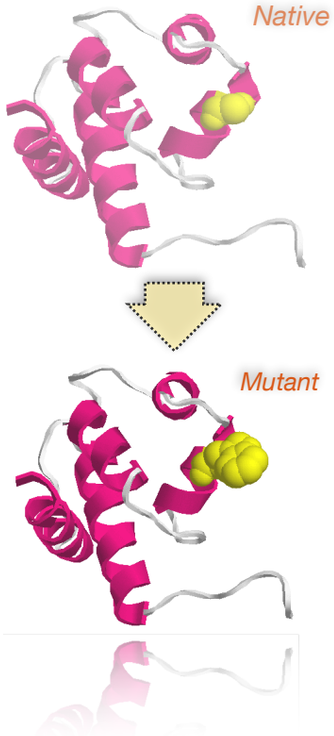
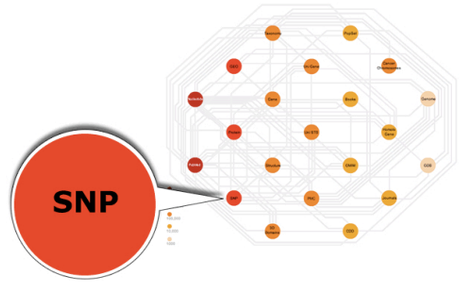
Computational methods for personalized medicine



SSS Carlo Urbani
University of Camerino, Camerino (MC)
October 24, 2018



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

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA
A.D. 1088

Presentation outline

- Human genome project:

  Sequencing, assembly, international consortiums

- Genetic variants:

  Variant databases and annotation

- Machine learning methods for variant interpretation:

  machine learning algorithms, prediction assessment

- Variations in cancer:

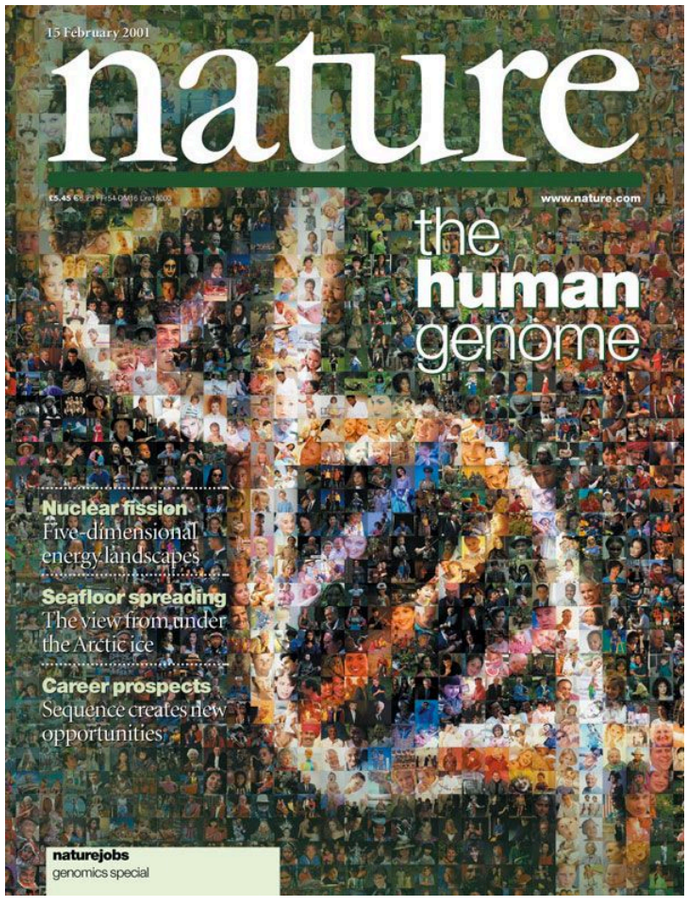
  Cancer data resources, gene prioritization

- Conclusions and future directions

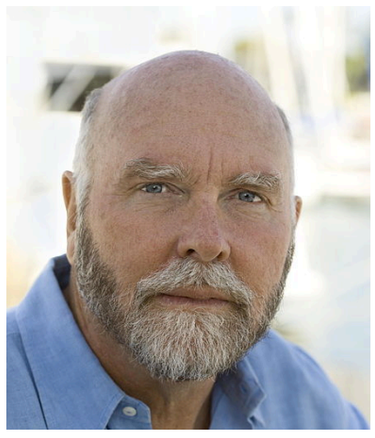
Human genome race

The first draft of the human genome was released in 2001.

The project was started 1990 and ended in 2003 and cost $3 billion



Int. HGS Consortium (2001).
Nature. 409: 860–921.

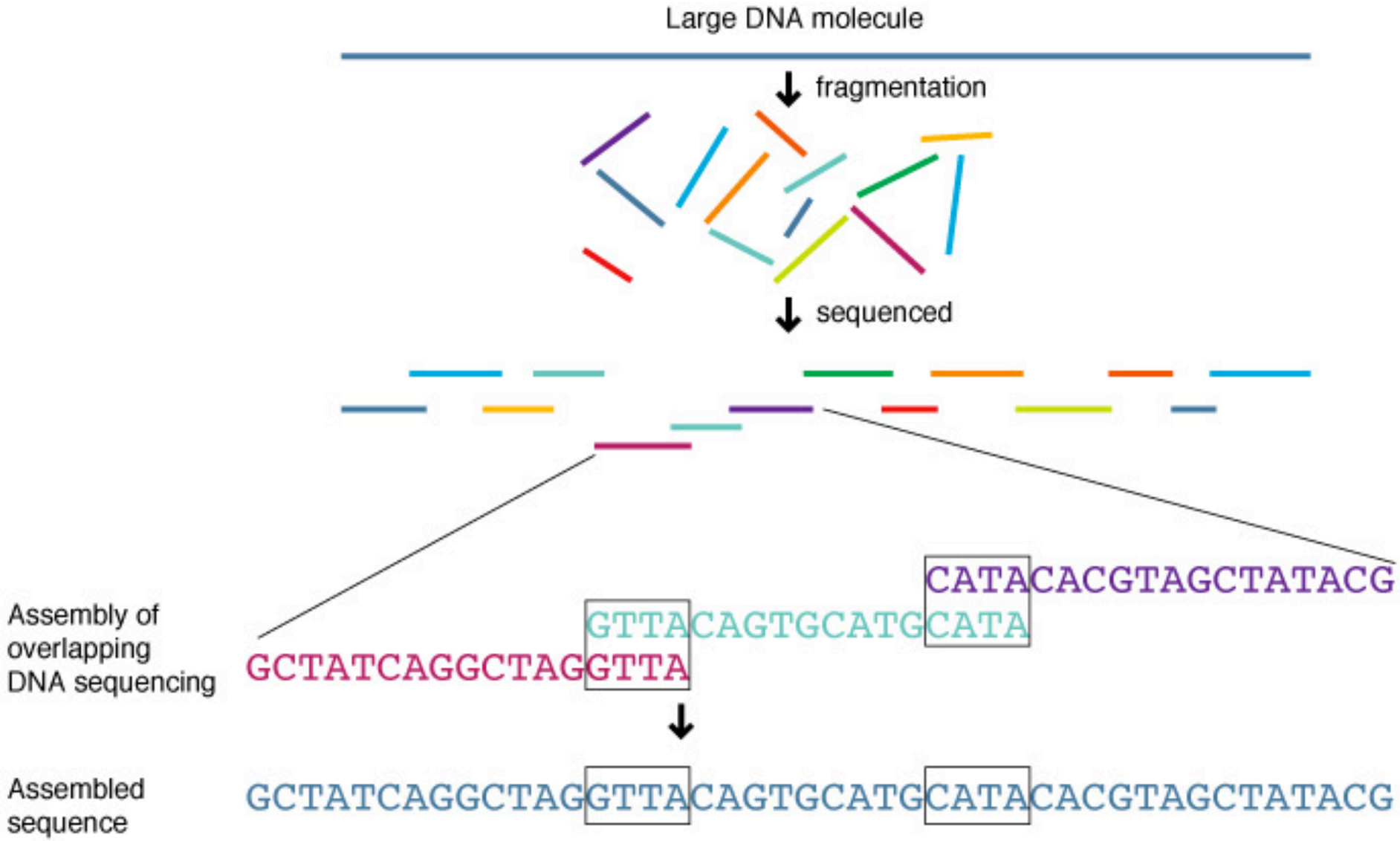


Cracking the Genome: Inside the Race to Unlock Human DNA.
by *Kevin Davies*

Venter et al (2001).
Science. 291: 1304-1351.

Sequencing method

Shotgun sequencing involves randomly breaking up DNA sequences into fragments (reads) and then reassembling the sequence by looking for regions of overlap.



Venter et al. (1996) Nature.381:364-366

The genome assembly

The assembly problem is to reconstruct as much of a genome as possible given a collection of reads or read pairs.

- the orientation of each read is not known

- one must allow a certain amount of error

- the entire genome is not covered by the read data

Different algorithms were developed for optimizing the genome assembly. An important contribution was given by Eugene Myers who significantly contributed to the determination of the Human, Mouse and Drosophila genomes

Myers (2016). Information Technology. 58: 126–132

Some numbers

- Size: ~3.23 Billon bases

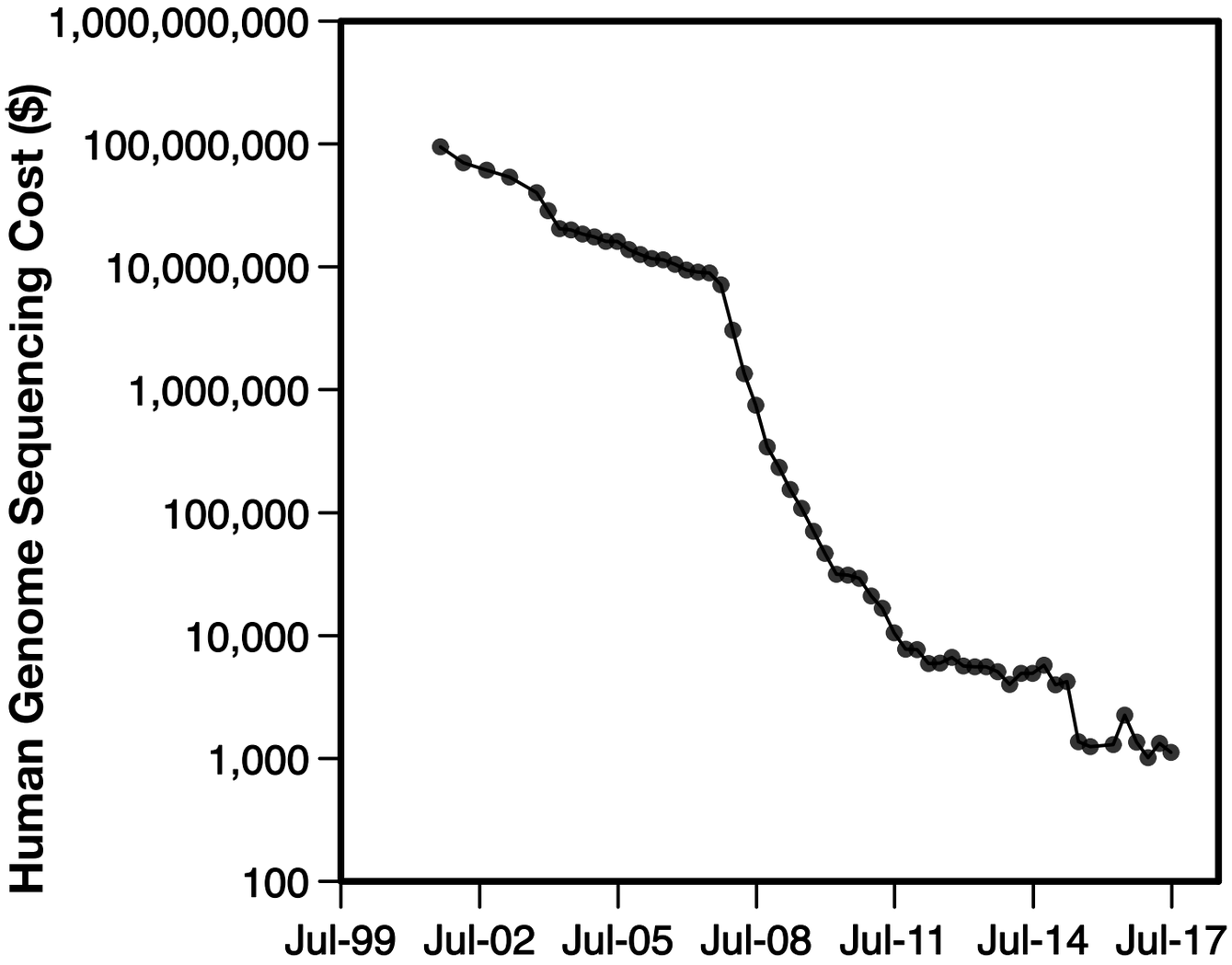
- 19,000-20,000 protein-coding genes

- Protein-coding sequences account ~1.5% of the genome, remaining part is associated with introns, non-coding, RNA molecules, regulatory DNA and sequences for which as yet no function has been determined.

- Differences among individuals on the order of ~0.1% while the differences with with chimpanzee is ~4%

Sequencing cost

During the last few years the sequencing cost of the human genome decreased significantly



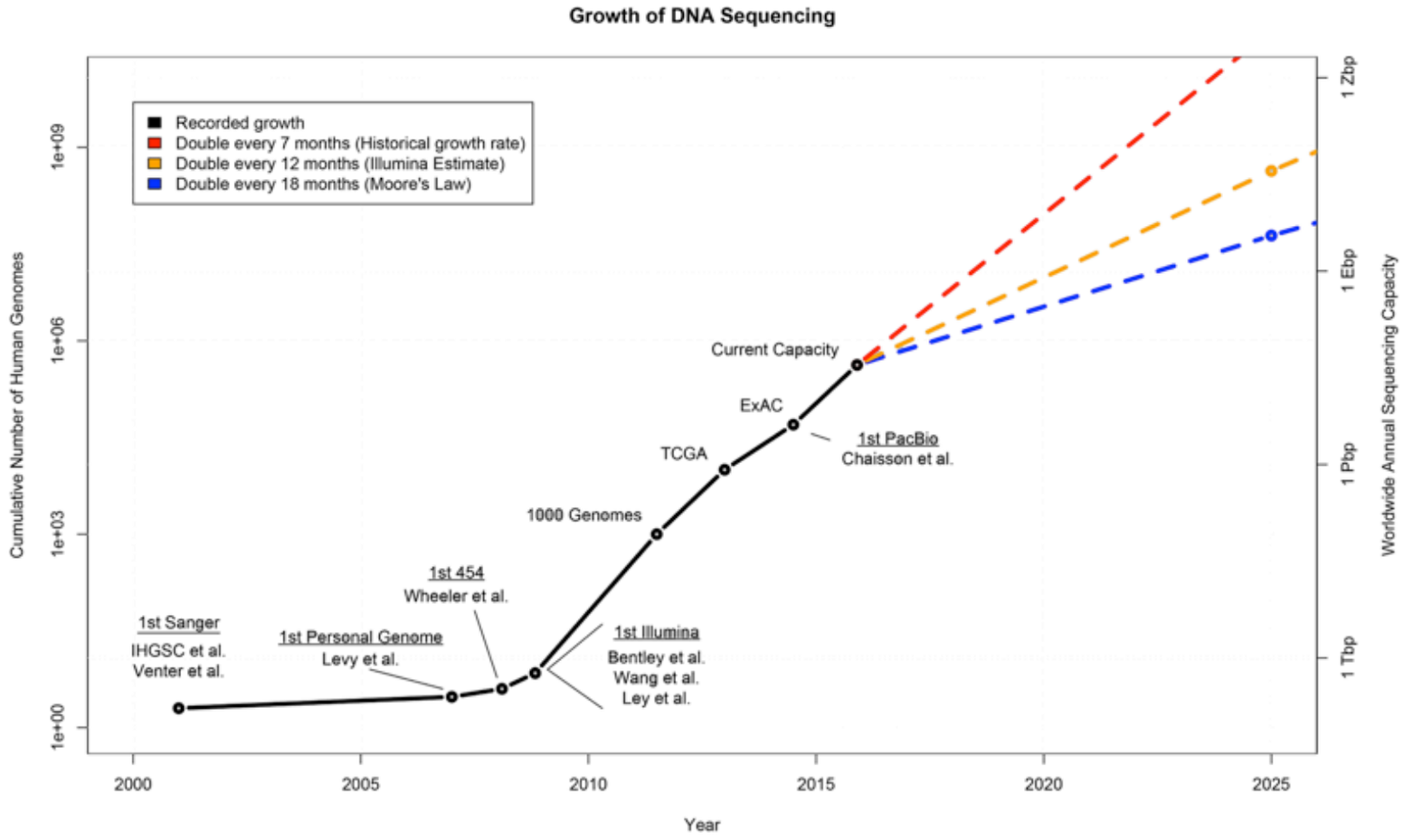
AB370A
>500 Kb/day
~16 years/Genome

HiSeq X
1.6-1.8 Tb/run
>18,000 Genomes/year
$10M

https://www.genome.gov/

Big Data in biomedicine

International consortiums generated a huge amount of sequencing data from human and genomes from many organisms



International consortiums

large-scale sequencing projects of the human genome

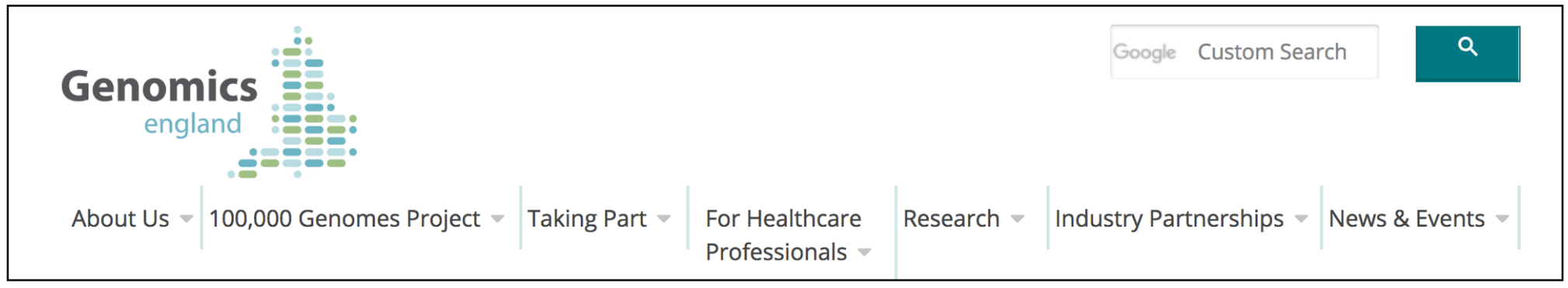
HapMap Project (2002-2009)

1000 Genomes Project (2008-2015)

http://www.internationalgenome.org/

100,000 Genomes Project (2012-)

https://www.genomicsengland.co.uk/



Single Nucleotide Variants

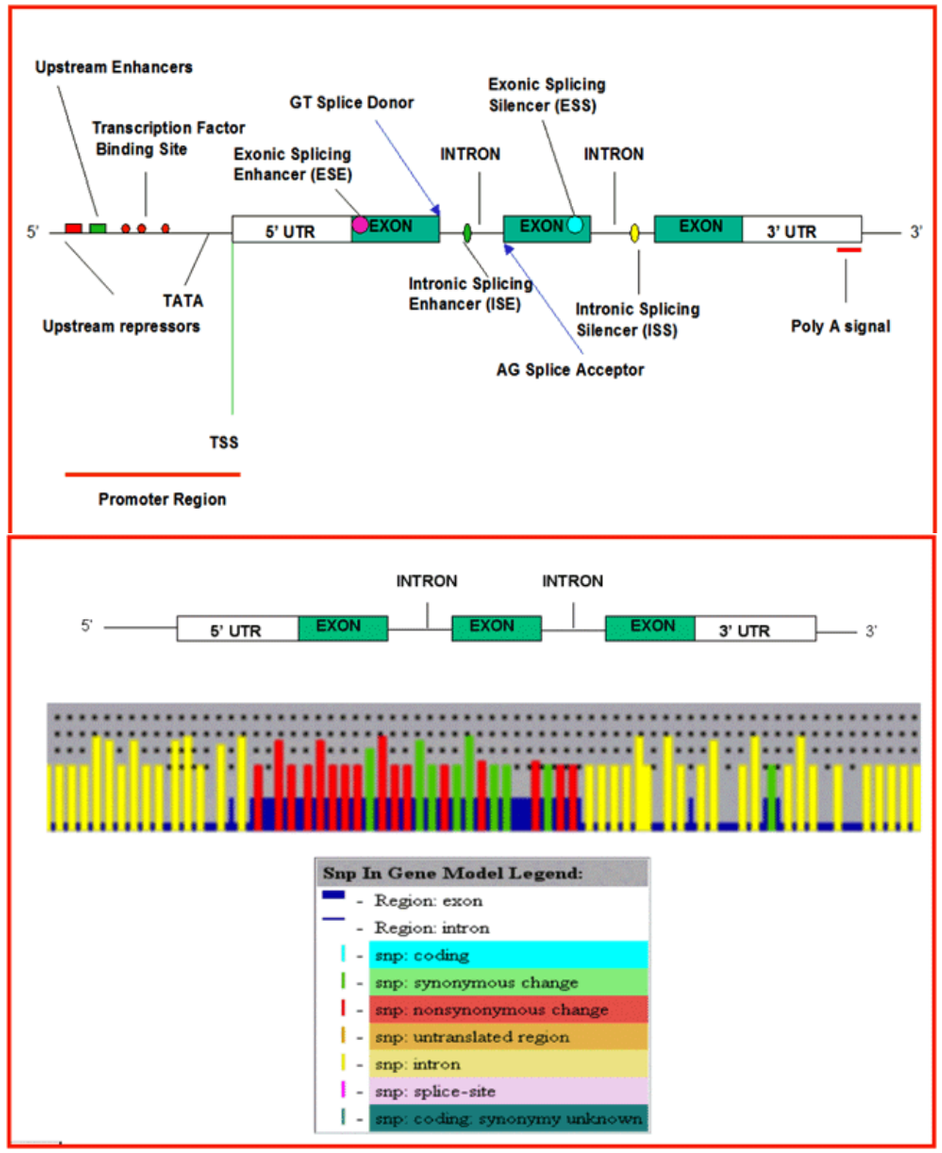
Single Nucleotide Variants (SNVs)
is a DNA sequence variation occurring when a single nucleotide A, T, C, or G in the genome differs between members of the species.
It is used to refer to Polymorphisms when the population frequency is ≥ 1%

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 versione 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

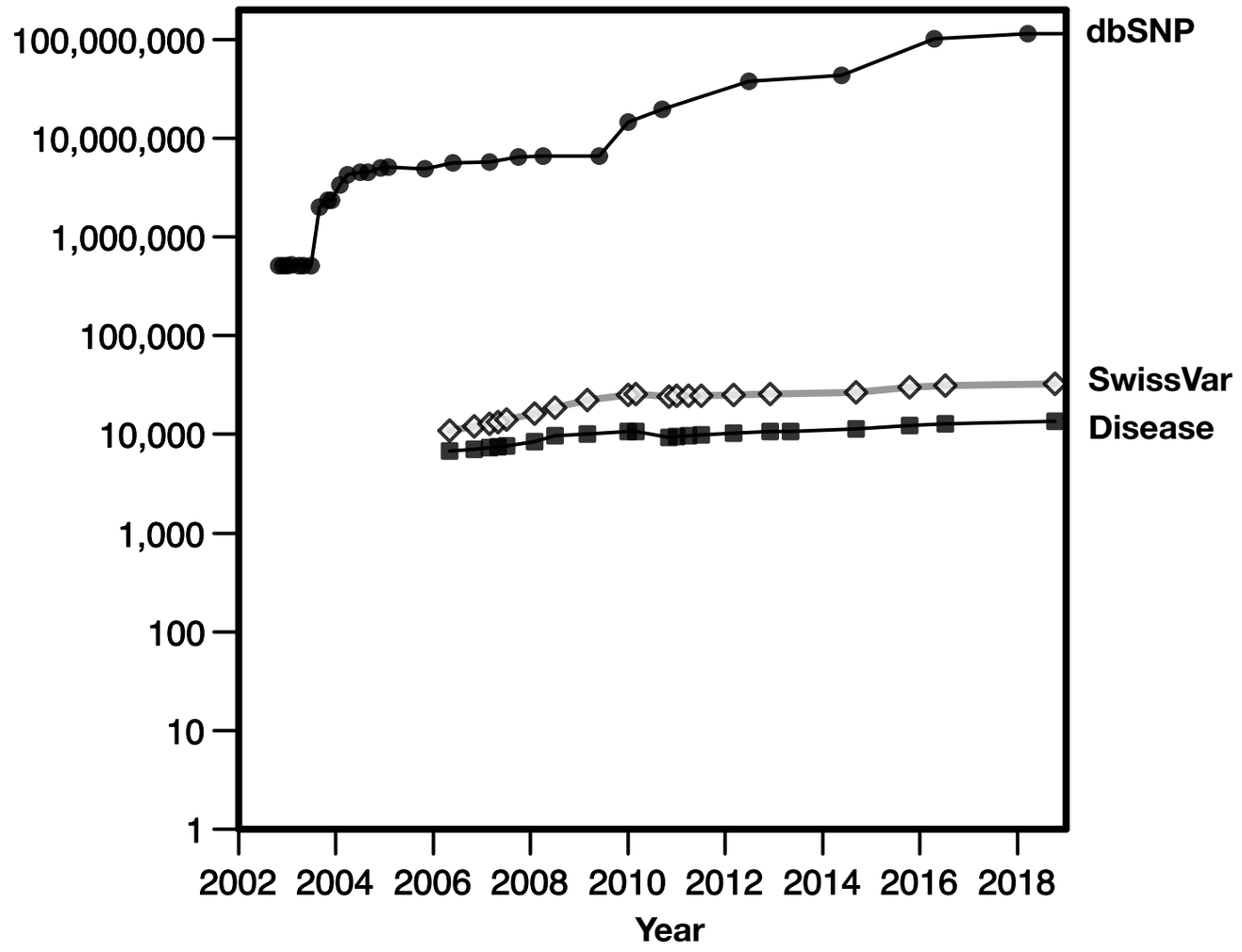
	Low coverage				Trios				
Statistic	CEU	YRI	CHB+JPT	Total	CEU	YRI	Total	Exon (total)	Union across projects
Samples	60	59	60	179	3	3	6	697	742
Total raw bases (Gb)	1,402	874	596	2,872	560	615	1,175	845	4,892
Total mapped bases (Gb)	817	596	468	1,881	369	342	711	56	2,648
Mean mapped depth (×)	4.62	3.42	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 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).

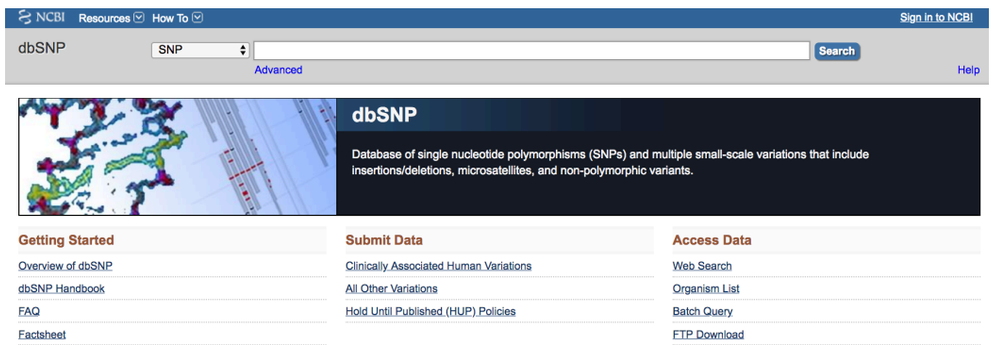


nonsynonymous SNVs

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

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/

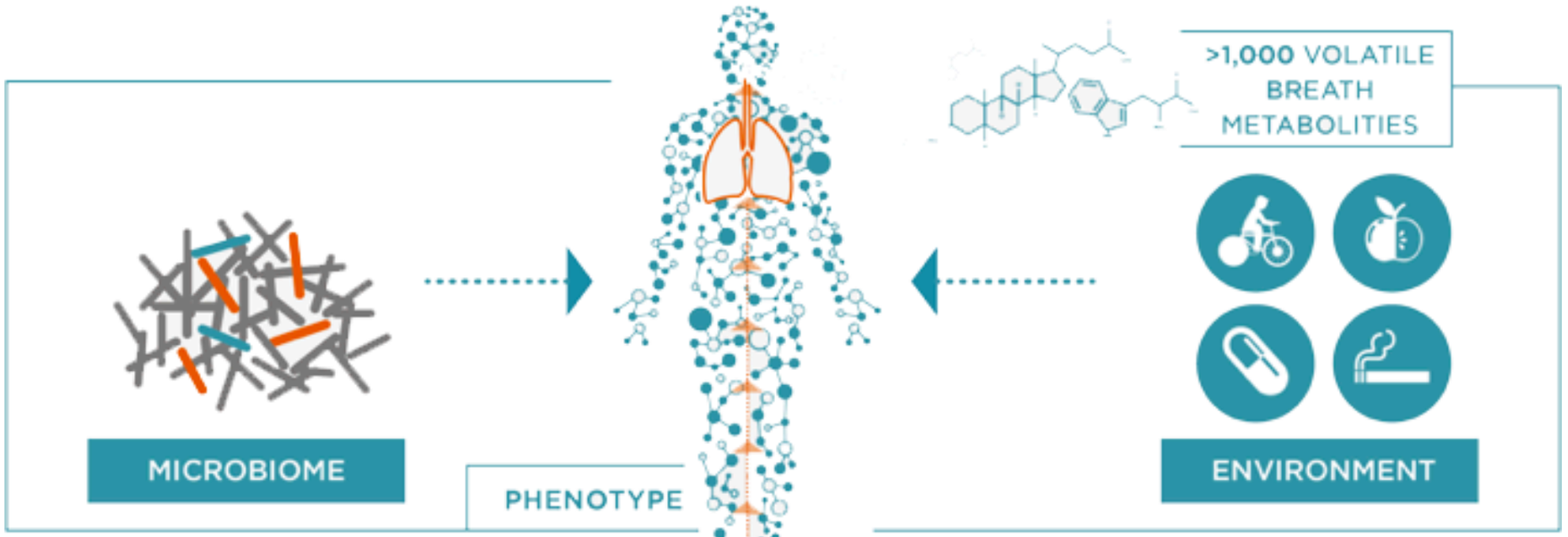
Single Amino acid Variants

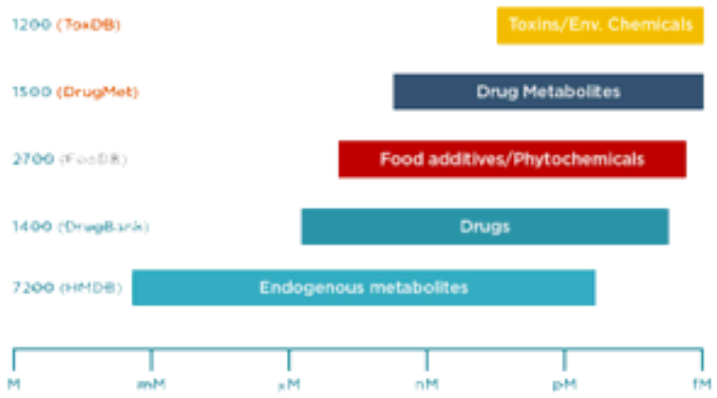
Homo sapiens	76,608
Disease	*29,529*
Polymorphisms	*39,779*

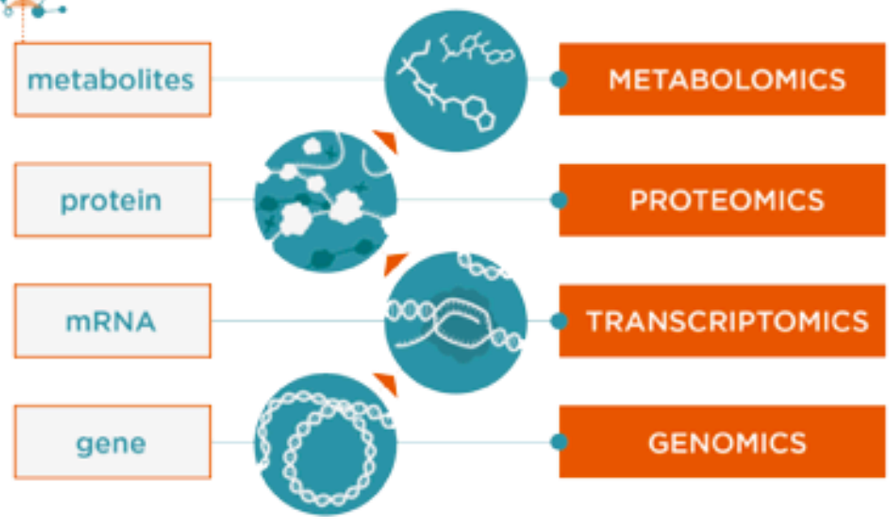
Oct 2018

Precision Medicine

The analysis of genomic data from healthy individuals and patients can be used to develop better diagnostic and personalized treatment strategies





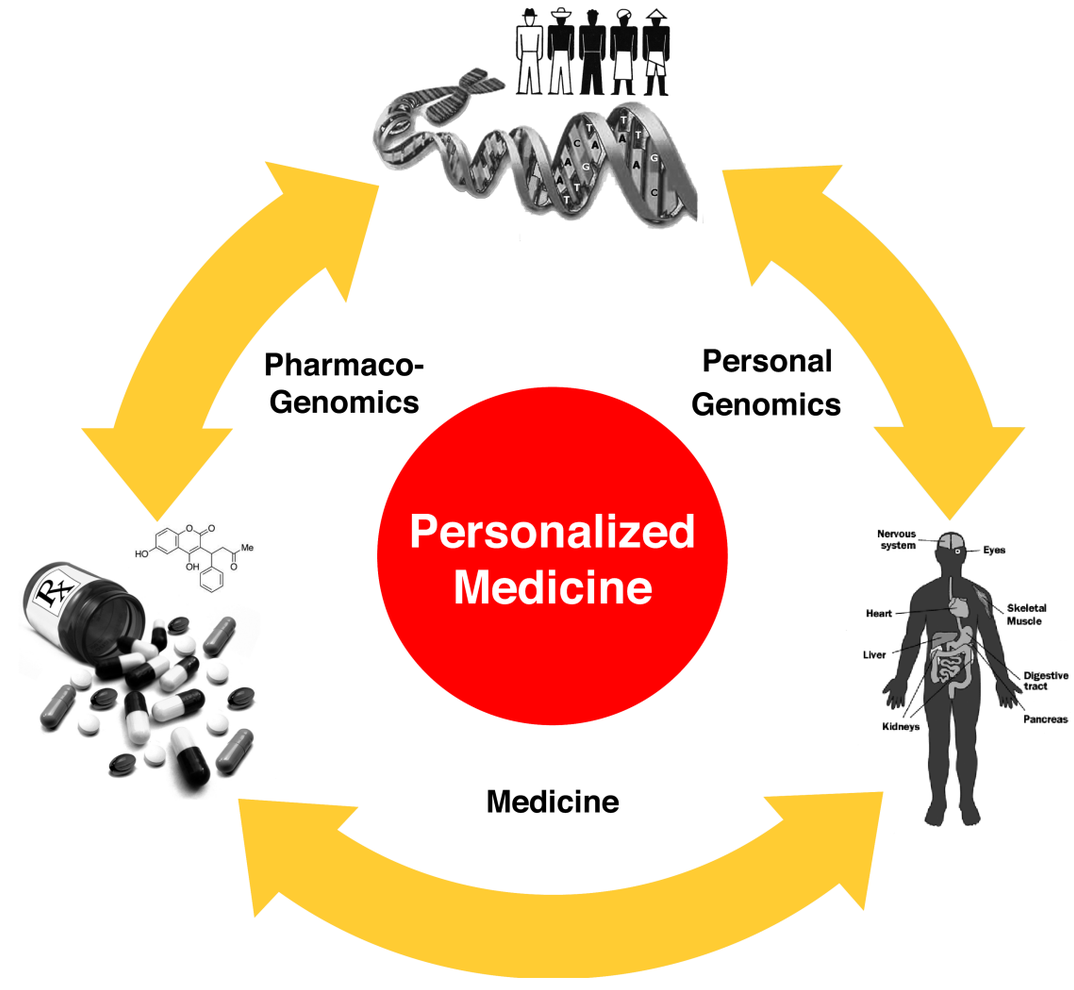


Personalized medicine

Direct to consumers company are performing genotype test on markers associated to genetic traits, and soon the full genome sequencing will cost ~$1,000.

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



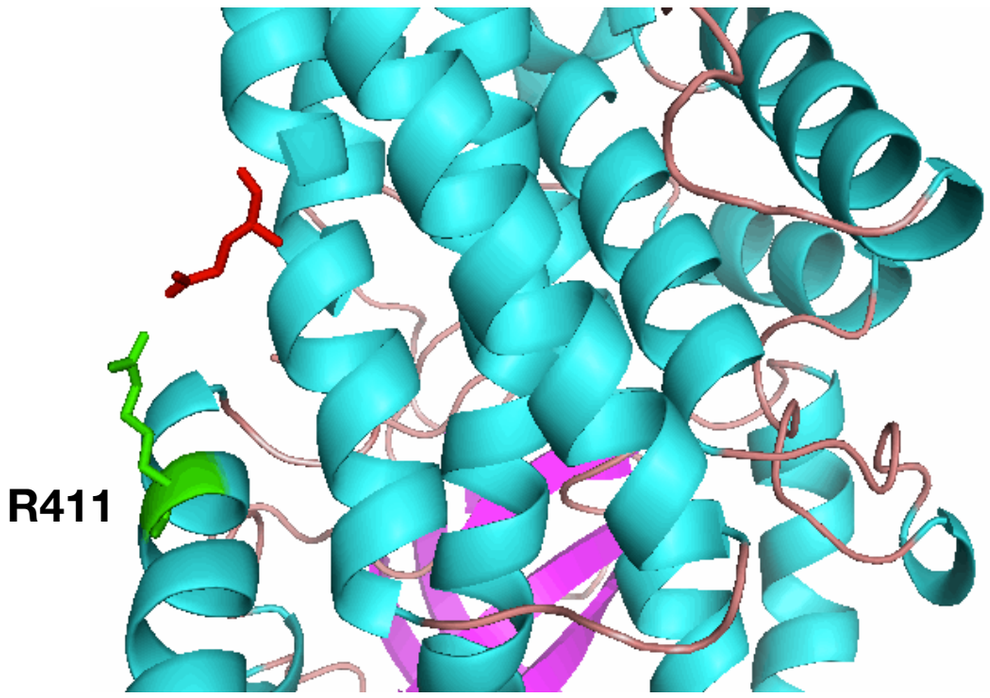
Fernald GH, et al (2011). Bioinformatics. 27: 1741-1748.

Variant Interpretation

Sequence, Structure & Function

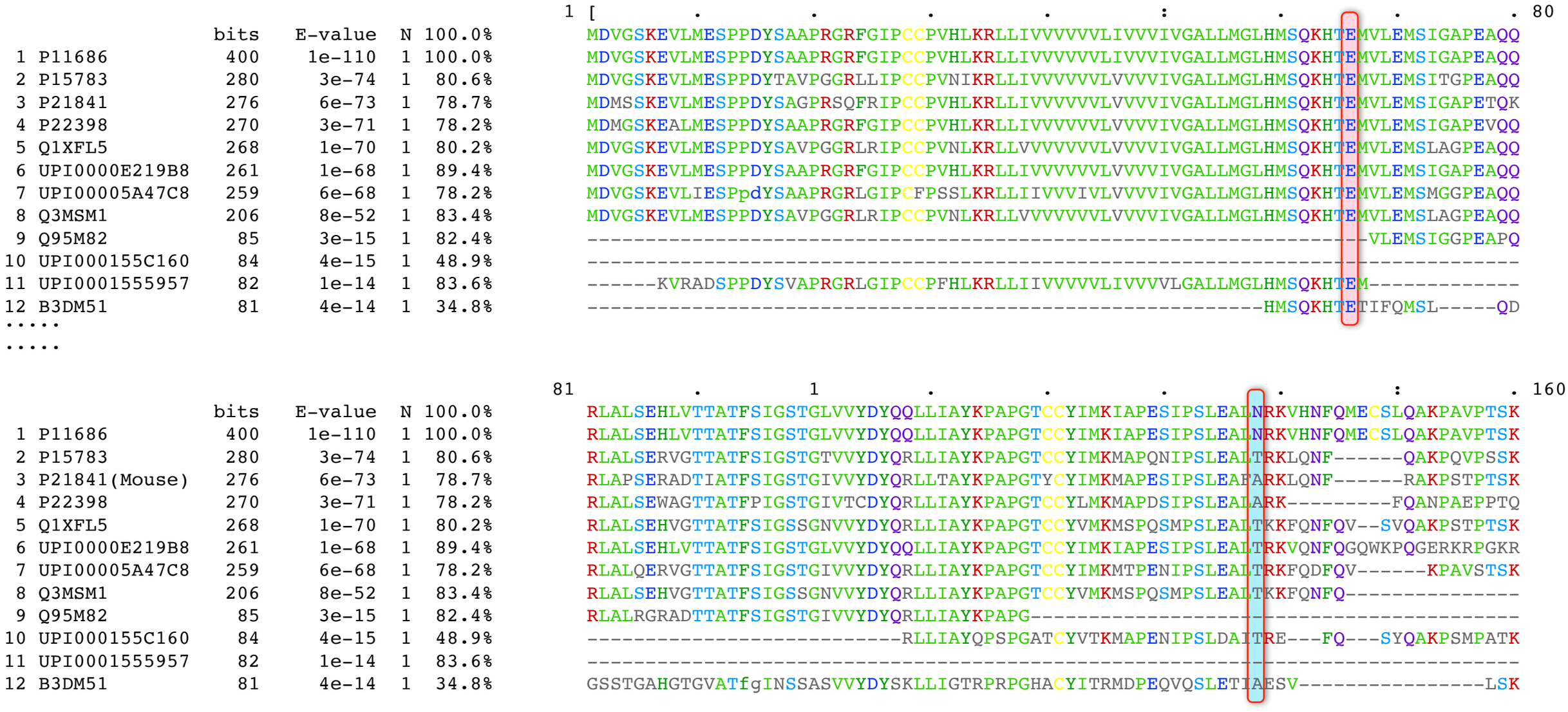
Genomic variants in sequence motifs could affect protein function.
Mutation S362A of P53 affect the interaction with hydrolase USP7 and the deubiquitination of the protein.

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?

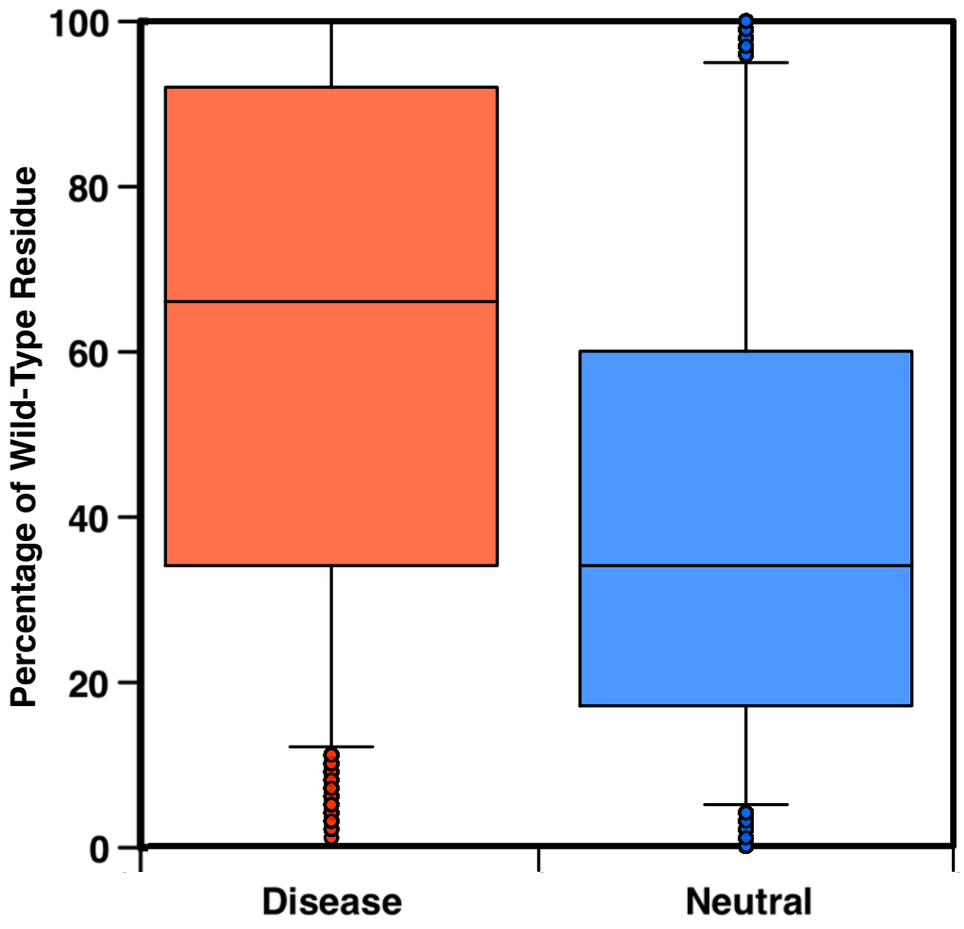
In positions 66 the Glutamic acid is highly conserved Asparagine in position 138 is mutated Threonine or Alanine



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.

Machine learning

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

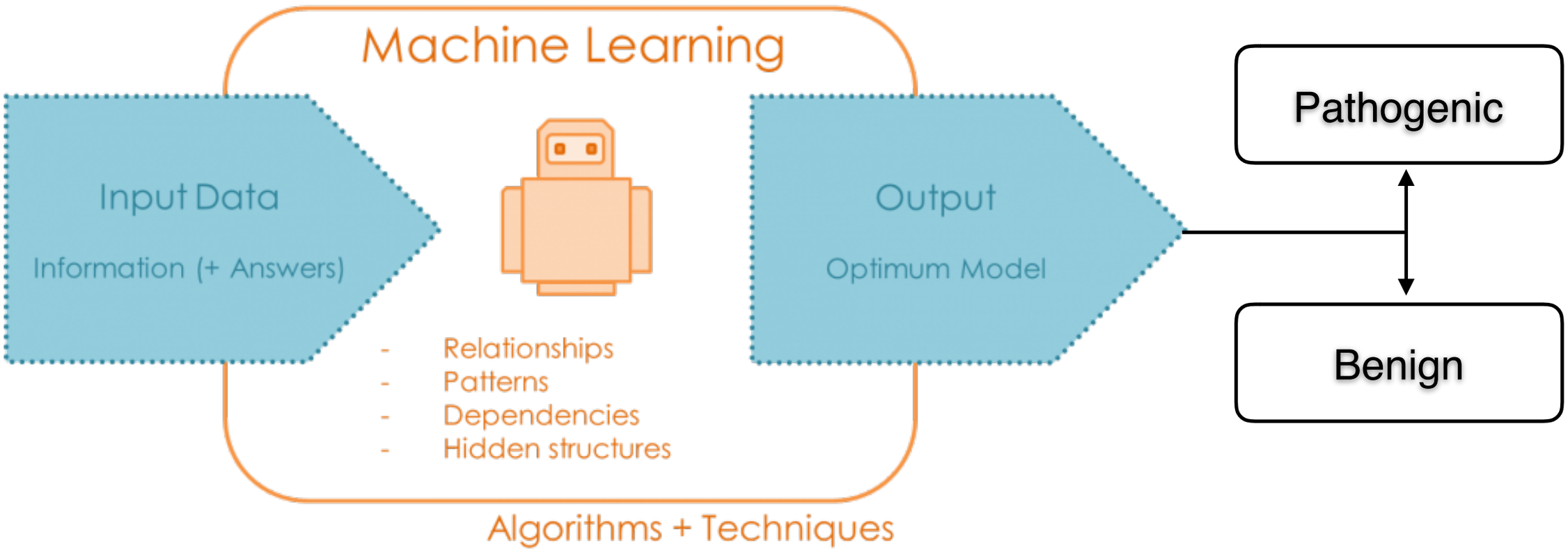
- Machine learning algorithms are suitable to address problems for which analytic solution does not exists and large amount of data are available.

- 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.

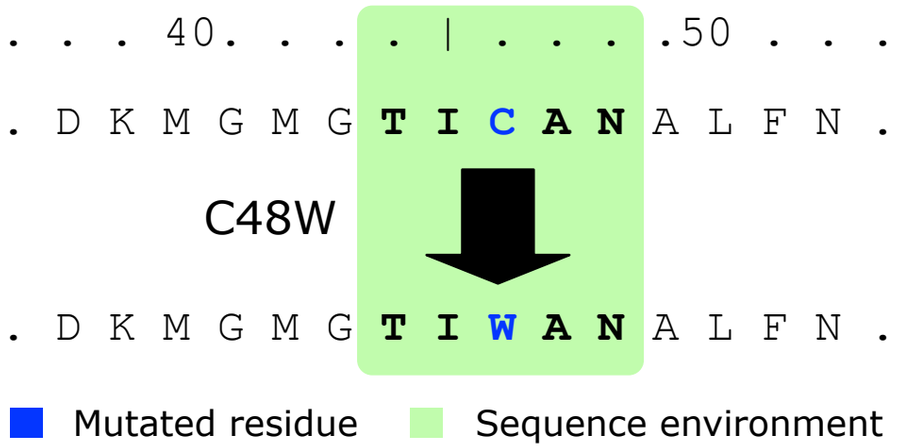
- Most popular machine learning approaches are in computational biology are Neural Networks, Support Vector Machines and Random Forest.

Variant interpretation

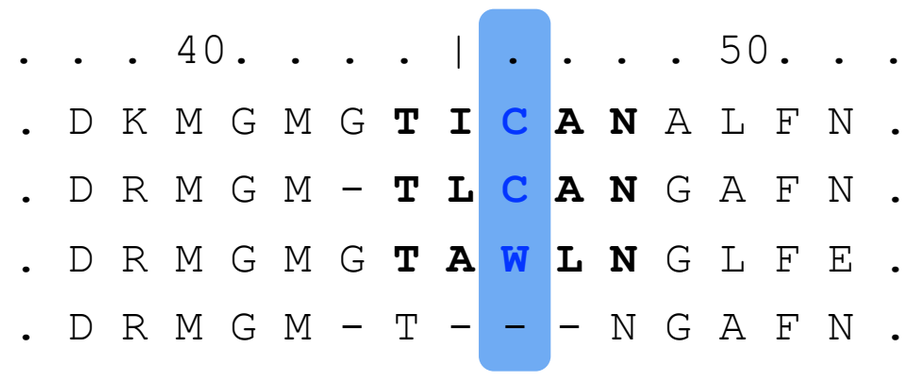
Usually based learning algorithm which takes in input features associated to the variants and returns a probability for the variant to be Pathogenic or Benign



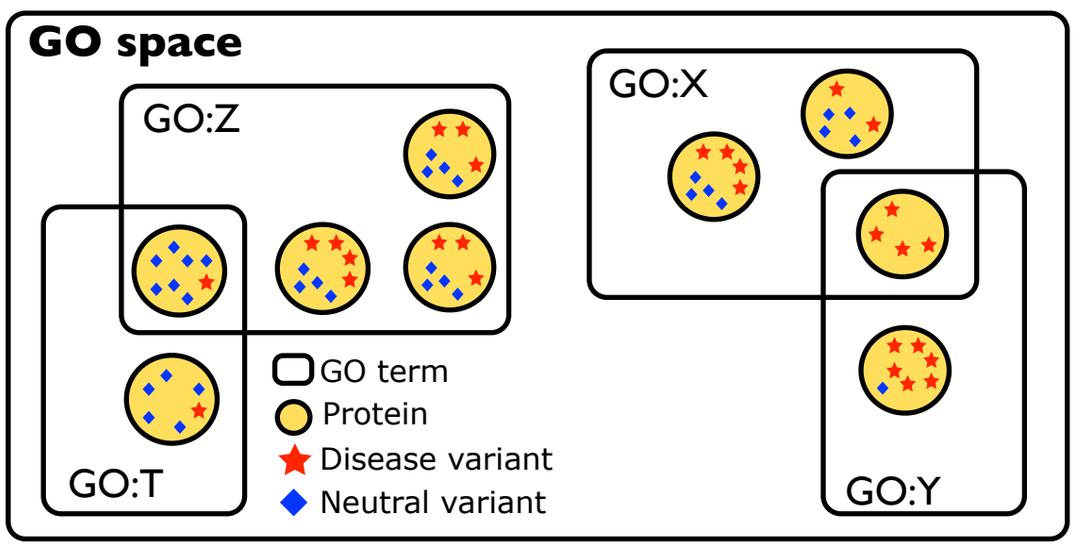
SNPs&GO input features



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



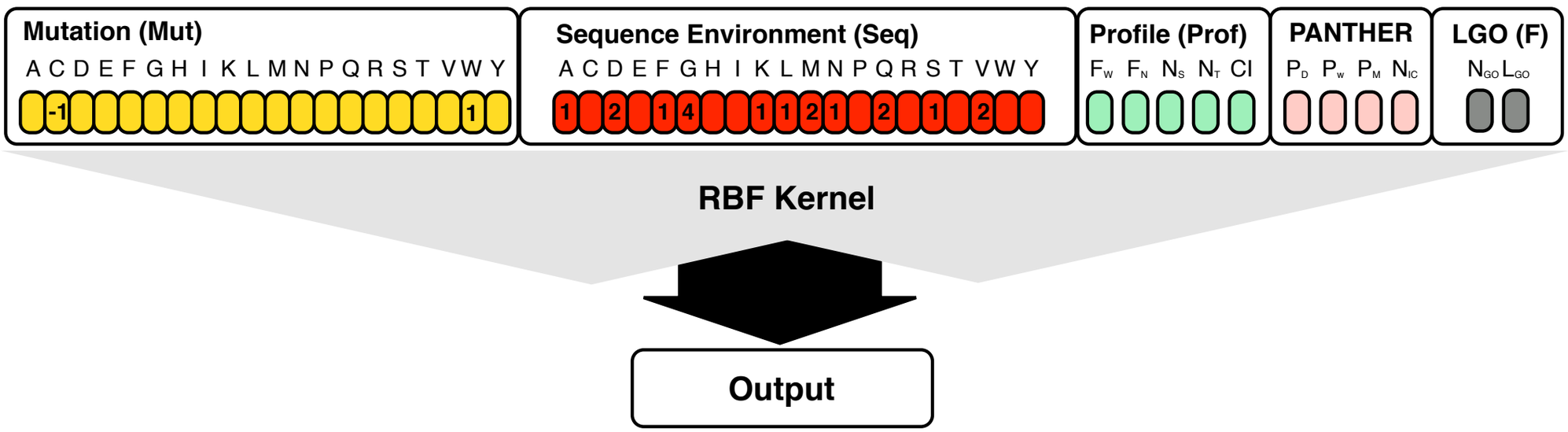
Protein sequence profile information derived from a multiple sequence alignment. It is encoded in a 5 elements vector corresponding to different features general and local features



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

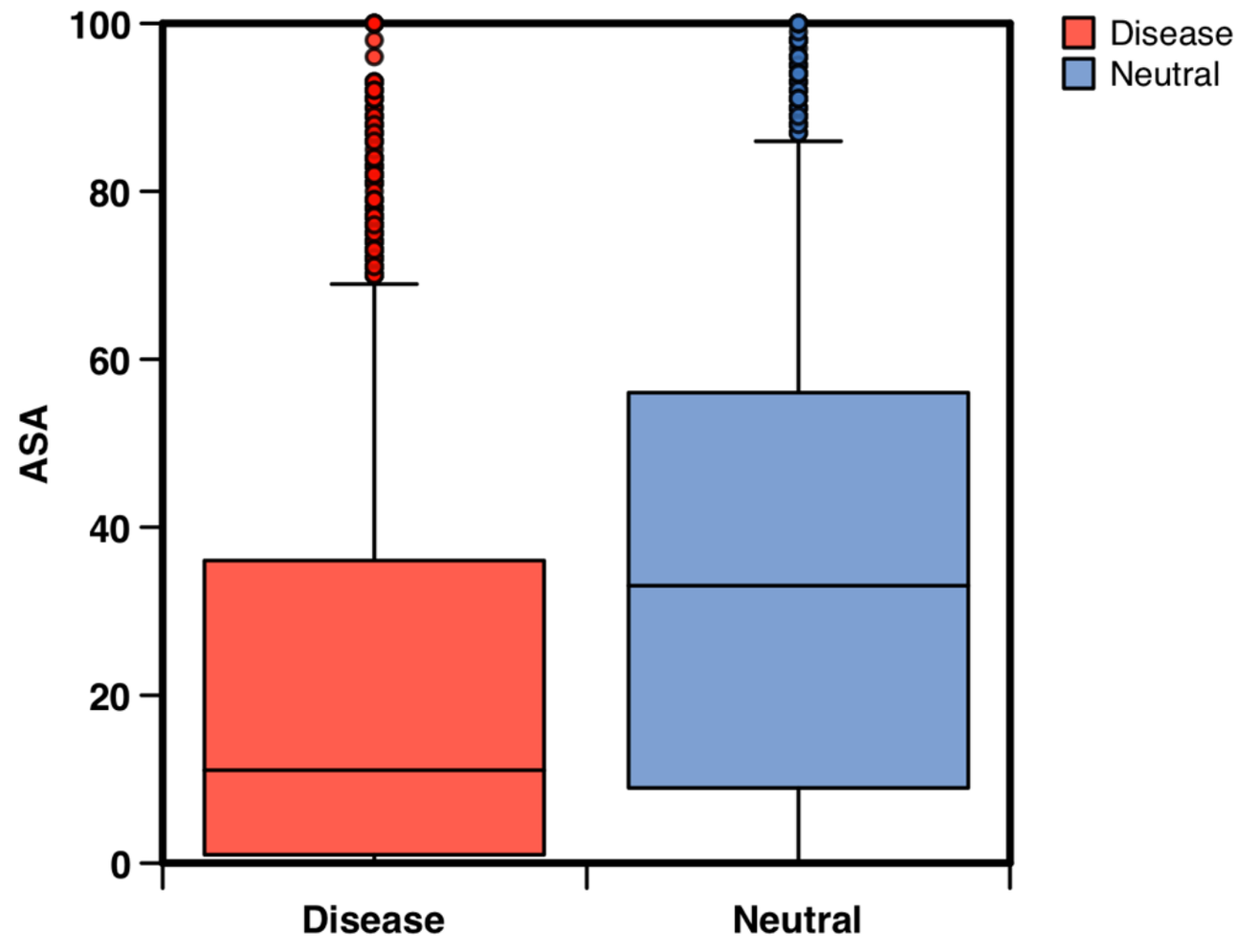
D = Disease related N = Neutral

DB= 33672 nsSNVs

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

Structure environment

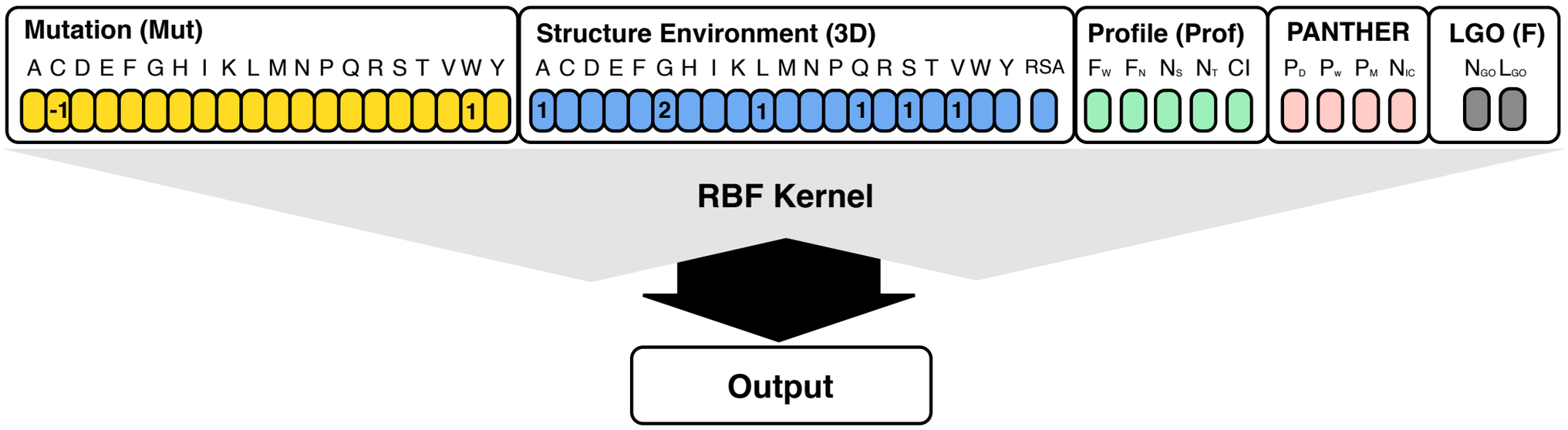
There is a significant difference (KS p-value = 2.8×10^{-71}) between the distributions of the relative Accessible Solvent Area for disease-related and neutral variants. Their mean values are respectively 20.6 and 35.7.

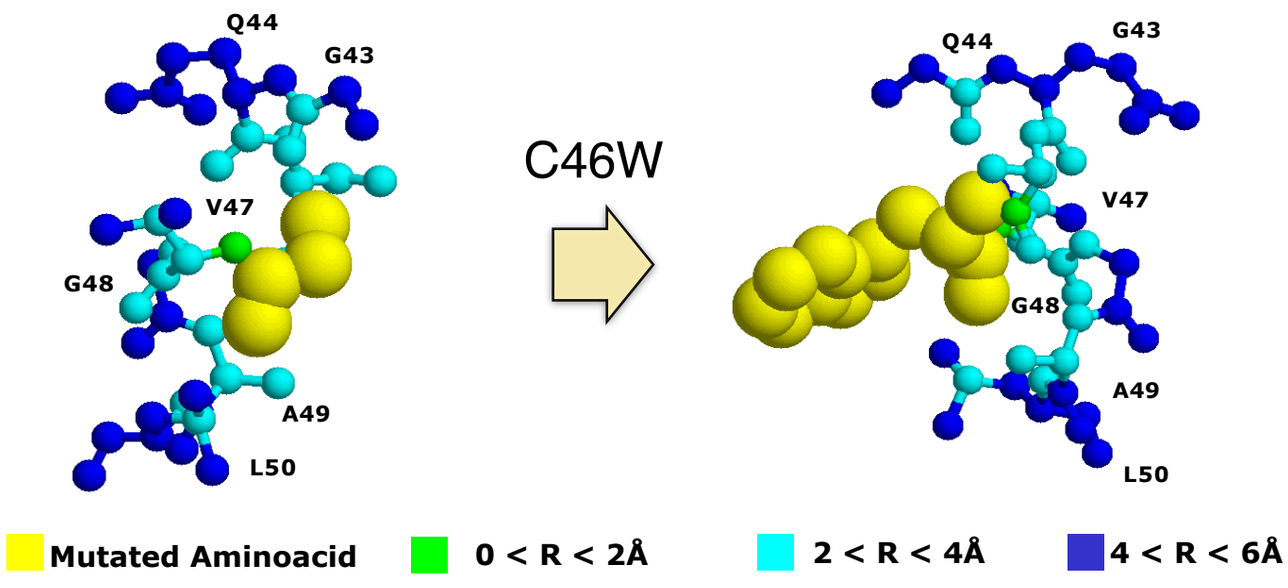


Capriotti and Altman. (2011) BMC Bioinformatics.12 (Suppl 4); S3.

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.

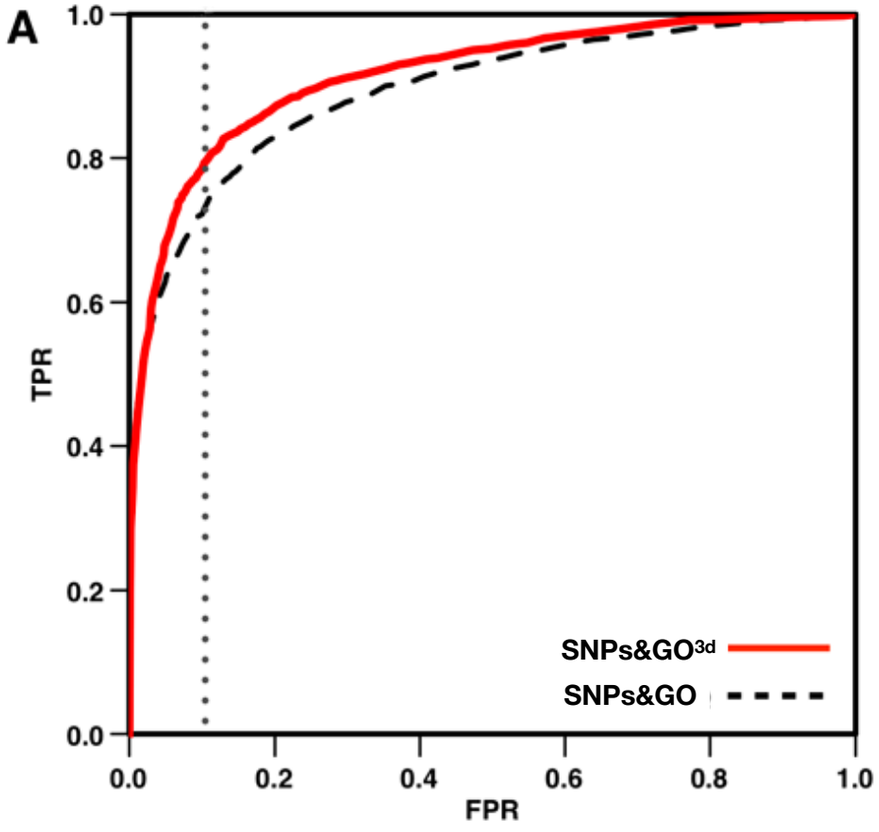


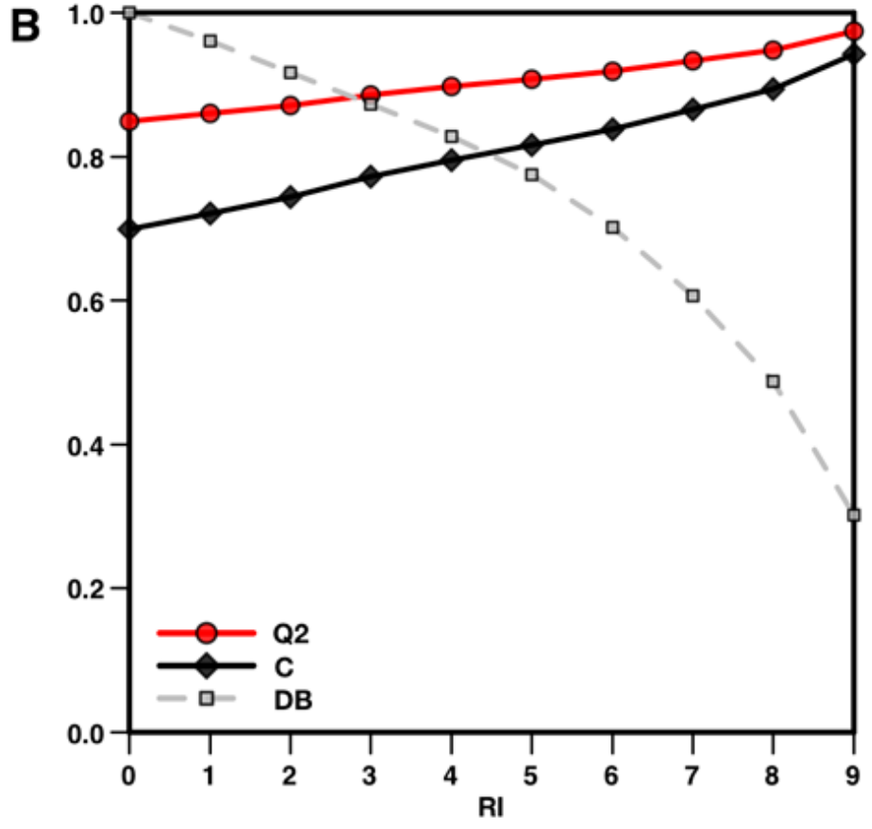


Sequence vs Structure

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



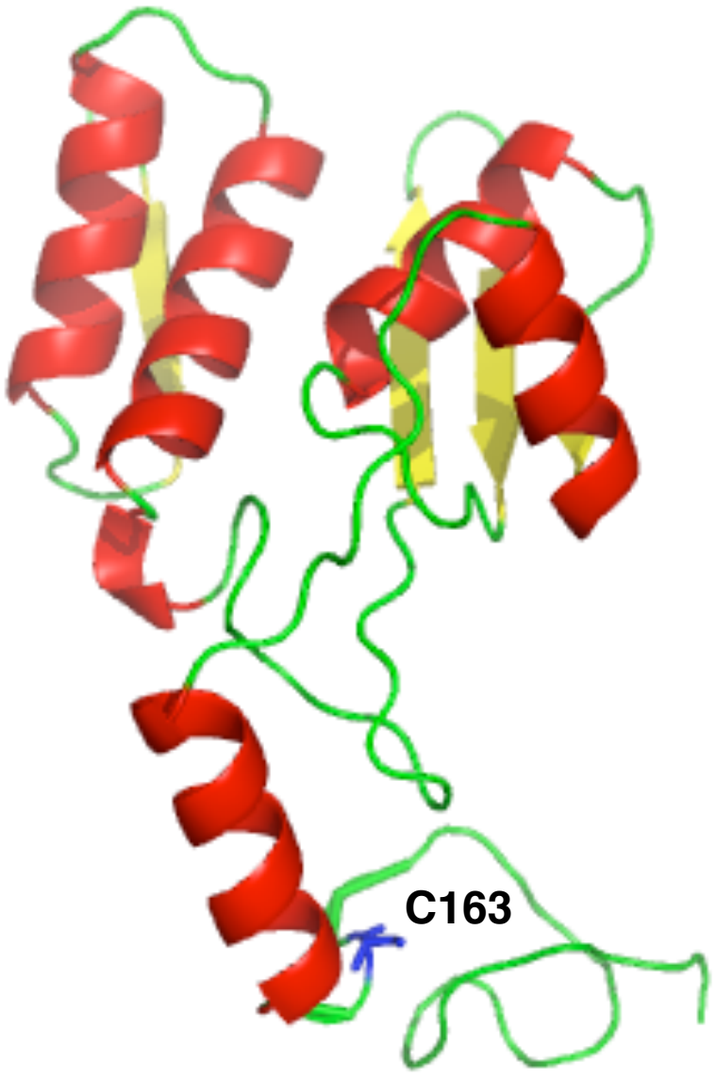


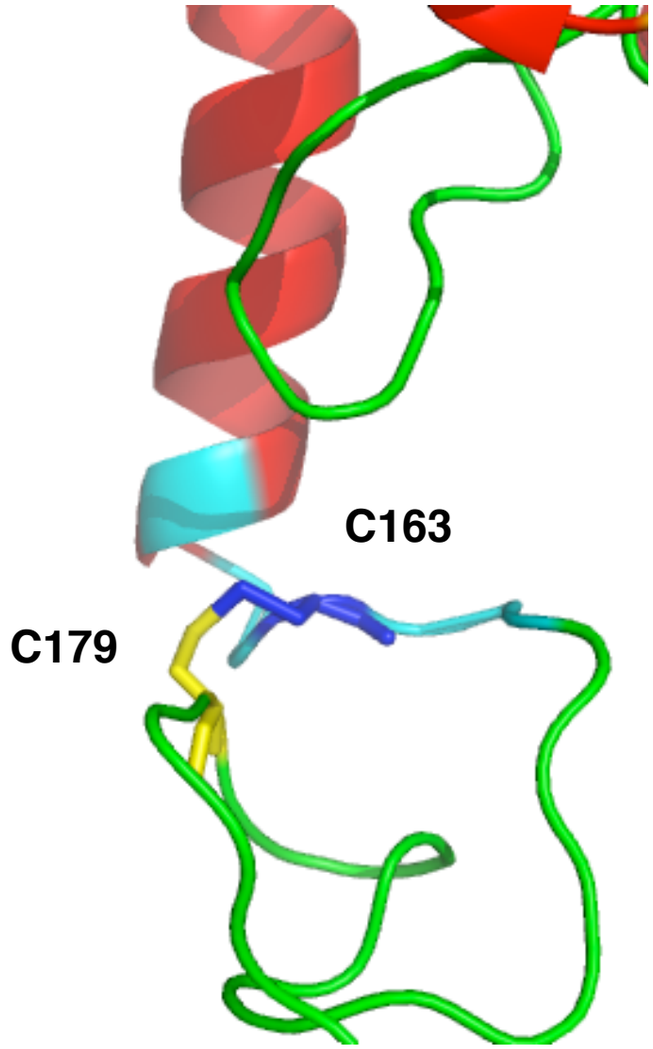
http://snps.biofold.org/snps-and-go

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.

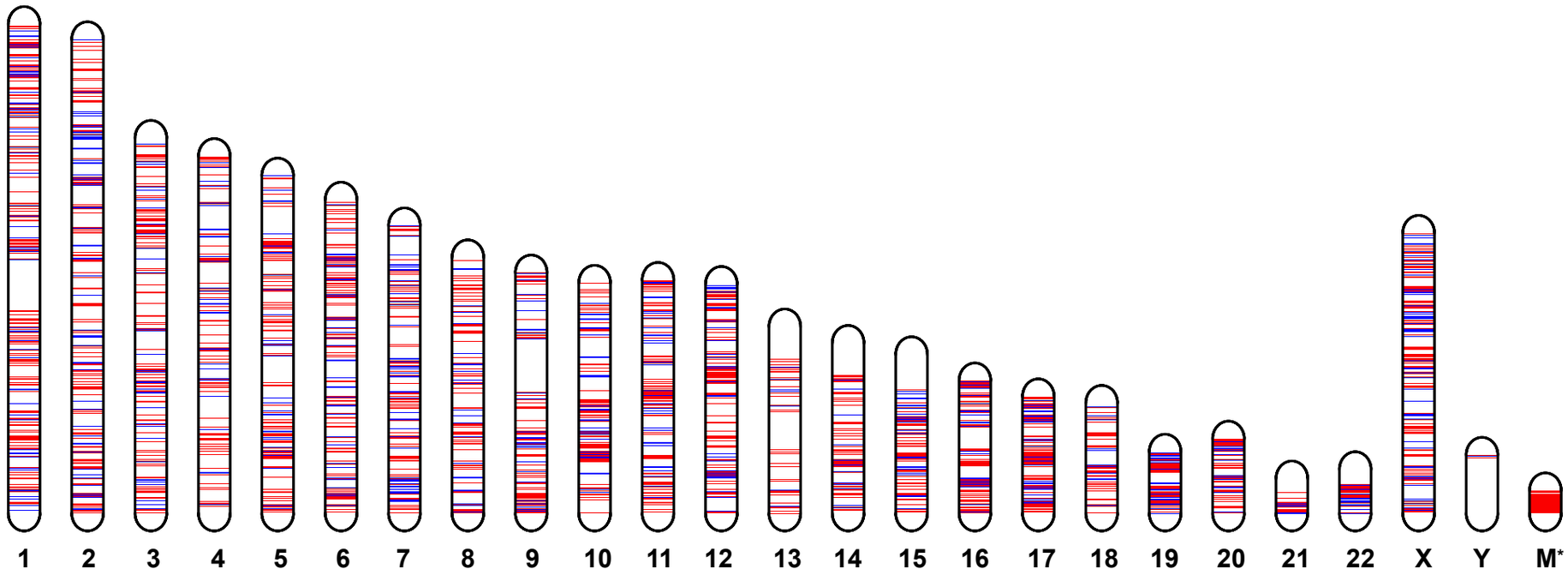
1APY: Chain A, Res: 2.0 Å





Whole-genome predictions

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

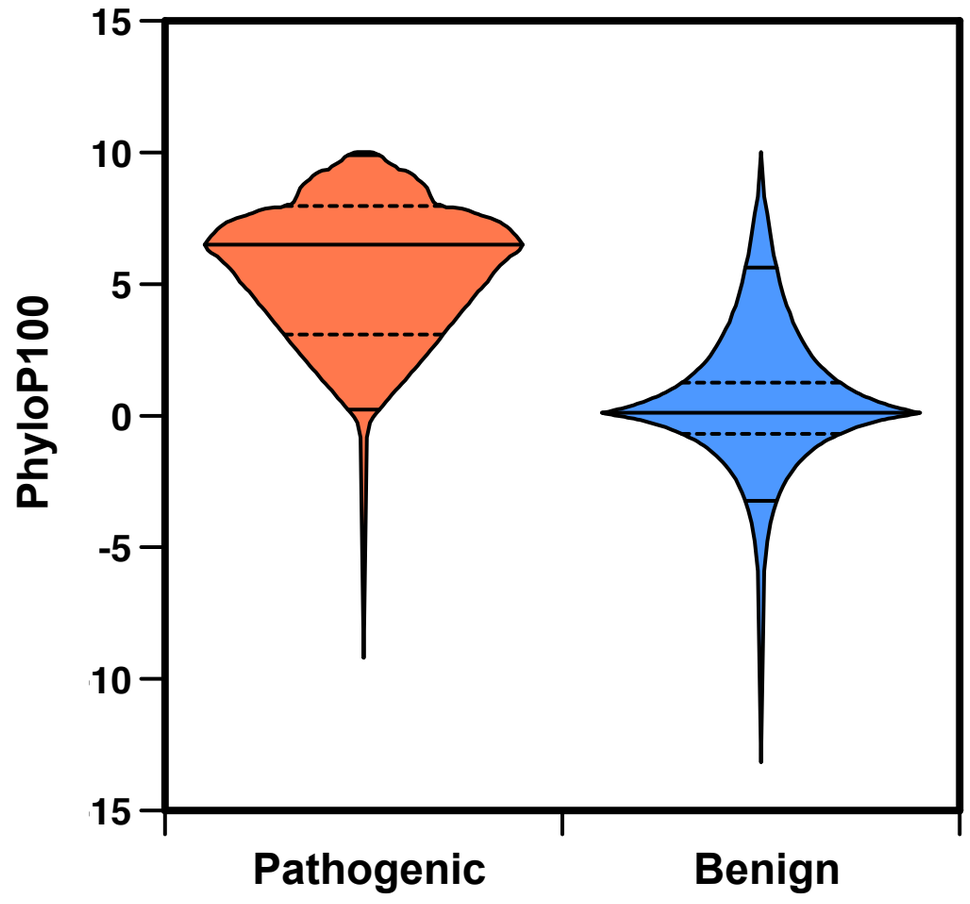


Predict the effect of SNVs in non-coding region is a challenging task because conservation is more difficult to estimate.

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

PhyloP100 score

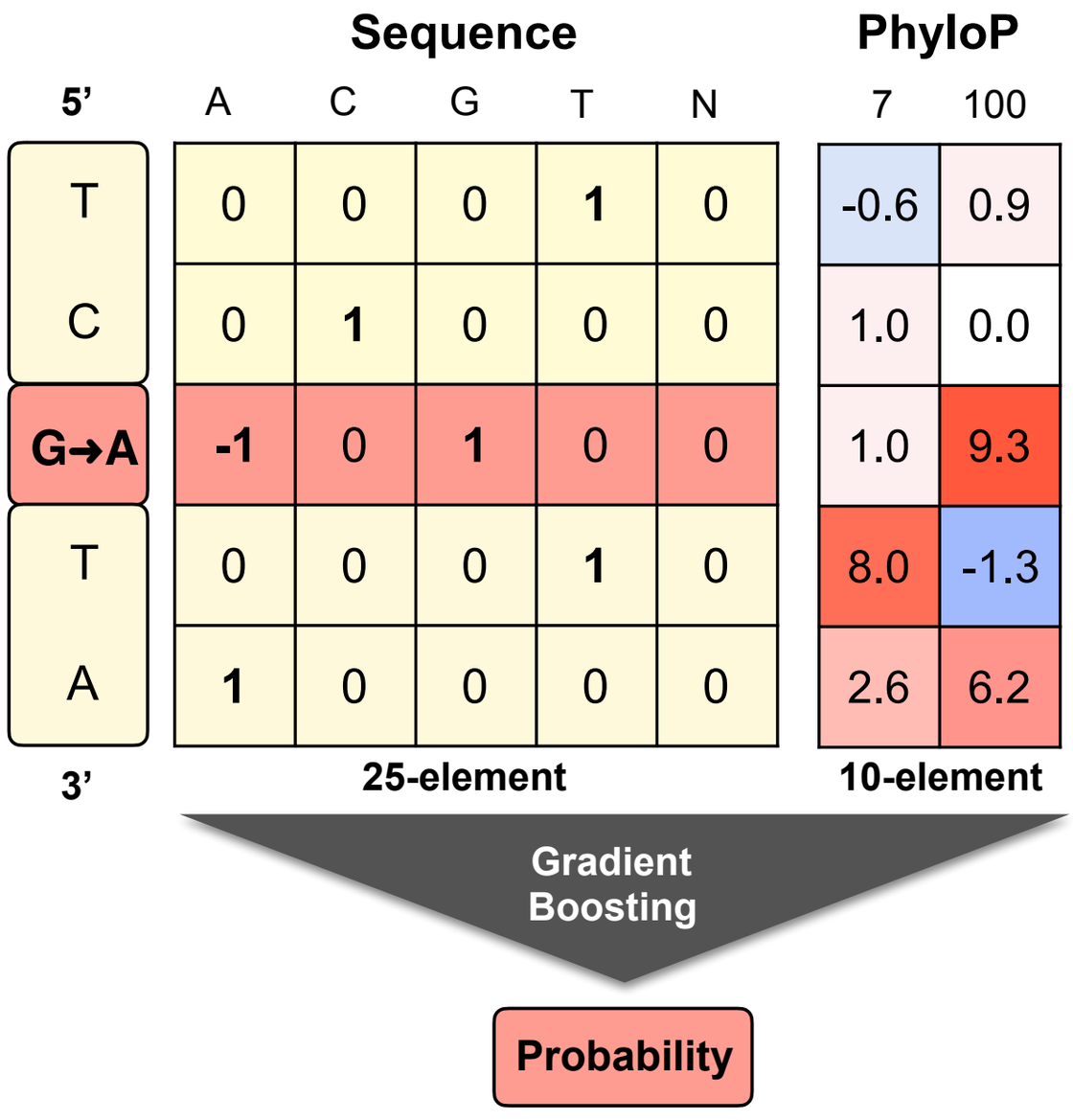
Conservation analysis based on the pre-calculated score available at the UCSC revealed a significant difference between the distribution of the PhyloP100 scores in Pathogenic and Benign SNVs.



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.

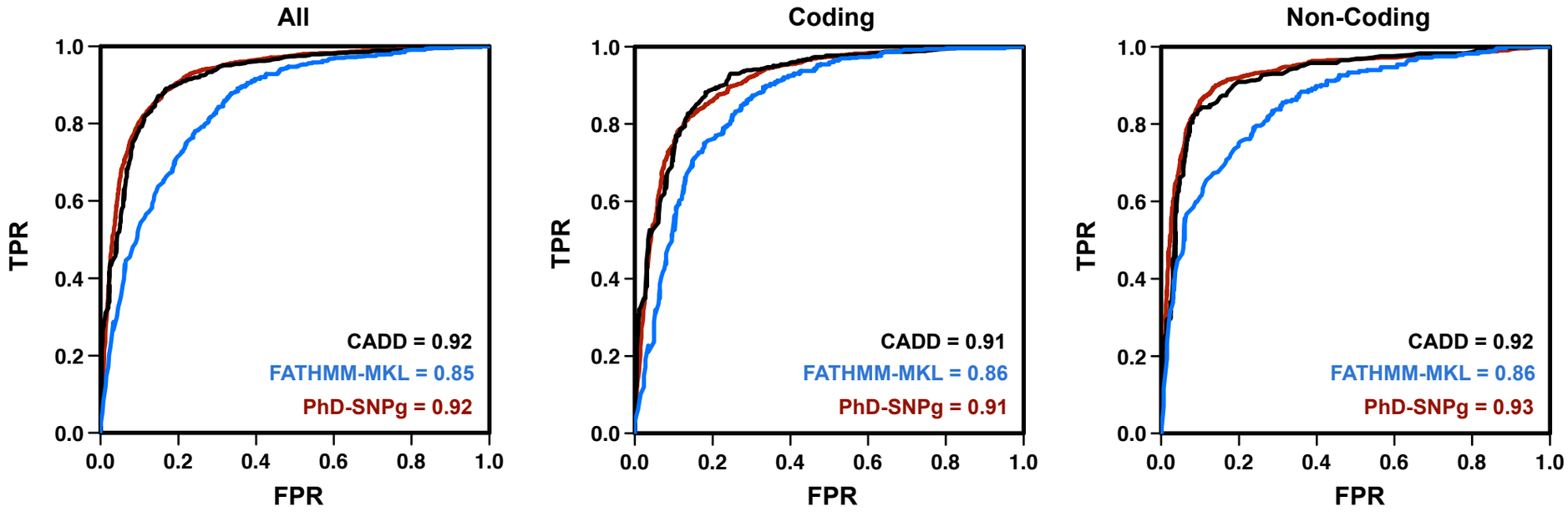


http://snps.biofold.org/phd-snpg/

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.

Blind Validation

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

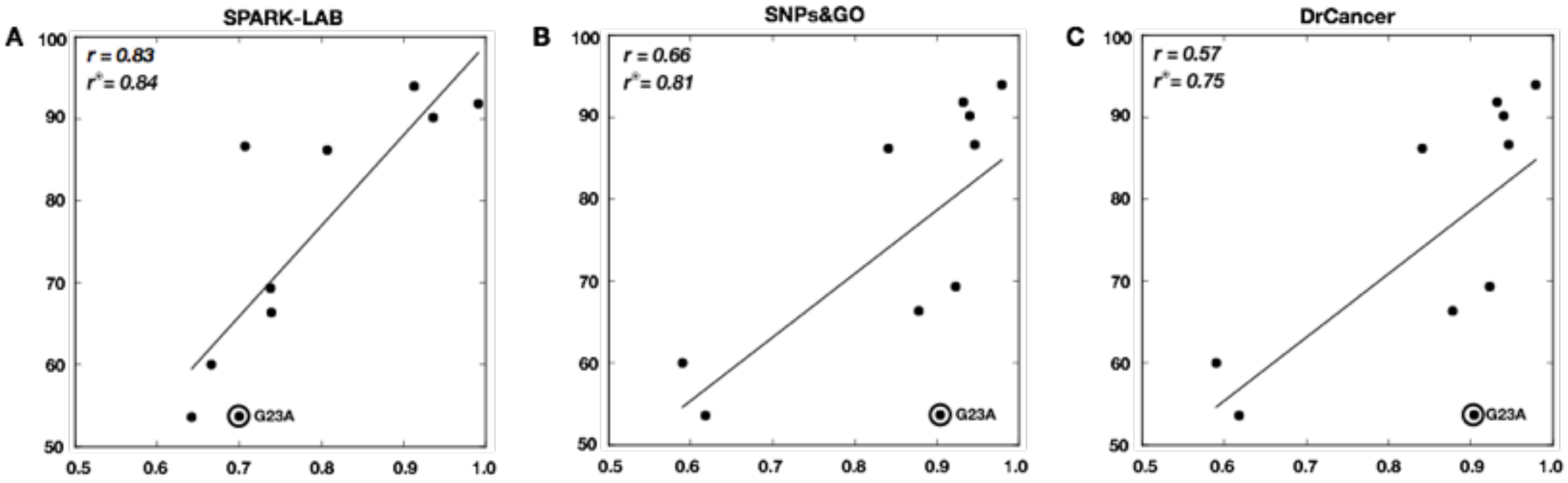
Proliferation rates predicted using the output of SNPs&GO without any optimization.

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

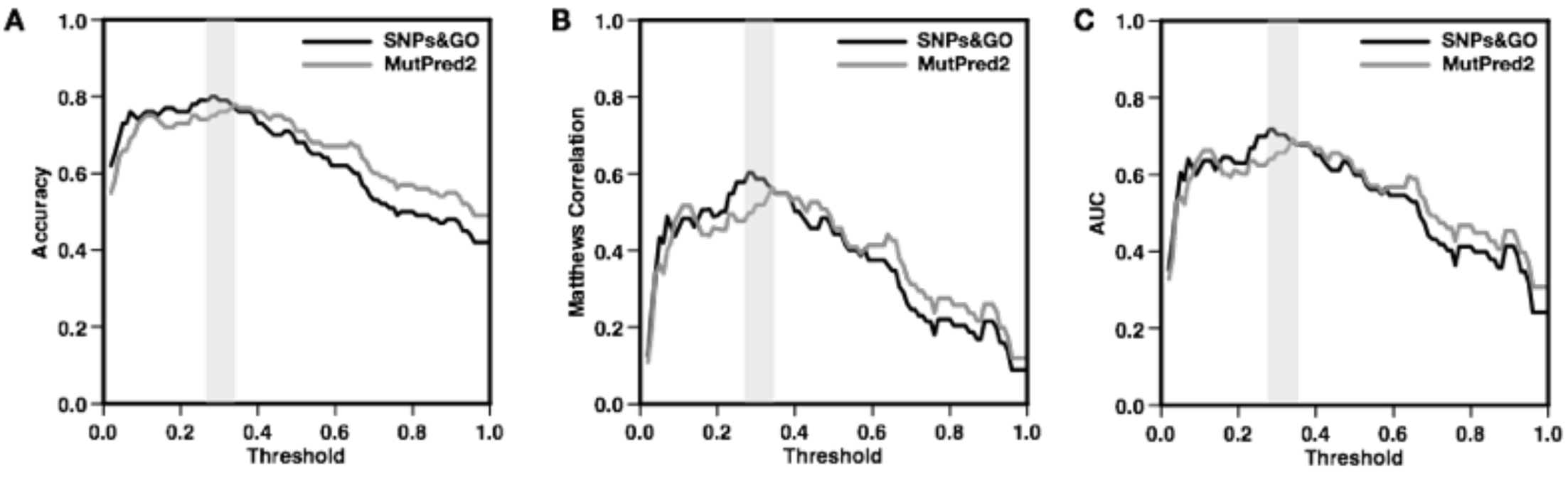
Challenge: Predict the effect of the 165 variants on NAGLU enzymatic activity.

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)

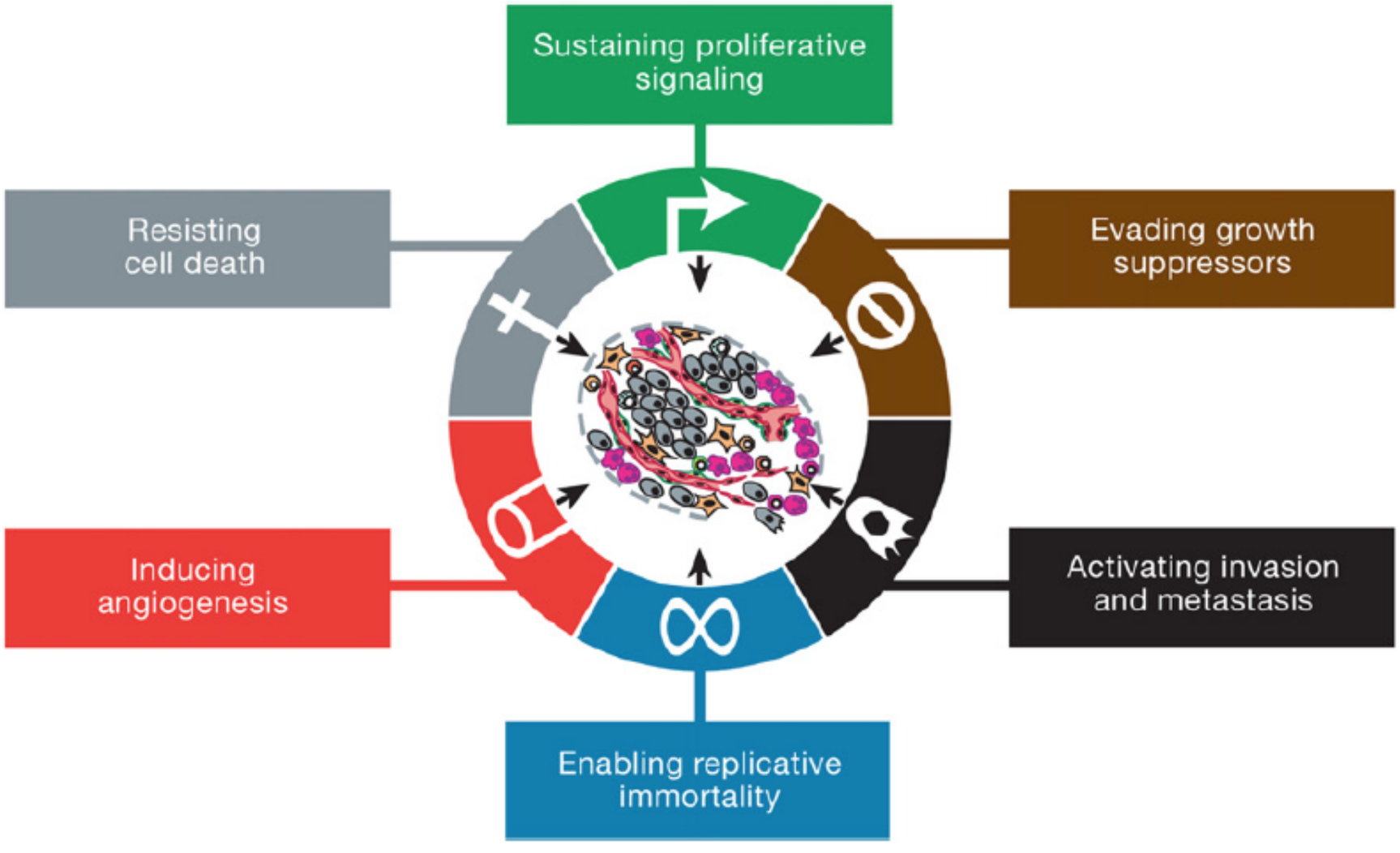
Method	Q2	AUC	MC	RMSE	$r_{Pearson}$	$r_{Spearman}$	$r_{KendallTau}$
MutPred2	0.780	0.850	0.565	0.30	0.595	0.619	0.443
SNPs&GO	0.800	0.854	0.603	0.33	0.575	0.616	0.445
SNPs&GO^{09}	0.750	0.749	0.499	0.46	0.477	0.495	0.409



Variations in Cancer

Hallmarks of cancer

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



Hanahan and Weinberg (2011) Cell, **144**:646

The complexity of cancer

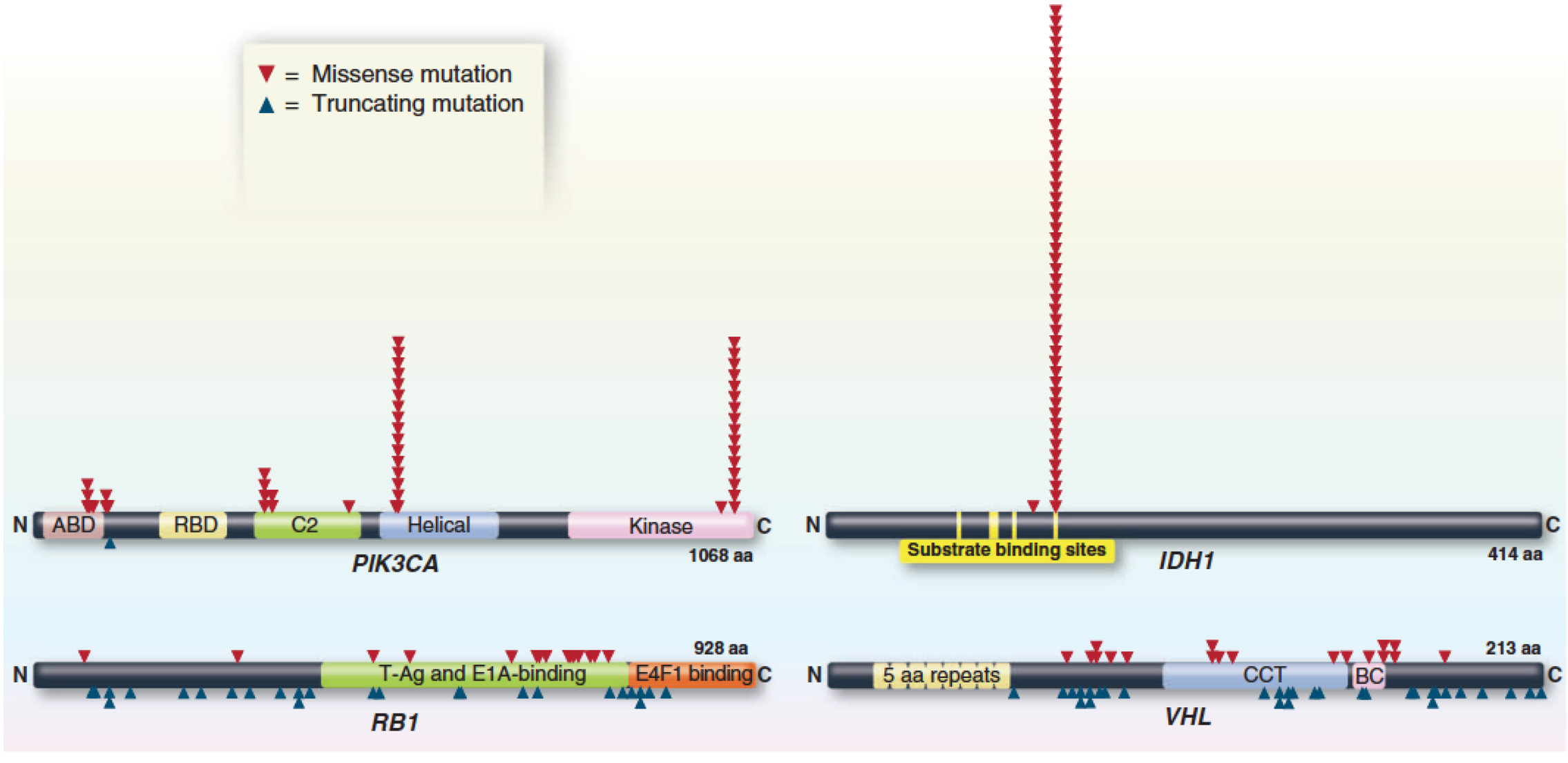
Cancer is **complex disorder** characterized by high level of mutation rate.

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.

Oncogene vs Suppressor

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



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

Main challenges

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

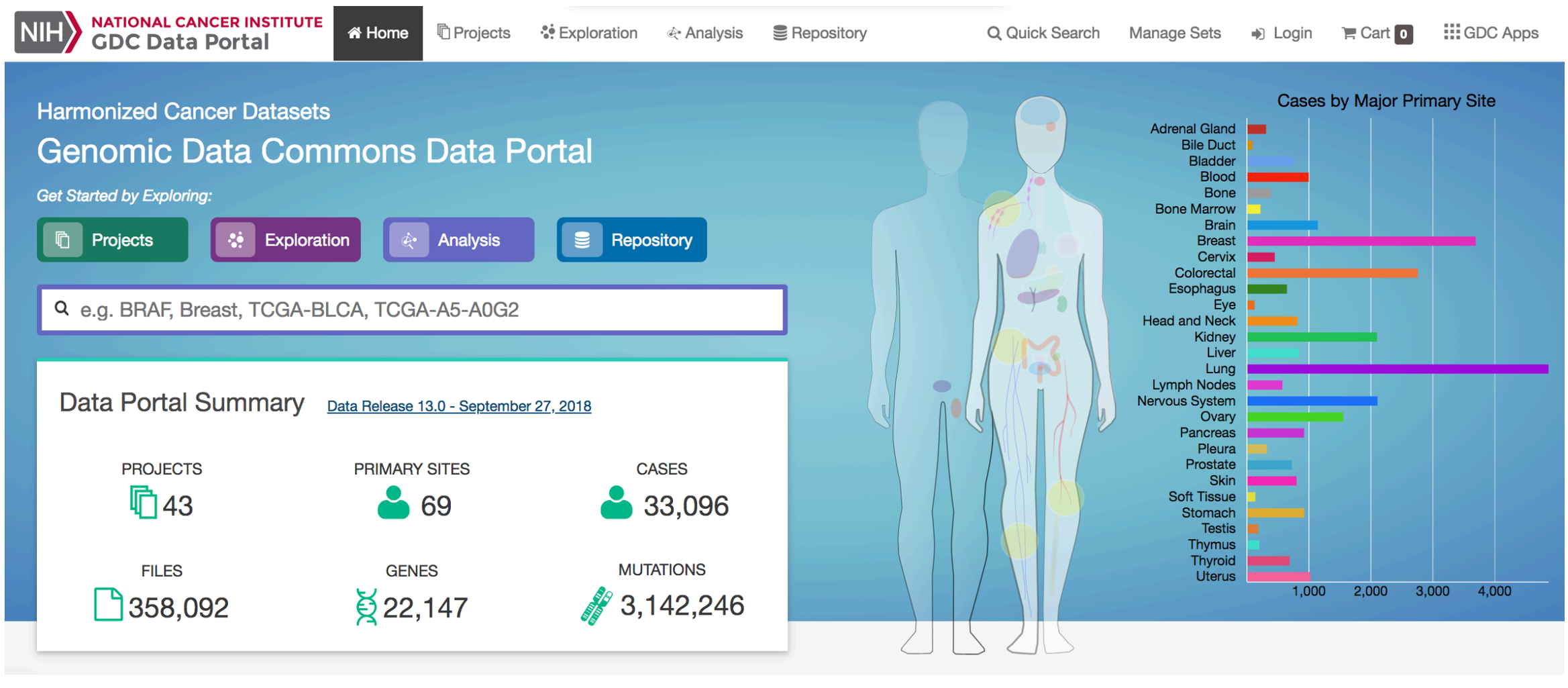
- Detection of recurrent somatic mutations and cancer driver genes;
- Prediction of driver variants and their functional impact;
- Estimate the impact of multiple variants at network and pathway level;
- Differentiate subclonal populations and their variation pattern.

The TCGA data

The Cancer Genome Atlas Consortium

Genomic Data Commons (https://portal.gdc.cancer.gov/)

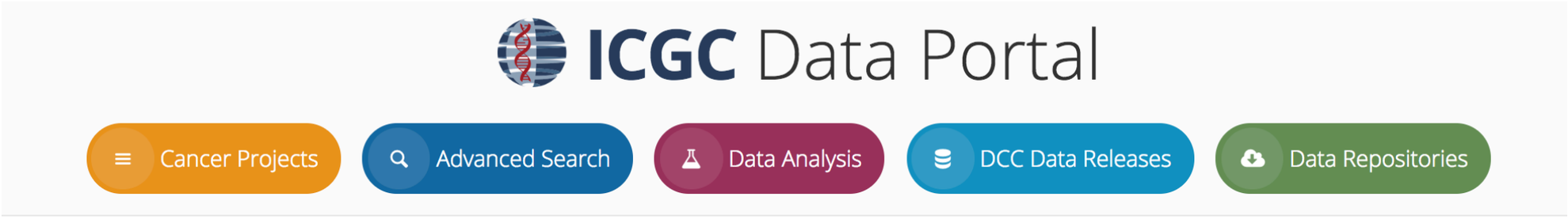
- 43 Projects
- 69 Primary sites



The ICGC data portal

The International Cancer Genome Consortium

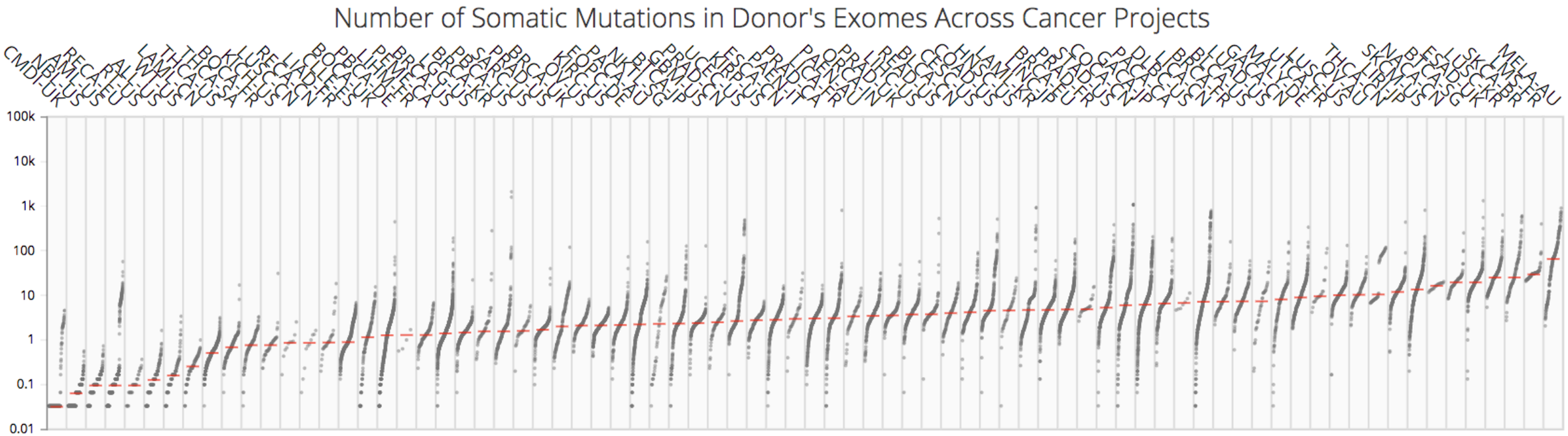
- ~24000 cancer patients
- 84 cancer projects in 22 primary sites
- more than 77 million simple somatic mutations.



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

Mutational landscape

The distribution of somatic variants varies significantly across cancer types

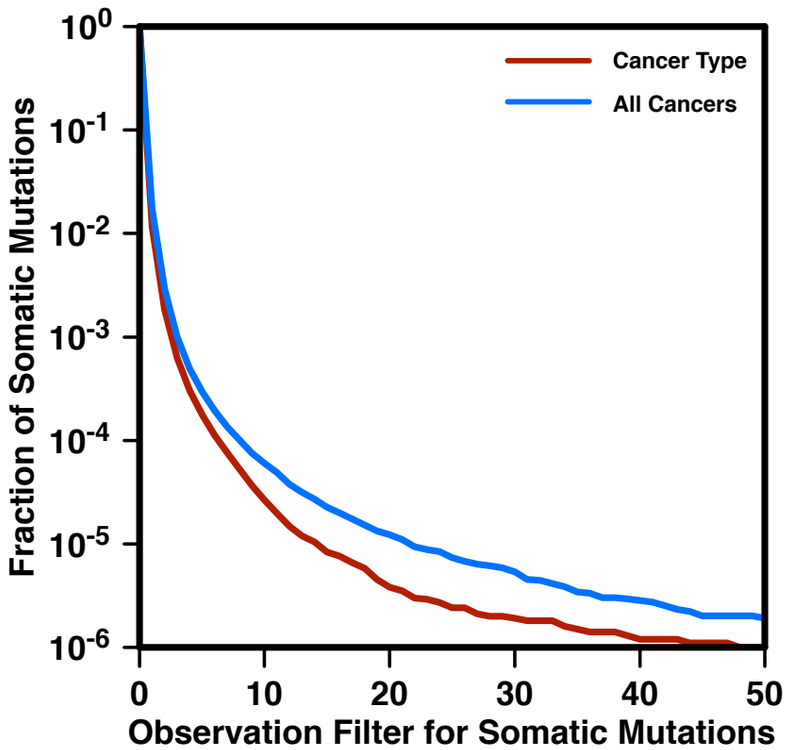


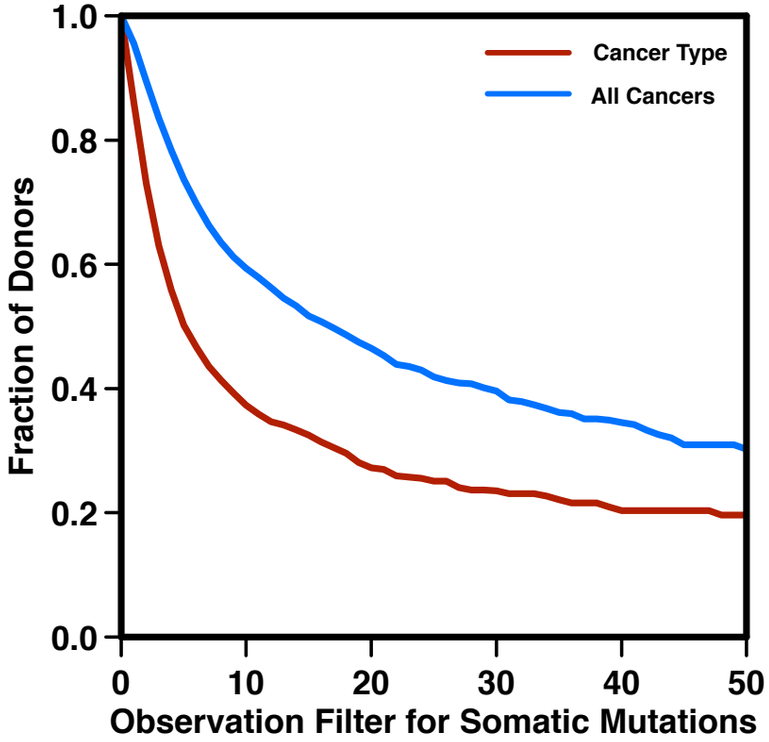
from *https://dcc.icgc.org/projects*

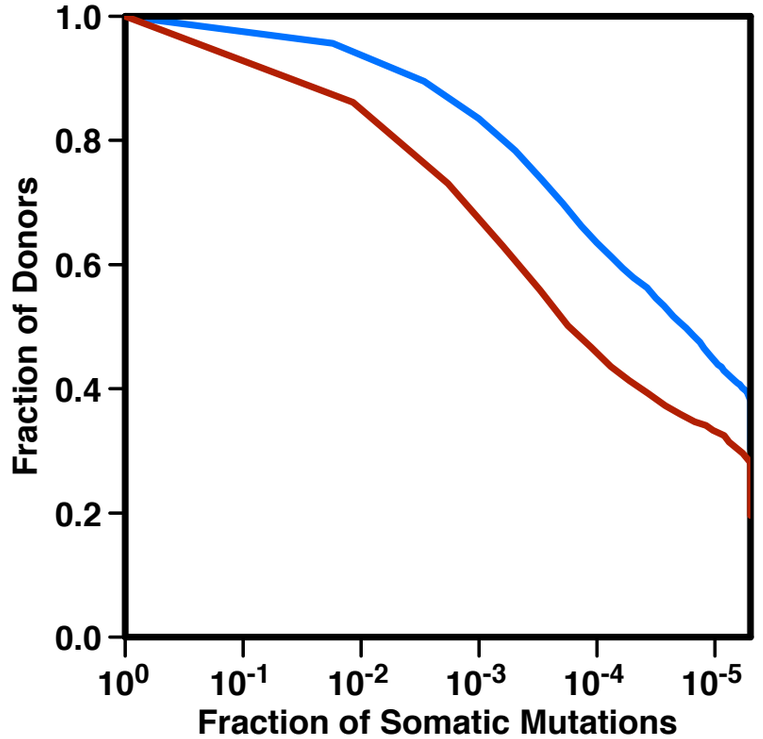
Driver vs Passenger

Number of recurrent mutations decrease exponentially.
On average a small fraction of variants is 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



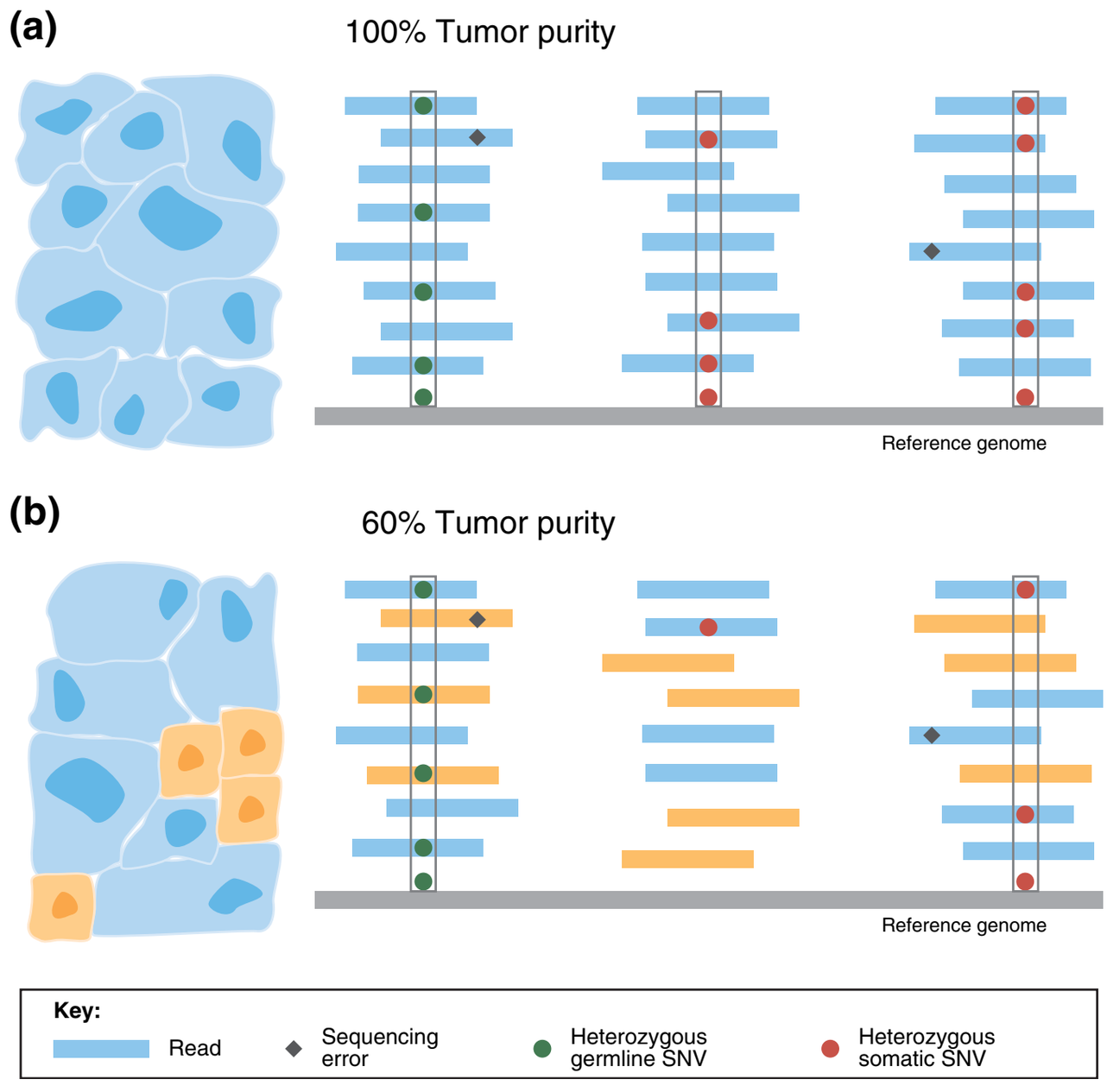




Tian R, Basu M, Capriotti E.(2015) BMC Genomics. 16 (Suppl. 8): S7.

Sample purity

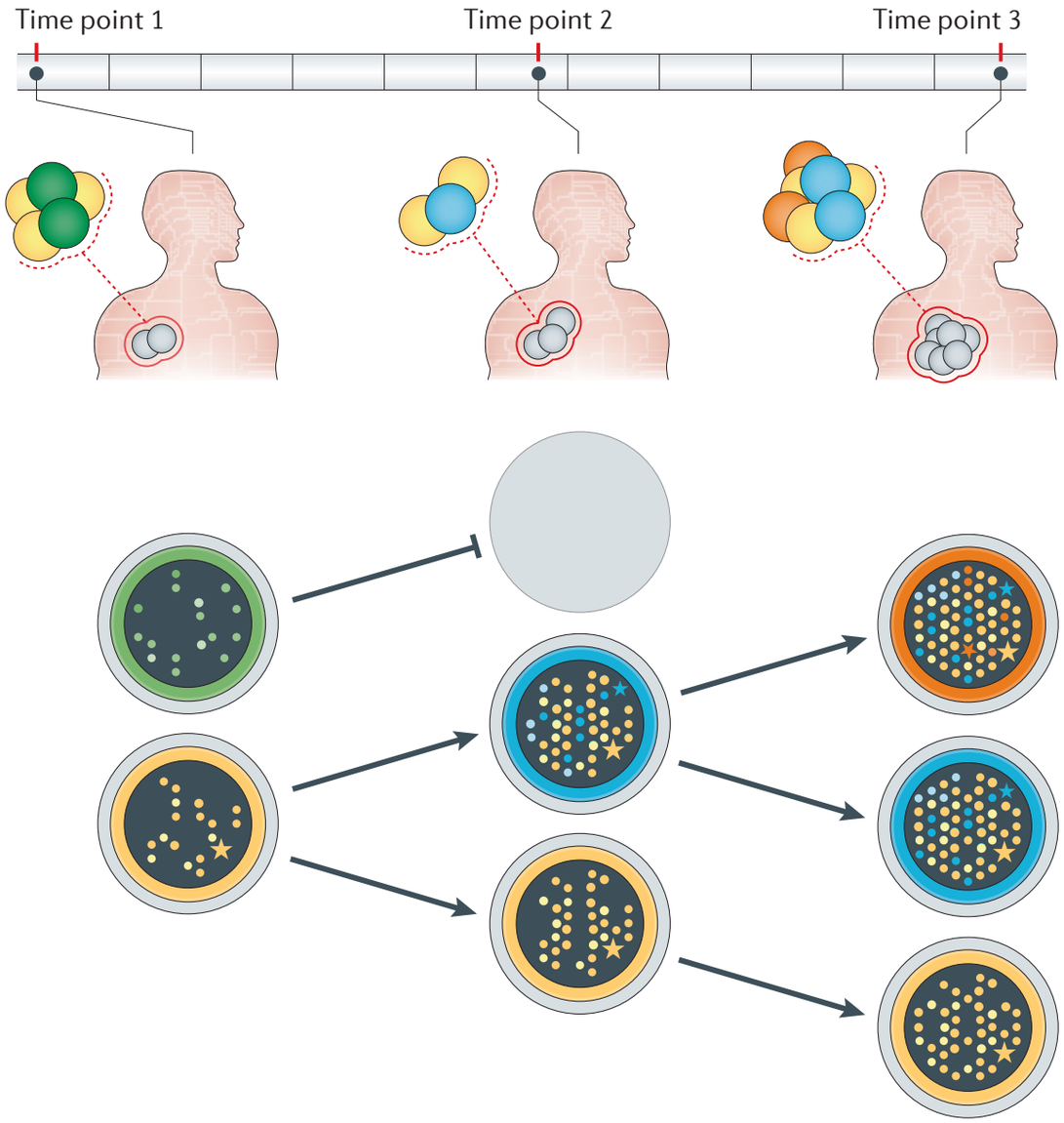
Impurity in the sample purity reduce the ability to detect variants

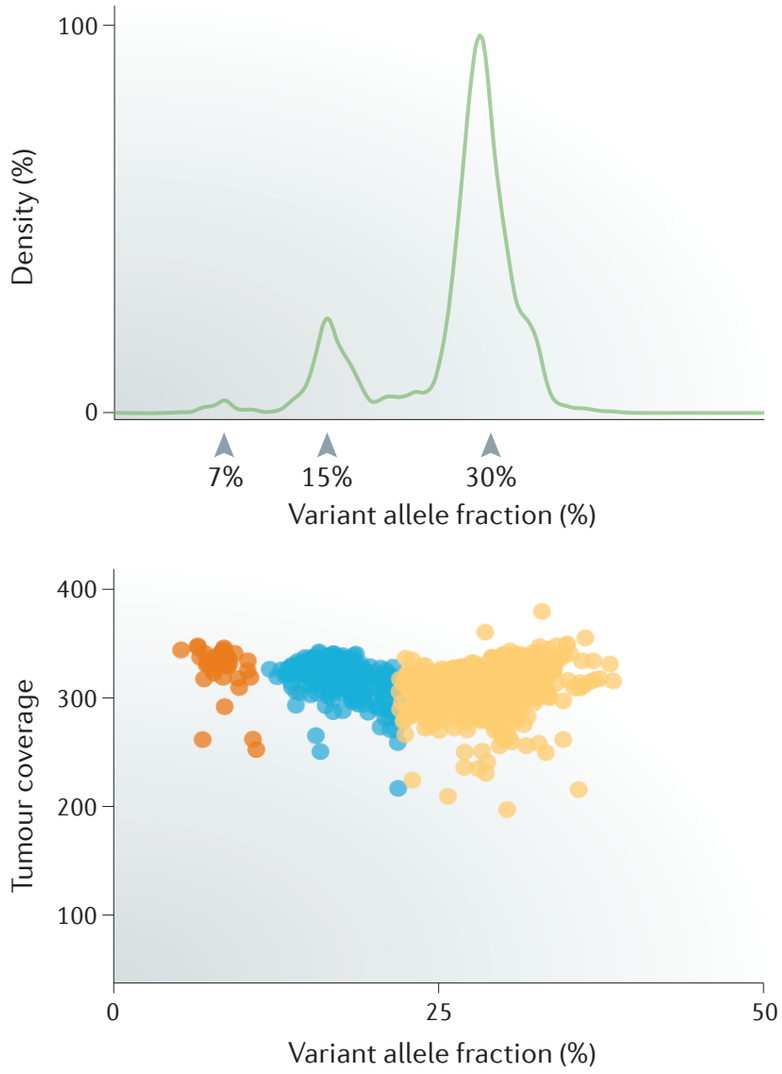


Raphael et al. (2014) Genome Medicine, **6**:5

Clonal evolution

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





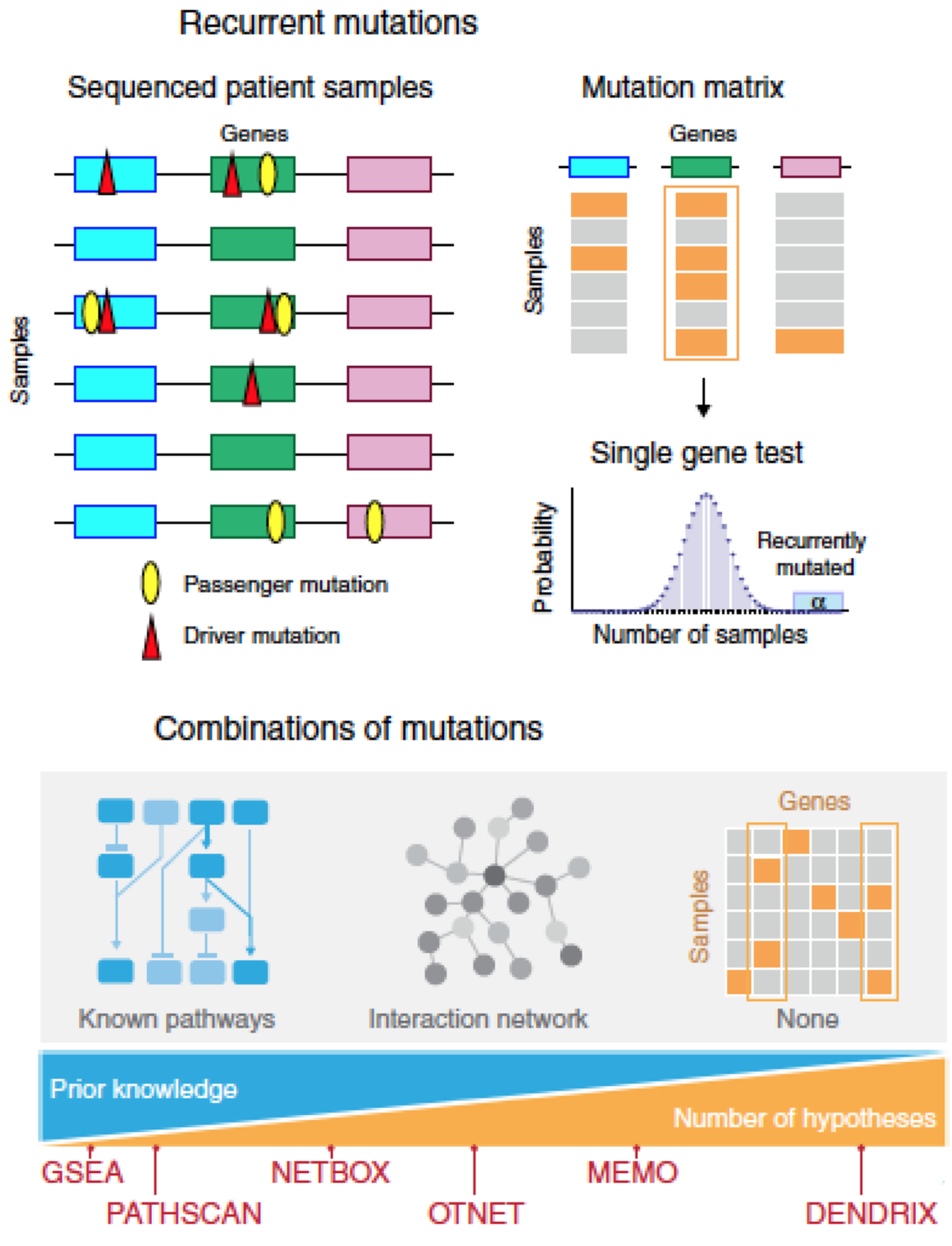
Ding et al (2014). Nat. Rev Genetics.

Recurrent variations

Recurrent mutations found in more samples than expected 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 of genes such as pathways and protein-protein interaction networks and de novo identification of combinations, without relying on a priori definition.

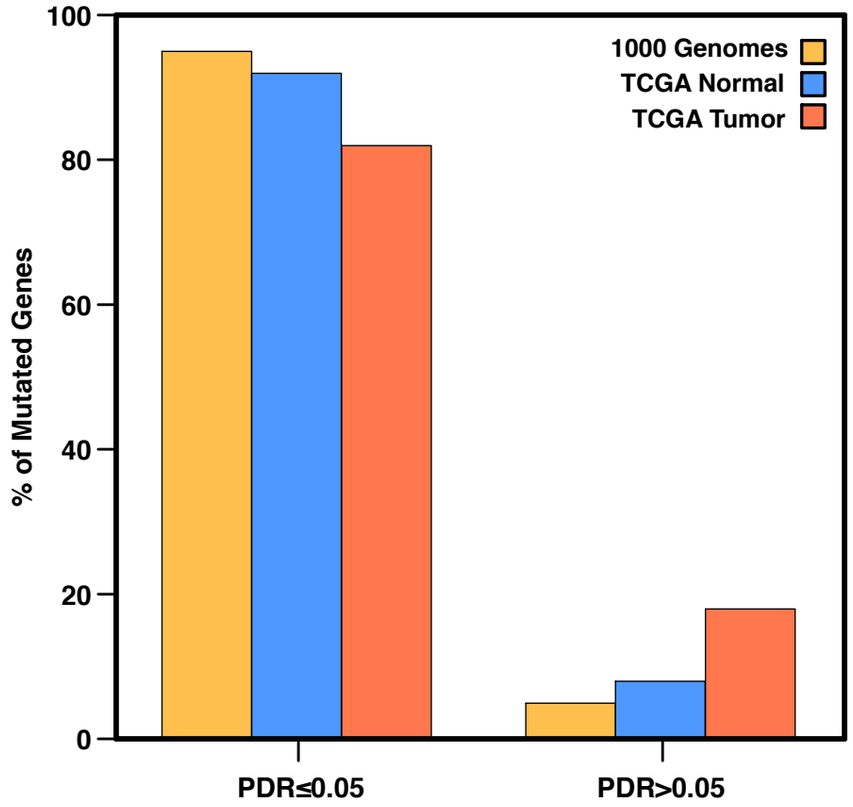


Raphael et al. Genome Medicine 2014, **6**:5

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

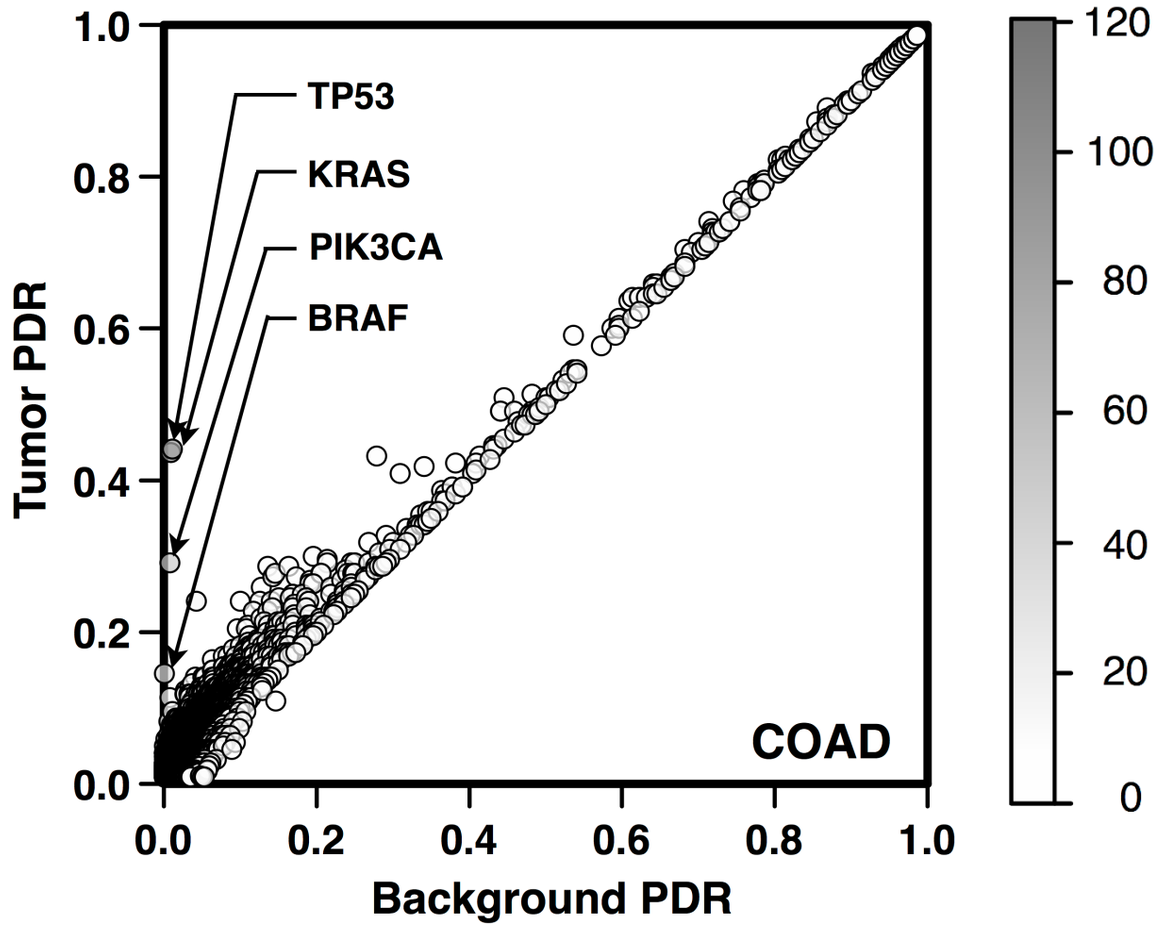
Fraction of samples in which a gene has ≥1 nonsynonymous variant with MAF≤0.5%

Gene prioritization

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)

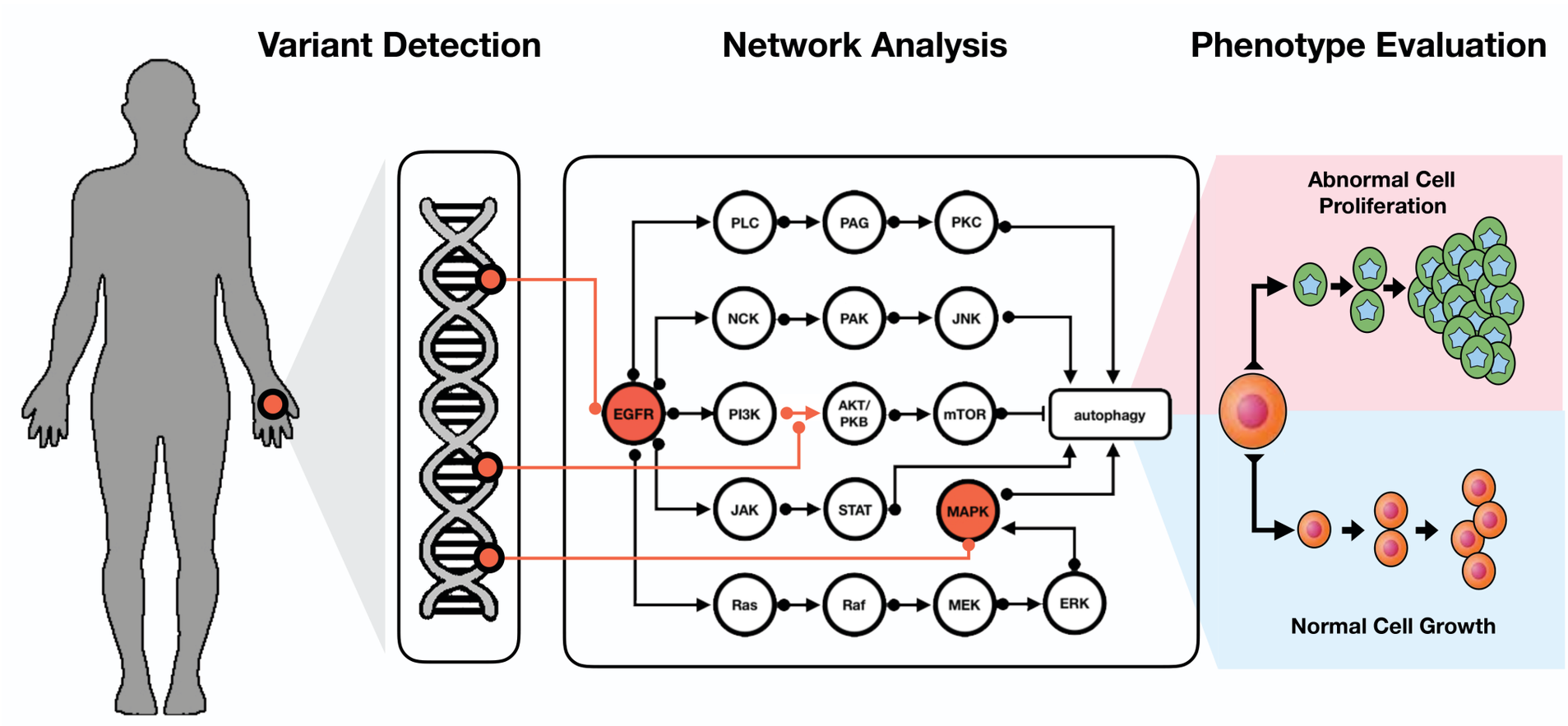


Tian R, Basu M, Capriotti E (2014). Bioinformatics. 30: i572-i578

Other Research Lines

Variants and networks

The simple one-variant one-phenotype model valid for many monogenic diseases does not capture the complexity of polygenic traits and disorders.



Capriotti E, Ozturk K, Carter H. (2018). WIREs Systems Biology and Medicine. Under review

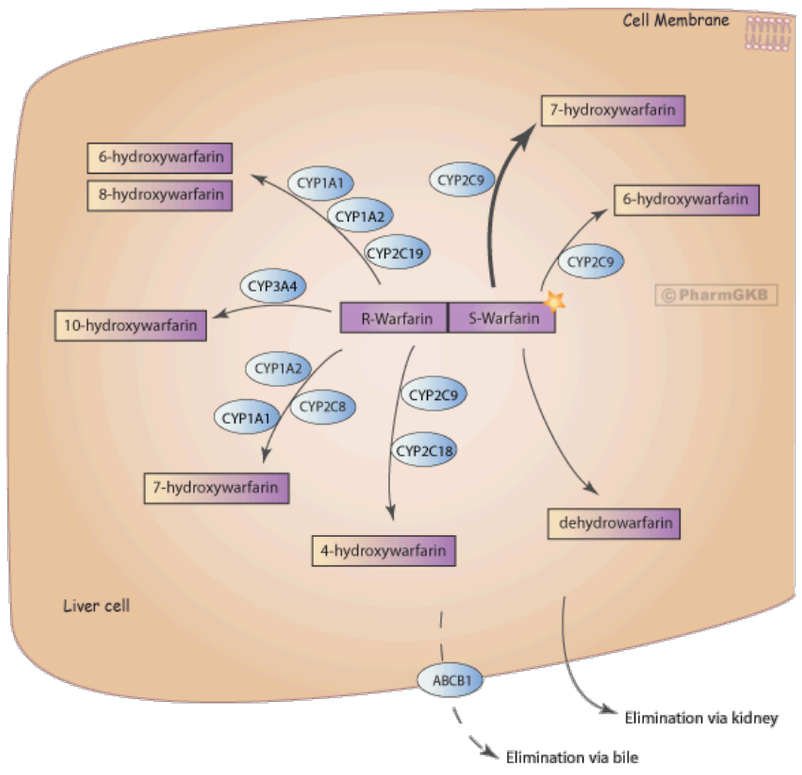
Variants and drug response

Pharmacogenomics aims at understanding how genetic variants influence drug efficacy and toxicity.

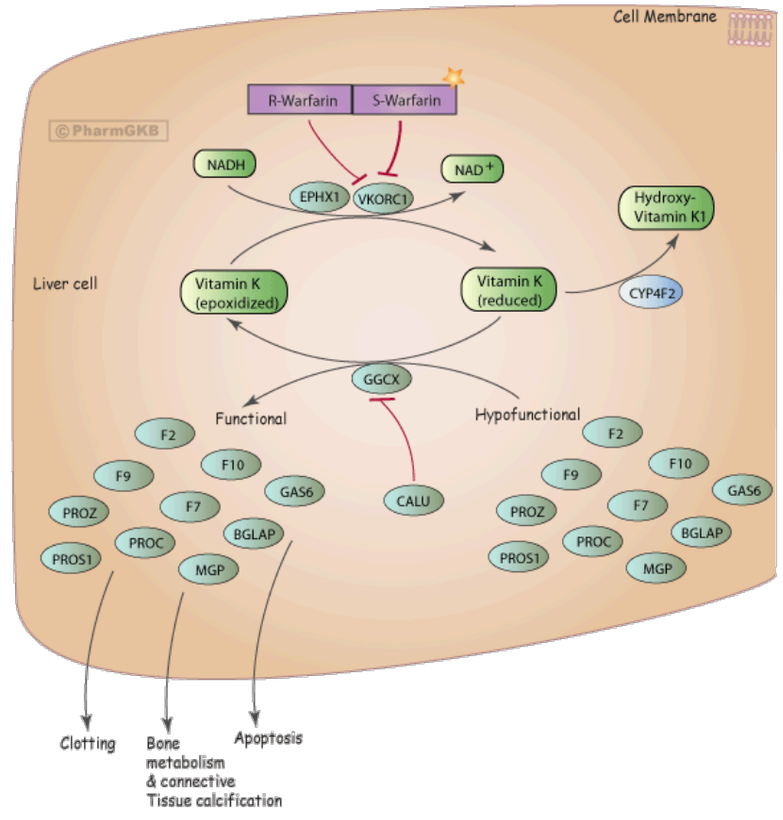
Pharmacokinetics variants: drug undergoes to bioinactivation via metabolic pathway. When the functionality of the pathway is compromised, a much higher concentrations of parent drug will accumulate.

Pharmacodynamics variants have an effect on the drug-receptor interactions and concentration. These variations have a directly impact on the dose-response relationships.

Warfarin and CYP2C9.



Warfarin and VKORC1



https://www.pharmgkb.org/

Conclusions

- The advances of the sequencing technology allowed to detect a huge amount of genetic variants whose function is unknown.

- Variant interpretation is a challenging task that can be solved by machine learning methods based on protein sequence, structure and function information.

- An important feature for variant interpretation is the sequence conservation. Variants in conserved regions are more likely to be pathogenic.This observation is valid also in noncoding regions.

- Statistical approaches for the analysis of genetic variations in cancer sample are important for developing gene prioritisation methods.

Future directions

- Development of computational methods for integration of omics data from different experimental techniques.

- Implement interoperable systems and software applications for storing and sharing genomic data.

- Detect genetic variants at single cell level. Test the effect of mutations using genome editing technique such as CRISPR-Cas9.

- Making all this information relevant at clinical level to improve health care system

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 Padova
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
NIH: 1R21 AI134027- 01A1
MIUR: FFABR
UNIBO: International Cooperation
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/