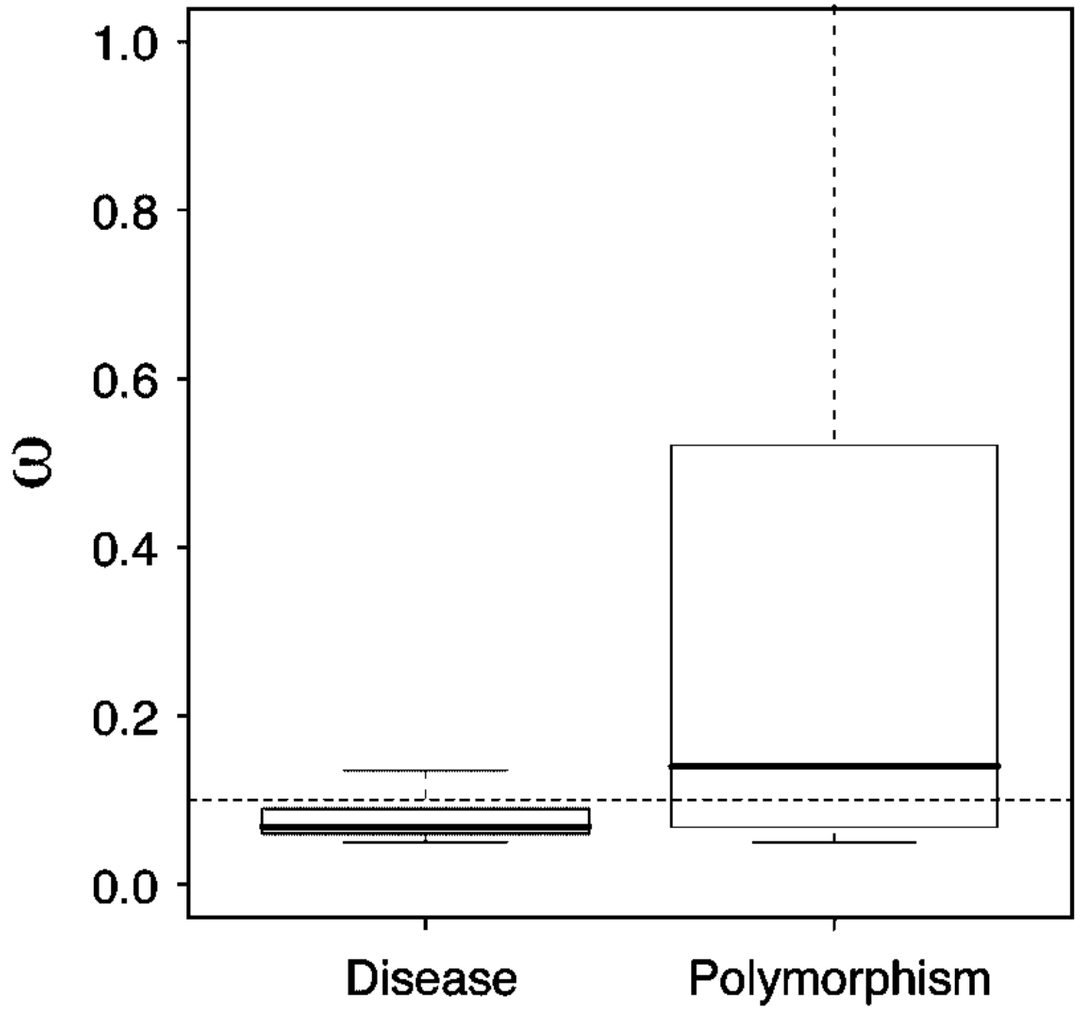
from Wikipedia

The omega values

In a previous work performed on 40 human disease genes, has been demonstrated that residues evolving under strong selective pressures (ω<0.1) are significantly associated with human disease (Arbiza et al. JMB, 2006).

We carried out a similar analysis on the dataset extracted from SwissProt and we found a statistically significant association between high selective pressures and disease in contrast to low selective pressures and neutral polymorphic variants in human.

$$\omega = \frac{dN}{dS}$$

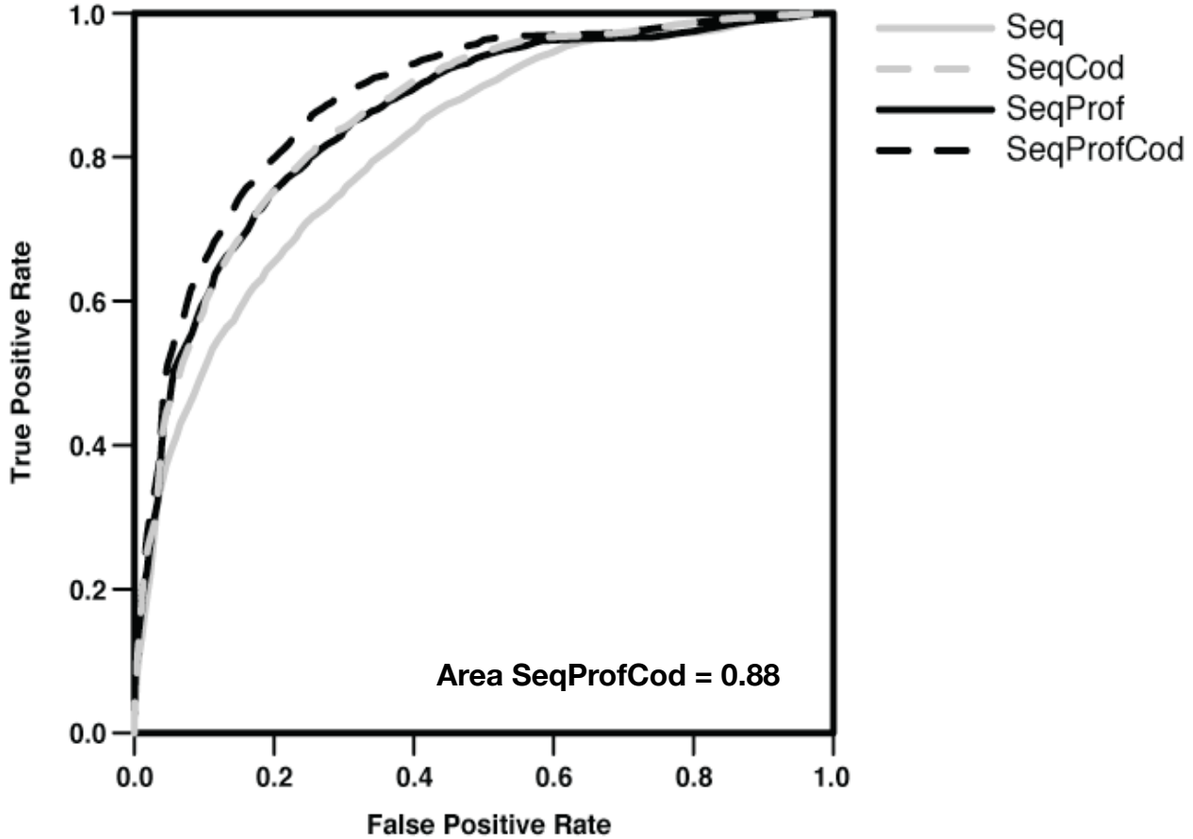


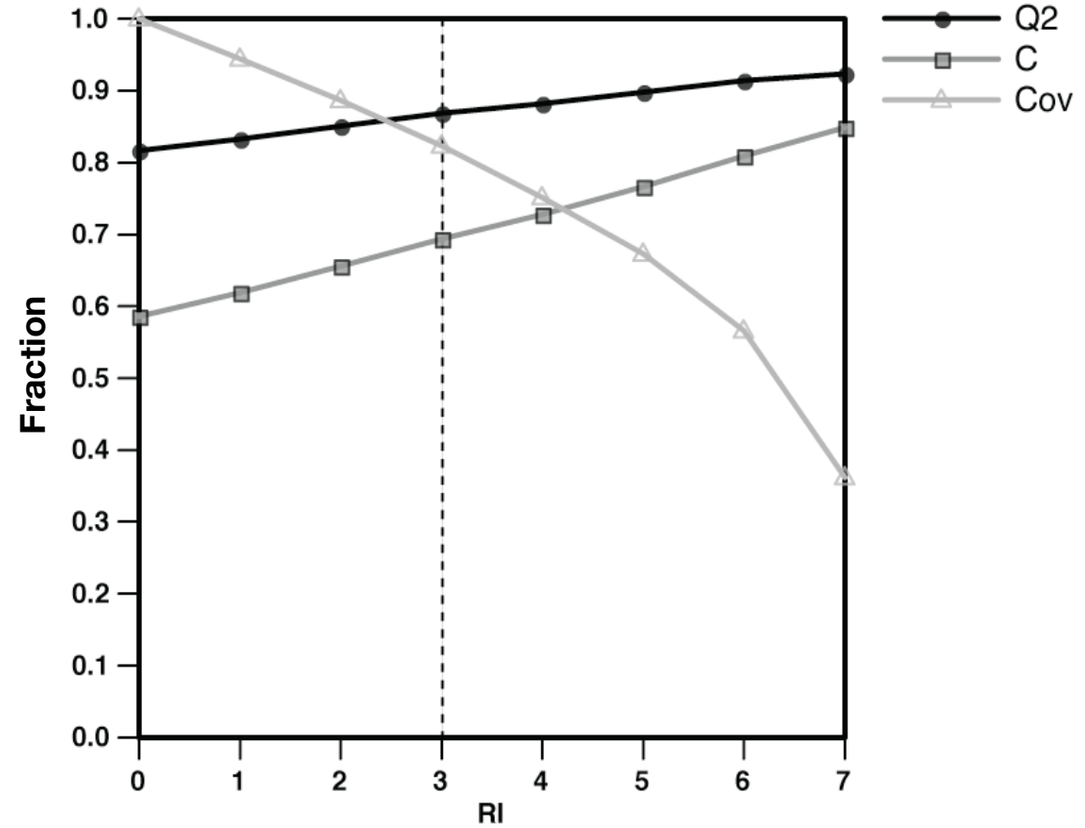
Capriotti et al. (2008) Human Mutation, 29: 198-204.

Omega-based method

SeqProfCod has higher accuracy than the previous two methods increasing the accuracy up to 82% and the correlation coefficient to 0.59.

	Q2	P[D]	Q[D]	P[N]	Q[N]	C
SeqProfCod	0.82	0.88	0.84	0.68	0.76	0.59





Q2: Overall Accuracy **C:** Correlation Coefficient **DB:** Fraction of database that are predicted with a reliability ≥ the given threshold

Gene Ontology

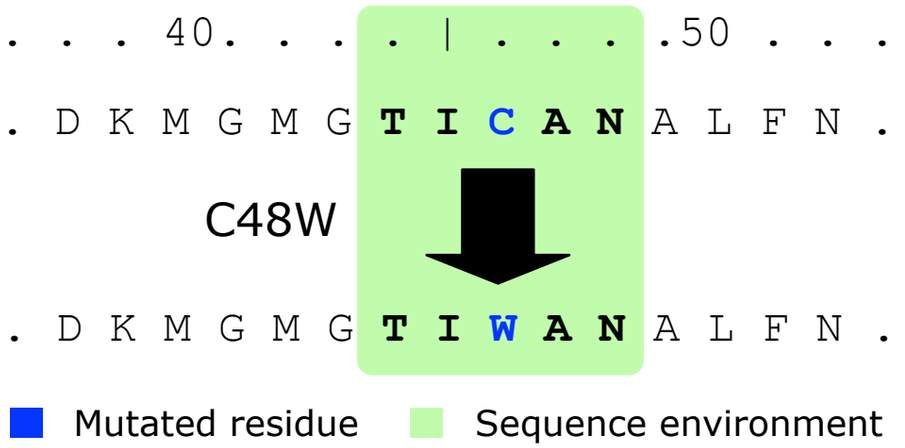
The Gene Ontology project is a major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases. The project provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation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

http://www.geneontology.org/

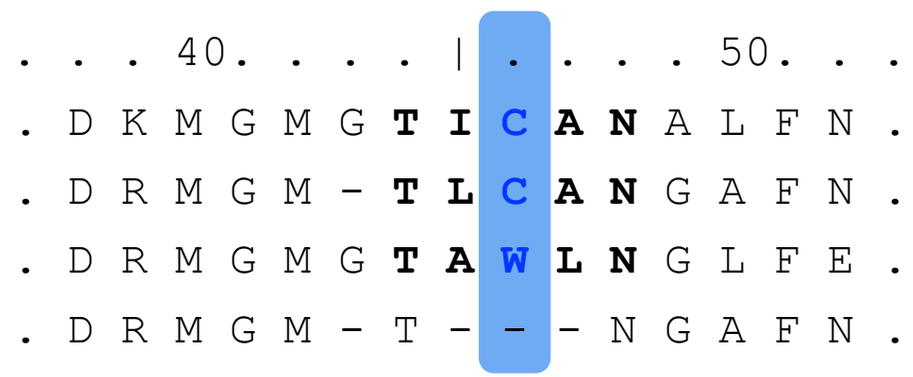
The ontology is represented by a direct acyclic graph covers three domains;

- 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.

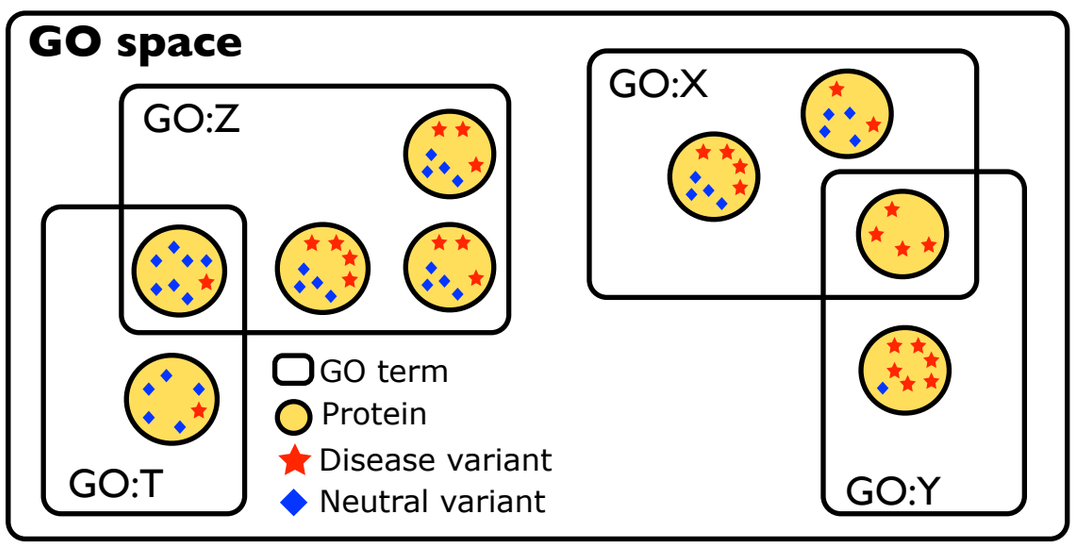
Prediction features



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



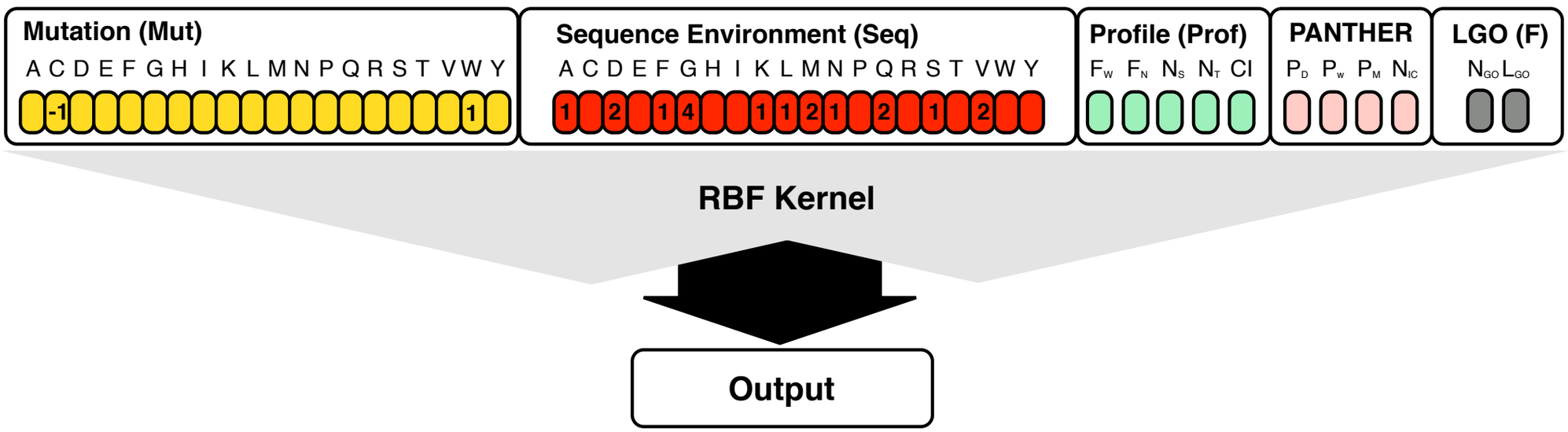
Protein sequence profile information derived from a multiple sequence alignment. It is encoded in a 5 elements vector corresponding to different features general and local features



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.

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

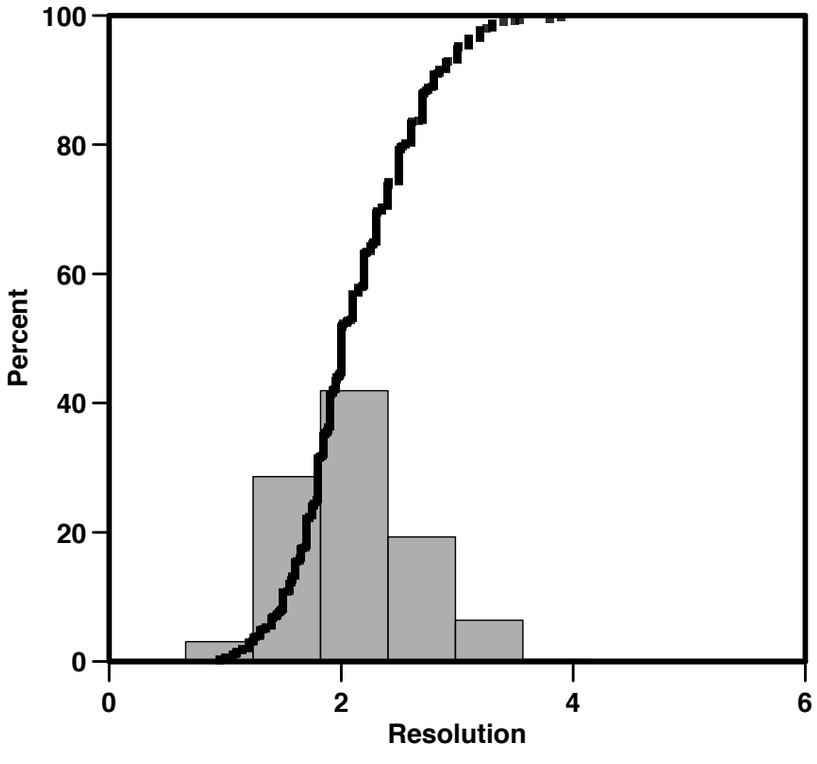
If one mutation maps on more than one PDB the one with lower resolution is selected

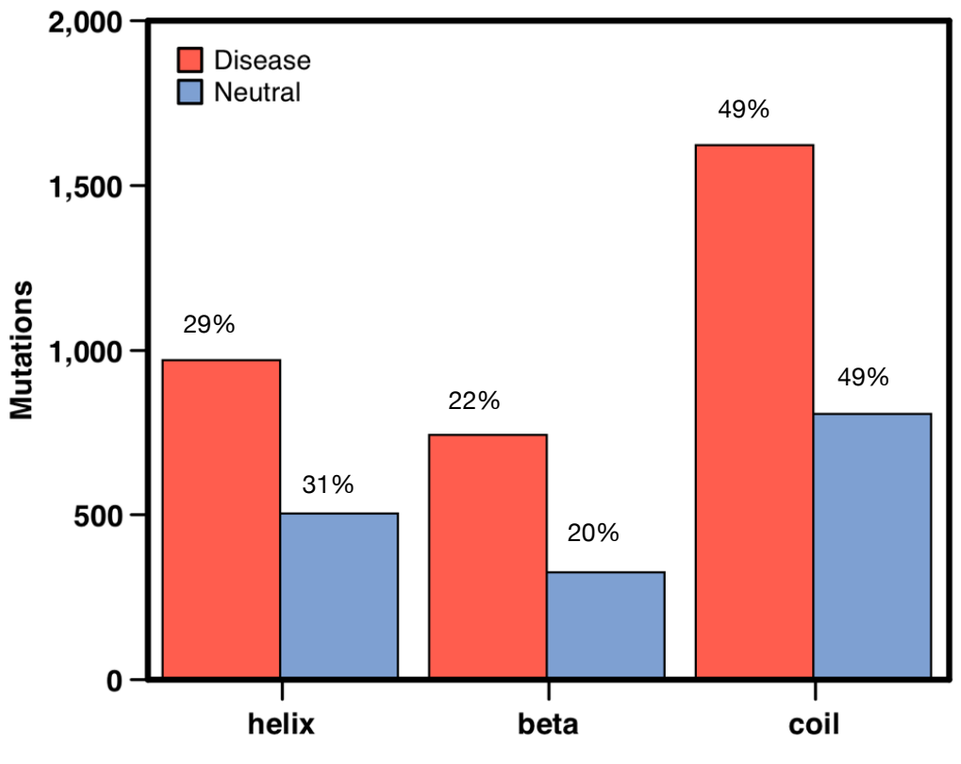
3D Structure Dataset

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**

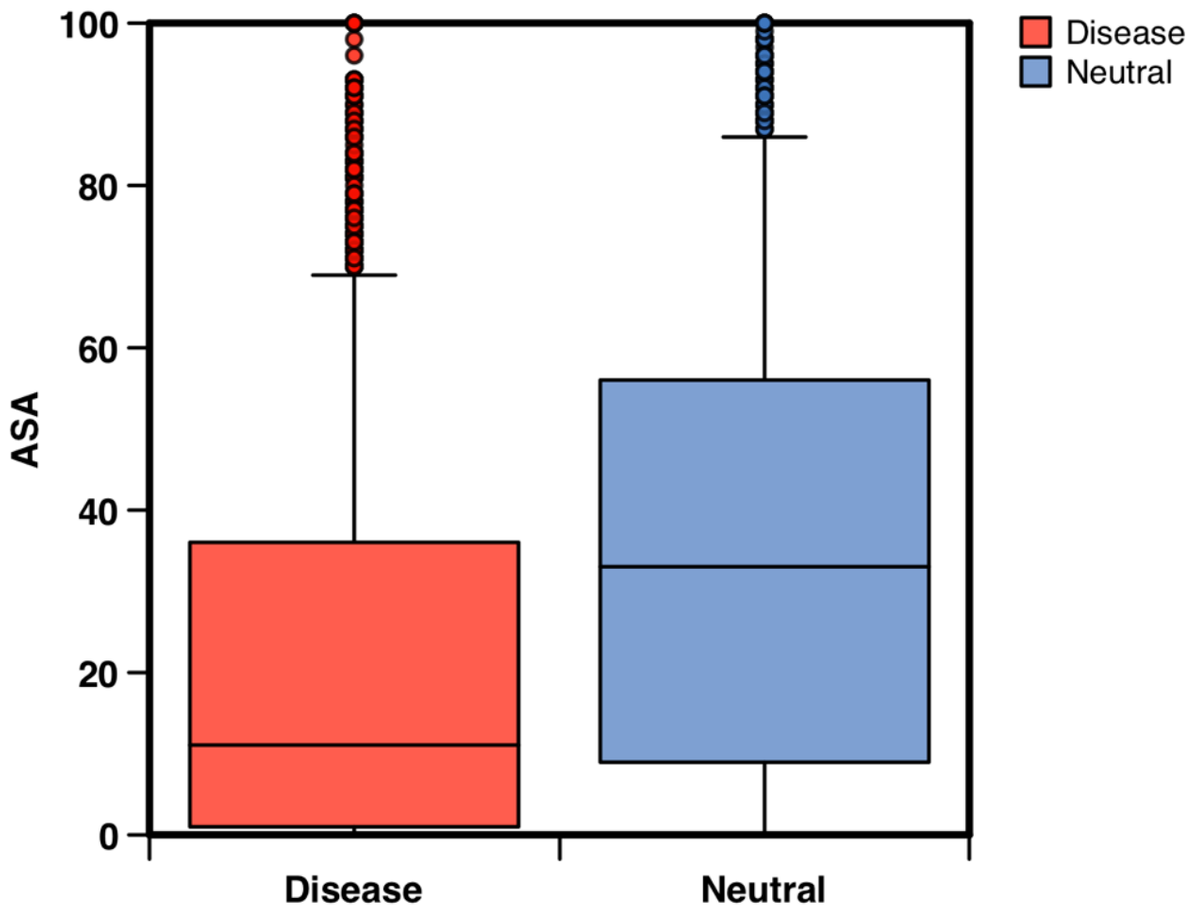
from 784 chains from 770 structures (584 X-ray, 92 NMR and 94 models).





Structure environment

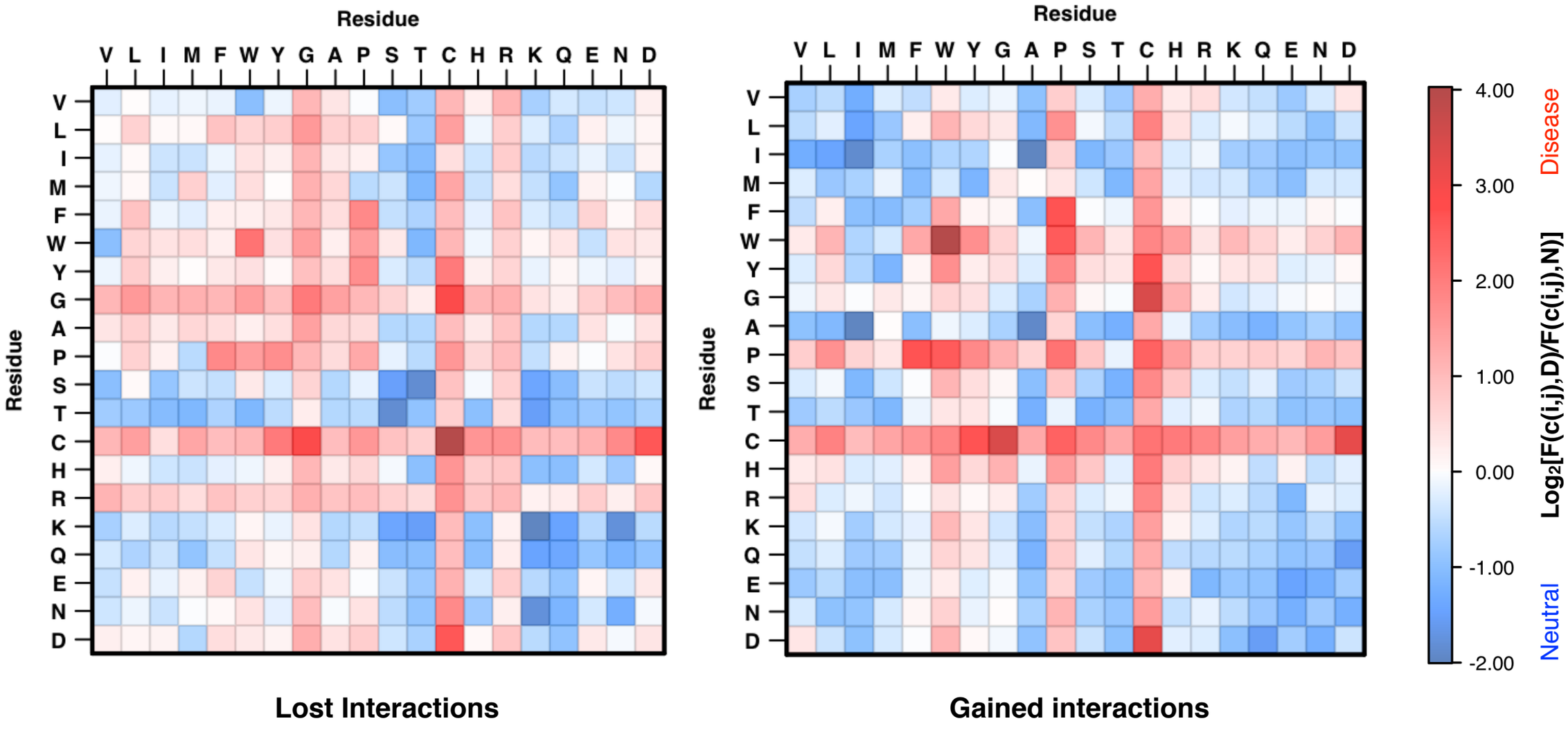
There is a significant difference (p-value KS < 0.001) between the distributions of the relative Accessible Solvent Area for disease-related and neutral variants.
Their mean values are respectively 20.6 and 35.7.



Analysis of the 3D interactions

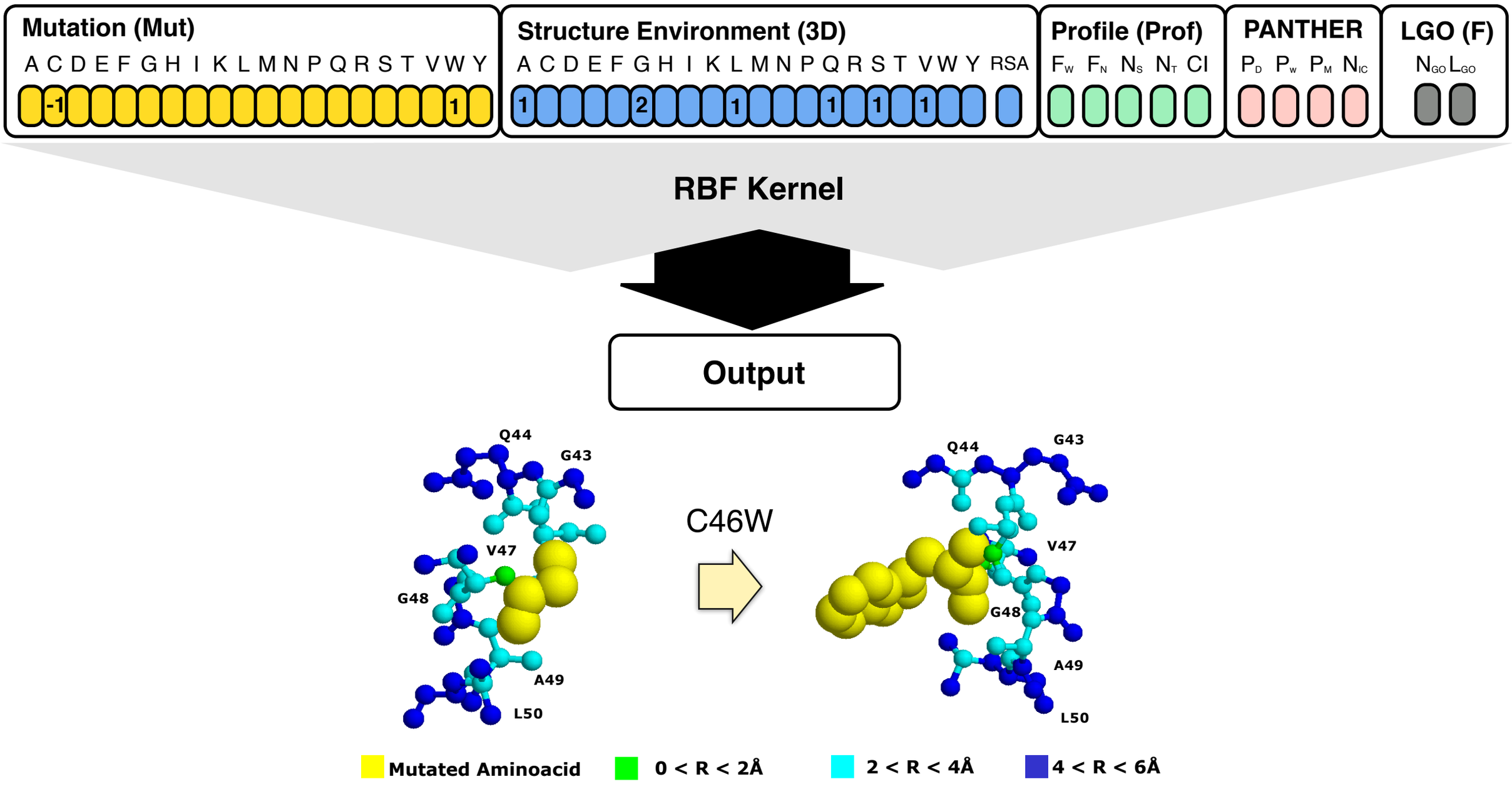
Using the whole set of SAVs with known structure, we calculate the log odd score of the ratio between the frequencies of the interaction between residue i and j for disease-related and neural variants.

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



The structure-based method

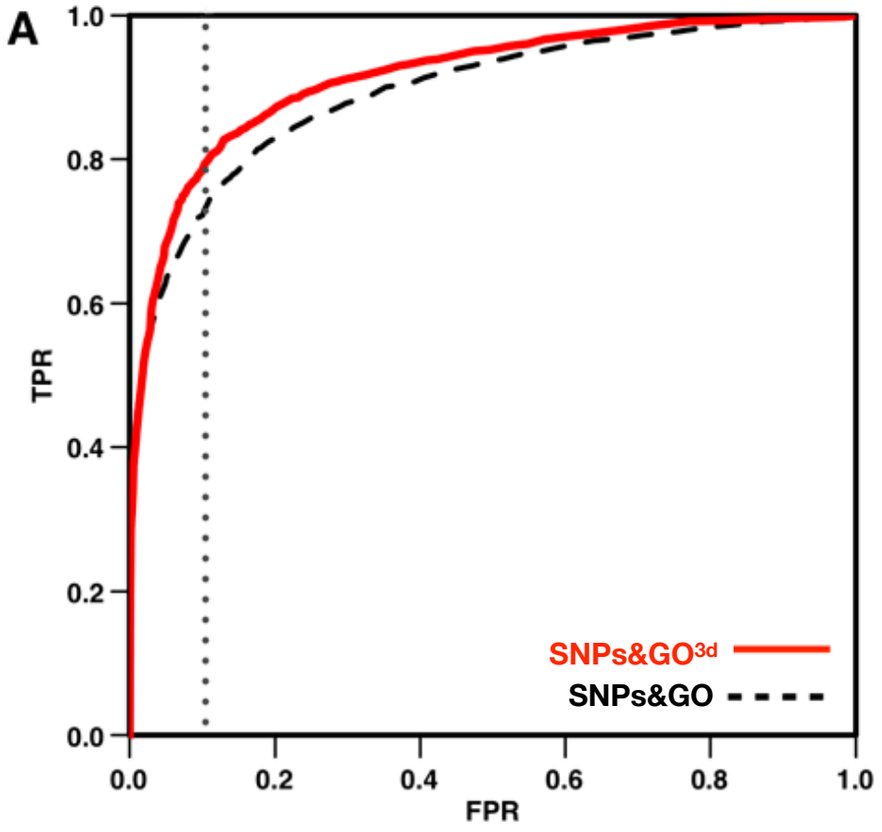
The method takes in to account 5 different types of information encoded in a 52 elements vector. The input features are: mutation data; structure environment, sequence profile and functional score based on GO terms.

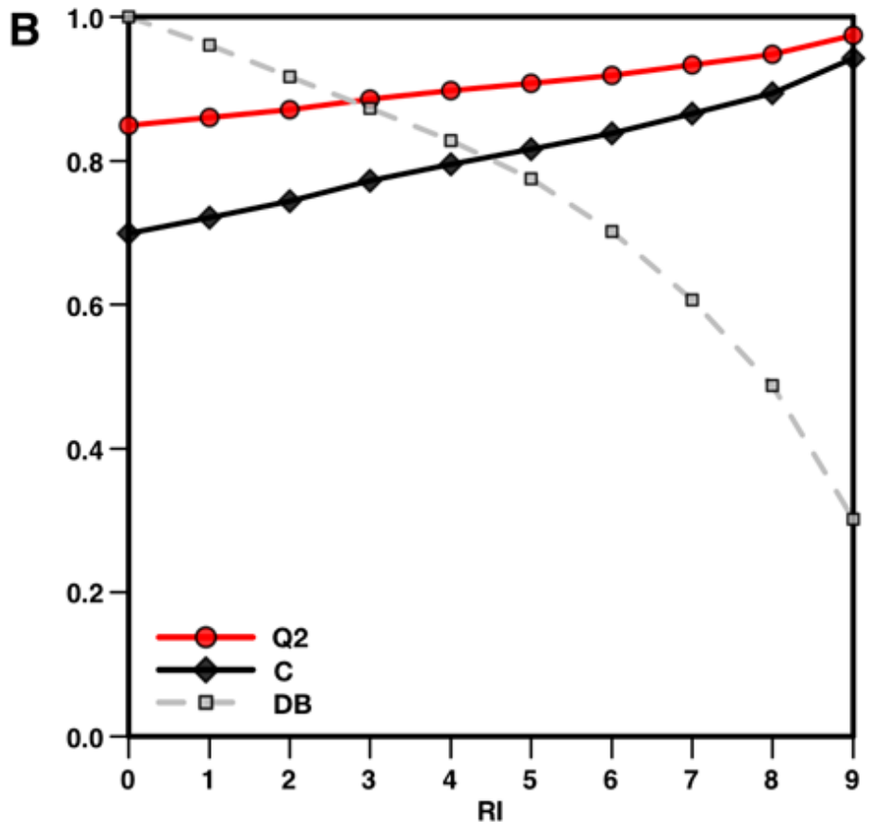


Sequence vs structure

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%.

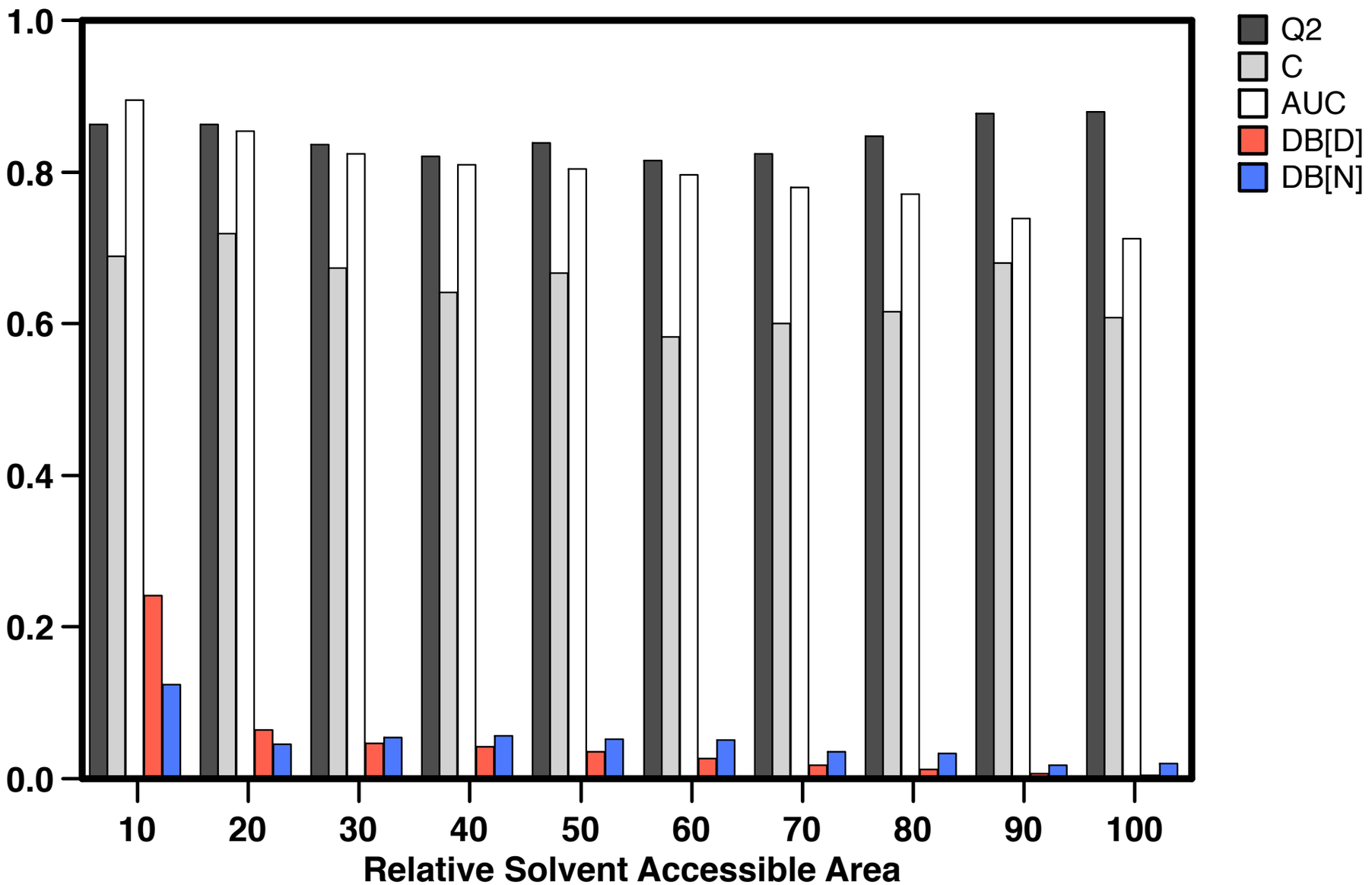
	Q2	P[D]	S[D]	P[N]	S[N]	C	AUC
SNPs&GO	0.82	0.81	0.83	0.82	0.81	0.64	0.89
SNPs&GO^{3d}	0.85	0.84	0.87	0.86	0.83	0.70	0.92





Accuracy vs Accessibility

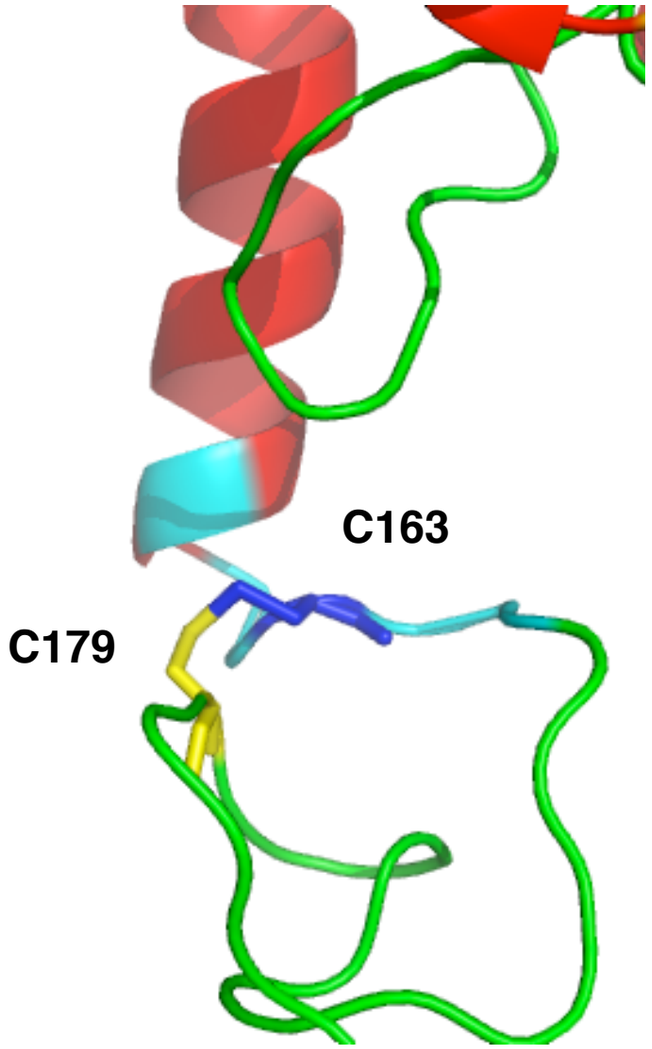
The predictions are more accurate for mutations occurring in buried region (0-30%). Mutations of exposed residues results in lower accuracy.



Prediction example

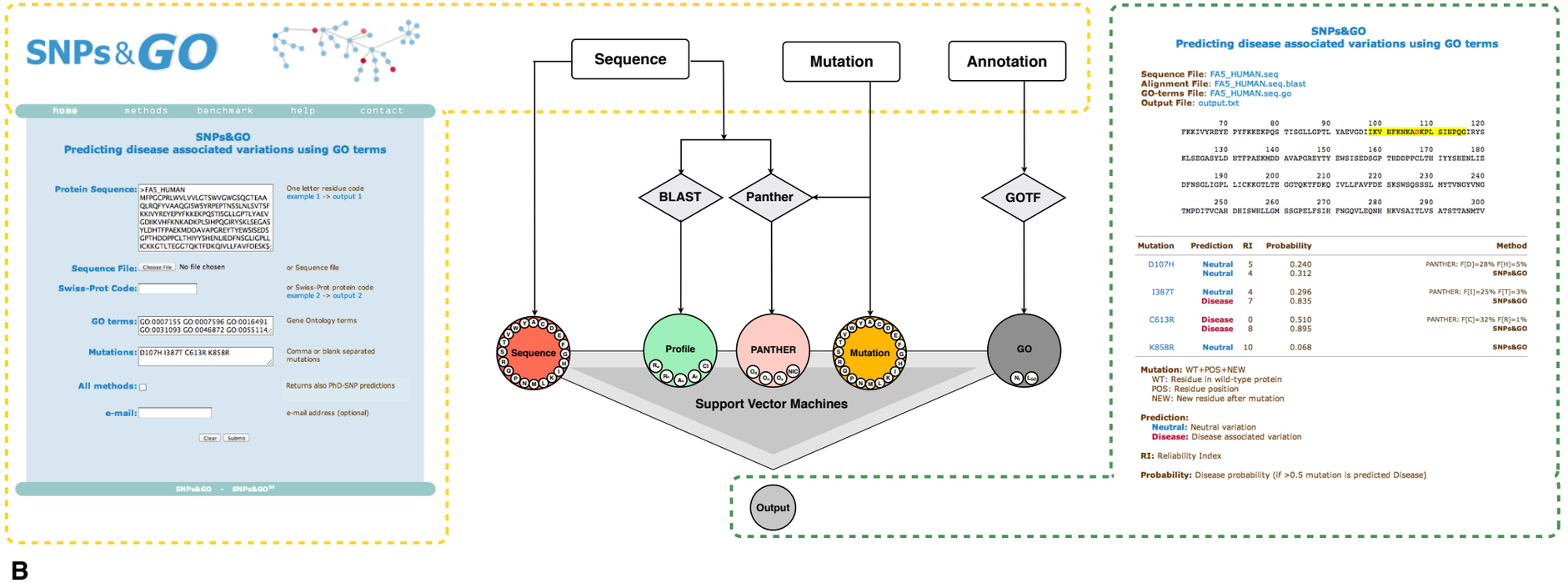
Damaging missing Cys-Cys interaction in the Glycosylasparaginase. The mutation p.Cys163Ser results in the loss of the disulfide bridge between Cys163 and Cys179. This SAP is responsible for Aspartylglucosaminuria.

1APY: Chain A, Res: 2.0 Å

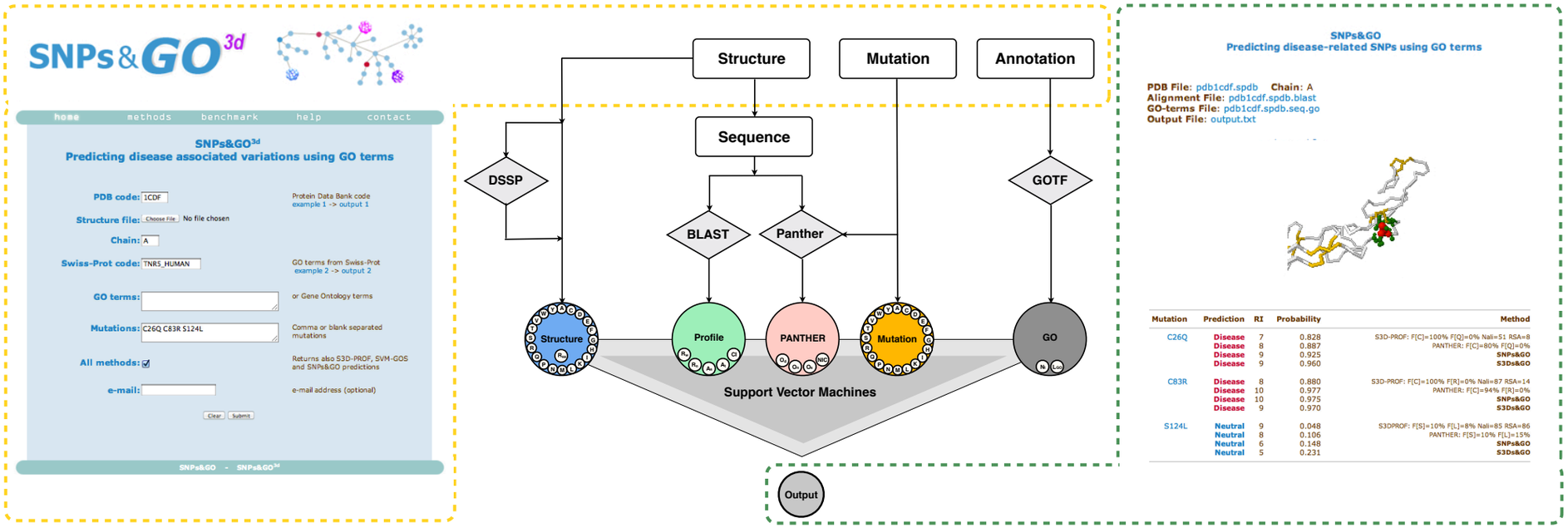


SNPs&GO web server

A



B



http://snps.biofold.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 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

SAVs Predictors

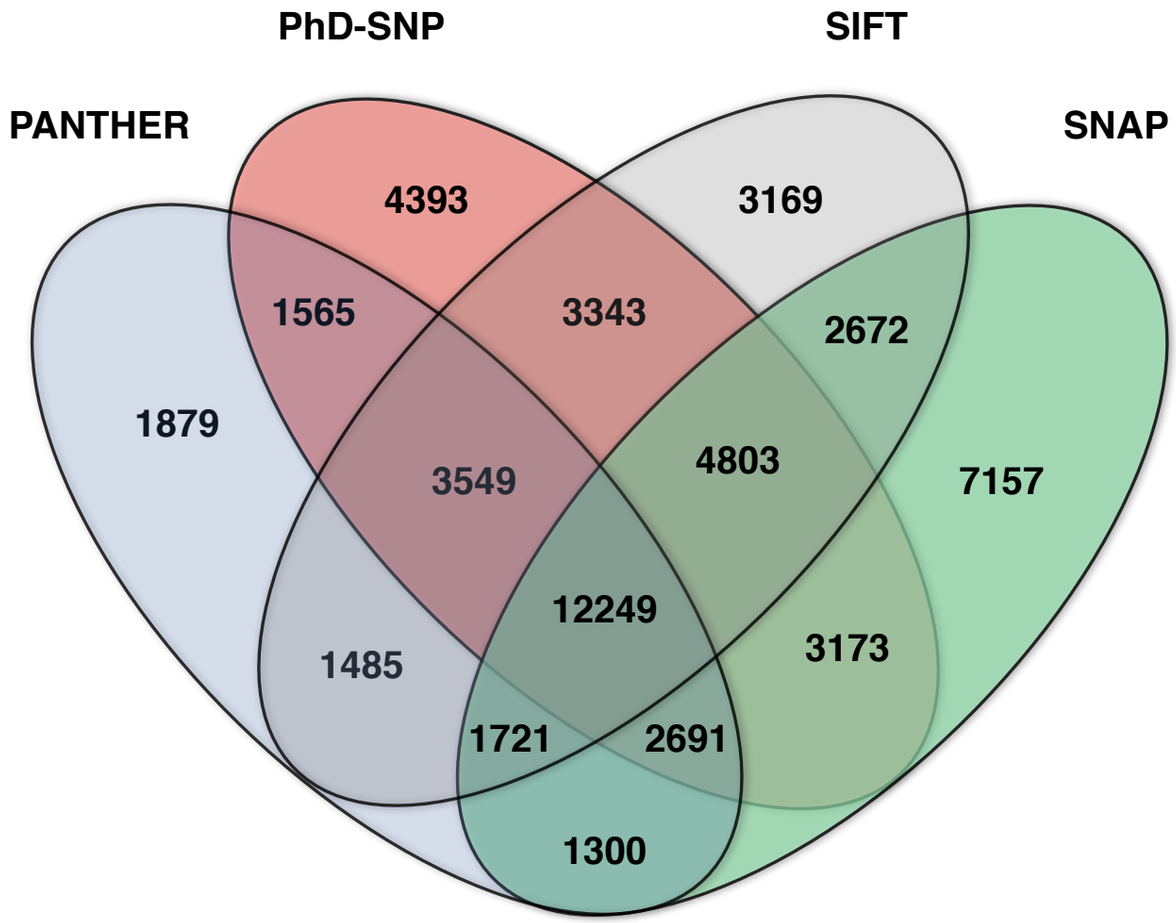
The higher correlation coefficient is between PANTHER and SIFT predictions. SNAP shows low correlation with PhD-SNP and PANTHER but higher correlation with SIFT which input is included in SNAP

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	-

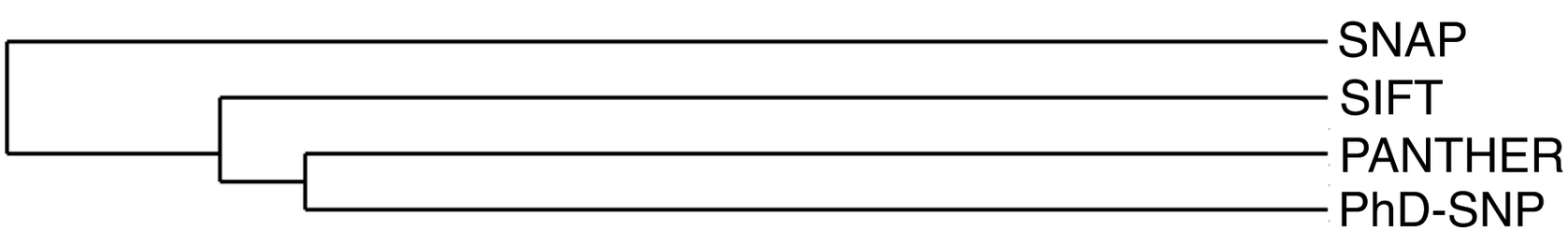
DB: Neutral 17993 and Disease 17993

Predictors tree

Using the prediction similarity we can build the predictors tree



UPGMA tree based on correlations



Prediction Analysis

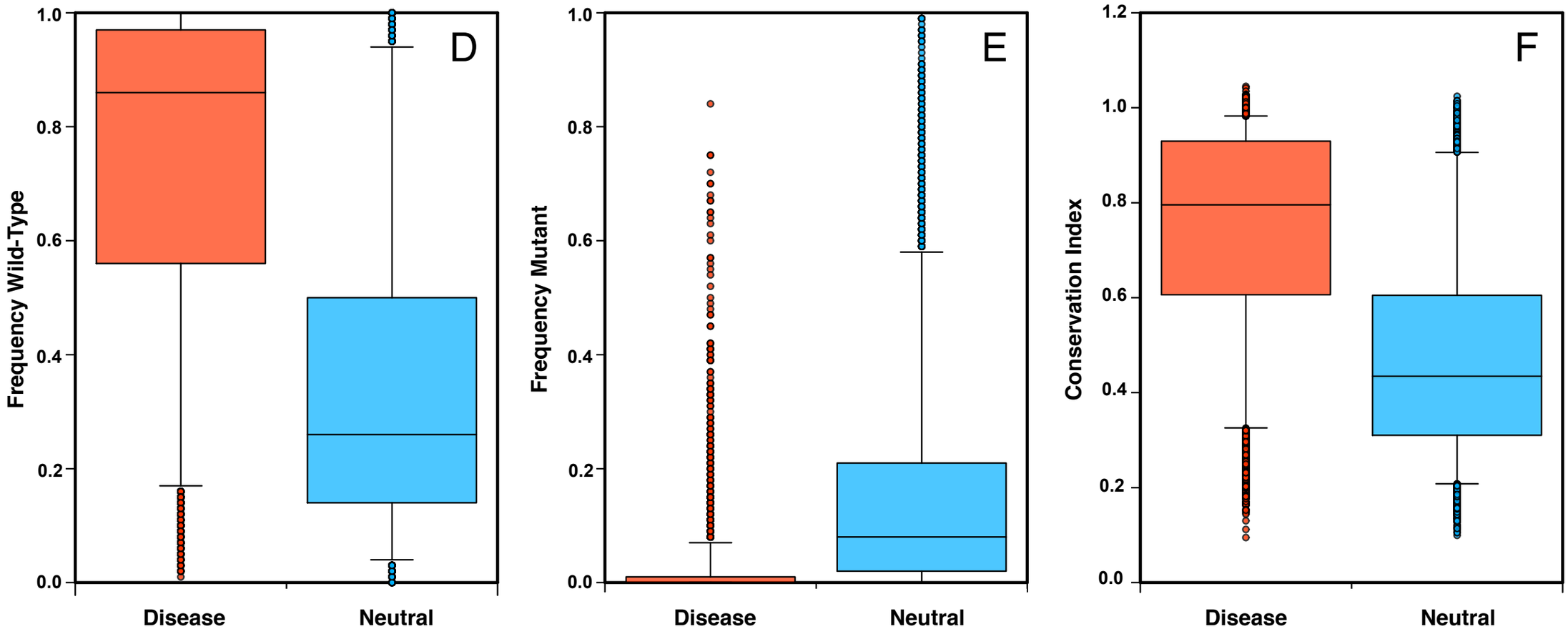
The accuracy of the predictions has been evaluated considering three different subset

- 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

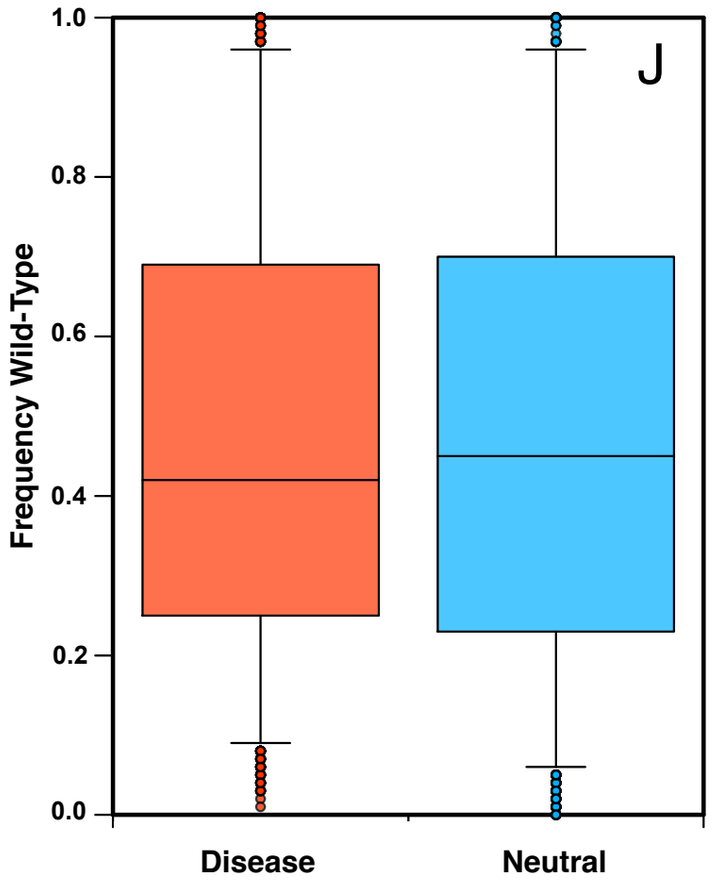
Consensus subset

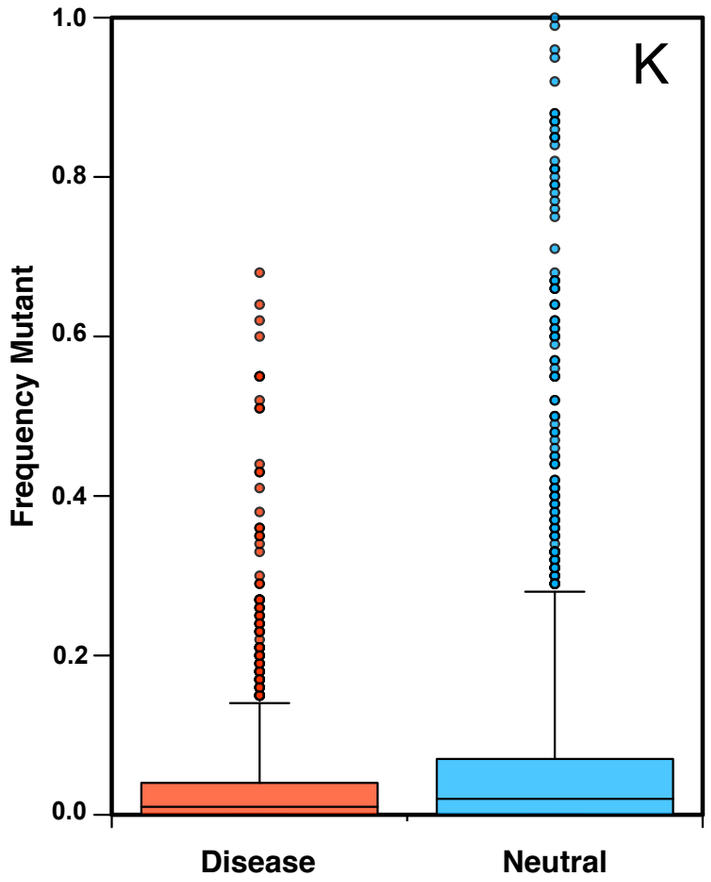
The distributions of the wild-type and new residues frequencies and CI for disease-related variants and polymorphisms on the *Consensus* subset have very little overlap.

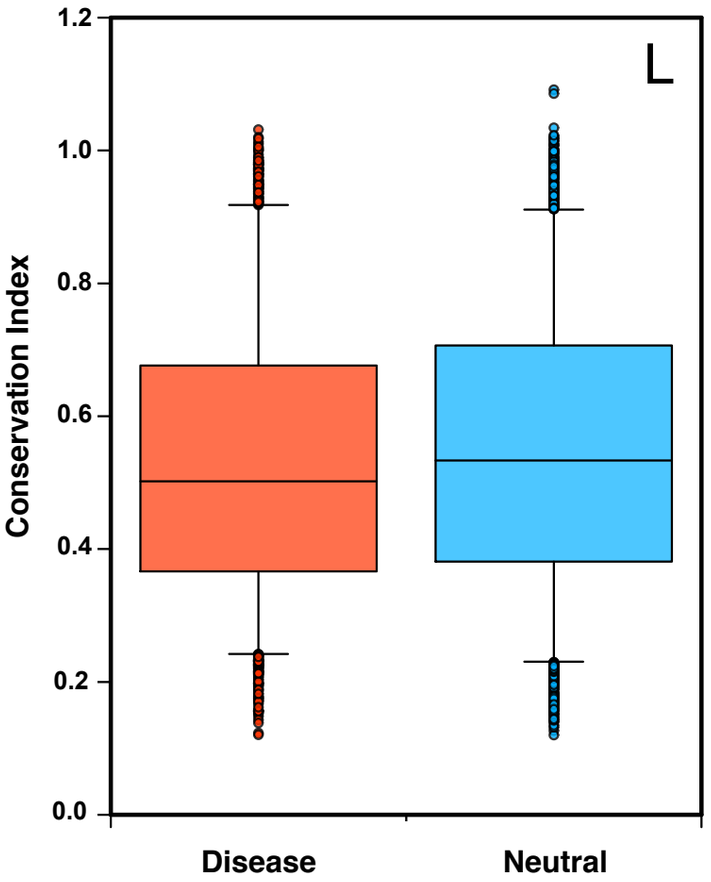


Tie subset

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.

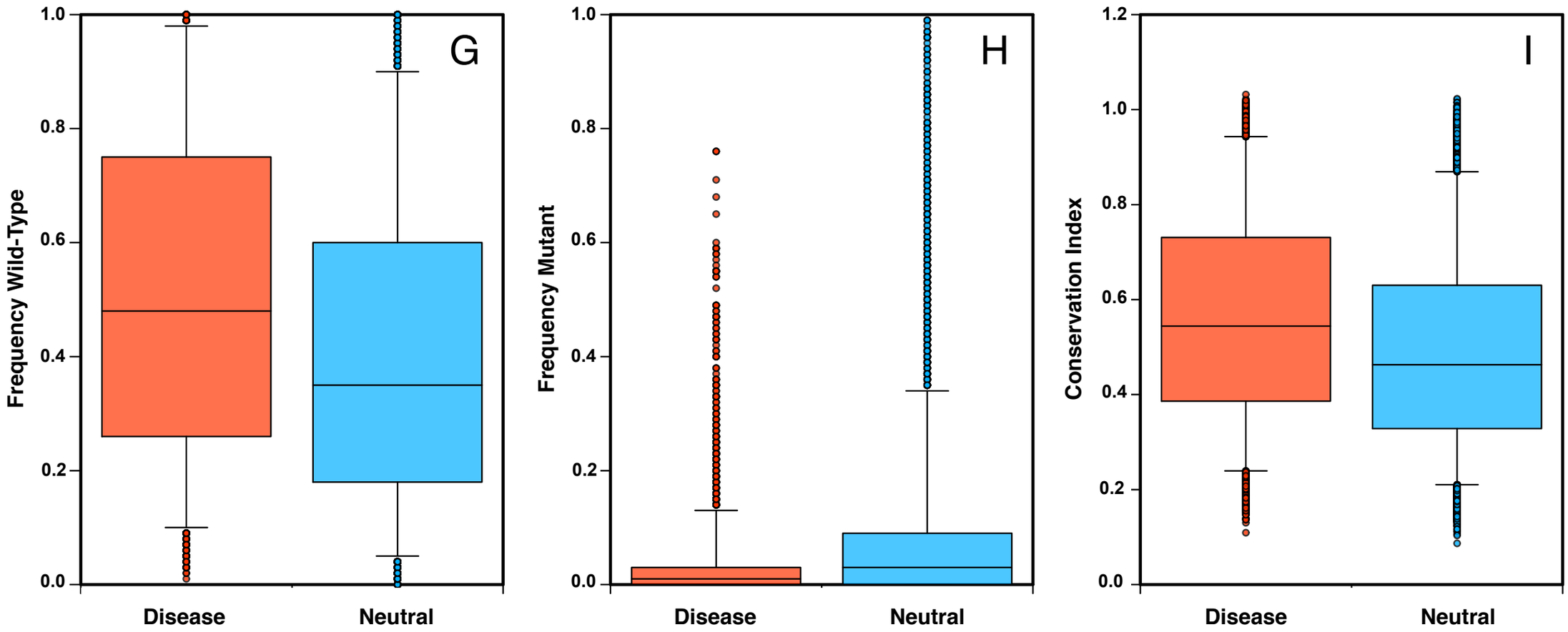






Majority subset

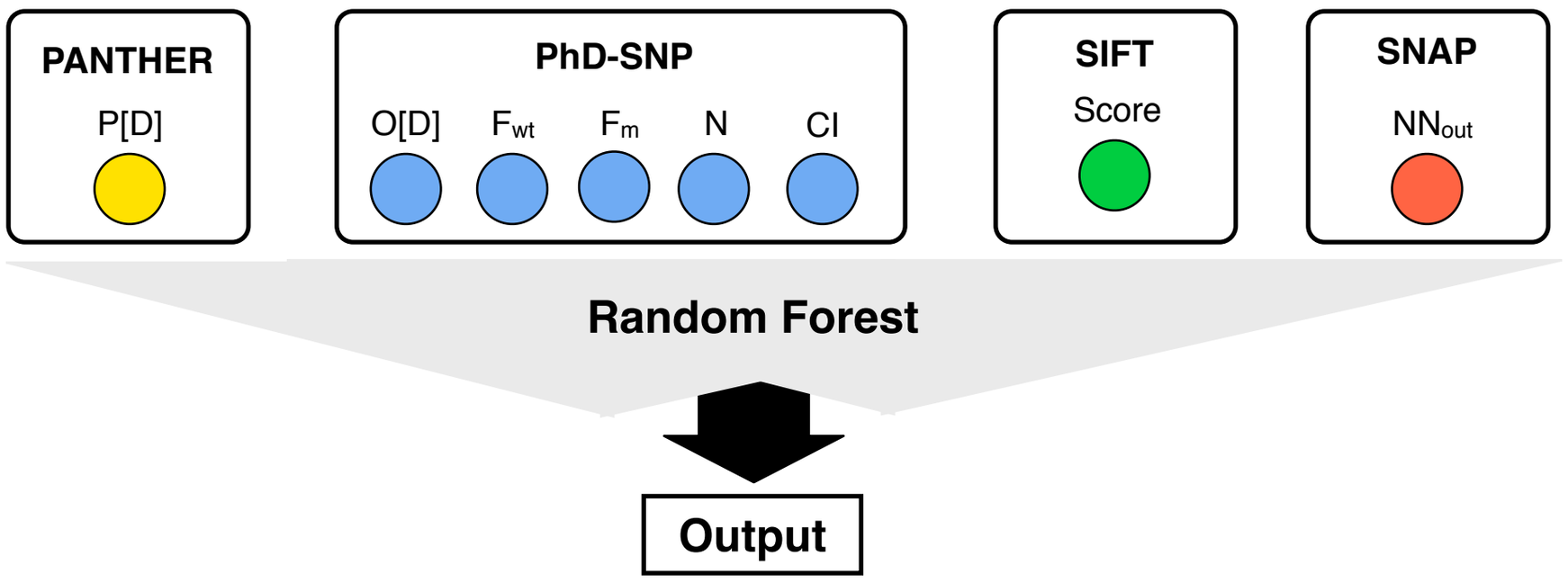
The distributions of the wild-type and new residues frequencies and CI for disease-related and polymorphism on the *Majority* subset are in an intermediate situation with respect to the previous cases.



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) .



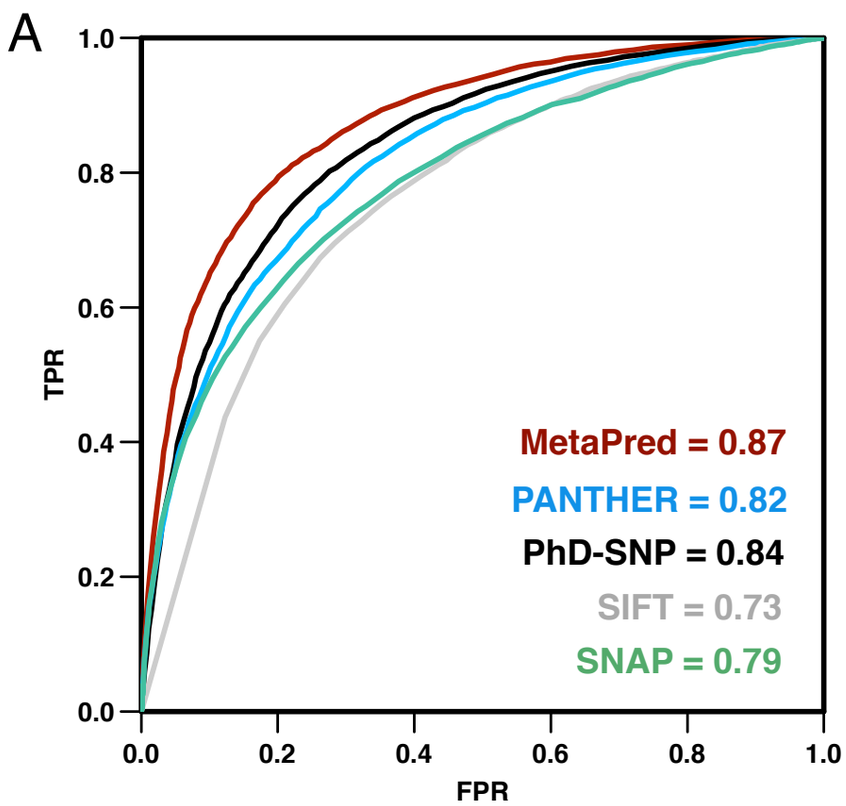
http://snps.biofold.org/meta-snp

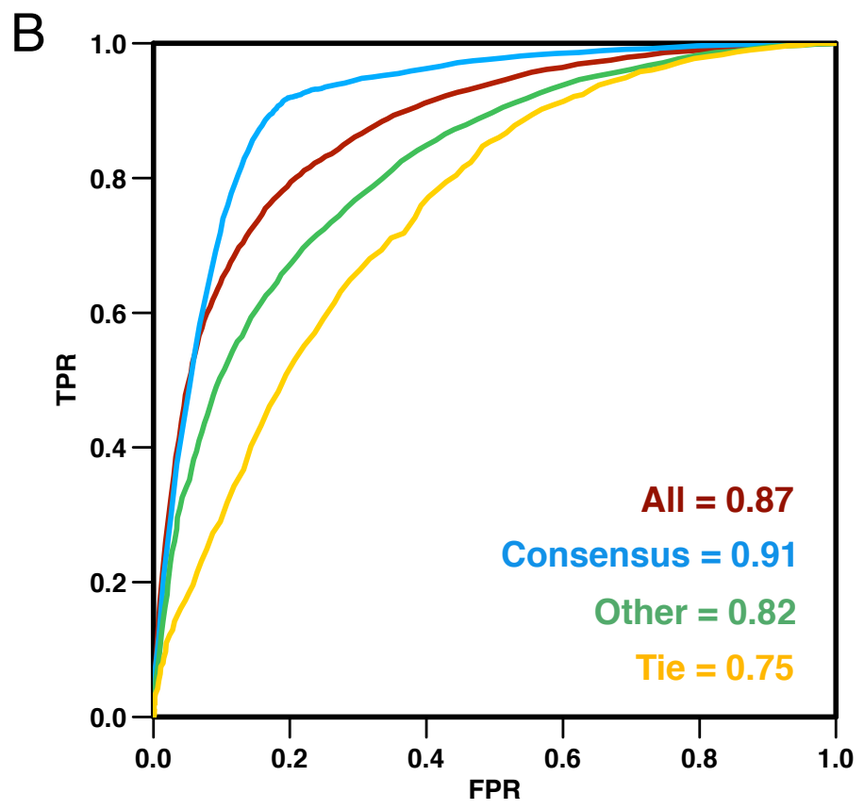
Meta-SNP accuracy

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



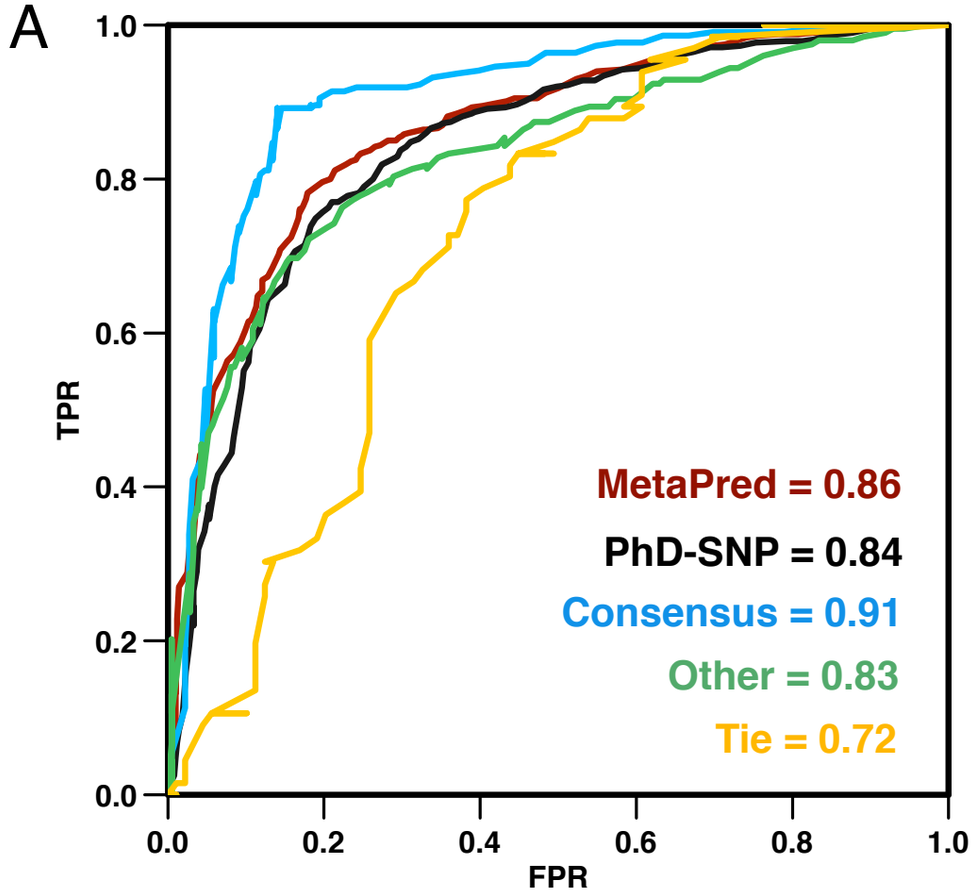


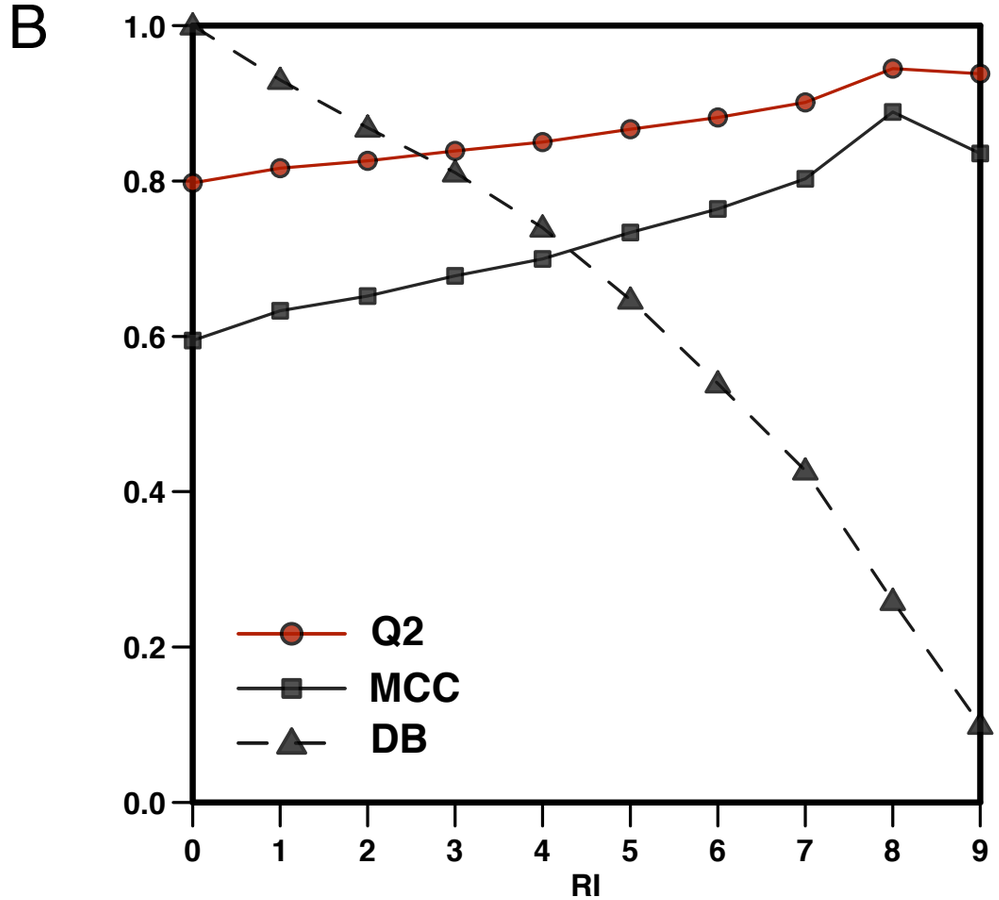
Testing Meta-SNP

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

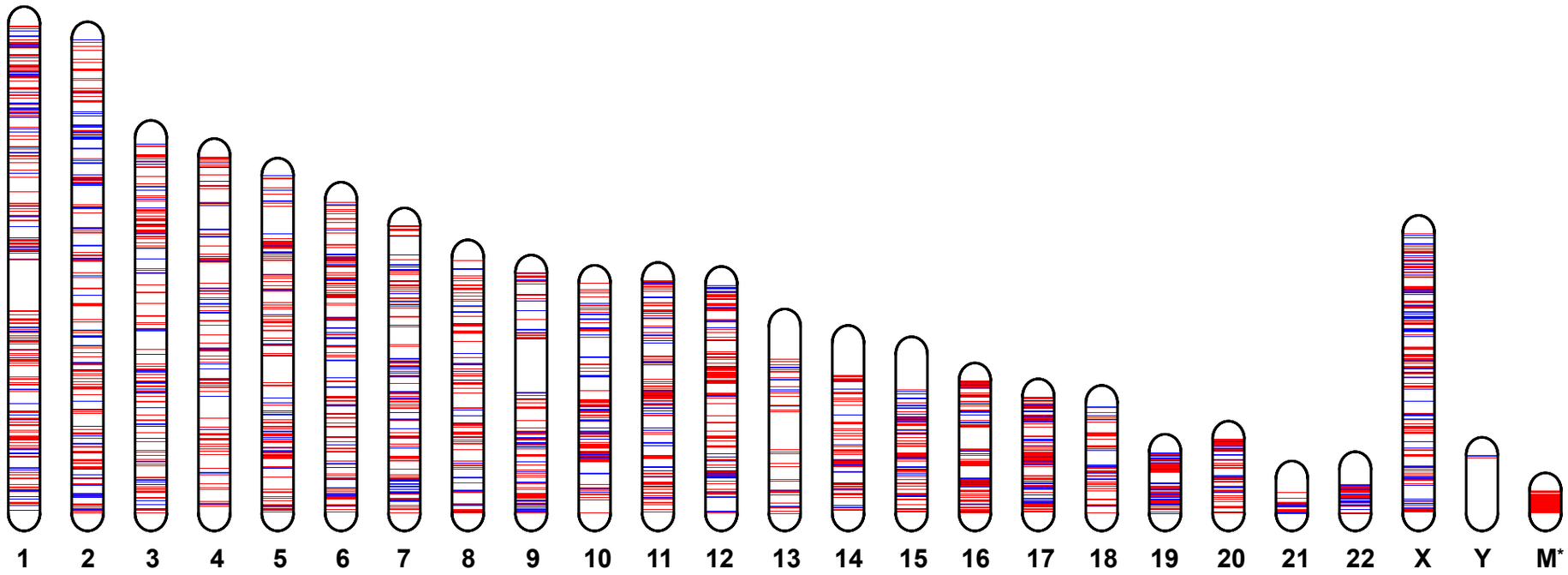




Capriotti et al. (2013). BMC Genomics. 14 (S3), In press.

Whole-genome predictions

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

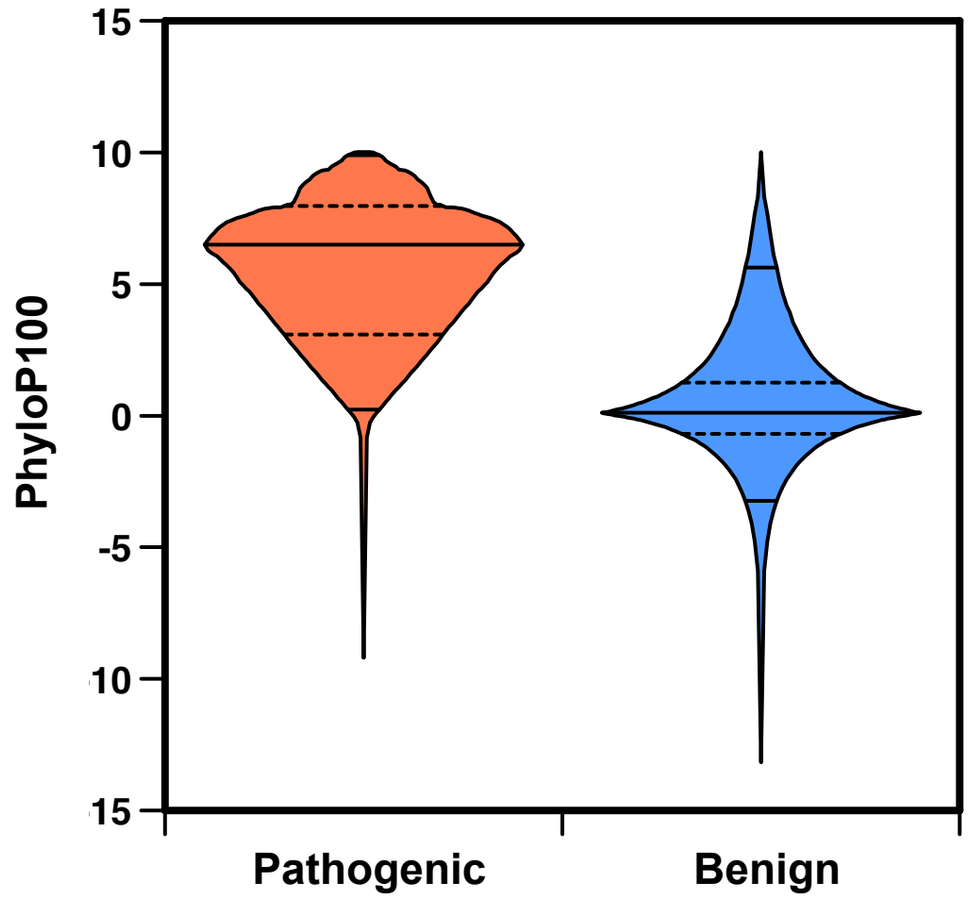


Predict the effect of SNVs in non-coding region is a challenging task because conservation is more difficult to estimate.

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

PhyloP100 score

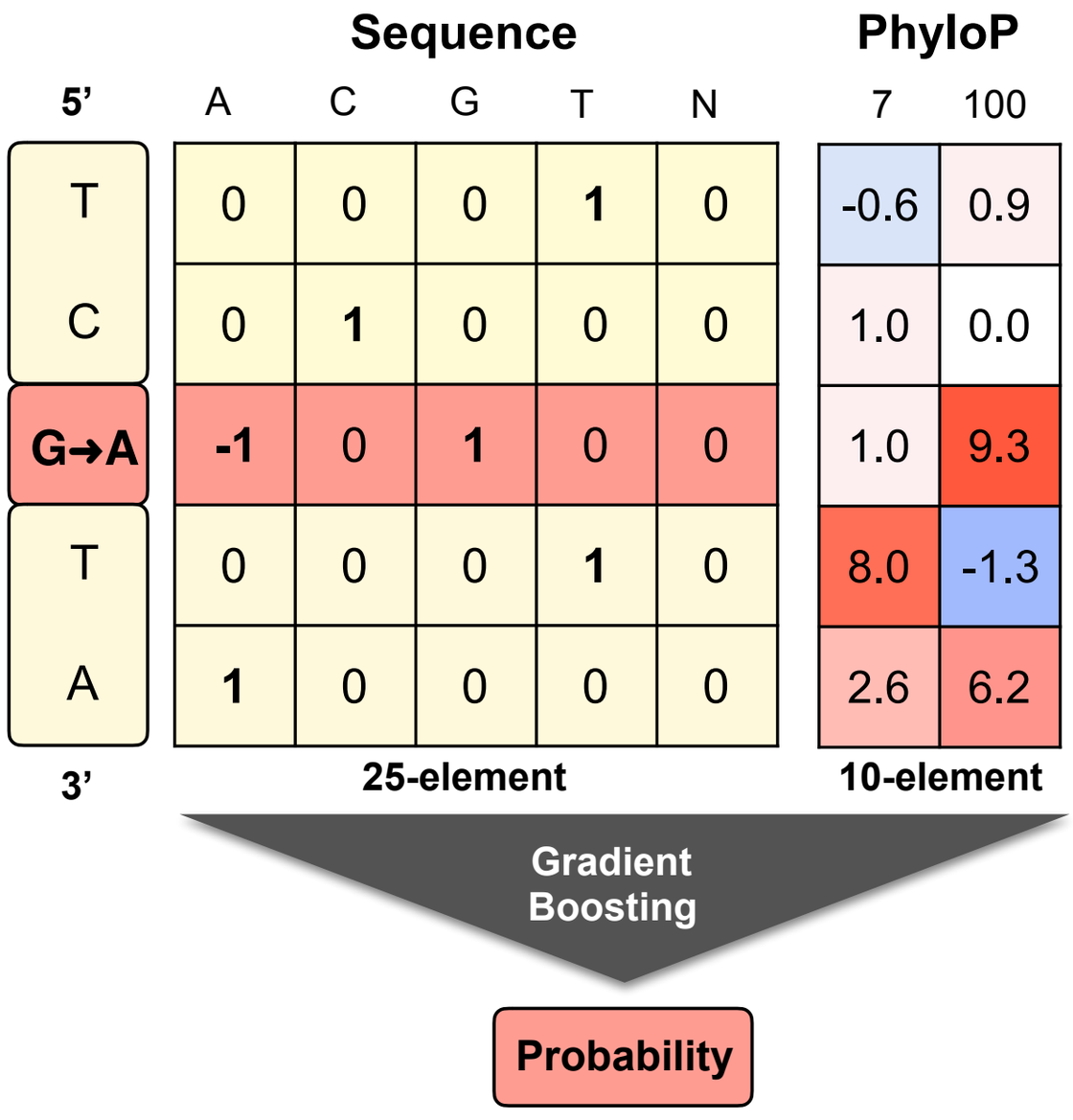
Conservation analysis based on the pre-calculated score available at the UCSC revealed a significant difference between the distribution of the PhyloP100 scores in Pathogenic and Benign SNVs.



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.

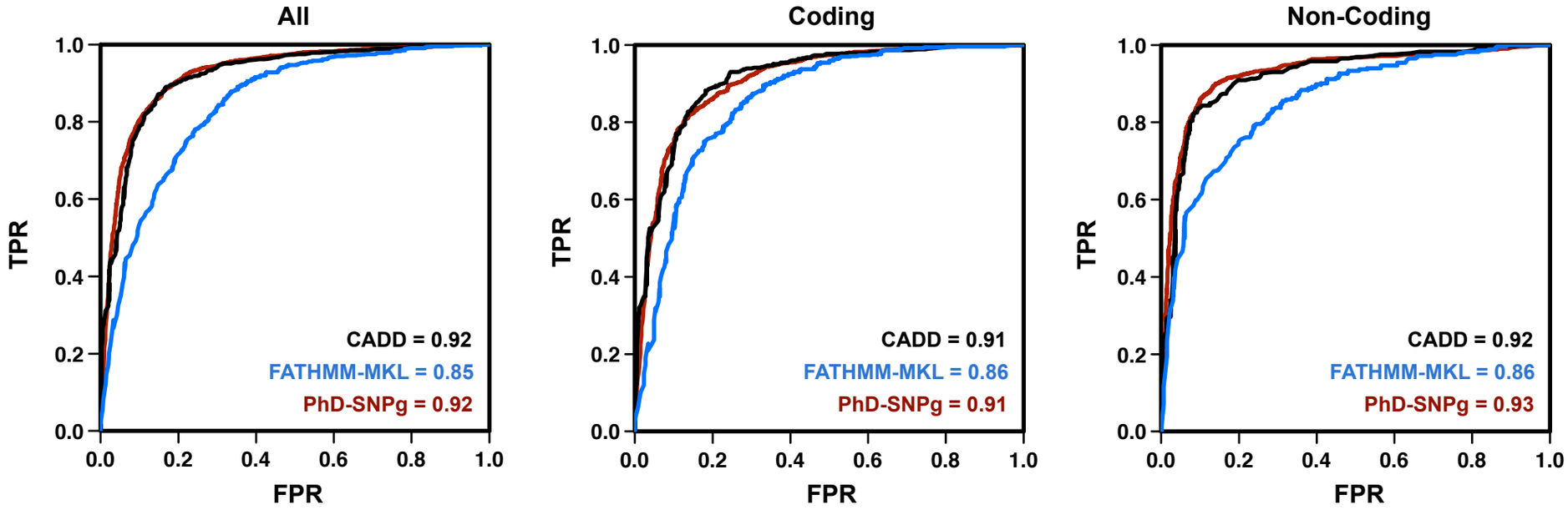


http://snps.biofold.org/phd-snpg/

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.

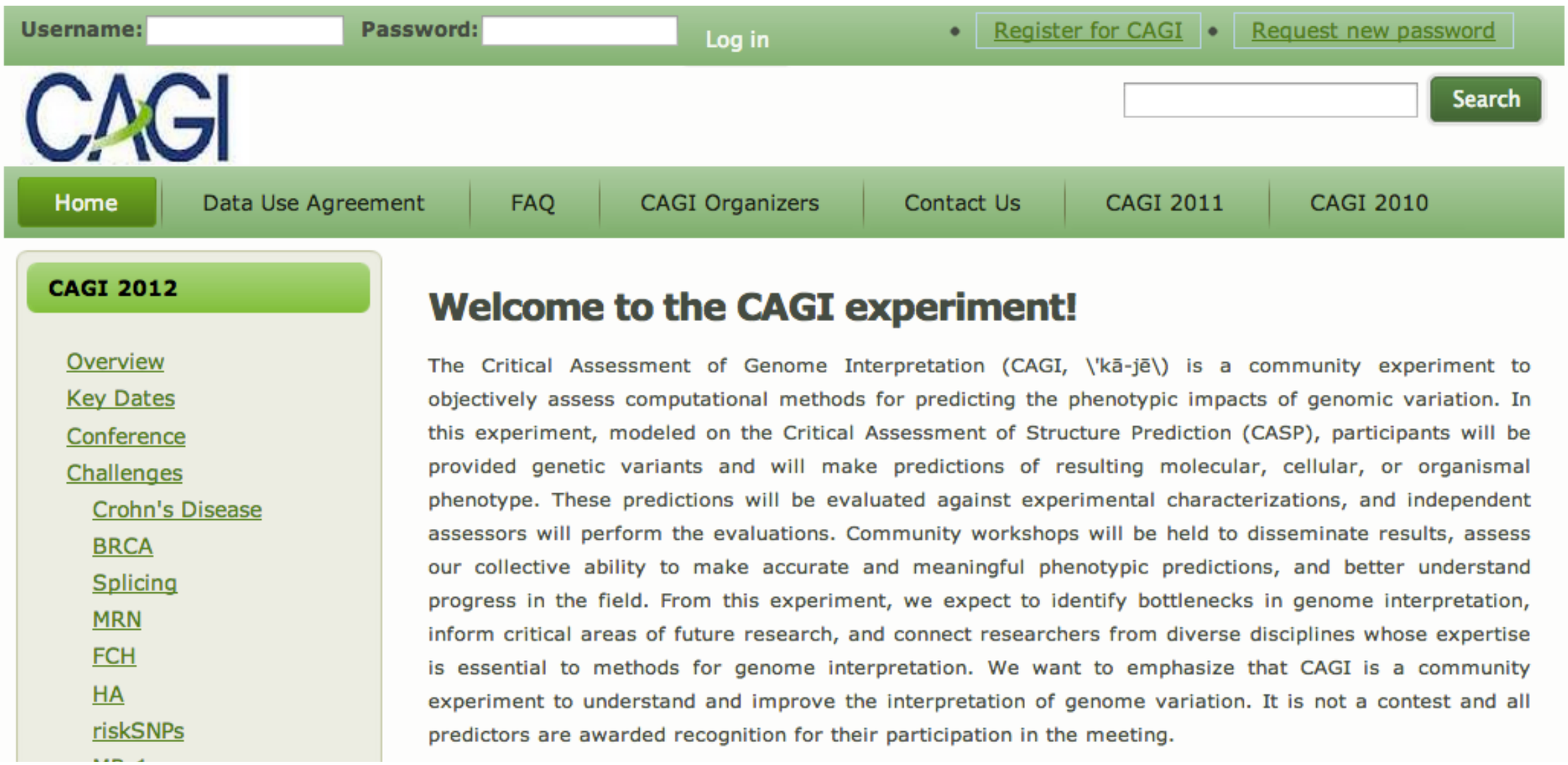
	Q2	TNR	NPV	TPR	PPV	MCC	F1	AUC
PhD-SNP^{g}	0.861	0.774	0.884	0.925	0.847	0.715	0.884	0.924
Coding	0.849	0.671	0.845	0.938	0.850	0.651	0.892	0.908
Non-Coding	0.876	0.855	0.911	0.901	0.839	0.753	0.869	0.930



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

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.



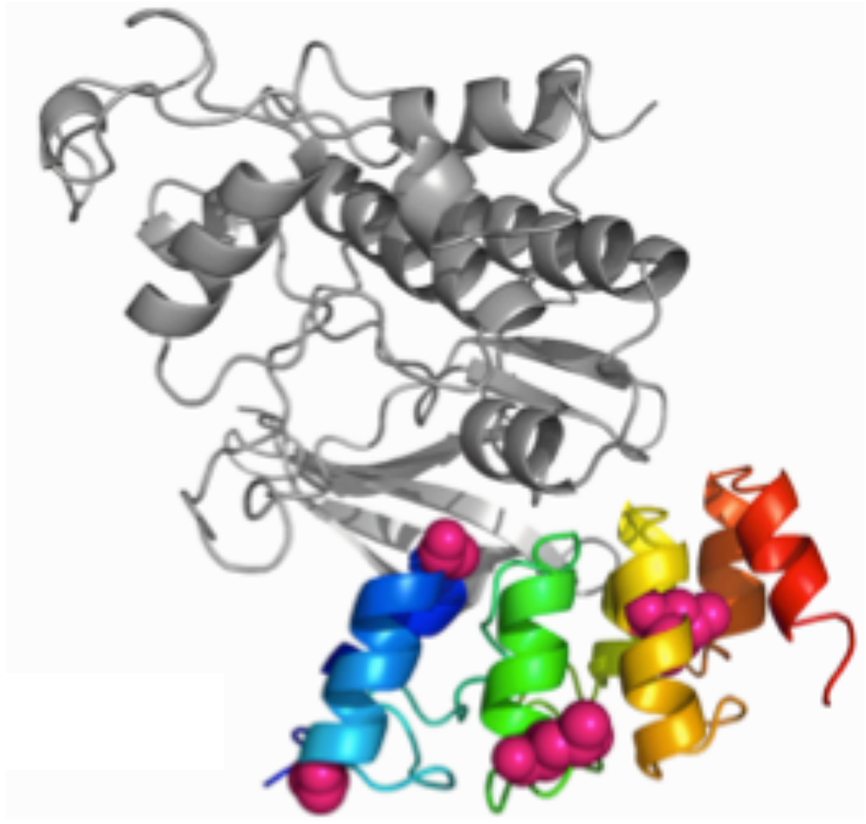
https://genomeinterpretation.org/

The 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

The complexity of cancer

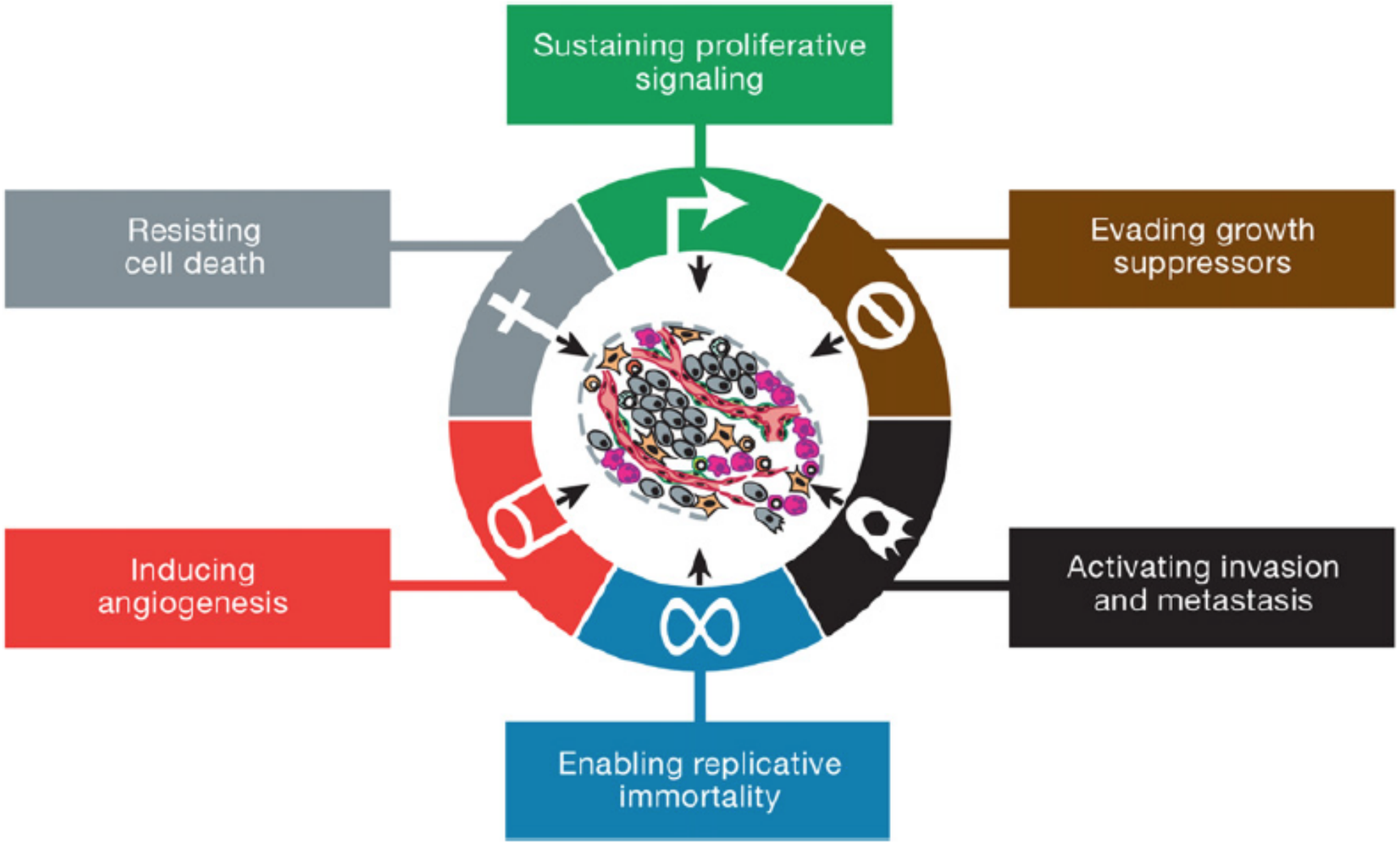
Cancer is **complex disorder** characterized by high level of mutation rate.

Mutations can be classified in germline and somatic whether they are inherited from parents or the result of error in DNA replication.

Another classification is between driver and passenger mutations whether they provide selective advantage with respect to normal cells increasing their proliferation rate or not.

Hallmarks of cancer

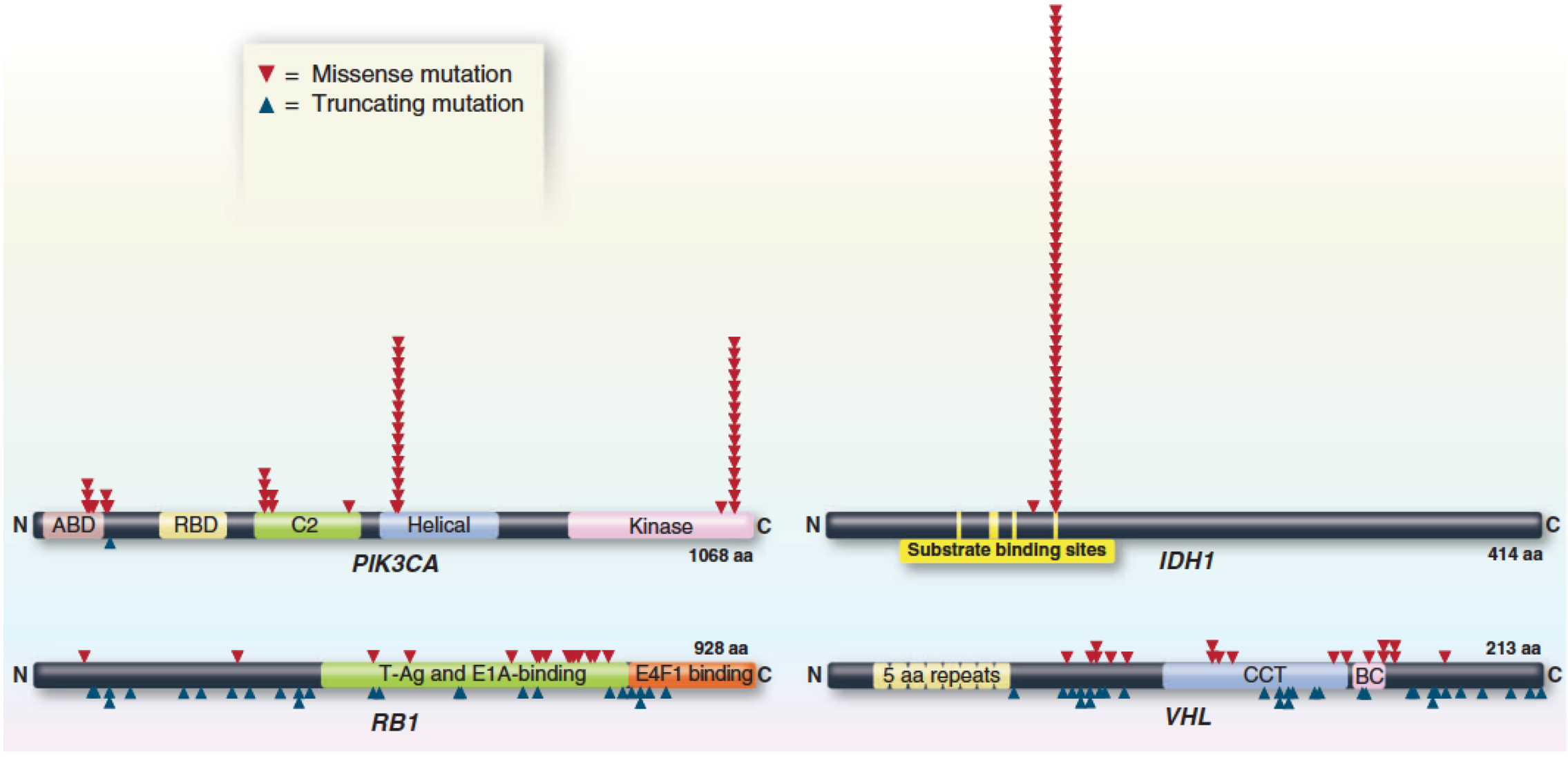
The six hallmarks of cancer - distinctive and complementary capabilities that enable tumor growth and metastatic dissemination.



Hanahan and Weinberg. Cell 2011, **144**:646

Oncogene vs Suppressor

Oncogenes have highly recurrent mutations, Tumor suppressors have sparse variants.

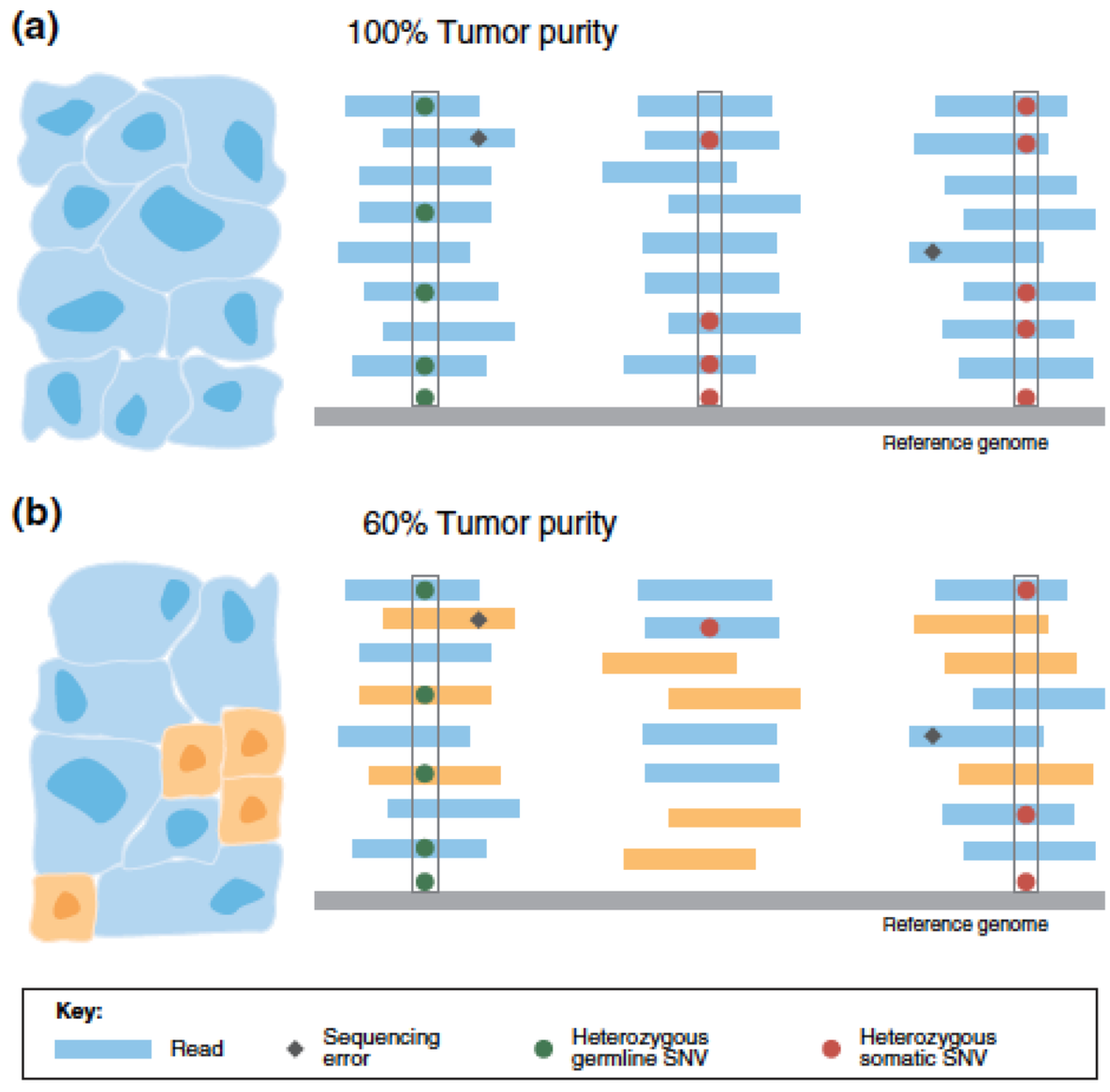


Vogelstein et al. Science 2013, **339**:1546

Main challenges

Computational methods for cancer genome interpretation have been developed to address the following issues:

- Detection of recurrent somatic mutations and cancer driver genes;
- Prediction of driver variants and their functional impact;
- Estimate the impact of multiple variants at network and pathway level;
- Differentiate subclonal populations and their variation pattern.



Raphael et al. Genome Medicine 2014, **6**:5

How data looks like?

Variant Calling File (VCF) with germline and somatic variants

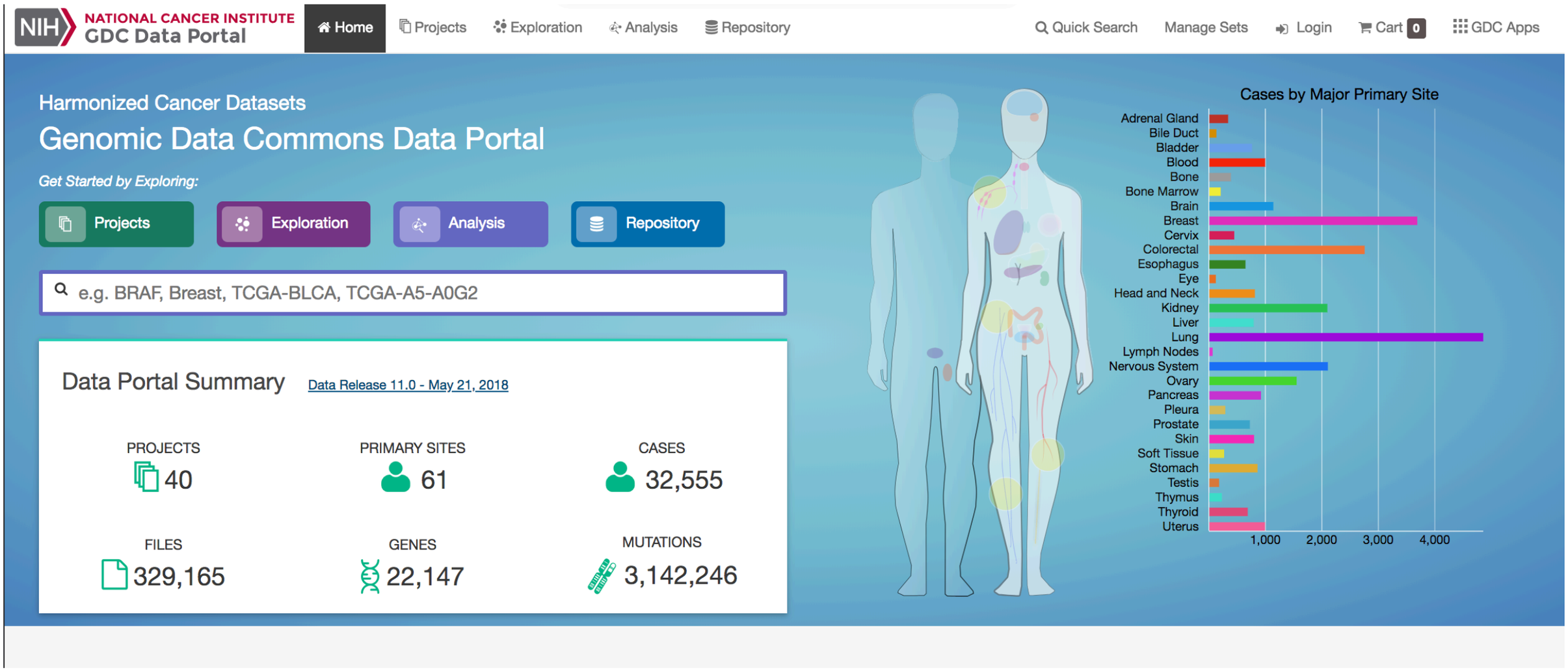
```
##fileformat=VCFv4.1
##tcgaversion=1.1
##reference=<ID=hg19,source=.>
##phasing=none
##geneAnno=none
##INFO=<ID=VT,Number=1,Type=String,Description="Variant type, can be SNP, INS or DEL">
##INFO=<ID=VLS,Number=1,Type=Integer,Description="Final validation status relative to non-adjacent Normal, ......”>
##FILTER=<ID=CA,Description="Fail Carnac (Tumor and normal coverage, tumor variant count, mapping quality, ......”>
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read depth at this position in the sample">
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Depth of reads supporting alleles 0/1/2/3...">
##FORMAT=<ID=BQ,Number=.,Type=Integer,Description="Average base quality for reads supporting alleles">
##FORMAT=<ID=SS,Number=1,Type=Integer,Description="Variant status relative to non-adjacent Normal,0=wildtype, ......">
##FORMAT=<ID=SSC,Number=1,Type=Integer,Description="Somatic score between 0 and 255">
##FORMAT=<ID=MQ60,Number=1,Type=Integer,Description="Number of reads (mapping quality=60) supporting variant">
#CHROM     POS         ID    REF   ALT   QUAL  FILTER     INFO             FORMAT                      NORMAL                  PRIMARY
1          10048       .     C     CCT   .     CA         VT=INS;VLS=5     GT:DP:AD:BQ:SS:SSC:MQ60     0/0:66:.,0:.:0:.:0      0/1:32:.,2:.:2:.:0
1          10078       .     CT    C     .     CA         VT=DEL;VLS=5     GT:DP:AD:BQ:SS:SSC:MQ60     0/0:25:.,0:.:0:.:0      0/1:13:.,2:.:2:.:0
1          10177       .     A     AC    .     CA         VT=INS;VLS=5     GT:DP:AD:BQ:SS:SSC:MQ60     0/0:57:.,0:.:0:.:0      0/1:22:.,2:.:2:.:0
. . . . .
. . . . .
1          900505      .     G     C     .     PASS       VT=SNP;VLS=5     GT:DP:AD:BQ:SS:SSC:MQ60     0/1:188:.,89:26:1:.:81  0/1:210:.,113:24:1:.:100
. . . . .
. . . . .
1          1991007     .     G     T     .     PASS       VT=SNP;VLS=5     GT:DP:AD:BQ:SS:SSC:MQ60     0/0:222:.,1:2:0:.:1     0/1:88:.,41:25:2:50:34
. . . . .
```

The TCGA data

The Cancer Genome Atlas Consortium

TCGA data (https://portal.gdc.cancer.gov/)

- 33 cancer projects (~11,300 cases)
- BAM files available



The ICGC data portal

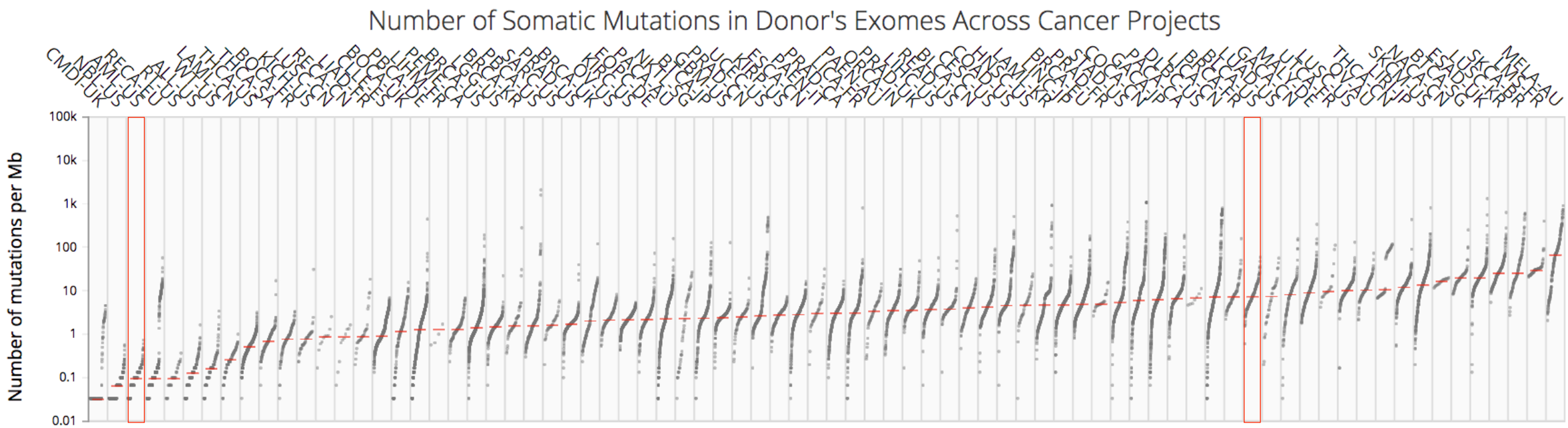
The International Cancer Genome Consortium

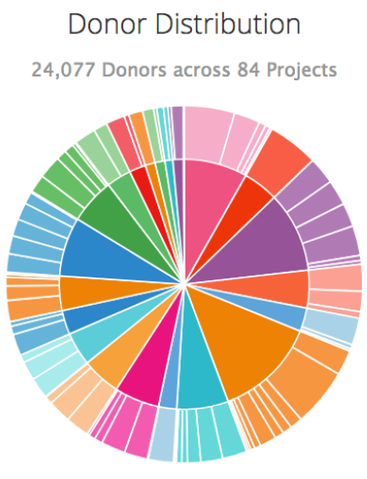
ICGC (https://dcc.icgc.org/)

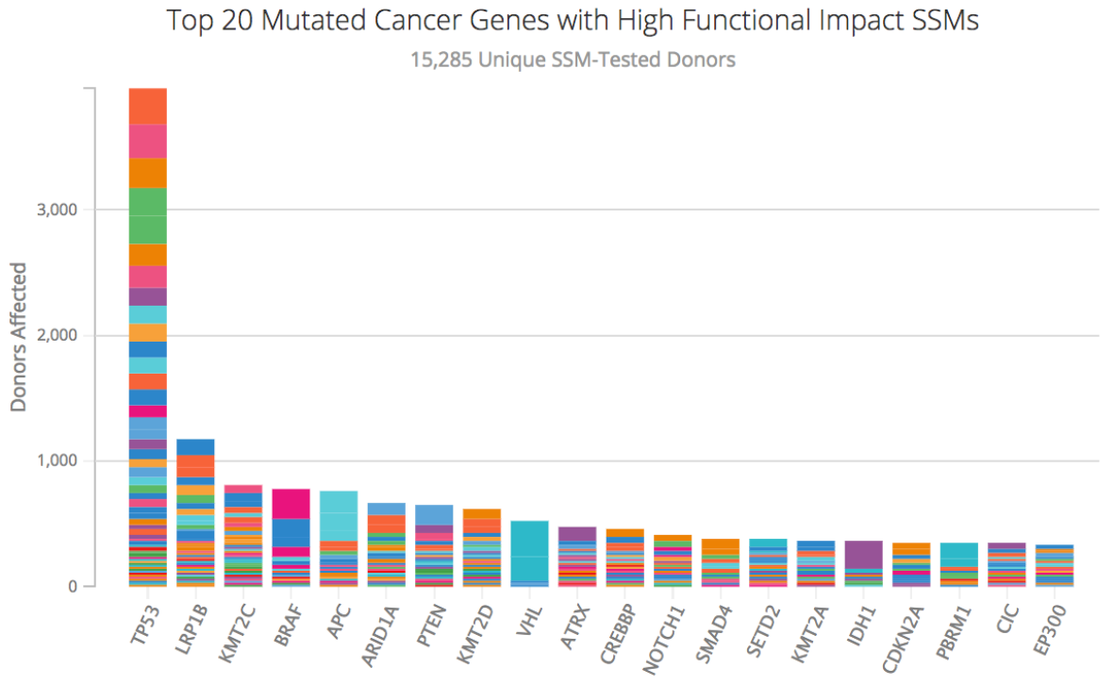
- 20,487 cancer patients
- 84 cancer types in 22 primary sites for which sequencing data are available
- 77.4 million simple somatic mutations.

Somatic Mutations

Number of somatic mutations per sample vary significantly across cancer types



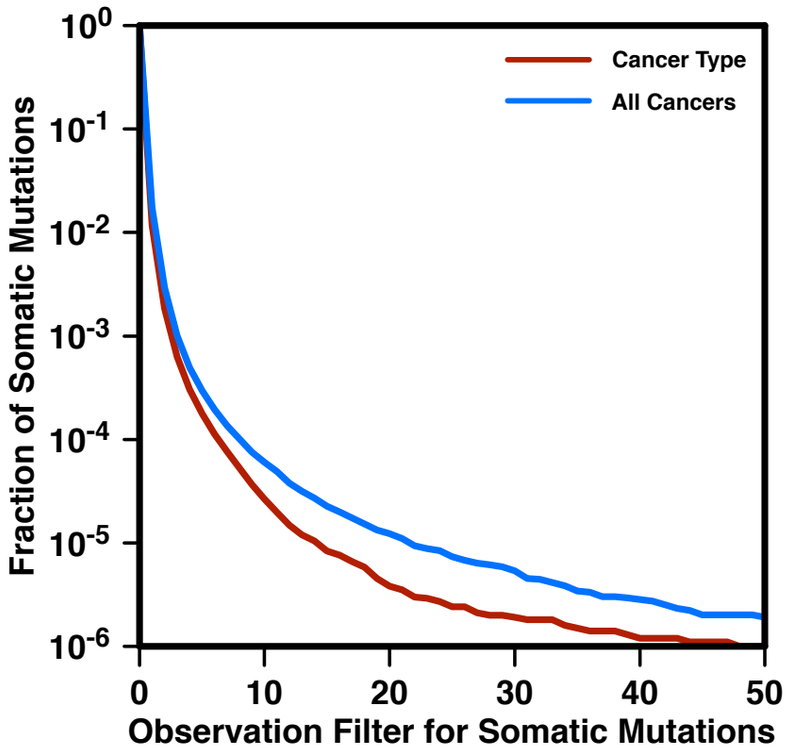


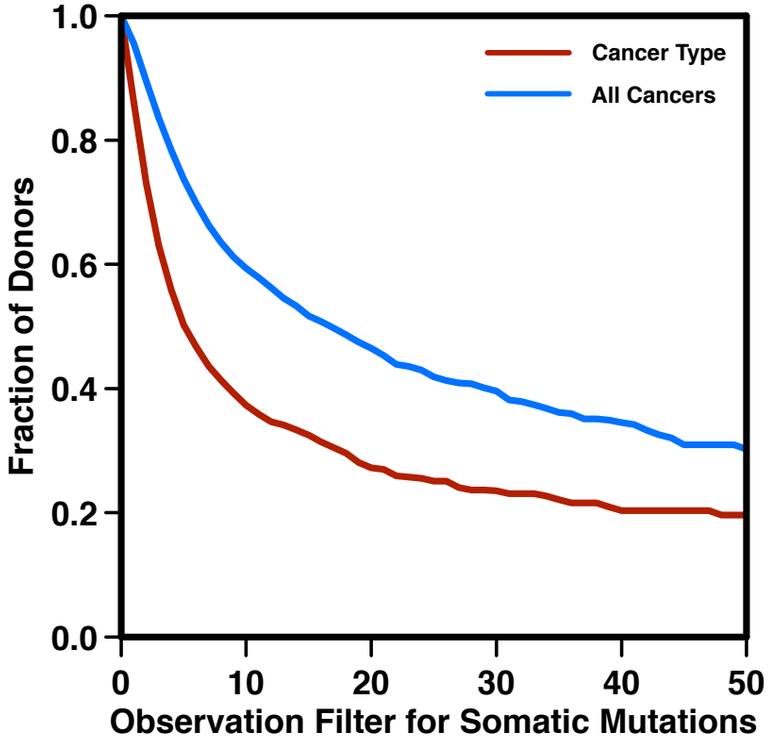


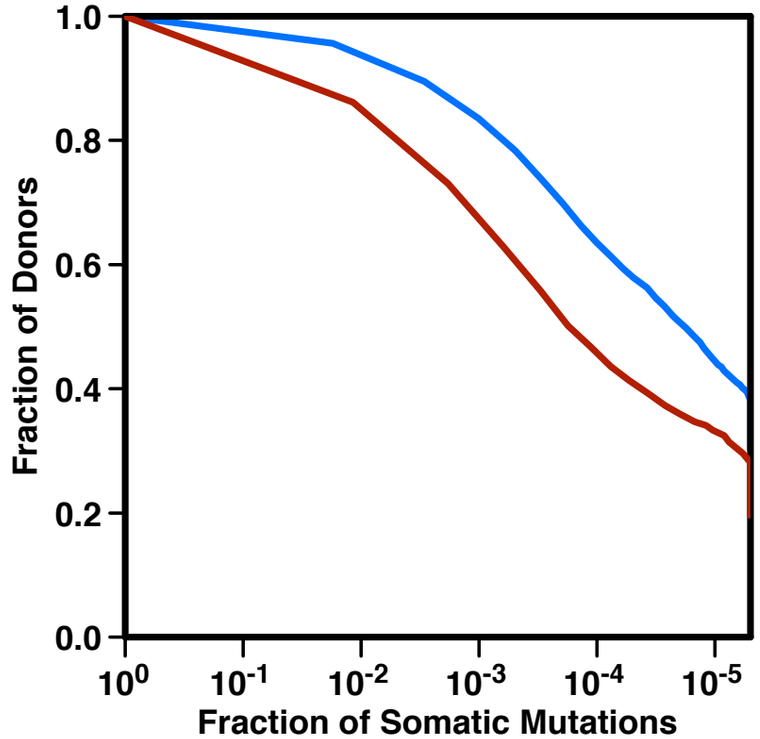
Driver vs Passenger

Number of recurrent mutations decrease exponentially.
On average a small fraction of variants a present in the majority of the samples.

Selecting mutations that are repeated at least twice we filter out ~98% mutations
and are still able to recover ~96% of the patients



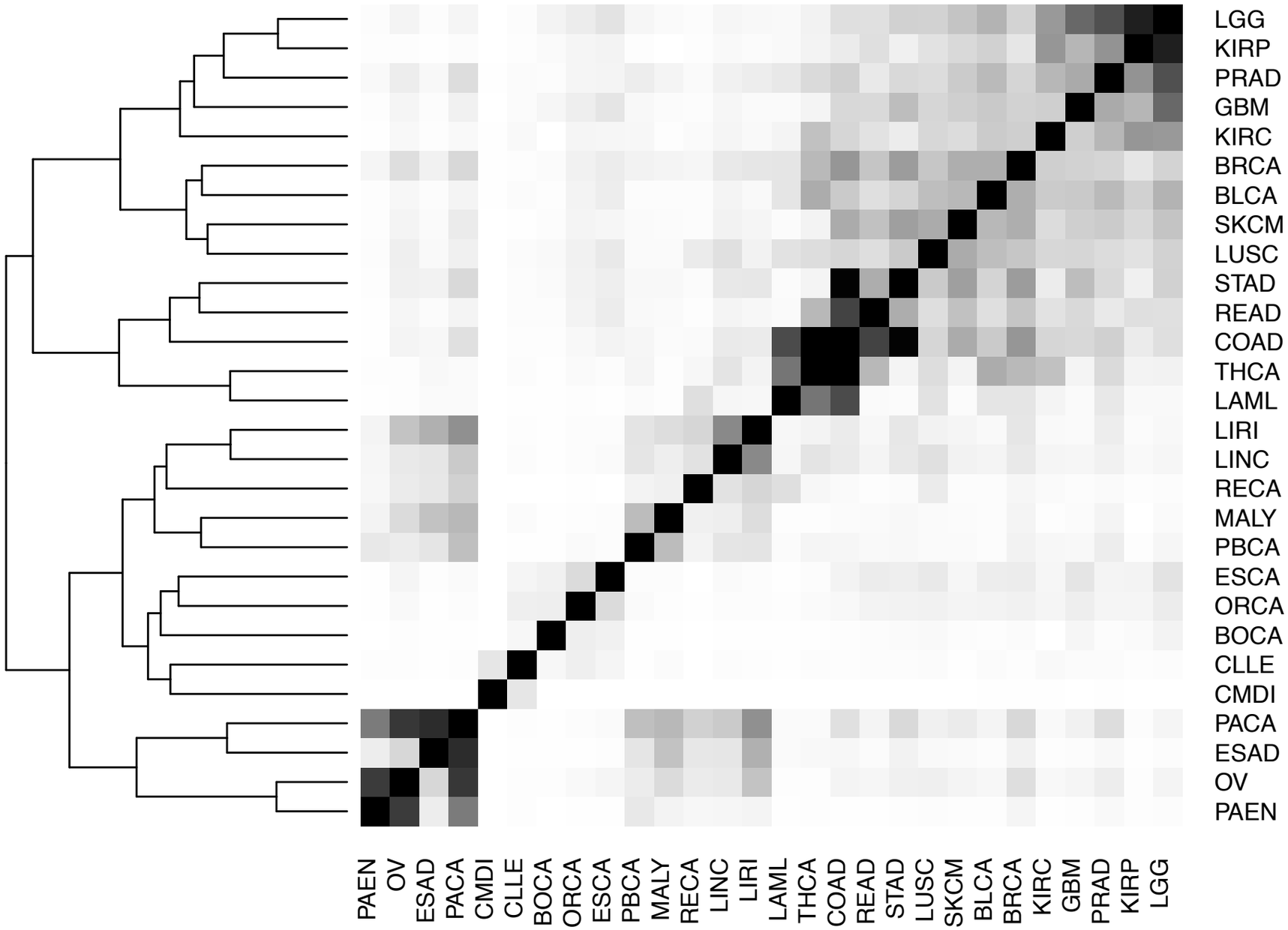




Tian R, Basu M, Capriotti E. BMC Genomics 2015

The Cancer Tree

The analysis of recurrent somatic mutations can be used to define similarities across cancer types.

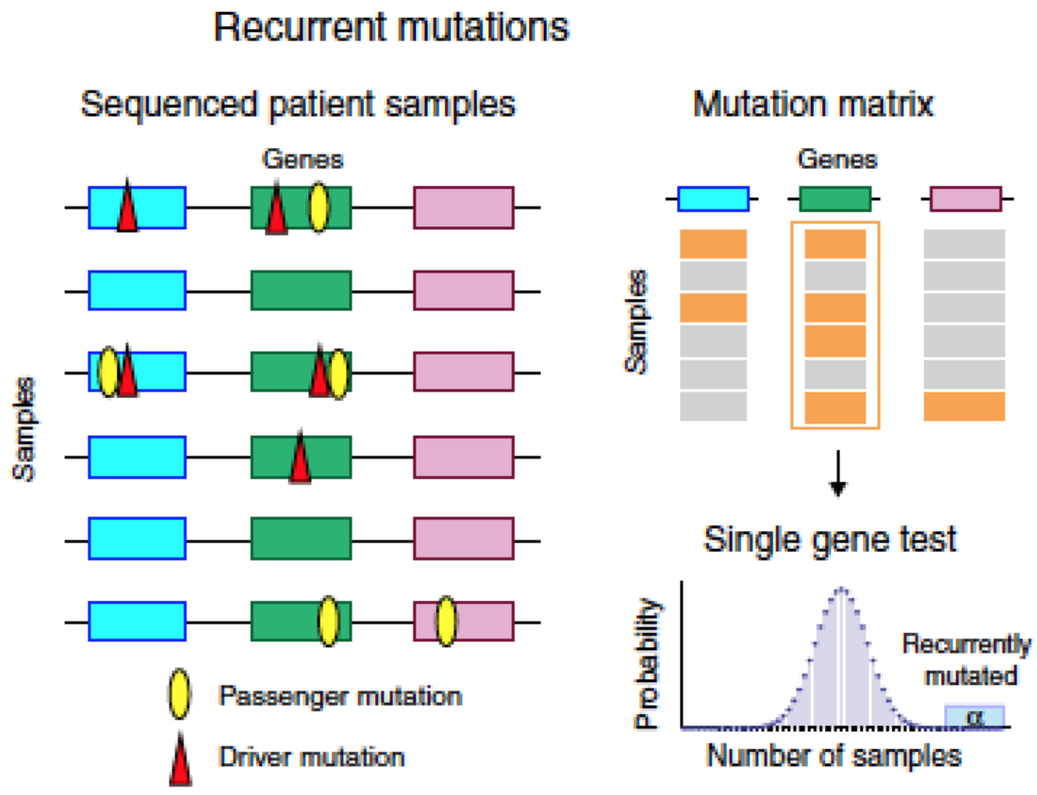


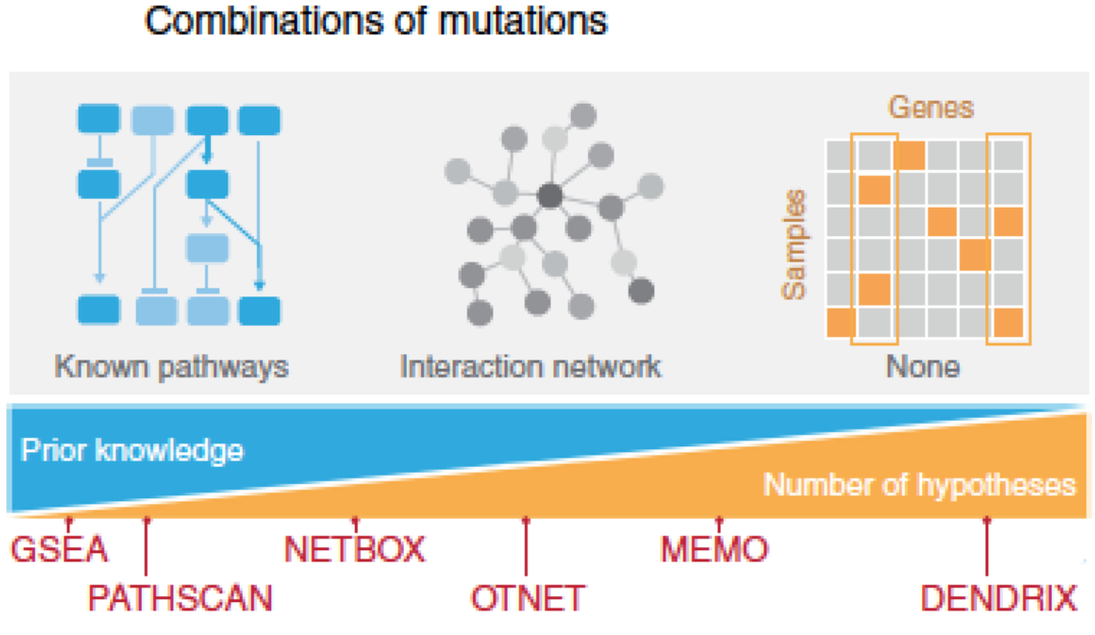
Recurrent variations

Recurrent mutations that are found in more samples than would be expected by chance are good candidates for driver mutations.

To identify such recurrent mutations, a statistical test is performed which usually collapses all the non-synonymous mutations in a gene.

Identification of recurrent mutations in predefined groups such as pathways and protein-protein interaction networks and de novo identification of combinations, without relying on a priori definition.





Raphael et al. Genome Medicine 2014, **6**:5

The main idea

Genes implicated in cancer should have **high mutation rate**

In comparison to normal, tumor cells should have higher occurrence of functional mutations in genes involved in the insurgence and progression of the disease.

Problem:

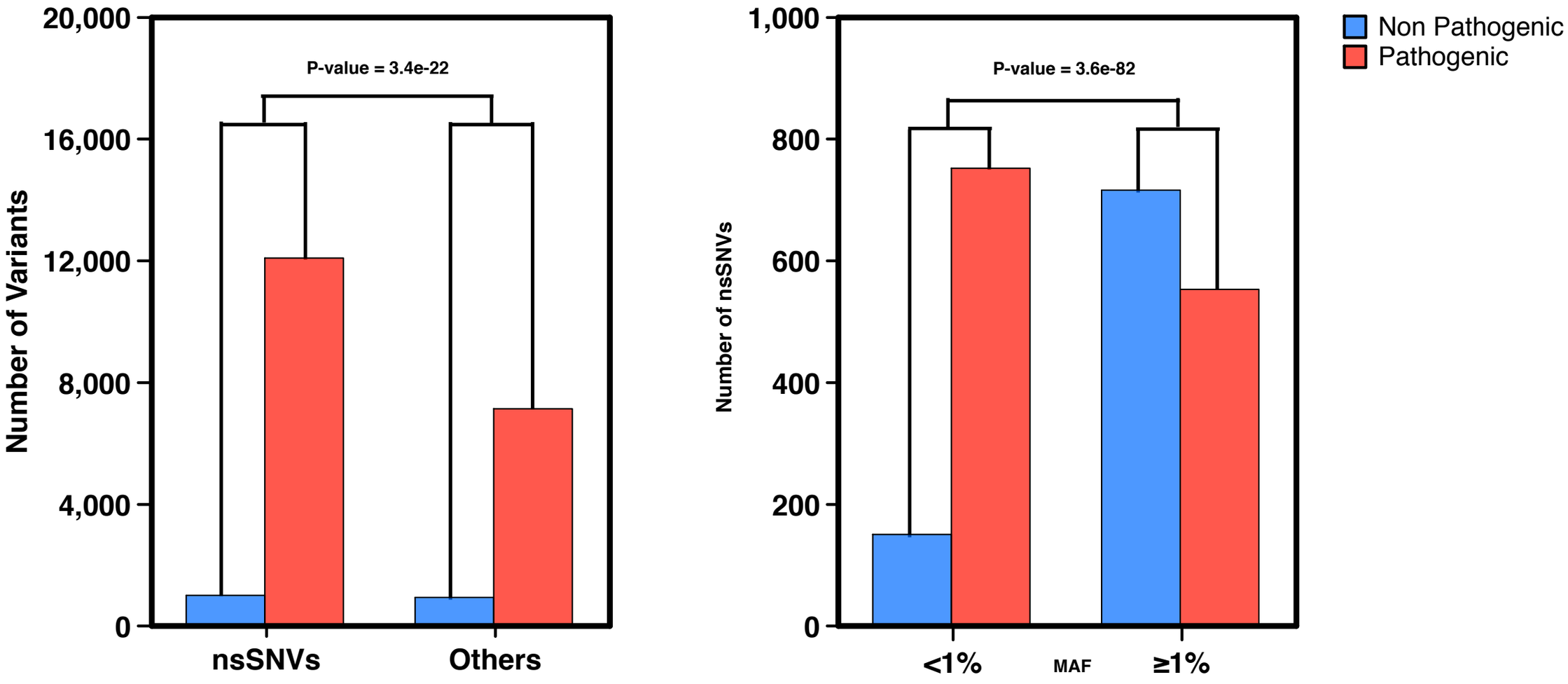
How can we select mutations with functional impact?

Average number of variants	~3,000,000
Average exome variants	~23,000
Average nonsynonymous single nucleotide variants	~10,000
Average rare (MAF≤0.5%) nonsynonymous single nucleotide variants	~300

The 1000 Genomes Project (2010). Nature. 467; 1062-1073.

Variants and MAF

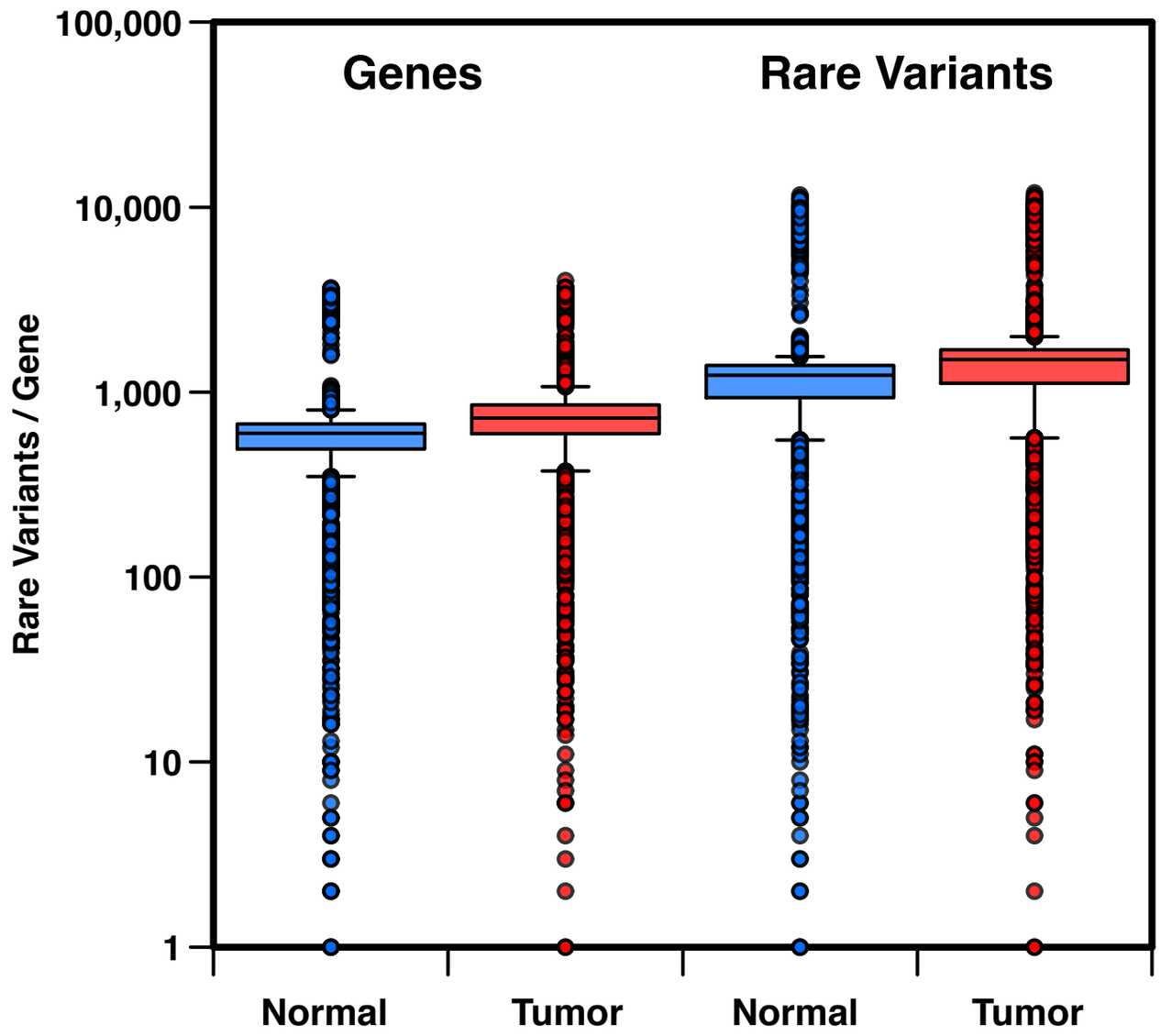
Rare variants are more likely to be associated to disease than high frequency variants



Tian R, Basu M, Capriotti E (2014). Bioinformatics. 30: i572-i578

Rate Variants and Genes

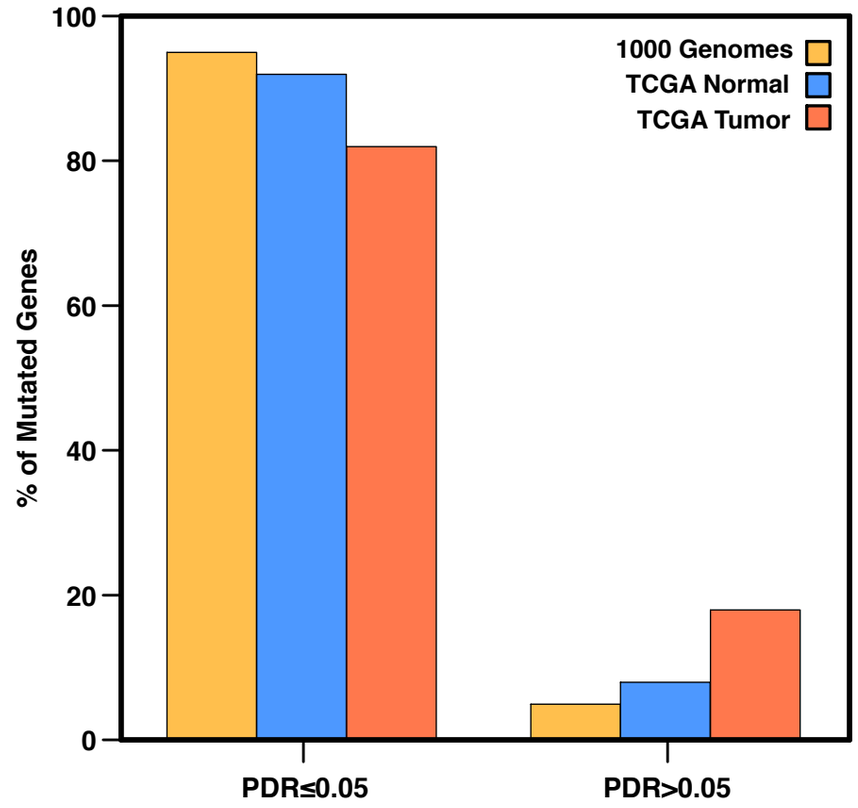
On average tumor samples (COAD) have ~150 more rare missense variants and mutated genes



Mutation rates

The analysis of **1000 Genomes, The Cancer Genome Atlas (TCGA)** normal and tumor samples shows an **increasing number of genes with rare nonsynonymous SNVs**.

Cohort	%Genes PDR≤0.05	%Genes PDR>0.05
1000 Genomes	95%	5%
TCGA Normal	92%	8%
TCGA Tumor	82%	18%



Tumor = Colon Adenocarcinoma

PDR = Gene Putative Defective Rate

Fraction of samples in which a gene has ≥1 nonsynonymous variant with MAF≤0.5%

ContrastRank score

The gene prioritization **score** is calculated using a **binomial distribution.**

$$b_g(k, N, \pi) = \frac{N!}{k!(N-k)!} \pi_g^k (1-\pi_g)^{N-k}$$

k: number of time a gene is observed to be a PIG across all the samples
N: total number of samples
π_g: probability of success

$$P_g(x \geq k, N, \pi) = 1 - \sum_{i=0}^{k-1} b_g(i, N, \pi) = 1 - \sum_{i=0}^{k-1} \frac{N!}{i!(N-i)!} \pi_g^i (1-\pi_g)^{N-i}$$

with k>0

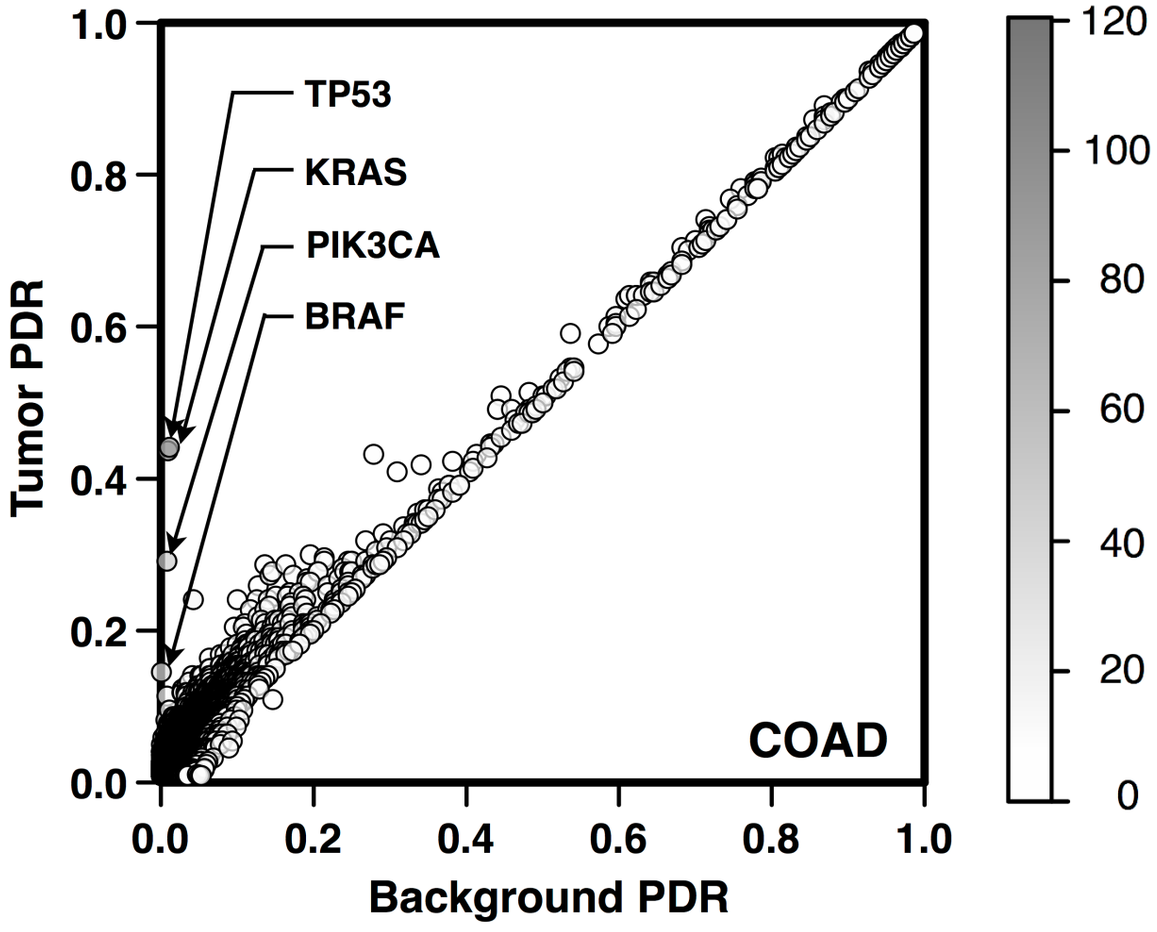
$$s_g = -\log_{10} P_g$$

Cancer Genome Analysis

New method for cancer gene prioritization based on the comparison of the mutation rates in tumor samples vs normal and 1000 Genomes samples.

Gene	PDR[T]	PDR[B]	Score
KRAS	0.436	0.009	72.6
TP53	0.441	0.011	63.7
PIK3CA	0.291	0.007	39.4
BRAF	0.146	0.001	29.9

Colon Adenocarcinoma
PDR[T] = Putative Defective Rate Tumor
PDR[B] = Putative Defective Rate Background
Background = Max (Normal and 1000 Genomes)



Whole Exome Score

The prioritization score can be used to **score the whole exome**

The score associated to the whole sample is the average score over the total number of putative impaired genes (M) in the sample

$$S = \frac{1}{M}\sum_{i=1}^{M} s_{g_i} = \frac{1}{M}\sum_{i=1}^{M} -\log_{10} P_{g_i}$$

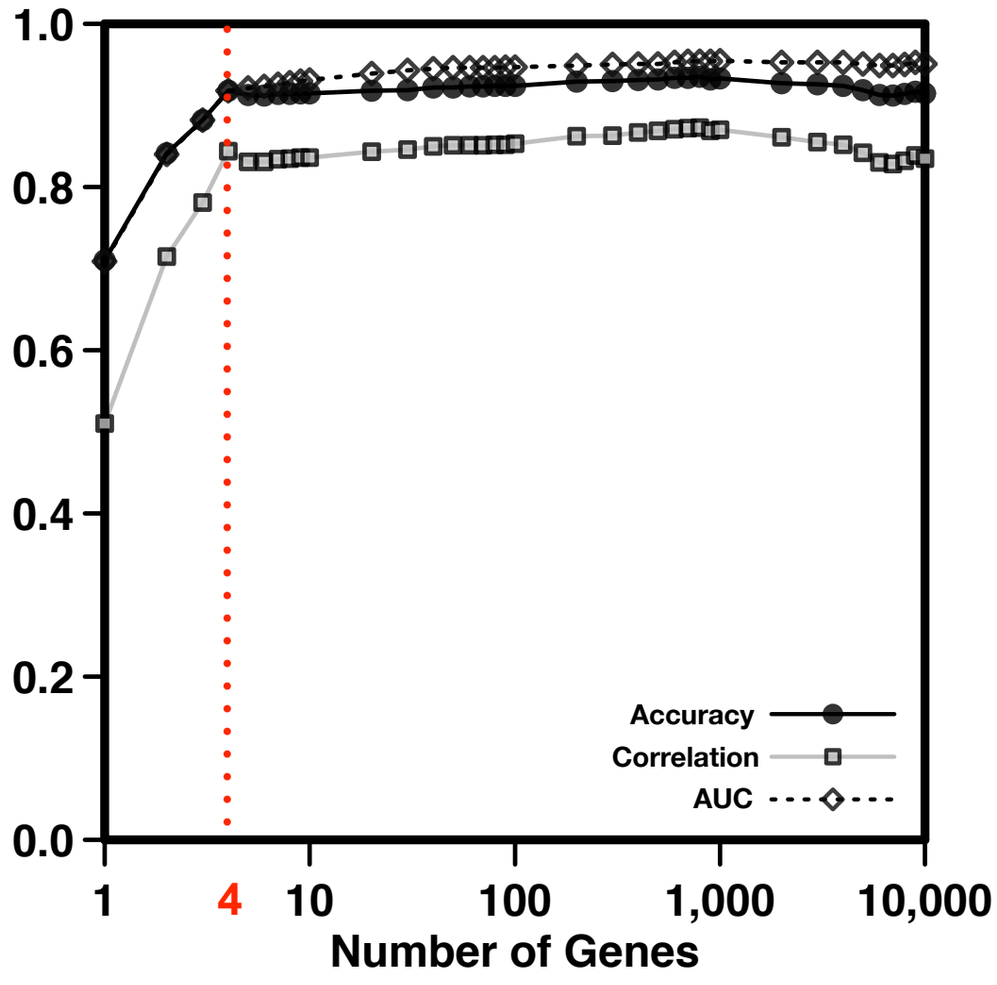
M: Total number of Putative Impaired Genes (PIGs) in the sample.

Scoring the risk of tumor

New method for discriminating normal from tumor samples scoring the genome with the prioritization approach based on the **background PDR from normal and 1000 Genomes samples.**

#Genes	Accuracy	Correlation	AUC
4	0.92	0.84	0.92

Colon Adenocarcinoma
Tumor vs Normal samples
First 4 Genes: KRAS, TP53, PIK3CA, BRAF



Discriminating tumor types

With three cancer types we tried to **discriminate tumor type A** from a **mixture of the remaining two (B +C).**

The new prioritization score (s_g) is the differences between the score of the gene calculated on both subsets.

$$s_g = s_g^A - s_g^{BC}$$

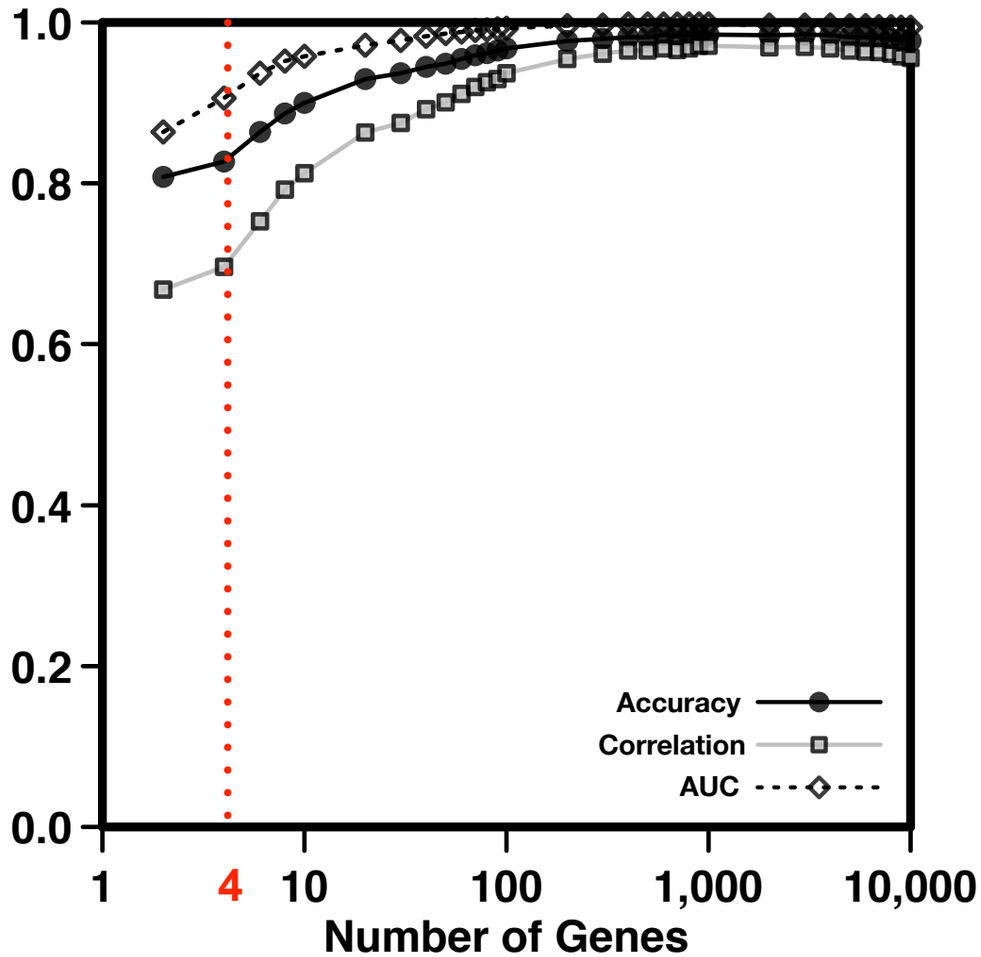
In this test we use the top ranking positively scored gene and lowest ranking negative scored genes to classify a specific cancer type.

Tumor Profiling

Profiling tumor mutations **comparing specific tumor samples against a mixture of other tumor types**.

#Genes	Accuracy	Correlation	AUC
4	0.83	0.70	0.91

Colon vs Lung and Prostate Adenocarcinomas
2 High Positive Genes: KRAS, TP53
2 High Negative Genes: GAGE2A, CT45A6

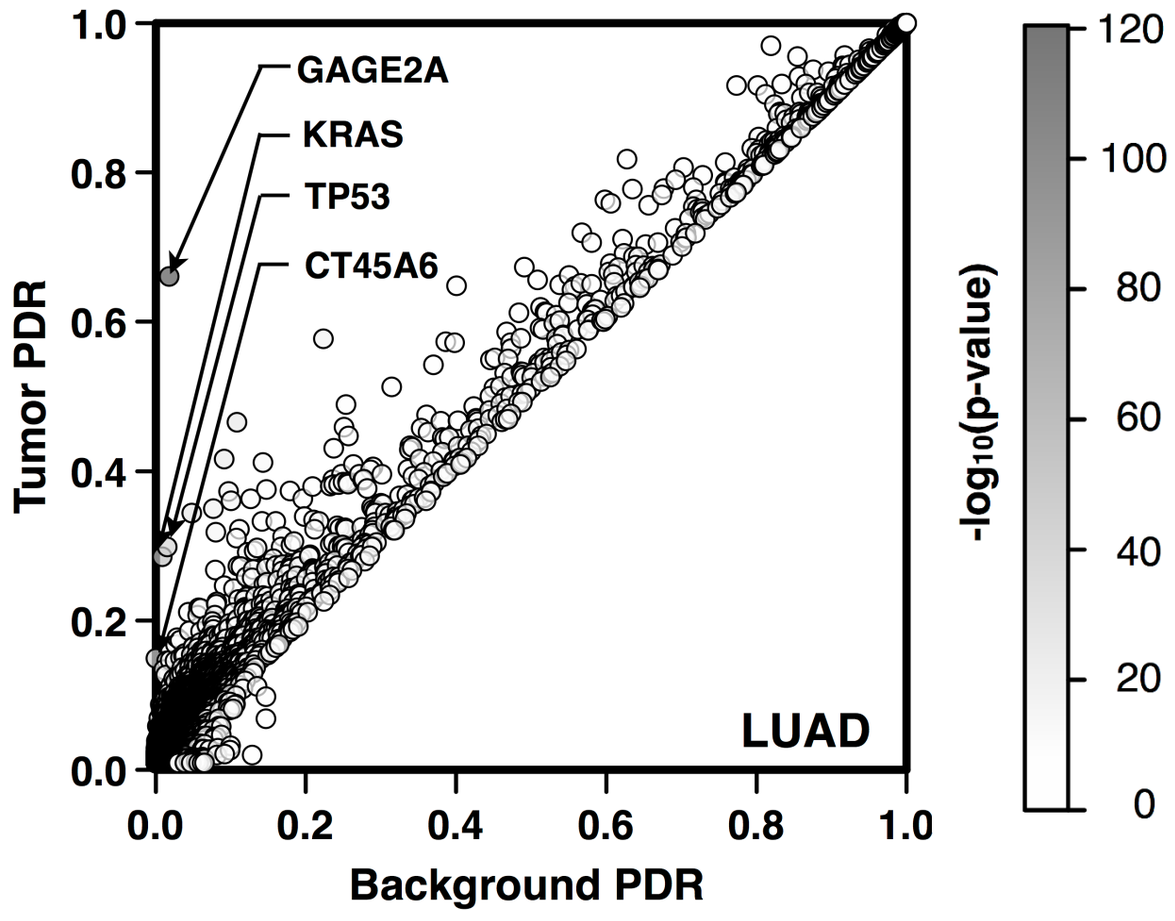


Another example

Prioritization of genes involved in lung adenocarcinoma

Gene	PDR[T]	PDR[B]	Score
GAGE2A	0.661	0.018	112.8
KRAS	0.286	0.008	46.3
CT45A6	0.0005	0.149	35.3
TP53	0.012	0.299	33.3

Lung Adenocarcinoma
PDR[T] = Putative Defective Rate Tumor
PDR[B] = Putative Defective Rate Background
Background = Max (Normal and 1000 Genomes)

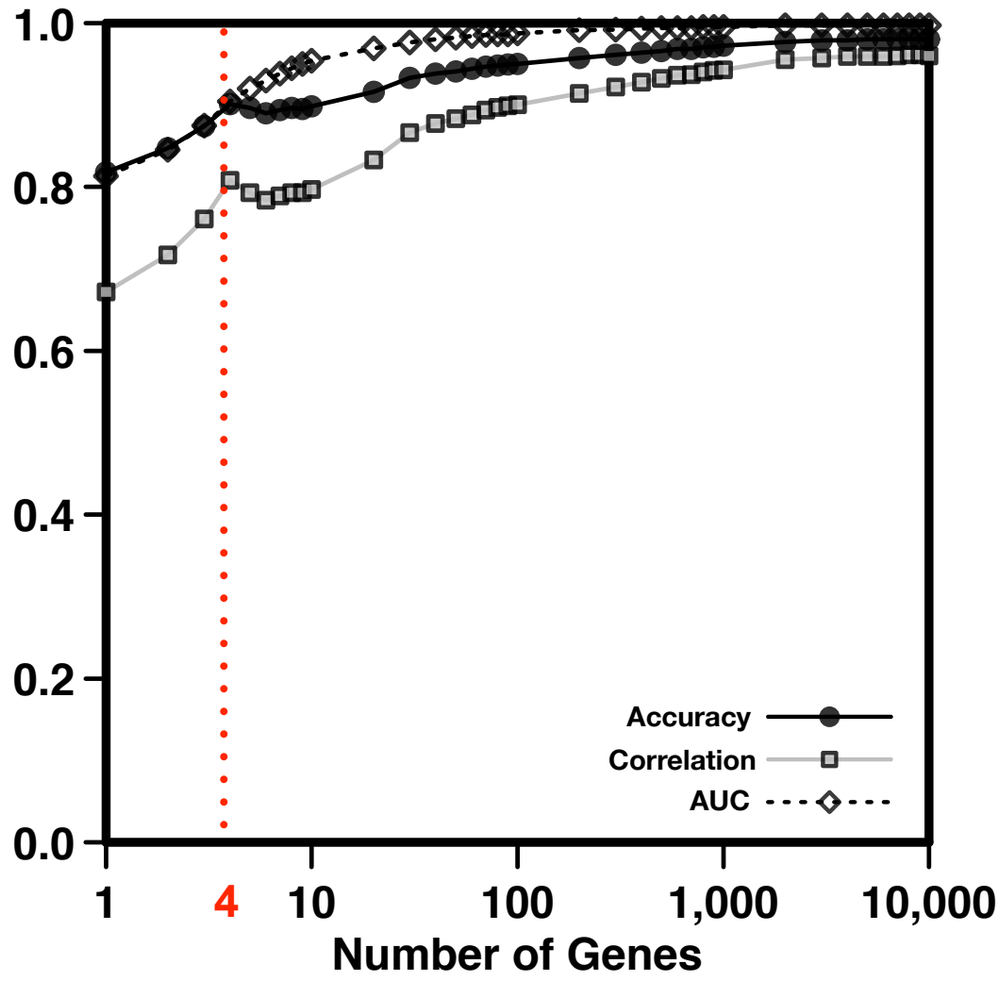


Tumor vs Normal

Scoring normal and tumor samples in lung adenocarcinoma.

#Genes	Accuracy	Correlation	AUC
4	0.90	0.81	0.90

Lung Adenocarcinoma
Tumor vs Normal samples
First 4 Genes: GEGA2, KRAS, CT45A6, TP53

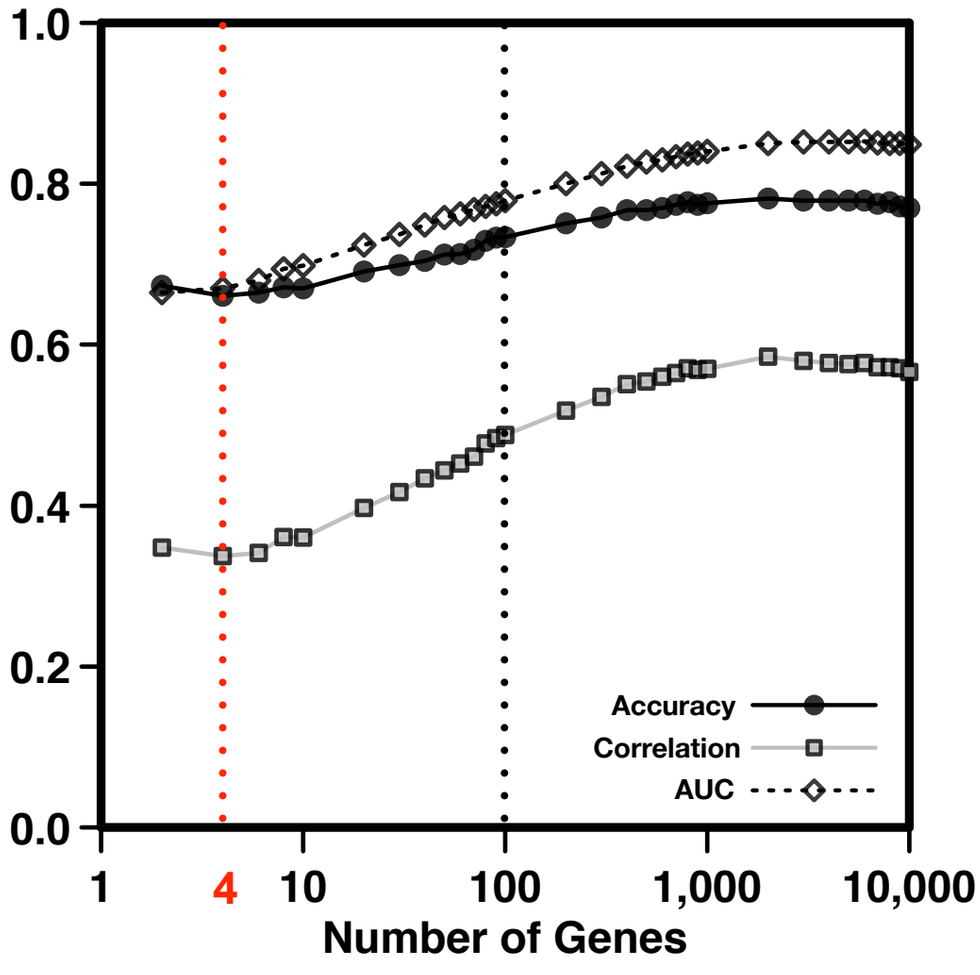


Lung adenocarcinoma

Comparing lung adenocarcinoma against a mixture of other tumor types.

#Genes	Accuracy	Correlation	AUC
4	0.66	0.34	0.67
100	0.73	0.49	0.78

Lung vs Colon and Prostate Adenocarcinomas
2 High Positive Genes: GAGE2A, CT45A6
2 High Negative Genes: SPOP, PIK3CA

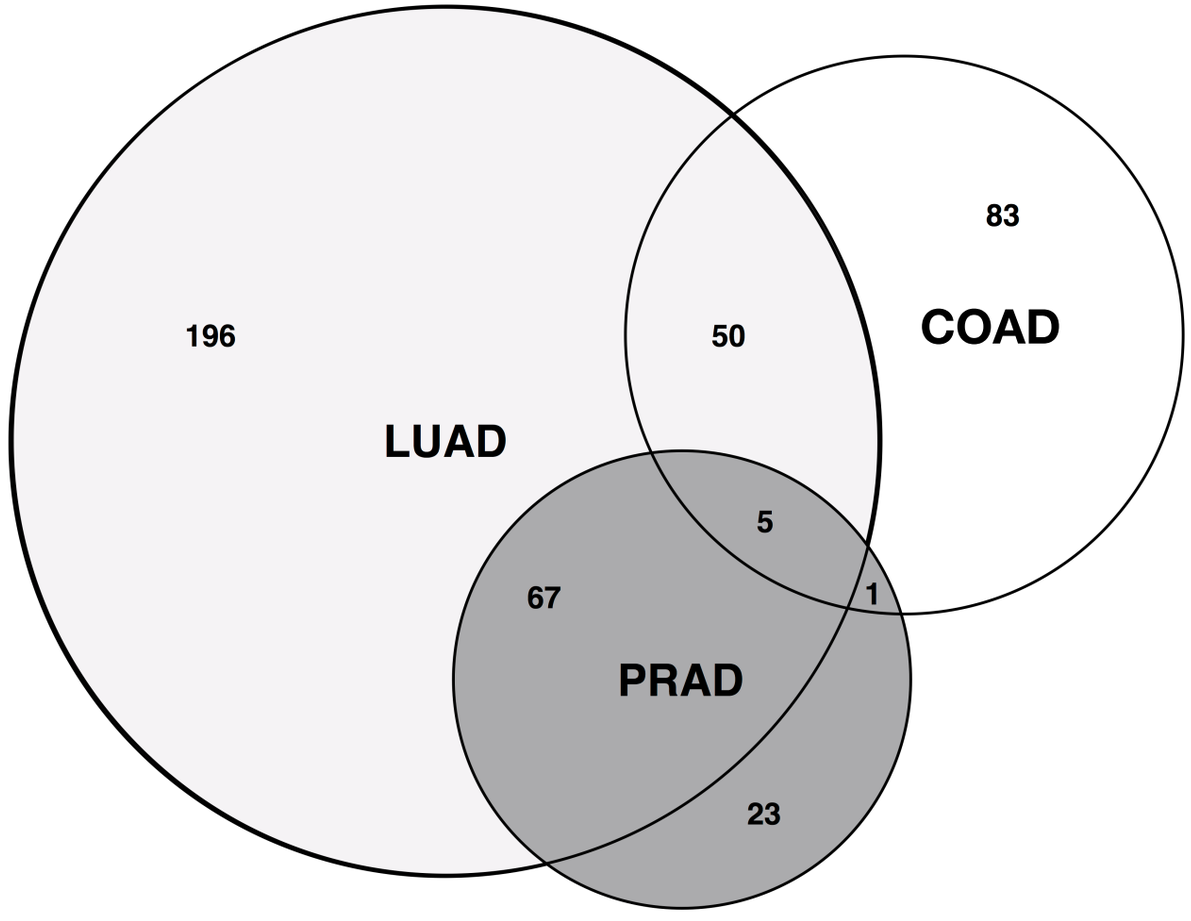


Comparing tumor types

Lung adenocarcinoma is more heterogenous than colon and prostate.
Significantly high scored genes for lung adenocarcinoma are also important for prostate and colon adenocarcinomas.

Lung (LUAD), Colon (COAD) and Prostate (PRAD) Adenocarcinomas
Respectively 318, 139 and 96 with score > 3

5 common genes are: TP53, BRAF, NBEA, AR, RNF145.



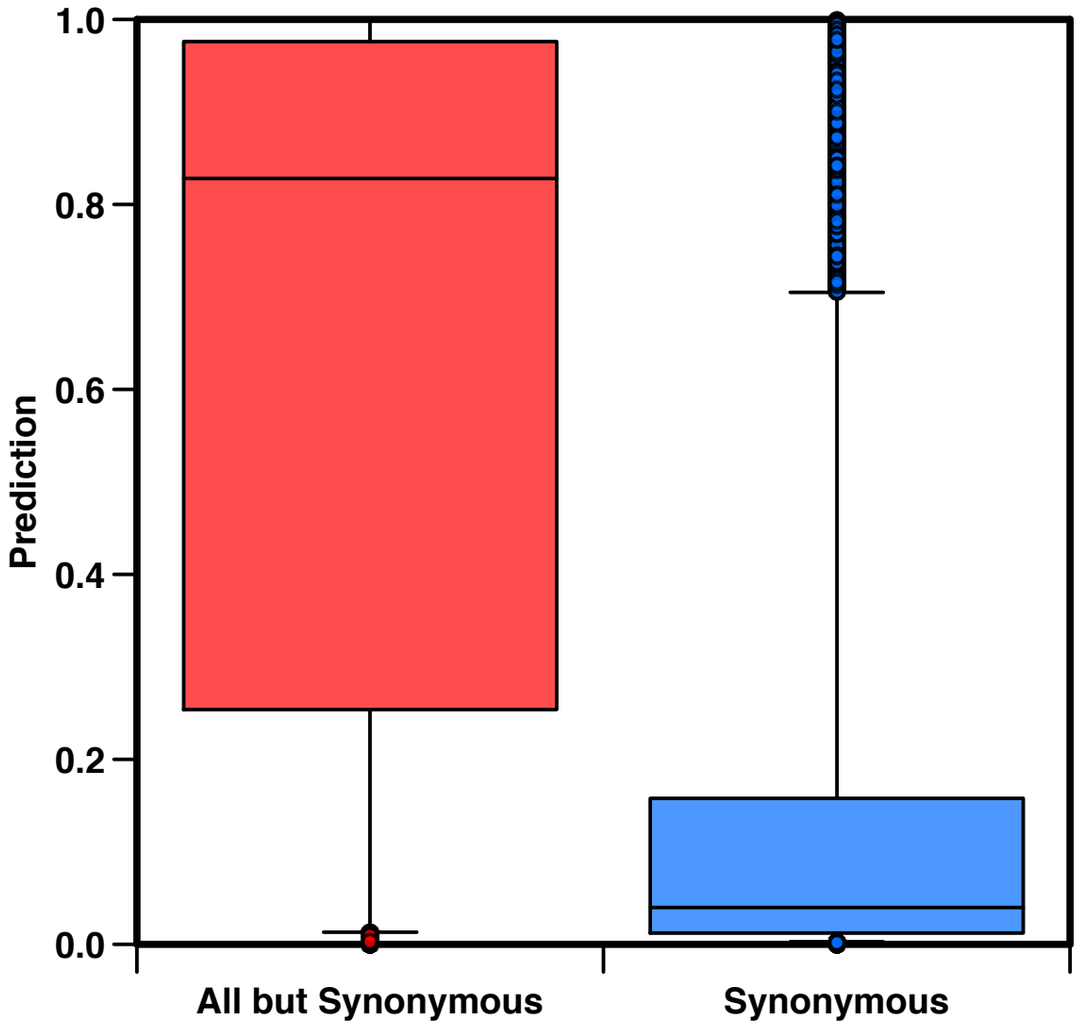
Improving Prioritization

Considering all but synonymous variants the method assigning the top ranking score to APC. When the raking procedure is performed the top genes are:

APC, TP53, KRAS, PIK3CA, BRAF.

Using PhD-SNP^{g} we predicted the impact of the variants

- 66% of the all but synonymous are predicted as Pathogenic
- 10% of the synonymous variants are predicted as Pathogenic



Exercise

Download the humsavar.txt file from UniProt

- Parse the file and extract variants annotated as Disease and Polymorphism

- Test the discrimination power different substitution matrices (BLOSUM, PAM, etc.)

- Calculate the performance of the method at the optimized classification threshold.