Tools and pipelines for interpreting the impacts of genetic variants

Niroula, Abhishek

2016

Tools and pipelines for interpreting the impacts of genetic variants

Abhishek Niroula

Logo

DOCTORAL DISSERTATION

by due permission of the Faculty of Medicine, Lund University, Sweden.

To be defended in GK Salen, BMC, Lund on 24th November, 2016 at 13:00.

Faculty opponent
Professor Bengt Persson

Organization LUND UNIVERSITY	Document name DOCTORAL DISSERTATION
	Date of issue 2016-11-24
Author(s) Abhishek Niroula	Sponsoring organization

Title and subtitle

Tools and pipelines for interpreting the impacts of genetic variants

Abstract

Next generation sequencing (NGS) methods have been widely used for diagnosis. The genome and exome sequencing projects produce huge amounts of variation data but clinical relevance of a large proportion of them are not known. Among various types of genetic variations, the single nucleotide variations (SNVs) that lead to amino acid substitutions are the most difficult to interpret. Since experimental methods are expensive and time consuming, these are not feasible for all identified variants. Computational tools can be used for scoring and ranking the variants and prioritizing them for experiments. Guidelines for interpreting clinical relevance of variants have recommended use of computational tools as one of several lines of evidence.

In this study, we developed four computational tools for interpreting the impacts of genetic variations. PON-P2, PON-MMR2, and PON-mt-tRNA predict pathogenicity of protein and RNA variations. PON-PS predicts the phenotypic severity due to genetic variations. All the tools use machine learning algorithms and have been tested extensively using independent datasets. All tools showed better performance when compared with state-of-the-art tools. These tools are freely accessible from our website.

We used the developed tools for analysing the impacts of variations in mismatch repair proteins, mitochondrial tRNAs, and amino acid substitutions in cancer. All possible amino acid substitutions in mismatch repair proteins and all possible SNVs in mitochondrial transfer RNAs were analysed by using PON-MMR2 and PON-mt-tRNA, respectively. We also analysed 5 million somatic variations from 7,042 genomes or exomes grouped into 30 types of cancer. The harmful somatic variations were identified using PON-P2. Several pathways previously associated with cancer and new pathways were identified in most of the cancer types.

The tools developed in this study are useful for early and reliable identification of harmful variations and can easily be integrated to high throughput data analysis pipelines. The findings from the analysis of genetic variations enable prioritization of experimental studies in various cancers as well as for interpreting the impacts of RNA and protein variations.

Key words

Genetic variation, variation interpretation, variation impact, mutation impact, disease severity, cancer, computational tool, machine learning, feature selection

Classification system and/or index terms (if any)

Supplementary bibliographical information		Language English
ISSN and key title 1652-8220		ISBN 978-91-7619-360-0
Recipient's notes	Number of pages	Price
	Security classification	

I, the undersigned, being the copyright owner of the abstract of the above-mentioned dissertation, hereby grant to all reference sourcespermission to publish and disseminate the abstract of the above-mentioned dissertation.

Signature _____________________________ Date 2016-11-24

Tools and pipelines for interpreting the impacts of genetic variants

Abhishek Niroula

Logo

Doctoral Thesis
2016

Protein Structure and Bioinformatics
Department of Experimental Medical Sciences
Faculty of Medicine
Lund University
Sweden

Faculty of Medicine, Lund University

ISBN 978-91-7619-360-0
ISSN 1652-8220

Printed in Sweden by Media-Tryck, Lund University
Lund 2016

To my family

Contents

Papers included in this thesis

Paper I

PON-P2: prediction method for fast and reliable identification of harmful variants
Abhishek Niroula, Siddhaling Urolagin, and Mauno Vihinen
PLoS ONE (2015), 10:e0117380

Paper II

Classification of amino acid substitutions in mismatch repair proteins using PON-MMR2
Abhishek Niroula and Mauno Vihinen
Human Mutation (2015), 36 (12):1128-1134

Paper III

PON-mt-tRNA: a multifactorial probability-based method for classification of mitochondrial tRNA variations
Abhishek Niroula and Mauno Vihinen
Nucleic Acids Research (2016), 44 (5):2020-2027

Paper IV

Predicting severity of disease-causing variants
Abhishek Niroula and Mauno Vihinen
(Submitted manuscript)

Paper V

Harmful somatic amino acid substitutions affect key pathways in cancers
Abhishek Niroula and Mauno Vihinen
BMC Medical Genomics (2015), 8:53

Abstract

Next generation sequencing (NGS) methods have been widely used for diagnosis. As time and cost of sequencing has reduced sharply during the last decade, genome and exome-wide sequencing have increasingly been used. The genome and exome projects produce large amounts of variation data and the clinical relevance of large proportions of them are not known. Among various types of genetic variations, the single nucleotide variations (SNVs) that lead to amino acid substitutions (AASs) are the most challenging to interpret. The best way to characterize the impacts of variations is by experimental studies. Since these experiments are expensive and time consuming, they cannot be performed for all identified variants. Computational tools can be used for scoring and ranking the variants and prioritizing them for experimental studies. Reliable and fast tools are necessary for accurate variation interpretation and to cope with the amounts of generated data. Several tools are available for predicting impacts of genetic variations. These tools use various types of information and have different performances. Various performance assessment studies have shown that most of the widely used tools have inconsistent and sub-optimal performance.

In this study, we implemented a systematic approach to develop four computational tools for interpreting the impacts of genetic variations. The tools are based on machine learning algorithm. Benchmark variation datasets were obtained from various sources for training and testing the tools. A systematic feature selection technique was employed to identify relevant and non-redundant features for predicting variation impact. The benchmark datasets and the features were used for training the tools. Finally, the tools were tested by using independent datasets to estimate their performance for unseen data. The tools PON-P2, PON-MMR2, and PON-PS predict impacts of AASs in human proteins and the PON-mt-tRNA tool predicts the impacts of SNVs in human mitochondrial transfer RNAs (mt-tRNAs). All the tools showed better performance when compared with state-of-the-art tools. These tools have consistently shown the best performance in our studies as well as in independent studies.

The tools developed in this study are useful for ranking variations and prioritizing the likely harmful ones for further evaluation. These tools were developed for different purposes. Three of the tools (PON-P2, PON-MMR2, and PON-mt-tRNA) predict pathogenicity of variations. While PON-P2 is a generic tool for predicting pathogenicity of AASs in all human proteins, PON-MMR2 and PON-mt-tRNA are specific tools for predicting pathogenicity of variations in mismatch repair proteins and mt-tRNA genes, respectively. PON-PS is the first tool for predicting disease severity due to AASs. Pathogenicity of variations indicate the relevance of variation to a disease but cannot predict severity of phenotype. Early identification of disease severity promotes personalized medicine by facilitating early interventions, such as preventive measures, clinical monitoring, and molecular tests, for patients and their family members.

The developed computational tools were used for analysing the impacts of variations in DNA mismatch repair proteins, mt-tRNA genes, and somatic variations in cancer. The impacts of all possible AASs in four mismatch repair proteins (MLH1, MSH2, MSH6, and PMS2) were predicted using PON-MMR2 and the impacts of all possible SNVs in 22 human mt-tRNAs were predicted using PON-mt-tRNA. We also studied the distribution of predicted pathogenic and benign variations in the protein domains and 3-dimensional structures of proteins and mt-tRNAs.

PON-P2 was used to identify harmful somatic AASs from among 5 million somatic variations from 7,042 genomes or exomes grouped into 30 types of cancer. Only a small fraction of the somatic variations were identified to be harmful. Although known cancer genes contained higher numbers of harmful variations, the proportion of harmful variations was only 40%. We prioritized the proteins that were implicated (containing harmful AASs) in the largest number of samples in each cancer type and studied the networks and pathways affected by them. In the functional interaction network, the prioritized proteins were centrally located. The significantly enriched pathways included several new pathways and previously known pathways implicated in cancer. Our findings facilitates prioritization of experimental studies in various cancer types as well as interpretation of variation impacts in mismatch repair proteins and mt-tRNA genes.

Abbreviations

3-D	3-Dimensional
AAS	Amino Acid Substitution
ACMG	American College of Medical Genetics and Genomics
API	Application Programming Interface
AUC	Area Under the Curve
AUROC	Area Under the Receiver Operating Characteristic curve
BACC	Balanced ACCuracy
BLAST	Basic Local Alignment Search Tool
CAGI	Critical Assessment of Genome Interpretation
cDNA	coding DNA
CFTR	Cystic Fibrosis Transmembrane conductance Regulator
CGC	Cancer Gene Census
CGP	Cancer Genome Project
COSMIC	Catalogue Of Somatic Mutations In Cancer
DoCM	Database of Curated Mutations
ESHG	European Society of Human Genetics
EVS	Exome Variant Server
ExAC	Exome Aggregation Consortium
FN	False Negative
FP	False Positive
GO	Gene Ontology
HGMD	Human Gene Mutation Database
HGP	Human Genome Project
HNC	Head and Neck Cancer
ICGC	International Cancer Genome Consortium
IDbase	Immunodeficiency Database
InSiGHT	International Society for Gastrointestinal Hereditary Tumors
LOVD	Leiden Open Variation Database
LR	Likelihood Ratio
LSDB	Locus Specific Database
MCC	Matthews Correlation Coefficient
ML	Machine Learning
MMR	Mismatch Repair
mRNA	messenger RNA
MSA	Multiple Sequence Alignment
mtDB	Human Mitochondrial Genome Database
mtDNA	mitochondrial DNA
mtSNP	Human Mitochondrial Genome Polymorphism Database
mt-tRNA	mitochondrial transfer RNA
MT2	MutationTaster2
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
NHLBI-ESP	National Heart, Lung, and Blood Institute Exome Sequencing Project

NPV	Negative Predictive Value
nsSNV	non-synonymous Single Nucleotide Variation
OMIM	Online Mendelian Inheritance in Man
OOB	Out Of Bag
OPM	Overall Performance Measure
PDB	Protein Data Bank
PPV	Positive Predictive Value
PP2	PolyPhen-2
PSSM	Position Specific Scoring Matrix
rCRS	revised Cambridge Reference Sequence
RefSeq	NCBI Reference Sequences
RF	Random Forests
ROC	Receiver Operating Characteristic
SNV	Single Nucleotide Variation
SVM	Support Vector Machines
TCGA	The Cancer Genome Atlas
TN	True Negative
TP	True Positive
tRNA	transfer RNA
UCSC	University of California, Santa Cruz
UMD	Universal Mutation Database
UniProtKB	UniProt Knowledgebase
VCF	Variant Call Format
VEP	Variant Effect Predictor
VIC	Variation Interpretation Committee

1. Background

1.1 Genetic variations

The Human Genome Project (HGP) (Lander, et al., 2001) sequenced a reference human genome along with key model organisms such as bacteria, yeast, worms, flies, and mice. Successful completion of the HGP in 2003 marked the beginning of the genomic era in biomedical research (Collins, et al., 2003; Hood and Rowen, 2013). Since the completion of the HGP, the capabilities of sequencing methods have increased by many fold (van Dijk, et al., 2014; Goodwin, et al., 2016). There has also been a significant reduction in the cost of sequencing a genome. It is now possible to sequence a whole genome using Next Generation Sequencing (NGS) technology at a cost of around $1,000. The progresses in the sequencing methods have made routine use of NGS methods possible. Various genome and exome sequencing projects have been initiated and some of them have already been completed. The 1000 Genomes Project (Abecasis, et al., 2010; The 1000 Genomes Project Consortium, 2012; The 1000 Genomes Project Consortium, 2015), the Singapore Genome Variation Project (Teo, et al., 2009), the Genome of the Netherlands (Genome of the Netherlands Consortium, 2014), the UK10K project (Walter, et al., 2015), the National Heart, Lung, and Blood Institute Exome Sequencing Project (NHLBI-ESP) (Fu, et al., 2013), The Cancer Genome Atlas (TCGA) (http://cancergenome.nih.gov/), and the International Cancer Genome Consortium (ICGC) (Hudson, et al., 2010) are some of the sequencing projects.

All human genomes are 99.9% identical. Variations in the remaining 0.1% of the genome make each of them unique. The diversity of genetic variations is wide: from small single nucleotide variations (SNVs) to large chromosomal duplications or deletions. Single nucleotide substitutions are the most common genetic variations. The 1000 Genomes Project estimated that every human genome contains about 3 million SNVs in comparison to a reference genome (Abecasis, et al., 2010). The frequencies of insertions and deletions and larger structural variations were much smaller compared to that of SNVs. The frequency of variations decreased with an increasing size of the variations (Abecasis, et al., 2010).

Variations can have different consequences at DNA, RNA, and protein levels. Variations in the non-coding regions do not directly alter the protein sequences. But variations in the coding regions can have various consequences. Due to the degeneracy of the genetic code, a single amino acid can be coded by more than one codon and SNVs may or may not alter a protein sequence. The SNVs that do not alter the protein sequences are called synonymous variations and those that alter protein sequences by amino acid substitutions (AASs) are called non-synonymous SNVs (nsSNVs). SNVs that terminate the protein sequences prematurely by substitution of an amino acid by a stop codon are called protein truncating variations. The SNVs at or near splicing sites can alter splicing and produce alternative messenger RNA (mRNA) transcripts. According to the 1000 Genomes Project, each genome codes for about 11,000 AASs and approximately 12,500 synonymous substitutions (Abecasis, et al., 2010). Insertions and deletions can change the translation frame and thus the protein sequence after the variant site. Such variations are caused by insertion or deletion of one or more nucleotides (length not divisible by 3) and are called amphigoric amino acid insertion and deletion. Insertion or deletion of nucleotides of length 3-mer (any number divisible by 3) inserts or deletes amino acid(s) at the variation site. Large variations can lead to multiple copies of genes due to duplication or absence of a gene due to deletion.

1.2 Variation interpretation

Large numbers of genetic variations are being detected from patients and healthy population in various sequencing projects. Many of the variations are novel, or without proper annotation, and their disease relevance is missing. Whole-genome or exome sequencing provides valuable information about an individual if the data can be interpreted in a reliable and meaningful way. Easy and fast access to the genetic data was expected to revolutionize medical care and enable personalized medicine. Personalized medicine refers to the individualized medical care based on personal data, both genetic and non-genetic. Reliable interpretation of the genomic data is one of the major challenges for personalized medicine. Improvements in the sequencing technologies have exposed the major deficits in our understanding of the clinical relevance of the variations. Data analysis and variation interpretation are the most time consuming steps in sequencing projects. The bottleneck of personalized medicine has shifted from obtaining the genome sequences to interpreting them.

The impacts of certain types of variants are often straightforward to explain. The variants that alter protein sequences by truncation, amphigoric amino acid insertions and deletions, and other types of variations (such as substitution of initiation codon and large insertions and/or deletions) are often deleterious. The synonymous variants that do not alter splicing are often benign. The most difficult variants to interpret are the SNVs leading to AASs. Experimental studies are the best ways to interpret the effects of variations and their relevance to disease. However, such methods are often expensive and time consuming and it is impractical to characterize experimentally all the variants identified by NGS methods. The European Society of Human Genetics (ESHG) and the American College of Medical Genetics and Genomics (ACMG) have developed guidelines for application of NGS to clinical practice and for interpretation of genetic variations (Matthijs, et al., 2015; Richards, et al., 2015). These guidelines are intended for inherited genetic variants in relation to monogenic diseases. The guidelines recommend the use of variation databases, computational predictions, and experimental and clinical data for interpreting the impacts of variants.

1.2.1 Variation databases

The collection and sharing of variation data can facilitate fast and improved variation interpretation. Various databases collect and share variations and corresponding annotations. These databases differ in their contents and structures. Population databases contain frequencies of variants in populations but often lack information about the disease relevance of variants. A variant is likely not harmful if the frequency of the variant is high among healthy individuals in the population. The population databases may contain pathogenic variants and variation data from non-healthy individuals (Richards, et al., 2015). Some of the population databases include the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015), the NHLBI-ESP Exome Variant Server (EVS) (Fu, et al., 2013), and the Exome Aggregation Consortium (ExAC) (Lek, et al., 2016). The 1000 Genomes Project contains variation data from 2,504 individuals from 26 populations. The EVS contains exome data from 200,000 individuals with specific traits related to blood, heart diseases, and lung diseases as well as controls from African-American and European-American populations (Fu, et al., 2013). The ExAC database contains data from 60,706 unrelated individuals from several projects including the 1000 Genomes Project and the NHLBI-ESP. The dbSNP is a database of short genetic variations and contains variations from several large sequencing projects regardless of their functional or clinical relevance (Sherry, et al., 2001).

Disease databases contain variants from patients and their relevance to disease. The disease databases can be generic or specific. Generic disease databases include variations from many genes, proteins, and diseases. The specific databases contain variants associated with specific diseases or variants at specific genomic regions, such as genes, proteins, or protein domains. Generic disease databases include the Online Mendelian Inheritance in Man (OMIM) (Hamosh, et al., 2005), the ClinVar database (Landrum, et al., 2014), the UniProt Knowledgebase (UniProtKB) (The UniProt Consortium, 2015), and the Human Gene Mutation Database (HGMD) (Stenson, et al., 2014). Locus-specific variation databases (LSDBs) contain variants in specific genes and are usually manually curated. The LSDBs can also contain other information including detailed clinical characteristics of the patients. The Leiden Open Variation Database (LOVD) system hosts LSDBs for all human genes (Fokkema, et al., 2011). Other large collections of LSDBs include immunodeficiency databases (IDbases) (Piirilä, et al., 2006) and those maintained at the Universal Mutation Databases (UMD) platform (Béroud, et al., 2000).

Some databases are dedicated to specific types of variations or to an effect or mechanism. For example, the ProTherm database contains variants affecting protein stability (Kumar, et al., 2006). Some tools are useful for searching various types of resources including genetic variants and their annotations from several sources. The University of California, Santa Cruz (UCSC) Genome Browser (Kent, et al., 2002), the National Center for Biotechnology Information (NCBI) Map Viewer (Wheeler, et al., 2003), the Ensembl Genome Browser (Stalker, et al., 2004), and others are useful for finding information about genes, their products, and sequence variants.

1.2.2 Variation impact prediction

Pathogenicity prediction

Variation databases are useful resources for filtering disease-causing and benign variations. However, numerous variants in the variation databases lack information about their clinical relevance. In addition, large numbers of novel variants are detected by genome and exome sequencing. Computational tools are useful for predicting the impacts of variants and for ranking and prioritizing them for experimental studies (Thusberg and Vihinen, 2009; Zhang, et al., 2012; Kucukkal, et al., 2014; Niroula and Vihinen, 2016). As experimental methods are impractical for characterizing large number of variations, computational tools are required for interpreting their impacts. Several computational tools have been developed for variation interpretation. They vary widely depending on the principle, implementation, and application (Karchin, 2009; Thusberg and Vihinen, 2009; Capriotti, et al., 2012; Niroula and Vihinen, 2016; Tang and Thomas, 2016). A large majority of these tools are for predicting the pathogenicity of AASs.

The prediction tools utilize various types of information. Evolutionary conservation is widely used in combination with other information for predicting the impact of variations. Disease-causing variants appear frequently at conserved positions and are underrepresented at positions that are variable during evolution (Miller and Kumar, 2001). The conserved positions are usually important for protein structure or function (Miller and Kumar, 2001; Vitkup, et al., 2003; Shen and Vihinen, 2004). Evolutionary conservation is estimated based on multiple sequence alignment (MSA) of related sequences, called homologous sequences. Homologous sequences have a shared ancestry and they are separated during evolution by speciation (orthologs) or by duplication (paralogs).

Homologous sequences often have a high similarity compared to unrelated sequences. The Basic Local Alignment Search Tool (BLAST) (Camacho, et al., 2009) is often used to find similar sequences but the identified sequences may not necessarily be homologous. Various approaches have been used to generate MSAs and several measures have been derived from them (Niroula and Vihinen, 2016). Some tools are completely based on evolutionary conservation scores (Ng and Henikoff, 2001; Choi, et al., 2012), while many others use conservation scores along with other information such as properties of amino acids, protein structure, sequence environment, etc. Some of the tools using diverse information include CADD (Kircher, et al., 2014), MutationTaster2 (Schwarz, et al., 2014), MutPred (Li, et al., 2009), nsSNPAnalyzer (Bao, et al., 2005), PolyPhen-2 (Adzhubei, et al., 2010), SNPs&GO (Calabrese, et al., 2009), and VEST (Carter, et al., 2013).

Disease-causing AASs often have more drastic changes in their physicochemical properties than benign AASs (Steward, et al., 2003; de Beer, et al., 2013). Physical and biochemical properties of amino acids (e.g. hydropathy, charge, size, secondary structure propensities) are often used for predicting the impacts of variations. The amino acids present in the surrounding of a variant site in the protein sequence are also used as features by some tools (Calabrese, et al., 2009). Other features used for variation impact prediction are the annotations at the variant site in the sequence and structure databases (Carter, et al., 2013; Yates, et al., 2014). Variants occurring at functionally or structurally important sites can have deleterious effects.

Some tools use features derived from 3-dimensional (3-D) protein structures (Adzhubei, et al., 2010; Capriotti and Altman, 2011a; Yates, et al., 2014). The structural features can improve the prediction performance when used together with sequence features (Capriotti and Altman, 2011a). Solvent accessibility of amino acid residues and distribution of amino acids in the periphery of a variant site in the 3-D protein structures have been used as features for predicting pathogenicity of variants. However, these features cannot be used for variations in all proteins since 3-D structures are not available for all of them. One way of computing the structural features for all proteins is to predict the structures or the features (Yates, et al., 2014).

Prediction tools have also used features specific for genes or proteins such as the feature derived from Gene Ontology (GO) annotations. Although such features are specific for proteins and have the same values for all variations in a protein, GO-based feature improved classification of deleterious and benign variations (Calabrese, et al., 2009).

As the tools differ in feature composition and implementation, their predictions for the same variation can be different. Although the overall performances of the tools are similar, the predictions disagree for numerous variations. To utilize the benefits of various tools, meta-predictors have been developed. They utilize the predictions of various independent tools. Some of the meta-predictors include Condel (Gonzalez-Perez and Lopez-Bigas, 2011), PON-P (Olatubosun, et al., 2012), Meta-SNP (Capriotti, et al., 2013a), and PredictSNP (Bendl, et al., 2014). The variants used for training the constituent tools cannot be used for training and testing the meta-predictors. The predictions of the constituent tools are biased for their training data and lead to overfitting of the meta-predictor (Niroula and Vihinen, 2016).

Although nsSNVs are the most common variants associated with disease, other types of variations including synonymous variations and insertions and/or deletions are associated with several diseases (Piirilä, et al., 2006; Krawczak, et al., 2007; Sauna and Kimchi-Sarfaty, 2011; Hunt, et al., 2014). Tools have been developed to predict the pathogenicity of synonymous variations (Buske, et al., 2013) and of insertions and/or deletions (Zia and Moses, 2011; Hu and Ng, 2012;

Hu and Ng, 2013; Zhao, et al., 2013; Bermejo-Das-Neves, et al., 2014; Liu, et al., 2014; Douville, et al., 2016). Synonymous variations at splice sites as well as in exonic splicing regulatory regions may lead to splicing defects. Tools are available for predicting the impacts of both intronic and exonic variations on splicing (Nalla and Rogan, 2005; Desmet, et al., 2009; Woolfe, et al., 2010; Mort, et al., 2014). Some tools can predict impacts of more than one type of variations (Choi, et al., 2012; Carter, et al., 2013; Kircher, et al., 2014; Schwarz, et al., 2014; Douville, et al., 2016). Although most tools are for variations in the protein coding regions, some tools have been developed for predicting the impact of variations in non-coding regions (Macintyre, et al., 2010; Manke, et al., 2010; Ritchie, et al., 2014; Lee, et al., 2015; Zhou and Troyanskaya, 2015).

Various tools and services collect and disseminate variation impact predictions from multiple predictors. The dbNSFP database contains predictions of 14 tools for 83 million nsSNVs (Liu, et al., 2016). Variation annotation tools such as ANNOVAR (Yang and Wang, 2015), AVIA (Vuong, et al., 2015), SnpEff (Cingolani, et al., 2012), Variant Effect Predictor (VEP) (McLaren, et al., 2010), etc. can provide predictions of several tools.

Specific pathogenicity predictors

Some genes, protein domains or regions have been widely studied in association with diseases. Databases, services, and tools for specific genes, protein domains, or regions have been developed. Some of the examples include the resources for primary immunodeficiency-causing genes (Piirilä, et al., 2006; Samarghitean, et al., 2007; Ortutay and Vihinen, 2009), DNA mismatch repair (MMR) genes (Thompson, et al., 2014), and protein kinase domain (Stenberg, et al., 2000; Ortutay, et al., 2005; Vazquez, et al., 2016). The resources provide useful information for interpretation of variants at specific locations or in relation to specific diseases. As the amounts of resources in specific areas are growing, it is possible to develop novel tools specific for many of them. Tools have been developed for predicting the impacts of variations in MMR genes (Chao, et al., 2008; Ali, et al., 2012; Thompson, et al., 2013b; Thompson, et al., 2014), cystic fibrosis transmembrane conductance regulator (CFTR) protein (Masica, et al., 2012), cytochrome P450 enzymes (Fechter and Porollo, 2014), hypertrophic cardiomyopathy related proteins (Jordan, et al., 2011), protein kinase domains (Torkamani and Schork, 2007; Väliaho, et al., 2015; Vazquez, et al., 2016), phosphorylation sites (Wagih, et al., 2015), signal peptides (Hon, et al., 2009), and many others.

Mechanism-specific prediction

Genetic variations can have various effects, consequences and mechanisms (Vihinen, 2015). The pathogenicity predictors do not provide information about the mechanism of variation impact. To understand the mechanism of pathogenicity, mechanism-specific tools are required. The pathogenic AASs often affect the stability of the protein (Wang and Moult, 2001; Ferrer-Costa, et al., 2002; Stefl, et al., 2013; Peng and Alexov, 2016). Various tools have been developed to predict the impact of variation on protein stability (Guerois, et al., 2002; Parthiban, et al., 2006; Capriotti, et al., 2008; Dehouck, et al., 2009; Masso and Vaisman, 2010; Yang, et al., 2013; Pires, et al., 2014; Fariselli, et al., 2015; Laimer, et al., 2015), protein localization (Laurila and Vihinen, 2011), protein disorder (Ali, et al., 2014), protein aggregation (Fernandez-Escamilla, et al., 2004; Conchillo-Sole, et al., 2007; Walsh, et al., 2014; Zambrano, et al., 2015), protein solubility (Tian, et al., 2010; Sormanni, et al., 2015; Yang, et al., 2016), and many others (Thusberg and Vihinen, 2009).

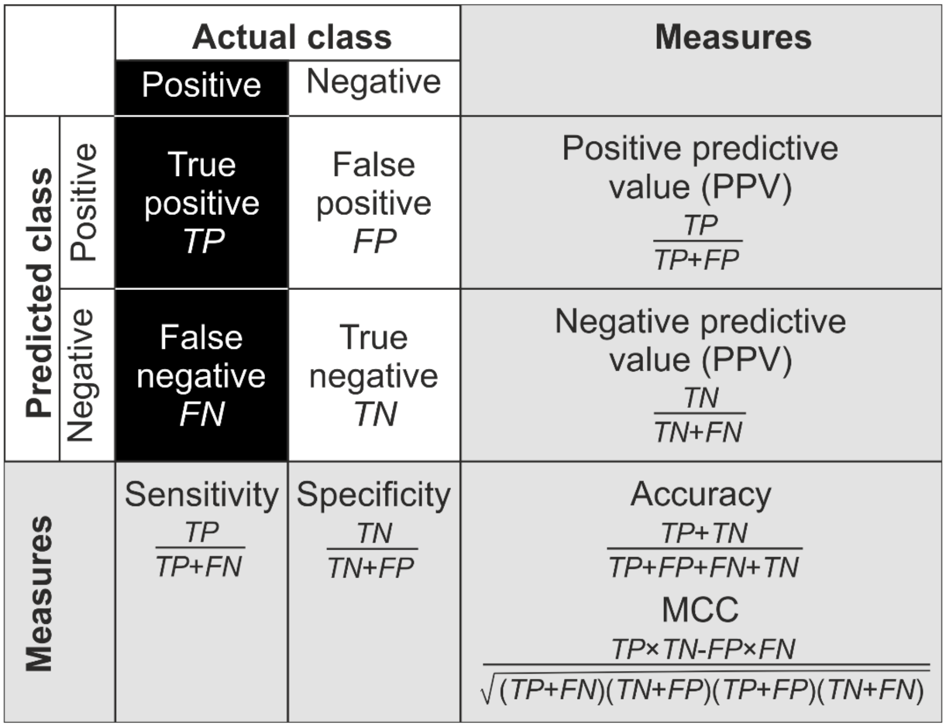
1.2.3 Performance assessment of prediction tools

Computational prediction tools are based on mathematical functions and statistics. The tools are optimized or trained using a training dataset consisting of samples with known outcomes. The performances of the tools depend on how well they can optimize or generalize from the training data. To estimate their reliability, they should be assessed using independent datasets. The performance of prediction methods can be assessed in three ways (Niroula and Vihinen, 2016). Variation interpretation challenges enable testing the capabilities for interpreting variants using available knowledge and tools. Critical Assessment of Genome Interpretation (CAGI, http://genomeinterpretation.org) organizes community-wide challenges to assess methods for interpreting the phenotypic impacts of genomic variations. CAGI provides unpublished experimentally characterized variation data and the participants are required to predict their impacts. The submissions from the participants are compared to the experimental findings and the performances of the methods applied by them are estimated. Although, such challenges enable assessment of methods for specific tasks, they do not provide systematic performance assessments due to small size of the test data.

The second way of assessing prediction tools is the performance assessment examined by the tool developers. The reliability of such a performance assessment depends on the quality of the test dataset and the method used for assessment. As developers tend to use the biggest possible data for training a new tool, test datasets are usually small in size. With the increasing size and quality of test datasets, such an assessment approach tends to be as reliable as a systematic performance assessment which is the third way of assessing prediction tools.

Systematic performance assessment is the most reliable way to estimate the overall performance of computational tools (Vihinen, 2012; Vihinen, 2013; Niroula and Vihinen, 2016). Benchmark datasets are required for systematic assessment. Databases of benchmark variation datasets have been established to provide gold standard datasets for development and assessment of prediction tools. VariBench (Nair and Vihinen, 2013) and VariSNP (Schaafsma and Vihinen, 2015) collect benchmark variation datasets from various sources and distribute them. Another requirement for the datasets used for systematic assessment is that they should be free from circularity which means that there should be no overlap between the tools' training and the test datasets. The performance of tools are overestimated in the presence of data circularity (Grimm, et al., 2015). Circularity may arise at various levels depending on the implementation of the tools. Overlapping variants in the training and test datasets is referred to as 'Type 1 circularity'. Circularity may occur even when variants are non-overlapping but the proteins are overlapping in the training and test datasets (Grimm, et al., 2015). For example, if a tool utilizes a protein-specific feature, the feature value will be same for all variants in a protein. In such a case, the presence of variants in the same protein in both training and test datasets leads to data circularity.

Additional requirements for a systematic performance assessment include assessing various tools together with state-of-the-art tools and reporting a wide-range of performance measures. Several performance measures are used to estimate the performance of prediction tools based on their implementations. Most variation impact predictors categorize variants into binary classes, while some tools predict continuous values. Results of binary classifiers can be presented in a contingency table or matrix which consists of four measures- true positive, false positive, false negative, and true negative. Based on these four measures, various performance measures can be computed (Fig. 1.1). Positive predictive value (PPV), negative predictive value (NPV), sensitivity, specificity, accuracy, and Matthews correlation coefficient (MCC) are the standard performance measures computed from the contingency table. Receiver operating characteristic (ROC) curves and area under the ROC curves (AUROC or AUC) are often used to assess the reliability of prediction methods. For continuous prediction and multi-class classification (classification with more than two classes) problems, different performance measures can be used to assess the performance of the tools (Pires, et al., 2014; Yang, et al., 2016). A single performance measure cannot reliably present the performance of prediction tools; therefore, various measures should be evaluated in the systematic performance assessments (Vihinen, 2012; Lever, et al., 2016).



		Actual class		Measures
		Positive	Negative	
Predicted class	Positive	True positive *TP*	False positive *FP*	Positive predictive value (PPV) $\frac{TP}{TP+FP}$
	Negative	False negative *FN*	True negative *TN*	Negative predictive value (PPV) $\frac{TN}{TN+FN}$
Measures		Sensitivity $\frac{TP}{TP+FN}$	Specificity $\frac{TN}{TN+FP}$	Accuracy $\frac{TP+TN}{TP+FP+FN+TN}$ MCC $\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)(TN+FP)(TP+FP)(TN+FN)}}$

Figure 1.1: Contingency matrix and six standard performance measures. The matrix shows the true and false predictions for data with known labels. Various performance measures can be computed based on the matrix. Sensitivity, specificity, PPV and NPV use two of the four cells in the matrix. Accuracy and MCC use all four cells in the matrix.

Performances of several prediction tools have been evaluated in various independent studies. Such assessments have been performed for tools predicting the impact of variations on pathogenicity (Thusberg, et al., 2011; Bendl, et al., 2014; Grimm, et al., 2015; Miosge, et al., 2015), protein stability (Potapov, et al., 2009; Khan and Vihinen, 2010), and splicing (Desmet, et al., 2010; Houdayer, et al., 2012; Jian, et al., 2014). These studies include different tools for assessment and their performances vary. The tools have inconsistent performances in different studies or datasets. Even a slight difference in the performance can lead to differences in the interpretation of large number of variants when applied to genome or exome-wide datasets.

1.3 Cancer

Cancer is characterized by uncontrolled cell growth which can invade surrounding tissues and spread to distant organs (Hanahan and Weinberg, 2011). Cancer cells have an increased mutation rate and large numbers of accumulated variations. While inherited variations in many genes increase cancer susceptibility, somatic variations are mainly involved in cancer development (Hindorff, et al., 2011; Garraway and Lander, 2013). The number of variations vary greatly depending on the type of cancer. Pediatric and hematologic cancers have a low variation frequency, while cancers prevalent at adulthood have higher frequencies of variations (Lawrence, et al., 2013; Watson, et al., 2013). In addition, mutagenic exposures increase the variation frequency in certain cancers, for example ultraviolet radiation in melanoma and smoking in lung cancer (Govindan, et al., 2012; Hodis, et al., 2012). Defects in the MMR genes is another reason for a high mutation rate leading to accumulation of large number of variations (Gryfe and Gallinger, 2001).

Among large number of somatic variations, some are drivers but the majority of them are passengers (Haber and Settleman, 2007). Driver variations develop a growth advantage and are responsible for the initiation, development, progression, and/or maintenance of tumors. Passenger variations are incidental and are carried along with the drivers. The number of driver variations can vary between cancers and each of them have small growth advantages (Stratton, et al., 2009; Bozic, et al., 2010). Genes containing driver variations are often known as driver genes and they are grouped as oncogenes and tumor suppressor genes. The driver variations in oncogenes are activating while those in tumor suppressor genes are inactivating. The oncogenes contain recurrent variations at the same position while the tumor suppressor genes contain variations throughout the protein sequences (Vogelstein, et al., 2013; Pon and Marra, 2015).

Large amounts of cancer genomic data are available from genomic projects such as the Cancer Genome Project (CGP, https://www.sanger.ac.uk/research/projects/cancergenome/), TCGA (http://cancergenome.nih.gov/), and the ICGC (Hudson, et al., 2010). These projects collect and provide various types of genetic data for large numbers of cancer samples. The Catalogue Of Somatic Mutations In Cancer (COSMIC) (Forbes, et al., 2011) stores cancer variations collected from the literature. These massive datasets provide unprecedented possibilities for data analysis. Various approaches have been used to study mechanisms of tumorigenesis and several genes and variations associated with cancers have been revealed. The Cancer Gene Census (CGC) lists genes causally implicated in cancer (Futreal, et al., 2004). Some databases collect cancer variants from various sources. The Database of Curated Mutations (DoCM) contains a curated list of harmful somatic variations (Ainscough, et al., 2016). The TP53 mutation database contains somatic variations in the *TP53* gene and their effect on the activity of tumor protein p53 encoded by the gene (Edlund, et al., 2012). Kin-Driver is a manually-curated database of validated driver variations (Simonetti, et al., 2014).

Several approaches have been employed to search for driver variations, genes, networks, and pathways (Gonzalez-Perez, et al., 2013; Ding, et al., 2014; Raphael, et al., 2014; Chen, et al., 2015; Tian, et al., 2015). Tolerance prediction tools are often applied for analysis of somatic variations in cancer genomes or for development of cancer-specific prediction tools. Several tools have been developed to identify driver variations and genes. These include CHASM (Wong, et al., 2011), transFIC (Gonzalez-Perez, et al., 2012), CanPredict (Kaminker, et al., 2007a), SPF-Cancer (Capriotti and Altman, 2011b), cancer-specific FATHMM (Shihab, et al., 2013), and CanDrA (Mao, et al., 2013). Most of these tools are trained using frequent somatic variations in the

COSMIC database, other cancer-related variations, and putative neutral variations from diverse sources.

Various other approaches have been used to identify driver genes, networks, and pathways in cancer. These methods are based on mutation rates (Dees, et al., 2012; Hodis, et al., 2012; Hua, et al., 2013; Lawrence, et al., 2013), functional impacts or patterns of variations (Gonzalez-Perez and Lopez-Bigas, 2012; Tamborero, et al., 2013; Vogelstein, et al., 2013; Korthauer and Kendziorski, 2015), or networks and pathways (The Cancer Genome Atlas Research Network, 2008; Cerami, et al., 2010; Vandin, et al., 2012; Ciriello, et al., 2013a; Wu, et al., 2015). As the large cancer genomic projects have collected heterogeneous data from large number of samples, it is possible to integrate different types of data from various sources. Some methods integrate different types of data e.g. genome, transcriptome, proteome, and epigenome (Bashashati, et al., 2012; Hou and Ma, 2014; Bertrand, et al., 2015; Verbeke, et al., 2015).

1.4 DNA mismatch repair (MMR)

The MMR system recognizes base pair mismatches and small insertions and deletions during DNA replication and repairs them (Jiricny, 2006). Besides repairing errors in the DNA, the MMR system also plays roles in cell cycle arrest and apoptosis (Li, 2008). Defects in the MMR mechanism leads to the spontaneous increase in the mutation rate and accumulation of variations in microsatellite repeats, a phenomenon known as microsatellite instability. Variations in the MMR genes are associated with Lynch syndrome (LS) and increase the risk of colorectal and various other cancers (Sijmons and Hofstra, 2016). LS is one of the most common hereditary cancer syndromes (Lynch, et al., 2015; Heinen, 2016).

Large numbers of variations have been identified in the MMR genes. Until recently, the MMR gene variations were stored in several databases. Conflicting interpretations of the disease relevance of some variants were reported in different studies (Ali, et al., 2012). By the efforts of the International Society for Gastrointestinal Hereditary Tumors (InSiGHT), several databases were merged to create a single LSDB for each MMR gene (Thompson, et al., 2014) and a Variant Interpretation Committee (VIC) was established to classify MMR gene variants. The VIC developed a multifactorial method and applied it to classify over 2,300 variations into five classes. The pathogenicity for approximately one-third of the variants could not be interpreted due to lack of evidence and thus were grouped as unclassified. The majority of the variants in the unclassified group were AASs.

Specific tools have been developed for classification of MMR variants. The MAPP-MMR tool is an optimized version of the MAPP tool for classifying variants in the MLH1 and MSH2 proteins (Chao, et al., 2008). PON-MMR is a meta-predictor for classification of MMR variants (Ali, et al., 2012). The tool utilizes the prediction of several generic variation impact predictors. Thompson et al. tested the combination of six different generic variation impact predictors for classification of MMR variants (Thompson, et al., 2013b). They found that the combination of MAPP and PolyPhen-2.1 performed the best. The method was integrated with additional evidence to predict a multifactorial posterior probability (Thompson, et al., 2013a) which was later used by the InSiGHT VIC for classification of MMR variants (Thompson, et al., 2014).

1.5 Transfer RNAs (tRNAs)

The genetic information is transferred from DNA to mRNA and is translated into proteins. During translation, tRNAs deliver amino acid residues to the ribosome for elongation of the polypeptide chain. Out of the 64 codons, 61 code for 20 amino acids and the remaining three are nonsense (stop) codons. Due to wobble base pairing (base pairing that does not follow Watson-Crick base pairing), a tRNA anti-codon can pair with multiple codons that code for the same amino acid. The numbers of tRNA genes vary between organisms. The human genome contains 597 nuclear-encoded tRNA genes and 22 mitochondrial tRNA (mt-tRNA) genes (Chan and Lowe, 2009). The human mitochondrial genome consists of a circular DNA which encodes for 13 protein-coding, 2 ribosomal RNA, and 22 tRNA genes. Unlike the nuclear DNA, the major portion (~93%) of the mitochondrial DNA (mtDNA) codes for genes. The mutation rate for mtDNA is several times (10-17x) higher than for the nuclear genome due to various reasons, including an inefficient MMR mechanism and the lack of histones (Khrapko, et al., 1997; Tuppen, et al., 2010). The nuclear tRNAs and the mt-tRNAs also differ in structure. While most tRNAs have a highly conserved cloverleaf structure, the human mt-tRNAs have one of the three non-canonical structures (Suzuki, et al., 2011).

Several copies of mtDNA co-exist in a cell since there are numerous mitochondria per cell. All copies of mtDNA in a cell may be identical, a condition known as homoplasmy, or there may be multiple variants of mtDNA, known as heteroplasmy. Heteroplasmy plays an important role in pathogenicity and disease severity of mitochondrial variations (Yarham, et al., 2010; Suzuki, et al., 2011; Abbott, et al., 2014). The mtDNA variants are tolerated unless a minimum proportion of variant copies are present in the cell (DiMauro and Schon, 2001; Yarham, et al., 2010). Numerous variations have been identified in tRNAs and several of them are associated with diseases. Thus far, all disease-associated tRNA variations have been found in the mt-tRNAs (Abbott, et al., 2014; Kirchner and Ignatova, 2015). The disease-causing and benign mtDNA variations are stored in various databases including the MITOMAP database (Lott, et al., 2013), the Human Mitochondrial Genome Database (mtDB) (Ingman and Gyllensten, 2006), Human Mitochondrial Genome Polymorphism Database (mtSNP) (Tanaka, et al., 2004), and Mammit-tRNA database (Putz, et al., 2007).

To classify the pathogenicity of mtDNA variants, four canonical criteria were derived (DiMauro and Schon, 2001). Using the canonical criteria and some additional criteria, a new scoring system for mt-tRNA variants was established (McFarland, et al., 2004). Evidence from functional studies such as biochemical, histochemical, single-fiber and *trans*-mitochondrial cybrid studies was added to the scoring system. Using the variants classified based on the scoring, a new method was developed to classify the pathogenicity of mt-tRNA variants (Kondrashov, 2005). The method used the evolutionary conservation and Watson-Crick base pairing in the stems of mt-tRNAs to predict the pathogenicity of all possible SNVs in the human mt-tRNAs. In 2011, the evidence-based scoring criteria was re-evaluated and the classification threshold was adjusted (Yarham, et al., 2011). The scoring system assigns certain scores when the results of functional studies are positive; negative results have a score of zero. Although the scoring system does not use the negative results of *trans*-mitochondrial cybrid studies, the system is widely used to classify the pathogenicity of mt-tRNA variants. To adjust the negative results of *trans*-mitochondrial cybrid studies, a modification to the scoring system has been suggested (González-Vioque, et al., 2014).

The evidence-based scoring system requires results from expensive and time-consuming experimental studies to score and classify the variants.

1.6 Disease severity due to genetic variations

The disease relevance of a large number of variations has been verified. The variants are associated with a wide range of diseases and clinical phenotypes. In most cases, protein truncating and amphigoric variations alter the protein sequences after the variation site and often cause severe phenotypes (Feucht, et al., 2008). The AASs only change the amino acid at the variation site and are associated with a wide range of disease severity, from benign to severe. Since monogenic Mendelian diseases are caused by variations in a single gene, it is possible to study variations associated with different disease severity in them. In many diseases, variations have been associated with similar phenotypes but with different severity (Guldberg, et al., 1998; Caldovic, et al., 2015). In some other cases, variations in the same protein can lead to different phenotypes (Massaad, et al., 2013; Demurger, et al., 2015).

The ability to correlate phenotype to genotype makes predictive medicine possible by improving prognosis and facilitating early clinical interventions (Dipple and McCabe, 2000). Genotype-phenotype correlation has been studied for variations in many proteins and diseases (Fu and Jinnah, 2012; Mannini, et al., 2013; Vincent, et al., 2013; Demurger, et al., 2015). However, the correlation between genotype and phenotype is inconsistent. Severity of variants in many proteins have been classified based on clinical and molecular data (Weinreb, et al., 2010; McCormick, et al., 2013). While many variants have been classified to have mild, moderate, or severe phenotypes, some variations are associated with phenotypic heterogeneity. These variants can have different phenotypes in different individuals. Genetic and non-genetic factors can influence the phenotypes of various monogenic disorders (Scriver and Waters, 1999; Cutting, 2010). Five different threshold models were proposed to explain the relationship between variations and disease severity (Dipple and McCabe, 2000). These thresholds distinguish the different groups of variants: severe, mild, and indeterminate. The relation between genotype and phenotype has also been studied in relation to protein sequence and structure, and endophenotypes (Robins, et al., 2006; Masica, et al., 2015; Reblova, et al., 2015; Sengupta, et al., 2015). Endophenotypes are the quantitative traits or risk factors associated with phenotypes through shared genetic influence (Masica and Karchin, 2016).

1.7 Machine learning

Machine learning (ML) is a form of artificial intelligence in which computer algorithms learn from given data and gain capability of predicting for new data. ML tasks are mainly categorized into two groups, i.e. supervised and unsupervised learning. Supervised learning requires a training dataset containing labels (true outcomes) for each data point. The task is to learn from the training dataset to predict the labels. The labels are categorical for classification and numerical for regression. Some of the widely used algorithms for supervised ML include random forests (RF), neural networks, support vector machines (SVM), Bayes classifier, logistic and linear regressions (https://www.kaggle.com/wiki/Algorithms) (Kotsiantis, et al., 2006). Unsupervised learning does not require any labels for the data. It is generally applied for exploring structure and patterns in the data. Clustering is the most common example of unsupervised learning.

Besides supervised and unsupervised learnings, there are other types of learning such as semi-supervised learning and reinforcement learning. In semi-supervised learning, the dataset consists of both types of data: with and without labels. The semi-supervised approach is typically used in areas where labeled data are scarce but large amounts of unlabeled data are available. The addition of unlabeled data may or may not improve the performance of a predictive model (Singh, et al., 2008). In reinforcement learning, the algorithm can interact with the environment and optimize its behavior based on the consequences of previous actions.

ML has been widely used in various research areas and applications including various bioinformatics applications (Larranaga, et al., 2006; Inza, et al., 2010; Libbrecht and Noble, 2015; Konig, et al., 2016). ML methods have been used for developing classification and regression models as well as for studying the data structure and finding patterns in the data by unsupervised learning. Supervised learning is widely used to develop predictive classification and regression models. Krishnan and Westhead introduced ML for predicting the impact of nsSNVs (Krishnan and Westhead, 2003). They used SVM and decision trees to predict the impact of variations. After their work, various ML algorithms including Bayesian framework, neural networks, SVM, and RF were used for predicting variation impact (Cai, et al., 2004; Ferrer-Costa, et al., 2004; Bao and Cui, 2005; Karchin, et al., 2005). In addition to predicting the disease association of variations, ML was applied for predicting effect of variations on mechanisms such as the stability of protein (Capriotti, et al., 2004). Most variation impact predictors developed in the last decade use ML algorithms. The performance of a supervised ML-model is highly dependent on the quality of the training data, the optimization of the algorithm parameters, and the features used to describe the data (Kotsiantis, et al., 2006; Vihinen, 2012).

1.7.1 Data preparation

ML algorithms are used to explore data and recognize patterns from the data. For a supervised ML method, the quality of training data is critical. Benchmark datasets are required for systematic training and testing ML models. The qualities of a benchmark dataset for ML are relevance to the problem, representativeness, reliable labeling, non-redundancy, scalability, and reusability (Nair and Vihinen, 2013).

'Missing data' is a common problem in almost all real world data. Missing values can have a significant impact on the conclusions. Various approaches are used to address the issue of missing data. Excluding the cases or variables with missing values is the simplest way to get rid of the missing data. However, such an approach can significantly reduce the size of training and test data. In addition, data exclusion can miss data structure or important features if the missing values have non-random distribution. Other approaches for handling missing data include mean or mode substitution, maximum likelihood, regression imputation, multiple imputation, and the special value method (Kotsiantis, et al., 2006; Graham, 2009; Kang, 2013).

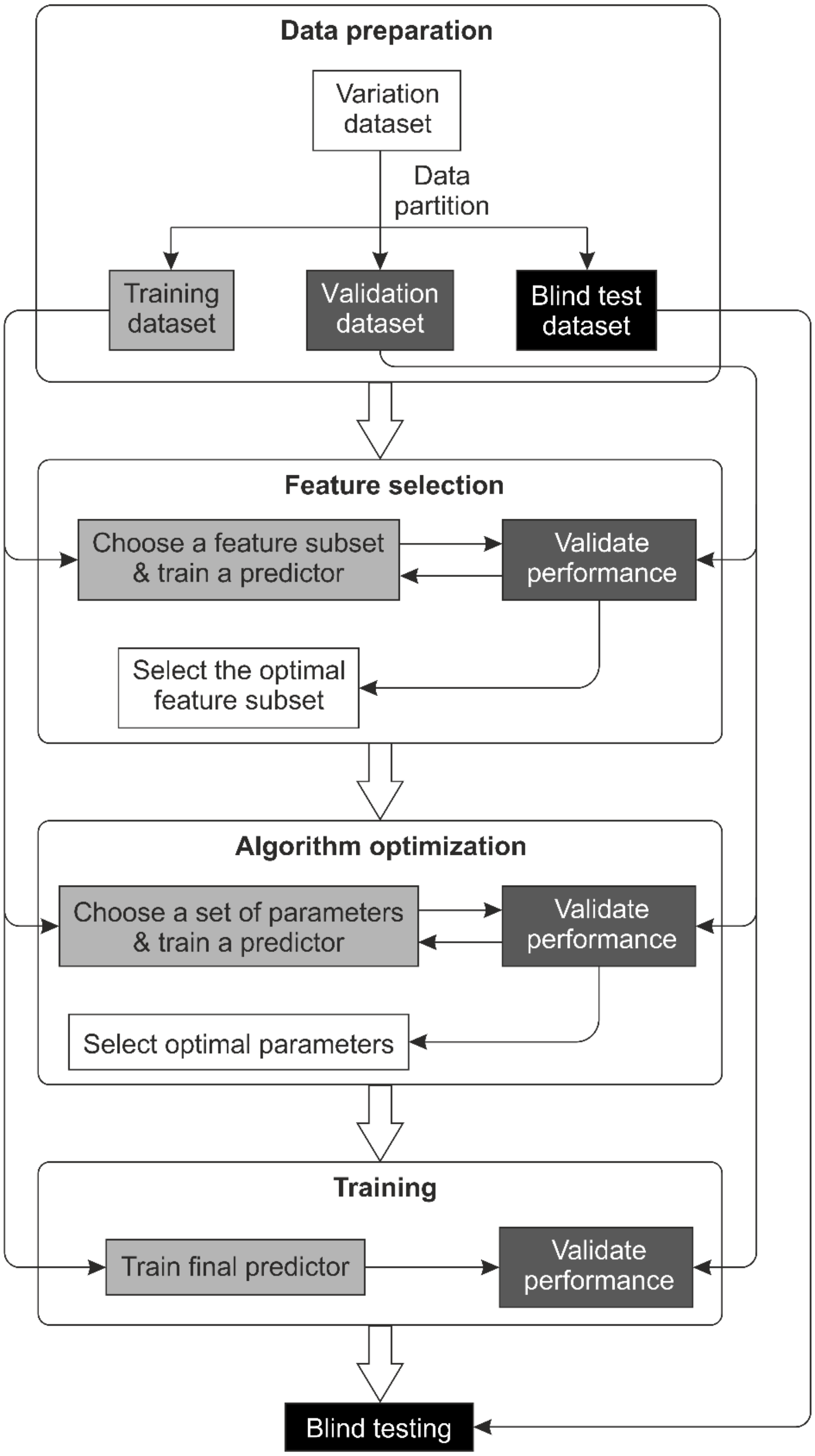


Figure 1.2: Framework for developing an ML tool. The framework is adapted from Niroula and Vihinen (2016). The data is split into training, validation, and blind test sets. The training and validation sets are used for feature selection, algorithm optimization, and training. The blind test set is used for testing the performance of the final ML model. Model selection should not be performed after blind testing.

A supervised ML model should be evaluated to estimate its performance. The same data cannot be used both for training and testing. Cross-validation is a common method for data partition and performance evaluation. In cross-validation, the data is split into several disjoint parts, one of which is used for testing the model trained by using the remaining parts. The method is repeated until all the disjoint parts are used for testing. The final performance is computed based on the performance of all the trained models. Another method for data partition is to split the data into two parts, one for training and another for testing. In some cases when the ML algorithms need to be optimized, the datasets are split into three parts- training, validation, and test datasets (Fig. 1.2).

One of the challenges in data partitioning is to represent the data structure in the partitions. Random sampling can be used to randomly select data points for partitioning and avoid systematic errors. Stratified random sampling is a method of splitting the data into strata based on one or more criteria and then selecting randomly from each stratum for data partitioning. In classification, an unbalanced data, i.e. containing different numbers of cases in different outcome classes, can lead to biased training. An unbalanced training dataset can reduce the performance of ML classifiers (Wei and Dunbrack, 2013). There are different methods for balancing the training data such as undersampling the majority class, oversampling the minority class, cost-sensitive learning, etc. (Vihinen, 2012; Wei and Dunbrack, 2013).

1.7.2 Algorithm optimization

Several supervised ML algorithms exist and their performances on various types of datasets have been compared (Wu, et al., 2003; Caruana and Niculescu-Mizil, 2006; Kotsiantis, et al., 2006; Statnikov, et al., 2008). The assessments show that the performances of the algorithms vary with the type of data and none of them are superior to the others on all types of data. RF, SVM, neural networks, and naïve Bayes methods often perform better than other methods. Along with the training data, each ML algorithm uses a set of pre-defined parameters known as hyper-parameters. As the hyper-parameters influence the learning process, optimizing them is an important step in developing an ML-based tool. The hyper-parameters vary according to the ML algorithm. The optimization step involves training several models using different sets of hyper-parameters and testing their performance using the validation data. The hyper-parameters showing the best performance are selected. Different approaches have been used for hyper-parameter optimization (Bergstra, et al., 2011; Bergstra and Bengio, 2012). Although the main aim of hyper-parameter optimization is to improve performance, it can also lead to overfitting.

1.7.3 Feature selection

Small sample size and high-dimensional feature space are the typical characteristics of biomedical data. As the number of features increases, the number of possible combinations of feature values increases exponentially. Thus, the sample size becomes sparse for describing the feature space, a phenomenon known as the ‘curse of dimensionality’. A model trained on such a dataset leads to overfitting and lacks generalization ability. Irrelevant and redundant features do not contribute to model performance but increase the model complexity and computation time. The data dimensionality can be reduced by applying a feature selection technique in which a subset of relevant and non-redundant features are selected from a large set of features. Feature selection reduces computation time, increases generalization ability and performance, and enables understanding of the feature importance (Saeys, et al., 2007; Ma and Huang, 2008). After feature selection, the final feature set consists of a subset of the original feature set. Feature extraction is another type of dimensionality reduction in which the initial high-dimensional data are transformed to final low-dimensional data (Bartenhagen, et al., 2010). In feature extraction, the identities of the original features are lost and a new set of features is generated.

The relevance of features for an ML task has been the most important criterion for feature selection. Feature relevance can be estimated either for each feature by ranking or for feature subsets by subset selection (Guyon and Elisseeff, 2003). Feature ranking deals with individual features and the feature dependencies are disregarded. Additionally, the redundancy between the features are unknown. In case of subset selection, all features in a subset are considered as a unit and their relevance is tested together. A generic feature selection algorithm consists of four steps as described below (Fig. 1.3).

i) Feature subset generation

First, a subset of features is generated to test its performance for a given ML task. There are several approaches to generate feature subsets including genetic algorithm (Yang and Honavar, 1998), simulated annealing (Debuse and Rayward-Smith, 1997), greedy hill climbing algorithms (Bordea, et al., 2015), and others (Kohavi and John, 1997). The greedy hill climbing algorithms are among the most widely used algorithms for feature subset generation. Sequential feature addition, backward elimination, and bi-directional selection are different versions of greedy hill climbing approach.

ii) Subset evaluation

The feature subset is used to train an ML-model and its performance is tested by using a validation dataset. The performance of the model is compared with the previous best performance and the best performing feature subset is chosen.

iii) Termination

After a feature subset has been evaluated, the algorithm checks if it meets any of the pre-defined termination criteria. If any of the criteria is met, the algorithm terminates. Otherwise, the algorithm starts a new iteration by generating a new feature subset. Some of the common termination criteria include the completion of a pre-defined number of iterations, the inability to improve the performance compared to previous iterations, the completion of all predefined feature subsets, etc. In case of lack of a suitable termination criterion, the algorithm may run exhaustively.

iv) Validation

Validation is performed after the feature selection has been completed. The selected feature subset is used to train an ML-model and its performance is evaluated using an independent test dataset.

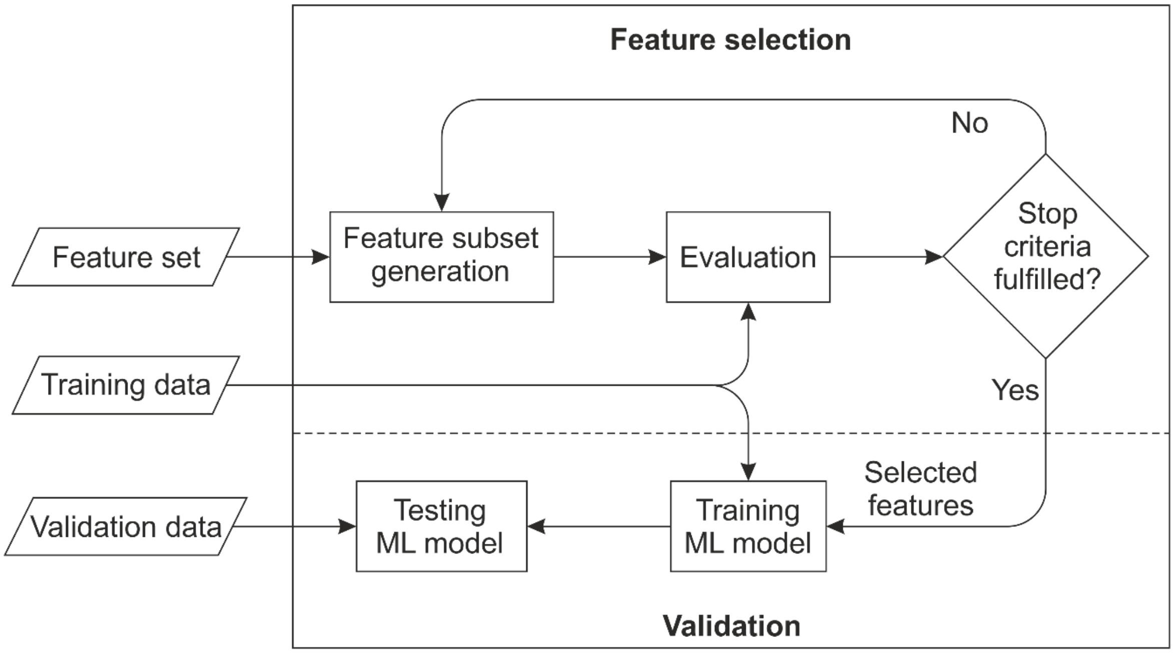


Figure 1.3: Schematic diagram of a standard feature selection approach. A feature subset is taken from the feature set and a prediction model is trained. The performance of the model is assessed and compared with the previous best performance. The feature subset with the best performance is selected. The stop criteria are tested and the algorithm iterates by generating a new feature subset unless a stop criterion is met. After a termination criterion is met, the best feature subset is used to train a model which is tested by using a validation dataset.

2. Aims of the study

The general aims of the study were to develop fast and accurate computational tools for predicting the impact of genetic variations and to apply them for analysing genetic variation datasets. More specific aims were as follows.

a) To identify useful features for classification of disease-causing and benign variations and use them to develop a fast and reliable tool for predicting the impact of AASs in human proteins (Paper I)

b) To develop a robust tool for classification of AASs in MMR proteins (Paper II)

c) To develop a tool for classification of mt-tRNA variations based on sequence information and additional evidence (Paper III)

d) To collect variations leading to different phenotypic severity and develop a tool for predicting the severity of disease-causing AASs in human proteins (Paper IV)

e) To study the impact of harmful AASs in cancer (Paper V)

3. Materials and Methods

3.1 Variation data

The variation datasets were collected from several variation databases and literature.

3.1.1 VariBench

VariBench is a database of benchmark variation datasets which contains variations collected from various sources (Nair and Vihinen, 2013). The datasets are widely used for training and testing prediction tools. The disease-causing and neutral variations used for training PON-P (Olatubosun, et al., 2012) were obtained from VariBench. The pathogenic dataset was collected from the PhenCode database (Giardine, et al., 2007), the IDbases (Piirilä, et al., 2006), and various LSDBs and the neutral dataset was collected from the dbSNP database (Sherry, et al., 2001).

An additional dataset which was used for training PON-MMR was also obtained from VariBench. The dataset contained 80 pathogenic and 88 neutral variations from MMR proteins (Ali, et al., 2012).

3.1.2 Locus specific databases

Pathogenic and neutral variations in specific proteins or genes were obtained from their respective LSDBs. The variants in the MMR proteins were obtained from the InSiGHT databases for the *MLH1*, *MSH2*, *MSH6*, and *PMS2* genes (Thompson, et al., 2014). The severe and less severe disease-causing variants in various genes/proteins were collected from the LSDBs hosted at LOVD (Fokkema, et al., 2011), UMD (Béroud, et al., 2000), and IDbases (Piirilä, et al., 2006).

3.1.3 Literature

Additional variations associated with pathogenicity, severity, and cancer were collected from literature. The somatic SNVs in 30 types of cancers from 7,042 samples were obtained from the Sanger Institute (ftp://ftp.sanger.ac.uk/pub/cancer/AlexandrovEtAl/). The variants were previously used for investigating the signatures of SNVs in cancer (Alexandrov, et al., 2013). The mt-tRNA variants classified by Yarham et al. and associated molecular evidence were collected (Yarham, et al., 2011). The severe and non-severe variants were collected from several publications containing case reports and genotype-phenotype correlations.

3.2 Sequences and structures

The DNA, RNA, and protein sequences for human genes were obtained from the Ensembl database (Yates, et al., 2016), the UniProtKB/SwissProt database (The UniProt Consortium, 2015), and the NCBI Reference Sequences (RefSeq) database (Pruitt, et al., 2014). The human mt-tRNA sequences were obtained from the mito-tRNAdb (Juhling, et al., 2009) and mapped to the revised Cambridge Reference Sequence (rCRS) of human mtDNA (NC_012920.1).

The 3-D structures for proteins and RNAs were obtained from the Protein Data Bank (PDB) (Berman, et al., 2000). The variants in the 3-D structures were visualized using the visualization software UCSF Chimera (Pettersen, et al., 2004).

3.3 Annotations, networks and pathways

The protein sequences were mapped to protein families from the Pfam database (Finn, et al., 2016) using the Ensembl BioMart tool (Kinsella, et al., 2011). The protein domains were obtained from the InterPro database (Mitchell, et al., 2015).

The GO annotations for human proteins were obtained from the Gene Ontology Consortium database (Ashburner, et al., 2000). The statistical analysis for enrichment of GO terms was performed by using topGO, an R statistical software package (Alexa, et al., 2006).

The functional protein interaction network was obtained from ReactomeFI (Wu, et al., 2010). The Cytoscape tool (Saito, et al., 2012) was used for visualizing the networks. The ReactomeFI plugin in Cytoscape was used for identifying enriched pathways in protein interaction network.

3.4 ML algorithm

The RF algorithm was chosen for developing the prediction tools. RF is a tree-based ensemble ML algorithm (Breiman, 2001). It consists of several trees each of which can predict the outcome. The final outcome of the algorithm is based on the votes obtained from all the trees. Each tree is generated by using a different training dataset selected by bootstrapping. Two-thirds of the cases in the training data are used for tree generation and the remaining one-third is used for estimating the error rate, a process known as out-of-bag (OOB) error estimation. The RF algorithm estimates the importance of each feature based on the OOB error estimation. The R statistical software package, randomForest (https://cran.r-project.org/web/packages/randomForest/index.html), was used for computing feature importance and developing prediction tools.

3.5 Features for ML

3.5.1 Evolutionary conservation features

We computed two sets of evolutionary conservation features. The first set of features was computed based on orthologous sequences. The orthologous sequences for human protein and coding DNA (cDNA) sequences were obtained from the Ensembl Compara database (Herrero, et al., 2016). The protein sequence for each protein was aligned with its ortholog sequences using ClustalW (Larkin, et al., 2007). Based on the MSA of protein sequences, a codon alignment of human cDNA sequence and its ortholog sequences were generated using the PAL2NAL tool (Suyama, et al., 2006). From the codon alignment, the codon-wise selective pressure was computed using the locally installed Selecton tool (Stern, et al., 2007). Additional features representing the frequency of reference and altered amino acids were computed from the MSA of the protein sequences.

Another set of evolutionary features was computed from the MSA of homologous sequences. The homologs of human protein sequences were obtained by running BLAST against a non-redundant protein sequence database. The homologous sequences for the mt-tRNA sequences were obtained from the Mammit-tRNA database (Putz, et al., 2007). The MSA of homologous sequences was generated using the ClustalW tool. From the MSA, the information content at each position in the alignment and the Position Specific Scoring Matrix (PSSM) were computed using the AlignInfo module in Biopython (http://biopython.org/DIST/docs/api/Bio.Align.AlignInfo-module.html).

3.5.2 GO terms-based feature

Features based on GO terms have been found to improve the performance of predicting variation impact (Kaminker, et al., 2007b; Calabrese, et al., 2009). All GO terms associated with human proteins were obtained and the ancestors for all of them were collected using the Bioconductor package GO.db (http://www.bioconductor.org/packages/2.13/data/annotation/html/GO.db.html). For each protein in the training dataset, the GO terms associated with a protein and their ancestors were collected.

Two bags of GO terms were generated, one containing the GO terms associated with proteins containing the pathogenic variations and another containing the GO terms associated with proteins containing the neutral variations. The GO terms associated with the proteins containing both pathogenic and neutral variations were present in both bags. The frequencies of each GO term in the pathogenic bag and in the neutral bag were computed and stored in a database. For each variation, the GO feature was computed by using the following formula

$$GO\ feature = \sum_{i=1}^{n} log\frac{f(P_i)+1}{f(N_i)+1},$$

where, n is the number of GO terms associated with the protein containing the variation, $f(P_i)$ is the frequency of the i^{th} GO term in the pathogenic bag, and $f(N_i)$ is the frequency of the i^{th} GO term in the neutral bag. One was added to the frequencies to avoid indeterminate ratios.

3.5.3 Biochemical properties of amino acids

The biochemical and physico-chemical properties of amino acids were used as features for training tools for interpreting impacts of protein variations. The biochemical properties of amino acids were obtained from the AAindex database (Kawashima, et al., 2008).

Several additional features were extracted based on the protein sequences and RNA sequences and structures. These are described in the respective publications included in this thesis (Papers III and IV).

3.6 Training and testing

The variation data were split into training and test datasets. Approximately one-tenth of the data was first separated for testing. The remaining data were used for training and feature selection. To avoid data circularity, disjoint training and test datasets were generated. The datasets were disjoint at different levels:

i) variant level: a variant was present in either the training or the test dataset (Papers II and III).

ii) protein level: variants in the same protein were present either in the training or the test dataset (Paper IV).

iii) protein-family level: variants in the proteins within the same protein family were present either in the training or the test dataset (Paper I).

Feature selection was performed by using the algorithm presented in Figure 1.3. Sequential feature addition and backward elimination were used for generating feature subsets. The evaluation step was performed using cross-validation. The dataset separated for training and feature selection was further split into five disjoint partitions. One partition was used as validation data and the remaining as training data. For each feature subset, a model was trained using the training data and the performance was computed using the validation data. The process was repeated until all partitions were used as validation data. The average performance of the models was used as the performance for the feature subset.

The tools were trained and validated using cross-validation and jackknife resampling methods. Cross-validation was used to train a model using certain proportion of the data and validate the model using the remaining data. Jackknife resampling was used to sample balanced training datasets (i.e. datasets containing equal numbers of variants in all the classes) and the remaining data were used for validation. To introduce variability to the training and validation datasets, the training and validation were performed multiple times (i.e. 2,000 times for PON-mt-tRNA and 100 times for PON-PS). All the tools were tested using blind test datasets.

3.7 Integration of ML prediction and evidence

The predictions from the ML-model were integrated with evidence from segregation, biochemical, and histochemical tests for classification of human mt-tRNA variants. The prediction obtained from the ML-predictor was used as a prior probability. The prior probability was integrated with the evidence from various sources to compute the posterior probability of pathogenicity based on which the variants were classified. The likelihood ratio (LR), posterior odds, and the posterior probability were computed using the following equations

$$LR = \frac{Probability\ of\ finding\ evidence\ for\ a\ pathogenic\ variation}{Probability\ of\ finding\ evidence\ for\ a\ neutral\ variation}$$

$$Posterior\ odds = LR \times \frac{Prior\ probability}{1 - Prior\ probability}$$

$$Posterior\ probability = \frac{Posterior\ odds}{1 + Posterior\ odds}$$

3.8 Performance evaluation measures

Performances of prediction tools were evaluated using various performance measures. Six standard performance measures were derived from a contingency matrix (Fig. 1.1). The measures were PPV, NPV, sensitivity, specificity, accuracy, and MCC. Additionally, when the numbers of positive and negative cases in the test dataset were unequal, the balanced accuracy (BACC) was used instead of accuracy. The ROC curves and AUC were also used to compare the performance of the tools. One additional performance measure was used to integrate all six performance scores, the overall performance measure (OPM). The relation between OPM and the six standard performance measures can be described by using an example of a cube. If O is the centroid of a cube, the six performance measures are represented along the six walls of the cube from O. PPV, NPV, sensitivity, and specificity represent two of the four cells in the contingency matrix. PPV and NPV are disjoint; sensitivity and specificity are also disjoint. These pairs are represented along the opposite walls of the cube. Accuracy and MCC represent all four cells in the contingency matrix and are represented along the remaining two walls of the cube. As the performance measures often have different values, they often form a cuboid instead of a cube. OPM is represented by the volume of the cuboid which is normalized to range from 0 (for total disagreement between prediction and actual class) to 1 (for total agreement between prediction and actual class). As MCC ranges from -1 to +1 and the remaining five measures range from 0 to 1, MCC is rescaled from 0 to 1 before computing OPM.

$$BACC = \frac{Sensitivity + Specificity}{2}$$

$$OPM = \frac{(PPV + NPV)(Sensitivity + Specificity)(Accuracy + \left(\frac{1 + MCC}{2}\right))}{8}$$

4. Summary of results

We developed generic and specific tools for predicting the impact of AASs in human proteins (Papers I, II, and IV) and SNVs in human mt-tRNAs (Paper III). We used the tools to predict the impact of somatic AASs in cancers (Paper V), all possible AASs in MMR proteins (Paper II), and all possible SNVs in mt-tRNAs (Paper III).

4.1 PON-P2: fast and reliable tool for identifying harmful variants

PON-P2 is a fast and reliable tool for predicting the pathogenicity of AASs in human proteins (Paper I). The tool is based on 8 features representing evolutionary conservation, GO annotations, and properties of amino acids which were identified by feature selection. PON-P2 predicts the pathogenicity of each variant by using 200 independent predictors and estimates the reliability of the prediction. The variations predicted with high reliability are classified as pathogenic or neutral and the remaining variants remain unclassified.

PON-P2 was trained and tested using benchmark variation datasets and had the best performance in the cross-validation as well as in the independent performance evaluation (Paper I). The tool consistently showed the best performance when tested with additional datasets (Table 4.1). The superior performance of PON-P2 has also been reported by independent studies (König, et al., 2016; Riera, et al., 2016). PON-P2 performed better than generic predictors as well as protein-specific predictors for variants in 70 out of 82 proteins (85.4%).

PON-P2 has been widely used since it became publicly available in July 2013. PON-P2 has received 2,688 queries from 580 unique users until 29 August 2016. The number of PON-P2 users is continuously increasing (Fig. 4.1a). Since December 2015, we are recording the number of variations predicted by PON-P2 for each submission. The tool has predicted pathogenicity for about 200,000 AASs during the last 8 months (Fig. 4.1a). Users can submit variations in four different formats – protein sequence identifier, genomic location, Variant Call Format (VCF) file, and protein sequence submission. The protein sequence identifier is the most widely used submission format (Fig. 4.1b). The number of submissions in the VCF format is low but they contain large numbers of variations. Recently, we developed an application programming interface (API) for PON-P2 and a plugin for the VEP tool (McLaren, et al., 2016). The API is useful for submitting queries to the tool and obtaining predictions programmatically. VEP is a tool for annotation of variations including the predictions of variation impacts.

Table 4.1: Performance comparison of PON-P2 with other tools on predictSNPSelected and SwissVarSelected datasets from Grimm et al. (2015).

	TP	TN	FP	FN	PPV	NPV	Sens	Spec	BACC	MCC
predictSNPSelected										
MT2[a]	50	502	274	15	0.15	0.97	0.77	0.65	0.71	0.23
PP2[a]	7941	4137	1961	2059	0.80	0.67	0.79	0.68	0.74	0.47
MASS[a]	7207	4353	1544	2714	0.82	0.62	0.73	0.74	0.74	0.45
SIFT[a]	7296	3914	1747	2287	0.81	0.63	0.76	0.69	0.73	0.45
LRT[a]	7573	3001	2207	2007	0.77	0.60	0.79	0.58	0.69	0.37
PON-P2[b]	5124	3173	345	590	0.94	0.84	0.90	0.90	0.90	0.79
PON-P2[c]	5116	3173	341	590	0.94	0.84	0.90	0.90	0.90	0.79
PON-P2[d]	1385	1243	186	210	0.88	0.86	0.87	0.87	0.87	0.74
SwissVarSelected										
MT2[a]	3391	4114	3180	829	0.52	0.83	0.80	0.56	0.68	0.36
PP2[a]	3086	5580	2623	1440	0.54	0.79	0.68	0.68	0.68	0.35
MASS[a]	2457	5214	2299	1943	0.52	0.73	0.56	0.69	0.63	0.25
SIFT[a]	2592	4828	2515	1617	0.51	0.75	0.62	0.66	0.64	0.26
LRT[a]	2985	3958	2675	1184	0.53	0.77	0.72	0.60	0.66	0.30
PON-P2[b]	1566	3412	818	773	0.66	0.82	0.67	0.81	0.74	0.47
PON-P2[c]	1551	3194	818	773	0.65	0.81	0.67	0.80	0.74	0.46
PON-P2[d]	737	1751	417	414	0.64	0.81	0.64	0.81	0.73	0.45

[a] Performance scores were obtained from Grimm et al. (2015).

[b] All variants predicted by PON-P2 tool

[c] All variants that were not present in the PON-P2 training data

[d] Variants in the proteins not present in the PON-P2 training data

MT2, MutationTaster2; PP2, PolyPhen-2; MASS, MutationAssessor; TP, True positive; TN, True negative; FP, False positive; FN, False negative; PPV, Positive predictive value; NPV, Negative predictive value; Sens, Sensitivity; Spec, Specificity; BACC, Balanced accuracy; MCC, Matthews correlation coefficient

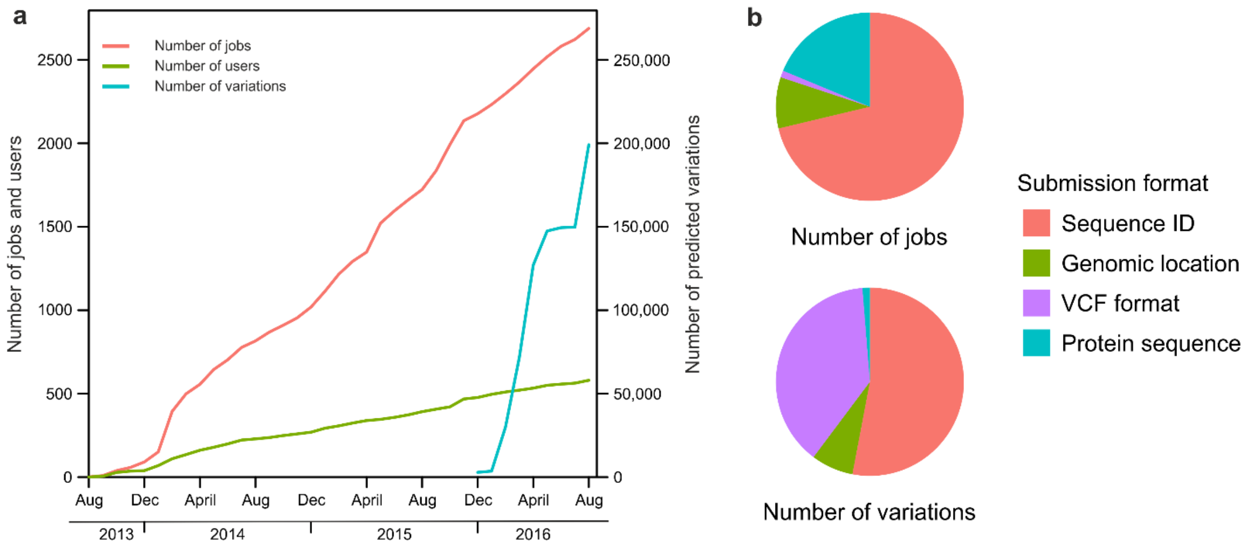
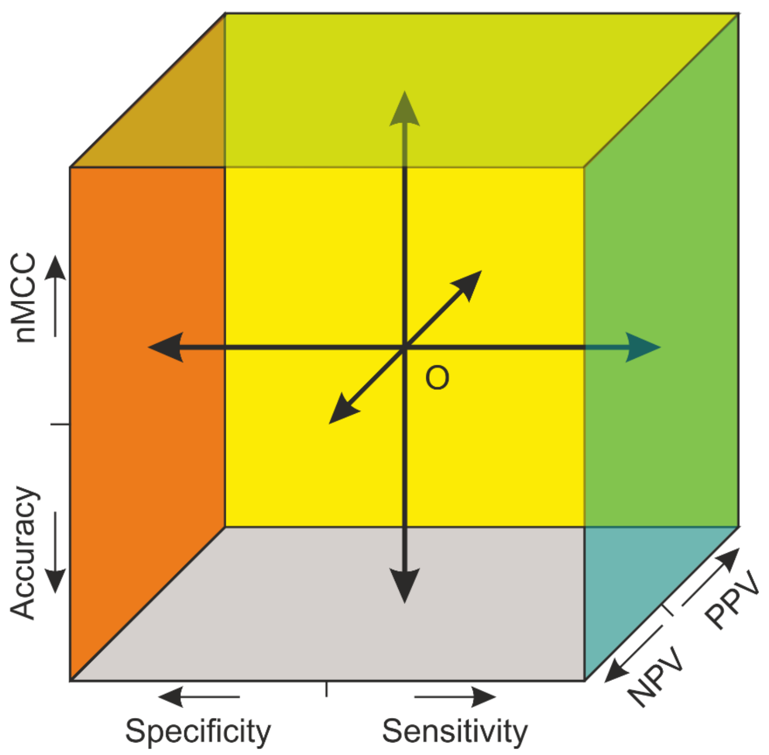


Figure 4.1: Usage statistics of PON-P2. a) The number of users and number of jobs submitted to PON-P2 is continuously increasing. The numbers of variations predicted by PON-P2 were recorded since December 2015. b) Number of jobs submitted to PON-P2 and number of variations predicted by PON-P2 for different submission formats. PON-P2 enables submission of variants in four formats. All the test submissions and other submissions from the members of our group are excluded.

In Paper I, we proposed a new performance measure, OPM, for assessing the performance of prediction tools. All performance measures do not represent all four cells in a contingency matrix. ML tools can have unbalanced performance scores due to various reasons. The tools can have high sensitivity but with a poor specificity or vice-versa. An unbalanced test data can result in unbalanced PPV and NPV. Trade-offs between sensitivity and specificity may be acceptable depending on the purpose of the tools. MCC is the only measure that handles these imbalances. Therefore, it is recommended to report all six performance measures (Vihinen, 2012). OPM integrates six standard performance measures- PPV, NPV, sensitivity, specificity, accuracy, and MCC. Figure 4.2 shows a framework of OPM with an example of a cube. The six performance measures are represented by the distance of the six walls from the centroid of the cube. OPM is given by the volume of the cube. As computational tools often have different scores for the performance measures, they generally form cuboids instead of the cubes.



$$\text{OPM} = \frac{(\text{PPV}+\text{NPV})\times(\text{Sensitivity}+\text{Specificity})\times(\text{Accuracy}+\text{nMCC})}{8}$$

Figure 4.2: Theoretical concept of OPM. Six standard performance measures are represented along the six walls of a cube from its centroid O. MCC is rescaled to adjust its range from 0 to 1. PPV and NPV, sensitivity and specificity, and accuracy and MCC are represented along the opposite walls on the same axes. OPM is obtained by computing the volume of the cuboid and rescaling the volume to range from 0 (total disagreement) to 1 (total agreement).

4.2 PON-MMR2 for classification of MMR variants

PON-MMR2 is a tool for classification of AASs in MMR system proteins. A total of 623 features were collected and a feature selection technique was applied to identify useful features for classifying MMR variants. Finally, 5 useful features were identified which represented evolutionary conservation and amino acid properties. The selected features were used to train a ML-based tool. The tool was tested using cross-validation as well as using an independent test dataset. In both tests, PON-MMR2 showed the best performance scores in comparison to generic prediction tools and other MMR-specific tools.

Using PON-MMR2, we classified all possible AASs at all positions in the four MMR proteins (MLH1, MSH2, MSH6, and PMS2). The proportion of pathogenic AASs varies between proteins. The proportion of predicted AASs is the lowest in PMS2 (22.3%) and the highest in MSH2

(55.3%). In addition, the proportion of pathogenic AASs was higher for the AASs that require more than one nucleotide substitution compared to the AASs caused by a single nucleotide substitution. In total, 44.6% of AASs that require multiple nucleotide substitutions are predicted to be pathogenic but only 28.5% of AASs caused by a single nucleotide substitution are predicted to be pathogenic. We mapped the AASs to protein domains and known 3-D protein structures. The pathogenic AASs are concentrated in the protein domains and in the α-helices and β-strands in the 3-D protein structures. Although the protein structures were not used for training the tool, the predicted pathogenicity is in line with the protein structure.

Table 4.2: PON-MMR2 predictions for AASs in MSH2 protein characterized by oligonucleotide-directed mutagenesis screening. Variants detected to be pathogenic or likely pathogenic by the mutagenesis screening and InSiGHT classification are listed.

			PON-MMR2 prediction	
AASs	Classification by mutagenesis study[a]	InSiGHT class[b]	Probability of pathogenicity	Classification
V63E	Partially pathogenic		0.98	Pathogenic
L93F	ND	4	0.81	Pathogenic
V161D	Pathogenic		0.20	Neutral
G162R	Partially pathogenic	5	0.768	Pathogenic
L173P	Pathogenic		1.00	Pathogenic
L173R	Pathogenic		0.99	Pathogenic
C333Y	Pathogenic		0.96	Pathogenic
L341P	Pathogenic	4	0.99	Pathogenic
V342I	Pathogenic		0.27	Neutral
P349L	Pathogenic	5	1.00	Pathogenic
P349R	Pathogenic	5	0.98	Pathogenic
D603N	Partially pathogenic		0.89	Pathogenic
G674A	Partially pathogenic		0.97	Pathogenic
G674R	Pathogenic	4	1.00	Pathogenic
G692R	Pathogenic	4	1.00	Pathogenic
P696L	Pathogenic	5	1.00	Pathogenic
C697Y	Pathogenic		0.96	Pathogenic
S723F	Pathogenic		0.98	Pathogenic
G759E	Partially pathogenic		0.99	Pathogenic
E878D	Pathogenic	4	0.03	Neutral

[a] Variants detected by screening method 1 are indicated as pathogenic, detected by screening method 2 are indicated as partially pathogenic, and those not detected by both methods are indicated as not determined (ND) (Houlleberghs, et al., 2016).

[b] The classification was taken from the InSiGHT database (Thompson, et al., 2014).

Variants for which the classification of the mutagenesis experiment, InSiGHT VIC, and PON-MMR2 do not agree are highlighted with grey shades.

PON-MMR2 is freely accessible at our website http://structure.bmc.lu.se/PON-MMR2. Users can either submit queries for one or more AASs or download the predicted pathogenicity for all AASs. Recently, a study used oligonucleotide-directed mutagenesis screening to characterize AASs in the MSH2 protein (Houlleberghs, et al., 2016). Among 59 AASs analysed, 19 were detected to be pathogenic or partially pathogenic. We used PON-MMR2 to predict their pathogenicity and found that 16 out of the 19 AASs were correctly classified (84.2%) (Table 4.2). When the study was published, 9 of the 59 variants were classified as pathogenic or likely pathogenic by the InSiGHT VIC (Thompson, et al., 2014). One of the nine classified variants could not be detected by the mutagenesis method which however is classified as pathogenic by PON-MMR2. On the other hand, PON-MMR2 incorrectly classified one of the nine variants which was detected by the mutagenesis method (Table 4.2).

4.3 PON-mt-tRNA for classification of mt-tRNA variants

PON-mt-tRNA is a tool for classification of human mt-tRNA variations (Paper III). The tool is based on a multifactorial probability and it consists of two parts: i) an ML predictor, and ii) LR based on evidence of segregation, biochemical and histochemical tests. The ML predictor is used to predict a prior probability of pathogenicity based on evolutionary conservation, base pairing, and mt-tRNA structures. If evidence from at least one of the three sources (segregation, biochemical test, histochemical test) is available, the prior probability of pathogenicity is integrated with evidence-based LR to compute the posterior probability of pathogenicity. The variants are classified into five classes (pathogenic, likely pathogenic, neutral, likely neutral, and unknown) based on the posterior probability of pathogenicity. If the evidence from the three sources is not known, the ML-based probability of pathogenicity is used to classify the variants. Both versions of PON-mt-tRNA performed better than the available prediction method. PON-mt-tRNA showed an accuracy of 99% when evidence from all three sources was used to classify variants and 69% when the evidence was not used (Paper III).

PON-mt-tRNA can be accessed at http://structure.bmc.lu.se/PON-mt-tRNA. Using PON-mt-tRNA, the pathogenicity of all possible single nucleotide substitutions in the 22 human mt-tRNAs were predicted. Approximately half of the variants (51%) were predicted as pathogenic. The proportion of predicted pathogenic variants was higher in the stems (61.5%) than in the loops (34.1%). The predictions for all possible substitutions can be downloaded from the website. The predictions are based on the ML predictor. If evidence from at least one of the three sources is known, the variants and the evidence can be submitted to PON-mt-tRNA for predicting the posterior probability of pathogenicity and classifying the variants.

4.4 PON-PS for predicting severity of disease-causing AASs

PON-PS is the first tool for predicting the severity of disease-causing AASs. A dataset containing 1,399 severe and 1,529 mild and moderate disease-causing AASs from 91 proteins was collected from various databases and literature. The variants in 8 proteins were separated for testing and the remaining variants were used for feature selection and training. Among 1,304 features collected from various sources, 10 features were identified as useful. These features represented evolutionary conservation, sequence environment, and properties of amino acids. We compared the predictions of available generic predictors for severe and non-severe variants. As the available tools do not classify severity of variants, their predictive performance could not be assessed. But the predicted scores were largely overlapping for most of them. MutationAssessor and PON-P2 showed the highest AUC, i.e. 0.64 and 0.63, respectively. These performance scores are far worse than their performances for distinguishing disease-causing and benign variations. Therefore, we developed PON-PS to predict the severity of disease-causing variations and to group them into severe and less severe. We compared the performance of PON-PS with MutationAssessor which showed the highest performance among available tools. PON-PS showed better performance in the cross-validation as well as in an independent test.

The performance of PON-PS was further validated by using variation datasets from four proteins encoded by *CFTR*, *BRCA1*, *VWF*, and *PAH* genes. The predicted severe variations in the protein encoded by *CFTR* gene have a higher salt chloride concentration compared to the non-severe variations. For the variants in *BRCA1* and *VWF*, the balanced accuracies of distinguishing the severe and non-severe variations were 75% and 66.7%, respectively. The severity of *PAH* variants follows closely the pattern of average phenylalanine levels in the individuals having the variations.

As PON-PS is trained on severe and less severe disease-causing variations, the benign variations have to be filtered before predicting severity. Therefore, the tool uses the PON-P2 tool for filtering out the neutral variations. PON-P2 was chosen because the tool has shown the best performance in several studies. PON-PS is available as a web tool at http://structure.bmc.lu.se/PON-PS.

4.5 Harmful somatic AASs in cancer

In paper V, we studied the impacts of somatic AASs in cancer. First, we assessed the performance of PON-P2 on validated cancer variation datasets. The cancer variations were predicted to have high probabilities of pathogenicity. The recurrent variations in the COSMIC database showed a similar pattern. However, the majority of the variants in the COSMIC database were predicted to have low probabilities of pathogenicity. Using PON-P2, we identified harmful somatic AASs in 30 types of cancer from 6,861 cancer samples (whole genome or exome sequences). The numbers of harmful variations vary between the cancers as well as within the individuals having the same type of cancer. Among 824,001 somatic AASs, only 14.2% were predicted to be harmful. The proportion of harmful AASs was higher i.e. 40% in the proteins encoded by the known cancer genes. We studied the landscape of all variations leading to AASs and those leading to harmful AASs at nucleotide, amino acid, and at protein domain levels. The landscapes were different for harmful AASs and all AASs.

As the mutation rate is high in cancer, harmful variations may have occurred by random chance and may not have any role in cancer development. Therefore, the proteins containing the harmful AASs were ordered and prioritized based on the number of samples affected by harmful AASs in them. The prioritized proteins were analysed in the context of a functional interaction network. The prioritized proteins are central in the network compared to other proteins containing harmful AASs and the average nodes in the network. The prioritized proteins had a higher degree of connectivity similar to the cancer proteins in a previous study (Sun and Zhao, 2010). The GO terms and pathways enriched in the prioritized proteins in 30 types of cancer were identified. Several of the identified GO terms and pathways were previously found to be implicated in cancer. Additionally, several new pathways were affected by the harmful AASs. The proteins involved in the enriched pathways affected different numbers of samples. As an example, the network of proteins containing the harmful AASs in head and neck cancer (HNC) is shown in Figure 4.3. The proteins involved in two pathways are marked by background colour. A pathway can be affected by harmful variations in any of the proteins involved in the pathway. Some proteins affect a large number of samples while others affect a smaller number of samples.

Several genes and pathways are often affected in various cancer types. We studied the similarities between the cancer types based on the overlapping proteins and pathways affected by the harmful AASs. The degree of overlap between the cancers were different at the protein level and at the pathway level. As several proteins are involved in a pathway, different proteins can affect the same pathway. On the other hand, a single protein is involved in several pathways, some of which can be significant in one cancer type and some other in the other type.

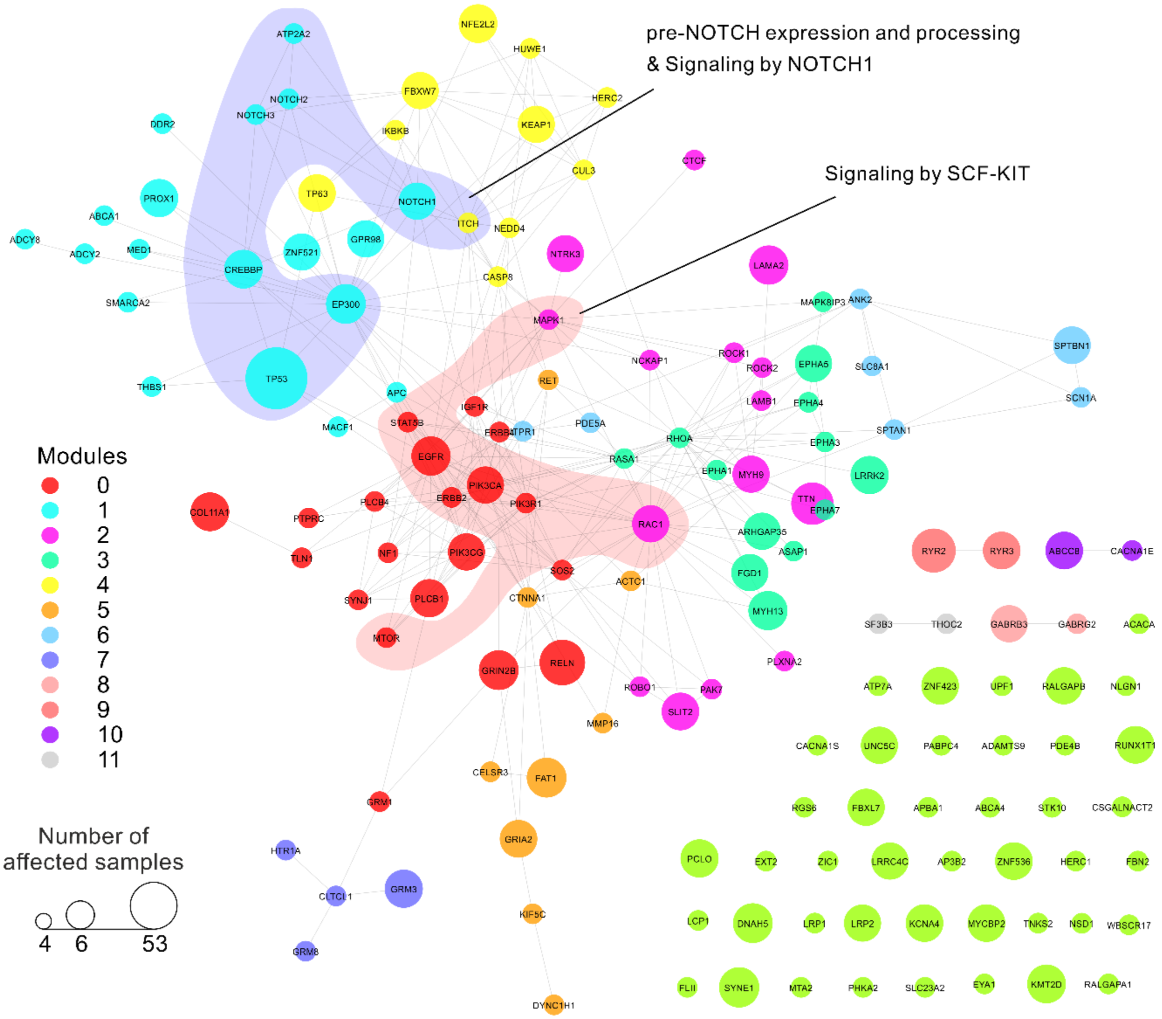


Figure 4.3: A functional interaction network of proteins containing harmful AASs in HNC. Two of the significantly enriched pathways are highlighted. These pathways are affected in the largest number of HNC samples. The network modules were identified by using the ReactomeFI plugin in Cytoscape.

5. Discussion

5.1 Generic and specific tools for variation interpretation

NGS methods are widely used to identify disease-causing variations. Early detection of harmful variations enables medical interventions for the patients and their relatives. However, large numbers of variations are identified in each individual. The disease-causing variation databases (Landrum, et al., 2014; Stenson, et al., 2014) are useful for annotating known disease-associated variations and the population genetic databases (Abecasis, et al., 2010; Fu, et al., 2013; Lek, et al., 2016) for excluding frequent variations in the population. Even after filtering variations from these databases, the disease-relevance of a large number of variations remain unknown. Guidelines for determining the pathogenicity of genetic variants have been developed (Thompson, et al., 2014; Richards, et al., 2015). These guidelines promote consistency in the classification of variations and harmonize quality of data. Using the guidelines, the classification for some of the variants has been re-assessed and corrected (Lek, et al., 2016; Walsh, et al., 2016). The ACMG guidelines and the InSiGHT VIC classification scheme recommend the use of computational predictions as one of several lines of evidence. Numerous computational tools have been developed for predicting variation impact; however, their performances are inconsistent in different studies (Grimm, et al., 2015; Masica and Karchin, 2016). Therefore, the choice of computational tools is critical. Systematic performance assessment of the available tools can provide useful information for choosing the best tools.

For clinical application, a tool must have a high reliability and it should be fast in order to handle the deluge of data. Various performance assessment studies have shown that most of the available tools have suboptimal performance (Thusberg, et al., 2011; Bendl, et al., 2014; Grimm, et al., 2015; Miosge, et al., 2015). In this study, we developed a fast and highly reliable tool, PON-P2, for ranking and prioritizing harmful AASs (Paper I). PON-P2 showed the best performance in our evaluation (Table 4.1 and Papers I, II and IV) as well as in independent studies (König, et al., 2016; Riera, et al., 2016).

Most computational tools classify variants into binary classes and some predict continuous scores for variants. PON-P2 estimates the reliability for its own predictions and groups variants into three classes: pathogenic, neutral, and unknown. The approach was previously applied to a meta-predictor, PON-P, developed in our group. The main advantage of this approach is that a certain proportion of variants can be predicted with a high reliability although a small fraction of variations remain unclassified. By grouping the variations predicted with a low reliability to the unknown class, PON-P2 reduces the chances of misinterpretation. The multifactorial methods, recommended by the VIC guidelines, classify variants into five classes and one of them is the unknown class (Thompson, et al., 2014). The variants are classified to one of the four classes (pathogenic, likely pathogenic, not-pathogenic, and likely not-pathogenic) when there is sufficient evidence and the variants are classified to unknown class when there is lack of sufficient evidence.

As the amounts of interpreted variation data are increasing, it has become possible to develop specific tools for various genes/proteins or diseases (Torkamani and Schork, 2007; Jordan, et al., 2011; Ali, et al., 2012; Masica, et al., 2015). We developed two specific tools: PON-MMR2 for classification of MMR variants (Paper II) and PON-mt-tRNA for classification of mt-tRNA variants (Paper III). The specific tools are trained by using a training dataset from specific genes/proteins or diseases. Therefore, the training data are likely to represent the mechanisms specific for the gene/protein or disease and the tools might have a higher performance compared to the generic tools (Jordan, et al., 2011; Ali, et al., 2012). On the other hand, the training and test datasets for specific tools are usually small which increases the chances of overfitting during training and over- or under-estimation of performance during testing. To address these issues, we trained and tested several prediction models by introducing variability to the training and the test datasets. We have tested the tools using independent datasets which were not used for training and feature selection.

For variations in many genes/proteins and diseases, both generic and specific tools can be used for interpretation. However, the use of generic tools may decrease with the increasing numbers of specific tools. Most available specific tools have shown similar or better performance than the generic tools. In this study, the generic tool PON-P2 and the specific tool PON-MMR2 showed similar performances for MMR variants (Paper II). However, PON-P2 could not reliably classify some variants. In a recent study, the performance of generic and specific prediction tools were compared for variants in 82 proteins. The results were mixed with generic tools performing better for some proteins and the specific tools for others (Riera, et al., 2016). Hence, both generic and specific tools are important and they can complement each other for reliable variation interpretation.

5.2 Predicting disease severity

Most diseases have a range of phenotypes, from mild to severe. Early identification of disease-causing variations and severity provides useful information for disease prognosis and clinical interventions for patients and their relatives. Knowledge of severity facilitates personalized medicine since it can be used for designing molecular tests, preventive interventions, and clinical monitoring. Individuals carrying severe variations may require immediate and intensive therapies to slow down disease progression or improve quality of life. On the other hand, individuals with milder variations can probably follow simpler preventive measures and get rid of unnecessary tests, therapies, and treatments.

Phenotypic severity due to genetic variation has been studied in relation to protein sequence and structure and endophenotypes (Robins, et al., 2006; Masica, et al., 2015; Reblova, et al., 2015; Sengupta, et al., 2015). These studies included variations in a single protein or disease. Such studies are important for studying the mechanisms of pathogenicity in specific diseases. However, data for performing such studies are available only for a small number of diseases. In this study, we collected variations associated with severe and less severe phenotypes from several proteins and diseases (Paper IV). There were no computational tools to predict disease severity due to variations. We tested the pathogenicity prediction tools and found that they cannot reliably distinguish severe from less severe variants. Although the majority of the tools often obtain an accuracy of over 75% for distinguishing disease-causing variations, they showed poor performance for predicting severity. Therefore, we developed a novel tool, PON-PS, for predicting severity due to AASs (Paper IV). The tool classifies the disease phenotype due to AASs into severe and less severe. The collected data and the developed tool will be of high importance for researchers and clinicians.

PON-PS is the first tool for distinguishing severe and less severe variants. The accuracy of predicting severity obtained is lower than the accuracy of predicting disease relevance. PON-PS tool showed a higher performance compared to the pathogenicity prediction tools. The evolutionary conservation features, which are powerful predictors of disease relevance, showed lower power to predict severity. Both severe and less severe variations are highly conserved. However, the evolutionary conservation features were among the useful ones identified during feature selection.

Several variations are associated with heterogeneous phenotypic severity. These variants were excluded from the training and the test datasets of PON-PS. The severity due to these variations are challenging to interpret. Several genetic and non-genetic factors are associated with pathogenicity and disease phenotype (Cutting, 2010; Cooper, et al., 2013). A recent pathogenicity model describes pathogenicity at a population level and consists of three components- severity, extent, and modulation (Vihinen, submitted). All three components are required to describe pathogenicity. The pathogenicity model at population level enables defining pathogenicity and phenotypic severity at individual level. Additional information about the patients is required for reliable interpretation of phenotypic severity. Disease specific tools capable of integrating multifactorial evidence will likely improve prediction of phenotypic severity. PON-PS can be integrated with other sources of evidence for developing disease-specific severity prediction tools. Since the additional information may vary with diseases and proteins, such tools can be developed only for certain diseases with sufficient data.

5.3 Useful features for variation impact prediction

Various types of information have been used for predicting impacts of variations. Features have mainly been derived from protein sequences and structures (Tang and Thomas, 2016). In this study, we used mostly features derived from the sequences since the 3-D structures are not known for most of the human proteins. Several features can be used to describe genetic variations. But the features may or may not be relevant to the mechanism of variation impact. Non-relevant features increase noise to the training dataset and may reduce performance of ML-models. Redundant features do not improve model performance but increase computation time. Feature selection techniques are useful for finding a set of relevant and non-redundant features. The feature selection technique reduces the complexity of the prediction model and reduces training and prediction time without decreasing model performance. We applied systematic feature selection techniques to find the most useful features (Papers I, II, and IV). Evolutionary conservation, GO-based feature, and properties of amino acids were among the most useful features.

Most tools for variation impact prediction use evolutionary conservation in one form or the other (Niroula and Vihinen, 2016; Tang and Thomas, 2016). We used two types of evolutionary conservation features. One set of evolutionary features were derived from the MSA of orthologous sequences and another set from the MSA of homologous sequences. The quality of MSA was found to impact the performance of predictors using evolutionary conservation (Ng and Henikoff, 2006; Thusberg, et al., 2011). As orthologous sequences retain their function, the MSA based on orthologous sequences is expected to be of higher quality than the MSA based on homologous sequences. In Papers II and IV, both sets of evolutionary features were tested. In paper II, the features derived from the MSA of homologous sequences were selected over the features derived from the MSA of orthologous sequences. However, a direct comparison between the MSAs cannot be made since different types of features were derived from the two MSAs. In Paper IV, the two sets of evolutionary features showed a complementary effect in predicting severity. However, the contribution of evolutionary conservation for predicting disease severity was not as high as in the case of predicting pathogenicity (Paper IV).

GO terms describe genes and gene products. Some studies found that a GO-based feature improved the performance of predicting the variation impact (Kaminker, et al., 2007b; Calabrese, et al., 2009; König, et al., 2016). In this study, we tested the importance of a GO-based feature on variation impact prediction and found similar results (Paper I). The GO-based feature is specific for proteins and all variants in a protein, both pathogenic and benign, have the same value. Therefore, the presence of variations from the same protein in the training and the test datasets can introduce bias in the training and testing process (Grimm, et al., 2015). We addressed this issue carefully by partitioning the training and the test datasets so that all variations from the proteins in a protein-family were kept together either in the training or the test dataset. Such an approach of data partition handles possible circularity from two sources. Firstly, the approach avoids any performance bias due to the GO-based feature and secondly, avoids bias due to variants in similar protein sequences.

The functionally and structurally important sites in protein sequences are annotated in different databases. These sites are highly conserved between the species and any variations at these sites are highly deleterious (Bartlett, et al., 2003). However, the number of variations at such functional sites is small. Variations at those sites are likely selected against and eliminated from nature. The known functional sites provide useful information for interpreting impacts of variations at those sites. However, our understanding of the structure and function of human proteins is incomplete and the functional and structural sites in many proteins are not known.

Although 3-D structures are not available for most of the human proteins, the structure-based features can improve performance when used together with sequence-based features (Capriotti and Altman, 2011a; Capriotti, et al., 2013b). However, there are limitations for using structure-based features. The size of training and test dataset and the applicability of the tool will be reduced significantly unless predicted protein structures are used. The features derived from the predicted protein structures have been used for predicting variation impact (Yates, et al., 2014). Even though the tools developed in this study do not use features based on protein structure, they perform better than those that use structural features. For developing PON-mt-tRNA, we have used features derived from the secondary and tertiary structures of mt-tRNAs. However, the two features derived from the structure were the least important features for classification of mt-tRNA variations (Paper III).

Several lines of evidence are required for reliable classification of variations (Yarham, et al., 2011; Thompson, et al., 2014; Richards, et al., 2015). Evidence from different sources can be integrated to predict posterior probability for classification (Lindor, et al., 2012). In Paper III, we have integrated ML prediction and evidence from three sources to predict posterior probability of pathogenicity. The performance of the integrated tool was almost perfect which is extremely high compared to the ML approach alone. However, the development of such tools is hindered by the lack of additional data or evidence. As data collection is becoming more systematic, the amount of additional data is likely to increase in the future.

5.4 Harmful variations in cancer

Cancer genomics is a rapidly expanding research area. Large cancer projects such as TCGA and the ICGC collect and share genomic, transcriptomic, proteomic, epigenetic, and other data from large numbers of cancer samples. The data generated by these projects have driven various discoveries (see (Tomczak, et al., 2015) for some examples). Different approaches have been taken to understand cancer development and to identify the implicated genes, networks, and pathways in cancer. Interpretation of the massive amounts of data has been challenging due to the large number of passenger variations. Recurrent variations and driver genes among the large number of samples have been identified in many cancer types. The identification of rare driver variations remains to be challenging. Additionally, it is difficult to identify causative variations in a cancer sample which is one of the limitations for applying precision medicine in cancer.

In this study, we exploited variation impact and frequency of protein impairment to identify affected pathways in 30 types of cancer (Paper V). The predicted variation impact facilitates prioritization of likely harmful variations. The PON-P2 prediction tool was used to identify the harmful variations. The tool was first validated using recurrent variants in COSMIC and additional cancer variants. We could filter out a large fraction of the AASs identified in the cancers using PON-P2 prediction. Although PON-P2 showed high performance during validation, there are some false positives and false negatives at a low rate. Additionally, several harmful variations may have occurred by random chance due to a high mutation rate. And random harmful variations may not be relevant to cancer despite being harmful for a normal cell. So, we ranked the proteins containing harmful AASs based on the number of samples containing harmful AASs in the proteins. The most frequently affected proteins were prioritized and were used to identify significantly enriched GO terms and pathways. The pathways identified in this study included several novel and previously known pathways.

Large scale genomic studies have revealed the heterogeneous nature of cancers. Variation patterns are diverse even in tumors originating from the same tissue or organ while similar patterns of genomic alterations are observed in cancers from different tissues of origin (Alexandrov, et al., 2013; Ciriello, et al., 2013b; Lawrence, et al., 2013). We studied the relation between cancers based on the proteins containing harmful AASs and pathways affected by them. The cancers have overlapping proteins and pathways; however, the overlaps are not consistent at protein and pathway level (Paper V). A pathway can be affected by an impaired function of any of the several proteins involved in the pathway. Therefore, the relationship between cancers can be better understood at pathway level than at the protein level.

Variation impact can be used for filtering variations in cancer. Several computational tools (both generic as well as cancer specific) are available for predicting the impacts of nsSNVs in cancer (Raphael, et al., 2014; Tian, et al., 2015; Niroula and Vihinen, 2016). However, the tools have varying performances and even minor differences in the performance lead to large numbers of differently predicted variations when applied to large datasets. Most tools predict the impact of individual variation as an independent event which however is not true in cancer. But, the combined impact of large number of variations cannot be reliably predicted with the available tools and data.

5.5 ML approach for variation interpretation

In this study, we used a systematic approach for developing four ML-based tools for variation interpretation. Benchmark datasets, systematic feature selection, and appropriate training and testing strategies were applied. Although the general approach was similar for all the tools, they differ in scopes and implementations. The available data and knowledge have influenced our approach to train and test these tools. The largest dataset was available for developing PON-P2. We partitioned the data for training and testing at the protein family level to avoid data circularity. Such a data partition enabled us to use GO feature (a protein-specific feature) without affecting the reliability of the test results. Such a strict data partition could not be applied to other datasets due to their small sizes. To avoid data circularity, we partitioned the data at protein (Paper IV) and at variant levels (Papers II and III). Further, we did not use GO feature and any other features specific for proteins or genes.

In Paper I, we used multiple predictors trained by using bootstrap datasets (data generated by random sampling with replacement). The predictions obtained from all the predictors were used to estimate the reliability of prediction and classify variants into three classes. Due to the small size of data, the bootstrap approach could not be implemented for other tools. In the bootstrap method, the same variants can be randomly selected multiple times and the repetition of cases in a small training data would have a larger impact. In Paper II, only one predictor was trained after testing the approach by cross-validation. In Papers III and IV, we used ensemble predictors by sampling different sets of training and test datasets by the jack-knife approach. The jack-knife approach introduced variability in the training and test datasets and enabled a reliable estimation of the tools' performance. In all cases, the tools were additionally tested by using independent test datasets.

Different features were used for protein variations and for RNA variations. For protein variations, we collected features from the protein sequences and biochemical properties of amino acids (Papers I, II, and IV). We tested several features known to improve performance as well as new features that could be relevant for variation interpretation but were never tested before. In PON-P2, we used features for functional and structural annotations at the variant site. Since variations at known functional and structural sites are likely deleterious, the information about such sites is important for recognizing harmful variations. However, we could not use these features for training ML predictor as these contained missing values. We integrated these features with the predictions of ML models using a probability rule. In PON-mt-tRNA, we collected 9 features from the RNA sequences and structures for training an ML predictor. In addition, experimental data were available for all variants. We integrated the ML predictor and the LR of pathogenicity based on the experimental data for classifying pathogenic and neutral variations. Such experimental data facilitate a reliable interpretation of variation impact as was observed in Paper III.

A single performance measure cannot represent overall performance of a prediction tool. Several performance measures are required to reliably assess performance of the tools. For performance assessment, we used six standard performance measures. When comparing various tools, the performance scores between the tools do not correlate. For example, tools can have a high sensitivity but a low specificity or vice versa. Therefore, we have proposed a new performance measure, OPM, which measures the overall performance of prediction tools. The OPM enables easy comparison of the prediction tools by computing a single measure based on the six standard performance measures.

6. Summary and conclusions

Variation interpretation is a highly active and dynamic topic. The amount of variation data is increasing rapidly. The biological databases are expanding with a bulk of information. Although the population and disease-causing variation databases are increasing, disease relevance of a large number of variations are not known. Interpreting the impacts of variations is critical for diagnosis and treatment of patients and their family members. Computational tools are useful for ranking variations and prioritizing likely harmful variations for characterizing their disease relevance. In this study, we have implemented a systematic approach for developing computational tools for variation interpretation. We developed four tools for interpreting the impacts of amino acid and nucleotide substitutions in proteins and mt-tRNAs (Papers I, II, III, and IV). Benchmark variation datasets were collected and were used for systematic feature selection, training, and testing. The developed tools have shown the best performance in various performance assessment studies. All the tools were validated and were used for analyzing AASs in MMR proteins, SNVs in mt-tRNA genes and somatic AASs in cancer.

ML algorithms are powerful for generalizing the patterns in data. We used RF algorithm for developing the tools. The reliability of ML-based tools depend on the training dataset, features used to describe data, and the approach of training and performance assessment. Benchmark datasets are the best option for training and testing ML-based tools. As large number of features can be extracted for variations, feature selection is important for choosing a relevant and non-redundant feature set. We used validated benchmark datasets and features identified by performing a systematic feature selection for training. The performance of the trained method should be assessed using an independent dataset. Circularity in the training and test datasets leads to biased performance scores (Grimm, et al., 2015). As circularity can occur at different levels, we used the strictest criteria possible for assessing the tools. The performance assessments were unbiased which is supported by the performance shown by PON-P2 and PON-MMR2 in independent studies. They show similar or better performance than obtained during our performance assessments.

Generic tools are trained by using variants from a wide range of proteins and diseases. They find patterns from variations in various proteins and diseases. The generic tools are important for scanning harmful variations in all proteins and diseases. On the other hand, specific tools are trained by using variation data from specific proteins or diseases. With increasing amounts of data, it will be possible to develop more specific tools in the future. However, both generic and specific tools are required for reliable variation interpretation because of their complementary roles (Riera, et al., 2016). Here, we developed two generic tools and two specific tools. PON-mt-tRNA uses the genetic information and evidence from patients and molecular tests to classify the disease relevance of variants (Paper III). Such multifactorial tools have shown high performances but the additional information required for developing them is scarce. Patient information along with genetic data will increase our understanding of pathogenicity and improve our abilities to interpret the consequences of variations. However, it is difficult to obtain patient information due to various reasons such as patient security and privacy (Shabani and Borry, 2015).

Early identification of harmful variations facilitates early diagnosis and clinical intervention. Patients and their family members can benefit from preventive interventions, clinical monitoring and prioritized molecular tests. The tools developed in this study promote early identification of harmful variations. As the tools are based on statistics, additional evidence is required to verify their disease relevance. The tools are important for scanning the most likely harmful variations and prioritizing them for experimental evaluation to obtain additional evidence. Although the tools showed the best performance when compared with other available tools, more accurate tools are required for predictive medicine. Availability of reliable variation data and patient information enables developing powerful tools.

We implemented a systematic method for developing ML-based tools for interpreting the impacts of SNVs and AASs. Such a method can be implemented to develop tools for diverse application areas. Several loci in the non-coding regions have been associated with various common diseases. Reliable tools are needed to interpret impacts of non-coding variations. Most variation impact tools including those developed in the present study interpret the impact of each variation as an independent event. However, variations at different sites in the same gene are common even in monogenic disorders. In multigenic or multifactorial disorders, several variations and factors contribute to pathogenicity. Tools to interpret the combined impact of several variations would be of high importance. Such tools will also be useful for whole genome and exome interpretation. Reliable genome and exome interpretation would facilitate precision medicine.

7. Acknowledgements

I would to like to thank everyone who have supported me during my PhD studies.

Firstly, I want to thank my supervisor Mauno Vihinen for your guidance and support before and during my PhD studies. You have been a source of encouragement and inspiration for me. This thesis has been possible only by your great support and inspiration. You have been an amazing supervisor. Thank you for believing in me and continuously motivating me.

I also express my gratitude to my co-supervisor, Jens Lagerstedt, for your support. You have been very motivating and inspiring. Thank you for the interesting discussions and your guidance.

Many thanks to all the past and present group members. Gabriel, thanks for everything. We have shared office since the first day in BMC. We have had long discussion on different things and I have to learn a lot from you. These discussions have been key to be optimistic at times of frustration. You have been an amazing colleague and a great teacher for me. Thank you Gerard for your support. It has been great working with you and sharing the office. Thanks for sharing various information. Special thanks for proofreading my lengthy manuscripts and this thesis. Thanks to Yang and Siddhaling for good times and your support.

I also want to thank Jouni for your technical support during the beginning days, the Late Ayodeji Olabutosun for his encouragement and simplifying my programming and machine learning lessons, and Tiina for interesting discussions and motivation. Many thanks to people working in BMC D10 and B13 for nice discussions. Thanks to the administrative staffs in the Faculty of Medicine and Department of Experimental Medical Sciences for the practical help and the IT service for technical help.

Thank you all my friends in Lund and Malmö. It would not have been so much fun to live in Lund without all of you. Sudip, Aruna, Rajendra, Shiva, Jasmine and Abhishek, thanks for the wonderful time from the beginning days in Lund. Those events and trips we organized together were awesome. Thank you Riju, Kiran, Beer, Sushma, Suraj, and Sheeva. Those moments we have shared are memorable. I have not missed much of our festivals because of all the events and trips. Lund has been a good place to live in and it is all because of you.

Kara, thank you very much for giving me a place to live when I arrived Lund. You are a nice and kind person. It would have been difficult to get used to in Lund without your help. Thank you for helping and motivating me to learn Swedish.

Thank you Subas Dai, Binita bhauju, Bala dai, and Mithila bhauju for your continuous motivation, support, and love. Subas dai, you have inspired me all the time. Bala dai, my interest in programming grew all because of you which is an integral part of this thesis. Also many thanks to Narayan, Prakash, and all my friends for your support.

I thank my family for your unconditional love and support. Anuj and Ranjana, thank you for supporting me. Nabina, I cannot appreciate your love and support in words. This thesis would not have been possible without your support. Thank you for everything. I would not stand where I am today without my parents' efforts. Thank you for believing in me and letting me grow. This work is for all of you.

8. References

Abbott JA, Francklyn CS, Robey-Bond SM. 2014. Transfer RNA and human disease. Front Genet 5:158.

Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. 2010. A map of human genome variation from population-scale sequencing. Nature 467(7319):1061-1073.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. Nat Methods 7(4):248-249.

Ainscough BJ, Griffith M, Coffman AC, Wagner AH, Kunisaki J, Choudhary MNK, McMichael JF, Fulton RS, Wilson RK, Griffith OL, Mardis ER. 2016. DoCM: a database of curated mutations in cancer. Nat Methods 13(10):806-807.

Alexa A, Rahnenfuhrer J, Lengauer T. 2006. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics 22(13):1600-1607.

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale A-L, Boyault S, Burkhardt B, et al. 2013. Signatures of mutational processes in human cancer. Nature 500(7463):415-421.

Ali H, Olatubosun A, Vihinen M. 2012. Classification of mismatch repair gene missense variants with PON-MMR. Hum Mutat 33(4):642-650.

Ali H, Urolagin S, Gurarslan O, Vihinen M. 2014. Performance of protein disorder prediction programs on amino acid substitutions. Hum Mutat 35(7):794-804.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, et al. 2000. Gene ontology: tool for the unification of biology. Nat Genet 25(1):25-29.

Bao L, Cui Y. 2005. Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. Bioinformatics 21(10):2185-2190.

Bao L, Zhou M, Cui Y. 2005. nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. Nucleic Acids Res 33(Web Server issue):W480-482.

Bartenhagen C, Klein HU, Ruckert C, Jiang X, Dugas M. 2010. Comparative study of unsupervised dimension reduction techniques for the visualization of microarray gene expression data. BMC Bioinformatics 11:567.

Bartlett GJ, Borkakoti N, Thornton JM. 2003. Catalysing new reactions during evolution: Economy of residues and mechanism. J Mol Biol 331(4):829-860.

Bashashati A, Haffari G, Ding J, Ha G, Lui K, Rosner J, Huntsman DG, Caldas C, Aparicio SA, Shah SP. 2012. DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. Genome Biol 13(12):R124.

Bendl J, Stourac J, Salanda O, Pavelka A, Wieben ED, Zendulka J, Brezovsky J, Damborsky J. 2014. PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. PLoS Comput Biol 10(1):e1003440.

Bergstra J, Bardenet R, Bengio Y, Kégl B. Algorithms for hyper-parameter optimization; 2011 2011-12-12; Granada, Spain. Neural Information Processing Systems Foundation.

Bergstra J, Bengio Y. 2012. Random search for hyper-parameter optimization. J Mach Learn Res 13(1):281-305.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. Nucleic Acids Res 28(1):235-242.

Bermejo-Das-Neves C, Nguyen HN, Poch O, Thompson JD. 2014. A comprehensive study of small non-frameshift insertions/deletions in proteins and prediction of their phenotypic effects by a machine learning method (KD4i). BMC Bioinformatics 15:111.

Béroud C, Collod-Béroud G, Boileau C, Soussi T, Junien C. 2000. UMD (Universal mutation database): a generic software to build and analyze locus-specific databases. Hum Mutat 15(1):86-94.

Bertrand D, Chng KR, Sherbaf FG, Kiesel A, Chia BK, Sia YY, Huang SK, Hoon DS, Liu ET, Hillmer A, Nagarajan N. 2015. Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles. Nucleic Acids Res 43(7):e44.

Bordea G, Panthong R, Srivihok A. 2015. Wrapper feature subset selection for dimension reduction based on ensemble learning algorithm. Procedia Comput Sci 72:162-169.

Bozic I, Antal T, Ohtsuki H, Carter H, Kim D, Chen S, Karchin R, Kinzler KW, Vogelstein B, Nowak MA. 2010. Accumulation of driver and passenger mutations during tumor progression. Proc Natl Acade Sci U S A 107(43):18545-18550.

Breiman L. 2001. Random Forests. Mach Learn 45(1):5-32.

Buske OJ, Manickaraj A, Mital S, Ray PN, Brudno M. 2013. Identification of deleterious synonymous variants in human genomes. Bioinformatics 29(15):1843-1850.

Cai Z, Tsung EF, Marinescu VD, Ramoni MF, Riva A, Kohane IS. 2004. Bayesian approach to discovering pathogenic SNPs in conserved protein domains. Hum Mutat 24(2):178-184.

Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R. 2009. Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum Mutat 30(8):1237-1244.

Caldovic L, Abdikarim I, Narain S, Tuchman M, Morizono H. 2015. Genotype-phenotype correlations in ornithine transcarbamylase deficiency: A mutation update. J Genet Genomics 42(5):181-194.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

Capriotti E, Altman RB. 2011a. Improving the prediction of disease-related variants using protein three-dimensional structure. BMC Bioinformatics 12 Suppl 4:S3.

Capriotti E, Altman RB. 2011b. A new disease-specific machine learning approach for the prediction of cancer-causing missense variants. Genomics 98(4):310-317.

Capriotti E, Altman RB, Bromberg Y. 2013a. Collective judgment predicts disease-associated single nucleotide variants. BMC Genomics 14 Suppl 3:S2.

Capriotti E, Calabrese R, Fariselli P, Martelli PL, Altman RB, Casadio R. 2013b. WS-SNPs&GO: a web server for predicting the deleterious effect of human protein variants using functional annotation. BMC Genomics 14 Suppl 3:S6.

Capriotti E, Fariselli P, Casadio R. 2004. A neural-network-based method for predicting protein stability changes upon single point mutations. Bioinformatics 20 Suppl 1:i63-68.

Capriotti E, Fariselli P, Rossi I, Casadio R. 2008. A three-state prediction of single point mutations on protein stability changes. BMC Bioinformatics 9 Suppl 2:S6.

Capriotti E, Nehrt NL, Kann MG, Bromberg Y. 2012. Bioinformatics for personal genome interpretation. Brief Bioinform 13(4):495-512.

Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. 2013. Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics 14 Suppl 3:S3.

Caruana R, Niculescu-Mizil A. 2006. An empirical comparison of supervised learning algorithms. Pittsburgh, Pennsylvania, USA: ACM. p 161-168.

Cerami E, Demir E, Schultz N, Taylor BS, Sander C. 2010. Automated network analysis identifies core pathways in glioblastoma. PLoS One 5(2):e8918.

Chan PP, Lowe TM. 2009. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. Nucleic Acids Res 37(Database issue):D93-97.

Chao EC, Velasquez JL, Witherspoon MS, Rozek LS, Peel D, Ng P, Gruber SB, Watson P, Rennert G, Anton-Culver H, Lynch H, Lipkin SM. 2008. Accurate classification of MLH1/MSH2 missense variants with multivariate analysis of protein polymorphisms-mismatch repair (MAPP-MMR). Hum Mutat 29(6):852-860.

Chen J, Sun M, Shen B. 2015. Deciphering oncogenic drivers: from single genes to integrated pathways. Brief Bioinform 16(3):413-428.

Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. 2012. Predicting the functional effect of amino acid substitutions and indels. PLoS One 7(10):e46688.

Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly (Austin) 6(2):80-92.

Ciriello G, Cerami E, Aksoy BA, Sander C, Schultz N. 2013a. Using MEMo to discover mutual exclusivity modules in cancer. Curr Protoc Bioinformatics Chapter 8:Unit 8 17.

Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C. 2013b. Emerging landscape of oncogenic signatures across human cancers. Nat Genet 45(10):1127-1133.

Collins FS, Green ED, Guttmacher AE, Guyer MS. 2003. A vision for the future of genomics research. Nature 422(6934):835-847.

Conchillo-Sole O, de Groot NS, Aviles FX, Vendrell J, Daura X, Ventura S. 2007. AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. BMC Bioinformatics 8:65.

Cooper DN, Krawczak M, Polychronakos C, Tyler-Smith C, Kehrer-Sawatzki H. 2013. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. Hum Genet 132(10):1077-1130.

Cutting GR. 2010. Modifier genes in Mendelian disorders: the example of cystic fibrosis. Ann NY Acad Sci 1214:57-69.

de Beer TA, Laskowski RA, Parks SL, Sipos B, Goldman N, Thornton JM. 2013. Amino acid changes in disease-associated variants differ radically from variants observed in the 1000 genomes project dataset. PLoS Comput Biol 9(12):e1003382.

Debuse JCW, Rayward-Smith VJ. 1997. Feature subset selection within a simulated annealing data ining algorithm. J Intell Inf Syst 9(1):57-81.

Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, Mooney TB, Callaway MB, Dooling D, Mardis ER, Wilson RK, Ding L. 2012. MuSiC: identifying mutational significance in cancer genomes. Genome Res 22(8):1589-1598.

Dehouck Y, Grosfils A, Folch B, Gilis D, Bogaerts P, Rooman M. 2009. Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. Bioinformatics 25(19):2537-2543.

Demurger F, Ichkou A, Mougou-Zerelli S, Le Merrer M, Goudefroye G, Delezoide A-L, Quelin C, Manouvrier S, Baujat G, Fradin M, Pasquier L, Megarbane A, et al. 2015. New insights into genotype-phenotype correlation for GLI3 mutations. Eur J Hum Genet 23(1):92-102.

Desmet FO, Hamroun D, Collod-Béroud G, Claustres M, Béroud C. 2010. Bioinformatics identification of splice site signals and prediction of mutation effects. In: Mohan RM, editor. Research Advances in Nucleic Acids Research. Kerala, India: Global Research Network. p 1-14.

Desmet FO, Hamroun D, Lalande M, Collod-Béroud G, Claustres M, Béroud C. 2009. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res 37(9):e67.

DiMauro S, Schon EA. 2001. Mitochondrial DNA mutations in human disease. Am J Med Genet 106(1):18-26.

Ding L, Wendl MC, McMichael JF, Raphael BJ. 2014. Expanding the computational toolbox for mining cancer genomes. Nat Rev Genet 15(8):556-570.

Dipple KM, McCabe ER. 2000. Phenotypes of patients with "simple" Mendelian disorders are complex traits: thresholds, modifiers, and systems dynamics. Am J Hum Genet 66(6):1729-1735.

Douville C, Masica DL, Stenson PD, Cooper DN, Gygax DM, Kim R, Ryan M, Karchin R. 2016. Assessing the pathogenicity of insertion and deletion variants with the Variant Effect Scoring Tool (VEST-indel). Hum Mutat 37(1):28-35.

Edlund K, Larsson O, Ameur A, Bunikis I, Gyllensten U, Leroy B, Sundstrom M, Micke P, Botling J, Soussi T. 2012. Data-driven unbiased curation of the TP53 tumor suppressor gene mutation database and validation by ultradeep sequencing of human tumors. Proc Natl Acad Sci U S A 109(24):9551-9556.

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

Fechter K, Porollo A. 2014. MutaCYP: Classification of missense mutations in human cytochromes P450. BMC Med Genomics 7:47.

Fernandez-Escamilla AM, Rousseau F, Schymkowitz J, Serrano L. 2004. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. Nat Biotechnol 22(10):1302-1306.

Ferrer-Costa C, Orozco M, de la Cruz X. 2002. Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. J Mol Biol 315(4):771-786.

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

Feucht M, Kluwe L, Mautner VF, Richard G. 2008. Correlation of nonsense and frameshift mutations with severity of retinal abnormalities in neurofibromatosis 2. Arch Ophthalmol 126(10):1376-1380.

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, et al. 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res 44(D1):D279-D285.

Fokkema IF, Taschner PE, Schaafsma GC, Celli J, Laros JF, den Dunnen JT. 2011. LOVD v.2.0: the next generation in gene variant databases. Hum Mutat 32(5):557-563.

Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A, Teague JW, Campbell PJ, et al. 2011. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res 39(Database issue):D945-950.

Fu R, Jinnah HA. 2012. Genotype-phenotype correlations in Lesch-Nyhan disease: moving beyond the gene. J Biol Chem 287(5):2997-3008.

Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Rieder MJ, Altshuler D, Shendure J, Nickerson DA, Bamshad MJ, et al. 2013. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493(7431):216-220.

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. 2004. A census of human cancer genes. Nat Rev Cancer 4(3):177-183.

Garraway Levi A, Lander Eric S. 2013. Lessons from the cancer genome. Cell 153(1):17-37.

Genome of the Netherlands Consortium. 2014. Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet 46(8):818-825.

Giardine B, Riemer C, Hefferon T, Thomas D, Hsu F, Zielenski J, Sang Y, Elnitski L, Cutting G, Trumbower H, Kern A, Kuhn R, et al. 2007. PhenCode: connecting ENCODE data with mutations and phenotype. Hum Mutat 28(6):554-562.

Gonzalez-Perez A, Deu-Pons J, Lopez-Bigas N. 2012. Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation. Genome Med 4(11):89.

Gonzalez-Perez A, Lopez-Bigas N. 2011. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet 88(4):440-449.

Gonzalez-Perez A, Lopez-Bigas N. 2012. Functional impact bias reveals cancer drivers. Nucleic Acids Res 40(21):e169.

Gonzalez-Perez A, Mustonen V, Reva B, Ritchie GR, Creixell P, Karchin R, Vazquez M, Fink JL, Kassahn KS, Pearson JV, Bader GD, Boutros PC, et al. 2013. Computational approaches to identify functional genetic variants in cancer genomes. Nat Methods 10(8):723-729.

González-Vioque E, Bornstein B, Gallardo ME, Fernandez-Moreno MA, Garesse R. 2014. The pathogenicity scoring system for mitochondrial tRNA mutations revisited. Mol Genet Genomic Med 2(2):107-114.

Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17(6):333-351.

Govindan R, Ding L, Griffith M, Subramanian J, Dees Nathan D, Kanchi Krishna L, Maher Christopher A, Fulton R, Fulton L, Wallis J, Chen K, Walker J, et al. 2012. Genomic landscape of non-small cell lung cancer in smokers and never-smokers. Cell 150(6):1121-1134.

Graham JW. 2009. Missing data analysis: making it work in the real world. Annu Rev Psychol 60:549-576.

Grimm DG, Azencott CA, Aicheler F, Gieraths U, MacArthur DG, Samocha KE, Cooper DN, Stenson PD, Daly MJ, Smoller JW, Duncan LE, Borgwardt KM. 2015. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. Hum Mutat 36(5):513-523.

Gryfe R, Gallinger S. 2001. Microsatellite instability, mismatch repair deficiency, and colorectal cancer. Surgery 130(1):17-20.

Guerois R, Nielsen JE, Serrano L. 2002. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol 320(2):369-387.

Guldberg P, Rey F, Zschocke J, Romano V, Francois B, Michiels L, Ullrich K, Hoffmann GF, Burgard P, Schmidt H, Meli C, Riva E, et al. 1998. A European multicenter study of phenylalanine hydroxylase deficiency: classification of 105 mutations and a general system for genotype-based prediction of metabolic phenotype. Am J Hum Genet 63(1):71-79.

Guyon I, Elisseeff A. 2003. An introduction to variable and feature selection. J Mach Learn Res 3:1157-1182.

Haber DA, Settleman J. 2007. Cancer: Drivers and passengers. Nature 446(7132):145-146.

Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. 2005. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 33(Database issue):D514-517.

Hanahan D, Weinberg RA. 2011. Hallmarks of cancer: the next generation. Cell 144(5):646-674.

Heinen CD. 2016. Mismatch repair defects and Lynch syndrome: The role of the basic scientist in the battle against cancer. DNA Repair (Amst) 38:127-134.

Herrero J, Muffato M, Beal K, Fitzgerald S, Gordon L, Pignatelli M, Vilella AJ, Searle SMJ, Amode R, Brent S, Spooner W, Kulesha E, et al. 2016. Ensembl comparative genomics resources. Database 2016.

Hindorff LA, Gillanders EM, Manolio TA. 2011. Genetic architecture of cancer and other complex diseases: lessons learned and future directions. Carcinogenesis 32(7):945-954.

Hodis E, Watson Ian R, Kryukov Gregory V, Arold Stefan T, Imielinski M, Theurillat J-P, Nickerson E, Auclair D, Li L, Place C, DiCara D, Ramos Alex H, et al. 2012. A landscape of driver mutations in melanoma. Cell 150(2):251-263.

Hon LS, Zhang Y, Kaminker JS, Zhang Z. 2009. Computational prediction of the functional effects of amino acid substitutions in signal peptides using a model-based approach. Hum Mutat 30(1):99-106.

Hood L, Rowen L. 2013. The Human Genome Project: big science transforms biology and medicine. Genome Med 5(9):1-8.

Hou JP, Ma J. 2014. DawnRank: discovering personalized driver genes in cancer. Genome Med 6(7):56.

Houdayer C, Caux-Moncoutier V, Krieger S, Barrois M, Bonnet F, Bourdon V, Bronner M, Buisson M, Coulet F, Gaildrat P, Lefol C, Leone M, et al. 2012. Guidelines for splicing analysis in molecular diagnosis derived from a set of 327 combined in silico/in vitro studies on BRCA1 and BRCA2 variants. Hum Mutat 33(8):1228-1238.

Houlleberghs H, Dekker M, Lantermans H, Kleinendorst R, Dubbink HJ, Hofstra RM, Verhoef S, Te Riele H. 2016. Oligonucleotide-directed mutagenesis screen to identify pathogenic Lynch syndrome-associated MSH2 DNA mismatch repair gene variants. Proc Natl Acad Sci U S A 113(15):4128-4133.

Hu J, Ng PC. 2012. Predicting the effects of frameshifting indels. Genome Biol 13(2):R9.

Hu J, Ng PC. 2013. SIFT Indel: predictions for the functional effects of amino acid insertions/deletions in proteins. PLoS One 8(10):e77940.

Hua X, Xu H, Yang Y, Zhu J, Liu P, Lu Y. 2013. DrGaP: a powerful tool for identifying driver genes and pathways in cancer sequencing studies. Am J Hum Genet 93(3):439-451.

Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabe RR, Bhan MK, Calvo F, Eerola I, Gerhard DS, Guttmacher A, Guyer M, et al. 2010. International network of cancer genome projects. Nature 464(7291):993-998.

Hunt RC, Simhadri VL, Iandoli M, Sauna ZE, Kimchi-Sarfaty C. 2014. Exposing synonymous mutations. Trends Genet 30(7):308-321.

Ingman M, Gyllensten U. 2006. mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences. Nucleic Acids Res 34(Database issue):D749-751.

Inza I, Calvo B, Armananzas R, Bengoetxea E, Larranaga P, Lozano JA. 2010. Machine learning: an indispensable tool in bioinformatics. Methods Mol Biol 593:25-48.

Jian X, Boerwinkle E, Liu X. 2014. In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Res 42(22):13534-13544.

Jiricny J. 2006. The multifaceted mismatch-repair system. Nat Rev Mol Cell Biol 7(5):335-346.

Jordan DM, Kiezun A, Baxter SM, Agarwala V, Green RC, Murray MF, Pugh T, Lebo MS, Rehm HL, Funke BH, Sunyaev SR. 2011. Development and validation of a computational method for assessment of missense variants in hypertrophic cardiomyopathy. Am J Hum Genet 88(2):183-192.

Juhling F, Morl M, Hartmann RK, Sprinzl M, Stadler PF, Putz J. 2009. tRNAdb 2009: compilation of tRNA sequences and tRNA genes. Nucleic Acids Res 37(Database issue):D159-162.

Kaminker JS, Zhang Y, Watanabe C, Zhang Z. 2007a. CanPredict: a computational tool for predicting cancer-associated missense mutations. Nucleic Acids Res 35(Web Server issue):W595-598.

Kaminker JS, Zhang Y, Waugh A, Haverty PM, Peters B, Sebisanovic D, Stinson J, Forrest WF, Bazan JF, Seshagiri S, Zhang Z. 2007b. Distinguishing cancer-associated missense mutations from common polymorphisms. Cancer Res 67(2):465-473.

Kang H. 2013. The prevention and handling of the missing data. Korean J Anesthesiol 64(5):402-406.

Karchin R. 2009. Next generation tools for the annotation of human SNPs. Brief Bioinform 10(1):35-52.

Karchin R, Kelly L, Sali A. 2005. Improving functional annotation of non-synonomous SNPs with information theory. Pac Symp Biocomput:397-408.

Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. 2008. AAindex: amino acid index database, progress report 2008. Nucleic Acids Res 36(Database issue):D202-205.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. Genome Res 12(6):996-1006.

Khan S, Vihinen M. 2010. Performance of protein stability predictors. Hum Mutat 31(6):675-684.

Khrapko K, Coller HA, André PC, Li XC, Hanekamp JS, Thilly WG. 1997. Mitochondrial mutational spectra in human cells and tissues. Proc Natl Acad Sci U S A 94(25):13798-13803.

Kinsella RJ, Kahari A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, Kersey P, Flicek P. 2011. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database (Oxford) 2011:bar030.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46(3):310-315.

Kirchner S, Ignatova Z. 2015. Emerging roles of tRNA in adaptive translation, signalling dynamics and disease. Nat Rev Genet 16(2):98-112.

Kohavi R, John GH. 1997. Wrappers for feature subset selection. Artif Intell 97(1-2):273-324.

Kondrashov FA. 2005. Prediction of pathogenic mutations in mitochondrially encoded human tRNAs. Hum Mol Genet 14(16):2415-2419.

Konig IR, Auerbach J, Gola D, Held E, Holzinger ER, Legault MA, Sun R, Tintle N, Yang HC. 2016. Machine learning and data mining in complex genomic data--a review on the lessons learned in Genetic Analysis Workshop 19. BMC Genet 17 Suppl 2:1.

Korthauer KD, Kendziorski C. 2015. MADGiC: a model-based approach for identifying driver genes in cancer. Bioinformatics 31(10):1526-1535.

Kotsiantis SB, Zaharakis ID, Pintelas PE. 2006. Machine learning: a review of classification and combining techniques. Artif Intell Rev 26(3):159-190.

Krawczak M, Thomas NS, Hundrieser B, Mort M, Wittig M, Hampe J, Cooper DN. 2007. Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing. Hum Mutat 28(2):150-158.

Krishnan VG, Westhead DR. 2003. A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. Bioinformatics 19(17):2199-2209.

Kucukkal TG, Yang Y, Chapman SC, Cao W, Alexov E. 2014. Computational and experimental approaches to reveal the effects of single nucleotide polymorphisms with respect to disease diagnostics. Int J Mol Sci 15(6):9670-9717.

Kumar MD, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, Uedaira H, Sarai A. 2006. ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. Nucleic Acids Res 34(Database issue):D204-206.

König E, Rainer J, Domingues FS. 2016. Computational assessment of feature combinations for pathogenic variant prediction. Mol Genet Genomic Med 4(4):431-446.

Laimer J, Hofer H, Fritz M, Wegenkittl S, Lackner P. 2015. MAESTRO - multi agent stability prediction upon point mutations. BMC Bioinformatics 16(1):116.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409(6822):860-921.

Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. 2014. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 42(Database issue):D980-985.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, et al. 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23(21):2947-2948.

Larranaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, Lozano JA, Armananzas R, Santafe G, Perez A, Robles V. 2006. Machine learning in bioinformatics. Brief Bioinform 7(1):86-112.

Laurila K, Vihinen M. 2011. PROlocalizer: integrated web service for protein subcellular localization prediction. Amino Acids 40(3):975-980.

Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, Kiezun A, Hammerman PS, et al. 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499(7457):214-218.

Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL, McCallion AS, Beer MA. 2015. A method to predict the impact of regulatory variants from DNA sequence. Nat Genet 47(8):955-961.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536(7616):285-291.

Lever J, Krzywinski M, Altman N. 2016. Points of significance: Classification evaluation. Nat Methods 13(8):603-604.

Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P. 2009. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics 25(21):2744-2750.

Li G-M. 2008. Mechanisms and functions of DNA mismatch repair. Cell Res 18(1):85-98.

Libbrecht MW, Noble WS. 2015. Machine learning applications in genetics and genomics. Nat Rev Genet 16(6):321-332.

Lindor NM, Guidugli L, Wang X, Vallee MP, Monteiro AN, Tavtigian S, Goldgar DE, Couch FJ. 2012. A review of a multifactorial probability-based model for classification of *BRCA1* and *BRCA2* variants of uncertain significance (VUS). Hum Mutat 33(1):8-21.

Liu M, Watson LT, Zhang L. 2014. Quantitative prediction of the effect of genetic variation using hidden Markov models. BMC Bioinformatics 15:5.

Liu X, Wu C, Li C, Boerwinkle E. 2016. dbNSFP v3.0: A one-stop database of functional predictions and annotations for human non-synonymous and splice site SNVs. Hum Mutat 37(3):235-241.

Lott MT, Leipzig JN, Derbeneva O, Xie HM, Chalkia D, Sarmady M, Procaccio V, Wallace DC. 2013. mtDNA variation and analysis using MITOMAP and MITOMASTER. Curr Protoc Bioinformatics 1(123):1.23.21-21.23.26.

Lynch HT, Snyder CL, Shaw TG, Heinen CD, Hitchins MP. 2015. Milestones of Lynch syndrome: 1895-2015. Nat Rev Cancer 15(3):181-194.

Ma S, Huang J. 2008. Penalized feature selection and classification in bioinformatics. Brief Bioinform 9(5):392-403.

Macintyre G, Bailey J, Haviv I, Kowalczyk A. 2010. is-rSNP: a novel technique for in silico regulatory SNP detection. Bioinformatics 26(18):i524-530.

Manke T, Heinig M, Vingron M. 2010. Quantifying the effect of sequence variation on regulatory interactions. Hum Mutat 31(4):477-483.

Mannini L, Cucco F, Quarantotti V, Krantz ID, Musio A. 2013. Mutation spectrum and genotype–phenotype correlation in Cornelia de Lange syndrome. Hum Mutat 34(12):1589-1596.

Mao Y, Chen H, Liang H, Meric-Bernstam F, Mills GB, Chen K. 2013. CanDrA: cancer-specific driver missense mutation annotation with optimized features. PLoS One 8(10):e77945.

Masica DL, Karchin R. 2016. Towards increasing the clinical relevance of in silico methods to predict pathogenic missense variants. PLoS Comput Biol 12(5):e1004725.

Masica DL, Sosnay PR, Cutting GR, Karchin R. 2012. Phenotype-optimized sequence ensembles substantially improve prediction of disease-causing mutation in cystic fibrosis. Hum Mutat 33(8):1267-1274.

Masica DL, Sosnay PR, Raraigh KS, Cutting GR, Karchin R. 2015. Missense variants in CFTR nucleotide-binding domains predict quantitative phenotypes associated with cystic fibrosis disease severity. Hum Mol Genet 24(7):1908-1917.

Massaad MJ, Ramesh N, Geha RS. 2013. Wiskott-Aldrich syndrome: a comprehensive review. Ann N Y Acad Sci 1285(1):26-43.

Masso M, Vaisman, II. 2010. AUTO-MUTE: web-based tools for predicting stability changes in proteins due to single amino acid replacements. Protein Eng Des Sel 23(8):683-687.

Matthijs G, Souche E, Alders M, Corveleyn A, Eck S, Feenstra I, Race V, Sistermans E, Sturm M, Weiss M, Yntema H, Bakker E, et al. 2015. Guidelines for diagnostic next-generation sequencing. Eur J Hum Genet 24(1):2-5.

McCormick EM, Hopkins E, Conway L, Catalano S, Hossain J, Sol-Church K, Stabley DL, Gripp KW. 2013. Assessing genotype-phenotype correlation in Costello syndrome using a severity score. Genet Med 15(7):554-557.

McFarland R, Elson JL, Taylor RW, Howell N, Turnbull DM. 2004. Assigning pathogenicity to mitochondrial tRNA mutations: when "definitely maybe" is not good enough. Trends Genet 20(12):591-596.

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. 2016. The Ensembl Variant Effect Predictor. Genome Biol 17(1):1-14.

McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26(16):2069-2070.

Miller MP, Kumar S. 2001. Understanding human disease mutations through the use of interspecific genetic variation. Hum Mol Genet 10(21):2319-2328.

Miosge LA, Field MA, Sontani Y, Cho V, Johnson S, Palkova A, Balakishnan B, Liang R, Zhang Y, Lyon S, Beutler B, Whittle B, et al. 2015. Comparison of predicted and actual consequences of missense mutations. Proc Natl Acad Sci U S A 112(37):E5189-5198.

Mitchell A, Chang H-Y, Daugherty L, Fraser M, Hunter S, Lopez R, McAnulla C, McMenamin C, Nuka G, Pesseat S, Sangrador-Vegas A, Scheremetjew M, et al. 2015. The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Res 43(D1):D213-D221.

Mort M, Sterne-Weiler T, Li B, Ball EV, Cooper DN, Radivojac P, Sanford JR, Mooney SD. 2014. MutPred Splice: machine learning-based prediction of exonic variants that disrupt splicing. Genome Biol 15(1):R19.

Nair PS, Vihinen M. 2013. VariBench: a benchmark database for variations. Hum Mutat 34(1):42-49.

Nalla VK, Rogan PK. 2005. Automated splicing mutation analysis by information theory. Hum Mutat 25(4):334-342.

Ng PC, Henikoff S. 2001. Predicting deleterious amino acid substitutions. Genome Res 11(5):863-874.

Ng PC, Henikoff S. 2006. Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet 7:61-80.

Niroula A, Vihinen M. 2016. Variation interpretation predictors: principles, types, performance and choice. Hum Mutat 37(6):579-597.

Olatubosun A, Väliaho J, Härkönen J, Thusberg J, Vihinen M. 2012. PON-P: integrated predictor for pathogenicity of missense variants. Hum Mutat 33(8):1166-1174.

Ortutay C, Vihinen M. 2009. Immunome knowledge base (IKB): an integrated service for immunome research. BMC Immunol 10:3.

Ortutay C, Väliaho J, Stenberg K, Vihinen M. 2005. KinMutBase: a registry of disease-causing mutations in protein kinase domains. Hum Mutat 25(5):435-442.

Parthiban V, Gromiha MM, Schomburg D. 2006. CUPSAT: prediction of protein stability upon point mutations. Nucleic Acids Res 34(Web Server issue):W239-242.

Peng Y, Alexov E. 2016. Investigating the linkage between disease-causing amino acid variants and their effect on protein stability and binding. Proteins 84(2):232-239.

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. 2004. UCSF Chimera - a visualization system for exploratory research and analysis. J Comput Chem 25(13):1605-1612.

Piirilä H, Väliaho J, Vihinen M. 2006. Immunodeficiency mutation databases (IDbases). Hum Mutat 27(12):1200-1208.

Pires DE, Ascher DB, Blundell TL. 2014. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Res 42(Web Server issue):W314-319.

Pon JR, Marra MA. 2015. Driver and passenger mutations in cancer. Annu Rev Pathol 10(1):25-50.

Potapov V, Cohen M, Schreiber G. 2009. Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. Protein Eng Des Sel 22(9):553-560.

Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM, Murphy MR, O'Leary NA, et al. 2014. RefSeq: an update on mammalian reference sequences. Nucleic Acids Res 42(Database issue):D756-763.

Putz J, Dupuis B, Sissler M, Florentz C. 2007. Mamit-tRNA, a database of mammalian mitochondrial tRNA primary and secondary structures. RNA 13(8):1184-1190.

Raphael BJ, Dobson JR, Oesper L, Vandin F. 2014. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. Genome Med 6(1):5.

Reblova K, Kulhanek P, Fajkusova L. 2015. Computational study of missense mutations in phenylalanine hydroxylase. J Mol Model 21(4):70.

Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL. 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. Genet Med 17(5):405-424.

Riera C, Padilla N, de la Cruz X. 2016. The complementarity between protein-specific and general pathogenicity predictors for amino acid substitutions. Hum Mutat 37(10):1013-1024.

Ritchie GR, Dunham I, Zeggini E, Flicek P. 2014. Functional annotation of noncoding sequence variants. Nat Methods 11(3):294-296.

Robins T, Carlsson J, Sunnerhagen M, Wedell A, Persson B. 2006. Molecular model of human CYP21 based on mammalian CYP2C5: structural features correlate with clinical severity of mutations causing congenital adrenal hyperplasia. Mol Endocrinol 20(11):2946-2964.

Saeys Y, Inza I, Larranaga P. 2007. A review of feature selection techniques in bioinformatics. Bioinformatics 23(19):2507-2517.

Saito R, Smoot ME, Ono K, Ruscheinski J, Wang PL, Lotia S, Pico AR, Bader GD, Ideker T. 2012. A travel guide to Cytoscape plugins. Nat Methods 9(11):1069-1076.

Samarghitean C, Väliaho J, Vihinen M. 2007. IDR knowledge base for primary immunodeficiencies. Immunome Res 3:6.

Sauna ZE, Kimchi-Sarfaty C. 2011. Understanding the contribution of synonymous mutations to human disease. Nat Rev Genet 12(10):683-691.

Schaafsma GC, Vihinen M. 2015. VariSNP, a benchmark database for variations from dbSNP. Hum Mutat 36(2):161-166.

Schwarz JM, Cooper DN, Schuelke M, Seelow D. 2014. MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods 11(4):361-362.

Scriver CR, Waters PJ. 1999. Monogenic traits are not simple: lessons from phenylketonuria. Trends Genet 15(7):267-272.

Sengupta M, Sarkar D, Ganguly K, Sengupta D, Bhaskar S, Ray K. 2015. In silico analyses of missense mutations in coagulation factor VIII: identification of severity determinants of haemophilia A. Haemophilia 21(5):662-669.

Shabani M, Borry P. 2015. Challenges of web-based personal genomic data sharing. Life Sci Soc Policy 11:3.

Shen B, Vihinen M. 2004. Conservation and covariance in PH domain sequences: physicochemical profile and information theoretical analysis of XLA-causing mutations in the Btk PH domain. Protein Eng Des Sel 17(3):267-276.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29(1):308-311.

Shihab HA, Gough J, Cooper DN, Day IN, Gaunt TR. 2013. Predicting the functional consequences of cancer-associated amino acid substitutions. Bioinformatics 29(12):1504-1510.

Sijmons RH, Hofstra RM. 2016. Clinical aspects of hereditary DNA mismatch repair gene mutations. DNA Repair (Amst) 38:155-162.

Simonetti FL, Tornador C, Nabau-Moreto N, Molina-Vila MA, Marino-Buslje C. 2014. Kin-Driver: a database of driver mutations in protein kinases. Database (Oxford) 2014:bau104.

Singh A, Nowak RD, Zhu X. Unlabeled data: Now it helps, now it doesn't; 2008; Vancouver, British Columbia, Canada.

Sormanni P, Aprile FA, Vendruscolo M. 2015. The CamSol method of rational design of protein mutants with enhanced solubility. J Mol Biol 427(2):478-490.

Stalker J, Gibbins B, Meidl P, Smith J, Spooner W, Hotz HR, Cox AV. 2004. The Ensembl Web site: mechanics of a genome browser. Genome Res 14(5):951-955.

Statnikov A, Wang L, Aliferis CF. 2008. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC Bioinformatics 9(1):1-10.

Stefl S, Nishi H, Petukh M, Panchenko AR, Alexov E. 2013. Molecular mechanisms of disease-causing missense mutations. J Mol Biol 425(21):3919-3936.

Stenberg KA, Riikonen PT, Vihinen M. 2000. KinMutBase, a database of human disease-causing protein kinase mutations. Nucleic Acids Res 28(1):369-371.

Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN. 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):1-9.

Stern A, Doron-Faigenboim A, Erez E, Martz E, Bacharach E, Pupko T. 2007. Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach. Nucleic Acids Res 35(Web Server issue):W506-511.

Steward RE, MacArthur MW, Laskowski RA, Thornton JM. 2003. Molecular basis of inherited diseases: a structural perspective. Trends Genet 19(9):505-513.

Stratton MR, Campbell PJ, Futreal PA. 2009. The cancer genome. Nature 458(7239):719-724.

Sun J, Zhao Z. 2010. A comparative study of cancer proteins in the human protein-protein interaction network. BMC Genomics 11 Suppl 3:S5.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res 34(Web Server issue):W609-612.

Suzuki T, Nagao A, Suzuki T. 2011. Human mitochondrial tRNAs: biogenesis, function, structural aspects, and diseases. Annu Rev Genet 45:299-329.

Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. 2013. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics 29(18):2238-2244.

Tanaka M, Takeyasu T, Fuku N, Li-Jun G, Kurata M. 2004. Mitochondrial genome single nucleotide polymorphisms and their phenotypes in the Japanese. Ann N Y Acad Sci 1011:7-20.

Tang H, Thomas PD. 2016. Tools for predicting the functional impact of nonsynonymous genetic variation. Genetics 203(2):635.

Teo YY, Sim X, Ong RT, Tan AK, Chen J, Tantoso E, Small KS, Ku CS, Lee EJ, Seielstad M, Chia KS. 2009. Singapore Genome Variation Project: a haplotype map of three Southeast Asian populations. Genome Res 19(11):2154-2162.

The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491(7422):56-65.

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. Nature 526(7571):68-74.

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

The UniProt Consortium. 2015. UniProt: a hub for protein information. Nucleic Acids Res 43(D1):D204-D212.

Thompson BA, Goldgar DE, Paterson C, Clendenning M, Walters R, Arnold S, Parsons MT, Michael DW, Gallinger S, Haile RW, Hopper JL, Jenkins MA, et al. 2013a. A multifactorial likelihood model for MMR gene variant classification incorporating probabilities based on sequence bioinformatics and tumor characteristics: a report from the Colon Cancer Family Registry. Hum Mutat 34(1):200-209.

Thompson BA, Greenblatt MS, Vallee MP, Herkert JC, Tessereau C, Young EL, Adzhubey IA, Li B, Bell R, Feng B, Mooney SD, Radivojac P, et al. 2013b. Calibration of multiple in silico tools for predicting pathogenicity of mismatch repair gene missense substitutions. Hum Mutat 34(1):255-265.

Thompson BA, Spurdle AB, Plazzer JP, Greenblatt MS, Akagi K, Al-Mulla F, Bapat B, Bernstein I, Capella G, den Dunnen JT, du Sart D, Fabre A, et al. 2014. Application of a 5-tiered scheme for standardized classification of 2,360 unique mismatch repair gene variants in the InSiGHT locus-specific database. Nat Genet 46(2):107-115.

Thusberg J, Olatubosun A, Vihinen M. 2011. Performance of mutation pathogenicity prediction methods on missense variants. Hum Mutat 32(4):358-368.

Thusberg J, Vihinen M. 2009. Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. Hum Mutat 30(5):703-714.

Tian R, Basu MK, Capriotti E. 2015. Computational methods and resources for the interpretation of genomic variants in cancer. BMC Genomics 16 Suppl 8:S7.

Tian Y, Deutsch C, Krishnamoorthy B. 2010. Scoring function to predict solubility mutagenesis. Algorithms Mol Biol 5:33.

Tomczak K, Czerwińska P, Wiznerowicz M. 2015. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol 19(1A):A68-A77.

Torkamani A, Schork NJ. 2007. Accurate prediction of deleterious protein kinase polymorphisms. Bioinformatics 23(21):2918-2925.

Tuppen HAL, Blakely EL, Turnbull DM, Taylor RW. 2010. Mitochondrial DNA mutations and human disease. Biochim Biophys Acta 1797(2):113-128.

Wagih O, Reimand J, Bader GD. 2015. MIMP: predicting the impact of mutations on kinase-substrate phosphorylation. Nat Methods 12(6):531-533.

Walsh I, Seno F, Tosatto SC, Trovato A. 2014. PASTA 2.0: an improved server for protein aggregation prediction. Nucleic Acids Res 42(Web Server issue):W301-307.

Walsh R, Thomson KL, Ware JS, Funke BH, Woodley J, McGuire KJ, Mazzarotto F, Blair E, Seller A, Taylor JC, Minikel EV, Exome Aggregation C, et al. 2016. Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. Genet Med:[In Press].

Walter K, Min JL, Huang J, Crooks L, Memari Y, McCarthy S, Perry JR, Xu C, Futema M, Lawson D, Iotchkova V, Schiffels S, et al. 2015. The UK10K project identifies rare variants in health and disease. Nature 526(7571):82-90.

van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. 2014. Ten years of next-generation sequencing technology. Trends Genet 30(9):418-426.

Vandin F, Upfal E, Raphael BJ. 2012. De novo discovery of mutated driver pathways in cancer. Genome Res 22(2):375-385.

Wang Z, Moult J. 2001. SNPs, protein structure, and disease. Hum Mutat 17(4):263-270.

Watson IR, Takahashi K, Futreal PA, Chin L. 2013. Emerging patterns of somatic mutations in cancer. Nat Rev Genet 14(10):703-718.

Vazquez M, Pons T, Brunak S, Valencia A, Izarzugaza JM. 2016. wKinMut-2: Identification and interpretation of pathogenic variants in human protein kinases. Hum Mutat 37(1):36-42.

Wei Q, Dunbrack RL, Jr. 2013. The role of balanced training and testing data sets for binary classifiers in bioinformatics. PLoS One 8(7):e67863.

Weinreb NJ, Cappellini MD, Cox TM, Giannini EH, Grabowski GA, Hwu WL, Mankin H, Martins AM, Sawyer C, vom Dahl S, Yeh MS, Zimran A. 2010. A validated disease severity scoring system for adults with type 1 Gaucher disease. Genet Med 12(1):44-51.

Verbeke LP, Van den Eynden J, Fierro AC, Demeester P, Fostier J, Marchal K. 2015. Pathway relevance ranking for tumor samples through network-based data integration. PLoS One 10(7):e0133503.

Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Tatusova TA, Wagner L. 2003. Database resources of the National Center for Biotechnology. Nucleic Acids Res 31(1):28-33.

Vihinen M. 2012. How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. BMC Genomics 13 Suppl 4:S2.

Vihinen M. 2013. Guidelines for reporting and using prediction tools for genetic variation analysis. Hum Mutat 34(2):275-282.

Vihinen M. 2015. Types and effects of protein variations. Hum Genet 134(4):405-421.

Vincent A, Robson AG, Neveu MM, Wright GA, Moore AT, Webster AR, Holder GE. 2013. A phenotype-genotype correlation study of X-linked retinoschisis. Ophthalmology 120(7):1454-1464.

Vitkup D, Sander C, Church GM. 2003. The amino-acid mutational spectrum of human genetic disease. Genome Biol 4(11):R72.

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Jr., Kinzler KW. 2013. Cancer genome landscapes. Science 339(6127):1546-1558.

Wong WC, Kim D, Carter H, Diekhans M, Ryan MC, Karchin R. 2011. CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. Bioinformatics 27(15):2147-2148.

Woolfe A, Mullikin JC, Elnitski L. 2010. Genomic features defining exonic variants that modulate splicing. Genome Biol 11(2):R20.

Wu B, Abbott T, Fishman D, McMurray W, Mor G, Stone K, Ward D, Williams K, Zhao H. 2003. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. Bioinformatics 19(13):1636-1643.

Wu G, Feng X, Stein L. 2010. A human functional protein interaction network and its application to cancer data analysis. Genome Biol 11(5):R53.

Wu H, Gao L, Li F, Song F, Yang X, Kasabov N. 2015. Identifying overlapping mutated driver pathways by constructing gene networks in cancer. BMC Bioinformatics 16 Suppl 5:S3.

Vuong H, Che A, Ravichandran S, Luke BT, Collins JR, Mudunuri US. 2015. AVIA v2.0: annotation, visualization and impact analysis of genomic variants and genes. Bioinformatics 31(16):2748-2750.

Väliaho J, Faisal I, Ortutay C, Smith CI, Vihinen M. 2015. Characterization of all possible single-nucleotide change caused amino acid substitutions in the kinase domain of bruton tyrosine kinase. Hum Mutat 36(6):638-647.

Yang H, Wang K. 2015. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. Nat Protoc 10(10):1556-1566.

Yang J, Honavar VG. 1998. Feature subset selection using a genetic algorithm. IEEE Intell Syst 13(2):44-49.

Yang Y, Chen B, Tan G, Vihinen M, Shen B. 2013. Structure-based prediction of the effects of a missense variant on protein stability. Amino Acids 44(3):847-855.

Yang Y, Niroula A, Shen B, Vihinen M. 2016. PON-Sol: prediction of effects of amino acid substitutions on protein solubility. Bioinformatics 32(13):2032-2034.

Yarham JW, Al-Dosary M, Blakely EL, Alston CL, Taylor RW, Elson JL, McFarland R. 2011. A comparative analysis approach to determining the pathogenicity of mitochondrial tRNA mutations. Hum Mutat 32(11):1319-1325.

Yarham JW, Elson JL, Blakely EL, McFarland R, Taylor RW. 2010. Mitochondrial tRNA mutations and disease. Wiley Interdiscip Rev RNA 1(2):304-324.

Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L, Girón CG, Gordon L, et al. 2016. Ensembl 2016. Nucleic Acids Res 44(D1):D710-D716.

Yates CM, Filippis I, Kelley LA, Sternberg MJ. 2014. SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features. J Mol Biol 426(14):2692-2701.

Zambrano R, Jamroz M, Szczasiuk A, Pujols J, Kmiecik S, Ventura S. 2015. AGGRESCAN3D (A3D): server for prediction of aggregation properties of protein structures. Nucleic Acids Res 43(W1):W306-313.

Zhang Z, Miteva MA, Wang L, Alexov E. 2012. Analyzing effects of naturally occurring missense mutations. Comput Math Methods Med 2012:805827.

Zhao H, Yang Y, Lin H, Zhang X, Mort M, Cooper DN, Liu Y, Zhou Y. 2013. DDIG-in: discriminating between disease-associated and neutral non-frameshifting micro-indels. Genome Biol 14(3):R23.

Zhou J, Troyanskaya OG. 2015. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods 12(10):931-934.

Zia A, Moses AM. 2011. Ranking insertion, deletion and nonsense mutations based on their effect on genetic information. BMC Bioinformatics 12:299.