Functional characterization of single amino acid variants

Víctor López Ferrando

Universitat de Barcelona

Programa de Doctorat en Biomedicina

Functional characterization of single amino acid variants

Memòria presentada per Víctor López Ferrando per optar al grau de doctor per la Universitat de Barcelona.

Doctoral Student	Advisor and Tutor	Advisor
Víctor López Ferrando	**Dr. Josep Lluís Gelpí Buchaca**	**Dr. Modesto Orozco López**

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Universitat de Barcelona

To Anaïs

Acknowledgements

First of all I want to thank my advisors Josep Lluís and Modesto. Your dedication and passion has been an inspiration during this journey. I would also like to thank Ramon, who was my first contact with the group and encouraged me to apply for the scholarship; and to Núria and Xavier, for your insightful comments in the tracking commissions.

I want to thank all the colleagues with whom I have collaborated during this work: Òscar, Juan, Pau, Jürgen, Ricard, Diana... it was really exciting to work with such talented young scientists.

Thanks to all INB group members. Romina, Laia, Dmitry, Genís, Lluís, it was great to share this adventure with you. Thank you, Adam, for always helping me —and everyone—, and to the latest reinforcements: Salva, Vicky... it was a pleasure working with you.

To all the mates from BSC: Lucía, Mireia, Miguel, Didier, Arti, Dani... thanks for so nice lunch breaks, for sharing hopes and miseries. Brian, thank you for sharing your enthusiasm for science and bioinformatics and for introducing me to the science outreach group. I wish you all the best luck in the world!

A special thanks needs to be addressed to Pau, you were the best desk partner, and friend, I could have wished for. Thank you for the countless coffee breaks and their great conversations.

Thanks to my parents and my sister Laura, for encouraging me to take challenges and give the best of myself.

And thank you, Anaïs, for making life so easy and joyful.

Abstract

Single amino acid variants (SAVs) are one of the main causes of Mendelian disorders, and play an important role in the development of many complex diseases. At the same time, they are the most common kind of variation affecting coding DNA, without generally presenting any damaging effect. With the advent of next generation sequencing technologies, the detection of these variants in patients and the general population is easier than ever, but the characterization of the functional effects of each variant remains an open challenge.

It is our objective in this work to tackle this problem by developing machine learning based *in silico* SAVs pathology predictors. Having the PMut classic predictor as a starting point, we have rethought the entire supervised learning pipeline, elaborating new training sets, features and classifiers. PMut2017 is the first result of these efforts, a new general-purpose predictor based on SwissVar and trained on 12 different conservation scores. Its performance, evaluated both

by cross-validation and different blind tests, was in line with the best predictors published to date.

Continuing our efforts in search for more accurate predictors, especially for those cases were general predictors tend to fail, we developed PMut-S, a suite of 215 protein-specific predictors. Similar to PMut in nature, PMut-S introduced the use of co-evolution conservation features and balanced training sets, and showed improved performance, specially for those proteins that were more commonly misclassified by PMut. Comparing PMut-S to other specific predictors we proved that it is possible to train specific predictors using a unique automated pipeline and match the results of most gene specific predictors released to date.

The implementation of the machine learning pipeline of both PMut and PMut-S was released as an open source Python module: PyMut, which bundles functions implementing the features computation and selection, classifier training and evaluation, plots drawing, among others. Their predictions were also made available in a rich web portal, which includes a precomputed repository with analyses of more than 700 million variants on over 100,000 human proteins, together with relevant contextual information such as 3D visualizations of protein structures, links to databases, functional annotations, and more.

Contents

List of Figures

List of Tables

Abbreviations

Biology and bioinformatics

BLAST	Basic local alignment search tool
CYPs	Cytochromes P450
DNA	Deoxyribonucleic acid
GWAS	Genome wide association studies
LoF	Loss-of-function
MAF	Minor allele frequency
MSA	Multiple sequence alignment
NGS	Next generation sequencing
RNA	Ribonucleic acid

SAV	Single amino acid variant
SNP	Single nucleotide polymorphism
WES	Whole exome sequencing
WGS	Whole genome sequencing

Databases, institutions and consortiums

1000G	1000 Genomes
ACMG	American College of Medical Genetics and Genomics
CAGI	Critical assessment of genome interpretation
dbNSFP	Database for nonsynonymous SNPs' functional predictions
dbSNP	Single nucleotide polymorphism database
EBI	European Bioinformatics Institute
ExAC	Exome Aggregation Consortium
gnomAD	Genome aggregation database
HGMD	Human gene mutation database
ICGC	International Cancer Genome Consortium
OMIM	Online Mendelian Inheritance in Man

TCGA	The Cancer Genome Atlas

Machine learning

Acc.	Accuracy
AUC	Area under the ROC curve
CART	Classification and regression trees
CV	Cross-validation
FPR	False positive rate
HMM	Hidden Markov model
MCC	Matthews correlation coefficient
ML	Machine learning
NPV	Negative predictive value
PPV	Positive predictive value
RF	Random forest
ROC	Receiver operating characteristic
Sens.	Sensitivity
SGD	Stochastic gradient descent
Spec.	Specificity

SVM	Support vector machine
TPR	True positive rate

Diseases

LQTS	Long QT syndrome
MODY	Maturity-onset diabetes of the young
NPCD	Niemann-pick disorder

1. Introduction

> Indeed, it is not intellect, but intuition which advances humanity. Intuition tells man his purpose in this life.
>
> Albert Einstein

1.1 Human DNA sequence variation and disease

DNA sequence variation is great: it makes each one of us different, and this diversity, which is a key component of evolution, strengthens our resiliency as a species as it does for all life beings (Ellegren and Galtier, 2016; Quintana-Murci and Clark, 2013).

Compared with the reference genome, a typical human genome diverges in around 4 to 5 million sites. 99.9% of this variation is caused by single nucleotide polymorphisms and short indels, whereas the rest is due to structural variants (such as copy number variations) affecting more than 1 kilobase (Stankiewicz and Lupski, 2010; The 1000 Genomes Project, 2015). This variation, far from being randomly distributed along the sequence, follows certain patterns that help us unveil the origin and function of our DNA.

Most human variation originated millions of years before people first emigrated out of Africa between 50,000 and 60,000 years ago (Cavalli-Sforza and Feldman, 2003). These ancient polymorphisms are shared by all human populations and account for about 90% of the variation in any person (McClellan and King, 2010). The rest of the variation is much more rare an diverse.

New mutations, which occur at an average rate of 175 per diploid genome generation (Nachman and Crowell, 2000), prevail due to the exponential growth of human population since the development of agriculture 10,000 years ago and urbanization in the last centuries (Coventry et al., 2010). All these recent, rare mutations collectively represent the vast majority of human variation nowadays (Tennessen et al., 2012).

Genetic variation is subjected to both positive and negative evolutionary pressure, which has shaped some of the patterns we mentioned before. For example, the exome has a mutation rate that is half of the general genome, with synonymous variants being more prevalent than missense or nonsense ones (Lek et al., 2016). Even if the median number of SNVs per gene is 24, this number may range between 0 and more than 700, depending on the gene (Tennessen et al., 2012). These numbers are influenced by different factors, such as the biochemical properties of the variant (Cooper and Youssoufian, 1988), the location in the genome (Vicoso and Charlesworth, 2006), the existence of related genes (MacArthur et al., 2012) or the ancestry of the individual (Lek et al., 2016).

The study of DNA variation is fundamental in biomedicine, as almost all diseases have a genetic component. In the case of Mendelian diseases, the link between genotype and phenotype is generally easier to establish than for complex diseases. The latter are generally caused by a combination of environmental factors and a genetic predisposition in the form of mostly rare variants in different genes (McClellan and King, 2010; Tennessen et al., 2012).

A great deal of our current knowledge on genetics and human DNA variation comes from recent methodological advancements in sequencing technologies, those commonly known as next generation sequencing (NGS).

1.1.1 Next generation sequencing

Knowledge on the complete human genome sequence is very recent: the Human Genome Project published its first draft of the human DNA in 2001 (International Human Genome Sequencing Consortium, 2001) and its final version in 2004 (International Human Genome Sequencing Consortium, 2004). This project, with a cost of billions of dollars and based on the Sanger method for sequencing DNA, set the basis for the development of cheaper and faster sequencing methods.

The second generation of sequencers, based on improvements of the *shotgun* method, in which lots of small DNA fragments are sequenced in parallel, has fostered a revolution in biomedicine and genomics. Known as next generation sequencing, these methods are behind the explosion in the number of species and human genomes sequenced in the past decade.

As an example of the amount of information generated by NGS, we can see the growth rate of UniRef100 (the reference sequence database used in this work), which has grown in a 4x factor during the last 5 years (Figure 1.1). Innovation in sequencing methodology hasn't stopped, the cost of sequencing a human genome is nowadays slightly above $1000 (Schwarze et al., 2018), and a third generation of sequencers is underway, promising even faster and cheaper DNA analyses (Schadt et al., 2010). This new availability of sequences has

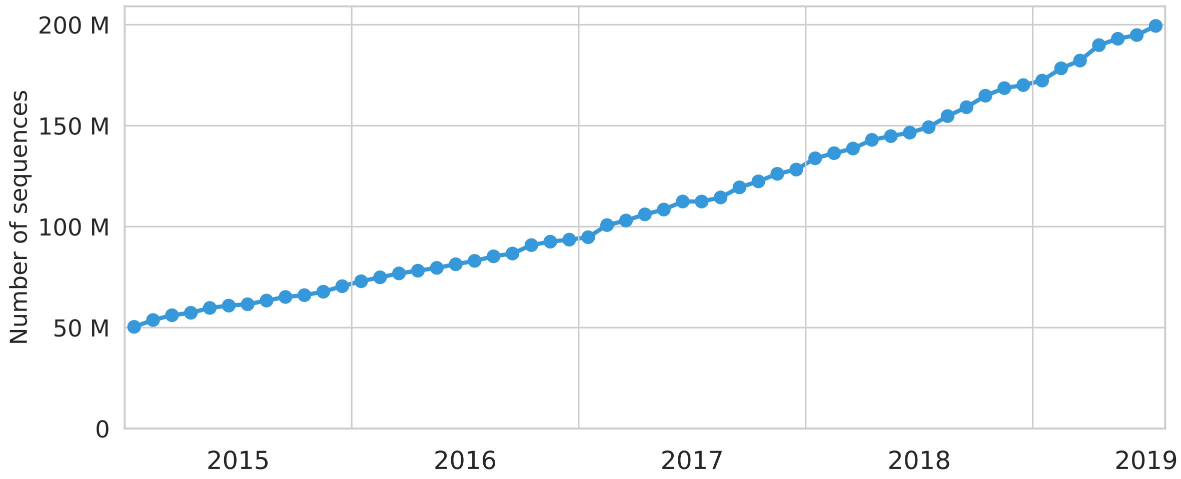


Figure 1.1 Monthly evolution of the number of sequences in the UniRef100 database during the work on this thesis.

shifted the focus of the research community from the sequencing of genomes to their interpretation (Mardis, 2010; Schrijver et al., 2012).

1.1.2 Large-scale sequencing studies

The first human genome sequence assembled was in fact a combination of sequences from different volunteer donors. Once this consensus sequence was released, the bases were established to start studying DNA variation among the population. The HapMap project (McVean et al., 2005; The International HapMap Project, 2003) was the first project working towards this goal, its main objective being the identification of common patterns in DNA sequence variation. For this purpose, they sequenced DNA samples from 270 unrelated individuals from 4 different geographic ancestry locations. As a result of this study, more than one million SNPs (single nucleotide polymorphisms, point variants with a minor allele frequency of at least 1%) were identified.

The next important milestone in the study of DNA variation came from the 1000 Genomes initiative (1000G, The 1000 Genomes Project, 2015), consistent on the low-coverage study of the whole genome and the deep exome sequencing of 2,504 healthy individuals from 26 different populations. In the same line as 1000G, but limited to exomes, are the ESP project (Exome Sequencing Project, Fu et al., 2013), which sequenced a total of 6,515 exomes, and more importantly, the ExAC project (Exome Aggregation Consortium, Lek et al., 2016), which analyzed the exomes of 60,706 humans of varied ancestries.

Different initiatives have also emerged aggregating data from these projects such as dbSNP (Sherry et al., 2001), the classic database collecting SNPs. More recently, the Genome aggregation database (gnomAD, Karczewski et al., 2019) has been released, aggregating variants from 141,456 human exomes and genomes.

Large-scale sequencing studies have not been limited to the sequencing of healthy individuals, but are a fundamental means for the study of complex diseases via Genome Wide Association Studies (GWAS, The Wellcome Trust Case Control Consortium, 2007). Two flagship projects on cancer such as The Cancer Genome Atlas (TCGA, The Cancer Genome Atlas Research Network, 2008) and the International Cancer Genome Consortium (ICGC, The International Cancer Genome Consortium, 2010) have also sequenced thousands of cancer genomes to better understand these diseases.

Sequencing technology is not only limited to DNA: RNA-Seq (Wang et al., 2009) allows to quantitatively measure the transcriptome, and thus quantify gene expression. The GTEx project (Genotype-Tissue Expression, The GTEx Consortium, 2015) is the first large-scale attempt to complement whole genome sequencing with RNA sequencing from 175 individuals across 43 different tissues. RNA-Seq data has also been collected in the ICGC and TCGA projects, and the joint analysis of DNA sequence and gene expression is one of the current challenges in biomedicine.

1.1.3 Genetic variation and Mendelian disorders

Mendelian disorders, also known as monogenic disorders, are diseases caused by alterations on a single gene. About 0.4% of newborns present a clinically recognized Mendelian phenotype (Baird et al., 1988). Mendelian diseases are usually caused by the inheritance of pathogenic alleles, although they may also originate from *de novo* variants (Veltman and Brunner, 2012), being this the prevalent cause for some rare diseases (Heinzen et al., 2012). The knockout of any non-redundant protein-coding gene compatible with live is suspected to cause a Mendelian phenotype, as suggested by studies in mice (Ayadi et al., 2012).

Research on Mendelian disorders has focused on identifying the links between genes and diseases. Traditionally, this was done by positional cloning and candidate-gene approaches (Bamshad et al., 2011; Collins, 1995). In recent years, whole exome and whole genome sequencing have replaced these methods and underlie nowadays the vast majority of gene discoveries (Chong et al., 2015; Ng et al., 2010; Tennessen et al., 2012). Exome sequencing has even proved valuable for the detection of undiagnosed genetic conditions (Need et al., 2012).

Mendelian phenotypes have clearly been associated with alterations in the normal coding sequence of proteins, and are not generally caused by synonymous variants (Botstein and Risch, 2003). In the most extreme case, the gene may be affected by a loss-of-function (LoF) variant. The loss-of-function of a gene can be caused by the introduction of a stop codon (nonsense mutation), a SNV disrupting a splice site, insertions or deletions disrupting the transcript's reading frame, or large deletions completely removing the first exon or a large proportion of the protein-coding sequence of the affected transcript. These kinds of mutations, which can be tolerated if affecting less-conserved genes with related paralogs, cause a disease if they affect a gene linked with a Mendelian disease (MacArthur et al., 2012).

There is a middle ground between damaging protein-truncating variants and harmless synonymous variants: missense variants that cause a change in the amino acid sequence of the protein. The consequence of these variants can range from having no effect at all, to the development of a high penetrance phenotype (Chakravorty and Hegde, 2017). Our work focuses on the distinction of these cases.

1.1.4 Single amino acid variants and disease

Single amino acid variants are the most common type of variant affecting protein-coding genes, and their effect can diverge from absolutely innocuous to completely impeding the protein function.

The substitution of a single residue can hamper the normal function of the protein by a number of reasons. A mutation in a protein-protein interface can prevent the formation of multiprotein complexes (Jubb et al., 2017), a substitution in the active site can affect the protein's biochemical function (Daudé et al., 2013), mutations can prevent the folding of the protein to its proper conformation (Valastyan and Lindquist, 2014), or they can negatively affect the protein dynamics, eg. by hindering hinges in protein structure (Sayılgan et al., 2019).

The effects of SAVs on proteins couldn't be systematically studied until recently. Experimental advancements now allow for parallel mutagenesis of all variants in a protein (Kitzman et al., 2015; Starita et al., 2017), which is a great step forward compared to traditional approaches that targeted only one site at a time (Kunkel, 1985). These *in vitro* experiments have been successfully applied to the mapping of distribution fitness effects of some proteins (Firnberg et al., 2014), but the number of proteins studied are still a very small fraction of the human proteome.

The identification of SAVs causing a disease is an arduous task carried out by hundreds of different researchers and physicians around the world. The results of these efforts, usually published on research papers, have also been aggregated over the years in varied databases of annotated variants.

1.1.5 Databases of annotated variants

Different databases exist that collect variants which have been proven to cause a disease. Even though it has a broader scope, it is necessary to mention the Online Mendelian Inheritance in Man (OMIM, Amberger et al., 2015), as it represents the most complete resource of curated genes related to Mendelian disorders, with comprehensive descriptions on the phenotypes developed, references to relevant studies, and also variants known to affect the function of their gene. Having a more specific focus on the aggregation of variants linked to disease, the three most relevant databases are SwissVar, ClinVar and HGMD.

SwissVar (Yip et al., 2008), also known as Humsavar, is a catalogue of single amino acid variants in which each variant is labeled as disease-causing, polymorphism or unclassified. These annotations are derived from literature reports, and variants are related to dbSNP or OMIM entries when possible. The database is quite balanced between neutral and disease mutations (40,043 polymorphisms, 30,632 disease variants and 8,100 unclassified; as of July 2019), and it is published under the Creative Commons Attribution License (CC BY 4.0).

ClinVar (Landrum et al., 2016) is a database with a broader scope in terms of variation: it collects both germline and somatic variants from any region of the human genome, and of any length or type: insertions, deletions, copy number variations, SNPs, cytogenetic rearrangements... In addition to variant information extracted from published reports, ClinVar also allows for the direct submission of variants, which account for 40% of its data. A lot of care is taken on linking annotations and their source, and so its records contain information about the submitter, the variation, the clinical condition, interpretation and evidence supporting it.

The Human Gene Mutation Database, (HGMD, Stenson et al., 2017) is another database collecting more than 200,000 germline gene lesions on 8,000 genes which underlie, or are closely associated with inherited human disease. These lesions, which include missense, nonsense or splicing substitutions, deletions, indels, etc. are extracted from scientific literature with the aid of text mining. It is worth remarking that HGMD does not contain neutral variants, its focus is on disease, and it classifies variants as either disease causing or possibly disease causing, depending on the evidence. HGMD has two versions of its database: a free version which is 3.5 years outdated, and a complete paid version.

The databases we have presented so far are mainly focused on germline mutations (ClinVar also contains somatic mutations, but these represent only about 1% of its records), and by linking single variants with a disease, they are mainly focused on monogenic or Mendelian diseases. But these are not the only mutations than underlie disease: somatic mutations have a prevalent role in some complex diseases, and specially cancer. Two of the largest databases collecting cancer mutations are those released by two consortium initiatives we presented above: the TCGA (The Cancer Genome Atlas) and the ICGC (International Cancer Genome Consortium). Their mutations are also aggregated in COSMIC (Forbes et al., 2015), a database that started in 2004 describing 4 cancer genes, and now contains information from more than 1.4 million samples, including more than 6 million protein coding variants.

Even though the size of these databases keeps increasing (SwissVar adds between 1,000 and 2,000 variants every year, and HGMD reports a steady increase of 17,000 records per annum), we are very far from having a complete catalogue of disease-causing mutations. As we commented earlier, most of the human DNA variation is rare and evolutionary recent; this is even more true for disease-causing variants, especially if they cause early-onset severe illness. The vast majority of disease causing variants identified in any sequencing assay will not be previously known or present in any of these databases, and the experimental determination of pathology is a difficult and costly process; which rises the need for a fast and cheap assessment of the pathology of single amino acid variants.

1.2 Single amino acid variants pathology prediction

In silico SAV's pathology predictors have been used in research settings for almost two decades. They provide a fast and cheap analysis of SAVs, enabling the researchers to prioritize the variants under study and focus their efforts on the SAVs that are most likely to cause a disease.

Before these methods were developed, the need of a preliminary assessment of the impact of an amino acid mutation was usually met by comparison of the physicochemical properties of the wild type and the mutated residue. Grantham scale (Grantham, 1974), which assigns a similarity score to each pair of amino acids based on chemical properties and correlated with observed substitution frequencies, was a relevant method for decades to approximate the impact of SAVs.

In the early 2000s, the foundations of SAVs pathology prediction were established in a series of publications which described ways of characterizing and discerning neutral from damaging mutations: Chasman and Adams (2001); Ng and Henikoff (2001); Sunyaev et al. (2001); Ferrer-Costa et al. (2002). Some of the teams behind these studies developed in the following years the first SAV pathology predictors: PolyPhen1 (Ramensky et al., 2002), SIFT (Ng and Henikoff, 2003), PMut (Ferrer-Costa et al., 2004). Interestingly, each of these methods used a different approach to predict the pathogenicity of variants:

PolyPhen1 is an ad-hoc method, based on expert, experimentally derived rules; SIFT is based on a sequence conservation; and PMut is a machine learning method trained on variants previously classified as neutral or pathogenic.

Since then, dozens of predictors have been developed and published. In the following sections we will present the main paradigms that underlie these predictors.

1.2.1 Conservation scores

The most crucial insight for assessing the deleteriousness of SAVs is the fact that disease-causing variants suffer a negative selection pressure, and thus are depleted from homologous sequences in other organisms. With this central idea was developed SIFT (Sorting Intolerant From Tolerant, Kumar et al. (2009); Ng and Henikoff (2003); Sim et al. (2012)), a predictor that assigns a conservation score to each mutation and labels them as deleterious when this score reaches a threshold. Despite the age and simplicity of the method, it is still widely used to predict pathology, to build metapredictors based on its scores, and as a baseline to compare new methods. Needless to say, the performance of the method is highly dependent on the multiple sequence alignment (MSA) used to compute the score.

Leveraging on the Hidden Markov Models (HMM) used to construct the alignments of the protein families database PANTHER (Protein Analysis Through Evolutionary Relationships, Thomas et al., 2003), the authors used the position-specific amino acid probabilities in the model as a measure of conservation, where less likely substitutions are classified as deleterious. This same approach was followed to compute the Log.R E-value (Clifford et al., 2004), using the HMMER (Eddy, 1998) software on the Pfam (Finn et al., 2016) motifs database. More recently, the FATHMM (Shihab et al., 2012) predictor applied this concept using newly available sequence data.

Other methods, such as MAPP (Multivariate Analysis of Protein Polymorphism, Stone and Sidow, 2005) and PASE (Prediction of AAS Effects, Li et al., 2013) have opted to dug into the physicochemical properties of amino acids in order to identify subtler conservation patterns. MAPP measures the deviation of the mutated amino acid with respect to the aligned position in terms of its hydropathy, polarity, charge, side-chain volume and free energy in α-helix and β-sheet formation. PASE relies on the conservation of 7 different physicochemical properties of amino acids: transfer of free energy from octanol to water, van der Waals volume, isoelectric point, polarity, frequency of turn, frequency of α-helix and free energy of solution in water.

Other elaborate statistical approaches have been used for deriving conservation scores from multiple sequence alignments, such as LRT (Likelihood Ratio Test, Chun and Fay, 2009); MutationAssessor (Reva et al., 2011), with its Functional Impact Score, based on entropy measures and subfamilies conservation; and PROVEAN (Protein Variation Effect Analyzer, Choi and Chan, 2015; Choi et al., 2012), with the particularity that it is able to score in-frame insertions and deletions in addition to single amino acid variants.

PANTHER-PSEP (position-specific evolutionary preservation, Tang and Thomas, 2016) uses a score based on evolutionary preservation (Marini et al., 2010). Closely related to conservation, preservation measures how long a site has been in the same state in its ancestors. PANTHER-PSEP uses phylogenetic trees and MSAs from the PANTHER database, together with the reconstruction of common ancestors to give an approximate measure of the time the site has been preserved through evolution.

Sequence co-variation

Co-variation conservation scores are a natural evolution of point-mutation scores: they account for variation trends of pairs of amino acids. These methods, which are more complex and computationally intensive, have thrived thanks to the availability of more sequences, and have led to great results in the prediction of amino acid contacts and the modelling of proteins 3D structure (Morcos et al., 2011; Weigt et al., 2009). They have recently been shown to better reproduce the experimental fitness landscapes (Cheng et al., 2016; Figliuzzi et al., 2016; Flynn et al., 2017; Louie et al., 2018), and to improve pathology prediction (Feinauer and Weigt, 2017; Hopf et al., 2017; Nielsen et al., 2017).

When applied to pathology prediction, co-evolution conservation scores have the advantage of being able to detect compensated mutations. These mutations are substitutions that would seem to be deleterious when studied individually, but which do not cause harm because another mutation on the same protein compensates that effect (Marín Sala, 2017). Co-evolution conservation analyses study the conservation patterns of each pair of amino acids in the protein, and thus are able to recover theses mutations that would be catalogued as pathogenic by traditional methods.

EVMutation (Hopf et al., 2017) is the first predictor based on co-evolution conservation or epistatic effect prediction, with promising results, as it compares well with other machine learning based predictors trained with richer data.

These methods, however, also have their drawbacks, as they require high quality alignments, and less conserved parts of protein sequences, or evolutionary young families may lack the data required for these methods to succeed.

1.2.2 Supervised learning

During the last two decades, we have seen a rise of machine learning. The combination of mathematical developments, tied with the increase in computing resources, storage capacity and connectivity, has allowed for unseen data collection and automated learning from this data. Machine learning has been successfully applied in many fields, and bioinformatics is, of course, one of them.

Some of the most relevant SAV pathology predictors based on machine learning methods are, in chronological order: PMut (Ferrer-Costa et al., 2005; Ferrer-Costa et al., 2004), nsSNPAnalyzer (Bao et al., 2005), LS-SNP (Karchin et al., 2005), PhD-SNP (Capriotti et al., 2006), SNAP (Screening of non-acceptable polymorphisms, Bromberg et al., 2008), SeqProfCod (Capriotti et al., 2008), SNPs&GO (Calabrese et al., 2009), CHASM (Cancer-Specific High-Throughput Annotation of Somatic Mutations, Carter et al., 2009), MutPred (Li et al., 2009), PolyPhen2 (Adzhubei et al., 2010), MutationTaster (Schwarz et al., 2010), MuD (Mutation Detector, Wainreb et al., 2010), CADD (Combined Annotation-Dependent Depletion (Kircher et al., 2014)), PON-P2 (Niroula et al., 2015), VIPUR (Variant Interpretation and Prediction Using Rosetta, Baugh et al., 2016), DEOGEN (Raimondi et al., 2016), PhD-SNP^g (Capriotti and Fariselli, 2017), PMut2017 (López-Ferrando et al., 2017), DEOGEN2 (Raimondi et al., 2017), Missense3D (Ittisoponpisan et al., 2019).

Some of these predictors are not only limited to SAVs, but are able to classify short indels (CADD), non-coding variants (FATHMM-MKL, PhD-SNP^{g}, CADD), synonymous variants (MutationTaster), or can differentiate gain-of-function and loss-of-function variants (MutPred).

We can better explain the differences between these methods by focusing on the key components of the supervised learning methodology: the training set, features and classifier. In the prediction of SAVs pathology, the training sets are lists of variants annotated as neutral or pathological, the features are numerical or categorical values that describe these variants, and classifiers are methods that infer the relationship between the features and the pathogenicity of variants.

Training set

The choice of a training set for these predictors has been greatly affected by the public availability of neutral and disease causing variants datasets. For example, the training of PMut required an exhaustive literature search in order to collect variants reported as pathological. The later publication of the SwissVar, HGMD, Clinvar or COSMIC databases (see section 1.1.5) has greatly eased this process, and most of the previous predictors rely on one of these databases as their training set.

About a decade ago, it was easier to find pathological variants for training than neutral variants because they were clinically more relevant and thus more studied. To balance the training sets, some predictors such as PolyPhen2, derived neutral variants by homology: amino acids found in the same position in homologous sequences were added to the training set as neutral variants.

The recent availability of sequences from large-scale sequencing studies such as ExAC has turned the situation around, and now many more neutral variants are available compared to deleterious ones. CADD adapted to this new circumstances and it is trained on a dataset of 14.7 million neutral variants from public variation datasets, and 14.7 million *de novo* simulated variants that are expected to have a more damaging effect, as they have not suffered selective pressure.

Features describing mutations

Features are a determining part of any predictor. By capturing the relevant information needed to discriminate neutral from pathological variants, they allow the classifier to implement this differentiation. Each of the previously named predictors uses a different set of features, but we can identify similar principles underlying all of them. We can group all features used in these categories: conservation scores, structural properties, functional and gene ontology annotations, physicochemical measures and protein-protein interaction information.

Almost all predictors rely on sequence conservation measures as part of their features (MuD is an exception, as it is solely based on structural features). These scores are generally computed following the methodologies already described in section 1.2.1, but sometimes a well-known score, such as SIFT, is directly used as a feature, as does nsSNPAnalyzer.

Structural properties, usually derived from manual annotation, but also from predictions when they are not available, are other common features. These include solvent accessibility, environmental polarity, secondary structure, flexibility, transmembrane helices, coiled-coil structures, stability, B-factor, intrinsic disorder..., some of which are used, for example, by nsSNPAnalyzer and LS-SNP. VIPUR predictor goes one step further and uses the Rosetta (Leaver-Fay et al., 2011) simulator to characterize the differences between the original and mutated proteins, as does Missense3D using the Phyre2 modelling software (Kelley et al., 2015).

In a similar fashion, functional annotation of amino acids or functional site predictions are used by some of these predictors. These features include: DNA-binding residues, catalytic residues, calmodulin-binding targets, phosphorylation sites, methylation sites, etc. In addition to functional annotations, Gene Ontology annotations are other common features used, for example, by SNPs&GO and PON-P2.

Predictors have also relied on the physicochemical properties of the wild type and the mutated residues. Some of these features such as hydropathy, polarity, charge and volume are used, for instance, in PON-P2, LS-SNP or classic PMut.

Finally, interaction network properties, such as the number of protein-protein interactions, or the protein's role in the interactome measured as the centrality or other graph metrics, are additional features that have been used in pathology prediction. For example, DEOGEN2 uses annotations on the protein residues known to participate in interactions as features for its training.

Oftentimes, the decision for choosing the features is done *a priori*, with some specific goal in mind. For example, some researchers decide to use structural features of the proteins, reducing the number of proteins on which the predictor can be applied, but hoping to capture information that is missing in the sequence. In other cases, authors decide to use manual annotations on protein function in order to boost the predictor performance, but accepting the performance will presumably worsen when applied to less studied proteins.

Feature selection Feature selection, that is, the selection of a subset of all computed features to train the predictor, is a key part of the design of a predictor. For example, PON-P2 is a predictor which uses 8 features, selected from a total of 622 computed features, to train a random forest classifier. The computation of lots of features and the selection of the most informative ones, allows for the best features to emerge in a less biased way. Selecting a small number of features has different advantages such as the prevention of overfitting and the obtaining of more efficient predictors (Hawkins, 2004).

Classifiers

The predictors we have named use a variety of methods as classifiers. The classifier choice is probably the least determining of the decisions when developing a pathology predictor; with the appropriate selection of features, typical classifiers behave with similar performance.

Nevertheless, different methods have been used, such as simple naive bayes classifiers (PolyPhen2, MutationTaster), shallow neural networks (PMut, SNAP), support vector machines (LS-SNP, SeqProfCod, CADD, Phd-SNP), and random forests (PON-P2, MutPred, MuD, nsSNPAnalyzer, CHASM). Each of these methods uses a different algorithmic approach to find the link between features and labels; check section 3.3 for a brief summary of these and other classifiers.

1.2.3 Meta-predictors

Given the variety and number of methods available, it is just natural that some meta-predictors have been developed, combining the predictions of other methods and trying to take advantage of their complementarity.

The first of such methods was CanPredict (Kaminker et al., 2007a,b), a method aimed at the distinction of driver and passenger mutations in cancer, which combined SIFT conservation scores, Log.R E-values from Pfam and Gene Ontology similarity scores using a random forest classifier.

Condel, (Consensus Deleteriousness, González-Pérez and López-Bigas, 2011), derives a consensus score combining the output of five popular predictors (MAPP, MutationAssessor, PolyPhen2, SIFT and PFam's Log.R E-value), which can then be used to classify variants as deleterious or neutral, and also to quantify the impact of the mutation.

Similar to CanPredict, MetaSNP (Capriotti et al., 2013) uses a random forest classifier based on the output of PANTHER, PhD-SNP, SIFT and SNAP; PredictSNP (Bendl et al., 2014) bases its prediction on a majority vote between 6 predictors: MAPP, PolyPhen1, PolyPhen2, SIFT, SNAP and Phd-SNP; and MetaSVM (Dong et al., 2015a) combines the output of up to 18 different predictors using a support vector machine.

Finally, M-CAP (Mendelian Clinically Applicable Pathogenicity, Jagadeesh et al., 2016), which is probably the most outstanding meta-predictor to date, focuses on the clinical application of its predictions, and so tries to reduce the number of pathological mutations that are classified as benign. With variants from HGMD and ExAC as a training set, it uses a gradient boosting tree classifier to combine the output of 9 predictors, 7 conservation scores, and 298 features derived from multiple sequence alignments.

1.2.4 Specific predictors

The predictors we have described so far target all human proteins in general, although some of them have a special focus on certain diseases, such as CanPredict and CHASM, which target cancer-related proteins. Even though they are general-purpose predictors, their performance on certain protein families is diverse (Leong et al., 2015; Li et al., 2014). In fact, we have shown that for some specific families, most of these predictors consistently report a bad performance (see Results section 4.1.5).

To improve the variant pathology prediction in some of these cases, several specific predictors have been developed, targeting specific proteins, genes or protein families. Typically developed by research groups with a special focus on a given disease, and with more training data than that publicly available, these predictors have shown their ability to improve the performance of other general predictors by detecting singular functional trends (Crockett et al., 2012; Riera et al., 2016; Torkamani and Schork, 2007).

As expected, sequence conservation plays a key role in all specific predictors. This is the case of SAVER (Single Amino Acid Variant Evaluator, Adebali et al., 2016), an ad-hoc method for classifying mutations related to Niemann-Pick disease. It is based on the manual analysis of multiple sequence alignments and phylogenetic trees, from which rules are derived to classify variants.

As in general-purpose predictors, supervised-learning methods are the most popular approach for building specific predictors, heavily relying on conservation features enriched with additional, often manual, annotations. Some examples are KvSNP (Stead et al., 2011), based on random forests and specialized on mutations in voltage-gated potassium channels; HApredictor (Hamasaki-Katagiri et al., 2013), a decision tree predictor for Coagulation Factor VII mutations causing Hemophilia A disease; MutaCYP (Fechter and Porollo, 2014), a neural network predictor for variants in Cytochrome P450 proteins; wKinMut-2 (Vazquez et al., 2015, an evolution of KinMut, Izarzugaza et al., 2012), a random forest predictor targeting protein kinases mutations; and the neural network based predictor built by Riera et al. (2015) for Fabry disease related variants.

We also find an example of meta-prediction in the work of Leong et al. (2015), who combine the output of 5 popular predictors (SIFT, PoplyPhen2, PROVEAN, SNPs&GO and SNAP) to create a meta-predictor targeting Long QT syndrome related genes.

1.2.5 Estimating the performance of predictors

Estimating the accuracy of predictors is still a big challenge. The metrics used to evaluate predictors are clear (see Materials and methods section 3.4.2), but having a fair comparison of methods is not straightforward.

Newer methods always have the advantage of having more data available. In the case of conservation scores, this means larger sequence databases, and in the case of supervised learning predictors, larger training sets. Rarely old methods are evaluated using newer data in order to get a fair comparison. In the case of machine learning predictors, it is also difficult to find blind test sets disjoint with the training set of each predictor, even more if we want these test sets to be big enough as to not be biased.

Recent initiatives such as the CAGI (Hoskins et al., 2017) challenges are a great opportunity to create a level playing field for comparing predictors. Similar to well established initiatives like CAPRI (Janin et al., 2003) in the protein-protein interactions field, CAGI proposes challenges to the research community related to the prediction of variants effects on function. These effects, which are privately studied by the organizers in experimental assays, allow for a fair ranking of the predictions submitted to the challenge.

2. Objectives

The main objective of this work is to contribute to the understanding of the functional implications of single amino acid variants (SAVs) in proteins, with specific interest in their influence in the development of disease. Part of the work is based in our previous experience in developing PMut, a well known predictor of the pathological consequences of SAVs.

To progress in such general objective we shall address the following specific questions.

1. Building, a new, revised and more powerful version of the classic PMut predictor. It is our goal to update the machine learning training data set and methods with presently available data and state-of-the-art algorithms. Automation of the training and validation process is essential, as it allows the generation of tailor-made classifiers, and opens the possibility to explore other scenarios like cancer, not previously available in PMut.

2. Development of protein-specific predictors and assessment if this is a valid approach to improve the predictions of general-purpose predictors like PMut. We want to build them using an automated standard pipeline which should allow us to target as many proteins as possible.

3. Creation of a new PMut web portal, making the new predictor accessible as a general-purpose single amino acid variant analyzer, putting in place the previous achievements. The website must include a training facility to prepare specific classifiers and a complete repository of precomputed data including human sequences and their known variants. To smooth user experience, compatibility and integration with common formats and tools should be provided.

3. Materials and methods

> Essentially, all models are wrong, but some are useful.
>
> George E. P. Box

The PMut2017 mutation pathology predictor and PMut-S (Specific PMut), a suite of protein-specific predictors, are two of the main results of this thesis. We have approached the prediction of pathology in single amino acid variants as a supervised learning problem. As such, it consists of the typical steps of a machine learning pipeline: the training data set collection, feature computation and selection, classifier training, evaluation and comparison to other predictors. In this chapter we present the materials and methods that underlie each of these steps. Then, we explain the technologies used to bundle and release all this methodology as a Python software module (PyMut) and made it accessible through a web interface: the PMut web portal.

3.1 Training sets

Our main source of annotated variants (as either neutral or pathological) is SwissVar (Yip et al., 2008). For the PMut2017 predictor training we used the October 2016 release, which included 27,203 disease and 38,078 neutral mutations on 12,141 proteins. For the training of PMut-S predictors we used the release from April 2018, which contained 28,790 disease and 38,490 neutral mutations on 12,234 proteins.

Protein-specific training sets

Mutations in SwissVar are not evenly distributed among all proteins. In fact, only 1,000 proteins account for more than 90% of all the pathological variants, and most of the neutral variants belong to thousands of other proteins, of which no disease-causing variant is reported.

For training PMut-S, it was necessary to have a significant amount of variants on a single protein, and it was desirable that the number of neutral and deleterious variants be balanced, to obtain more robust predictors. We decided to select proteins with more than 30 disease causing variants in SwissVar (215 in total), and balance this training set with variants reported in ExAC (Lek et al., 2016) —which we considered neutral—, picking the most common variants first until they matched the number of pathological mutations.

Blind test sets

We performed two different blind tests on PMut2017. First, by training the predictor with an older release of SwissVar (December 2015), and evaluating it on variants reported in a newer release (December 2016), that is, a total of 3,166 variants (1,656 pathological and 1,510 neutral) on 762 proteins. Second, we used ClinVar's (Landrum et al., 2016) variants that were missing in SwissVar, 20,308 in total.

3.2 Features

After collecting a training set, the next step of building a predictor is the calculation of features that describe the variants. The machine learning classifier will try to differentiate neutral from pathological variants by identifying particular feature trends. Set to build a powerful predictor, we decided to compute a wide variety of features (215 in total), and then design and apply an automated feature selection algorithm to choose the most relevant ones.

The decision was made not to use categorical annotations from databases (such as Gene Ontology or functional annotations from UniProt) as a way of building a more general predictor avoiding a bias towards better-known proteins.

The first 8 features we computed account for physicochemical properties of the amino acids. We computed the absolute and relative change in volume (Chothia, 1975), hydrophobicity (Wimley and White, 1996), free energy transfer octanol-water (Fauchère et al., 1988) and Kyte-Doolittle hydropathy index (Kyte and Doolittle, 1982). Additionally, we took the amino acid position in the protein sequence as another feature.

Second, we used 5 different substitution matrix scores for the mutation, taken from the matrices BLOSUM50, BLOSUM62, BLOSUM80 (Henikoff and Henikoff, 1992), PAM60 (Dayhoff et al., 1978) and Miyata (Miyata et al., 1979).

Third, we added 6 descriptors of the protein in the interactome graph (provided by Patrick Aloy's Structural Bioinformatics and Network Biology group in the Institute for Research in Biomedicine). These features were the degree (number of interactions), and five measures of graph centrality: betweenness, cross-clique, closeness, eigenvector centrality and degree centrality.

Finally, 196 other features account for sequence conservation, which we describe in the following section.

3.2.1 Conservation features

Sequence conservation features are the most informative descriptors available to predict the pathology of protein mutations. Virtually every predictor published relies directly or indirectly in protein sequence conservation data. Being it so important, we decided to calculate sequence conservation using different methods in order to get as much as possible out of it.

Similar sequences search

To characterize sequence conservation, first we need to find sequences similar to the one of interest. We did this by searching the UniRef100 and UniRef90 cluster databases (Suzek et al., 2015) using PSI-BLAST (Position-specific iterated BLAST, Altschul et al., 1997) with a limit of 10,000 results and an E-value of 10^{-5}. Even though doing two searches may seem redundant, we found that due to the huge size of current sequence databases, many searches on UniRef100 reach the 10,000 sequences limit, and it was useful to search also UniRef90 in order to find a wider variety of similar proteins.

Multiple sequence alignments

The previous searches are local (small chunks of the original protein are found in other proteins) and matched pairwise. To get more relevant evolutionary conservation information, we retrieved the full sequences of each protein found in the search, and built a multiple sequence alignment. We used Kalign2 (Lassmann and Sonnhammer, 2005) for this purpose, as we found it best to handle big alignments after evaluating MUSCLE (multiple sequence comparison by log-expectation, Edgar, 2004), T-Coffee (tree-based consistency objective function for alignment evaluation, Notredame et al., 2000), and MAFFT (multiple alignment using fast Fourier transform, Katoh and Standley, 2013).

Sequences filtering

At this point, we had 4 different alignments —one local search, one MSA, over UniRef90 and UniRef100. Still, we believed we could get more out of them if we filtered each of these alignments following different criteria.

As we said before, we often hit the 10,000 sequences limit, specially when searching UniRef100. This means that in some cases we would only compare our protein to other very similar proteins, and in other cases we would compare it to farther evolutionary relatives. To have a more consistent comparison, we chose two E-value thresholds (10^{-75} for UniRef100 and 10^{-45} for UniRef90), which reduced the size of the alignments and kept most of them under the 10,000 limit.

In another vein, we also filtered alignments to either keep only human proteins, or to reject all human proteins. By rejecting human proteins we can get rid of paralogs, which may not share the same function of our target protein and thus not reflect the desired conservation we want to capture. On the other hand, by keeping only human proteins we may keep sequences that are evolutionarily closer.

Column features

In the end, given an aligned column of amino acids coming from any of the previous alignments, we need to compute some numerical feature to describe the feasibility of the substitution. We computed:

1. The number of sequences in the alignment.
2. The number of amino acids in the aligned position (number of sequences minus gaps in the column).
3. The total and relative number of wild type amino acids in the aligned position.
4. The total and relative number of mutated amino acids in the aligned position.
5. The Position Weight Matrix score, defined in Eq. 3.1, where $\mathrm{Freq}[mt]$ and $\mathrm{Freq}[wt]$ are the relative presence of the mutated and the wild type amino acids in the complete Swiss-Prot (Boeckmann et al., 2003) database.

$$PWM_{wt\rightarrow mt} = \log\left(\frac{\text{number of } mt}{\mathrm{Freq}[mt]}\right) - \log\left(\frac{\text{number of } wt}{\mathrm{Freq}[wt]}\right) \tag{3.1}$$

Finally, each of these 5 values was computed as is and in a weighted fashion. When weighted, we gave more importance to sequences which are more similar to the original sequence. In the case of PSI-BLAST searches, we weighted each value by the bit score of the match, and for multiple sequence alignments we weighted them by sequence similarity.

3.2.2 Features selection

Not all 215 features described above were used in the final PMut2017 model, only 12 of them were selected by running an automated feature selection algorithm, described in length in the Results section 4.1.2.

3.2.3 PMut-S features

We used 11 out of 12 features from PMut to train PMut-S (we discarded 1 feature that had the same value for all variants in the same protein), and we also added 4 features from the EVmutation model (Hopf et al., 2017), which is based on multiple sequence alignments of Pfam domains (Finn et al., 2016).

These 4 EVmutation features consist of a classic column conservation measure, the frequency of the substitution, the effect according to an independent conservation model and the predicted effect by an epistatic model, which adds pairwise epistatis to the estimation of the substitution effect. These features were only available for 51% of the variants used for PMut-S training, with the rest of the variants lying in parts of the sequence that were not covered by the Pfam alignments.

3.3 Classifiers

We did not have any preference *a priori* on which classifier to use, so different machine learning methods were considered for powering our predictors:

Gaussian naive Bayes Simple classifier based on the Bayes Theorem under the assumption that features are independent and follow a Gaussian distribution. A simple and fast predictor that can perform well in some cases but is usually used as a baseline to compare to other methods.

Decision tree The prediction of the label associated with a given sample is decided by traveling a tree from the root to a leaf, choosing the child of each node based on some splitting rules. Some popular algorithms are CART (classification and regression tree, Leo et al. (1984)), ID3 (Iterative Dichotomiser 3, Quinlan (1986)) and its successors C4.5 and C5.0.

Logistic regression Linear model based on the composition of a linear function with the logistic function, a sigmoid or S-shaped function with many convenient properties.

Stochastic Gradient Descent (Robbins and Monro, 1951) Linear model that is fitted by minimizing the loss function iteratively, following at each step a randomly chosen subgradient of this function.

AdaBoost (Freund and Schapire, 1997) Short for adaptative boosting, AdaBoost is a meta-estimator that starts fitting a classifier (previously cited

CART, in our case). Then, it iteratively fits additional copies of the classifier, giving more importance at each step to the training samples that were incorrectly classified in previous iterations.

Random forest (Breiman, 2001) Collection of independent decision trees in which the prediction is decided by a majority vote on the trees' output. We describe them with more detail in the next section.

Extremely randomized trees (Geurts et al., 2006) Similar to random forest, it consists on a set of decision trees where a randomly chosen subset of features is used at each split in the trees. Instead of using the most discriminative threshold, the threshold chosen is the best from a random sample.

3.3.1 Random forests

As we will see in section 4.1.1, random forest was the classifier chosen to implement the PMut2017 and PMut-S predictors. It is thus interesting to describe in more detail how this method works.

A random forest predictor trains an ensemble of trees, and chooses its prediction by performing a majority vote on the output of each tree (Figure 3.1). Trees are different from each other due to two randomization steps in their construction:

1. The training set of each tree is a randomized sample (with replacement) of the original training set, and of the same size as the original set. This process is usually named bootstrapping.

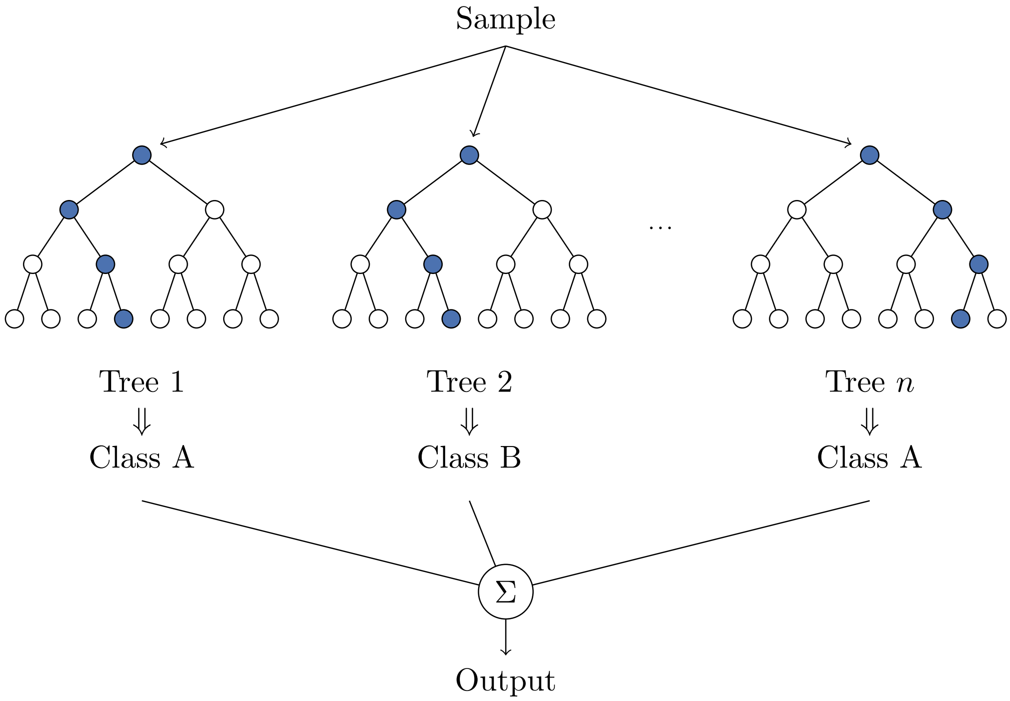


Figure 3.1 Random forest algorithm run scheme. The output is decided by a majority vote on the predictions of several decision trees.

2. At each node of the tree, only a random subsample of all the features are considered for choosing the split.

Each tree is built using the same algorithm, CART (classification and regression tree) in our case. This algorithm builds a tree, recursively splitting the sample in two, until the tree reaches a maximum depth or the number of samples in a node goes below a certain threshold.

At each node, the best split is chosen in a greedy fashion: by trying all possible splits for each of the features sampled. The quality of each split is measured by the Gini impurity measure (Gini, 1912), which is 0 in case of a pure sample (all elements have the same label), and 0.5 in the worst case (half the elements have one label, and the other half have the other).

For any given set of samples, the Gini index is measured as:

$$G = 1 - P[x = \text{Disease}]^2 - P[x = \text{Neutral}]^2 \tag{3.2}$$

The best split minimizes the weighted sum of the impurity of the two splits it generates.

3.3.2 Hyper-parametrization tuning

Each of the previously described methods depends on a set of parameters, which can greatly affect their performance. For example, the number of iterations in AdaBoost or logistic regression, the trees depth in random forests or extremely randomized trees, the loss function used in stochastic gradient descent, etc.

We did a randomized search on the parameter space of each of these predictors, performing a k-fold cross-validation (later explained) for each parametrization, and keeping the best for each classifier. This way, we were able to compare which classifier gave better predictions when performing at their best.

3.4 Model evaluation

When training a predictor it is as important to quantify the quality of the models as it is to build them. In this section we will discuss what strategies we followed to evaluate the models, and which metrics help us understand the performance of our predictors.

3.4.1 Cross-validation

The classic and most straightforward way of evaluating a predictor is by *k-fold cross-validation* (Figure 3.2), typically with $k \in [3, 10]$. This process consists in splitting the training set in k disjoint sets; then, each of the folds is predicted using a predictor trained with the other $k-1$ folds. Finally, the metrics that are computed for each fold (such as accuracy or sensitivity), are averaged. To make sure the evaluation is robust it is interesting to compute the standard deviation of the metrics across folds, as a big variance might signal undesired random behavior of the predictor.

In the case of unbalanced datasets, such as training sets with more neutral variants than pathological ones, we have used *stratified cross-validation*, which preserves the classes proportion across folds, yielding a more consistent evaluation.

To evaluate the PMut2017 predictor we used stratified cross-validation, but keeping a 50% sequence identity exclusion between folds, i.e., mutations on

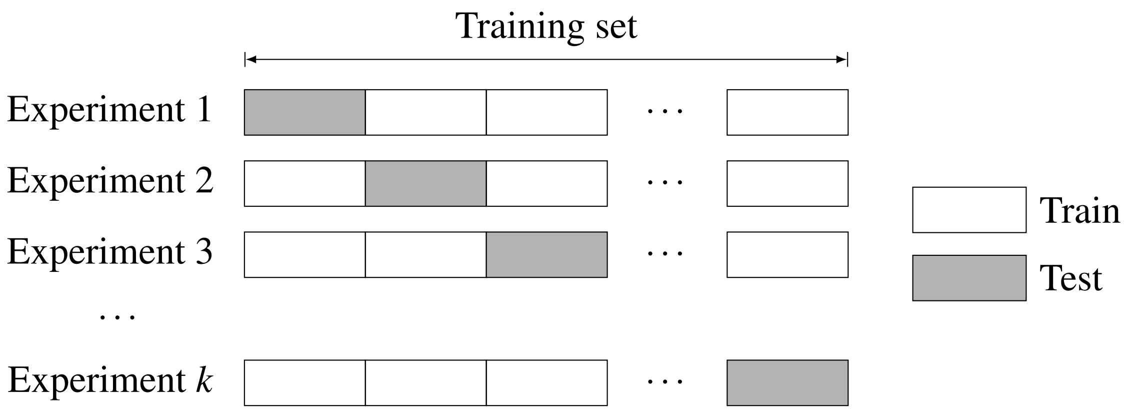


Figure 3.2 k-fold cross-validation. The training set is divided in k subsets; then, in k different experiments, a predictor is trained using all subsets except one, and the predictor is used to predict the labels for that missing subset.

similar proteins are grouped in the same fold. This way, we evaluated how the predictor performs when faced with variants from proteins families unknown to it, which is a stricter evaluation but closer to a real scenario.

In the case of PMut-S, where specific predictors are trained with only tens or hundreds of mutations, we resorted to *leave-one-out cross-validation* (Kearns and Ron, 1999), in which a predictor is trained for each variant using all variants except that one, and used to get that single prediction. The same methodology was applied to PMut2017 on these variants in order to get a fair comparison of the two predictors.

3.4.2 Model evaluation metrics

Here we present the metrics we used to evaluate our binary classification models. Dividing our samples by their label and their prediction, we can build

		Prediction outcome		
		p	**n**	**total**
Actual value	**p′**	True Positive (TP)	False Negative (FN)	P
	n′	False Positive (FP)	True Negative (TN)	N

Figure 3.3 Confusion matrix of a binary prediction model. Across all this work we have used the convention of considering pathological mutations as positives, and neutral mutations as negatives.

a confusion matrix (Figure 3.3). We will define most of our metrics based on P, N, TP, TN, FP and FN.

The simplest metric is *accuracy* (Eq. 3.3), which is the proportion of predictions that are correct —either positives or negatives.

$$\text{Accuracy} = \frac{TP+TN}{P+N} = \frac{TP+TN}{TP+FN+TN+FN} \tag{3.3}$$

Although accuracy is a very easy to grasp metric, it is also a limited one. For example, given a training set with 50% positives and negatives, a predictor that classified all samples as positives and another one that predicted all as negatives, would both have a 50% accuracy but be very different in nature. Even more, if the training set were unbalanced, with 90% positives and 10% negatives, a trivial predictor classifying all samples as positives would reach a 90% accuracy, giving the false impression of good performance.

These problems can be addressed computing *sensitivity* (also known as recall or True Positive Rate, Eq. 3.4) and *specificity* (False Positive Rate, Eq. 3.5). A high sensitivity predictor will be able to detect most of the positives correctly. A highly specific predictor will detect most of the negatives correctly.

$$\text{Sensitivity} = \frac{TP}{P} = \frac{TP}{TP+FN} \tag{3.4}$$

$$\text{Specificity} = \frac{TN}{N} = \frac{TN}{TN+FP} \tag{3.5}$$

In our case, we are interested in a predictor that has both high sensitivity and high specificity. This is better captured by the Matthews correlation coefficient (Equation 3.6), the most commonly used metric in the field, which favors this balance between TPR and FPR.

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{3.6}$$

Two other metrics of interest are the *positive predictive value* (PPV, also named *precision*, Eq. 3.7) and the *false predictive value*. The PPV is the proportion of samples predicted as positives which are effectively positive, and the NPV is the negative counterpart. We will use these metrics when evaluating the reliability of PMut2017's predictions (see section 4.1.3).

$$\text{Positive predictive value} = \frac{TP}{TP+FP} \tag{3.7}$$

$$\text{Negative predictive value} = \frac{TN}{TN+FN} \tag{3.8}$$

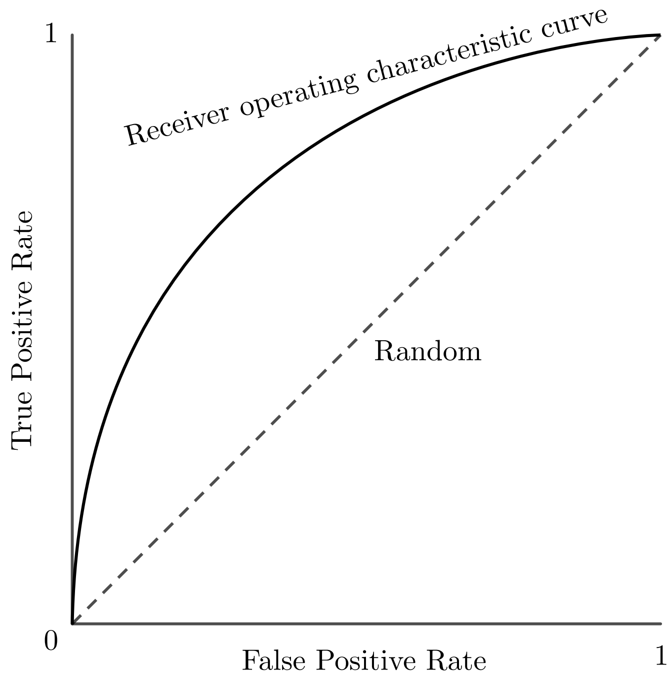


Figure 3.4 Receiver operating characteristic (ROC) curve. At different thresholds of the predictor, it plots the TPR with respect to the FPR. A random predictor would draw a line near $y = x$; the bigger the area under the ROC curve, the more accurate is the predictor.

Apart from these numerical metrics, we will also draw *receiver operating characteristic* (ROC) curves to compare models. A ROC curve plots the true positive rate in the y axis against the false positive rate at various threshold settings (Fig. 3.4). The higher the area under this curve, the better the predictions will be, and so the *area under the curve* (AUC) is another commonly used metric to evaluate predictors.

Although not a metric itself, we will often mention *coverage* when evaluating or comparing models. Coverage is the percentage of samples for which the predictor outputs a verdict.

Table 3.1 PyMut software dependencies.

Python module	Version	URL	Description
NumPy	1.10	numpy.org	Fast numerical computing library.
SciPy	0.17	scipy.org	Scientific computing library.
Pandas	0.17	pandas.pydata.org	Python data analysis library.
Matplotlib	1.5	matplotlib.org	Python plotting library.
Seaborn	0.8	seaborn.pydata.org	Statistical data visualization library.
Scikit-learn	0.17	scikit-learn.org	Machine learning methods.

These dependencies are documented in the official repository package (https://pypi.org/project/pymut) and are installed automatically by the standard Python package manager (pip).

3.5 PyMut Python module

The PyMut Python module is a software framework that provides all the functionality needed to train, evaluate and predict the pathology of protein mutations. It is based on the standard Python scientific stack, and Table 3.1 holds a list of all the dependencies and versions used. The full list of functions exported by the module is detailed in the Results section 4.3.

3.6 PMut web portal

PMut's predictions were made available via a rich web interface. This web portal is developed in Python, based on the Django (djangoproject.com) framework. Django, which embraces the Model-View-Controller pattern, helped us build a modular web application, where View components (HTML, CSS, JavaScript) are separate from the Controller (URL routing, form data processing...) and have an isolated Model, mapped to a MySQL relational database, implementing the bulk of the business logic. Being based in Python, the web portal could directly integrate all PyMut functionality, and use it to parse user input, generate plots and make calculations.

Figure 3.5 shows an overview of the software architecture powering the web portal. The web application runs on a virtual server, which uses Nginx to serve static assets and proxy dynamic pages to uWSGI, a supervisor that keeps a pool of processes running the Django application. The MySQL database used to store the jobs and user information runs in the same server.

A precomputed repository with millions of predicted variants is stored in a No-SQL MongoDB database running in another cluster of servers (the Results section 4.4.1 contains the details on the computation and contents of this repository). Finally, user submitted jobs are handled by a Sun Grid Engine job queue with three worker servers.

3.7 Other predictors

At different times in this work we have compared our predictions to those of other published predictors. When possible, we have accessed them via a web service, and have also relied on ANNOVAR (Wang et al., 2010), a database containing lots of precomputed predictions. Table 3.2 contains a list of the predictors we have evaluated and the source we used to retrieve their predictions.

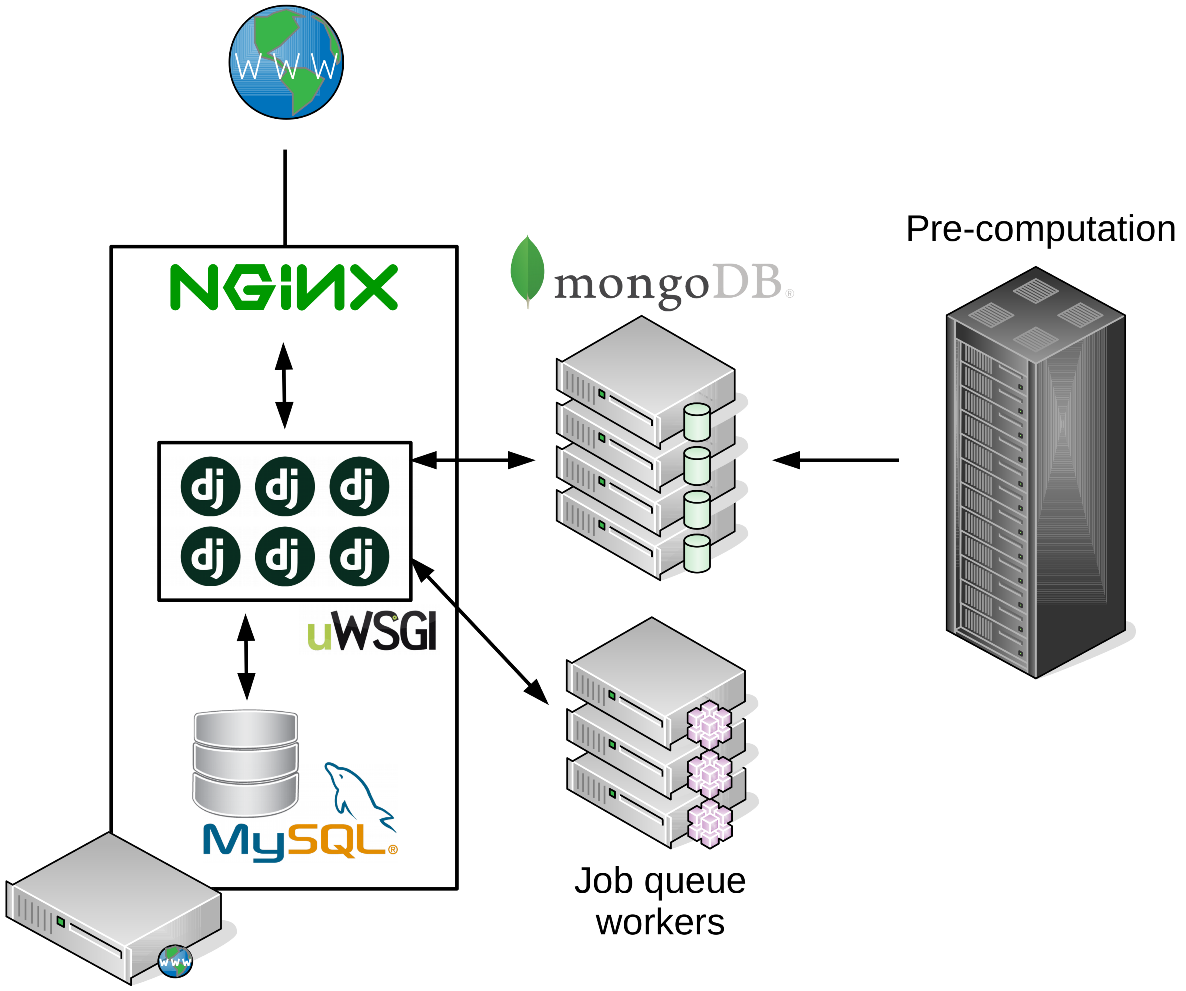


Figure 3.5 PMut web portal architecture. In a virtual server, the web application runs behind NGINX and uses a local MySQL database to store user and job information. It accesses precomputed repository data from a MongoDB cluster and send jobs to a job queue running on other virtual servers.

Table 3.2 External predictors.

Predictor	Reference	Source
Polyphen-2	Adzhubei et al. (2010)	genetics.bwh.harvard.edu/pph2/
PROVEAN	Choi and Chan (2015)	provean.jcvi.org
FATHMM	Shihab et al. (2012)	fathmm.biocompute.org.uk
PON-P2	Niroula et al. (2015)	structure.bmc.lu.se/PON-P2/
CADD	Kircher et al. (2014)	cadd.gs.washington.edu
M-CAP	Jagadeesh et al. (2016)	bejerano.stanford.edu/mcap/
Condel	González-Pérez and López-Bigas (2011)	bbglab.irbbarcelona.org/fannsdb/
SIFT	Ng and Henikoff (2003)	ANNOVAR
LRT	Chun and Fay (2009)	ANNOVAR
MutationAssessor	Reva et al. (2011)	ANNOVAR
MetaSVM	Dong et al. (2015a)	ANNOVAR
MetaLR	Dong et al. (2015a)	ANNOVAR
MutationTaster	Schwarz et al. (2010)	ANNOVAR

List of predictors we have used to compare our models and source from where we obtained them.

4. Results

We have divided the results chapter of this thesis in four sections. First, we describe PMut2017, an up-to-date general-purpose predictor, analyse what conforms the method and evaluate its performance. Second, we present PMut-S, the set of protein-specific predictors that complements PMut2017 and introduces new methods that improve the predictions. In the third section we describe PyMut, the Python module that is the computational foundation of all our work. Finally, we present the PMut web portal, which gives access to all the previous predictions in a handy and contextualized manner.

4.1 PMut2017 predictor

In this section we will detail how the PMut2017 predictor is built, the methods that conform the final model, evaluate its performance and compare it to other popular predictors.

4.1.1 Classifier choice

Once we settled on SwissVar as the training set for our predictor, and had 215 features describing these mutations, the next choice we had to make when developing PMut2017 was the machine learning method it would be based on. We evaluated many methods (described in section 3.3), and compared them using a classic 10-fold cross-validation. This comparison led to the comparisons in Figure 4.1 and Table 4.1. Each of these predictors was optimized using randomized hyper-parametrization tuning, and we compared the best performance of each of them.

Table 4.1 Performance comparison of classifiers for PMut2017.

Classifier	Accuracy	Sens.	Spec.	AUC	MCC
Random forest	0.81	0.75	0.86	0.81	**0.62**
AdaBoost	0.81	0.75	0.86	0.80	**0.61**
Extremely randomized trees	0.80	0.72	0.86	0.79	**0.59**
Logistic regression	0.77	0.70	0.83	0.76	**0.53**
Stochastic gradient descent	0.77	0.69	0.82	0.76	**0.52**
Gaussian naive Bayes	0.67	0.91	0.49	0.70	**0.42**

In this comparison we observed that the two predictors with better performance were AdaBoost and random forest. Both predictors are very similar in nature, as they both are ensemble methods based on decision trees. However, AdaBoost required a total of 258 trees to achieve these results, while only 13 trees were needed for the random forest. This difference made the random forest classifier the most efficient and convenient and was therefore chosen as the predictor for PMut2017.

4.1.2 Feature selection

The previous random forest predictor was trained with all 215 computed features (described in the Materials and methods section 3.2). Although we validated that the random forest classifier is resilient to overfitting, we were interested in reducing the number of features: the classifiers would be faster and we could save time by skipping the computation of uninformative features.

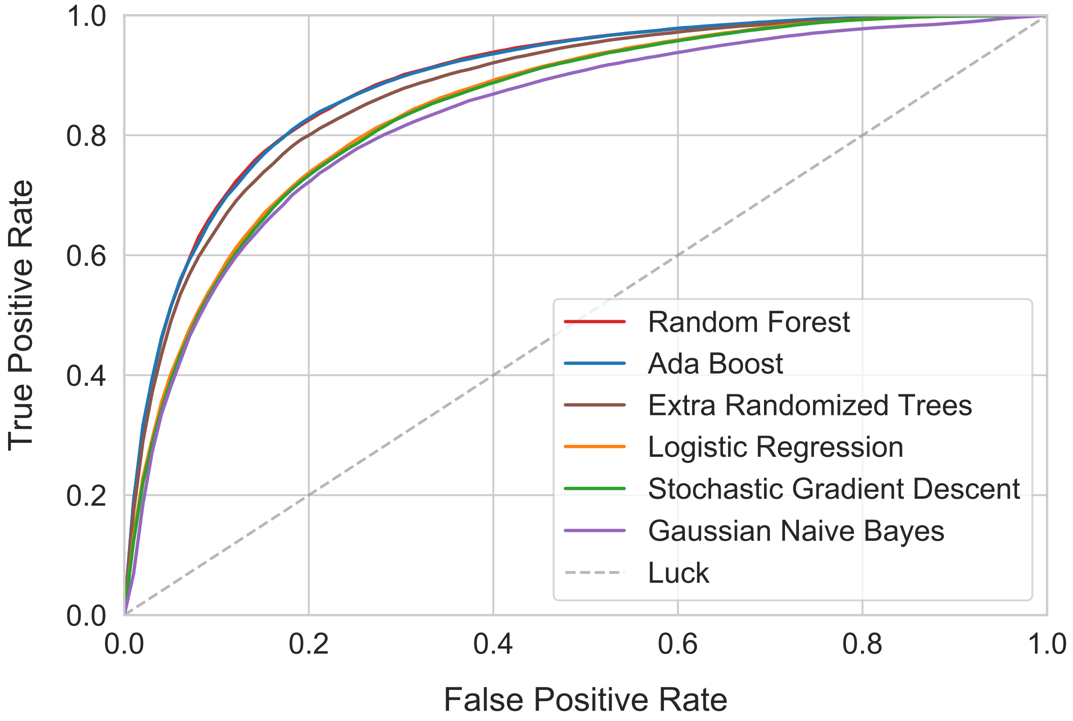


Figure 4.1 ROC curves comparison of classifiers for PMut2017. Random forest presents the best performance, closely followed by AdaBoost.

Table 4.2 PMut2017 random forest parameters.

Parameter	Value
Tree max depth	9
Number of features in split	50%
Minimum samples in leaf	10
Number of trees	13
Split criterion	Gini index

Most of our features were very correlated —specially all the conservation ones—, and we suspected that the boost in predictive power would come from subtleties from different alignments, rather than by choosing the least correlated features, as it is often done. For this reason we devised a simple iterative algorithm, described in Figure 4.2, which adds one feature at a time until the maximum performance is reached.

The run of this algorithm is summarized in Figure 4.3: compared to the predictor using all 215 features (in green), we see the increasing performance of the new model (in blue), when it is trained on $1, 2, 3$... features. Finally, when using 12 features, the predictor matched the performance of the predictor trained using all features.

Table 4.3 holds an account of the 12 features that were selected for the final model. Except for the Miyata substitution matrix score, all the features were derived from various alignments described in detail in section 3.2.1. The most important features were based on UniRef100 searches and alignments (78.4% importance in total), but 4 UniRef90-derived features, with a relative importance of 20.6%, were also selected. Also, we see that a variety of our different filtering criteria (such as limiting alignments by the BLAST E-value, or taking only human sequences) was automatically selected.

The distribution of these features is plotted in Figure 4.4, where it is clear that the most important features for the model are those that better separate neutral variants from disease causing ones.

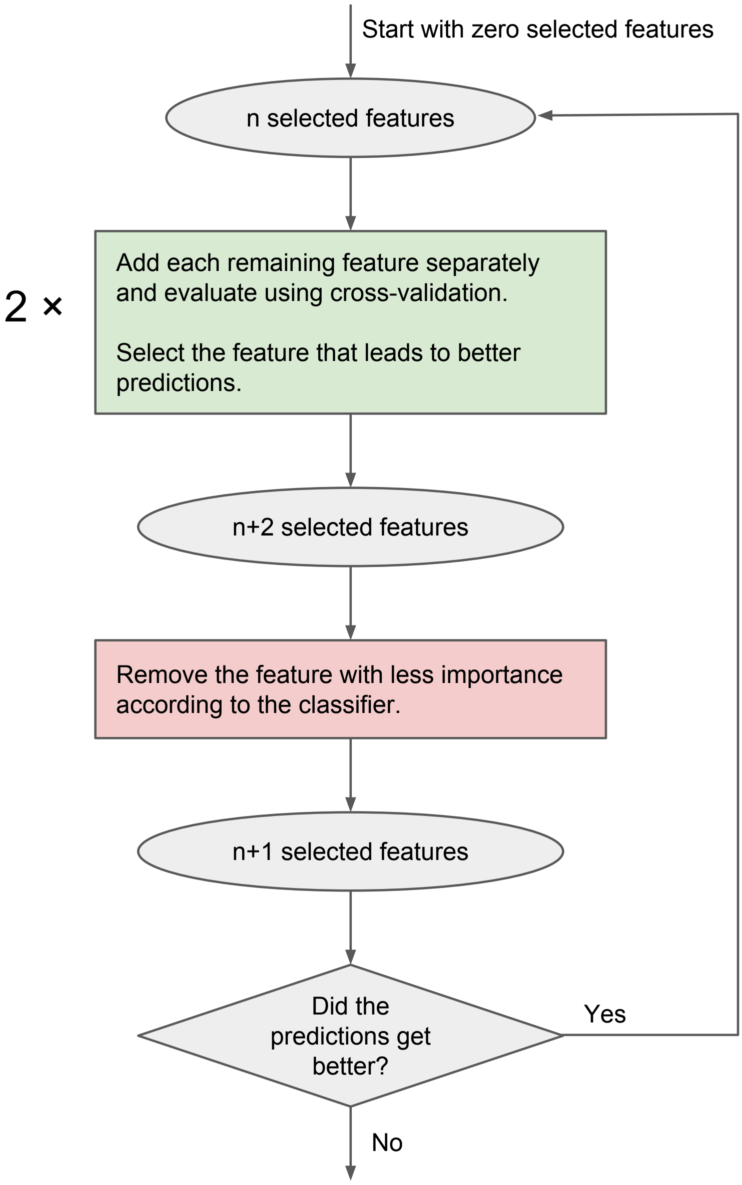

Start with zero selected features
n selected features
2 ×
Add each remaining feature separately and evaluate using cross-validation.
Select the feature that leads to better predictions.
n+2 selected features
Remove the feature with less importance according to the classifier.
n+1 selected features
Did the predictions get better?
Yes
No

Figure 4.2 Iterative feature selection algorithm. Features are added to the selected set until the performance increase in terms of the Matthews correlation coefficient is negligible. At each step, the two features that increase the MCC the most are added and then the least important feature is removed. This second step is designed to skip local minima in performance.

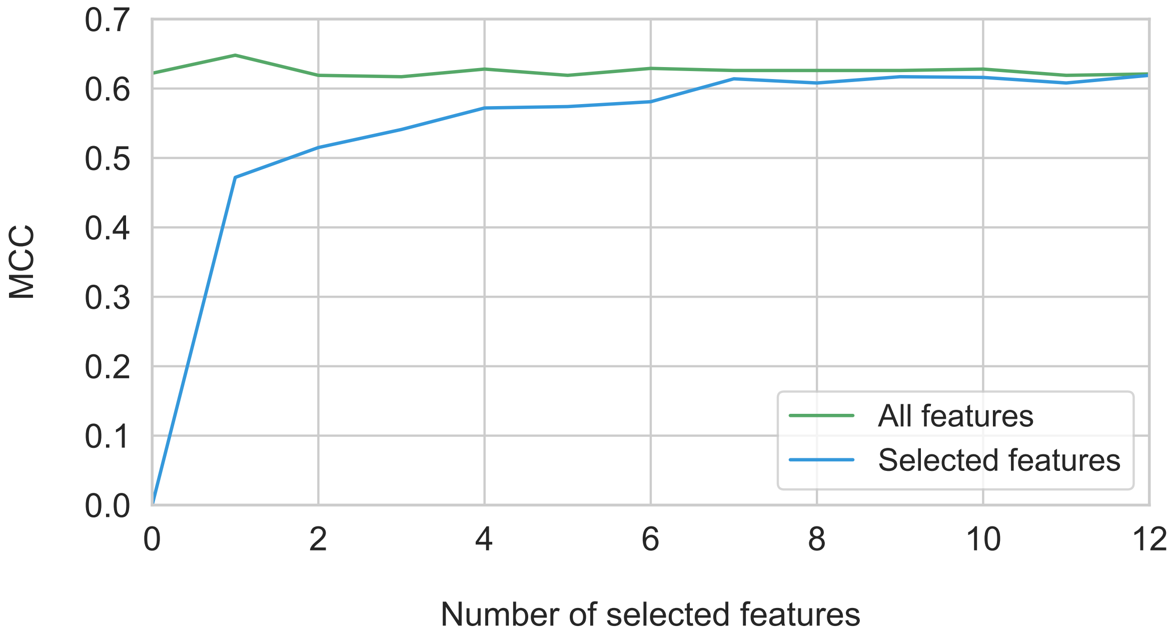


Figure 4.3 Feature selection algorithm run for the PMut2017 classifier. The predictor performance increased with each feature added to the selection, and it matched the performance of the predictor using all 215 features with 12 selected features. Note that the variation of the target MCC at each step is due to changes in the cross-validation folds, which are different and randomly chosen at each step.

Table 4.3 List of PMut2017 selected features sorted by importance.

#	Importance	Feature
1	24.9%	Position Weight Matrix score in the MSA over UniRef100 (weighted by sequence similarity).
2	24.1%	PWM score in the PSI-BLAST search over UniRef100 with E-value $< 10^{75}$ (weighted by E-value).
3	12.9%	Number of wild-type amino acids in the MSA over UniRef100 (weighted by sequence similarity).
4	11.2%	Number of wild-type amino acids in the aligned position of a MSA over UniRef90 (weighted by sequence similarity).
5	5.4%	Ratio of wild-type amino acids in the PSI-BLAST search over UniRef100 with E-value $< 10^{75}$.
6	4.5%	Number of sequences in the MSA alignment over UniRef100 search results.
7	4.2%	Number of amino acids in the aligned position of a PSI-BLAST search over UniRef100 with E-value $< 10^{75}$.
8	3.5%	Number of amino acids of human sequences in the aligned position of a MSA over UniRef90 (weighted by sequence similarity.
9	2.9%	Number of amino acids in the aligned position of a MSA over UniRef90 (weighted by sequence similarity).
10	2.9%	Number of amino acids in the aligned position of a PSI-BLAST search over UniRef90 with E-value $< 10^{45}$.
11	2.4%	Number of amino acids of human sequences in the MSA over UniRef100 search results.
12	1.0%	Miyata substitution matrix score.

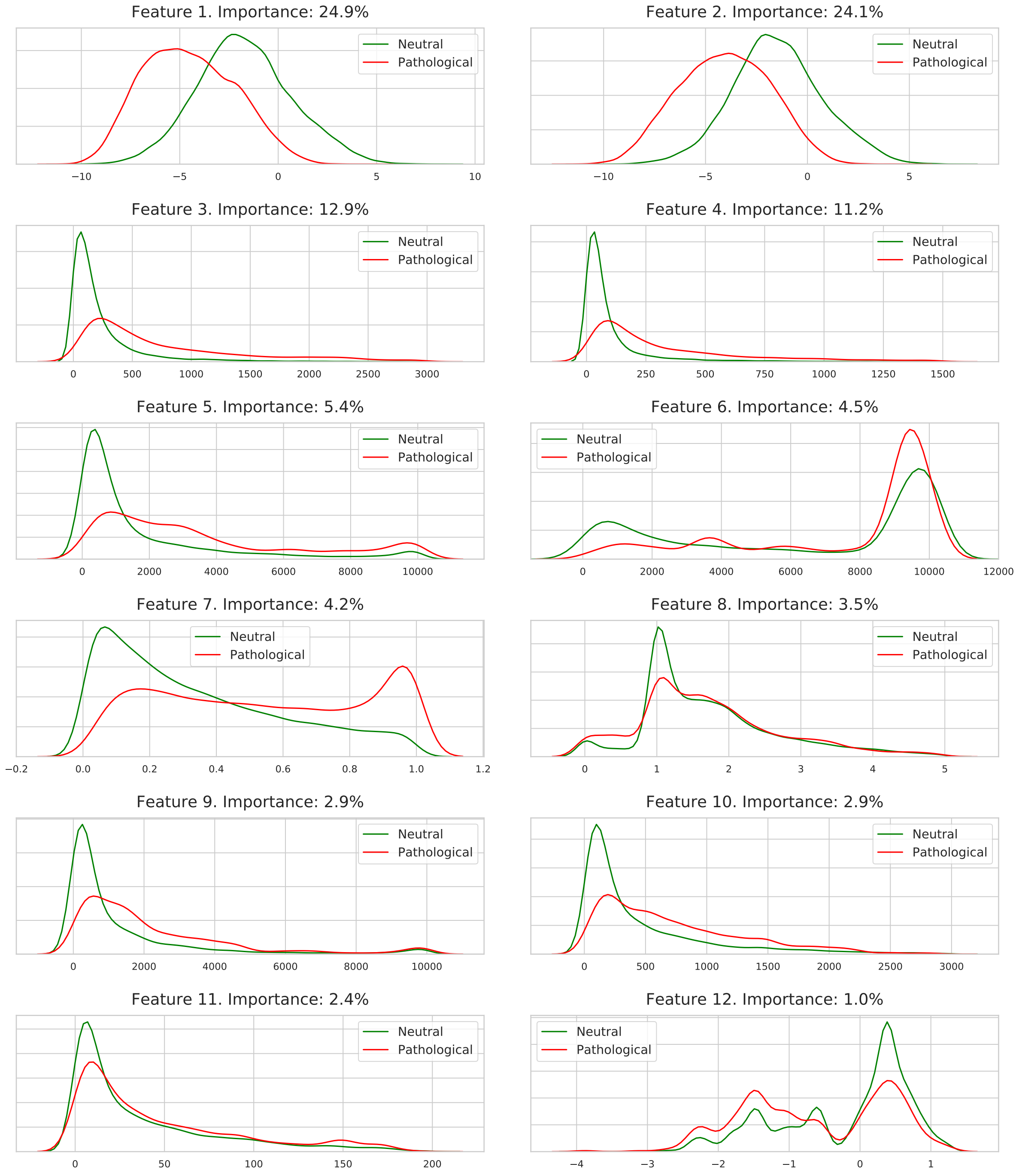


Figure 4.4 Distribution plots of the 12 features that configure the PMut2017 model for the 65,281 SwissVar variants used to train the model. The features are sorted by importance, and it is evident that the first 4 features (with a cumulative importance of 74.3%) are the ones more clearly separating neutral variants from pathological ones. See Table 4.3 for a description of the features.

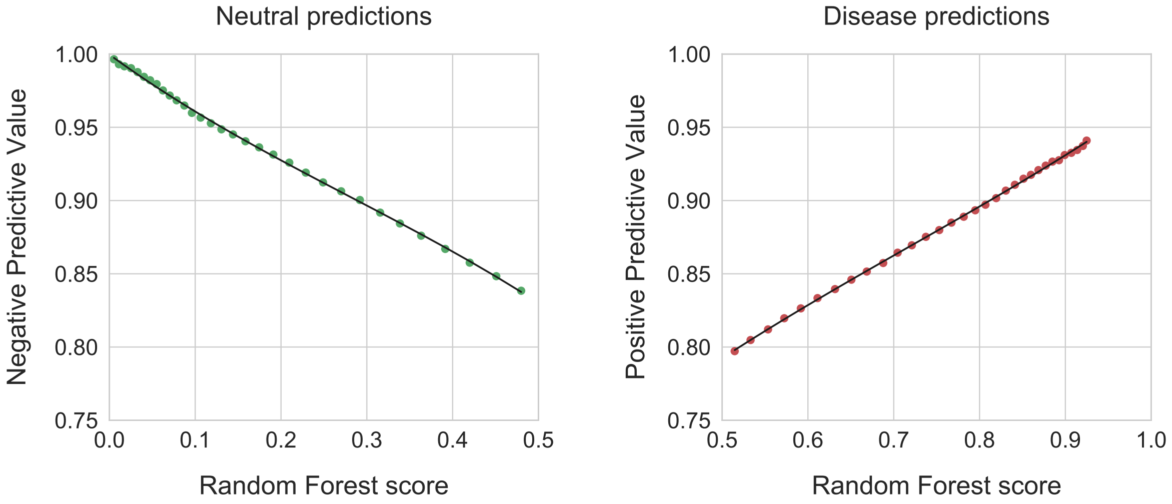


Figure 4.5 Reliability score regression for PMut2017. We plot the Negative Predictive Value (left), and the Precision or Positive Predictive Value (right) for different thresholds of the random forest score. Metrics are obtained from a 10-fold cross-validation with 50% sequence identity exclusion.

4.1.3 Prediction reliability score

The random forest classifier we have trained does not only provide a binary classification output, but a score in the range [0,1], where scores $<$ 0.5 are neutral predictions and scores $>$ 0.5 are pathological predictions. We wanted to see if we could map this score to some statistical measure of the reliability of the predictions, as this is very convenient when using predictors to prioritize further analysis of mutations.

We tackled this problem by analyzing the PPV and NPV (Positive/Negative Predictive Value) for different thresholds of the random forest score. The PPV (NPV) is the probability that a Positive (Negative) prediction is correct; if we saw that the PPV is higher when random forest scores are higher, we could output the PPV as a reliability measure of the prediction.

In Figure 4.5 we see the confirmation of this link between PPV/NPV and random forest score, which we mapped using an univariate spline regression. Later on, when we evaluate PMut2017 predictions, we will filter predictions with a reliability >85% or >90% and see how by limiting the coverage we can obtain more accurate predictions.

4.1.4 Performance evaluation

Once we had chosen our model —a random forest classifier trained with SwissVar mutations described by the 12 features from Table 4.3— we were ready to quantify the quality of its predictions and compare it to other methods.

Cross-validation

Our first evaluation was a 10-fold cross-validation, with each fold having a 50% sequence identity exclusion, that is, with similar proteins belonging on the same fold and the predictor being evaluated on proteins significantly different from the training ones. The numerical results of this evaluation are summarized in Table 4.5, where we also separate predictions with a higher reliability score. Figure 4.6 shows the ROC curves of this evaluation when we reduced the coverage of the predictor to the two thirds and one third more reliable predictions.

With the same cross-validation technique we also drew a *learning curve* (Figure 4.7), which shows us the improvement in accuracy we can expect by adding more variants to the training set. The training score in the plot, which is the MCC obtained when predicting the whole training set, is an upper bound of the MCC of the predictor.

Table 4.5 PMut2017 performance metrics in a 10-fold cross-validation.

Confidence	Coverage	Accuracy	Sensitivity	Specificity	AUC	MCC
All	100	0.82	0.76	0.86	0.81	0.62
$> 85\%^{*}$	85.9	0.85	0.75	0.92	0.83	0.69
$> 90\%^{*}$	64.9	0.90	0.80	0.95	0.87	0.77

* Only predictions with higher reliability scores were considered (see Figure 4.5).

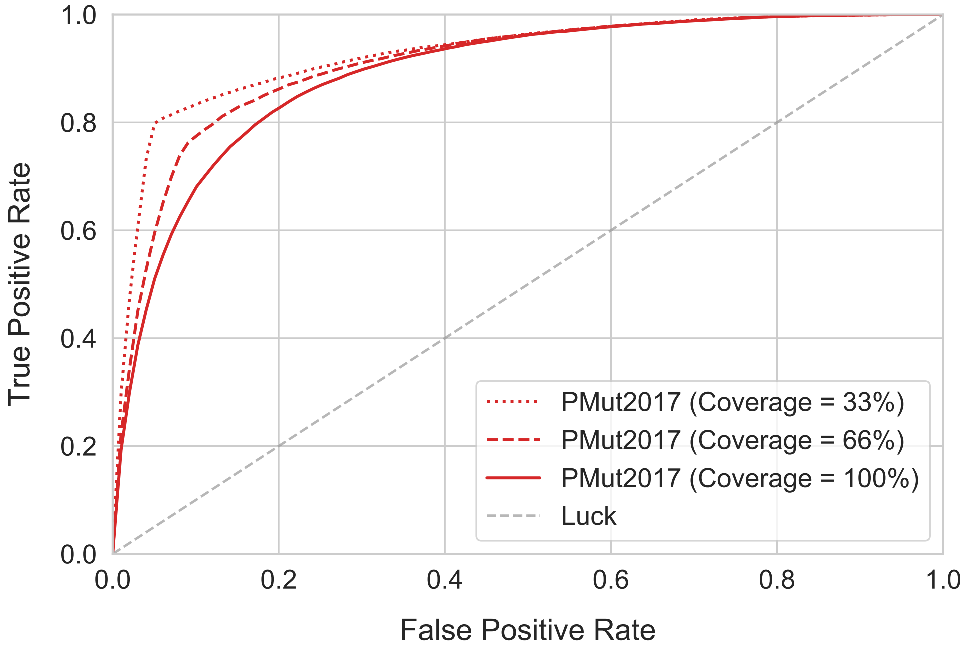


Figure 4.6 ROC curves for the PMut2017 10-fold cross-validation. No sequence in the validation set has more than 50% identity with sequences in the training set. Additional curves correspond to subsets with greater confidence in the predictions.

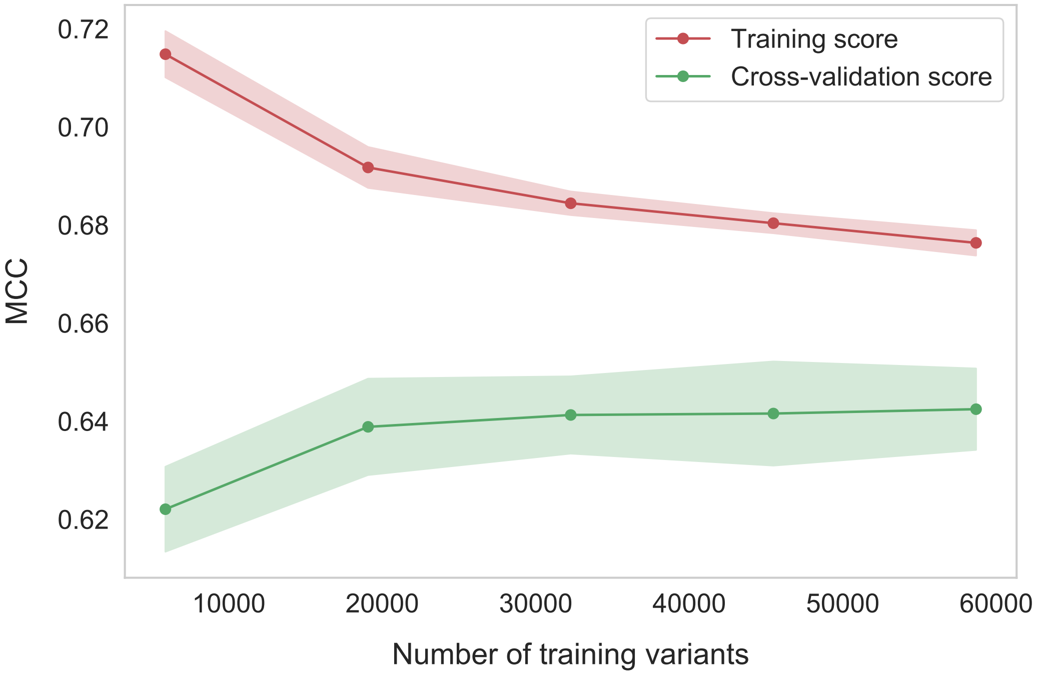


Figure 4.7 PMut2017 learning curve. Evaluation of the predictor for increasing training set sizes. The red line (training score) is the MCC obtained when evaluating the predictor on its own training set, and the green line is the average MCC in a 10-fold cross-validation. Colored areas around the lines are the standard deviation on these measures.
The training score represents an upper-bound on the performance of the predictor, and both lines serve as an indication of the improvement we can expect from expanding the training set.

Comparison to other predictors

The second evaluation performed was a blind test based on new SwissVar variants. We trained a predictor using the SwissVar database as of December 2015, and evaluated this predictor on the entries added to SwissVar during 2016. The results are shown in Table 4.6, where we also compare PMut to 13 other predictors. PMut showed one of the best performances (MCC = 0.42), slightly surpassed by LRT (MCC = 0.45) and PON-P2 (MCC = 0.47, but with a much lower coverage of 42.4%). In this evaluation we confirmed that PMut predictions with higher reliability score, effectively present better predictions. This evaluation has special interest because by using recently added variants to SwissVar we minimize the odds that these variants had been used for training the other methods in the comparison.

ClinVar blind test

We carried out a third evaluation on PMut2017, based on ClinVar mutations that are not present in SwissVar, a total of 20,308 variants (see Table 4.7). In this evaluation, PMut2017 reports an MCC of 0.49, which lies between the 0.42 and 0.62 that we obtained in the blind SwissVar test and the cross-validation, respectively.

4.1.5 Comparison on selected genes

To this point, PMut's performance was evaluated globally, but we are also interested in seeing how the predictor performs in specific cases. We evaluated

Table 4.6 PMut2017 comparison on SwissVar blind test.

Method	Coverage	Acc.	Spec.	Sens.	AUC	MCC
SIFT	89.6	0.61	0.33	0.88	0.60	0.25
Polyphen2	92.1	0.64	0.35	0.91	0.63	0.32
PROVEAN	91.5	0.64	0.41	0.87	0.64	0.31
FATHMM	90.5	0.55	0.45	0.64	0.55	0.09
PON-P2	42.4	0.72	0.52	0.9	0.71	0.45
CADD	95.0	0.65	0.33	0.94	0.64	0.35
M-CAP	91.5	0.60	0.19	0.95	0.57	0.22
Condel	91.0	0.63	0.40	0.84	0.62	0.26
LRT	95.1	0.73	0.58	0.87	0.73	0.47
MutationAssessor	95.1	0.63	0.46	0.78	0.62	0.26
MetaSVM	95.1	0.63	0.51	0.74	0.62	0.26
MetaLR	95.1	0.6	0.46	0.73	0.60	0.20
MutationTaster	95.1	0.65	0.31	0.96	0.64	0.36
PMut	**100.0**	**0.71**	**0.65**	**0.76**	**0.71**	**0.42**
PMut (85%)[*]	**81.0**	**0.76**	**0.76**	**0.77**	**0.76**	**0.53**
PMut (90%)[*]	**51.2**	**0.81**	**0.78**	**0.84**	**0.81**	**0.62**

Blind validation based on new variants added to SwissVar during 2016 (3,166 variants), CADD predictor has been evaluated using a threshold of 20.
[*] Analysis restricted to most reliable PMut predictions (reliability level in parentheses).

Table 4.7 PMut2017 blind test on ClinVar.

Coverage	Accuracy	Specificity	Sensitivity	AUC	MCC
100%	0.73	0.88	0.85	0.75	0.49

the predictor on 27 different genes, and compared PMut's performance to other predictors (Table 4.8).

This comparison showed how PMut's performance is not equal across different genes, as its MCC varies uniformly from 0.86 to 0.12, and even more interestingly, we observe that this trend holds for all the compared predictors. This was the result that boosted our interest on specific predictors and the development of PMut-S.

4.1.6 CAGI 5 participation

PMut2017's predictions were submitted to the CAGI 5 (Critical Assessment of Genome Interpretation) *Annotate all missense* challenge, which consisted on the pathology prediction of the 81,084,849 variants from dbNSFP. The evaluation of this challenge is still in progress; the predictions submitted will be assessed by comparing them with experimental annotations that will be added in the future on an ongoing basis.

Table 4.8 PMut2017 performance comparison on selected genes.

Gene	Disease	#D	#N	PMut	SIFT	Polyphen	LRT	Mut. Taster	Mut. Assessor	PROVEAN
MECP2	Rett syndrome	46	22	0.86	0.66	0.85	0.69	0.64	0.41	0.53
COL1A2	Osteogenesis Imperfecta	78	20	0.77	0.74	0.62	0.55	0.55	0.74	0.74
SLC4A1	Distal Renal Tubular Acidosis	38	36	0.69	0.65	0.65	0.54	0.55	0.68	0.60
ADAMTS13	Upshaw-Schulman syndrome	43	17	0.62	0.76	0.46	0.00	0.71	0.54	0.62
ATM	Hereditary cancer-predisposing syndrome	46	54	0.60	0.53	0.57	0.32	0.42	0.48	0.55
ATP7B	Wilson disease	195	25	0.48	0.34	0.49	0.37	0.43	0.29	0.52
MLH1, MLH2, MLH6, PMS2	Lynch syndrome	159	78	0.48	0.32	0.31	0.23	0.16	0.43	0.32
MYOC	Primary open angle glaucoma	57	24	0.47	0.37	0.45	0.38	0.50	0.47	0.49
TTC21B	Jeune thoracic dystrophy	16	28	0.42	0.20	0.22	0.18	0.16	0.28	0.26

Gene	Disease	#D	#N	PMut	SIFT	Polyphen	LRT	Mut. Taster	Mut. Assessor	PROVEAN
SCN5A	Brugada syndrome	154	46	0.40	0.32	0.26	0.43	0.31	0.34	0.34
KCNH2, SCN5A	Congenital long QT syndrome	270	54	0.38	0.32	0.28	0.36	0.32	0.30	0.38
ABCA1	Tangier disease	32	31	0.37	0.43	0.31	0.32	0.47	0.43	0.47
PKHD1, PKD1	Polycystic kidney disease	197	96	0.37	0.43	0.37	0.30	0.41	0.36	0.45
FBN1	Marfan syndrome	385	20	0.35	0.31	0.25	0.21	0.33	0.32	0.30
RYR1	Central core disease	147	25	0.34	0.27	0.31	0.00	0.36	0.28	0.34
LDLR	Familial hypercholesterolemia	103	23	0.32	0.29	0.08	0.17	0.09	0.26	0.25
DYSF	Limb-Girdle Muscular Dystrophy	48	16	0.31	0.35	0.27	0.15	0.21	0.41	0.39
BRCA2	Breast-ovarian cancer, familial 2	43	61	0.31	0.10	0.18	0.18	0.14	0.19	0.01
BRCA1	Breast-ovarian cancer, familial 1	27	36	0.31	0.24	0.20	0.38	0.29	0.30	0.17
WFS1	WFS1-Related Spectrum Disorders	40	17	0.30	0.25	0.35	0.20	0.18	0.16	0.26

Gene	Disease	#D	#N	PMut	SIFT	Polyphen	LRT	Mut. Taster	Mut. Assessor	PROVEAN
PINK1	Parkinson Disease	23	39	0.25	0.33	0.48	0.40	0.41	0.44	0.30
LRRK2	Parkinson Disease	21	24	0.19	0.06	0.14	0.01	0.13	0.09	0.14
CFTR	Cystic fibrosis	146	32	0.15	0.06	0.20	0.21	0.12	0.20	0.27
PROC	Thrombophilia	36	28	0.12	-0.15	-0.08	0.07	0.14	0.08	-0.01

4.2 PMut-S

The second result we present in this section is PMut-S (Specific PMut), a collection of 215 protein-specific predictors. We will start presenting the features that form these predictors, to then compare their performance to other specific and general-purpose predictors. We will complete this analysis by checking the predictions of common variants from healthy people.

4.2.1 Training sets

Each of the predictors in PMut-S is based on a training set derived from Swiss-Var. We selected those proteins having more than 30 disease variants reported there, and complemented each training set with neutral variants extracted from ExAC (see section 3.1 for more details).

4.2.2 Features

As commented in the Materials and methods section 3.2.3, PMut-S is based on 11 features from PMut2017 (Table 4.3) and 4 additional features from EVmutation (Hopf et al., 2017). In Table 4.9 we describe these features, sorted by descending importance.

It is important to note that EVmutation's features are only available for 51% of the variants, and still they are very relevant. Interestingly, the most important of EVmutation's feature is the epistatic model prediction, the one that *a priori* seemed more promising as it involves a richer model than simple column conservation.

Table 4.9 List of PMut-S features sorted by importance.

#	Importance*	Source	Feature description
1	10.0%	PMut	Number of wild-type amino acids in the MSA over UniRef100 (weighted by sequence similarity).
2	10.0%	PMut	Number of wild-type amino acids in the aligned position of a MSA over UniRef90 (weighted by sequence similarity).
3	9.5%	PMut	PWM score in the PSI-BLAST search over UniRef100 with E-value $< 10^{75}$ (weighted by E-value).
4	9.3%	PMut	Position Weight Matrix score in the MSA over UniRef100 (weighted by sequence similarity).
5	8.6%	PMut	Ratio of wild-type amino acids in the PSI-BLAST search over UniRef100 with E-value $< 10^{75}$.
6	6.5%	PMut	Number of amino acids in the aligned position of a PSI-BLAST search over UniRef90 with E-value $< 10^{45}$.
7	6.4%	EVmutation	Epistatic model prediction.
8	6.3%	PMut	Number of amino acids in the aligned position of a PSI-BLAST search over UniRef100 with E-value $< 10^{75}$.
9	5.9%	PMut	Number of amino acids in the MSA over UniRef90 (weighted by sequence similarity).
10	5.4%	EVmutation	Independent conservation model prediction
11	5.2%	EVmutation	Frequency of the substitution.
12	4.3%	PMut	Miyata substitution matrix score.
13	4.2%	EVmutation	Column conservation.
14	3.9%	PMut	Number of amino acids of human sequences in the aligned position of a MSA over UniRef90 (weighted by sequence similarity.
15	3.8%	PMut	Number of amino acids of human sequences in the aligned position of a PSI-BLAST search over UniRef100.

* The importance is the average importance that each of the 215 predictors assigns to each feature.

4.2.3 PMut-S versus PMut

We started our analysis of PMut-S performance by comparing it to PMut. This comparison was specially interesting because we could evaluate PMut using the same leave-one-out methodology and were able to capture the improvement that came from using specific predictors.

Figure 4.8A plots the average MCC for each of the 215 genes under study; Figure 4.8B summarizes this same plot in a boxplot and shows how, in average, PMut-S improved by 0.1 points PMut's MCC. It is to be noted that this improvement is not evenly distributed; Figure 4.8C shows how this improvement is more important for those proteins that are commonly misclassified by PMut.

It was interesting to analyse which factor was responsible for this boost in accuracy between PMut and PMut-S. We can identify three differences between the two predictors:

1. The addition of common variants from ExAC as neutral variants to balance the training set.

2. The use of 4 features from EVmutation, including one epistatic effect estimation.

3. The training of protein-specific predictors, instead of a single, general-purpose one.

Figure 4.9 summarizes the contribution of each of these additions. Comparing the two first boxplots we conclude that adding ExAC variants to PMut does not improve its predictive power. This is probably due to the fact that PMut

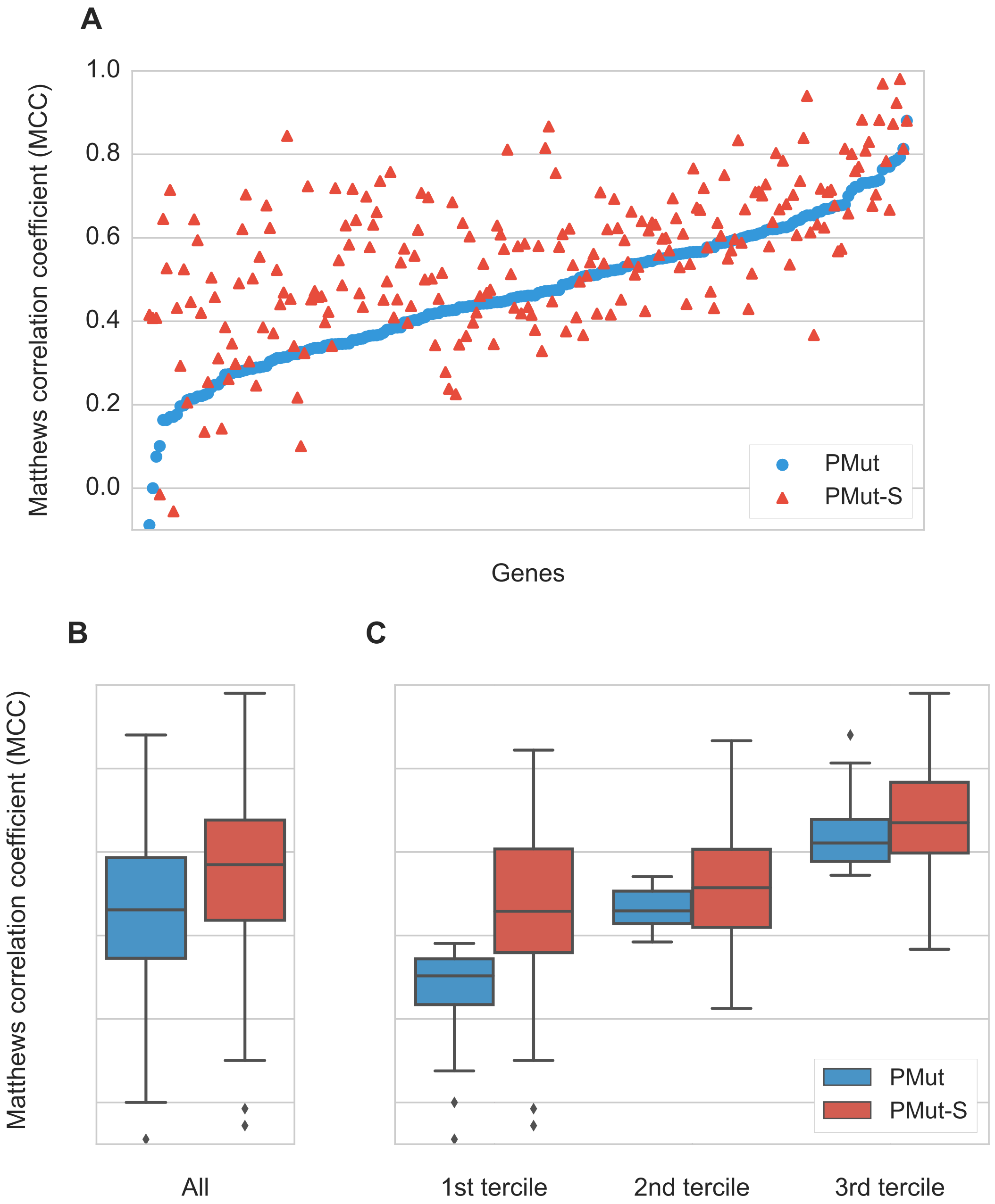


Figure 4.8 PMut-S compared to PMut. A. Per-protein Matthews correlation coefficient (MCC) comparison between PMut and PMut-S (plotted in increasing PMut score). B. Summary of the PMut and PMut-S per-protein MCC distribution in a boxplot. C. Boxplots split in PMut-S' MCC terciles; we can appreciate that PMut-S improvement over PMut-S is much more relevant in the first tercile, that is, PMut-S improves PMut predictions specially in the cases where PMut performs worse.

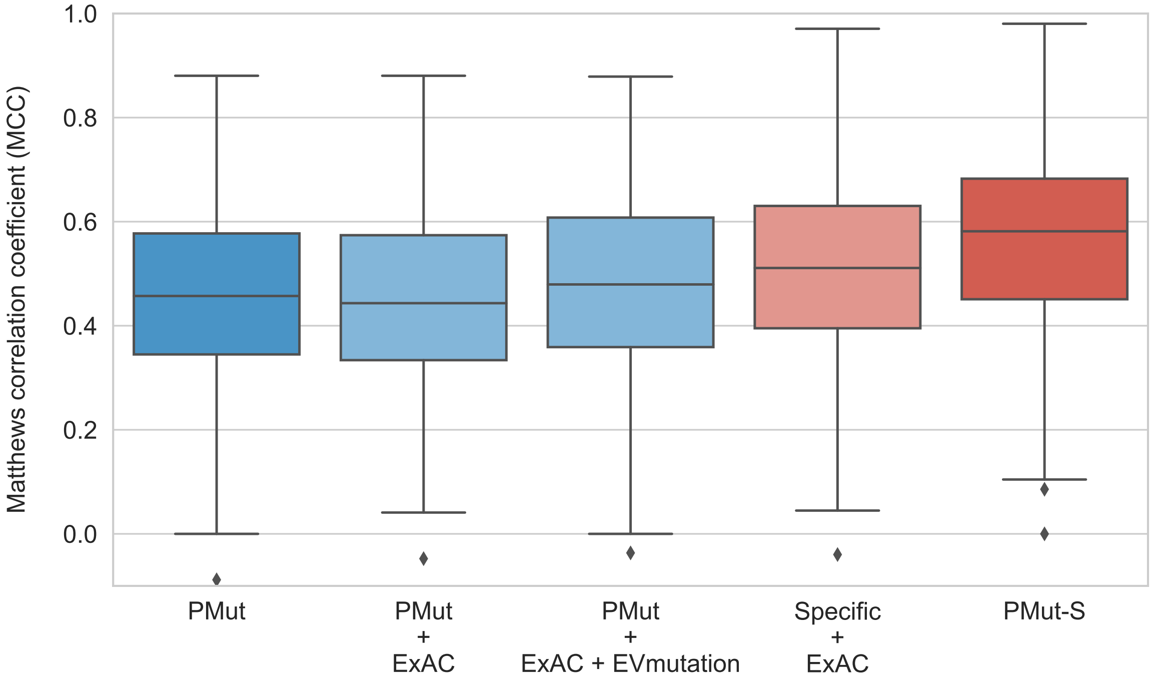


Figure 4.9 Comparison of per-protein MCC distribution using
1) general-purpose PMut predictor,
2) PMut predictor trained with the addition of neutral variants form ExAC,
3) PMut predictor with the addition of neutral ExAC variants and trained using co-evolution model features from EVmutation,
4) specific predictors trained with neutral variants from ExAC and classic PMut features,
5) PMut-S, protein-specific predictors using balanced datasets and co-evolution related features.

was already trained using a balanced training set, whereas for PMut-S this was a necessary step, as it lacked neutral variants.

The addition of epistatic effect predictions to PMut improves slightly its accuracy, as seen in the third boxplot, which confirms that co-variation conservation models add valuable information. It is important to note that PMut is based on 12 features selected from a total of 215 computed features; to improve its performance by simply adding one feature is a remarkable accomplishment.

Moving to specific predictors, the 4th boxplot shows that specific predictors trained with balanced datasets already improve the performance of our general-purpose predictors. Finally, PMut-S, trained with co-variation features from EVmutation achieves the highest performance in this detailed comparison with PMut, even though for the majority of these proteins, less than half their variants have co-evolution features computed.

4.2.4 PMut-S versus other general-purpose predictors

The initial insight that brought us to build PMut-S was the realization that many general-purpose predictors consistently performed badly for some genes. We can now compare PMut-S with 12 other predictors (Figure 4.10A), and see that it outperforms all of them. It is to note that this is a conservative comparison, as there is no guarantee that any of these predictors wasn't trained with variants that we are now evaluating (in fact we have the certainty that this is the case for some of them).

This same comparison, limited to the 100 proteins with worse average MCC, is done in Figure 4.10B, where we confirm the results we obtained comparing

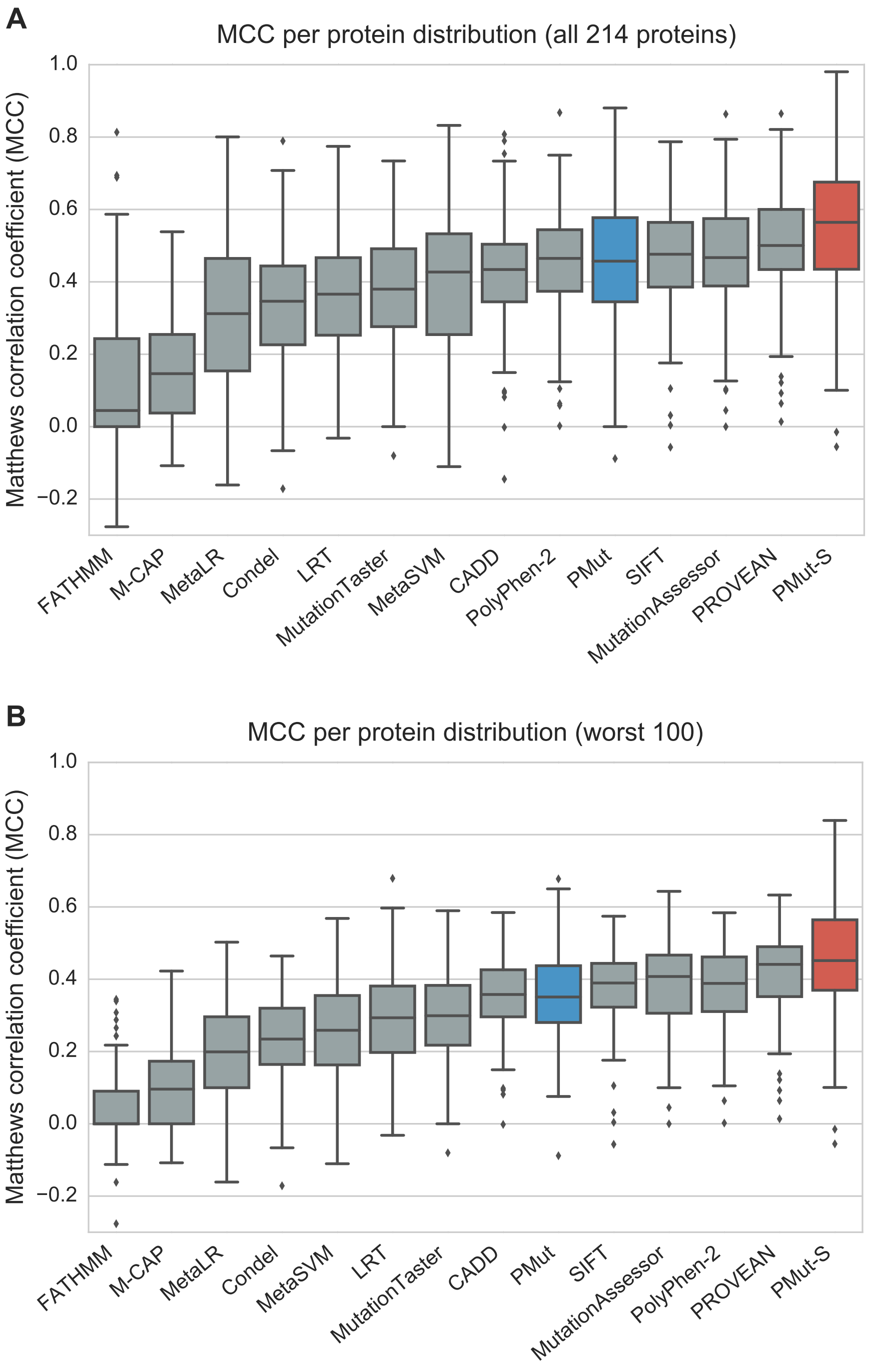


Figure 4.10 PMut-S compared to general-purpose predictors.
A. Per-protein MCC comparison between PMut-S (red), PMut (blue), and twelve other general-purpose predictors (gray);
B. Same analysis but limiting the comparison to the 100 proteins with worse predictions in general.

to PMut: PMut-S is even more powerful, compared to other predictors, when targeting the proteins commonly misclassified.

Another interesting consideration we can derive from this comparison is that, in general, unsupervised approaches such as those of conservation scores (SIFT, MutationAssessor, PROVEAN), are more consistent than supervised learning ones, specially in the case of meta-predictors (M-CAP, MetaLR, Condel, MetaSVM).

4.2.5 PMut-S versus other specific predictors

We have compared PMut-S' performance on a set of genes for which different specific predictors have been developed. Table 4.11 contains the complete comparison of PMut-S to 8 other predictors that we will now comment.

We start our comparison with maturity-onset diabetes of the young related genes. We evaluated PMut-S over 5 MODY genes (GCK, HNF1A, INS, ABCC8, KCNJ11 —namely MODY2, MODY3, MODY10, MODY12 and MODY13). We compared our results to those reported in Li et al. (2014), where eleven general-purpose pathology predictors were evaluated, being RadialSVM (Dong et al., 2015b) the most accurate predicting pathology in MODY genes. Overall, PMut-S has a better MCC (0.595) compared to RadialSVM's (0.474), and yields more accurate predictions for 4 of the 5 genes. This is a clear example in which a specific predictor is able to outperform all the general predictors.

Table 4.11 Comparison between PMut-S and eight specific predictors.

Family	Associated disease	Genes	PMut-S					Literature		
			Acc.	Sens.	Spec.	AUC	MCC	MCC	Method	References
MODY genes	Maturity-onset diabetes of the young	ABCC8	0.74	0.78	0.71	0.74	0.49	0.44	RadialSVM	Li et al. (2014), Dong et al. (2015b)
		GCK	0.86	0.89	0.82	0.86	0.72	0.60		
		HNF1A	0.75	0.70	0.80	0.75	0.50	0.37		
		INS	0.70	0.68	0.73	0.70	0.41	0.76		
		KCNJ11	0.86	0.89	0.83	0.86	0.72	0.57		
		Overall	0.80	0.82	0.78	0.80	0.60	0.47		
Kinase superfamily	Chronic myelogenous leukaemia, gastrointestinal stromal tumours, etc.	BTK	0.89	0.92	0.80	0.86	0.72		wKinMut-2	Vazquez et al. (2015)
		FGFR1	0.82	0.81	0.84	0.82	0.64			
		FGFR2	0.86	0.83	0.89	0.86	0.72			
		Overall	0.86	0.86	0.85	0.86	0.71	0.69		

Family	Associated disease	Genes	PMut-S					Literature		
			Acc.	Sens.	Spec.	AUC	MCC	MCC	Method	References
LQTS 1-3 genes	Long QT syndrome	KCNH2	0.81	0.74	0.89	0.81	0.63	0.62	Combination of 5 predictors	Leong et al. (2015)
		KCNQ1	0.80	0.77	0.84	0.80	0.61	0.70		
		SCN5A	0.69	0.64	0.75	0.69	0.39	0.32		
		Overall	0.76	0.71	0.82	0.76	0.53	0.44		
Voltage-gated potassium (Kv) channels	Cardiac arrhythmogenesis, LQTS, epilepsy, etc.	KCNH2	0.81	0.74	0.89	0.81	0.63		kvSNP	Stead et al. (2011)
		KCNQ1	0.80	0.77	0.84	0.80	0.61			
		KCNQ2	0.94	0.94	0.94	0.94	0.88			
		Overall	0.82	0.77	0.87	0.82	0.65	0.70		
Cytochrome P450	Congenital adrenal hyperplasia, etc.	CYP11B1	0.71	0.64	0.78	0.71	0.42		MutaCYP	Fechter and Porollo (2014)
		CYP17A1	0.75	0.81	0.69	0.75	0.50			
		CYP1B1	0.72	0.68	0.78	0.72	0.45			
		CYP21A2	0.88	0.88	0.88	0.88	0.76			
		Overall	0.78	0.76	0.80	0.78	0.56	0.70		

Family	Associated disease	Genes	PMut-S					Literature		
			Acc.	Sens.	Spec.	AUC	MCC	MCC	Method	References
NPC1 gene	Niemann-Pick disorder	NPC1	0.77	0.74	0.79	0.77	0.54	0.59	SAVER	Adebali et al. (2016)
Coagulation Factor VII	Hemophilia A	F8	0.85	0.90	0.81	0.85	0.71	Sens. = 0.85 Spec. = 0.79	HApredictor	Hamasaki-Katagiri *et al.* (2013)
Alpha-galactosidase A	Fabry disease	GLA	0.85	0.95	0.52	0.74	0.55	0.56 – 0.72	V7, V8	Riera et al. (2015)

In a recent study, Leong et al. (2015), evaluate five general-purpose predictors on the three main Long QT syndrome (LQTS) related genes: KCNQ1, KCNH2 and SCN5A. They propose a consensus method based on the best combination of these predictors. As seen in Table 4.11, PMut-S achieved a similar performance for each of these genes, and improved the overall MCC score. The case of SCN5A, for which PMut-S' MCC is the lowest in the table, is relevant to be discussed. This gene can be the cause of two different diseases: when affected by loss-of-function mutations, it causes LQTS, but if it has a gain-of-function mutation, Brugada syndrome may develop. This is a classic example where predictors such as PMut-S show their limits, as they generally tend to consider gain-of-function mutations as neutral mutations.

In a study of mutations in voltage-gated potassium (Kv) channels genes, which can also cause LQTS, Stead et al. (2011) present KvSNP, a random forest predictor trained on conservation, physicochemical and structural features. KvSNP was shown to improve the performance of general-purpose predictors, performing with an MCC of 0.7, which is a slightly higher value than PMut-S' MCC of 0.65.

Also based on a random forest classifier, wKinMut-2 (Vazquez et al., 2015) is a predictor specialized in protein kinases family mutations. Its features are derived from Pfam domains annotations, Gene Ontology terms, physicochemical properties of residues and sequence conservation. Its overall performance (MCC of 0.69) is very similar to PMut-S' (MCC of 0.70). PMut-S' performance holds for all three genes (BTK, FGFR1, FGFR1), with MCCs in the 0.65−0.75 range.

We compared PMut-S' performance on a set of Cytochrome P450 monooxygenases (CYPs) to the specific predictor MutaCYP (Fechter and Porollo, 2014). CYPs are well suited for specific predictors, as they have singular evolution patterns such as highly variable regions —substrate recognition sites— that may be considered unimportant by typical conservation analyses. In this case, PMut-S obtains an MCC of 0.56, lower than MutaCYP's MCC of 0.70. This difference is probably caused by the fact that PMut-S is trained on a smaller dataset. Because we discarded all the proteins with less than 30 disease variants were reported —thus keeping only 4 CYPs, whereas MutaCYP is trained using all variants from 15 CYPs.

Finally, we evaluated PMut-S' performance when predicting pathology in three monogenic diseases: Niemann-Pick disorder, caused by mutations on the NPC1 gene; Hemophilia A, due to a malfunction of Coagulation Factor VII (F8); and Fabry disease, caused by loss-of-function mutations in Alpha-galactosidase A (GLA). Regarding NPC1, in Adebali et al. (2016), the gene's evolutionary history was derived from carefully crafted multiple sequence alignments, and the ad-hoc SAVER algorithm was proposed as a specific pathology predictor for this gene, reporting an MCC of 0.59. PMut-S' MCC is 0.54, which hints the importance of the multiple sequence alignments quality for identifying evolutionary trends and pathology, as explained in Adebali et al. (2016). Compared to HApredictor, the F8-specific decision-tree classifier presented in Hamasaki-Katagiri et al. (2013), PMut-S obtained sligthly better results both in terms of sensitivity (0.90 vs. 0.85) and specificity (0.81 vs. 0.79). On the other hand, PMut-S' performance didn't match that of the GLA-specific neural network predictor described in Riera et al. (2015), which obtains an

MCC in the range of 0.56 – 0.72, higher than PMut'S MCC of 0.55. The use of first hand variation data from clinical settings probably has allowed them to build a more performant predictor.

In summary, we compared PMut-S to eight other specific predictors and saw it performed similarly to 4 of them, improved the predictions of 2 others, and had poorer performance than the other 2. However, it is important to note that these specific predictors generally relied on larger training sets than those available in public databases.

Although these studies confirmed the value of predictors relying on expert knowledge such as hand crafted sequence alignments, functional annotation of amino acids or the addition of relevant evolutionary information, it is remarkable that PMut-S, a set of predictors built in an automated and systematic way, performed at similar or even higher levels of accuracy in most of the cases analyzed.

4.2.6 Prediction of ExAC variants

As discussed in the Introduction, most of the single amino acid variants found in healthy humans are neutral variants that do not cause any disease (Kobayashi et al., 2017). It was interesting then to evaluate the predictions of PMut-S on ExAC variants, and to compare these predictions to those of PMut.

In Figure 4.11A, we see the PMut score distribution on more than 7 million variants. The curve is clearly shifted toward the neutral side of the plot, and shows that PMut correctly classifies most of the variants as neutral.

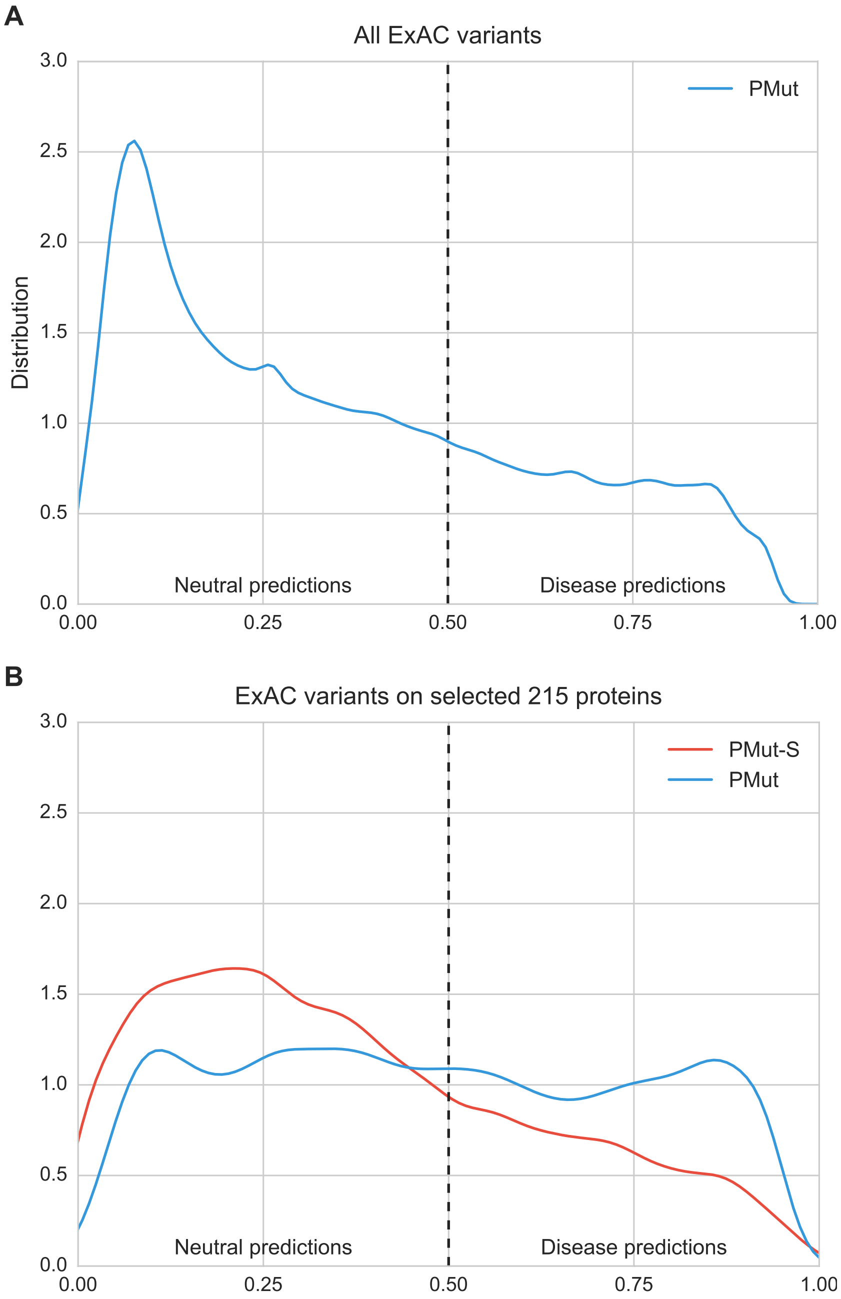


Figure 4.11 Score distribution on ExAC variants.
A. Distribution of the PMut score for about 7 million ExAC variants.
B. PMut and PMut-S scores distribution across 137,206 variants from the ExAC database on the 215 proteins studied (variants used in PMut-S' training were excluded from the analysis).

However, when limiting the analysis to the 215 proteins under study, we see that the distribution of PMut's score is almost flat (Figure 4.11B), that is, it evenly classifies variants as either neutral or pathological. As we discussed earlier, these 215 proteins are the ones that have more than 30 disease variants in SwissVar, and don't have, in general, as many neutral variants in this dataset. The overrepresentation of disease variants in these proteins is probably the cause for PMut's bad performance in this case.

PMut-S, which is trained on a balanced dataset, performs better in this case (Figure 4.11B), and classifies most of ExAC's variants as neutral. These results highlight the importance of using balanced datasets for training.

4.3 PyMut Python module

The PyMut Python module is a software framework based on the standard Python scientific and machine learning stack that provides functionality to apply it to the SAV pathology prediction problem.

It consists of about 2,000 lines of Python code, and offers functions that cover the whole supervised learning process: features computation, analysis and selection; model training and evaluation; and lots of other helper functions such as the parsing of several bioinformatics file formats like FASTA, BLAST XML results, etc.

PyMut is based on Pandas data frames, a matrix-like data structure based on efficient NumPy arrays and similar to R data frames. All PyMut variants data frames share a common list of columns (see Table 4.13), and each row represents a mutation. This standard data structure is then used by all the functions in PyMut, either as parameters o return values.

Next we detail a list of the most important variables and functions that this module exports:

`CLASSIFIERS` Data structure containing the six available classifiers, including their default parameters and parameter search space.

`FEATURES` List of all 215 features that can be computed by PyMut.

`PMUT_FEATURES` List of the 12 features selected in the PMut2017 predictor.

`FOLDS` List of different fold generation strategies for cross-validation: *k*-fold, stratified *k*-fold, label-exclusive and leave-one-out.

`compute_features` Compute all or selected features for a list of variants.

`features_distribution` Plot feature histograms, distinguishing neutral (green) and disease mutations (red).

`features_selection` Run the iterative feature selection algorithm described in Figure 4.2.

`cross_validate` Perform a cross-validation with the provided annotated variants, a classifier and a fold generation strategy.

`evaluate` Evaluate predictions and compute a complete set of metrics: accuracy, precision, sensitivity, specificity, AUC and MCC.

`roc_curve` Plot a ROC curve given the results of a cross-validation.

Table 4.13 PyMut variants data frame columns.

Column	Type	Description
protein_id	String	Protein identifier (usually UniProt).
sequence	String	String with the amino acid sequence of the protein.
position	Integer	Position of the variant in the sequence (1-based).
wt	String	Wild type amino acid in the variant position.
mt	String	Mutated amino acid.
disease	Boolean	Indication of whether the variant causes a disease.
feature_*	Float	Numerical features describing the mutation.
pred_score	Float	Predicted score in the range $[0,1]$.
pred_disease	Boolean	Prediction of pathology.

PyMut variants data frames all share this list of columns. Using this same data structure as input, different PyMut functions allow the user to perform tasks such as computing features, training a predictor, evaluating the quality of the predictions, etc.

`train` Build a predictor pipeline, which includes an *imputer* (for handling empty values), an *scaler* (to normalize each feature's distribution) and the classifier.

`predict` Predict the pathology of variants using a predictor pipeline.

`get_learning_curve` Plot learning curve (to estimate how better a predictor can be expected to get by making the training set bigger).

This software is available in the public Python package repository (https://pypi.org/project/pymut/), where it has been downloaded 2,402 times. Its source code is published in Github (https://github.com/vlopezferrando/pymut) under the open source MIT license, where it has been starred 6 times and has been forked in 5 ocasions.

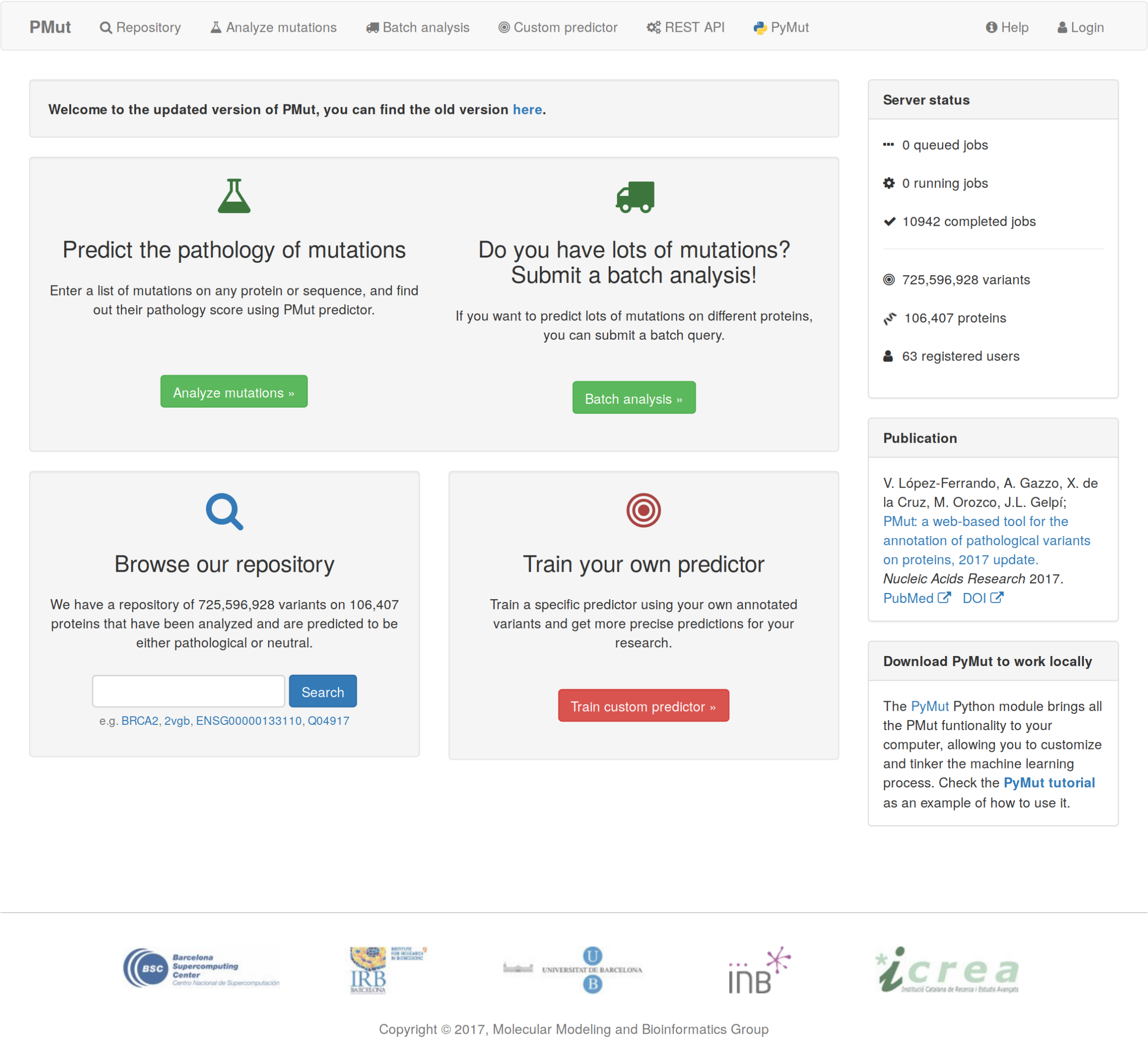
4.4 PMut web portal

The PMut web portal, available at http://mmb.irbbarcelona.org/PMut, is a web interface that gives access to the PMut2017 predictor, while contextualizing the predictions with other relevant information from other sources.

The home page (Figure 4.12) summarizes the services offered by the portal, which we itemize in the following sections.

4.4.1 Precomputed repository

It was clear when we set to develop the PMut web portal that we wanted PMut2017 predictions to be easily retrievable. Knowing that the vast majority



PMut   Repository   Analyze mutations   Batch analysis   Custom predictor   REST API   PyMut   Help   Login

Welcome to the updated version of PMut, you can find the old version here.

Predict the pathology of mutations

Enter a list of mutations on any protein or sequence, and find out their pathology score using PMut predictor.

Analyze mutations »

Do you have lots of mutations? Submit a batch analysis!

If you want to predict lots of mutations on different proteins, you can submit a batch query.

Batch analysis »

Browse our repository

We have a repository of 725,596,928 variants on 106,407 proteins that have been analyzed and are predicted to be either pathological or neutral.

Search

e.g. BRCA2, 2vgb, ENSG00000133110, Q04917

Train your own predictor

Train a specific predictor using your own annotated variants and get more precise predictions for your research.

Train custom predictor »

Server status

0 queued jobs

0 running jobs

✓ 10942 completed jobs

725,596,928 variants

106,407 proteins

63 registered users

Publication

V. López-Ferrando, A. Gazzo, X. de la Cruz, M. Orozco, J.L. Gelpí; PMut: a web-based tool for the annotation of pathological variants on proteins, 2017 update. *Nucleic Acids Research* 2017. PubMed DOI

Download PyMut to work locally

The PyMut Python module brings all the PMut funtionality to your computer, allowing you to customize and tinker the machine learning process. Check the **PyMut tutorial** as an example of how to use it.

Copyright © 2017, Molecular Modeling and Bioinformatics Group

Figure 4.12 PMut web portal home page.

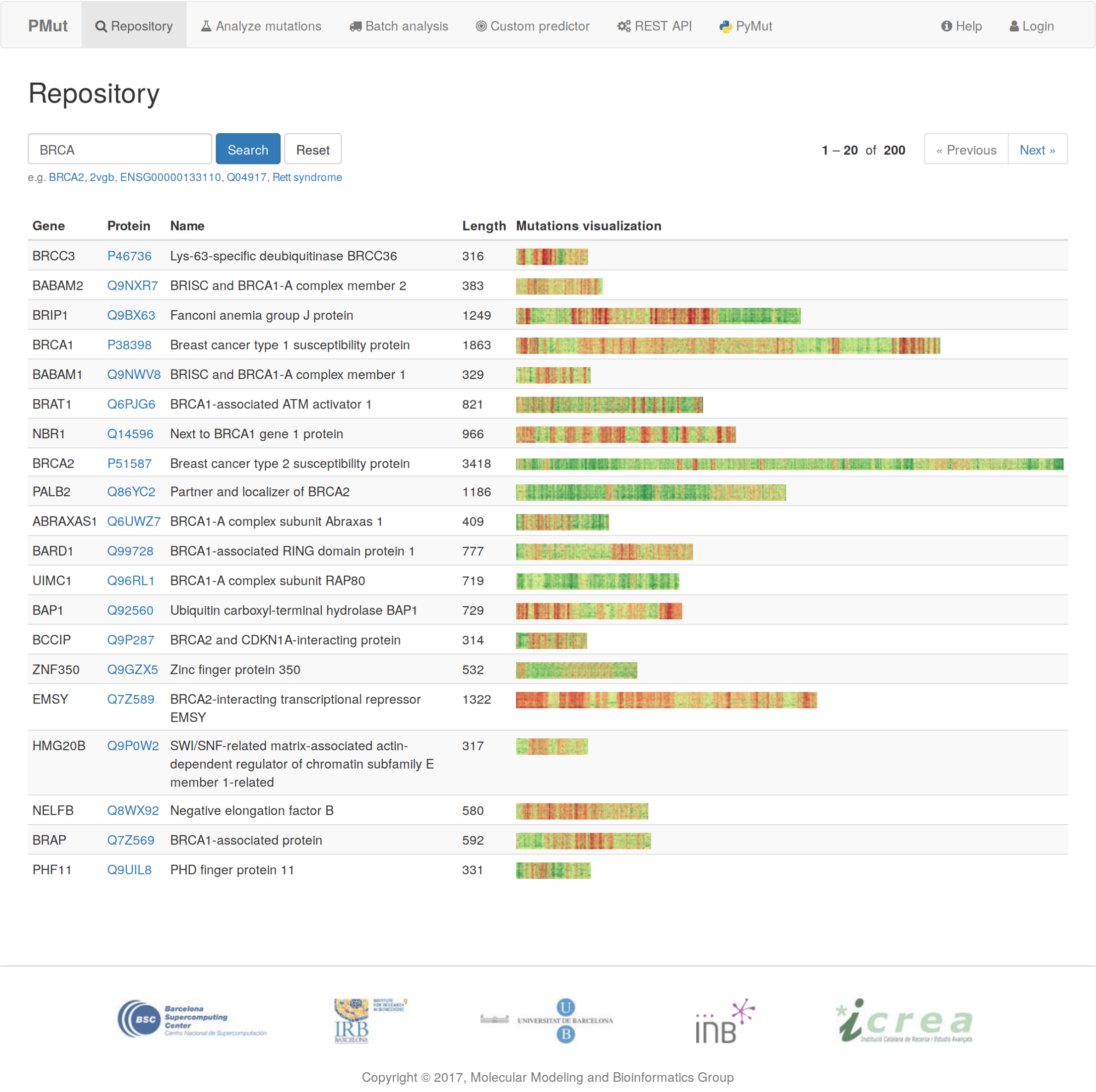
Table 4.15 PMut precomputed repository computations.

Step	Compute time	Description
1. PSI-BLAST	3,256 days	Search of similar sequences over UniRef100 and UniRef90.
2. Search of sequences	184 days	Retrieval of complete sequences from previous PSI-BLAST hits.
3. Kalign2 MSA	2,496 days	Multiple sequence alignment using Kalign2.
4. Features and predictions	152 days	Computation of features and prediction of pathology.
5. Protein images	32 days	Generation of images with all predictions of a protein for the repository.

Description of the five steps followed to produce the pre-computed repository.
Compute time is measured for single cores. The computation was done on MareNostrum III nodes consisting of two 8-core Intel Xeon processors E5-2670 at 2.6 GHz, 20 MB cache memory, and 8x4 GB DDR3-1600 DIMMS RAM.
It is worth noting that a total of 10,369 computation days were spent doing multiple sequence alignments using the MUSCLE program (Edgar, 2004), which were later discarded. Some of the alignments didn't succeed to finish in less than 72 hours, and others resulted in the program crashing during its execution. After this setback, we decided to use Kalign2, which was more stable and fast, at the expense of building less compact MSAs.

of our users would be studying human proteins, we decided to precompute the features and predictions of all possible mutations in all human proteins known to that date.

By pre-populating the repository with the features and predictions of 725,596,928 variants on 106,407 human proteins, the PMut web portal is able to answer most of the users' queries immediately. The computation of this repository required huge computational resources, which were provided by the Barcelona Supercomputing Center as compute time in the MareNostrum III supercomputer. Table 4.15 shows the steps and time spent for this computation.



Repository

BRCA

e.g. BRCA2, 2vgb, ENSG00000133110, Q04917, Rett syndrome

1 – 20 of 200

Gene	Protein	Name	Length	Mutations visualization
BRCC3	P46736	Lys-63-specific deubiquitinase BRCC36	316	
BABAM2	Q9NXR7	BRISC and BRCA1-A complex member 2	383	
BRIP1	Q9BX63	Fanconi anemia group J protein	1249	
BRCA1	P38398	Breast cancer type 1 susceptibility protein	1863	
BABAM1	Q9NWV8	BRISC and BRCA1-A complex member 1	329	
BRAT1	Q6PJG6	BRCA1-associated ATM activator 1	821	
NBR1	Q14596	Next to BRCA1 gene 1 protein	966	
BRCA2	P51587	Breast cancer type 2 susceptibility protein	3418	
PALB2	Q86YC2	Partner and localizer of BRCA2	1186	
ABRAXAS1	Q6UWZ7	BRCA1-A complex subunit Abraxas 1	409	
BARD1	Q99728	BRCA1-associated RING domain protein 1	777	
UIMC1	Q96RL1	BRCA1-A complex subunit RAP80	719	
BAP1	Q92560	Ubiquitin carboxyl-terminal hydrolase BAP1	729	
BCCIP	Q9P287	BRCA2 and CDKN1A-interacting protein	314	
ZNF350	Q9GZX5	Zinc finger protein 350	532	
EMSY	Q7Z589	BRCA2-interacting transcriptional repressor EMSY	1322	
HMG20B	Q9P0W2	SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily E member 1-related	317	
NELFB	Q8WX92	Negative elongation factor B	580	
BRAP	Q7Z569	BRCA1-associated protein	592	
PHF11	Q9UIL8	PHD finger protein 11	331	

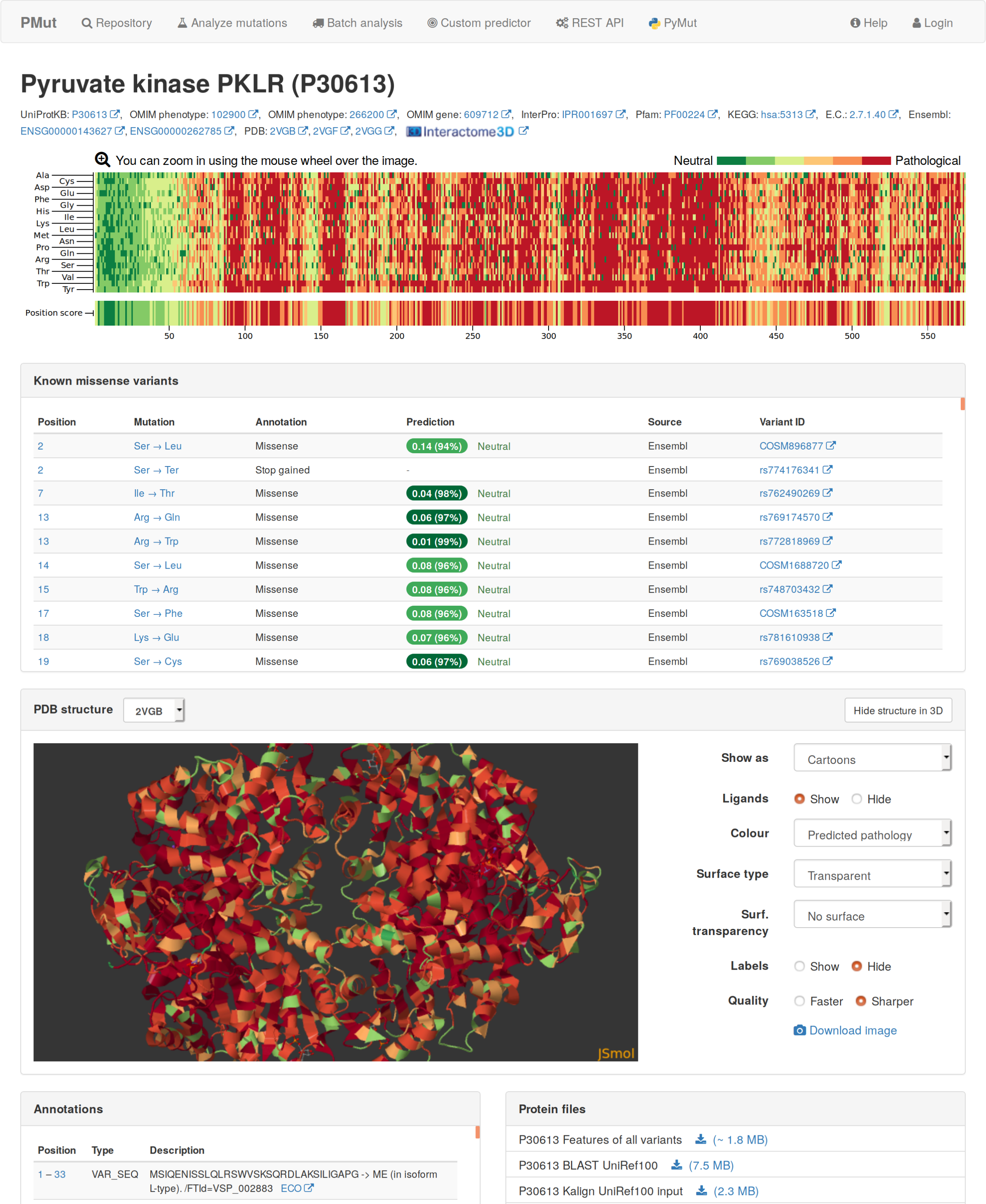
Copyright © 2017, Molecular Modeling and Bioinformatics Group

Figure 4.13 PMut web portal repository page, which allows the user to browse or search on 106,407 human proteins. The search is powered by UniProt's search engine.

Table 4.17 PMut repository statistics.

Proteins (from human UniRef100)	106,407
Analyzed variants	725,596,928
Analyzed variants (> 85% prediction reliability)	586,383,428 (80%)
Analyzed variants (> 90% prediction reliability)	370,444,279 (51%)

The repository can be browsed and searched (Figure 4.13) and each protein has a page (Figure 4.14) including all possible mutations' predicted pathology score. Also, the page offers links to different databases (UniProt, OMIM, InterPro, Pfam, KEGG, Ensembl, PDB, Interactome3D), highlights known variants from databases such as COSMIC, dbSNP or SwissVar, lists functional and structural annotations from UniProt, and maps the pathology scores on the protein's 3D structure if it is available in the PDB (protein data bank, Berman et al., 2003).



Pyruvate kinase PKLR (P30613)

UniProtKB: P30613, OMIM phenotype: 102900, OMIM phenotype: 266200, OMIM gene: 609712, InterPro: IPR001697, Pfam: PF00224, KEGG: hsa:5313, E.C.: 2.7.1.40, Ensembl: ENSG00000143627, ENSG00000262785, PDB: 2VGB, 2VGF, 2VGG, Interactome3D

You can zoom in using the mouse wheel over the image.

Known missense variants

Position	Mutation	Annotation	Prediction	Source	Variant ID
2	Ser → Leu	Missense	0.14 (94%) Neutral	Ensembl	COSM896877
2	Ser → Ter	Stop gained	-	Ensembl	rs774176341
7	Ile → Thr	Missense	0.04 (98%) Neutral	Ensembl	rs762490269
13	Arg → Gln	Missense	0.06 (97%) Neutral	Ensembl	rs769174570
13	Arg → Trp	Missense	0.01 (99%) Neutral	Ensembl	rs772818969
14	Ser → Leu	Missense	0.08 (96%) Neutral	Ensembl	COSM1688720
15	Trp → Arg	Missense	0.08 (96%) Neutral	Ensembl	rs748703432
17	Ser → Phe	Missense	0.08 (96%) Neutral	Ensembl	COSM163518
18	Lys → Glu	Missense	0.07 (96%) Neutral	Ensembl	rs781610938
19	Ser → Cys	Missense	0.06 (97%) Neutral	Ensembl	rs769038526

PDB structure 2VGB

Hide structure in 3D

Show as: Cartoons

Ligands: ● Show ○ Hide

Colour: Predicted pathology

Surface type: Transparent

Surf. transparency: No surface

Labels: ○ Show ● Hide

Quality: ○ Faster ● Sharper

Download image

Annotations

Position	Type	Description
1 – 33	VAR_SEQ	MSIQENISSLQLRSWVSKSQRDLAKSILIGAPG -> ME (in isoform L-type). /FTId=VSP_002883 ECO

Protein files

P30613 Features of all variants (~ 1.8 MB)

P30613 BLAST UniRef100 (7.5 MB)

P30613 Kalign UniRef100 input (2.3 MB)

Figure 4.14 PMut web portal protein page. In this page the user can retrieve the predictions on all possible variants on the protein, find a list on all known variants, visualize the predictions on the 3D protein structure, and view a list of annotations from different databases.

4.4.2 Pathology prediction

The PMut web portal offers the usual functionality of providing the pathology prediction for a list of variants. If these predictions have already been computed, the results are retrieved immediately from the repository; if not, a pipeline is launched to compute the sequence searches, alignments, features computations and predictions.

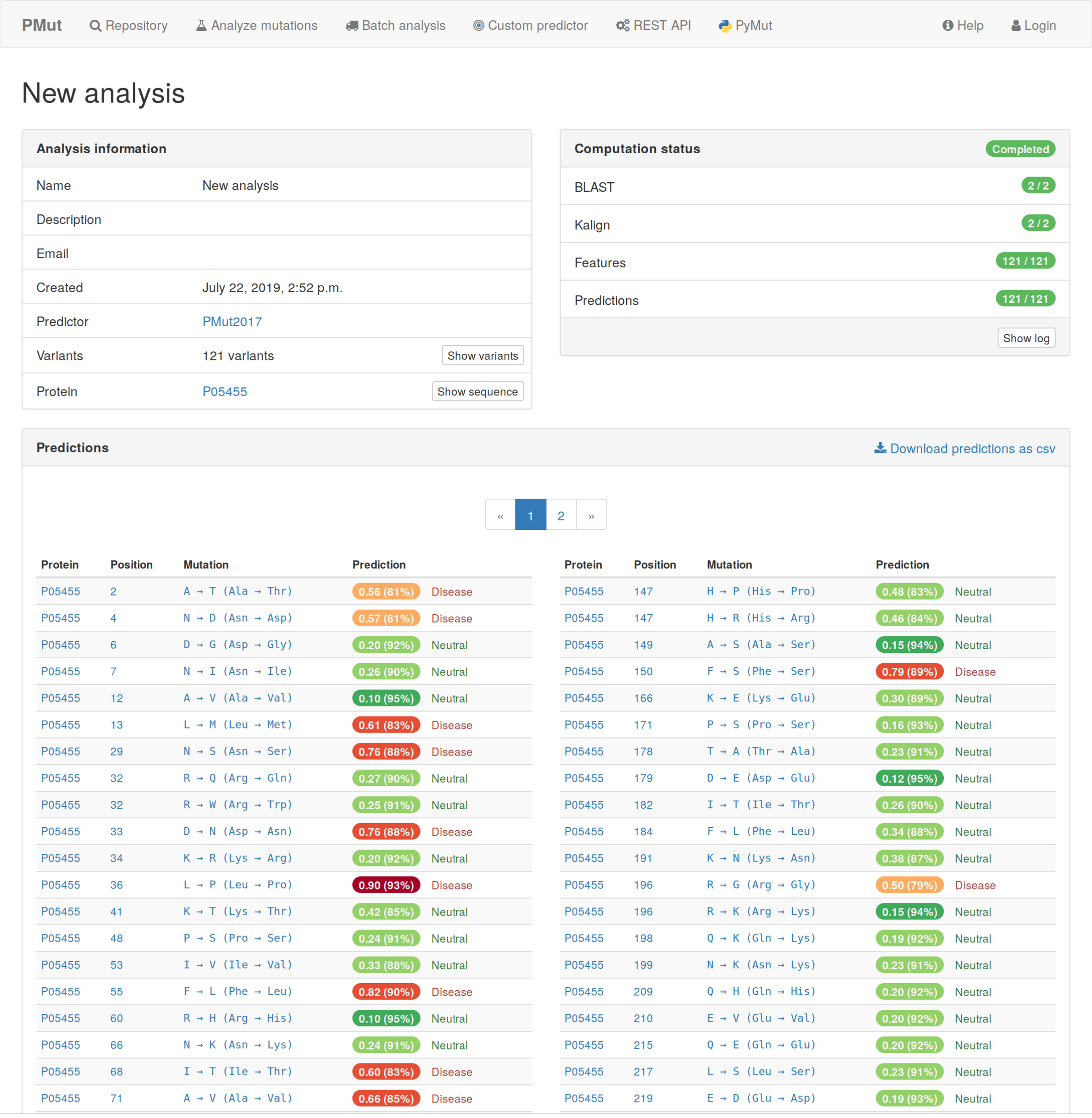
The results are shown in an intuitive interface (Figure 4.15), and are also available to be downloaded in CSV format.

4.4.3 Train your own predictor

One of the most advanced features of the web portal is the possibility for the users to train their own predictors. We provide a front-end to the PyMut engine, which allows users to visualize the features' distributions (Figure 4.16), train and cross-validate the predictor (Figure 4.17), and then use this predictor to predict pathology of variants of their choice.

4.4.4 API

Many users require automated access to predictions, either because they are interested in lots of data for large scale analyses, or because they may use these predictions as part of an automated pipeline. We ease this kind of use of our services by offering programatic access to the precomputed repository of the PMut web portal. These are the URLs available:



New analysis

Analysis information	
Name	New analysis
Description	
Email	
Created	July 22, 2019, 2:52 p.m.
Predictor	PMut2017
Variants	121 variants — Show variants
Protein	P05455 — Show sequence

Computation status	Completed
BLAST	2 / 2
Kalign	2 / 2
Features	121 / 121
Predictions	121 / 121

Predictions

Protein	Position	Mutation	Prediction	
P05455	2	A → T (Ala → Thr)	0.56 (81%)	Disease
P05455	4	N → D (Asn → Asp)	0.57 (81%)	Disease
P05455	6	D → G (Asp → Gly)	0.20 (92%)	Neutral
P05455	7	N → I (Asn → Ile)	0.26 (90%)	Neutral
P05455	12	A → V (Ala → Val)	0.10 (95%)	Neutral
P05455	13	L → M (Leu → Met)	0.61 (83%)	Disease
P05455	29	N → S (Asn → Ser)	0.76 (88%)	Disease
P05455	32	R → Q (Arg → Gln)	0.27 (90%)	Neutral
P05455	32	R → W (Arg → Trp)	0.25 (91%)	Neutral
P05455	33	D → N (Asp → Asn)	0.76 (88%)	Disease
P05455	34	K → R (Lys → Arg)	0.20 (92%)	Neutral
P05455	36	L → P (Leu → Pro)	0.90 (93%)	Disease
P05455	41	K → T (Lys → Thr)	0.42 (85%)	Neutral
P05455	48	P → S (Pro → Ser)	0.24 (91%)	Neutral
P05455	53	I → V (Ile → Val)	0.33 (88%)	Neutral
P05455	55	F → L (Phe → Leu)	0.82 (90%)	Disease
P05455	60	R → H (Arg → His)	0.10 (95%)	Neutral
P05455	66	N → K (Asn → Lys)	0.24 (91%)	Neutral
P05455	68	I → T (Ile → Thr)	0.60 (83%)	Disease
P05455	71	A → V (Ala → Val)	0.66 (85%)	Disease

Protein	Position	Mutation	Prediction	
P05455	147	H → P (His → Pro)	0.48 (83%)	Neutral
P05455	147	H → R (His → Arg)	0.46 (84%)	Neutral
P05455	149	A → S (Ala → Ser)	0.15 (94%)	Neutral
P05455	150	F → S (Phe → Ser)	0.79 (89%)	Disease
P05455	166	K → E (Lys → Glu)	0.30 (89%)	Neutral
P05455	171	P → S (Pro → Ser)	0.16 (93%)	Neutral
P05455	178	T → A (Thr → Ala)	0.23 (91%)	Neutral
P05455	179	D → E (Asp → Glu)	0.12 (95%)	Neutral
P05455	182	I → T (Ile → Thr)	0.26 (90%)	Neutral
P05455	184	F → L (Phe → Leu)	0.34 (88%)	Neutral
P05455	191	K → N (Lys → Asn)	0.38 (87%)	Neutral
P05455	196	R → G (Arg → Gly)	0.50 (79%)	Disease
P05455	196	R → K (Arg → Lys)	0.15 (94%)	Neutral
P05455	198	Q → K (Gln → Lys)	0.19 (92%)	Neutral
P05455	199	N → K (Asn → Lys)	0.23 (91%)	Neutral
P05455	209	Q → H (Gln → His)	0.20 (92%)	Neutral
P05455	210	E → V (Glu → Val)	0.20 (92%)	Neutral
P05455	215	Q → E (Gln → Glu)	0.20 (92%)	Neutral
P05455	217	L → S (Leu → Ser)	0.23 (91%)	Neutral
P05455	219	E → D (Glu → Asp)	0.19 (93%)	Neutral

Figure 4.15 PMut web portal analysis results.

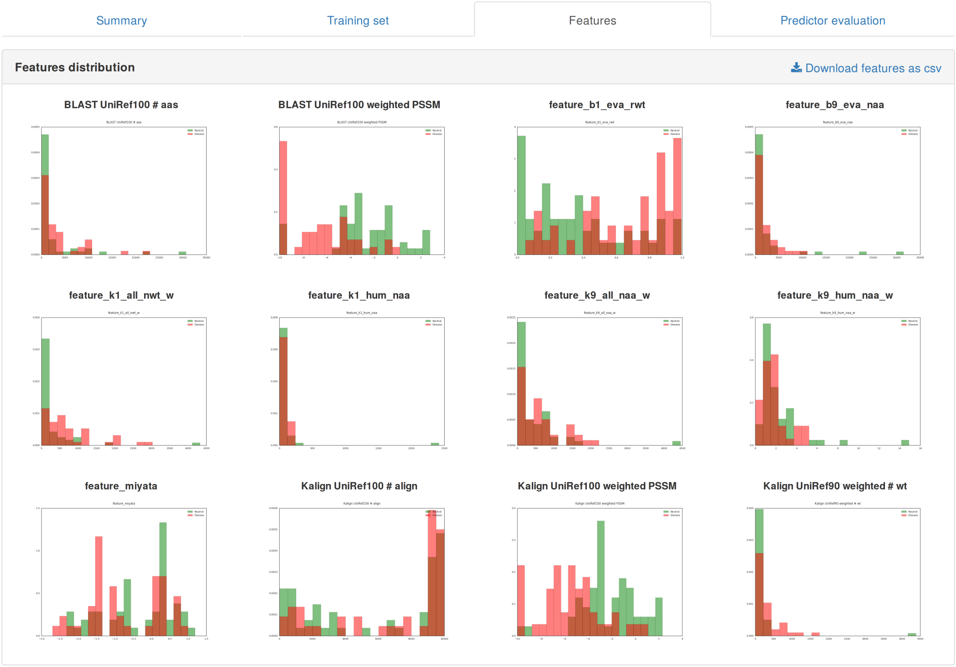
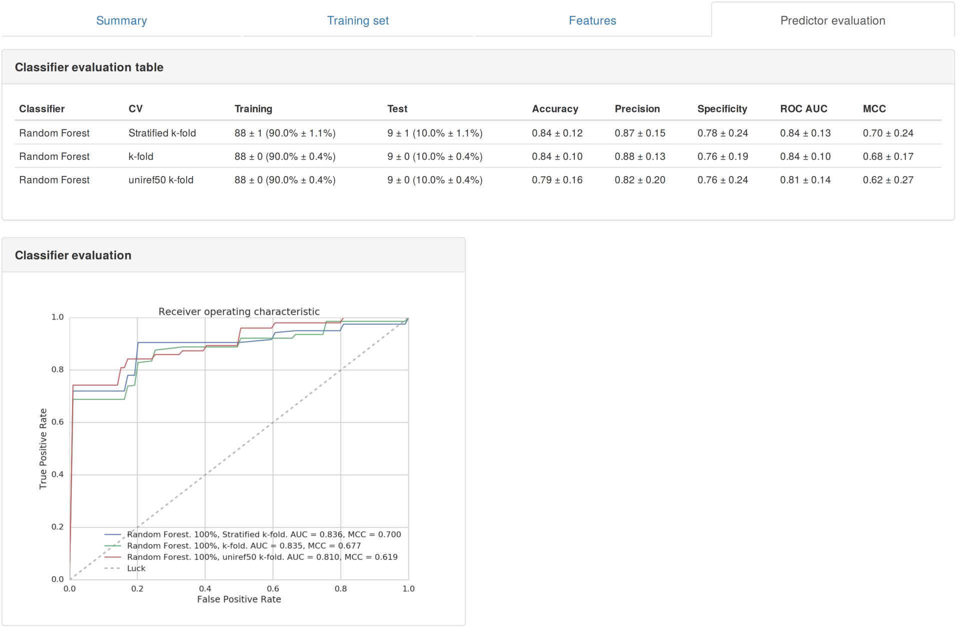


Figure 4.16 Histograms plotting the distribution of features used to train a custom predictor.



Classifier evaluation table

Classifier	CV	Training	Test	Accuracy	Precision	Specificity	ROC AUC	MCC
Random Forest	Stratified k-fold	88 ± 1 (90.0% ± 1.1%)	9 ± 1 (10.0% ± 1.1%)	0.84 ± 0.12	0.87 ± 0.15	0.78 ± 0.24	0.84 ± 0.13	0.70 ± 0.24
Random Forest	k-fold	88 ± 0 (90.0% ± 0.4%)	9 ± 0 (10.0% ± 0.4%)	0.84 ± 0.10	0.88 ± 0.13	0.76 ± 0.19	0.84 ± 0.10	0.68 ± 0.17
Random Forest	uniref50 k-fold	88 ± 0 (90.0% ± 0.4%)	9 ± 0 (10.0% ± 0.4%)	0.79 ± 0.16	0.82 ± 0.20	0.76 ± 0.24	0.81 ± 0.14	0.62 ± 0.27

Classifier evaluation

Receiver operating characteristic
True Positive Rate
False Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Random Forest. 100%, Stratified k-fold. AUC = 0.836, MCC = 0.700
Random Forest. 100%, k-fold. AUC = 0.835, MCC = 0.677
Random Forest. 100%, uniref50 k-fold. AUC = 0.810, MCC = 0.619
Luck

Figure 4.17 Evaluation metrics and ROC curves using different cross-validation strategies on a custom predictor built in the web.

`/PMut/uniprot/<UniProtID>/?<Position>/?<Mt>/features.csv`

Get the features and predictions for variants of a given protein. Optionally, we can limit our retrieval to a specific position in the sequence or a specific variant.

`/PMut/uniprot/<UniProtID>/?<Position>/?<Mt>.json`

Get all the information of a protein (or position or mutation), including all the links to other databases, annotations and known variants.

The most important data underlying the features computation are the PSI-BLAST search results and the multiple sequence alignments. This data can be downloaded from the following URLs:

`/PMut/blast/<UniRefID>.<90|100>.xml.gz`

PSI-BLAST search results on UniRef90/UniRef100 in XML format.

`/PMut/fa/<UniRefID>.<90|100>.fa.gz`

Fasta files including the whole sequences of proteins found in the PSI-BLAST search. This file is the input for the multiple sequence alignment program.

`/PMut/kalign/<UniRefID>.<90|100>.afa.gz`

MSA result in Fasta format as generated by Kalign2.

4.4.5 Usage

The web portal has been publicly available since January 2017, and has benefitted since the first day from the users of classic PMut, which were redirected to

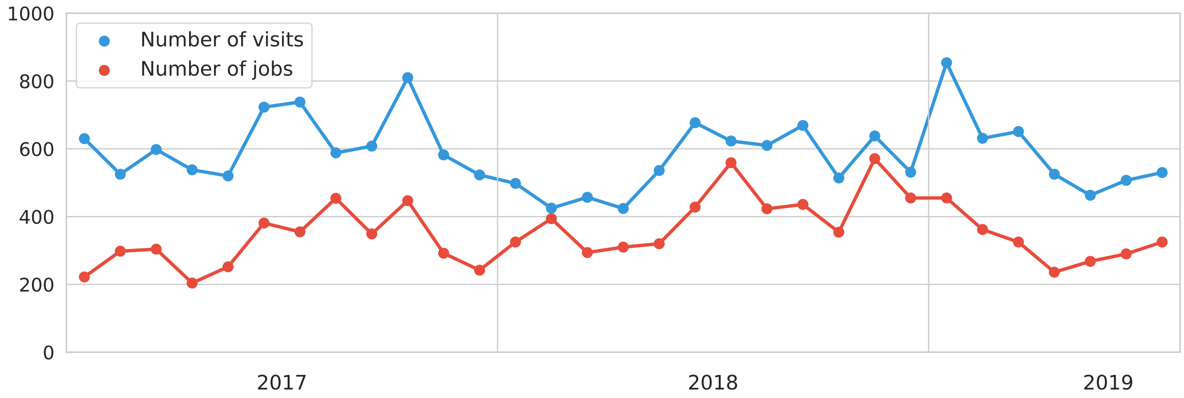


Figure 4.18 PMut web portal monthly visits and jobs submitted since its launch in January 2017.

the new page. In Figure 4.18 we see the monthly evolution of visits and jobs submitted.

In the PMut web portal it is not necessary to submit a job in order to retrieve a prediction, as millions of predictions are directly accessible in the precomputed repository. In the period between January 2017 and July 2019, and excluding all bot visits, we have counted 161,557 unique protein page views, covering 21,278 unique proteins. We have also recorded 110,758 downloads of CSV files with features and predictions for 19,032 different proteins.

During this period of time, 63 users have registered to the service, which allows them to save their analyses. Table 4.18 contains a list of the number of visits per country, which shows that the portal has users from all around the world.

Table 4.18 Country breakdown of PMut web portal visits.

Country	Number of visits*
United Kingdom	6,048
Spain	1,571
India	1,494
United States of America	1,204
Italy	910
China	483
Brazil	395
Iran	392
Germany	361
Japan	350
Canada	336
Other	4,417

* Number of visits between January 2017 and July 2019.

5. Discussion

> Science may be described as the art of systematic over-simplification — the art of discerning what we may with advantage omit.
>
> Karl Popper

> Science, in the very act of solving problems, creates more of them.
>
> Abraham Flexner

5.1 PMut2017 and PMut-S pathology predictors

5.1.1 PMut2017, a state-of-the-art pathology predictor

One of the main results of this thesis, the PMut2017 predictor, has proven to be a state-of-the-art predictor, with a performance comparable to that of the most accurate predictors published.

The basis of the efficiency of this predictor are its 12 features, selected from a total of 215, which measure sequence conservation using varied metrics on different alignments. These 12 features, selected by our automated feature selection algorithm, combined all the different criteria we had used in the initial computation of the features (Table 4.3). The choice of a random forest classifier was a natural one, as it was the best performing predictor (see Figure 4.1 and Table 4.1), offering fast train and prediction times. As we saw in the

Introduction, random forests are among the most common classifiers chosen to power SAVs pathology predictors.

One of the most useful features for PMut2017 users is the confidence score it assigns to each prediction. We measured in cross-validation tests that the MCC can go up to 0.77 (Table 4.5) for the two thirds more confident predictions, and from 0.42 to 0.62 for more than half the predictions in the blind test, leaving the other variants as unclassified. This confidence score can help prioritize the study of variants and gives a clear notion of the reliability of the prediction.

When studying PMut's behavior on predicting specific genes, we discovered that its performance varied greatly, ranging from MCCs of 0.86 to 0.12 (Table 4.8), and more interestingly, all the other predictors we evaluated behaved in a similar way. These results made us think that pathology and sequence conservation may follow different patterns depending on the gene, which led to the development of PMut-S, a set of protein-specific predictors.

5.1.2 PMut-S, suite of protein-specific predictors

The second main result of this thesis was PMut-S, a collection of 215 protein-specific predictors trained using a similar methodology to that of PMut.

With PMut-S we were able to show that specific predictors are able to improve predictions of general-purpose predictors. In our analyses, we observed an average improvement of 0.1 points in the MCC between PMut-S and PMut, enhancement which was identified to be even greater for genes with worse PMut predictions in general (Figure 4.8). This trend was later confirmed when compared to other general-purpose predictors (Figure 4.10). Part of the improvement in predictions came from the addition of co-evolution features and the configuration of training sets (Figure 4.9), additions that don't lead to such improvements in the case of general-purpose predictors. By comparing PMut-S to eight other protein-specific predictors we were able to see that it is possible to build specific predictors in a systematic fashion and achieve a performance similar those reported in these studies.

PMut-S also presented much better behavior than PMut when predicting neutral variants from the ExAC database (Figure 4.11). While PMut tended to evenly label these variants as neutral or pathological, PMut-S correctly classified most of them as neutrals. This lack of specificity in PMut is most probably due to biases in its training set, which had an enrichment of disease causing variants in the genes we evaluated.

5.2 Data, software and web services in Science

5.2.1 Open data and open source software

Reproducibility in science is a major concern (Baker, 2016; Open Science Collaboration, 2015). In the case of SAVs pathology prediction, the best way to assure the reproducibility of findings is the open sharing of data and software. This is starting to become commonplace, and for example M-CAP (Jagadeesh et al., 2016) and CADD (Kircher et al., 2014) both provide a precomputed dataset of predictions, which greatly helps to compare their performance to other methods. CADD authors also released the software implementing the method, as did EVmutation authors, who also published the MSAs used to fit their model. Still, many other predictors, and most of the specific predictors we compared PMut-S to, lack enough data to correctly assess their performance.

We applaud initiatives such as those promoting the FAIR principles (Findability, Accessibility, Interoperability, and Reusability, Wilkinson et al., 2016). Some services that conform to this standard, like UniProt (Consortium, 2018), have been essential for this work in providing protein sequences, variants and annotations.

Research on variant pathology prediction would greatly benefit from the adoption of these principles. Only by releasing all data used in the machine learning process: the training set, the features, the classifier software, and the predictions, it is possible to successfully reproduce the entire ML pipeline.

5.2.2 Web portals in biomedicine

The PMut web portal was developed having in mind the research community that uses predictors, rather than the one that develops them. Thousands of users in two years prove that the web is a great gateway to access SAVs predictions in a user-friendly manner. However, this accessibility comes at a cost, which is the effort made in building the web portal, and the non-negligible maintenance cost.

The nature of web development does not always fit well with research environments, where students come and go, and responsibilities tend to shift to others than the authors. I believe these matters are better handled under the umbrella of projects such as ELIXIR (the European infrastructure for biological information, Crosswell and Thornton, 2012), which promote the creation of the technical infrastructure to fill the gap between researchers and users. Services like BioTools (Ison et al., 2019, 2015), an aggregator of bioinformatics programs, or MuGVRE (Multiscale Genomics Virtual Research Environment, Codó et al., 2019), a web portal that integrates dozens of tools for working on 3D/4D genomics, are great examples of such initiatives.

5.2.3 Scientific software

It is obvious that software plays a central role in bioinformatics. We have stated before that it is desirable to publish the code (services like Github make this extremely easy), but it is important to emphasize that building good software is fundamental for doing good science. Just like we need to follow good practices in our assays design, statistical analyses, the plotting of data, and the citation of bibliography, it is crucial to follow software development good practices when building our software (Wilson et al., 2014).

We tried to do this in this work, and the release of PyMut is a result of this conviction. We published the source code in Github under an open source license, added it to the standard Python modules repository, and documented a tutorial in the form of a Jupyter Notebook to walk the user through all of its functionality (http://mmb.irbbarcelona.org/PMut/PyMut-tutorial).

It was a stretch to convert the programs powering PMut2017 to a generic prediction library, but this helped us build the PMut-S predictors in a much faster way.

Despite the publication of the code and tutorial, and the thousands of downloads, we are not aware of any use of PyMut in projects other than PMut or PMut-S.

5.3 Challenges and opportunities in SAVs pathology prediction

5.3.1 Training sets of annotated variants

The most important part of a supervised learning based SAVs pathology predictor is the training set, that is, the list of variants annotated as either neutral or disease causing that it builds upon. The quality of this dataset limits the maximum accuracy a predictor can achieve no matter the methodology used.

Wrongly annotated variants are an inevitable burden in mutation databases. It is common in SwissVar for a few variants to switch from neutral to pathological or *vice versa* with every release. The best approach to prevent, or reduce these troubles is the adoption of standard guidelines for the annotation of variants such as those proposed by MacArthur et al. (2014) and Richards et al. (2015), which define clear protocols to determine the pathogenicity of variants.

ClinVar's decision to pair each variant with information regarding the source, the level of confidence on the annotation, etc. and to foster a debate between experts whenever a conflict arises, is the best approach for building a robust database of variants. This traceability is also provided by SwissVar, which links each variant with relevant publications or databases used as sources in the UniProt web portal. In the case of HGMD, which is another high quality database of pathological mutations, it has the drawback that it is a private database, with its public version being 3.5 years outdated. Even if we were able to use the private database to train our predictors, we wouldn't be allowed to publish the full training set, which would hamper the reproducibility of our findings.

Balancing the training sets

Training sets are the base of any supervised learning predictor, and the lack of complete and balanced training sets has been a constant struggle for researchers in this field. Traditionally, disease-causing variants from databases were compensated by variants generated from homology models, considered as neutrals.

The recent availability of large databases with thousands of sequences from healthy individuals, such as gnomAD (Karczewski et al., 2019), offers a better source for neutral variants. We successfully used ExAC (Lek et al., 2016) variants as neutrals to train our PMut-S predictors, but saw how this addition didn't improve, by itself, the performance of the PMut2017 general-purpose predictor (Figure 4.9).

In contrast, CADD (Kircher et al., 2014) is trained to distinguish 14.7 million high-frequency human-derived variants from 14.7 million simulated variants. This is a very interesting approach, as it builds on the large data sets collected in large scale sequencing studies, which confers the method an *unsupervised* nature that allows for it to have very fair evaluations against annotated variants databases.

5.3.2 General versus specific predictors

The distinction between general and specific predictors is often not so clear. In predictors based solely on conservation or statistical scores, these scores are usually derived from a single multiple sequence alignment, which almost converts them in specific predictors. Even in the case of general supervised-learning approaches, the training set is general, but the feature computation for each variant also has a specific nature.

Nevertheless, we have seen with PMut-S that specific predictors, trained with variants from a single protein, and applicable only to that same protein, can achieve better accuracies than general-purpose predictors. This improvement is even greater for proteins that are generally wrongly classified by general predictors, hinting the presence of particular patterns in the way these proteins relate to disease.

We have also proved that building a specific predictor is a process that can be successfully automated. The most important step in this case is the training set collection and the crafting of high-quality multiple sequence alignments; the rest of the process can be stream-lined without much complication.

5.3.3 Conservation and co-variation scores

We have repeatedly shown in this work how sequence conservation underlies all pathology predictors, and the link between conservation and pathogenicity of a mutation is the most important insight used to tackle this problem. The sequencing of more organisms creates the opportunity to build richer multiple sequence alignments and refine conservation scores, but also introduces the methodological difficulties of aligning tens of thousands of sequences. This complexity has sometimes been bypassed by focusing on Pfam domains instead of whole sequences (Hopf et al., 2017) or using phylogenetic trees (Tang and Thomas, 2016); however, the calculation of conservation measures from tens or hundreds of thousands of sequences still poses a challenge.

In line with recent studies (Figliuzzi et al., 2016; Hopf et al., 2017), we have confirmed with PMut-S that co-variation is able to improve the accuracy of methods that were already highly optimized. Sequence co-variation analysis is widely regarded as a key feature for future predictors (Feinauer and Weigt, 2017), but for the moment they can only be considered as complementary to classic conservation, as the high computational cost of these methods makes them applicable to a fraction of protein sequences.

5.3.4 Supervised learning in pathology prediction

By reviewing the literature, it is easy to conclude that supervised learning predictors achieve higher accuracies than conservation scores. Only the new co-variation score EVmutation was able to match the performance of other supervised learning methods, but with significant lower coverage. It is obvious, however, that a supervised learning predictor including EVmutation scores as features (such as PMut-S) would obtain more accurate predictions, just like many predictors did by integrating SIFT scores as input features.

Still, it was surprising to find out that simple conservation scores yielded better results than supervised learning predictors for the genes in the PMut-

S comparison with general predictors (Figure 4.10). This rises concerns on overfitting for many supervised learning based predictors.

Machine learning has achieved unprecedented success rates in many fields in the past years. Deep learning, i.e. multi-layer neural networks, are behind great advances in speech recognition (Amodei et al., 2016), computer vision (Szegedy et al., 2016) or chess playing (Silver et al., 2017). There is great interest in the application of deep learning to genomics (Eraslan et al., 2019), and it has already been used to infer the natural history of populations (Flagel et al., 2019) and predict the impact of variants on splicing (Cheng et al., 2019) or gene expression (Zhou et al., 2018). We are not aware of any application of deep learning to SAV pathology prediction for the moment, but it will surely be a method of choice once sequence and variants data are abundant enough.

5.3.5 Predictor evaluation

Predictor evaluation and comparison is still an open problem in the field. We have seen predictors reporting accuracies of $80-90\%$ and MCCs higher than 0.6 for over a decade, but consistently performing worse when faced with independent test sets (Riera Ribas, 2016). Even though this also makes it seem as if we have reached a performance plateau, we can also presume that the accuracy of many predictors has been often overestimated.

These setbacks, although inevitable to some extent, need to be frontally addressed through publication of training sets, features, classifiers and predictions, as we mentioned earlier. Databases like dbNSFP (Liu et al., 2011b) or ANNOVAR (Wang et al., 2010), which contain millions of variants together with functional and pathology predictions are perfect evaluation sets to compare to other predictors.

CAGI challenges (like the *Annotate all missense* challenge, to which we submitted the PMut2017 predictions for all dbSNFP variants), are a great initiative to compare, in equal conditions, the performance of different predictors. These comparisons require the collection of first-hand experimental data, and it is through collective enterprises that they can succeed.

5.3.6 The future of SAVs pathology prediction

Predictors have succeeded, with almost 90% accuracy, at the prediction of pathology for SAVs involved in Mendelian diseases. However, many challenges remain open around pathology predictors.

The application of SAV predictors to complex diseases, for example, is still a current issue. Predictors have been developed to differentiate between passenger and driver mutations (Carter et al., 2009), but due to the intricate nature of this problem, they haven't reached the accuracies needed for clinical application. The integration of pathways and interactions will surely help in the understanding of the relation between variants and complex diseases.

We discussed in the Introduction how coding variants explain the vast majority of Mendelian phenotypes, but they don't explain all of them. The range of effects of in-frame indels on phenotype is as varied as that of SAVs. It has also been shown that non-coding variants, even far removed from the coding sequence, can alter gene expression and cause a severe phenotype

(Makrythanasis and Antonarakis, 2013). Characterizing the effect of all these kinds of variation is a key step in any DNA analysis pipeline.

For many pathologically classified variants, it remains to be described how they do cause a disease. There are many mechanisms that can lead a SAV to cause a loss-of-function of the protein (preventing folding, impeding a protein-protein interaction...), but it is a much harder problem to describe the effect of a gain-of-function variant. It would be desirable for predictors to shed some light on the specific effects of the mutation.

When envisioning an ideal predictor, one can imagine an ensemble of predictors rather than a single one. Conservation, co-variation, structure, dynamics, annotations of all kind... all are useful features to characterize variants, but are unevenly available depending on the protein. A machine learning method able to use all information available for each protein, but falling back to a general predictor when needed would probably yield the highest accuracies. This approach, which is similar to that of many meta-predictors, calls for a careful implementation, as it is prone to overfitting.

Throughout this work we have revolved around the neutral-pathological dichotomy of single amino acid variants; which, of course, is a simplification. Variants can cause a disease with variable penetrance, epigenetic changes and gene-environment interactions can alter the phenotype for equal genotypes and variants never occur in isolation. Even if they did, we would still have trillions of different cells in our bodies. In the end, we need to keep in mind the complexities of disease, while tackling one problem at a time.

6. Conclusions

1. We have built PMut2017, a completely renewed version of the PMut predictor.

2. PMut2017's accuracy, evaluated through cross-validation and two blind tests, is in line —or even greater— than that of the most accurate state-of-the-art predictors.

3. PMut2017 predictor outputs a reliability score fitted to the task of prioritizing mutations that are more likely to be pathological.

4. PMut-S, a suite of 215 protein-specific predictors, validates the hypothesis that specific predictors can reach better accuracies than general ones, specially for some proteins consistently misclassified by general-purpose predictors.

5. Epistatic coevolution models have been confirmed to enrich the classic conservation analysis and improve the accuracy of pathology predictors, specially in the case of specific predictors.

6. In the new PMut web portal, users can submit the variants they want analyzed and they can find a precomputed repository with the predictions for all possible mutations on human proteins, together with 3D visualizations and sequence annotations related to their proteins of interest.

7. The PyMut Python module encapsulates the functionalities we used in the machine learning methods of this work. Its source code is publicly available and can be used to easily train predictors with custom data.

A. Science outreach

One of the most rewarding activities that I carried out during my PhD was giving talks in different high schools as part of the *Camins Infinits* (Infinite Paths) project, from the Universistat de Barcelona outreach group (*La UB divulga*).

In these talks to students between 12 and 18 years old I introduced them to bioinformatics, supercomputing and machine learning. This is the list of centers I visited between November 2015 and February 2018:

- Institut Sant Just Desvern, Sant Just Desvern.
- Institut de Viladecans, Viladecans.
- Institut Badalona VII, Badalona.
- Escola Ntra. Senyora de Montserrat, Rubí.
- CS Jaume Viladoms, Sabadell.
- PFI-PTT de Viladecans, Viladecans.
- Col·legi Urgell, Barcelona.
- Institut Esteve Terradas i Illa, Cornellà de Llobregat.
- Escola Proa, Barcelona.
- IES Sos Baynat, Castelló de la Plana.

B. PMut: a web-based tool for the annotation of pathological variants on proteins, 2017 update

doi: 10.1093/nar/gkx313

PMut: a web-based tool for the annotation of pathological variants on proteins, 2017 update

Víctor López-Ferrando[1,2], Andrea Gazzo[2,3], Xavier de la Cruz[4,5], Modesto Orozco[2,3,6,*] and Josep Ll Gelpí[1,2,6,*]

[1]Barcelona Supercomputing Center (BSC), Barcelona, Spain, [2]Joint Program BSC-CRG-IRB Research Program for Computational Biology, Barcelona, Spain, [3]Institute for Research in Biomedicine (IRB) Barcelona, The Barcelona Institute of Science and Technology, Barcelona. Spain, [4]Vall d'Hebron Institute of Research (VHIR), Universitat Autònoma de Barcelona, Barcelona, Spain, [5]ICREA, Barcelona, Spain and [6]Dept. of Biochemistry and Molecular Biomedicine, University of Barcelona, Barcelona, Spain

Received February 19, 2017; Revised March 31, 2017; Editorial Decision April 08, 2017; Accepted April 18, 2017

ABSTRACT

We present here a full update of the PMut predictor, active since 2005 and with a large acceptance in the field of predicting Mendelian pathological mutations. PMut internal engine has been renewed, and converted into a fully featured standalone training and prediction engine that not only powers PMut web portal, but that can generate custom predictors with alternative training sets or validation schemas. PMut Web portal allows the user to perform pathology predictions, to access a complete repository of pre-calculated predictions, and to generate and validate new predictors. The default predictor performs with good quality scores (MCC values of 0.61 on 10-fold cross validation, and 0.42 on a blind test with SwissVar 2016 mutations). The PMut portal is freely accessible at http://mmb.irbbarcelona.org/PMut. A complete help and tutorial is available at http://mmb.irbbarcelona.org/PMut/help.

INTRODUCTION

Single nucleotide variants (SNVs) are responsible for ∼90% of human variability (1). When mapped on coding regions (single amino acid variants, SAVs) may affect the function of the transcribed proteins (2), leading to phenotype variations, and often to pathology. Last generation sequencing and genotyping techniques are reporting a large amount of human genetic variation data (3), fueling initiatives to derive links between genome alterations and pathologies. Thus, the HapMap consortium (4) is characterizing common variation and linkage disequilibrium patterns that can be related to common diseases (5). The Human Variation Project (6) collects, curates, and makes accessible information on genetic variations affecting human health. The 1000 Genomes Project (www.1000genomes.org) is expected to produce the most complete catalog of genetic variations in human population (7). Already in 2005, the Wellcome Trust Case Control Consortium genotyped ∼14 000 patients for seven common diseases performing one of the largest Genome-Wide Association Study (GWAS) (8) to date. As a result of these and other projects the dbSNP database at the NCBI (9) collects, nowadays, ∼20 million of validated human SNPs. The manually curated SwissVar database (10) reports on the pathological effect of ∼61 000 missense SNPs, the public version of the HGMD database (11) includes >78 000 missense mutations causing, or associated with human inherited diseases, plus disease-associated/functional polymorphisms and ClinVar reports over 125 000 clinically relevant variants (12). Systematic sequencing through NGS of cancer patients (projects like ICGC, www.icgc.org, and TGCA, cancergenome.nih.gov) expanded the range of mutations in the human genome. For example, the present public version of ICGC reports ∼520 000 new somatic SAVs.

Despite the amount of data available, the issue of predicting the functional consequences of SAVs is still open, and there is a continuous effort in developing more accurate and flexible predictors (13–15). There is no consensus on the type of approach used to obtained predictions. PMUT (16), one of the oldest and still widely used methods, uses neural networks, as so does for instance SNAP (17); SIFT (18), Polyphen (19), PROVEAN (20), LRT (21) and MutationAssessor (22; Hidden Markov Models are used in PANTHER (23) and FATHMM (24); Random Forests are used in PON-P2 (25), CHASM (26), CanPredict (27) and MuD (28); Support Vector Machines in CADD (29), SNPs&GO (30), SeqProfCod (31), LS-SNP (32), SNPs3D (33), MetaSVM (34) and MetaLR (34); Naïve Bayes is used

[*]To whom correspondence should be addressed. Tel: +34 934034009; Fax: +34 934021559; Email: gelpi@ub.edu
Correspondence may also be addressed to Modesto Orozco. Tel: +34 934037156; Fax: +34 934037157; Email: modesto.orozco@irbbarcelona.org
Present address: Andrea Gazzo, Interuniversity Institute for Bioinformatics, ULB-VUB, Brussels, 1050 Brussels, Belgium.

in MutationTaster (35), and a gradient boosting tree is used in M-CAP (36). Also predictors giving consensus predictions like Condel (37), are available. All of them use input features that represent the change in amino acid sequence, structure and evolutionary properties resulting from the amino acid replacement.

PMut (16), first released in 2005, predicted the pathological nature of a given SAV, and also hot-spot positions on protein sequences. PMut used a neural network-based classifier trained by a manually curated dataset extracted from SwissProt (38), and used sequence conservation and predicted physico-chemical properties as main features. Here, we present a major update of the PMut predictor and web portal. While maintaining the philosophy of the original application, the backend classification engine has been completely renewed and automated, taking advantage of the increase in the amount of data available for training. The engine powering PMut is provided as a separate software package (PyMut) that allows users to prepare their own predictors for specific families of proteins. The new PMut web site also provides access to a complete data repository, including all possible SAVs on human known proteins. The web portal is available at http://mmb.irbbarcelona.org/PMut and has already received more than 900 requests between 1 January 2017 and 1 April 2017.

PMut PREDICTION ENGINE

PMut prediction engine (PyMut) is prepared as a Python 3 module. PyMut is based on the widely used libraries NumPy (www.numpy.org) and Scipy (www.scipy.org), for fast numerical computing, Pandas (data management, pandas.pydata.org), Scikit-learn (machine learning, scikit-learn.org), Matplotlib (matplotlib.org*)*, and Seaborn (graphical representation, seaborn.pydata.org). PyMut performs all operations regarding calculation of features, selection of classifiers, validation and results analysis. It is distributed as a separate software module that can be downloaded and installed locally. Specific functions available are:

- Compute protein features and plot their distribution (see Supplementary Tables S1 and S2 in the Supplementary Material for a complete list, and details of the procedures).
- Select the most informative features. Selection of features is performed in an iterative way, following the improvement of MCC obtained in cross-validation. Supplementary Figure S1 shows a schema of the algorithm used, and Supplementary Figure S2 a plot of such MCC evolution.
- Train classifiers, evaluate them using several cross-validation protocols, and obtain their Receiver Operating Characteristic (ROC) curves (see Supplementary Table S4, and Supplementary Figure S3 for a list of the available classifiers and its comparative performance).
- Prepare and evaluate a pathology predictor.
- Predict the pathology of mutations.

PyMut module covers all operations required to generate new predictors in a fully automated way (see Supplementary Table S5 for a detailed list of software functions, and Supplementary Table S6 for a list of software dependencies), allowing the user to easily explore alternative datasets, classifiers or collections of features, enabling to fine tune the predictor to cover not only pathology, but other structural or functional characteristics of the proteins. As the module can be downloaded and run locally, it allows the user to analyze private data to derive tailored predictors without uploading it to a server.

PyMut source code is available at https://github.com/inab/pymut and in the official Python package repository (https://pypi.python.org/pypi/pymut), and can be downloaded from the PMut Web portal, where a tutorial following the main functions of PyMut is also available.

PMut2017 PREDICTOR

PMut2017 default predictor was trained using the manually curated variation database SwissVar (10) (October 2016 release), which contains 27 203 disease and 38 078 neutral mutations on 12 141 proteins. Two hundred fifteen numerical features were first computed for each mutation, accounting for (i) physical property differences between wild type and mutated amino acids, (ii) protein interactome information and (iii) amino acid conservation. The conservation features are derived from local searches over UniRef100 and UniRef90 cluster databases (39), using PSI-Blast (40), and multiple sequence alignments generated using Kalign2 (41). After evaluation of the different machine learning algorithms (Supplementary Table S4 and Supplementary Figure S3), the chosen predictor is based on a Random Forest (42) classifier, trained with only 12 selected features (see Supplementary Table S3). The classifier outputs a prediction score between 0 and 1; mutations scoring from 0 to 0.5 are classified as neutral, and those scoring from 0.5 to 1 are classified as pathological. To evaluate the confidence degree of such score, we have analyzed the accuracy of the predictions based on their score (Supplementary Figure S4). It can be seen that accuracy increases with extreme score values. The analysis of these results allows us to qualify the prediction with a statistically meaningful reliability score.

PMut predictor has been validated following several approaches:

1) *A traditional 10-fold cross-validation on protein families with 50% sequence identity exclusion*. No sequence in the testing set shares >50% sequence identity with any protein in the training set. Figure 1 shows the corresponding ROC curves. Detailed performance metrics are summarized in Table 1. Restriction of the analysis to most confident predictions lead to a significant increase in the performance of the prediction. See Supplementary Figure S4 for a measured confidence of PMut scores.
2) *A blind validation using new SwissVar entries*. To perform this analysis, PMut has been trained using the same protocol, but limited to the data available at the SwissVar December 2015 release, and tested with SwissVar 2016 entries (3166 new mutations on 762 proteins. 1656 mutations were tagged as pathological and 1510 as neutral). For comparative purposes, we have also performed a complete series of analyses of the same test-set using other prediction methods. Table 2 shows the good performance of PMut when applied to the entire test set.

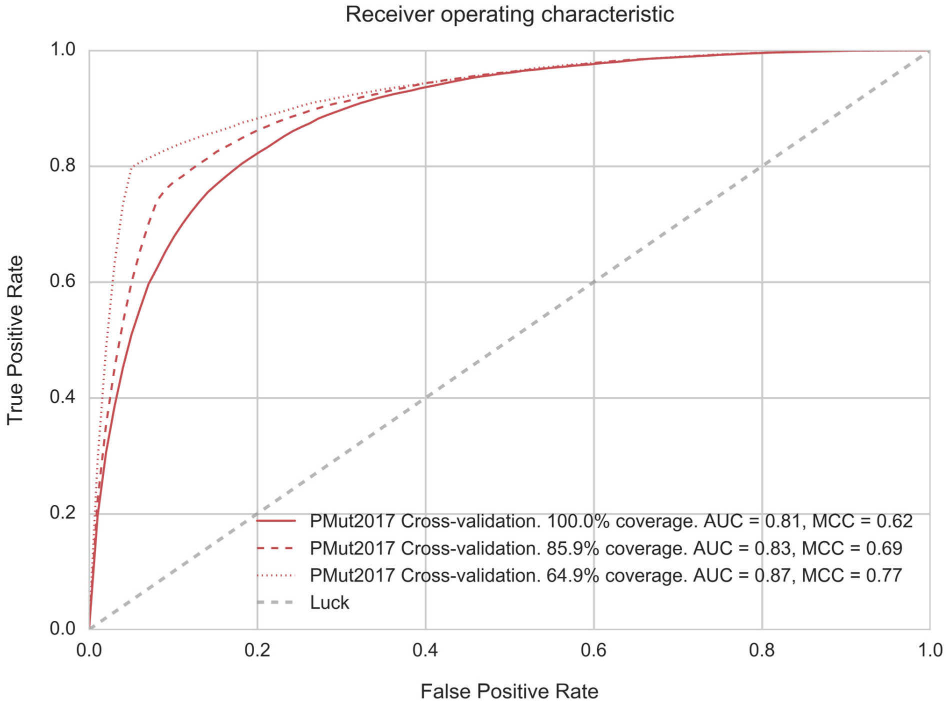


Figure 1. ROC Curves corresponding to PMut2017 10-fold cross-validation based on protein families. No sequence in the validation set has more than 50% identity with sequences in the training set. Additional curves correspond to the subsets of prediction with 85% and 90% confidence. AUC: area under the curve, MCC: Matthews Correlation Coefficient.

Table 1. Summary of performance metrics for PMut2017 predictor

Confidence	Coverage	Accuracy	Sensitivity	Specificity	AUC	MCC
All	100	0.82	0.76	0.86	0.81	0.62
>85% [a]	85.9	0.85	0.75	0.92	0.83	0.69
>90% [a]	64.9	0.90	0.80	0.95	0.87	0.77

[a] Only predictions with scores corresponding to higher confidence levels are considered (see Supplementary Figure S4).

The new engine used in PMut2017 represents a large improvement over the original PMut predictor (MCC 0.03 in the blind validation, data not shown). If predictions are limited to those cases where more reliable scores are obtained (PMut 85%), MCC value raises to 0.53, what puts PMut as the most accurate predictor within the test performed, covering yet >80% of the mutations. This can be further improved taking only >90% confidence scores (MCC 0.62), but in this case predictions can be obtained only in half of the cases. The availability of PyMut module allows for a seamless update of PMut predictor when new releases of SwissVar become available.

3) *A blind validation using ClinVar entries not included in the training set*. ClinVar (12) is an alternative well known source for disease related variants. ClinVar includes a much larger set of variants, although only 34 024 can be directly mapped to the protein sequences used in PMut. From those, 13 716 were already present in the SwissVar training set. To further analyze the performance of the PMut2017 predictor we have analyzed the variants reported in ClinVar that were not used in the training set (20 308 variants). Results are shown in Table 2. MCC value (0.49) is slightly better but comparable with those obtained in the SwissVar 2016 delta release. This behavior further confirms the prediction power of PMut2017.
4) *Comparative test on selected genes*. Global validation schemas provide an averaged estimation of the performance of a prediction method. Applying the predictor to specific protein families may result in a degraded performance due to the individual features of such families. We have selected a number of genes to evaluate PMut2017 performance in comparison with some other methods. To avoid a statistical bias, genes have been selected in a way that the number of neutral and pathological mutations were equilibrated and reasonably large. Results are reported in Table 3. As expected, MCC values obtained are specific to the analyzed gene and range widely around the average value obtained in the global test. Although some of them show a clearly poorer result, it can be seen that the behavior is consistent with the other methods assayed, and shows a good predicting power. These differences additionally support the need to develop specific predictors for protein families showing non-standard behavior.

Table 2. Comparative performance of PMut2017 predictor

Method	Coverage (%)	Accuracy	Specificity	Sensitivity	AUC	MCC
SIFT (18)	89.6	0.61	0.33	0.88	0.60	0.25
Polyphen2 (19)	92.1	0.64	0.35	0.91	0.63	0.32
PROVEAN (20)	91.5	0.64	0.41	0.87	0.64	0.31
FATHMM (24)	90.5	0.55	0.45	0.64	0.55	0.09
PON-P2 (25)	42.4	0.72	0.52	0.9	0.71	0.45
CADD (29)	95.0	0.65	0.33	0.94	0.64	0.35
M-CAP (36)	91.5	0.60	0.19	0.95	0.57	0.22
Condel (37)	91.0	0.63	0.40	0.84	0.62	0.26
LRT (21)[a]	95.1	0.73	0.58	0.87	0.73	0.47
MutationAssessor (22) [a]	95.1	0.63	0.46	0.78	0.62	0.26
MetaSVM (34) [a]	95.1	0.63	0.51	0.74	0.62	0.26
MetaLR (34) [a]	95.1	0.6	0.46	0.73	0.60	0.20
MutationTaster (35) [a]	95.1	0.65	0.31	0.96	0.64	0.36
PMut	**100.0**	**0.71**	**0.65**	**0.76**	**0.71**	**0.42**
PMut (85%)[b]	**81.0**	**0.76**	**0.76**	**0.77**	**0.76**	**0.53**
PMut (90%)[b]	**51.2**	**0.81**	**0.78**	**0.84**	**0.81**	**0.62**
PMut (ClinVar)[c]	**100.0**	**0.73**	**0.88**	**0.85**	**0.75**	**0.49**

Blind validation based on new variants added to SwissVar during 2016 (3166 variants), CADD predictor has been evaluated using a threshold of 20. AUC: area under the ROC curve, MCC: Matthews correlation coefficient.

[a] Analysis performed from ANNOVAR data (42).

[b] Analysis restricted to most reliable PMut predictions (reliability level in parentheses).

[c] Blind validation based on variants reported on ClinVar (43), not present in the SwissVar dataset (20,308 variants). Indicated coverage is calculated on ClinVar dataset.

Table 3. Comparative performance of the PMut2017 predictor on selected genes

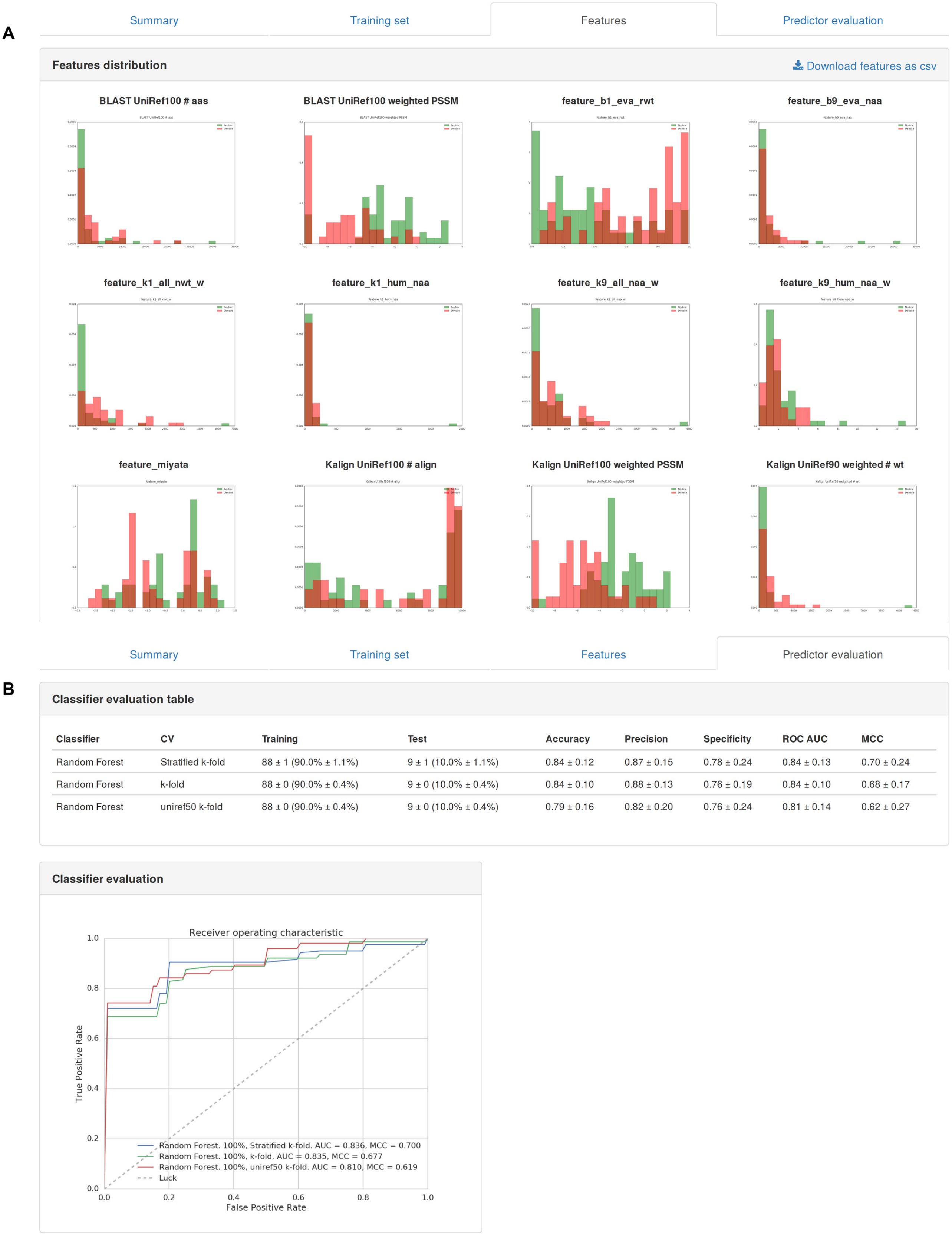
Gene	Disease	#D	#N	PMut	SIFT	Polyphen	LRT	Mut. Taster	Mut. Assessor	PROVEAN
MECP2	Rett syndrome	46	22	**0.86**	0.66	0.85	0.69	0.64	0.41	0.53
COL1A2	Osteogenesis Imperfecta	78	20	**0.77**	0.74	0.62	0.55	0.55	0.74	0.74
SLC4A1	Distal Renal Tubular Acidosis	38	36	**0.69**	0.65	0.65	0.54	0.55	0.68	0.60
ADAMTS13	Upshaw-Schulman syndrome	43	17	**0.62**	0.76	0.46	0.00	0.71	0.54	0.62
ATM	Hereditary cancer-predisposing syndrome	46	54	**0.60**	0.53	0.57	0.32	0.42	0.48	0.55
ATP7B	Wilson disease	195	25	**0.48**	0.34	0.49	0.37	0.43	0.29	0.52
MLH1+MSH2+MSH6+PMS2	Lynch syndrome	159	78	**0.48**	0.32	0.31	0.23	0.16	0.43	0.32
MYOC	Primary open angle glaucoma	57	24	**0.47**	0.37	0.45	0.38	0.50	0.47	0.49
TTC21B	Jeune thoracic dystrophy	16	28	**0.42**	0.20	0.22	0.18	0.16	0.28	0.26
SCN5A	Brugada syndrome	154	46	**0.40**	0.32	0.26	0.43	0.31	0.34	0.34
KCNH2+SCN5A	Congenital long QT syndrome	270	54	**0.38**	0.32	0.28	0.36	0.32	0.30	0.38
ABCA1	Tangier disease	32	31	**0.37**	0.43	0.31	0.32	0.47	0.43	0.47
PKHD1+PKD1	Polycystic kidney disease	197	96	**0.37**	0.43	0.37	0.30	0.41	0.36	0.45
FBN1	Marfan syndrome	385	20	**0.35**	0.31	0.25	0.21	0.33	0.32	0.30
RYR1	Central core disease	147	25	**0.34**	0.27	0.31	0.00	0.36	0.28	0.34
LDLR	Familial hypercholesterolemia	103	23	**0.32**	0.29	0.08	0.17	0.09	0.26	0.25
DYSF	Limb-Girdle Muscular Dystrophy	48	16	**0.31**	0.35	0.27	0.15	0.21	0.41	0.39
BRCA2	Breast-ovarian cancer, familial 2	43	61	**0.31**	0.10	0.18	0.18	0.14	0.19	0.01
BRCA1	Breast-ovarian cancer, familial 1	27	36	**0.31**	0.24	0.20	0.38	0.29	0.30	0.17
WFS1	WFS1-Related Spectrum Disorders	40	17	**0.30**	0.25	0.35	0.20	0.18	0.16	0.26
PINK1	Parkinson Disease	23	39	**0.25**	0.33	0.48	0.40	0.41	0.44	0.30
LRRK2	Parkinson Disease	21	24	**0.19**	0.06	0.14	0.01	0.13	0.09	0.14
CFTR	Cystic fibrosis	146	32	**0.15**	0.06	0.20	0.21	0.12	0.20	0.27
PROC	Thrombophilia	36	28	**0.12**	-0.15	-0.08	0.07	0.14	0.08	-0.01

MCC values obtained restraining the analysis to variants on the indicated genes. Analysis for non-PMut methods performed from ANNOVAR data (42). #N Neutral mutations, #D Disease causing mutations.

PMut WEB PORTAL

The access to PMut is possible through a Web portal (http://mmb.irbbarcelona.org/PMut/) which is implemented in Python using the Django Web framework (www.djangoproject.com). The variants' features and predictions are stored in a MongoDB database (www.mongodb.com). All calculations are performed under the control of a SGE queuing system configured to deploy additional back-end workers on peaks of demand. Id mapping and keyword searches are performed using the appropriate services at EBI (www.ebi.ac.uk). Sequences, features, variants and 3D structures are obtained from MMB-IRB data repository (mmb.irbbarcelona.org/api). Most functionalities of the portal are available anonymously, but the users may register to keep records of the activity in the server. After registering, a private workspace is created with links to the prediction requests, and to their customly trained predictors.

The portal is divided in four sections:



Classifier	CV	Training	Test	Accuracy	Precision	Specificity	ROC AUC	MCC
Random Forest	Stratified k-fold	88 ± 1 (90.0% ± 1.1%)	9 ± 1 (10.0% ± 1.1%)	0.84 ± 0.12	0.87 ± 0.15	0.78 ± 0.24	0.84 ± 0.13	0.70 ± 0.24
Random Forest	k-fold	88 ± 0 (90.0% ± 0.4%)	9 ± 0 (10.0% ± 0.4%)	0.84 ± 0.10	0.88 ± 0.13	0.76 ± 0.19	0.84 ± 0.10	0.68 ± 0.17
Random Forest	uniref50 k-fold	88 ± 0 (90.0% ± 0.4%)	9 ± 0 (10.0% ± 0.4%)	0.79 ± 0.16	0.82 ± 0.20	0.76 ± 0.24	0.81 ± 0.14	0.62 ± 0.27

Classifier evaluation
Receiver operating characteristic
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Random Forest. 100%, Stratified k-fold. AUC = 0.836, MCC = 0.700
Random Forest. 100%, k-fold. AUC = 0.835, MCC = 0.677
Random Forest. 100%, uniref50 k-fold. AUC = 0.810, MCC = 0.619
Luck

Figure 2. Partial screenshots of output of Predictor's training section. (**A**) Comparative plot of the selected protein features. (**B**) ROCs curves of performance evaluation.

Table 4. Statistics of PMut repository (January 2017)

Proteins available (from human UniRef)	106 407
Analysed variants	725 596 928
Analysed variants (>85% prediction reliability)	586 383 428 (80%)
Analysed variants (>90% prediction reliability)	370 444 279 (51%)

Data repository, which allows the user to access the set of pre-calculated PMut2017 predictions covering all human protein sequences in UniRef100. Search options available include protein name and id (UniprotKB (44) or PDB (45)), gene id (Ensembl (46)), dbSNP id (9) and free keywords. Data in the repository can be accessed programmatically through a REST API. The repository is continuously updated with new information appearing in UniRef100 (39). Detailed statistics of the repository can be found in Table 4.

Pathology prediction, which allows the user to evaluate the pathological profile of SAVs, input options include protein id(s), or uploaded sequences. PMut Data Repository is used to speed up analysis in the case of known protein sequences. The default predictor is PMut2017, but Custom Predictors can also be used. In the case of mutations mapping on known sequences all possible single SAVs are precomputed. The output includes a variety of graphical and numerical results which are presented in different formats. In all cases, the output combines the information with known variants and sequence features of the protein, giving a comprehensive context of the mutations. When 3D structure is available, the mutations are also mapped onto the structure, using a JsMol visualizer (www.jmol.com). Several public databases (UniprotKB (44), PDB (45), PFam (47), InterPro (48) and Interactome3D (49)) are linked to the results card. All intermediate data including alignments and calculated features, are available for download in the appropriate formats. See representative screenshots at Help pages on PMut portal, http://mmb.irbbarcelona.org/PMut/help).

Batch predictions, users requesting larger series of predictions can send them in a single batch. The options available, and output are equivalent to single requests. The user is informed when the work is finished and results are stored in his/her private workspace.

Train you own predictor, this section provides a frontend to the PyMut engine. Users can specify a training set, select a classifier and a validation procedure. Figure 2 shows some screenshots of the output. Available information includes the original training set, a graphical view of the calculated features (Figure 2A), and a summary of the evaluation results (Figure 2B). The newly trained predictor becomes automatically available in the prediction section of the portal. Please note that the use of trained predictors requires to log in the personal workspace and is restricted to the user developing it.

CONCLUSIONS

The 2017 new release of the PMut portal constitutes a novel approach that largely improves our previous 2005 PMut server. The new portal offers not only a generally trained predictor that performs in a competitive manner with current available methods, but allows the user to access an automatic procedure to train new predictors with specific datasets or features. The possibility of enriching the analysis with alternative predictors, or training predictors with specific information of a single protein family, largely increases the scope of usability of the portal. Overall, the 2017 release of PMut is a powerful tool to approach the issue of predicting functional consequences of protein sequence variants, and will surely contribute to improve the quality of the annotation of pathological variants. The server and platform are already available and accessible without restrictions.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Dr Patrick Aloy for providing protein interactome data. We thank Jose A. Alcantara, and Dr. Adam Hospital, for technical assistance.

FUNDING

Spanish Ministry of Science [BIO2015-64802-R-32868, SEV-2015-0493, TIN2012-34557, TEC2015-67774-C2-2-R, SAF2016-80255-R]; Catalan Government [2014-SGR-134, 2014-SGR-1051]; Instituto de Salud Carlos III-Instituto Nacional de Bioinformática [INB; PT13/0001/0019, PT13/0001/0028]; European Union, H2020 programme [Elixir-Excelerate: 676559; BioExcel: 674728 and MuG: 676566]; La Caixa Foundation fellowship [to V.L.F.]. Funding for open access charge: European Union and Spanish Ministry of Science.

Conflict of interest statement. None declared.

REFERENCES

1. Collins,F.S., Brooks,L.D. and Chakravarti,A. (1998) A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.*, **8**, 1229–1231.
2. Cargill,M., Altshuler,D., Ireland,J., Sklar,P., Ardlie,K., Patil,N., Shaw,N., Lane,C.R., Lim,E.P., Kalyanaraman,N. *et al.* (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.*, **22**, 231–238.
3. Fernald,G.H., Capriotti,E., Daneshjou,R., Karczewski,K.J. and Altman,R.B. (2011) Bioinformatics challenges for personalized medicine. *Bioinformatics (Oxford, England)*, **27**, 1741–1748.
4. Frazer,K.A., Ballinger,D.G., Cox,D.R., Hinds,D.A., Stuve,L.L., Gibbs,R.A., Belmont,J.W., Boudreau,A., Hardenbol,P., Leal,S.M. *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.
5. Wang,D.G., Fan,J.B., Siao,C.J., Berno,A., Young,P., Sapolsky,R., Ghandour,G., Perkins,N., Winchester,E., Spencer,J. *et al.* (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science (New York, N.Y.)*, **280**, 1077–1082.

6. Cotton,R.G., Auerbach,A.D., Axton,M., Barash,C.I., Berkovic,S.F., Brookes,A.J., Burn,J., Cutting,G., den Dunnen,J.T., Flicek,P. *et al.* (2008) GENETICS. The human variome project. *Science (New York, N.Y.)*, **322**, 861–862.
7. Abecasis,G.R., Altshuler,D., Auton,A., Brooks,L.D., Durbin,R.M., Gibbs,R.A., Hurles,M.E. and McVean,G.A. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.
8. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.
9. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
10. Yip,Y.L., Famiglietti,M., Gos,A., Duek,P.D., David,F.P., Gateau,A. and Bairoch,A. (2008) Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase. *Hum. Mut.*, **29**, 361–366.
11. Stenson,P.D., Mort,M., Ball,E.V., Shaw,K., Phillips,A. and Cooper,D.N. (2014) The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet.*, **133**, 1–9
12. Landrum,M.J., Lee,J.M., Benson,M., Brown,G., Chao,C., Chitipiralla,S., Gu,B., Hart,J., Hoffman,D., Hoover,J. *et al.* (2016) ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.*, **44**:D862–D868.
13. Karchin,R. (2009) Next generation tools for the annotation of human SNPs. *Brief. Bioinform.*, **10**, 35–52.
14. Mooney,S. (2005) Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. *Brief. Bioinform.*, **6**, 44–56.
15. Tavtigian,S.V., Greenblatt,M.S., Lesueur,F. and Byrnes,G.B. (2008) In silico analysis of missense substitutions using sequence-alignment based methods. *Hum. Mut.*, **29**, 1327–1336.
16. Ferrer-Costa,C., Gelpi,J.L., Zamakola,L., Parraga,I., de la Cruz,X. and Orozco,M. (2005) PMUT: a web-based tool for the annotation of pathological mutations on proteins. *Bioinformatics (Oxford, England)*, **21**, 3176–3178.
17. Bromberg,Y., Yachdav,G. and Rost,B. (2008) SNAP predicts effect of mutations on protein function. *Bioinformatics (Oxford, England)*, **24**, 2397–2398.
18. Ng,P.C. and Henikoff,S. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, **31**, 3812–3814.
19. Adzhubei,I.A., Schmidt,S., Peshkin,L., Ramensky,V.E., Gerasimova,A., Bork,P., Kondrashov,A.S. and Sunyaev,S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.
20. Choi,Y. and Chan,A.P. (2015) PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics (Oxford, England)*, **31**, 2745–2747.
21. Chun,S. and Fay,J.C. (2009) Identification of deleterious mutations within three human genomes. *Genome Res.*, **19**, 1553–1561.
22. Reva,B., Antipin,Y. and Sander,C. (2011) Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.*, **39**, e118.
23. Thomas,P.D. and Kejariwal,A. (2004) Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 15398–15403.
24. Shihab,H.A., Gough,J., Cooper,D.N., Stenson,P.D., Barker,G.L., Edwards,K.J., Day,I.N. and Gaunt,T.R. (2013) Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mut.*, **34**, 57–65.
25. Niroula,A., Urolagin,S. and Vihinen,M. (2015) PON-P2: prediction method for fast and reliable identification of harmful variants. *PLoS One*, **10**, e0117380.
26. Carter,H., Chen,S., Isik,L., Tyekucheva,S., Velculescu,V.E., Kinzler,K.W., Vogelstein,B. and Karchin,R. (2009) Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res.*, **69**, 6660–6667.
27. Kaminker,J.S., Zhang,Y., Watanabe,C. and Zhang,Z. (2007) CanPredict: a computational tool for predicting cancer-associated missense mutations. *Nucleic Acids Res.*, **35**, W595–W598.
28. Wainreb,G., Ashkenazy,H., Bromberg,Y., Starovolsky-Shitrit,A., Haliloglu,T., Ruppin,E., Avraham,K.B., Rost,B. and Ben-Tal,N. (2010) MuD: an interactive web server for the prediction of non-neutral substitutions using protein structural data. *Nucleic Acids Res.*, **38**, W523–W528.
29. Kircher,M., Witten,D.M., Jain,P., O'Roak,B.J., Cooper,G.M. and Shendure,J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310–315.
30. Calabrese,R., Capriotti,E., Fariselli,P., Martelli,P.L. and Casadio,R. (2009) Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum. Mut.*, **30**, 1237–1244.
31. Capriotti,E., Arbiza,L., Casadio,R., Dopazo,J., Dopazo,H. and Marti-Renom,M.A. (2008) Use of estimated evolutionary strength at the codon level improves the prediction of disease-related protein mutations in humans. *Hum. Mut.*, **29**, 198–204.
32. Karchin,R., Diekhans,M., Kelly,L., Thomas,D.J., Pieper,U., Eswar,N., Haussler,D. and Sali,A. (2005) LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. *Bioinformatics (Oxford, England)*, **21**, 2814–2820.
33. Yue,P., Melamud,E. and Moult,J. (2006) SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics*, **7**, 166.
34. Dong,C., Wei,P., Jian,X., Gibbs,R., Boerwinkle,E., Wang,K. and Liu,X. (2015) Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.*, **24**, 2125–2137
35. Schwarz,J.M., Rodelsperger,C., Schuelke,M. and Seelow,D. (2010) MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods*, **7**, 575–576.
36. Jagadeesh,K.A., Wenger,A.M., Berger,M.J., Guturu,H., Stenson,P.D., Cooper,D.N., Bernstein,J.A. and Bejerano,G. (2016) M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat. Genet.*, **48**, 1581–1586
37. Gonzalez-Perez,A. and Lopez-Bigas,N. (2011) Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am. J. Hum. Genet.*, **88**, 440–449.
38. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
39. Suzek,B.E., Wang,Y., Huang,H., McGarvey,P.B. and Wu,C.H. (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics (Oxford, England)*, **31**, 926–932.
40. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402
41. Lassmann,T. and Sonnhammer,E.L. (2005) Kalign–an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*, **6**, 298.
42. Breiman,L., (2001) Random Forests. *Mach. Learn.*, **45**, 5–32
43. Wang,K., Li,M. and Hakonarson,H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.
44. Uniprot (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.
45. Berman,H., Henrick,K., Nakamura,H. and Markley,J.L. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.*, **35**, D301–D303.
46. Aken,B.L., Achuthan,P., Akanni,W., Amode,M.R., Bernsdorff,F., Bhai,J., Billis,K., Carvalho-Silva,D., Cummins,C., Clapham,P. *et al.* (2017) Ensembl 2017. *Nucleic Acids Res.*, **45**, D635–D642.
47. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M., Sangrador-Vegas,A. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
48. Mitchell,A., Chang,H.Y., Daugherty,L., Fraser,M., Hunter,S., Lopez,R., McAnulla,C., McMenamin,C., Nuka,G., Pesseat,S. *et al.* (2015) The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.*, **43**, D213–D221.
49. Mosca,R., Ceol,A. and Aloy,P. (2013) Interactome3D: adding structural details to protein networks. *Nat. Methods*, **10**, 47–53.

C. PMut-S: protein-specific mutation pathology predictors

PMut-S: protein-specific mutation pathology predictors

Víctor López-Ferrando [1,2], Juan Rogriguez-Rivas [1], Xavier de la Cruz [3,4], Alfonso Valencia [1,4] Modesto Orozco [2,4,5,6], and Josep Ll. Gelpí [1,2,6]*

[1] Barcelona Supercomputing Center (BSC), Barcelona, Spain.
[2] Joint Program BSC-CRG-IRB Research Program for Computational Biology, Barcelona, Spain.
[3] Vall d'Hebron Institute of Research (VHIR), Universitat Autònoma de Barcelona, Barcelona, Spain.
[4] ICREA, Barcelona, Spain.
[5] Institute for Research in Biomedicine (IRB) Barcelona, The Barcelona Institute of Science and Technology, Barcelona. Spain.
[6] Dept. of Biochemistry and Molecular Biomedicine, University of Barcelona, Barcelona, Spain.

*To whom correspondence should be addressed.

Abstract

Motivation: Mutation pathology predictors are a key tool for the interpretation of the ever growing variation data available. However, the still limited performance of these predictors hampers its use in clinical settings. It is generally accepted that the use of protein-specific predictors can improve pathology prediction in well-known protein families. Recent advancements in sequencing technology have fostered the collection of huge variant catalogs and the newly developed epistatic models offer an opportunity to test the performance of protein-specific predictors as a more general solution for pathology assessment.
Results: We trained and evaluated 215 protein-specific predictors using PMut-S, a Random Forest predictor based on 15 features accounting for sequence conservation and co-evolutive models fitness. PMut-S improves the prediction over 13 general-purpose predictors we compared it to, and greatly improves the worse predicted proteins in a one-on-one comparison with PMut predictor. Compared to 8 published family specific predictors, we find PMut-S matches their performance in most cases, and improves their accuracy in some of them. Finally, we show PMut-S behaves better when used to predict the pathology of common variants from healthy individuals.
Availability: Precomputed predictions for all variants in the 215 studied proteins: http://mmb.irbbarcelona.org/PMut-S/.
Contact: gelpi@ub.edu

1 Introduction

In silico single amino acid variants (SAVs) pathogenicity predictors have extensively been used in research, and have become a fundamental tool in the interpretation of the large amounts of genetic variation data reported by last generation sequencing techniques. However, their accuracies are still below the requirements of clinical application (Riera *et al.*, 2014).

Predictors have followed different approaches: starting from rule-based predictors such as SIFT (Ng and Henikoff, 2003) and PROVEAN (Choi and Chan, 2015), recent studies have shifted to a supervised learning paradigm, with predictors such as CADD (Kircher *et al.*, 2014), LRT (Chun and Fay, 2009), FATHMM (Shihab *et al.*, 2012), M-CAP (Jagadeesh *et al.*, 2016), MetaSVM (Dong *et al.*, 2015a), MutationAssessor (Reva *et al.*, 2011), MutationTaster (Schwarz *et al.*, 2010), PolyPhen-2 (Adzhubei *et al.*, 2010), PON-P2 (Niroula *et al.*, 2015), RadialSVM (Dong *et al.*, 2015b) and PMut2017 (López-Ferrando *et al.*, 2017). Also, different meta-predictors, wich derive a consensus score from other predictors' output have been developed, like Condel (González-Pérez and López-Bigas, 2011), PredictSNP (Bendl *et al.*, 2014) and MetaSNP (Capriotti *et al.*, 2013). All these general-purpose predictors have been trained on wide variantion datasets and are meant to be applied to any human protein. Since general-purpose predictors target an ideal average protein, their performance in specific protein families is diverse (Li *et al.*, 2014; Leong *et al.*, 2015), and it has been found that for some families all of them consistently performed badly (López-Ferrando *et al.*, 2017).

An alternative approach to variant pathology prediction is the building of predictors specific for each gene, protein or protein family. Several studies have confirmed that specific predictors can detect singular functional trends and perform better than general-purpose predictors (Riera *et al.*, 2016; Crockett *et al.*, 2012; Torkamani and Schork, 2007). For example, wKinMut-2 (Vazquez *et al.*, 2015) is a Random Forest (Breiman, 2001) predictor specialized on protein kinases mutations; kvSNP (Stead *et al.*, 2011) is a predictor aimed at predicting the effect of variants in voltage-gated potassium channels; MutaCYP (Fechter and Porollo, 2014) –based on neural networks– specializes in mutations affecting Cytochrome P450 proteins; also based on neural networks, Riera *et al.* (2015) build a specific predictor targeting Fabry disease related mutations; an ad-hoc method named SAVER (Adebali *et al.*, 2016) specializes in mutations related to Niemann-Pick disease; HApredictor (Hamasaki-Katagiri *et al.*, 2013) is specialized on the Coagulation Factor VII mutations causing Hemophilia A disease. The generation of such predictors often requires an extensive analysis of the additional information available and implies a strong manual effort on data curation.

Most of the previously cited predictors, general or specific, rely heavily on sequence conservation features, derived from multiple sequence alignments (MSA). Recent works have shown that statistical models which consider not only conservation but residue covariation reproduce more exactly experimental fitness landscapes (Figliuzzi *et al.*, 2016; Cheng *et al.*, 2016; Hopf *et al.*, 2017; Flynn *et al.*, 2017; Louie *et al.*, 2018) and can improve the prediction of pathogenicity (Flynn *et al.*, 2017; Nielsen *et al.*, 2017). The inclusion of covariation data permits to capture non-trivial dependencies between positions that might be valuable in the prediction of the mutational effect. For instance, one amino acid may be common on a protein family in a given position only when a second position is populated by a specific amino acid, and uncommon otherwise. A key aspect of these models, first introduced with high success in the context of three dimensional contact prediction (Weigt *et al.*, 2009; Morcos *et al.*, 2011), is their ability to minimize the influence of widespread transitive correlations (Weigt *et al.*, 2009; Morcos *et al.*, 2011). The downside is that a large number of parameters are required and, thus, these methods are typically restricted to large protein families for which abundant sequence data is available (Weigt *et al.*, 2009; Morcos *et al.*, 2011; Hopf *et al.*, 2017). Still, the information provided by these models can be beneficial when enough sequence data is available and, therefore, we expect its inclusion to improve the prediction performance in this range of cases.

In this work, we have explored the usability of protein-specific predictors generated in a automatic manner (PMut-S), as a replacement for protein families where general predictors tend to give poor results. Also, we have evaluated the usefullness of covariation data in this context. PMut-S is based on a Random Forest classifier, trained using 15 features obtained from both single amino acid conservation and epistatic effect predictions. We trained 215 protein-specific predictors using this protocol, compared them to both general-purpose and specific predictors, and assessed their behaviour when predicting the pathogenicity of variants identified in large sequencing studies.

2 Materials and methods

2.1 Datasets

The main source of annotated variants used in the study where taken from the SwissVar database (Yip *et al.*, 2008, April 2018 release). For the purpose of this study, we selected all the variants belonging to 215 genes for which SwissVar contains 30 or more mutations annotated as deleterious (Li *et al.*, 2014; Fechter and Porollo, 2014). To balance the number of disease and neutral mutations for each protein family we used variants reported in ExAC (Lek *et al.*, 2016) as needed to match the number of pathological mutations already present in the dataset. See Supplementary Figures 1 and 2 for a detailed description of the dataset.

2.2 Features

A total of 15 numerical features to describe each variant were computed. 11 of these features are column conservation measures derived from sequence alignments obtained from PSI-Blast (Altschul *et al.*, 1997) searches over UniRef100 and UniRef90 cluster databases (Suzek *et al.*, 2015), and multiple sequence alignments computed by Kalign2 (Lassmann and Sonnhammer, 2005) over the results of the PSI-Blast previous searches. The alignments are filtered using different criteria such as having a PSI-Blast e-value below a threshold, or keeping only human sequences, and thence some features are computed, such as the proportion of wild-type or mutated amino acids in the column.

The remaining 4 features are derived from the EVmutation model (Hopf *et al.*, 2017), which is based on Pfam domain alignments (Finn *et al.*, 2016). These features include classic column conservation, frequency of the substitution, effect according to a classic independent conservation model and predicted effect in the novel epistatic model, which incorporates pairwise epistasis to the estimation of mutation effects.

See Supplementary Table 1 to find the details of all the features calculated and their average relative importance in the final predictors.

2.3 Predictor

PMut-S (for Specific PMut), encloses a set of predictors (one for each protein) based on the Random Forest (Breiman, 2001) classifier trained on the previous data set and features. Building and evaluation of the predictors was performed with the package pyMut (López-Ferrando *et al.*, 2017). Random Forest was provided by the scikit-learn Python package (Pedregosa *et al.*, 2011) and the standard scientific Python stack including SciPy, NumPy, Pandas, IPython and Matplotlib (Jones *et al.*, 2001; van der Walt *et al.*, 2011; McKinney, 2010; Pérez and Granger, 2007; Hunter, 2007). The classifier was optimized by searching the parameter space and the selected parametrization can be found in Supplementary Table 2. The ouptput of the predictor is a decimal number in the range $[0, 1]$, where scores < 0.5 are neutral predictions and scores > 0.5 are pathological predictions.

2.4 Performance assessment

Predictors are evaluated following a leave-one-out approach (Kearns and Ron, 1999). Assessment metrics include accuracy, specificity, sensitivity, and the Matthews correlation coefficient (Matthews, 1975). The same process is executed with the general-purpose PMut2017 predictor(López-Ferrando *et al.*, 2017), in order to get a consistent comparison using the same methodology.

For comparison purposes, we obtained predictions for our dataset from different general-purpose predictors using their respective web services: PROVEAN (Choi and Chan, 2015), PolyPhen-2 (**?**), CADD (Kircher *et al.*, 2014), Condel (González-Pérez and López-Bigas, 2011), M-CAP (Jagadeesh *et al.*, 2016) and FATHMM (Shihab *et al.*, 2012). Using ANNOVAR (Wang *et al.*, 2010), which includes a database with functional annotations of variants, we also obtained predictions from MutationTaster (Schwarz *et al.*, 2010), LRT (Chun and Fay, 2009), SIFT (Ng and Henikoff, 2003), MutationAssessor (Reva *et al.*, 2011), MetaSVM and MetaLR (Dong *et al.*, 2015a).

2.5 PMut-S website

The 215 predictors were used to train all possible variants of their respective protein (a total of 4 035 353 variants). All these predictions, along with the features that lead to these predictions and the training data are available at http://mmb.pcb.ub.edu/PMut-S/.

3 Results and discussion

The performance of the PMut-S predictor has been assessed at three levels: 1) Comparison with the PMut2017 general predictor, following the same leave-one-out validation strategy; 2) Comparison with PMut-S to 12 general, and 8 specific predictors; and 3) analyzing PMut-S' behavior when predicting SAVs from healthy individuals.

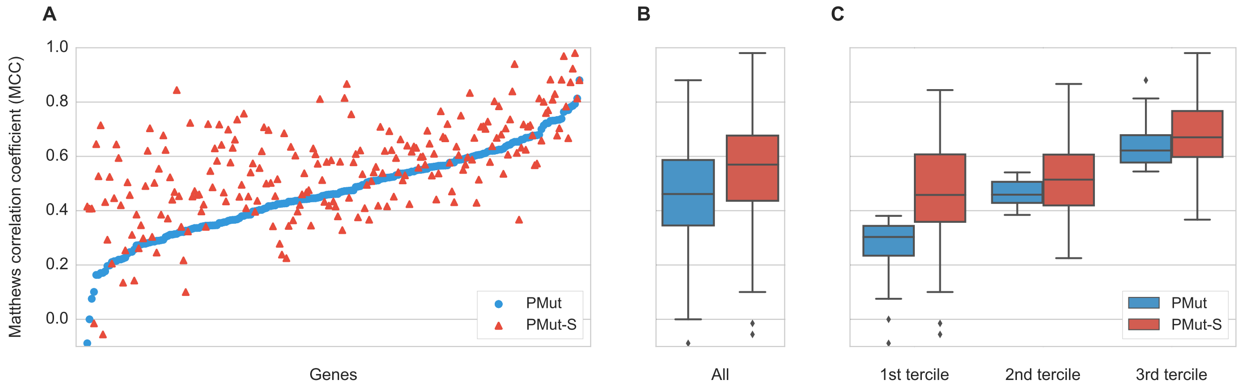


Fig. 1. A: Per-protein Matthews correlation coefficient (MCC) comparison between PMut and PMut-S (plotted in increasing PMut score). B. Summary of the PMut and PMut-S per-protein MCC distribution in a boxplot. C. Comparison splited in PMut'S MCC terciles; we can appreciate that PMut-S improvement over PMut-S is much more relevant in the first tercile, that is, PMut-S improves PMut predictions specially in the cases where PMut performs worse.

3.1 Dataset building

We used the SwissVar database (Yip *et al.*, 2008) as the main source of annotated variants. SwissVar is comprised of 38,490 neutral and 28,790 disease variants from 12,234 human proteins. The analysis of such variants shows that they are not evenly distributed among all the proteins. 90% of the disease causing mutations are concentrated in only 1,000 proteins, while most of the neutral variants belong to other proteins, of which no disease causing variant is reported (Supplementary Figure 1A). In order to maximize the number of valid mutations to be included in the study, we selected all the variants belonging to 215 genes for which SwissVar contains 30 or more deleterious mutations (Li *et al.*, 2014; Fechter and Porollo, 2014). In order to improve the predictor accuracy and also the robustness of its preformance evaluatoin, it is desirable to have a balanced number of disease and neutral mutatoins for each protein (He and Garcia, 2009). To make up the lack of neutral mutations in these genes, two main approaches are usually used: either adding variants found in homologs with high sequence similarity (Riera *et al.*, 2015; Stead *et al.*, 2011) or considering common SNPs as neutral mutations (Kobayashi *et al.*, 2017). The recent availability of large variant databases enables us to take the second approach. We resorted to ExAC (Lek *et al.*, 2016), a database containing single amino acid variants found in 60,706 healthy individuals. As reported by Kobayashi *et al.* (2017) and confirmed in Supplementary Figure 2, ExAC variants with higher allele frequency can be reliably considered neutral. We added as many ExAC variants (the most common first) as needed to match the number of pathological mutations in our dataset, hence obtaining the balanced training set we were seeking (Supplementary Figure 1B).

3.2 PMut-S versus PMut

Figure 1A shows the comparison of PMut versus PMut-S for each of the 215 genes evaluated. In average, we see an improvement of 0.1 in the MCC for each gene; Figure 1B sumarizes the distribution of these differences in a boxplot. But this improvement is not evenly distributed; Figure 1C shows a greater PMut-S' performance increase in the cases where PMut shows the worst behaviour (1st tercile).

Three main features in PMut-S can contribute to the enhanced performance: 1) the addition of neutral variants from the ExAC database, 2) the use of epistatic effect model features, and 3) the use of protein-specific predictors.

Figure 2 evaluates the contribution of each addition. The second boxplot shows that adding neutral variants from ExAC worsens slightly the performance of PMut. This can be attributed to the fact that PMut was trained with an already wide and well balanced training set, while the addition of these variants would make a stronger effect when training specific predictors due to their lack of balance. The third boxplot indicates a small improvement to the general-purpose PMut predictor upon the introduction of epistatic effects.

The fourth boxplot in Figure 2 shows the performance of gene-specific predictors lacking coevolution features in their training. Finally, the last boxplot shows that the addition of epistatic effect features to the training is decisive in the performance improvement shown by PMut-S.

As commented, it is clear that the addition of coevolutive conservation features enhances the prediction, but it's important to note that for more than half of the proteins, less than 50% of their variants have these features computed. Supplementary Figure 3 shows the same analysis restricted to proteins with most of their coevolution features available.

3.3 PMut-S versus other general-purpose predictors

We compared the per-protein performance of PMut-S to 12 other general-purpose predictors (Figure 3A). We observed that PMut-S obtains the best results in this comparison. It is to be noted that this is a conservative comparison, as some of these predictors may have been trained using

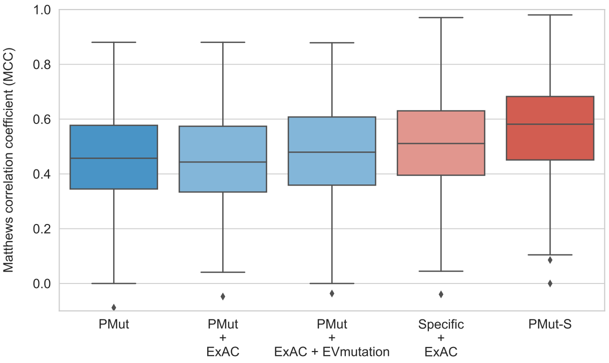


Fig. 2. Comparison of per-protein MCC distribution using 1) the general-purpose PMut predictor, 2) the PMut predictor trained with the addtion of neutral variants form ExAC, 3) PMut predictor with the addition of neutral ExAC variants and trained using co-evolution model derived features, and 4) PMut-S, protein-specific predictors using balanced datasets and co-evolution related features.

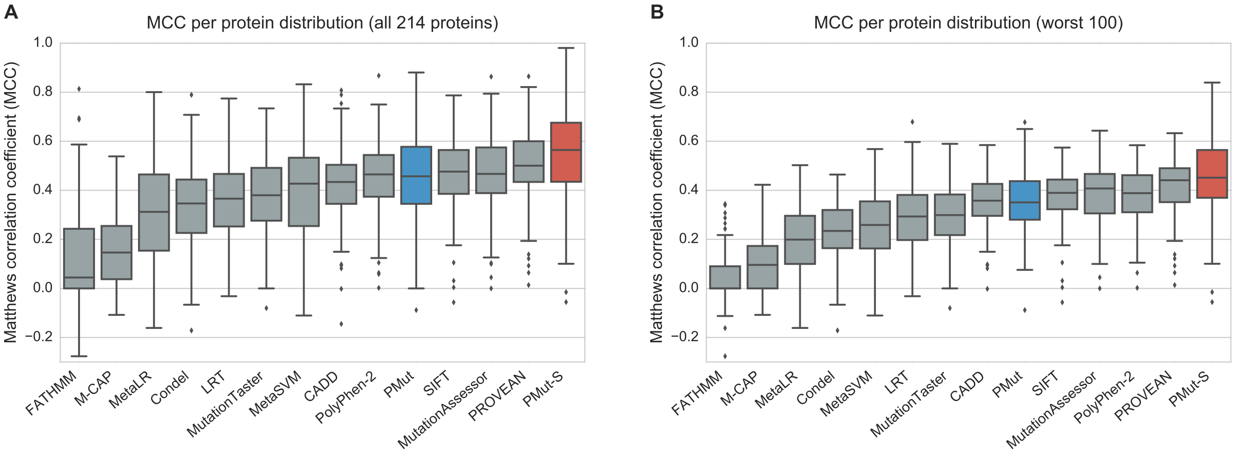


Fig. 3. A. Per-protein MCC comparison between PMut-S (red), PMut (blue), and twelve othe general-purpose predictors (gray); Supplementary Table 3 holds the numerical values of these boxplots. B. Same analysis but limiting the comparison to the 100 proteins with worse predictions in general.

some of the mutations evaluated and so the benchmark may overestimate their accuracy (Grimm *et al.*, 2015). We repeated the comparison for those genes that show poorer predictions in general (Figure 3B) and checked that in these cases, PMut-S also obtains more accurate predictions than the rest.

3.4 PMut-S versus other specific predictors

We also evaluated PMut-S' performance on a set of genes for which different protein-specific or family-specific predictors have been developed. In general, we saw that PMut's MCC, evaluated using the leave-one-out validation scheme explained before compared well with the reported MCC of these methods (Table 1).

Our first analysis was on maturity-onset diabetes of the young related genes. We evaluated PMut-S over 5 MODY genes (GCK, HNF1A, INS, ABCC8, KCNJ11 –namely MODY2, MODY3, MODY10, MODY12 and MODY13). We compared the results to those reported in Li *et al.* (2014), where eleven general-purpose pathology predictors were compared, and RadialSVM (Dong *et al.*, 2015b) was found to be the most accurate predicting pathology in MODY genes. We saw that overall, PMut-S had a better MCC (0.595) compared to RadialSVM's (0.474), and also improved the per-gene MCC in 4 out of 5 cases. This is an example of the enhancement in accuracy that specific predictors can produce.

In a recent study (Leong *et al.*, 2015), five different general-purpose predictos were evaluated on the main three Long QT syndrome (LQTS) related genes: KCNQ1, KCNH2 and SCN5A. A specific consensus method is proposed for each of these genes by combining some of the 5 predictors' outputs. As seen in Table 1, PMut-S yielded a very similar performance for each of these genes, and improved the overall MCC score. Note that in the case of SCN5A, PMut-S' MCC is the lowest in the table. This gene has the particularity of being the cause of two diseases: LQTS when having a loss-of-function mutation, and the Brugada syndrome when affected by a gain-of-function mutation. This is a classic example in which most predictors and also PMut-S might fail, as they may tend to classify gain-of-function mutations as neutral.

Closely related to the previous LQTS-causing genes, Stead *et al.* (2011) studied mutations in voltage-gated potassium (Kv) channels. They presented KvSNP, a Random Forest predictor based on 5 physicochemical, structural and conservational features. KvSNP improves the performance of general-purpose predictors, achieving an MCC of 0.7, slightly higher than PMut'S MCC of 0.65.

In the same fashion as KvSNP was developed wKinMut-2 (Vazquez *et al.*, 2015), a predictor specialized in mutations of protein kinases. It is a Random Forest predictor relying on annotation with Gene Ontology terms and Pfam domains, physicochemical features of aminoacids and sequence conservation. wKinMut-2 reports an overall MCC of 0.69, very similar to PMut-S' MCC of 0.70. By comparing the MCC for each gene, we see that PMut-S achieves good performance (MCC in the 0.65-0.75 range) for the three proteins studied.

We evaluated PMut-S over a set of Cytochrome P450 monooxygenases (CYPs), and compared its performance to MutaCYP (Fechter and Porollo, 2014). CYPs are a good fit for applying specific predictors, as they present singular evolution patterns such as important highly variable regions –substrate recognition sites– thay may pass unnoticed by typical conservation analyses. Also, methods based on structural features that consider surface mutations less important may also fail at recognizing their role in protein-protein interactions and electron transfer from a redox partner. PMut-S obtains an MCC of 0.56, significantly lower thant MutaCYP's MCC of 0.70. This difference can be attributed to the fact that PMut-S was trained on a smaller dataset, as we discarded all the proteins for which less than 30 disease variants were reported –keeping only 4, whereas MutaCYP is trained using all variants from 15 CYPs.

Finally, we evaluated PMut-S' performance when predicting pathology in three monogenic diseases: Niemann-Pick disorder, caused by mutations on the NPC1 gene; Hemophilia A, due to a malfunction of Coagulation Factor VII (F8); and Fabry disease, caused by loss-of-function mutations in Alpha-galactosidase A (GLA).

Regarding NPC1, in Adebali *et al.* (2016), the gene's evolutionary history was derived from carefully crafted multiple sequence alignments, and the ad-hoc SAVER algorithm was proposed as a specific pathology predictor for this gene, reporting an MCC of 0.59. PMut-S' MCC is 0.537, which hints the importance of the multiple sequence alignments quality for identifying evolutionary trends and pathology, as explained in Adebali *et al.* (2016).

Compared to HApredictor, the F8-specific decision-tree classifier proposed in Hamasaki-Katagiri *et al.* (2013), Pmut-S' obtained better results both in terms of sensitivity (0.90 vs. 0.85) and specificity (0.81 vs. 0.79). On the other hand, PMut-S' performance didn't match that of the GLA-specific neural network based predictor described in Riera *et al.* (2015), which obtains an MCC in the range of $0.56 - 0.72$, higher than

Table 1. Performance comparison between PMut-S and eight other specific predictors.

Family	Associated diseases	Genes	PMut-S					Literature		
			Acc.	Sens.	Spec.	AUC	MCC	MCC	Method	References
MODY genes	Maturity-onset diabetes of the young	ABCC8	0.74	0.78	0.71	0.74	0.49	0.44	RadialSVM	Li *et al.* (2014), Dong *et al.* (2015b)
		GCK	0.86	0.89	0.82	0.86	0.72	0.60		
		HNF1A	0.75	0.70	0.80	0.75	0.50	0.37		
		INS	0.70	0.68	0.73	0.70	0.41	0.76		
		KCNJ11	0.86	0.89	0.83	0.86	0.72	0.57		
		Overall	0.80	0.82	0.78	0.80	0.60	0.47		
Kinase superfamily	Chronic myelogenous leukaemia, gastrointestinal stromal tumours, etc.	BTK	0.89	0.92	0.80	0.86	0.72		wKinMut-2	Vazquez *et al.* (2015)
		FGFR1	0.82	0.81	0.84	0.82	0.64			
		FGFR2	0.86	0.83	0.89	0.86	0.72			
		Overall	0.86	0.86	0.85	0.86	0.71	0.69		
LQTS 1-3 genes	Long QT syndrome	KCNH2	0.81	0.74	0.89	0.81	0.63	0.62	Combination of 5 predictors	Leong *et al.* (2015)
		KCNQ1	0.80	0.77	0.84	0.80	0.61	0.70		
		SCN5A	0.69	0.64	0.75	0.69	0.39	0.32		
		Overall	0.76	0.71	0.82	0.76	0.53	0.44		
Voltage-gated potassium (Kv) channels	Cardiac arrhythmogenesis, LQTS, epilepsy, etc.	KCNH2	0.81	0.74	0.89	0.81	0.63		kvSNP	Stead *et al.* (2011)
		KCNQ1	0.80	0.77	0.84	0.80	0.61			
		KCNQ2	0.94	0.94	0.94	0.94	0.88			
		Overall	0.82	0.77	0.87	0.82	0.65	0.70		
Cytochrome P450	Congenital adrenal hyperplassia, etc.	CYP11B1	0.71	0.64	0.78	0.71	0.42		MutaCYP	Fechter and Porollo (2014)
		CYP17A1	0.75	0.81	0.69	0.75	0.50			
		CYP1B1	0.72	0.68	0.78	0.72	0.45			
		CYP21A2	0.88	0.88	0.88	0.88	0.76			
		Overall	0.78	0.76	0.80	0.78	0.56	0.70		
NPC1 gene	Niemann-Pick disorder	NPC1	0.77	0.74	0.79	0.77	0.54	0.59	SAVER	Adebali *et al.* (2016)
Coagulation Factor VII	Hemophilia A	F8	0.85	0.90	0.81	0.85	0.71	Sens. = 0.85 Spec. = 0.79	HApredictor	Hamasaki-Katagiri *et al.* (2013)
Alpha-galactosidase A	Fabry disease	GLA	0.85	0.95	0.52	0.74	0.55	0.56 − 0.72	V7, V8	Riera *et al.* (2015)

PMut'S MCC of 0.55. In this case, the use of first hand variation data by the authors undoubtedly contributes to their improved performance.

In the eight cases that we have evaluated, we saw that PMut-S improves (with an MCC of at least 0.1 points higher) the performance of two specific predictors, performs similar to other 4, and performs worse than two specific predictors. It is important to note, however, that these specific predictors were generally trained on wider mutation sets, not entirely contained in current mutations databases such as SwissVar. Even more, some of these predictors used for their training hand crafted alignments, functional annotations of amino acids, relevant evolutionary information and other expert knowledge. It is thus remarkable that PMut-S is able to match most of these specific predictors' performance following the same systematic method ot train the predictors. Although these results confirm the extraordinary validity of manual-crafted predictors, it is relevant to note that PMut-s, a set of completely automated predictors, performs at similar or even higher level in most of the analyzed cases.

3.5 Prediction of ExAC variants

Most of the single amino acid variants (SAVs) found in healthy humans are neutral variants –not causing any diease (Kobayashi *et al.*, 2017). It is thus interesting to check how predictors classify these variants. Figure 4A shows the prediction score distribution of PMut for 7 100 034 variants from ExAC. As expected, most of the variants are predicted as neutral (score < 0.5). However, when we limit our analysis to the 215 genes of our present study, we see that PMuts score distribution is almost uniform (Figure 4B), that is, variants are evenly classified as either neutral or pathological. We believe this abnormal distribution is due to the overrepresentation of deleterious SAVs in the training of PMut, specially in the case of these proteins. On the other hand, we also see in Figure 4B that PMut-S' scores distribution is shifted towards the neutral side, conforming with the expected distribution.are This results stress the relevance of using balanced datasets in the training of this type of predictors.

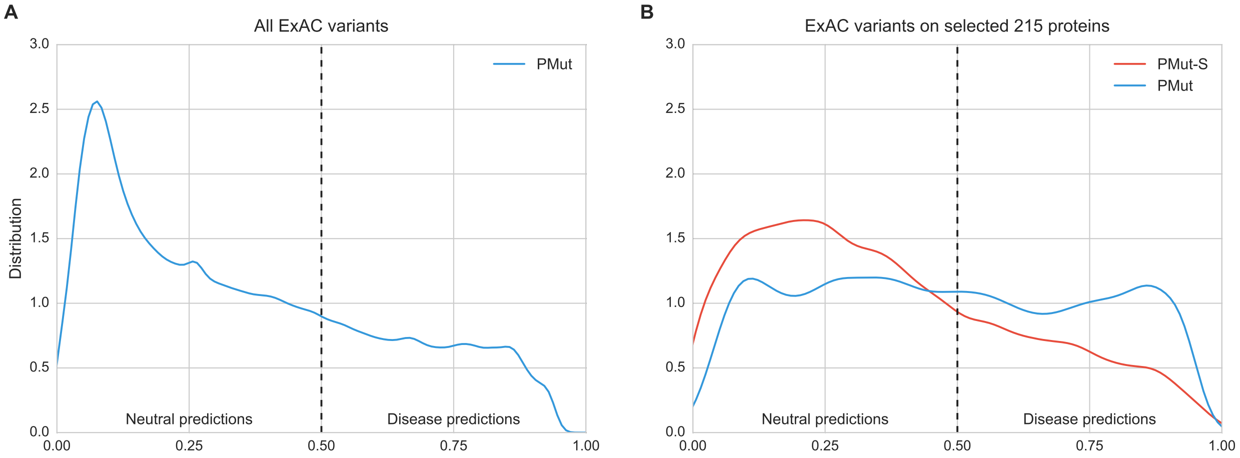


Fig. 4. A. Distribution of the PMut score for all ExAC variants. B. PMut and PMut-S scores distribution across 63 339 variants from the ExAC database on the 215 proteins studied (variants used in PMut-S' training were excluded from the analysis).

4 Conclusions

In this study we present PMut-S, a set of protein-specific mutation pathology predictors, which we have evaluated with a set of 4 035 353 variants over 215 proteins. In a thorough comparison with state-of-the-art general-purpose predictor PMut2017, we observe that PMut-S can achieve better predictions, specially in the cases where PMut performs worse.

By comparing PMut-S to 12 other popular general predictors, we confirmed that PMut-S yields more robust predictions overall, especially with protein families that traditionaly fail. We also evaluated PMut-S on 8 protein families for which specific predictors are available. We found that our automated approach to train specific predictors obtains significantly similar results compared to these specific predictors at a fraction of the time required to prepare them.

We believe this improvement in prediction accuracy is due to two main factors. First, PMut-S is trained using carefully balanced datasets, where most of the neutral variants where obtained from recently released human variation databases. We think this means is more reliable than traditional homology based neutral variant inference. Second, the addition of information provided by covariation based models is a key feature that improved overall prediction performance. The application of these models in this context is recent, and therefore, methodological improvements are foreseeable. In addition, given the large amount of sequence data required by this type of methods, the rapid growth of sequence databases will likely extend the scope of application and improve the prediction performance.

All the data used in this study, together with the code and a dataset of 5,479,087 precomputed predictions on the 215 proteins studied is available for the research community in http://mmb.irbbarcelona.org/PMut-S/.

Acknowledgements

We thank Adam Hospital and Pau Andrio for their technical assistance.

Funding

Spanish Ministry of Science [BIO2015-64802-R-32868, SEV-2015-0493, TIN2012-34557, TEC2015-67774-C2-2-R, SAF2016-80255-R]; Catalan Government [2014-SGR-134, 2014-SGR-1051]; Instituto de Salud Carlos III-Instituto Nacional de Bioinformática [INB; PT13/0001/0019, PT13/0001/0028]; European Union, H2020 programme [Elixir-Excelerate: 676559; BioExcel: 674728 and MuG: 676566]; La Caixa Foundation fellowship [to V.L.F.]. Funding for open access charge: European Union and Spanish Ministry of Science.

References

Adebali, O., Reznik, A. O., Ory, D. S., and Zhulin, I. B. (2016). Establishing the precise evolutionary history of a gene improves prediction of disease-causing missense mutations. *Genetics in Medicine*, **18**(10), 1029–1036.

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., and Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature methods*, **7**(4), 248–249.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, **25**(17), 3389–3402.

Bendl, J., Stourac, J., Salanda, O., Pavelka, A., Wieben, E. D., Zendulka, J., Brezovsky, J., and Damborsky, J. (2014). PredictSNP: Robust and Accurate Consensus Classifier for Prediction of Disease-Related Mutations. *PLoS Computational Biology*, **10**(1).

Breiman, L. (2001). Random Forests. *Machine Learning*, **45**(1), 5–32.

Capriotti, E., Altman, R. B., and Bromberg, Y. (2013). Collective judgment predicts disease-associated single nucleotide variants. *BMC Genomics*, **14**(Suppl 3), S2.

Cheng, R. R., Nordesjö, O., Hayes, R. L., Levine, H., Flores, S. C., Onuchic, J. N., and Morcos, F. (2016). Connecting the Sequence-Space of Bacterial Signaling Proteins to Phenotypes Using Coevolutionary Landscapes. *Molecular Biology and Evolution*, **33**(12), 3054–3064. 00009.

Choi, Y. and Chan, A. P. (2015). PROVEAN web server: A tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*, **31**(16), 2745–2747.

Chun, S. and Fay, J. C. (2009). Identification of deleterious mutations within three human genomes. *Genome Research*, **19**(9), 1553–1561.

Crockett, D. K., Lyon, E., Williams, M. S., Narus, S. P., Facelli, J. C., and Mitchell, J. A. (2012). Utility of gene-specific algorithms for predicting pathogenicity of uncertain gene variants. *Journal of the American Medical Informatics Association*, **19**(2), 207–211.

Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., and Liu, X. (2015a). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Human Molecular Genetics*, **24**(8), 2125–2137.

Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., and Liu, X. (2015b). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Human Molecular Genetics*, **24**(8), 2125–2137. 00211.

Fechter, K. and Porollo, A. (2014). MutaCYP: Classification of missense mutations in human cytochromes P450. *BMC Medical Genomics*, **7**, 47.

Figliuzzi, M., Jacquier, H., Schug, A., Tenaillon, O., and Weigt, M. (2016). Coevolutionary Landscape Inference and the Context-Dependence of Mutations in Beta-Lactamase TEM-1. *Molecular Biology and Evolution*, **33**(1), 268–280.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J., and Bateman, A. (2016). The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Research*, **44**(D1), D279–D285.

Flynn, W. F., Haldane, A., Torbett, B. E., and Levy, R. M. (2017). Inference of Epistatic Effects Leading to Entrenchment and Drug Resistance in HIV-1 Protease. *Molecular Biology and Evolution*, **34**(6), 1291–1306. 00004.

González-Pérez, A. and López-Bigas, N. (2011). Improving the Assessment of the Outcome of Nonsynonymous SNVs with a Consensus Deleteriousness Score, Condel. *American Journal of Human Genetics*, **88**(4), 440–449.

Grimm, D. G., Azencott, C.-A., Aicheler, F., Gieraths, U., MacArthur, D. G., Samocha, K. E., Cooper, D. N., Stenson, P. D., Daly, M. J., Smoller, J. W., Duncan, L. E., and Borgwardt, K. M. (2015). The Evaluation of Tools Used to Predict the Impact of Missense Variants Is Hindered by Two Types of Circularity. *Human Mutation*, **36**(5), 513–523.

Hamasaki-Katagiri, N., Salari, R., Wu, A., Qi, Y., Schiller, T., Filiberto, A. C., Schisterman, E. F., Komar, A. A., Przytycka, T. M., and Kimchi-Sarfaty, C. (2013). A Gene-Specific Method for Predicting Hemophilia-Causing Point Mutations. *Journal of molecular biology*, **425**(21), 4023–4033.

He, H. and Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, **21**(9), 1263–1284.

Hopf, T. A., Ingraham, J. B., Poelwijk, F. J., Schärfe, C. P. I., Springer, M., Sander, C., and Marks, D. S. (2017). Mutation effects predicted from sequence co-variation. *Nature Biotechnology*, **35**(2), 128–135.

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, **9**(3), 90–95.

Jagadeesh, K. A., Wenger, A. M., Berger, M. J., Guturu, H., Stenson, P. D., Cooper, D. N., Bernstein, J. A., and Bejerano, G. (2016). M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nature Genetics*, **48**(12), 1581–1586.

Jones, E., Oliphant, T., Peterson, P., and others (2001). SciPy: Open source scientific tools for Python. http://www.scipy.org/. [Online; accessed <today>].

Kearns, M. and Ron, D. (1999). Algorithmic Stability and Sanity-Check Bounds for Leave-One-Out Cross-Validation. *Neural Computation*, **11**(6), 1427–1453.

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*, **46**(3), 310–315.

Kobayashi, Y., Yang, S., Nykamp, K., Garcia, J., Lincoln, S. E., and Topper, S. E. (2017). Pathogenic variant burden in the ExAC database: An empirical approach to evaluating population data for clinical variant interpretation. *Genome Medicine*, **9**.

Lassmann, T. and Sonnhammer, E. L. (2005). Kalign – an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*, **6**, 298.

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., Cooper, D. N., Deflaux, N., DePristo, M., Do, R., Flannick, J., Fromer, M., Gauthier, L., Goldstein, J., Gupta, N., Howrigan, D., Kiezun, A., Kurki, M. I., Moonshine, A. L., Natarajan, P., Orozco, L., Peloso, G. M., Poplin, R., Rivas, M. A., Ruano-Rubio, V., Rose, S. A., Ruderfer, D. M., Shakir, K., Stenson, P. D., Stevens, C., Thomas, B. P., Tiao, G., Tusie-Luna, M. T., Weisburd, B., Won, H.-H., Yu, D., Altshuler, D. M., Ardissino, D., Boehnke, M., Danesh, J., Donnelly, S., Elosua, R., Florez, J. C., Gabriel, S. B., Getz, G., Glatt, S. J., Hultman, C. M., Kathiresan, S., Laakso, M., McCarroll, S., McCarthy, M. I., McGovern, D., McPherson, R., Neale, B. M., Palotie, A., Purcell, S. M., Saleheen, D., Scharf, J. M., Sklar, P., Sullivan, P. F., Tuomilehto, J., Tsuang, M. T., Watkins, H. C., Wilson, J. G., Daly, M. J., MacArthur, D. G., and Consortium, E. A. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**(7616), 285–291.

Leong, I. U., Stuckey, A., Lai, D., Skinner, J. R., and Love, D. R. (2015). Assessment of the predictive accuracy of five in silico prediction tools, alone or in combination, and two metaservers to classify long QT syndrome gene mutations. *BMC Medical Genetics*, **16**.

Li, Q., Liu, X., Gibbs, R. A., Boerwinkle, E., Polychronakos, C., and Qu, H.-Q. (2014). Gene-Specific Function Prediction for Non-Synonymous Mutations in Monogenic Diabetes Genes. *PLoS ONE*, **9**(8).

López-Ferrando, V., Gazzo, A., de la Cruz, X., Orozco, M., and Gelpí, J. L. (2017). PMut: A web-based tool for the annotation of pathological variants on proteins, 2017 update. *Nucleic Acids Research*, **45**(W1), W222–W228.

Louie, R. H. Y., Kaczorowski, K. J., Barton, J. P., Chakraborty, A. K., and McKay, M. R. (2018). Fitness landscape of the human immunodeficiency virus envelope protein that is targeted by antibodies. *Proceedings of the National Academy of Sciences*, page 201717765. 00003.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, **405**(2), 442–451.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In S. van der Walt and J. Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 51–56.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., and Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, **108**(49), E1293–E1301. 00538.

Ng, P. C. and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*, **31**(13), 3812–3814.

Nielsen, S. V., Stein, A., Dinitzen, A. B., Papaleo, E., Tatham, M. H., Poulsen, E. G., Kassem, M. M., Rasmussen, L. J., Lindorff-Larsen, K., and Hartmann-Petersen, R. (19-Apr-2017). Predicting the impact of Lynch syndrome-causing missense mutations from structural calculations. *PLOS Genetics*, **13**(4), e1006739. 00007.

Niroula, A., Urolagin, S., and Vihinen, M. (2015). PON-P2: Prediction method for fast and reliable identification of harmful variants. *PloS one*, **10**(2), e0117380. 00062.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

Pérez, F. and Granger, B. E. (2007). IPython: A System for Interactive Scientific Computing. *Computing in Science & Engineering*, **9**(3), 21–29.

Reva, B., Antipin, Y., and Sander, C. (2011). Predicting the functional impact of protein mutations: Application to cancer genomics. *Nucleic Acids Research*, **39**(17), e118.

Riera, C., Lois, S., and de la Cruz, X. (2014). Prediction of pathological mutations in proteins: The challenge of integrating sequence conservation and structure stability principles. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, **4**(3), 249–268.

Riera, C., Lois, S., Domínguez, C., Fernandez-Cadenas, I., Montaner, J., Rodríguez-Sureda, V., and de la Cruz, X. (2015). Molecular damage in Fabry disease: Characterization and prediction of alpha-galactosidase A pathological mutations. *Proteins: Structure, Function, and Bioinformatics*, **83**(1), 91–104.

Riera, C., Padilla, N., and de la Cruz, X. (2016). The Complementarity Between Protein-Specific and General Pathogenicity Predictors for Amino Acid Substitutions. *Human Mutation*, **37**(10), 1013–1024.

Schwarz, J. M., Rödelsperger, C., Schuelke, M., and Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. *Nature Methods*, **7**(8), 575–576.

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L. A., Edwards, K. J., Day, I. N. M., and Gaunt, T. R. (2012). Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. *Human Mutation*, **34**(1), 57–65.

Stead, L. F., Wood, I. C., and Westhead, D. R. (2011). KvSNP: Accurately predicting the effect of genetic variants in voltage-gated potassium channels. *Bioinformatics*, **27**(16), 2181–2186.

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., and UniProt Consortium (2015). UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics (Oxford, England)*, **31**(6), 926–932.

Torkamani, A. and Schork, N. J. (2007). Accurate prediction of deleterious protein kinase polymorphisms. *Bioinformatics*, **23**(21), 2918–2925. 00054.

van der Walt, S., Colbert, S. C., and Varoquaux, G. (2011). The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science and Engg.*, **13**(2), 22–30.

Vazquez, M., Pons, T., Brunak, S., Valencia, A., and Izarzugaza, J. M. G. (2015). wKinMut-2: Identification and Interpretation of Pathogenic Variants in Human Protein Kinases. *Human Mutation*, **37**(1), 36–42.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, **38**(16), e164–e164.

Weigt, M., White, R. A., Szurmant, H., Hoch, J. A., and Hwa, T. (2009). Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, **106**(1), 67–72. 00558.

Yip, Y. L., Famiglietti, M., Gos, A., Duek, P. D., David, F. P. A., Gateau, A., and Bairoch, A. (2008). Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase. *Human Mutation*, **29**(3), 361–366.

Bibliography

Adebali, O., Reznik, A. O., Ory, D. S., and Zhulin, I. B. (2016). Establishing the Precise Evolutionary History of a Gene Improves Prediction of Disease-Causing Missense Mutations. *Genetics in Medicine*, 18(10):1029–1036.

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., and Sunyaev, S. R. (2010). A Method and Server for Predicting Damaging Missense Mutations. *Nature methods*, 7(4):248–249.

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J. (2002). *Molecular Biology of the Cell*. Garland, 4th edition. 00125.

Ali, H., Olatubosun, A., and Vihinen, M. (2012). Classification of Mismatch Repair Gene Missense Variants with PON-MMR. *Human Mutation*, 33(4):642–650.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Research*, 25(17):3389–3402.

Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., and Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Research*, 43(D1):D789–D798.

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., et al. (2016). Deep speech 2: End-to-end speech recognition in english and mandarin. In *International Conference on Machine Learning*, pages 173–182.

Antonarakis, S. E. and Beckmann, J. S. (2006). Mendelian disorders deserve more attention. *Nature Reviews Genetics*, 7(4):277.

Ayadi, A., Birling, M.-C., Bottomley, J., Bussell, J., Fuchs, H., Fray, M., Gailus-Durner, V., Greenaway, S., Houghton, R., Karp, N., et al. (2012). Mouse large-scale phenotyping initiatives: Overview of the European mouse disease clinic (EUMODIC) and of the Wellcome Trust Sanger Institute Mouse Genetics Project. *Mammalian genome*, 23(9-10):600–610.

Baird, P. A., Anderson, T. W., Newcombe, H. B., and Lowry, R. (1988). Genetic disorders in children and young adults: A population study. *American journal of human genetics*, 42(5):677.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature News*, 533(7604):452.

Baldi, P. and Brunak, S. (2001). *Bioinformatics: The Machine Learning Approach*. MIT press. 00027.

Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. A., and Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics*, 12(11):745–755.

Bao, L., Zhou, M., and Cui, Y. (2005). nsSNPAnalyzer: Identifying disease-associated nonsynonymous single nucleotide polymorphisms. *Nucleic Acids Research*, 33(suppl_2):W480–W482.

Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., Bely, B., et al. (2017). UniProt: The Universal Protein Knowledgebase. *Nucleic Acids Research*, 45(D1):D158–D169. 00000.

Baugh, E. H., Simmons-Edler, R., Müller, C. L., Alford, R. F., Volfovsky, N., Lash, A. E., and Bonneau, R. (2016). Robust Classification of Protein Variation Using Structural Modelling and Large-Scale Data Integration. *Nucleic Acids Research*, 44(6):2501–2513.

Bendl, J., Stourac, J., Salanda, O., Pavelka, A., Wieben, E. D., Zendulka, J., Brezovsky, J., and Damborsky, J. (2014). PredictSNP: Robust and Accurate Consensus Classifier for Prediction of Disease-Related Mutations. *PLoS Computational Biology*, 10(1).

Berman, H. M., Bourne, P. E., Westbrook, J., and Zardecki, C. (2003). The protein data bank. In *Protein Structure*, pages 394–410. CRC Press.

Blanchard, M. G., Willemsen, M. H., Walker, J. B., Dib-Hajj, S. D., Waxman, S. G., Jongmans, M. C., Kleefstra, T., et al. (2015). De novo gain-of-function and loss-of-function mutations of SCN8A in patients with intellectual disabilities and epilepsy. *Journal of Medical Genetics*, 52(5):330–337.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. (2003). The SWISS-PROT Protein Knowledgebase and Its Supplement TrEMBL in 2003. *Nucleic Acids Research*, 31(1):365–370.

Botstein, D. and Risch, N. (2003). Discovering genotypes underlying human phenotypes: Past successes for mendelian disease, future approaches for complex disease. *Nature Genetics*, 33(3s):228–237.

Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, 32(3):314–331.

Breen, M. S., Kemena, C., Vlasov, P. K., Notredame, C., and Kondrashov, F. A. (2012). Epistasis as the primary factor in molecular evolution. *Nature*, 490(7421):535–538.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Bromberg, Y., Yachdav, G., and Rost, B. (2008). SNAP predicts effect of mutations on protein function. *Bioinformatics*, 24(20):2397–2398.

Bustamante, C. D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M. T., Glanowski, S., Tanenbaum, D. M., et al. (2005). Natural selection on protein-coding genes in the human genome. *Nature*, 437(7062):1153.

Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P. L., and Casadio, R. (2009). Functional annotations improve the predictive score of human disease-related mutations in proteins. *Human Mutation*, 30(8):1237–1244.

Capriotti, E., Altman, R. B., and Bromberg, Y. (2013). Collective Judgment Predicts Disease-Associated Single Nucleotide Variants. *BMC Genomics*, 14(Suppl 3):S2.

Capriotti, E., Arbiza, L., Casadio, R., Dopazo, J., Dopazo, H., and Marti-Renom, M. A. (2008). Use of estimated evolutionary strength at the codon level improves the prediction of disease-related protein mutations in humans. *Human Mutation*, 29(1):198–204.

Capriotti, E., Calabrese, R., and Casadio, R. (2006). Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics*, 22(22):2729–2734.

Capriotti, E. and Fariselli, P. (2017). PhD-SNPg: A webserver and lightweight tool for scoring single nucleotide variants. *Nucleic Acids Research*, 45(W1):W247–W252.

Capriotti, E., Montanucci, L., Profiti, G., Rossi, I., Giannuzzi, D., Aresu, L., and Fariselli, P. (2019). Fido-SNP: The first webserver for scoring the impact of single nucleotide variants in the dog genome. *Nucleic Acids Research*, 47(W1):W136–W141.

Carter, H., Chen, S., Isik, L., Tyekucheva, S., Velculescu, V. E., Kinzler, K. W., Vogelstein, B., and Karchin, R. (2009). Cancer-Specific High-Throughput Annotation of Somatic Mutations: Computational Prediction of Driver Missense Mutations. *Cancer Research*, 69(16):6660–6667.

Cavalli-Sforza, L. L. and Feldman, M. W. (2003). The application of molecular genetic approaches to the study of human evolution. *Nature Genetics*, 33(3s):266–275.

Chakravorty, S. and Hegde, M. (2017). Gene and Variant Annotation for Mendelian Disorders in the Era of Advanced Sequencing Technologies. *Annual Review of Genomics and Human Genetics*, 18(1):229–256.

Chalmers, A. F. (2013). *What Is This Thing Called Science?* Hackett Publishing. 04074.

Chang, F. and Li, M. M. (2013). Clinical Application of Amplicon-Based next-Generation Sequencing in Cancer. *Cancer Genetics*, 206(12):413–419. 00075.

Chasman, D. and Adams, R. M. (2001). Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: Structure-based assessment of amino acid variation11Edited by F. Cohen. *Journal of Molecular Biology*, 307(2):683–706.

Cheng, J., Nguyen, T. Y. D., Cygan, K. J., Çelik, M. H., Fairbrother, W. G., Avsec, ž., and Gagneur, J. (2019). MMSplice: Modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biology*, 20(1):48.

Cheng, R. R., Nordesjö, O., Hayes, R. L., Levine, H., Flores, S. C., Onuchic, J. N., and Morcos, F. (2016). Connecting the Sequence-Space of Bacterial Signaling Proteins to Phenotypes Using Coevolutionary Landscapes. *Molecular Biology and Evolution*, 33(12):3054–3064. 00009.

Choi, Y. and Chan, A. P. (2015). PROVEAN Web Server: A Tool to Predict the Functional Effect of Amino Acid Substitutions and Indels. *Bioinformatics*, 31(16):2745–2747.

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., and Chan, A. P. (2012). Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLOS ONE*, 7(10):e46688.

Chong, J. X., Buckingham, K. J., Jhangiani, S. N., Boehm, C., Sobreira, N., Smith, J. D., Harrell, T. M., et al. (2015). The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *The American Journal of Human Genetics*, 97(2):199–215.

Chothia, C. (1975). Structural Invariants in Protein Folding. *Nature*, 254(5498):304–308.

Chun, S. and Fay, J. C. (2009). Identification of Deleterious Mutations within Three Human Genomes. *Genome Research*, 19(9):1553–1561.

Clifford, R. J., Edmonson, M. N., Nguyen, C., and Buetow, K. H. (2004). Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. *Bioinformatics*, 20(7):1006–1014.

Cline, M. S. and Karchin, R. (2011). Using Bioinformatics to Predict the Functional Impact of SNVs. *Bioinformatics*, 27(4):441–448. 00065.

Codó, L., Bayarri, G., Cid-Fuentes, J. A., Conejero, J., Hospital, A., Royo, R., Repchevsky, D., et al. (2019). MuGVRE. A virtual research environment for 3D/4D genomics. *bioRxiv*, page 602474.

Collins, F. S. (1995). Positional cloning moves from perditional to traditional. *Nature Genetics*, 9(4):347.

Consortium, U. (2018). UniProt: A worldwide hub of protein knowledge. *Nucleic acids research*, 47(D1):D506–D515.

Cooper, D. N. and Youssoufian, H. (1988). The CpG dinucleotide and human genetic disease. *Human Genetics*, 78(2):151–155.

Coventry, A., Bull-Otterson, L. M., Liu, X., Clark, A. G., Maxwell, T. J., Crosby, J., Hixson, J. E., et al. (2010). Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nature Communications*, 1:131.

Crockett, D. K., Lyon, E., Williams, M. S., Narus, S. P., Facelli, J. C., and Mitchell, J. A. (2012). Utility of Gene-Specific Algorithms for Predicting Pathogenicity of Uncertain Gene Variants. *Journal of the American Medical Informatics Association*, 19(2):207–211.

Crosswell, L. C. and Thornton, J. M. (2012). ELIXIR: A distributed infrastructure for European biological data. *Trends Biotechnol*, 30(5):241–242.

Daudé, D., Topham, C. M., Remaud-Siméon, M., and André, I. (2013). Probing impact of active site residue mutations on stability and activity of Neisseria polysaccharea amylosucrase. *Protein Science : A Publication of the Protein Society*, 22(12):1754–1765.

Dayhoff, M., Schwartz, R., and Orcutt, B. (1978). 22 a Model of Evolutionary Change in Proteins. *Atlas of protein sequence and structure*, pages 345–352.

Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., and Liu, X. (2015a). Comparison and Integration of Deleteriousness Prediction Methods for Nonsynonymous SNVs in Whole Exome Sequencing Studies. *Human Molecular Genetics*, 24(8):2125–2137.

Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., and Liu, X. (2015b). Comparison and Integration of Deleteriousness Prediction Methods for Nonsynonymous SNVs in Whole Exome Sequencing Studies. *Human Molecular Genetics*, 24(8):2125–2137. 00211.

Drmanac, R., Sparks, A. B., Callow, M. J., Halpern, A. L., Burns, N. L., Kermani, B. G., Carnevali, P., et al. (2009). Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. *Science*. 01018.

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, 14(9):755–763.

Edgar, R. C. (2004). MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput. *Nucleic Acids Research*, 32(5):1792–1797.

Ellegren, H. and Galtier, N. (2016). Determinants of genetic diversity. *Nature Reviews Genetics*, 17(7):422–433.

Eraslan, G., Avsec, Ž., Gagneur, J., and Theis, F. J. (2019). Deep learning: New computational modelling techniques for genomics. *Nature Reviews Genetics*, page 1.

Eswarakumar, V. P., Horowitz, M. C., Locklin, R., Morriss-Kay, G. M., and Lonai, P. (2004). A gain-of-function mutation of Fgfr2c demonstrates the roles of this receptor variant in osteogenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 101(34):12555–12560.

Ezkurdia, I., Juan, D., Rodriguez, J. M., Frankish, A., Diekhans, M., Harrow, J., Vazquez, J., Valencia, A., and Tress, M. L. (2014). Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Human Molecular Genetics*, 23(22):5866–5878.

Fariselli, P., Martelli, P. L., Savojardo, C., and Casadio, R. (2015). INPS: Predicting the impact of non-synonymous variations on protein stability from sequence. *Bioinformatics*, 31(17):2816–2821.

Fauchère, J.-L., Charton, M., Kier, L. B., Verloop, A., and Pliska, V. (1988). Amino Acid Side Chain Parameters for Correlation Studies in Biology and Pharmacology. *International Journal of Peptide and Protein Research*, 32(4):269–278.

Fechter, K. and Porollo, A. (2014). MutaCYP: Classification of Missense Mutations in Human Cytochromes P450. *BMC Medical Genomics*, 7:47.

Feinauer, C. and Weigt, M. (2017). Context-Aware Prediction of Pathogenicity of Missense Mutations Involved in Human Disease. *arXiv:1701.07246 [q-bio]*.

Ferrer-Costa, C., Gelpí, J. L., Zamakola, L., Parraga, I., de la Cruz, X., and Orozco, M. (2005). PMUT: A web-based tool for the annotation of pathological mutations on proteins. *Bioinformatics*, 21(14):3176–3178.

Ferrer-Costa, C., Orozco, M., and de la Cruz, X. (2002). Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties11Edited by J. Thornton. *Journal of Molecular Biology*, 315(4):771–786.

Ferrer-Costa, C., Orozco, M., and de la Cruz, X. (2004). Sequence-based prediction of pathological mutations. *Proteins: Structure, Function, and Bioinformatics*, 57(4):811–819.

Fessenden, M. (2017). Protein Maps Chart the Causes of Disease. *Nature*, 549:293. 00001.

Figliuzzi, M., Barrat-Charlaix, P., and Weigt, M. (2018). How Pairwise Coevolutionary Models Capture the Collective Residue Variability in Proteins? *Molecular Biology and Evolution*, 35(4):1018–1027.

Figliuzzi, M., Jacquier, H., Schug, A., Tenaillon, O., and Weigt, M. (2016). Coevolutionary Landscape Inference and the Context-Dependence of Mutations in Beta-Lactamase TEM-1. *Molecular Biology and Evolution*, 33(1):268–280.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., et al. (2016). The Pfam Protein Families Database: Towards a More Sustainable Future. *Nucleic Acids Research*, 44(D1):D279–D285.

Firnberg, E., Labonte, J. W., Gray, J. J., and Ostermeier, M. (2014). A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape. *Molecular Biology and Evolution*, 31(6):1581–1592.

Flagel, L., Brandvain, Y., and Schrider, D. R. (2019). The Unreasonable Effectiveness of Convolutional Neural Networks in Population Genetic Inference. *Molecular Biology and Evolution*, 36(2):220–238.

Flexner, A. (2017). *The Usefulness of Useless Knowledge*. Princeton University Press. 00082.

Flynn, W. F., Haldane, A., Torbett, B. E., and Levy, R. M. (2017). Inference of Epistatic Effects Leading to Entrenchment and Drug Resistance in HIV-1 Protease. *Molecular Biology and Evolution*, 34(6):1291–1306. 00004.

Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., et al. (2015). COSMIC: Exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research*, 43(D1):D805–D811.

Fox, G., Sievers, F., and Higgins, D. G. (2015). Using de Novo Protein Structure Predictions to Measure the Quality of Very Large Multiple Sequence Alignments. *Bioinformatics*, page btv592.

Freund, Y. and Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1):119–139.

Fu, W., O'Connor, T. D., Jun, G., Kang, H. M., Abecasis, G., Leal, S. M., Gabriel, S., et al. (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, 493(7431):216–220.

Garzón, J. I., Deng, L., Murray, D., Shapira, S., Petrey, D., and Honig, B. (2016). A Computational Interactome and Functional Annotation for the Human Proteome. *Elife*, 5:e18715. 00016.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely Randomized Trees. *Machine Learning*, 63(1):3–42.

Gini, C. (1912). Variabilità e mutabilità. *Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi*.

Goldstein, D. B., Allen, A., Keebler, J., Margulies, E. H., Petrou, S., Petrovski, S., and Sunyaev, S. (2013). Sequencing studies in human genetics: Design and interpretation. *Nature Reviews Genetics*, 14(7):460–470.

González-Pérez, A. and López-Bigas, N. (2011). Improving the Assessment of the Outcome of Nonsynonymous SNVs with a Consensus Deleteriousness Score, Condel. *American Journal of Human Genetics*, 88(4):440–449.

Gotea, V., Gartner, J. J., Qutob, N., Elnitski, L., and Samuels, Y. (2015). The Functional Relevance of Somatic Synonymous Mutations in Melanoma and Other Cancers. *Pigment cell & melanoma research*, 28(6):673–684.

Goymer, P. (2007). Synonymous Mutations Break Their Silence. *Nature Reviews Genetics*, 8:92. 00053.

Grantham, R. (1974). Amino Acid Difference Formula to Help Explain Protein Evolution. *Science*, 185(4154):862–864.

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., Davies, H., et al. (2007). Patterns of Somatic Mutation in Human Cancer Genomes. *Nature*, 446(7132):153–158. 02595.

Gress, A., Ramensky, V., and Kalinina, O. V. (2017). Spatial Distribution of Disease-Associated Variants in Three-Dimensional Structures of Protein Complexes. *Oncogenesis*, 6(9):e380.

Hamasaki-Katagiri, N., Salari, R., Wu, A., Qi, Y., Schiller, T., Filiberto, A. C., Schisterman, E. F., Komar, A. A., Przytycka, T. M., and Kimchi-Sarfaty, C. (2013). A Gene-Specific Method for Predicting Hemophilia-Causing Point Mutations. *Journal of molecular biology*, 425(21):4023–4033.

Hawkins, D. M. (2004). The Problem of Overfitting. *Journal of Chemical Information and Computer Sciences*, 44(1):1–12.

He, H. and Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., and Scholkopf, B. (1998). Support Vector Machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28.

Heinzen, E. L., Swoboda, K. J., Hitomi, Y., Gurrieri, F., Nicole, S., de Vries, B., Tiziano, F. D., et al. (2012). *De Novo* mutations in *ATP1A3* cause alternating hemiplegia of childhood. *Nature Genetics*, 44(9):1030–1034.

Henikoff, S. and Henikoff, J. G. (1992). Amino Acid Substitution Matrices from Protein Blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919.

Hopf, T. A., Green, A. G., Schubert, B., Mersmann, S., Schärfe, C. P., Ingraham, J. B., Toth-Petroczy, A., Brock, K., Riesselman, A. J., Palmedo, P., et al. (2018). The EVcouplings Python framework for coevolutionary sequence analysis. *Bioinformatics*, 35(9):1582–1584.

Hopf, T. A., Ingraham, J. B., Poelwijk, F. J., Schärfe, C. P. I., Springer, M., Sander, C., and Marks, D. S. (2017). Mutation Effects Predicted from Sequence Co-Variation. *Nature Biotechnology*, 35(2):128–135.

Hoskins, R. A., Repo, S., Barsky, D., Andreoletti, G., Moult, J., and Brenner, S. E. (2017). Reports from CAGI: The Critical Assessment of Genome Interpretation. *Human Mutation*, 38(9):1039–1041.

Huang, K.-l., Mashl, R. J., Wu, Y., Ritter, D. I., Wang, J., Oh, C., Paczkowska, M., et al. (2018). Pathogenic Germline Variants in 10,389 Adult Cancers. *Cell*, 173(2):355–370.e14.

Hunt, R. C., Simhadri, V. L., Iandoli, M., Sauna, Z. E., and Kimchi-Sarfaty, C. (2014). Exposing Synonymous Mutations. *Trends in Genetics*, 30(7):308–321. 00114.

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3):90–95.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931.

Ison, J., Ienasescu, H., Chmura, P., Rydza, E., Ménager, H., Kalaš, M., Schwämmle, V., et al. (2019). The bio.tools registry of software tools and data resources for the life sciences. *Genome Biology*, 20(1):164.

Ison, J., Rapacki, K., Ménager, H., Kalaš, M., Rydza, E., Chmura, P., Anthon, C., Beard, N., Berka, K., Bolser, D., et al. (2015). Tools and data services registry: A community effort to document bioinformatics resources. *Nucleic acids research*, 44(D1):D38–D47.

Itan, Y., Shang, L., Boisson, B., Ciancanelli, M. J., Markle, J. G., Martinez-Barricarte, R., Scott, E., et al. (2016). The mutation significance cutoff: Gene-level thresholds for variant predictions. *Nature Methods*, 13(2):109–110.

Ittisoponpisan, S., Islam, S. A., Khanna, T., Alhuzimi, E., David, A., and Sternberg, M. J. E. (2019). Can Predicted Protein 3D Structures Provide Reliable Insights into whether Missense Variants Are Disease Associated? *Journal of Molecular Biology*, 431(11):2197–2212.

Izarzugaza, J. M., del Pozo, A., Vazquez, M., and Valencia, A. (2012). Prioritization of Pathogenic Mutations in the Protein Kinase Superfamily. *BMC Genomics*, 13(4):S3.

Jagadeesh, K. A., Wenger, A. M., Berger, M. J., Guturu, H., Stenson, P. D., Cooper, D. N., Bernstein, J. A., and Bejerano, G. (2016). M-CAP Eliminates a Majority of Variants of Uncertain Significance in Clinical Exomes at High Sensitivity. *Nature Genetics*, 48(12):1581–1586.

Janin, J., Henrick, K., Moult, J., Eyck, L. T., Sternberg, M. J. E., Vajda, S., Vakser, I., and Wodak, S. J. (2003). CAPRI: A Critical Assessment of PRedicted Interactions. *Proteins: Structure, Function, and Bioinformatics*, 52(1):2–9.

Jones, E., Oliphant, T., Peterson, P., et al. (2001). SciPy: Open Source Scientific Tools for Python. [Online; accessed <today>].

Jordan, D. M., Kiezun, A., Baxter, S. M., Agarwala, V., Green, R. C., Murray, M. F., Pugh, T., Lebo, M. S., Rehm, H. L., Funke, B. H., and Sunyaev, S. R. (2011). Development and Validation of a Computational Method for Assessment of Missense Variants in Hypertrophic Cardiomyopathy. *American Journal of Human Genetics*, 88(2):183–192.

Jubb, H. C., Pandurangan, A. P., Turner, M. A., Ochoa-Montaño, B., Blundell, T. L., and Ascher, D. B. (2017). Mutations at protein-protein interfaces: Small changes over big surfaces have large impacts on human health. *Progress in Biophysics and Molecular Biology*, 128:3–13.

Kaminker, J. S., Zhang, Y., Watanabe, C., and Zhang, Z. (2007a). CanPredict: A computational tool for predicting cancer-associated missense mutations. *Nucleic Acids Research*, 35(suppl_2):W595–W598.

Kaminker, J. S., Zhang, Y., Waugh, A., Haverty, P. M., Peters, B., Sebisanovic, D., Stinson, J., Forrest, W. F., Bazan, J. F., Seshagiri, S., and Zhang, Z. (2007b). Distinguishing Cancer-Associated Missense Mutations from Common Polymorphisms. *Cancer Research*, 67(2):465–473.

Karchin, R., Diekhans, M., Kelly, L., Thomas, D. J., Pieper, U., Eswar, N., Haussler, D., and Sali, A. (2005). LS-SNP: Large-scale annotation of coding non-synonymous SNPs based on multiple information sources. *Bioinformatics*, 21(12):2814–2820.

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., et al. (2019). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv*, page 531210.

Katoh, K. and Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, 30(4):772–780.

Katsanis, S. H. and Katsanis, N. (2013). Molecular Genetic Testing and the Future of Clinical Genomics. *Nature Reviews Genetics*, 14(6):415–426. 00212.

Kearns, M. and Ron, D. (1999). Algorithmic Stability and Sanity-Check Bounds for Leave-One-Out Cross-Validation. *Neural Computation*, 11(6):1427–1453.

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., and Sternberg, M. J. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nature protocols*, 10(6):845.

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., and Shendure, J. (2014). A General Framework for Estimating the Relative Pathogenicity of Human Genetic Variants. *Nature genetics*, 46(3):310–315.

Kitzman, J. O., Starita, L. M., Lo, R. S., Fields, S., and Shendure, J. (2015). Massively parallel single-amino-acid mutagenesis. *Nature Methods*, 12(3):203–206.

Kobayashi, Y., Yang, S., Nykamp, K., Garcia, J., Lincoln, S. E., and Topper, S. E. (2017). Pathogenic Variant Burden in the ExAC Database: An Empirical Approach to Evaluating Population Data for Clinical Variant Interpretation. *Genome Medicine*, 9.

Krebs, J. E., Goldstein, E. S., and Kilpatrick, S. T. (2017). *Lewin's Genes XII*. Jones & Bartlett Learning. 00132.

Kuhn, T. S. (2012). *The Structure of Scientific Revolutions*. University of Chicago press. 00778.

Kumar, P., Henikoff, S., and Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, 4(7):1073–1081.

Kunkel, T. A. (1985). Rapid and efficient site-specific mutagenesis without phenotypic selection. *Proceedings of the National Academy of Sciences*, 82(2):488–492.

Kyte, J. and Doolittle, R. F. (1982). A Simple Method for Displaying the Hydropathic Character of a Protein. *Journal of Molecular Biology*, 157(1):105–132.

Laddach, A., Gautel, M., and Fraternali, F. (2017). TITINdb—a Computational Tool to Assess Titin's Role as a Disease Gene. *Bioinformatics*, 33(21):3482–3485. 00003.

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., et al. (2016). ClinVar: Public Archive of Interpretations of Clinically Relevant Variants. *Nucleic Acids Research*, 44(D1):D862–D868.

Lassmann, T. and Sonnhammer, E. L. (2005). Kalign – an Accurate and Fast Multiple Sequence Alignment Algorithm. *BMC Bioinformatics*, 6:298.

Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K. W., Renfrew, P. D., Smith, C. A., Sheffler, W., et al. (2011). ROSETTA3: An object-oriented software suite for the simulation and design of macromolecules. In *Methods in Enzymology*, volume 487, pages 545–574. Elsevier.

Lehner, B. (2013). Genotype to phenotype: Lessons from model organisms for human genetics. *Nature Reviews Genetics*, 14(3):168–178.

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., et al. (2016). Analysis of Protein-Coding Genetic Variation in 60,706 Humans. *Nature*, 536(7616):285–291.

Leo, B., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). Classification and Regression Trees. *Wadsworth International Group*.

Leong, I. U., Stuckey, A., Lai, D., Skinner, J. R., and Love, D. R. (2015). Assessment of the Predictive Accuracy of Five in Silico Prediction Tools, Alone or in Combination, and Two Metaservers to Classify Long QT Syndrome Gene Mutations. *BMC Medical Genetics*, 16.

LeWinter, M. M. and Granzier, H. L. (2013). Titin Is a Major Human Disease Gene. *Circulation*, 127(8):938–944. 00058.

Li, B., Krishnan, V. G., Mort, M. E., Xin, F., Kamati, K. K., Cooper, D. N., Mooney, S. D., and Radivojac, P. (2009). Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics*, 25(21):2744–2750.

Li, Q., Liu, X., Gibbs, R. A., Boerwinkle, E., Polychronakos, C., and Qu, H.-Q. (2014). Gene-Specific Function Prediction for Non-Synonymous Mutations in Monogenic Diabetes Genes. *PLoS ONE*, 9(8).

Li, X., Kierczak, M., Shen, X., Ahsan, M., Carlborg, Ö., and Marklund, S. (2013). PASE: A novel method for functional prediction of amino acid substitutions based on physicochemical properties. *Frontiers in Genetics*, 4.

Liu, L., Okada, S., Kong, X.-F., Kreins, A. Y., Cypowyj, S., Abhyankar, A., Toubiana, J., et al. (2011a). Gain-of-function human STAT1 mutations impair IL-17 immunity and underlie chronic mucocutaneous candidiasis. *Journal of Experimental Medicine*, 208(8):1635–1648.

Liu, X., Jian, X., and Boerwinkle, E. (2011b). dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. *Human Mutation*, 32(8):894–899.

López-Ferrando, V., Gazzo, A., de la Cruz, X., Orozco, M., and Gelpí, J. L. (2017). PMut: A web-based tool for the annotation of pathological variants on proteins, 2017 update. *Nucleic Acids Research*, 45(W1):W222–W228.

Lott, M. T., Leipzig, J. N., Derbeneva, O., Xie, H. M., Chalkia, D., Sarmady, M., Procaccio, V., and Wallace, D. C. (2013). mtDNA Variation and Analysis Using Mitomap and Mitomaster. *Current Protocols in Bioinformatics*, 44(1):1.23.1–1.23.26.

Louie, R. H. Y., Kaczorowski, K. J., Barton, J. P., Chakraborty, A. K., and McKay, M. R. (2018). Fitness Landscape of the Human Immunodeficiency Virus Envelope Protein That Is Targeted by Antibodies. *Proceedings of the National Academy of Sciences*, page 201717765. 00003.

MacArthur, D. G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., et al. (2012). A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. *Science*, 335(6070):823–828.

MacArthur, D. G., Manolio, T. A., Dimmock, D. P., Rehm, H. L., Shendure, J., Abecasis, G. R., Adams, D. R., et al. (2014). Guidelines for investigating causality of sequence variants in human disease. *Nature*, 508(7497):469–476.

Makrythanasis, P. and Antonarakis, S. E. (2013). Pathogenic variants in non-protein-coding sequences. *Clinical Genetics*, 84(5):422–428.

Mardis, E. R. (2010). The 1,000 genome, the 100,000 analysis? *Genome medicine*, 2(11):84.

Marín Sala, Ò. (2017). *Caracterització bioinformàtica de la relació entre l'impacte molecular de les variants patogèniques i el fenotip clínic*. Ph.D. Thesis, Universitat Autònoma de Barcelona.

Marini, N. J., Thomas, P. D., and Rine, J. (2010). The Use of Orthologous Sequences to Predict the Impact of Amino Acid Substitutions on Protein Function. *PLOS Genetics*, 6(5):e1000968.

Masica, D. L., Sosnay, P. R., Cutting, G. R., and Karchin, R. (2012). Phenotype-Optimized Sequence Ensembles Substantially Improve Prediction of Disease-Causing Mutation in Cystic Fibrosis. *Human Mutation*, 33(8):1267–1274.

Matthews, B. W. (1975). Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451.

McCandlish, D. M., Shah, P., and Plotkin, J. B. (2016). Epistasis and the Dynamics of Reversion in Molecular Evolution. *Genetics*, 203(3):1335–1351.

McClellan, J. and King, M.-C. (2010). Genetic Heterogeneity in Human Disease. *Cell*, 141(2):210–217.

McGuffin, L. J., Bryson, K., and Jones, D. T. (2000). The PSIPRED Protein Structure Prediction Server. *Bioinformatics*, 16(4):404–405.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In van der Walt, S. and Millman, J., editors, *Proceedings of the 9th Python in Science Conference*, pages 51–56.

McVean, G., Spencer, C. C. A., and Chaix, R. (2005). Perspectives on Human Genetic Variation from the HapMap Project. *PLOS Genetics*, 1(4):e54.

Meyerson, M., Gabriel, S., and Getz, G. (2010). Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews Genetics*, 11(10):685–696.

Miyata, T., Miyazawa, S., and Yasunaga, T. (1979). Two Types of Amino Acid Substitutions in Protein Evolution. *Journal of Molecular Evolution*, 12(3):219–236.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., and Weigt, M. (2011). Direct-

Coupling Analysis of Residue Coevolution Captures Native Contacts across Many Protein Families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301. 00538.

Nachman, M. W. and Crowell, S. L. (2000). Estimate of the Mutation Rate per Nucleotide in Humans. *Genetics*, 156(1):297–304.

Need, A. C., Shashi, V., Hitomi, Y., Schoch, K., Shianna, K. V., McDonald, M. T., Meisler, M. H., and Goldstein, D. B. (2012). Clinical application of exome sequencing in undiagnosed genetic conditions. *Journal of Medical Genetics*, 49(6):353–361.

Nelson, D. L., Lehninger, A. L., and Cox, M. M. (2008). *Lehninger Principles of Biochemistry*. Macmillan. 14534.

Ng, P. C. and Henikoff, S. (2001). Predicting Deleterious Amino Acid Substitutions. *Genome Research*, 11(5):863–874.

Ng, P. C. and Henikoff, S. (2003). SIFT: Predicting Amino Acid Changes That Affect Protein Function. *Nucleic Acids Research*, 31(13):3812–3814.

Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., Huff, C. D., Shannon, P. T., Jabs, E. W., Nickerson, D. A., Shendure, J., and Bamshad, M. J. (2010). Exome Sequencing Identifies the Cause of a Mendelian Disorder. *Nature Genetics*, 42(1):30–35.

Nielsen, S. V., Stein, A., Dinitzen, A. B., Papaleo, E., Tatham, M. H., Poulsen, E. G., Kassem, M. M., Rasmussen, L. J., Lindorff-Larsen, K., and Hartmann-Petersen, R. (19-Apr-2017). Predicting the Impact of Lynch Syndrome-Causing Missense Mutations from Structural Calculations. *PLOS Genetics*, 13(4):e1006739. 00007.

Niroula, A., Urolagin, S., and Vihinen, M. (2015). PON-P2: Prediction Method for Fast and Reliable Identification of Harmful Variants. *PLOS ONE*, 10(2):e0117380.

Niroula, A. and Vihinen, M. (2015). Classification of Amino Acid Substitutions in Mismatch Repair Proteins Using PON-MMR2. *Human Mutation*, 36(12):1128–1134.

Niroula, A. and Vihinen, M. (2017). Predicting Severity of Disease-Causing Variants. *Human Mutation*, 38(4):357–364.

Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence alignment11Edited by J. Thornton. *Journal of Molecular Biology*, 302(1):205–217.

Ohta, T. and Kimura, M. (1973). A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population*. *Genetics Research*, 22(2):201–204.

Oliver, G. R., Hart, S. N., and Klee, E. W. (2014). Bioinformatics for Clinical Next Generation Sequencing. *Clinical Chemistry*, page clinchem.2014.224360.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., et al. (2011). Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pérez, F. and Granger, B. E. (2007). IPython: A System for Interactive Scientific Computing. *Computing in Science & Engineering*, 9(3):21–29.

Pestre, D. (2003). *Science, Argent et Politique: Un Essai d'interprétation*. Editions Quae. 00263.

Plotkin, J. B. and Kudla, G. (2011). Synonymous but not the same: The causes and consequences of codon bias. *Nature Reviews Genetics*, 12(1):32–42.

Ponzoni, L. and Bahar, I. (2018). Structural Dynamics Is a Determinant of the Functional Significance of Missense Variants. *Proceedings of the National Academy of Sciences*, page 201715896. 00000.

Popper, K. (2005). *The Logic of Scientific Discovery*. Routledge. 28953.

Porta-Pardo, E., Garcia-Alonso, L., Hrabe, T., Dopazo, J., and Godzik, A. (2015). A Pan-Cancer Catalogue of Cancer Driver Protein Interaction Interfaces. *PLoS computational biology*, 11(10):e1004518. 00036.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106.

Quintana-Murci, L. and Clark, A. G. (2013). Population genetic tools for dissecting innate immunity in humans. *Nature Reviews Immunology*, 13(4):280–293.

Raimondi, D., Gazzo, A. M., Rooman, M., Lenaerts, T., and Vranken, W. F. (2016). Multilevel biological characterization of exomic variants at the protein level significantly improves the identification of their deleterious effects. *Bioinformatics*, 32(12):1797–1804.

Raimondi, D., Tanyalcin, I., Ferté, J., Gazzo, A., Orlando, G., Lenaerts, T., Rooman, M., and Vranken, W. (2017). DEOGEN2: Prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. *Nucleic Acids Research*, 45(W1):W201–W206.

Ramensky, V., Bork, P., and Sunyaev, S. (2002). Human non-synonymous SNPs: Server and survey. *Nucleic Acids Research*, 30(17):3894–3900.

Reva, B., Antipin, Y., and Sander, C. (2011). Predicting the Functional Impact of Protein Mutations: Application to Cancer Genomics. *Nucleic Acids Research*, 39(17):e118.

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., and Rehm, H. L. (2015). Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*, 17(5):405–423.

Riera, C., Lois, S., Domínguez, C., Fernandez-Cadenas, I., Montaner, J., Rodríguez-Sureda, V., and de la Cruz, X. (2015). Molecular Damage in Fabry Disease: Characterization and Prediction of Alpha-Galactosidase A Pathological Mutations. *Proteins: Structure, Function, and Bioinformatics*, 83(1):91–104.

Riera, C., Padilla, N., and de la Cruz, X. (2016). The Complementarity Between Protein-Specific and General Pathogenicity Predictors for Amino Acid Substitutions. *Human Mutation*, 37(10):1013–1024.

Riera Ribas, C. (2016). *Novel Approaches in the Identification of Pathogenic Variants in the Clinical Diagnosis*. Ph.D. Thesis, Universitat Autònoma de Barcelona.

Ritchie, G. R. S., Dunham, I., Zeggini, E., and Flicek, P. (2014). Functional annotation of noncoding sequence variants. *Nature Methods*, 11(3):294–296.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

Rodriguez, L. L., Brooks, L. D., Greenberg, J. H., and Green, E. D. (2013). The Complexities of Genomic Identifiability. *Science*, 339(6117):275–276. 00095.

Sauna, Z. E. and Kimchi-Sarfaty, C. (2011). Understanding the Contribution of Synonymous Mutations to Human Disease. *Nature Reviews Genetics*, 12:683. 00493.

Savojardo, C., Fariselli, P., Martelli, P. L., and Casadio, R. (2016). INPS-MD: A Web Server to Predict Stability of Protein Variants from Sequence and Structure. *Bioinformatics*, 32(16):2542–2544.

Sayılgan, J. F., Haliloğlu, T., and Gönen, M. (2019). Protein dynamics analysis reveals that missense mutations in cancer-related genes appear frequently on hinge-neighboring residues. *Proteins*, 87(6):512–519.

Schaafsma, G. C. P. and Vihinen, M. (2017). Large Differences in Proportions of Harmful and Benign Amino Acid Substitutions between Proteins and Diseases. *Human Mutation*, 38(7):839–848.

Schadt, E. E., Turner, S., and Kasarskis, A. (2010). A window into third-generation sequencing. *Human Molecular Genetics*, 19(R2):R227–R240.

Schrijver, I., Aziz, N., Farkas, D. H., Furtado, M., Gonzalez, A. F., Greiner, T. C., Grody, W. W., et al. (2012). Opportunities and Challenges Associated with Clinical Diagnostic Genome Sequencing: A Report of the Association for Molecular Pathology. *The Journal of Molecular Diagnostics*, 14(6):525–540.

Schwarz, J. M., Rödelsperger, C., Schuelke, M., and Seelow, D. (2010). MutationTaster Evaluates Disease-Causing Potential of Sequence Alterations. *Nature Methods*, 7(8):575–576.

Schwarze, K., Buchanan, J., Taylor, J. C., and Wordsworth, S. (2018). Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *Genetics in Medicine*, 20(10):1122.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge university press. 00687.

Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001). dbSNP: The NCBI Database of Genetic Variation. *Nucleic Acids Research*, 29(1):308–311.

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L. A., Edwards, K. J., Day, I. N. M., and Gaunt, T. R. (2012). Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions Using Hidden Markov Models. *Human Mutation*, 34(1):57–65.

Shihab, H. A., Rogers, M. F., Gough, J., Mort, M., Cooper, D. N., Day, I. N. M., Gaunt, T. R., and Campbell, C. (2015). An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics*, 31(10):1536–1543.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. (2017). Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*.

Sim, N.-L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., and Ng, P. C. (2012). SIFT web server: Predicting effects of amino acid substitutions on proteins. *Nucleic Acids Research*, 40(W1):W452–W457.

Sobreira, N. L. M., Cirulli, E. T., Avramopoulos, D., Wohler, E., Oswald, G. L., Stevens, E. L., Ge, D., et al. (2010). Whole-Genome Sequencing of a Single Proband Together with Linkage Analysis Identifies a Mendelian Disease Gene. *PLOS Genetics*, 6(6):e1000991.

Sonnhammer, E. L. L., Gabaldón, T., da Silva, S., W, A., Martin, M., Robinson-Rechavi, M., Boeckmann, B., Thomas, P. D., and Dessimoz, C. (2014). Big Data and Other Challenges in the Quest for Orthologs. *Bioinformatics*, 30(21):2993–2998. 00067.

Stankiewicz, P. and Lupski, J. R. (2010). Structural Variation in the Human Genome and its Role in Disease. *Annual Review of Medicine*, 61(1):437–455.

Starita, L. M., Ahituv, N., Dunham, M. J., Kitzman, J. O., Roth, F. P., Seelig, G., Shendure, J., and Fowler, D. M. (2017). Variant Interpretation: Functional

Assays to the Rescue. *The American Journal of Human Genetics*, 101(3):315–325.

Stead, L. F., Wood, I. C., and Westhead, D. R. (2011). KvSNP: Accurately Predicting the Effect of Genetic Variants in Voltage-Gated Potassium Channels. *Bioinformatics*, 27(16):2181–2186.

Stenson, P. D., Mort, M., Ball, E. V., Evans, K., Hayden, M., Heywood, S., Hussain, M., Phillips, A. D., and Cooper, D. N. (2017). The Human Gene Mutation Database: Towards a Comprehensive Repository of Inherited Mutation Data for Medical Research, Genetic Diagnosis and next-Generation Sequencing Studies. *Human Genetics*, 136(6):665–677. 00109.

Stone, E. A. and Sidow, A. (2005). Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Research*, 15(7):978–986.

Sun, H. and Yu, G. (2019). New insights into the pathogenicity of non-synonymous variants through multi-level analysis. *Scientific Reports*, 9(1):1667.

Sunyaev, S., Ramensky, V., Koch, I., Lathe III, W., Kondrashov, A. S., and Bork, P. (2001). Prediction of deleterious human alleles. *Human Molecular Genetics*, 10(6):591–597.

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., and UniProt Consortium (2015). UniRef Clusters: A Comprehensive and Scalable Alternative for Improving Sequence Similarity Searches. *Bioinformatics (Oxford, England)*, 31(6):926–932.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826.

Szurmant, H. and Weigt, M. (2018). Inter-residue, inter-protein and inter-family coevolution: Bridging the scales. *Current Opinion in Structural Biology*, 50:26–32.

Tang, H. and Thomas, P. D. (2016). PANTHER-PSEP: Predicting disease-causing genetic variants using position-specific evolutionary preservation. *Bioinformatics*, 32(14):2230–2232.

Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., McGee, S., et al. (2012). Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. *Science*, 337(6090):64–69.

The 1000 Genomes Project (2015). A Global Reference for Human Genetic Variation. *Nature*, 526(7571):68–74.

the 1000 Genomes Project, Conrad, D. F., Keebler, J. E. M., DePristo, M. A., Lindsay, S. J., Zhang, Y., Casals, F., et al. (2011). Variation in genome-wide mutation rates within and between human families. *Nature Genetics*, 43(7):712–714.

The Cancer Genome Atlas Research Network (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068.

The GTEx Consortium (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660.

The International Cancer Genome Consortium (2010). International network of cancer genome projects. *Nature*, 464(7291):993–998.

The International HapMap Project (2003). The International HapMap Project. *Nature*, 426(6968):789.

The Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678.

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., and Narechania, A. (2003). PANTHER: A Library of Protein Families and Subfamilies Indexed by Function. *Genome Research*, 13(9):2129–2141.

Torkamani, A. and Schork, N. J. (2007). Accurate Prediction of Deleterious Protein Kinase Polymorphisms. *Bioinformatics*, 23(21):2918–2925.

Valastyan, J. S. and Lindquist, S. (2014). Mechanisms of protein-folding diseases at a glance. *Disease Models & Mechanisms*, 7(1):9–14.

van der Walt, S., Colbert, S. C., and Varoquaux, G. (2011). The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science and Engg.*, 13(2):22–30.

Vazquez, M., Pons, T., Brunak, S., Valencia, A., and Izarzugaza, J. M. G. (2015). wKinMut-2: Identification and Interpretation of Pathogenic Variants in Human Protein Kinases. *Human Mutation*, 37(1):36–42.

Veltman, J. A. and Brunner, H. G. (2012). *De Novo* mutations in human genetic disease. *Nature Reviews Genetics*, 13(8):565–575.

Vicoso, B. and Charlesworth, B. (2006). Evolution on the X chromosome: Unusual patterns and processes. *Nature Reviews Genetics*, 7(8):645.

Vihinen, M. (2017). How to Define Pathogenicity, Health, and Disease? *Human Mutation*, 38(2):129–136.

Wainreb, G., Ashkenazy, H., Bromberg, Y., Starovolsky-Shitrit, A., Haliloglu, T., Ruppin, E., Avraham, K. B., Rost, B., and Ben-Tal, N. (2010). MuD: An interactive web server for the prediction of non-neutral substitutions using protein structural data. *Nucleic Acids Research*, 38(suppl_2):W523–W528.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: Functional Annotation of Genetic Variants from High-Throughput Sequencing Data. *Nucleic Acids Research*, 38(16):e164–e164.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63.

Weigt, M., White, R. A., Szurmant, H., Hoch, J. A., and Hwa, T. (2009). Identification of Direct Residue Contacts in Protein–Protein Interaction by Message Passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72. 00558.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018.

Wilson, G., Aruliah, D. A., Brown, C. T., Hong, N. P. C., Davis, M., Guy, R. T., Haddock, S. H. D., et al. (2014). Best Practices for Scientific Computing. *PLOS Biology*, 12(1):e1001745.

Wimley, W. C. and White, S. H. (1996). Experimentally Determined Hydrophobicity Scale for Proteins at Membrane Interfaces. *Nature Structural & Molecular Biology*, 3(10):842–848.

Xue, Y., Ankala, A., Wilcox, W. R., and Hegde, M. R. (2015). Solving the Molecular Diagnostic Testing Conundrum for Mendelian Disorders in the Era of Next-Generation Sequencing: Single-Gene, Gene Panel, or Exome/Genome Sequencing. *Genetics in Medicine*, 17(6):444–451.

Yip, Y. L., Famiglietti, M., Gos, A., Duek, P. D., David, F. P. A., Gateau, A., and Bairoch, A. (2008). Annotating Single Amino Acid Polymorphisms in the UniProt/Swiss-Prot Knowledgebase. *Human Mutation*, 29(3):361–366.

Yue, P., Melamud, E., and Moult, J. (2006). SNPs3D: Candidate gene and SNP selection for association studies. *BMC Bioinformatics*, 7(1):166.

Zhou, J., Theesfeld, C. L., Yao, K., Chen, K. M., Wong, A. K., and Troyanskaya, O. G. (2018). Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature Genetics*, 50(8):1171.