REVIEW

Human Mutation

OFFICIAL JOURNAL
HGVS
HUMAN GENOME VARIATION SOCIETY
www.hgvs.org

Variation Interpretation Predictors: Principles, Types, Performance, and Choice

Abhishek Niroula and Mauno Vihinen*

Department of Experimental Medical Science, Lund University, BMC B13, Lund SE-22184, Sweden

For the 25th Anniversary Commemorative Issue

Received 16 November 2015; accepted revised manuscript 7 March 2016.
Published online 14 March 2016 in Wiley Online Library (www.wiley.com/humanmutation). DOI: 10.1002/humu.22987

ABSTRACT: Next-generation sequencing methods have revolutionized the speed of generating variation information. Sequence data have a plethora of applications and will increasingly be used for disease diagnosis. Interpretation of the identified variants is usually not possible with experimental methods. This has caused a bottleneck that many computational methods aim at addressing. Fast and efficient methods for explaining the significance and mechanisms of detected variants are required for efficient precision/personalized medicine. Computational prediction methods have been developed in three areas to address the issue. There are generic tolerance (pathogenicity) predictors for filtering harmful variants. Gene/protein/disease-specific tools are available for some applications. Mechanism and effect-specific computer programs aim at explaining the consequences of variations. Here, we discuss the different types of predictors and their applications. We review available variation databases and prediction methods useful for variation interpretation. We discuss how the performance of methods is assessed and summarize existing assessment studies. A brief introduction is provided to the principles of the methods developed for variation interpretation as well as guidelines for how to choose the optimal tools and where the field is heading in the future.

Hum Mutat 37:579–597, 2016.

KEY WORDS: variation interpretation; variation prediction; variation effect; mutation effect prediction; computational tools; prediction methods

Introduction

Next-generation sequencing (NGS) methods produce a wealth of detailed information, including large variation datasets. Every human genome contains about 3 million single nucleotide substitutions in comparison with a reference genome, but only a few of those are disease-related. In addition, there are other more complex variations, fewer in number but challenging to interpret. The 1000 Genomes project estimated that each genome codes for about 11,000 amino acid substitutions, approximately 12,500 synonymous substitutions, and hundreds of insertions and deletions [Abecasis et al., 2010].

DNA sequencing has been automated very effectively and thus is fast and inexpensive, especially in the case of resequencing the human genome. The $1,000 genome sequence is already a reality. Sequencing is widely used for diagnostic purposes, but is often done for individual genes. As the price of exome and complete genome sequencing is falling below the cost of sequencing one or a few specific genes, there is increased interest in using genome sequencing in health care.

The fall in sequencing costs has exposed the major deficits in our understanding of the pathogenic significance of most variants. The situation reduces clinical utility and greatly affects the potential for disease treatment and prevention. The problematic lack of knowledge arises because large numbers of the identified variations are novel, or without proper annotation, and knowledge about disease association is missing. Data analysis and variation interpretation are the most time-consuming steps in sequencing projects. Whole-genome or whole-exome sequencing will provide valuable new information about an individual if this information could be utilized and interpreted in a reliable and meaningful fashion. The genome-wide interpretation of variants is only feasible with computational prediction methods. The bottleneck in the utilization of sequence information has shifted from obtaining the sequences to understanding and interpreting them.

Several groups are working on developing guidelines for how to utilize sequencing information for clinical purposes. Recommendations by both the European Society of Human Genetics (ESHG) and the American College of Medical Genetics (ACMG) include the use of prediction methods [Matthijs et al., 2015; Richards et al., 2015]. For certain diseases, sequencing is the only method for differential diagnosis. Diagnosis and prognosis in many diseases is highly dependent on sequence information. Numerous computational tools have been developed for the analysis, prioritization, and interpretation of variations and their effects. End users of these tools are often puzzled by not knowing which method to choose, which ones are the most reliable, and how to choose the most suitable applications. This paper aims at describing what types of predictions are currently possible, how the tools can be compared, and their performance be assessed, as well as some principles that are necessary to understand and to be able to evaluate methods. Finally, tips are provided for how to choose the most suitable computational methods and what we can expect to have in the future. Although computational approaches are very helpful and vital for variation interpretation, one has to bear in mind that these methods alone cannot define whether a variant is pathogenic or not, and other types of information (such as patient clinical information, family history, etc.) are also needed

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Mauno Vihinen, Protein Structure and Bioinformatics Group, Department of Experimental Medical Science, BMC B13, Lund University, SE-22, 184 Lund, Sweden. E-mail: mauno.vihinen@med.lu.se

Contract grant sponsors: Lund University; Barncancerfonden; Vetenskapsrådet.

for making such decisions. The type(s) of additional information varies for different diseases.

Several organizations are working toward providing and promoting reliable variation information. The Human Genome Variation Society (HGVS, http://www.hgvs.org/) has been the scientific society for the variation database and interpretation community, variation nomenclature being one of their major achievements [den Dunnen and Antonarakis, 2001]. The Human Variome Project (HVP, http://www.humanvariomeproject.org/) "is working to ensure that all information on genetic variation and its effect on human health can be collected, curated, interpreted, and shared freely and openly." Recently, the Global Alliance for Genomics and Health (GA4GH, http://genomicsandhealth.org/) was founded to enable sharing of genomic and clinical data.

Variations can have their effects on many levels. The most studied are amino acid substitutions; however, also other types of variations, such as insertions and deletions, are important. Amino acid substitutions can have numerous effects and the mechanisms behind them are diverse [Steward et al., 2003; Thusberg and Vihinen, 2009; Vihinen, 2015c]. They may disrupt or block critical sites for protein function, such as catalytic residues or ligand-binding pockets. They can lead to alterations in the protein structural properties, causing, for example, abnormal folding, structural instability, or aggregation of the protein. Even small changes in the size or chemical or physical nature of an amino acid side chain can be critical. Amino acid substitutions may affect protein posttranslational modifications by generating or deleting such sites, for example, for phosphorylation or glycosylation, or protease cleavage, or altering protein localization targeting signals. DNA and RNA variations often affect important functional sites such as binding sites, for example, for transcription factors or microRNAs or in the case of RNA, the structure of the molecule. For detailed systematic organization of the variation types and effects, see the Variation Ontology (VariO) [Vihinen, 2014b, 2015c].

Most studies are concentrated on loss-of-function variations; however, there are also gain-of-function alterations due to irregular or altered binding of ligands or loss of specificity. Many proteins are robust and can tolerate alterations and insertions to numerous sites without any or just a minor effect on protein function [Poussu et al., 2004]; on the other hand, there are sites that do not tolerate any variations. Recently, some computational studies have been published in which all possible amino acid substitutions or those caused by single-nucleotide changes were analyzed. These include kinase domain variants in Bruton tyrosine kinase (BTK) [Väliaho et al., 2015], variants in four mismatch repair (MMR) proteins [Niroula and Vihinen, 2015a], and predicted protein solubility affecting variations in human interleukin-1 β [Yang et al., 2016]. In addition, similar studies have been performed for mitochondrial tRNA molecules [Kondrashov, 2005; Niroula and Vihinen, 2016]. These studies predicted the effects of all possible amino acid/nucleotide substitutions in the proteins/RNAs and provided important information about the types of variations and their distribution along the molecule sequence and structure highlighting both structural and functional aspects. The frequency of predicted harmful variants varies greatly. Depending on the location and the type of variations, the outcomes are different. The BTK kinase domain has the highest published ratio of harmful variants, with 67% of the studied cases [Väliaho et al., 2015]. In MMR proteins, the proportions of harmful amino acid substitutions varied from 14.6% (in PMS2) to 40.4% (in MSH2) [Niroula and Vihinen, 2015a]. There are only a few experimental studies for massive amounts of amino acid substitutions in single proteins. The ratios of harmful variations vary widely in them, see Väliaho et al. [2015].

Several reviews have been published about prediction methods and their use [Thusberg and Vihinen, 2009; Cline and Karchin, 2011; Capriotti et al., 2012; Peterson et al., 2013], but none of them have covered the full scope of the discussion as done here.

Variation Databases

When investigating the effects of variations, one has to find out whether the identified alteration has been previously found in patients or in healthy individuals. If a variant is frequent in healthy individuals in populations, possibly in several ethnic groups, it is likely not harmful, although there are some exceptions. Central variation databases, such as the Human Gene Mutation Database (HGMD) [Stenson et al., 2014], Online Mendelian Inheritance in Man (OMIM) [Hamosh et al., 2005], and ClinVar [Landrum et al., 2014], include variants for many human genes (Table 1). The UniProtKB/SwissProt [Uniprot Consortium, 2015] database contains manually annotated protein entries with some variants [Yip et al., 2008] available at the SwissVar service (http://swissvar.expasy.org/). For a comprehensive review of the current state and future challenges of genotype–phenotype databases, see Brookes and Robinson [2015].

Locus-specific variation databases (LSDBs) list variants in specific genes/diseases and are typically manually annotated. Many LSDBs contain also other information, sometimes even very detailed clinical data. LSDBs exist at the Leiden Open Variation Database (LOVD) system for all human genes [Fokkema et al., 2011]. Many of them are still empty and without curators. One of the problems is that there can be more than one database for some genes and diseases and they can contain overlapping data [Mitropoulou et al., 2010]. LSDBs are usually the primary and most trusted variation information sources as they are curated and maintained by experts in the genes and diseases. The currently existing LSDBs are listed at The Human Genome Variation Society (HGVS) Website (http://www.hgvs.org/locus-specific-mutation-databases), the LOVD site (http://grenada.lumc.nl/LSDB_list/lsdbs), the GEN2PHEN server (http://www.gen2phen.org/data/lsdbs) [Thorisson et al., 2009], and at the Web Analysis of the Variome (http://bioinformatics.ua.pt/WAVe/) [Lopes et al., 2011]. Other large collections of LSDBs include IDbases [Piirilä et al., 2006], and those maintained at Universal Mutation Databases (UMD) platform [Beroud et al., 2000].

Several guidelines and recommendations have been published for LSDBs including how to establish a database [Vihinen et al., 2012], their curation [Celli et al., 2012], overall contents [Kohonen-Corish et al., 2010], ethics [Cotton et al., 2005; Povey et al., 2010], data collection [Cotton et al., 2007; Cotton et al., 2009], somatic variations [Olivier et al., 2009], interpretation and reporting of variants [Plon et al., 2008; Richards et al., 2015], data sharing [den Dunnen et al., 2009], and nomenclature [den Dunnen and Antonarakis, 2001; Taschner and den Dunnen, 2011]. Beside variant descriptions, LSDBs can also contain clinically relevant information, including where and by whom the variant was found (contact information, publication details), how it was detected (screening methodology), and whether it influences the function of the gene or its product. Some LSDBs contain details of the phenotype of the patient.

The dbSNP database [Sherry et al., 2001] is a very large database for short variants and contains data from The 1000 Genomes Project (http://www.1000genomes.org/). Note that dbSNP contains both disease-causing and benign variants. The NHLBI Exome Sequencing Project (ESP) Exome Variant Server (EVS, http://evs.gs.washington.edu/EVS/) provides exome data from 200,000 individuals from African-American

Table 1. Representative list of variation databases

Database	URL	Reference
General variation databases		
ClinVar	http://www.ncbi.nlm.nih.gov/clinvar/	Landrum et al. [2014]
Database of immunodeficiency-causing variations (IDbases)	http://structure.bmc.lu.se/idbase/	Piirilä et al. [2006]
Database of short genetic variations (dbSNP)	http://www.ncbi.nlm.nih.gov/SNP/	Sherry et al. [2001]
Ensembl Variation Database	http://www.ensembl.org/info/genome/variation/index.html	
Exome Aggregation Consortium (ExAC)	http://exac.broadinstitute.org/	
Exome Variant Server (EVS)	http://evs.gs.washington.edu/EVS/	Fu et al. [2013]
Human Gene Mutation Database (HGMD)	http://www.hgmd.cf.ac.uk/ac/index.php	Stenson et al. [2014]
Leiden Open Variation Databases (LOVD)	http://www.lovd.nl/3.0/home	Fokkema et al. [2011]
MITOMAP	http://www.mitomap.org/MITOMAP	Lott et al. [2013]
Online Mendelian Inheritance in Man (OMIM)	http://www.ncbi.nlm.nih.gov/omim	Hamosh et al. [2005]
SwissVar	http://swissvar.expasy.org/	Yip et al. [2008]
Universal Mutation Database (UMD)	http://www.umd.be/	Beroud et al. [2000]
Databases of genomic structural variations		
dbVar	http://www.ncbi.nlm.nih.gov/dbvar/	Lappalainen et al. [2013]
Database of Genomic Variants (DGV)	http://dgv.tcag.ca/dgv/app/home	MacDonald et al. [2014]
Database of Genomic Variants archive (DGVa)	http://www.ebi.ac.uk/dgva	Lappalainen et al. [2013]
Human Polymorphic Inversion Database (InvFEST)	http://invfestdb.uab.cat/	Martinez-Fundichely et al. [2014]
Database of cancer variations		
Catalogue of Somatic Mutations in Cancer (COSMIC)	http://cancer.sanger.ac.uk/cosmic	Forbes et al. [2011]
Database or Curated Mutations (DoCM)	http://docm.genome.wustl.edu/	
Kin-Driver	http://kin-driver.leloir.org.ar/	Simonetti et al. [2014]
Databases of benchmark datasets		
VariBench	http://structure.bmc.lu.se/VariBench/	Nair and Vihinen [2013]
VariSNP	http://structure.bmc.lu.se/VariSNP/	Schaafsma and Vihinen [2015]
Specific databases		
Frequency INherited Disorders database (FINDbase)	http://www.findbase.org/	Papadopoulos et al. [2014]
Thermodynamic Database for Proteins and Mutants (ProTherm)	http://www.abren.net/protherm/	Kumar et al. [2006]

and European-American populations [Fu et al., 2013]. Individuals with specific traits related to blood, heart diseases, and lung diseases as well as controls are included in the EVS project. The Exome Aggregation Consortium (ExAC, http://exac.broadinstitute.org) has data from 60,706 unrelated individuals and the Allele Frequency Community (AFC, http://www.allelefrequencycommunity.org/) currently contains data for about 100,000 exomes/genomes. Some databases annotate human variation data with phenotype variations and protein structural and functional information such as KMDB/MutationView (http://mutview.dmb.med.keio.ac.jp/MutationView/jsp/index.jsp). 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 provide information about genes, their products, and sequence variants. The Encyclopedia of DNA Elements (ENCODE) project [Sloan et al., 2016] aims at identifying and mapping functional and regulatory elements in the human genome also outside protein coding regions. PhenCode [Giardine et al., 2007] connects human phenotype and clinical data in LSDBs with genomic data from the ENCODE project and other resources in the UCSC Genome Browser.

Additional types of variation databases include national and ethnic databases [Patrinos, 2006] such as ETHNOS [van Baal et al., 2010], variation frequency databases including FINDbase [Papadopoulos et al., 2014], chromosomal structural variation databases such as The Database of Genomic Variants (DGV) [MacDonald et al., 2014], Database of Genomic Variants archive (DGVa, http://www.ebi.ac.uk/dgva), dbVar (http://www.ncbi.nlm.nih.gov/dbvar) [Lappalainen et al., 2013], InvFEST database of inversions [Martinez-Fundichely et al., 2014], Mitelman Database of chromosome aberrations and gene fusions (http://cgap.nci.nih.gov/Chromosomes/Mitelman), and European Cytogeneticists Association Register of Unbalanced Chromosome Aberrations (ECARUCA) database containing cytogenetic and clinical information of rare chromosomal disorders including microdeletions and microduplications [Feenstra et al., 2006] (Table 1). Databases dedicated to certain types of variations or to an effect or mechanism are available, for example, in the ProTherm database for protein stability affecting variations [Kumar et al., 2006], and The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/) and the Catalogue of Somatic Mutations in Cancer (COSMIC, http://cancer.sanger.ac.uk/cosmic) [Forbes et al., 2011] databases for genetic variations in cancer (Table 1).

From the Cafe Variome [Lancaster et al., 2015], one can find whether a variant has been previously published, but without revealing further information. For more information about the variant, one needs to contact the relevant data owner/provider. One additional type of databases is highly relevant for variation effect predictions, namely, benchmark databases. VariBench [Nair and Vihinen, 2013] and VariSNP [Schaafsma and Vihinen, 2015] collect benchmark variation datasets from different sources and distribute datasets that provide gold standards for method developers and assessors.

It is recommended to use systematics in variation databases whenever possible, including those for phenotype such as the Human Phenotype Ontology (HPO) [Robinson et al., 2008] and the Variation Ontology (VariO) [Vihinen, 2014b] for a systematic description of effects, mechanisms, and consequences of variants. The usage of these systematic approaches facilitates efficient searches, data integration, as well as the development of computational systems for data analysis and interpretation. Several de facto standards are useful for variation databases. These include systematic gene names available from the HUGO Gene Nomenclature Committee (HGNC) [Gray et al., 2015]. Reference sequences have to be specified including version numbers, unless Locus Reference Genomic (LRG) [Dalgleish et al., 2010] entries are used. LRGs are recommended for

human sequences as they are stable and allow unambiguous mapping of positions. For naming variations, the HGVS nomenclature [den Dunnen and Antonarakis, 2001] should be used. It is widely used in the literature and also by several computer programs including those interpreting variations. The Mutalyzer tool can generate systematic names automatically and performs a number of consistency checks [Wildeman et al., 2008].

The HVP has released quality assessment criteria to evaluate the quality of genetic variation databases and to stimulate database curators to make improvements where they are needed (http://www.humanvariomeproject.org/finish/19/255.html) [Vihinen et al., 2016]. Once the quality scheme is implemented, the quality information will allow users to check how reliable, up-to-date, and user-friendly a certain database is.

Principles

It is important for users of any method, whether utilized in laboratories or run with computers, to know how they work. Only by doing so can one evaluate the significance of the observed findings. Prediction methods can be rather complex and complicated, especially when considering the details; however, the general principles of these methods can be understood by any scientist. Bioinformaticians seek to find the best solutions for their methods. This means that novel approaches are implemented to improve the performance of the methods and new methods are released frequently.

Machine Learning

A majority of the recent methods in the variation interpretation field are based on machine learning (ML). ML is a form of artificial intelligence where a computer system can learn from given data (Fig. 1). Among the different types of ML techniques, supervised ML classification is widely used for variation interpretation. In supervised ML classification, a computer model is trained to distinguish between two or more classes using given examples. The most common ML algorithms include neural networks, support vector machines, and random forests. The ML methods are highly dependent on the quality of data used for their training. Typically, the methods have been binary, that is, providing two classes of predictions. Some of the methods have more than two classes as variation pathogenicity is far from the binary, black and white, pathogenic or not dichotomy. Binary classifiers can still be useful for many applications. Some methods do not categorize the variations into classes but provide continuous scores.

ML methods are widely used to tackle complex phenomena, which would be otherwise difficult to handle. For a more detailed discussion, see Vihinen [2012], for example. Variations can have large numbers of effects that would be difficult to take into account in other types of predictors. An important aspect of ML method development is the choice of the approach. Among the numerous approaches, there is not one superior architecture. The quality of a predictor depends on how the training is done, which features are used to explain the phenomenon, and how the method is optimized.

Features are used to describe the characteristics of the investigated phenomenon. If several features are available, it is important to choose those that best capture the phenomenon (Fig. 1). This is partly due to the curse of dimensionality, which means that much more data are needed when the number of features increases. In fact, the volume of the feature space grows exponentially with the dimensionality such that the data become sparse and insufficient to adequately describe the feature space. Another problem is overfitting, which means that the algorithm describes noise or random features in the training dataset instead of the real phenomenon due to sparse data, the complexity of the model, an excessive learning procedure or any combination of them. Overfitting leads to a decreased performance on real cases.

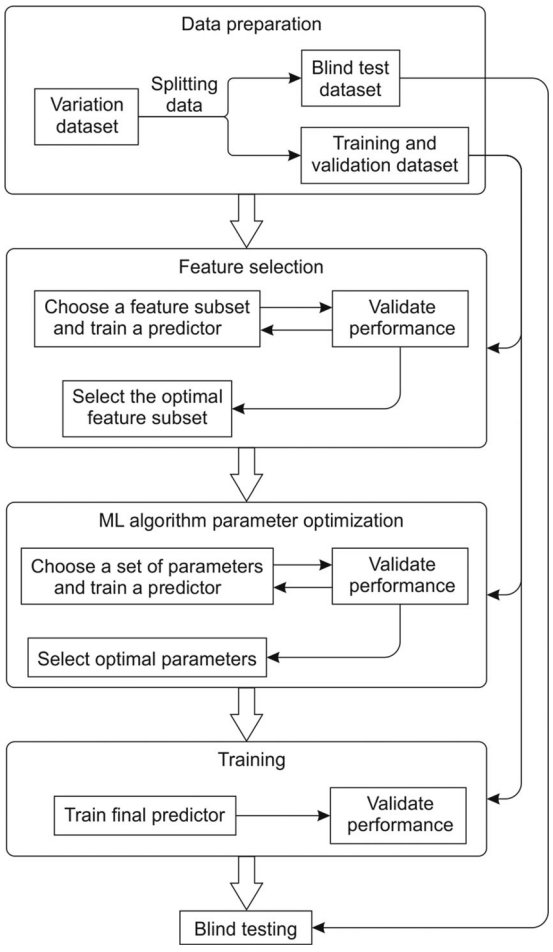

Figure 1. General framework for developing a machine learning-based method. A blind test dataset should be separated before performing feature selection and algorithm parameter optimization. The splitting should be done so that all variations in the same protein or preferably protein family should be kept in either the training or the test dataset.

ML methods are trained to detect differences in datasets. For that purpose, a good quality training dataset is required. The dataset should represent the space of possible cases [Nair and Vihinen, 2013]. This space is huge for genetic variations as there are so many different effects and underlying mechanisms. ML predictors are trained with known positive and negative instances in an approach called supervised learning that leads to reorganization of the system—learning (Fig. 1). Once the method has learned to distinguish between the given examples, it can be used to classify unknown cases. The training set should contain about equal numbers of cases in each class. An imbalance in the numbers of cases in the classes can reduce the performance of the method [Wei and Dunbrack, 2013].

The same data cannot be used both for method training and testing (Fig. 1). The trick is to partition the dataset. This can be done in different ways, cross-validation being the most popular. The dataset is divided into disjoint partitions, one of which is used for testing and the others for training. This is repeated until all the partitions

have been used as test set. The average performance measures computed from the splits are used to describe the overall prediction performance. The prediction performance would be biased if the algorithm parameters and feature sets are optimized using the same or a part of the data. Therefore, the use of an independent test data set is highly recommended. The data in this set should not be used during any of the previous steps of method development (Fig. 1).

Conservation Scores

Evolutionary conservation-based features are powerful for predicting the impact of variations. Disease-causing variants appear frequently at evolutionarily conserved positions. These positions are usually essential for the structure or function of the protein [Miller and Kumar, 2001; Vitkup et al., 2003; Shen and Vihinen, 2004]. On the other hand, disease-associated variations are underrepresented at positions that are variable during evolution [Miller and Kumar, 2001]. Evolutionary conservation is estimated based on multiple sequence alignment (MSA) of evolutionarily related sequences, called homologous sequences. Homologous sequences have a common ancestor and are separated by speciation (called orthologs) or by duplication (called paralogs) during evolution. Homologous sequences often have a high sequence similarity compared with unrelated sequences. The Basic Local Alignment Search Tool (BLAST) is often used to find similar sequences; however, the obtained sequences may not necessarily be evolutionarily related. Sequences annotated as homologs and orthologs in sequence databases are also used to obtain MSA. Several measures have been derived from MSAs including normalized frequencies of amino acids [Ng and Henikoff, 2001], evolutionary rates [Cooper et al., 2005], selective pressures [Niroula et al., 2015], sequence profiles [Calabrese et al., 2009], position-specific scoring matrix scores [Johansen et al., 2013], and position-specific independent count scores [Sunyaev et al., 1999]. Other conservation measures include pairwise alignment scores of homologous sequences [Choi et al., 2012], and scores derived from alignments of protein families such as substitution position-specific evolutionary conservation (subPSEC) [Thomas and Kejariwal, 2004] and hidden Markov Model (HMM) PHC scores [Karchin et al., 2005]. Most methods developed to predict the impact of variations use conservation scores as features. Some methods are completely based on conservation scores by optimizing the conservation scores to predict the impact of variations [Ng and Henikoff, 2001; Choi et al., 2012].

Energy Functions

Proteins form complex structures that are stabilized by various interactions. Variations can disrupt or form these interactions and therefore change the stability of a protein. Several stability predictors use energy functions, whereas others are ML-based methods. The energy function-based methods can be grouped into physical potential function, statistical potential function, and empirical potential function methods [Guerois et al., 2002; Capriotti et al., 2004; Parthiban et al., 2006]. These methods calculate the free energy change ($\Delta\Delta G$), that is, the free energy difference between the normal and variant protein. The most common applications utilize statistical potential calculations since physical potential calculations are very time-consuming [Gilis and Rooman, 2000]. The empirical potential methods combine and weigh physical and statistical energy components together with, for example, structure-related features [Guerois et al., 2002]. Some methods have also combined ML and energy function for stability prediction [Dehouck et al., 2009; Masso and Vaisman, 2010; Laimer et al., 2015].

Hybrid Methods

Hybrid methods combine diverse features. Here, we discuss hybrid methods that utilize evidence from experimental tests and clinical features. Hybrid methods integrate evidence for clinical features with scores obtained from predictors. The first step involves predicting scores for variation impact. In the second step, the scores from the first step are used as priors and integrated with evidence to estimate the posterior probability. The final classification of variations is based on the posterior probability. This approach has been used to classify variants in *BRCA1* and *BRCA2* genes in breast cancer [Goldgar et al., 2004, 2008; Lindor et al., 2012] as well as MMR genes in Lynch syndrome [Thompson et al., 2013a; Thompson et al., 2014]. Recently, this approach was implemented for classification of mitochondrial tRNA variations [Niroula and Vihinen, 2016].

Types of Tools

Before starting variation interpretation, one has to assume that the sequencing-related steps have been professionally and properly performed. Steps such as alignment, variation calling, and assessment of read quality are essential prerequisites. Variation interpretation cannot correct mistakes made in earlier steps; therefore, it is essential that the users of these tools are aware of the quality of the detected variants. In this respect, issues such as coverage and sequencing technologies used are very important; however, these are outside the scope of this article. Bear in mind that the interpretations of variation callers do not always agree [O'Rawe et al., 2013].

The effects of many deleterious variants are straightforward to explain, such as large deletions, protein truncations, amphigoric amino acid insertions and deletions that change the sequence after the variation site, and some other types. Usually, the most difficult ones to interpret are minor alterations, most often single-nucleotide substitutions and consequent amino acid changes. The effects on proteins are more difficult to study experimentally than those on DNA or RNA. Hybridization and sequence complementarity based methods work (almost) equally well for all nucleotide sequences, whereas a dedicated analysis protocol is needed for every individual protein and quite often even for individual variants, see Storz and Zera [2011], Yates and Sternberg [2013], and Perniola and Musco [2014], for example. Experimental methods are the first choice to study the detailed effects of variants. Such methods are often expensive and time-consuming and cannot be performed in the scale that exome and complete genome sequencing methods produce variants. Computational methods are thus useful for prioritizing variants for experimental studies [Thusberg and Vihinen, 2009; Zhang et al., 2012; Kucukkal et al., 2014]. Several computational tools have been developed for variation prioritization. The methods vary widely in the principle, implementation, and application.

Tolerance Predictors

Predictors for amino acid substitutions

The largest number of predictors is for tolerance prediction, that is, pathogenicity of variants. They vary widely in their concept and implementation [Karchin, 2009; Thusberg and Vihinen, 2009; Capriotti et al., 2012]. Most predictors are available for amino acid substitutions that are often erroneously called as missense variations [Vihinen, 2015a]. The methods utilize evolutionary information (alone or together with other features) from MSAs to estimate the conservation of the variant position. Practically, all the existing

methods utilize sequence comparison data in one way or other. Some methods compute conservation scores and optimize them to classify the variations [Ng and Henikoff, 2001; Choi et al., 2012]. Other methods use the conservation scores as features for training. Additional features used by methods include physicochemical properties of the amino acids (e.g., hydropathy, charge, size, secondary structure propensities), annotations in sequence and structure databases, information from three-dimensional protein structures, gene ontology (GO) annotations, and so on. These methods 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], PON-P2 [Niroula et al., 2015], SNPs&GO [Calabrese et al., 2009], and VEST [Carter et al., 2013]. Some representative methods are listed in Supp. Table S1. When considering the types of amino acid changes, disease-causing variants typically have more drastic changes in their physicochemical properties than tolerated variants [Steward et al., 2003; de Beer et al., 2013]. Variations at functionally and structurally important sites have a more severe impact than those at nonfunctional sites. Although GO annotation is a protein-specific feature, it has been shown to be useful for classifying deleterious and benign variations [Calabrese et al., 2009; Niroula et al., 2015].Some methods use features derived from three-dimensional protein structures. However, since experimental structures are not available for all proteins, these features cannot be used for variations in all proteins unless the structure or the features are predicted that might introduce problems and reduce method performance. Some metapredictors that utilize the results of other tools have been implemented including Condel [Gonzalez-Perez and Lopez-Bigas, 2011], PON-P [Olatubosun et al., 2012], Meta-SNP [Capriotti et al., 2013], and PredictSNP [Bendl et al., 2014].

The majority of these tools utilize ML algorithms. They are trained and tested on variations with known impact. Usually, the datasets have been different but recently benchmark datasets of sufficient size have become available and are being used increasingly. The VariBench database [Nair and Vihinen, 2013] contains positive and negative datasets for several types and effects of variants, whereas the VariSNP database contains only benign variants extracted from dbSNP [Schaafsma and Vihinen, 2015].

Methods intended to handle NGS datasets should, in addition to being accurate, have the capability to handle large datasets in reasonable time. Only some of the tools can handle very large datasets in reasonable time. In a test, the elapsed time ranged from about 10 min to more than 2 weeks for 30,000 variants [Niroula et al., 2015]. In addition, the application of such methods also depends on the possibility of submitting data in several formats, including those used by sequencers, such as variant call format (VCF) files. A programmatic interface (API) available for PON-P2 [Niroula et al., 2015] and possibility for local installation in some other programs allow seamless integration to variation analysis pipelines.

Predictors for synonymous variations

Although amino acid substitutions are the most common variants associated with disease, other types of variations including synonymous variations (that do not alter amino acid sequence) are associated with several diseases [Sauna and Kimchi-Sarfaty, 2011], for example, by altering splicing, changing mRNA structure, or changing translation speed [Hunt et al., 2014]. Synonymous variations at the splice sites as well as in exonic splicing regulatory regions may lead to splicing defects. Several splice-site effect predictors (discussed later) predict the impact of both intronic and exonic variations on splicing. The Silent Variation Analyzer (SilVA) is a method for prioritization of harmful synonymous variations [Buske et al., 2013]. The majority of the variations in the training data lead to a splicing defect but there are also variations that alter the methylation pattern or translational efficiency. The method suffers from the limitation of small numbers of known disease-causing variations, just 33 and six variants for training and validation, respectively.

Predictors for insertions and/or deletions

Insertions and/or deletions in coding regions can lead to three types of variations in a protein sequence: amphigoric alteration of the amino acid sequence due to RNA frameshift; insertion or deletion of one or more amino acid(s) without changing the reading frame; or termination of amino acid sequence by introducing an mRNA stop codon at the variation site or soon after it. Insertions or deletions of length divisible by 3 do not lead to translation frame alteration (i.e., do not change the amino acid sequence after the variation site). Tools in this domain include KD4i [Bermejo-Das-Neves et al., 2014] and SIFT Indel [Hu and Ng, 2013] that predict the impact of nonframeshifting insertions and deletions. Some other tools predict the impact of frameshifting as well as nonframeshifting insertions and deletions [Zia and Moses, 2011; Hu and Ng, 2012; Zhao et al., 2013; Liu et al., 2014; Douville et al., 2016] (Supp. Table S1). CADD [Kircher et al., 2014], MutationTaster2 [Schwarz et al., 2014], PROVEAN [Choi et al., 2012], and VEST [Carter et al., 2013; Douville et al., 2016] can predict impacts of insertions, deletions, as well as amino acid substitutions.

Predictors for variations in noncoding regions

Although variation interpretation methods largely focus on coding regions and protein sequences, some methods have been developed to predict the impact of variations in noncoding regions including regulatory regions [Macintyre et al., 2010; Manke et al., 2010; Ritchie et al., 2014; Lee et al., 2015; Zhou and Troyanskaya, 2015] (Supp. Table S1). Certain generic tolerance prediction tools including CADD [Kircher et al., 2014] and MutationTaster2 [Schwarz et al., 2014] can predict the impact of variations in coding as well as noncoding regions. The ENCODE project [Sloan et al., 2016] has compiled experimental and predicted information for vast amounts of noncoding sites, which can be used for prediction purposes.

Numerous tools and services have been established to collect and disseminate predictions of multiple predictors. The database of human nonsynonymous single-nucleotide variations (dbNSFP v3.0) contains results for 14 predictors, altogether for 83 million variants [Liu et al., 2016]. Annotation tools including ANNOVAR [Yang and Wang, 2015], AVIA v2.0 [Vuong et al., 2015], SnpEff [Cingolani et al., 2012], Variant Effect Predictor (VEP) [McLaren et al., 2010], and so on provide predictions of several predictors.

Domain-, Protein-, and Disease-Specific Predictors

Several databases and annotation tools are specific for certain genes, protein domains, or regions, for example, primary immunodeficiency-causing genes [Piirilä et al., 2006; Samarghitean et al., 2007; Ortutay and Vihinen, 2009], MMR genes [Thompson et al., 2014], protein kinase domain [Stenberg et al., 2000; Ortutay et al., 2005; Vazquez et al., 2016], phosphorylation sites [Safaei et al., 2011; Hornbeck et al., 2012], and many others. Specific predictors are available for variations in MMR genes [Chao et al., 2008; Ali et al., 2012; Thompson et al., 2013b; Thompson et al., 2014; Niroula and Vihinen, 2015a], the cystic fibrosis transmembrane conductance regulator protein [Masica

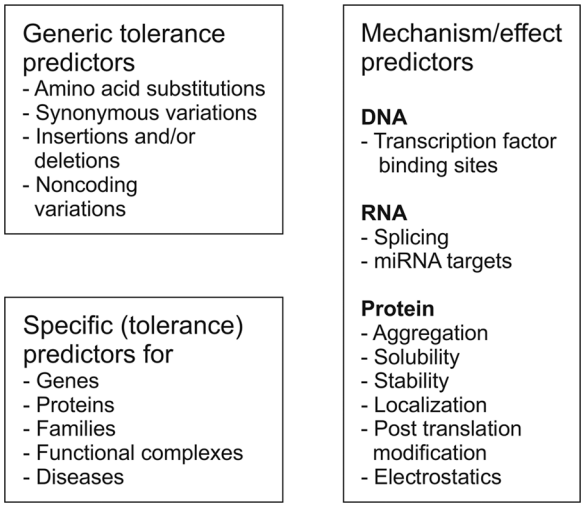

Figure 2. Types of variation effects and predictions.

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 others (Supp. Table S1). The size of the training dataset for such specific predictors is usually small, which may cause overfitting unless specifically taken care of. As it is difficult to find sufficient numbers of validated cases for evaluating specific predictors, it is essential to check how the performance of the methods has been determined. The methods should be tested on independent test datasets and all the recommended performance scores (see below) should be reported.

Prediction of the Effect or Mechanism of a Variant

Tolerance predictors detect variants that can be harmful but cannot explain what is wrong with them and in which way. Variations can have various effects, consequences, and mechanisms (Fig. 2); for those in amino acids, see Vihinen [2015c]. Variation effect analyses have to be done separately as there is no tool to combine the different predictions and interpret their results together. As the methods have widely varying speeds, it is beneficial to run a fast tolerance predictor first and then use tools to predict the effects and mechanisms for the potential harmful variants. This is also warranted by the fact that method and effect-specific tools typically have higher error rates than the best tolerance tools and they can be substantially slower to run. Therefore, filtering variants by tolerance predictors could improve their performance.

For some of the effects, including protein stability, aggregation, and solubility, there are several tools, some with tested performance. The biggest problem for the end-user with many of these predictors is that the method performance has often been presented in a somewhat biased or selective way and without comparison to related tools. Thus, it may be impossible to interpret the quality of the obtained predictions. The situation is improving due to the use of benchmark databases.

Protein stability predictions

Stability affects function, activity, and regulation of biomolecules. Incorrect folding and decreased stability are the major consequences of pathogenic amino acid substitutions [Wang and Moult, 2001; Ferrer-Costa et al., 2002; Stefl et al., 2013; Peng and Alexov, 2016]. Substitutions can cause, for example, a reduction in the hydrophobic area, over packing, backbone strain, loss of electrostatic interactions, and thereby affecting protein stability. Alterations to intramolecular interactions affect the free energy difference ($\Delta\Delta G$) between the folded and unfolded states of proteins.

Several methods predict stability effects by comparing $\Delta\Delta G$ between a wild-type protein and its variants. Some of these methods rely on energy functions, whereas others apply ML approaches. Methods utilizing energy functions can be subdivided to physical potential approaches, statistical potential approaches, and empirical potential energy approaches [Guerois et al., 2002; Parthiban et al., 2006]. All the ML-based methods are trained with data from ProTherm database [Kumar et al., 2006]. They use features derived from tertiary protein structure, evolutionary conservation, sequence environment, network topology, and properties of amino acid residues to describe each variation [Capriotti et al., 2005; Cheng et al., 2006; Capriotti et al., 2008; Teng et al., 2010; Chen et al., 2013; Yang et al., 2013; Folkman et al., 2014; Giollo et al., 2014; Pires et al., 2014; Fariselli et al., 2015]. It is evident that different features are important for tolerance and stability predictors. Some methods combine both ML algorithms as well as energy functions [Dehouck et al., 2009; Masso and Vaisman, 2010; Laimer et al., 2015]. While most of the methods predict the stability changes for individual variations, some predict the combined effect of variation pairs [Huang and Gromiha, 2009].

Protein localization predictions

To function properly, a protein must be translocated to the appropriate cellular compartment. Proteins are directed to the locations by short targeting signal peptide sequences. Substitutions in signal peptides can disrupt or alter the signal and prevent localization or redirect the produced protein to a wrong cellular part. When a protein is not transported to the correct subcellular location, central reactions can be inactivated or signaling pathways are misregulated. An active mislocalized protein is likely to have harmful effects by exerting its function in the wrong environment.

Despite the fact that numerous predictors are available for predicting the protein subcellular localization, they do not predict the impact of variations on protein localization. Using WoLF PSORT [Horton et al., 2007] and a pipeline of several tools [Emanuelsson et al., 2007], disease-causing variations were studied in relation to protein localization [Laurila and Vihinen, 2009]. Among the 22,416 disease-causing variations tested, hundreds of variations are likely to change protein localization. PRObalizer was later developed to predict the cellular localization as well as the impact of variations on localization [Laurila and Vihinen, 2011].

Splice-site effect predictors

Splicing of nascent premessenger RNA (pre-mRNA) is an essential step in mRNA maturation. Alternative splicing increases mRNA and protein diversity by generating multiple mRNA products from the same pre-mRNA. Hence, one gene can code for multiple protein products. At least about 95% of human genes have more than one transcript [Pan et al., 2008]; however, many of them are extremely rare and their biological significance has not been established. Among the variations in the HGMD, about 10% of the disease-causing variations occur within splice sites [Krawczak et al., 2007], which is likely to be an underestimation since the exonic variations outside the splice sites and variants at deep intronic sites could

potentially disrupt splicing [Cartegni et al., 2002; Hunt et al., 2014]. Among the primary immunodeficiency-causing variants in IDbases, 15% are splice-site variants [Piirilä et al., 2006]. Exonic RNA substitutions are classified as missense, nonsense, or silent variations and their role in splicing is not generally analyzed. Several methods have been developed to predict splice sites [Hebsgaard et al., 1996; Pertea et al., 2001; Yeo and Burge, 2004] and splicing regulatory sites [Fairbrother et al., 2002; Wang et al., 2004; Goren et al., 2006; Smith et al., 2006]. There are also some tools to predict the effects of variations on splicing (Supp. Table S1). The Human Splicing Finder (HSF) combines results from several splice site and splice regulatory site prediction tools to estimate the impact of variations on splicing motifs [Desmet et al., 2009]. HSF and a commercial program ASSEDA [Nalla and Rogan, 2005] can predict the splice-site defect of exonic as well as intronic variations. MutPredSplice [Mort et al., 2014] and SKIPPY [Woolfe et al., 2010] predict splicing defects for exonic variations only.

Protein disorder predictors

Although proteins are often considered as static molecules, they undergo all kinds of fluctuations at different time scales. Some proteins or parts of proteins do not have an ordered structure at all. These intrinsically disordered regions and proteins are involved in several activities [Dunker et al., 2002]. The disordered regions are difficult to crystallize because they are highly flexible. Although numerous methods are available for predicting protein disorder, tools to predict the impact of variations on protein disorder were released only recently. Disease-causing amino acid substitutions are more often caused by disordered-to-ordered changes compared with the neutral amino acid substitutions [Vacic and Iakoucheva, 2012]. PON-Diso is a ML-based method for prediction of variation effects on protein disorder [Ali et al., 2014]. The method uses features representing evolutionary conservation and biochemical properties of amino acids.

Protein aggregation predictors

Protein aggregation is characterized by an increased level of β-sheet structure. Amyloid fibrils and amorphous aggregates are formed by protein aggregation. These fibrils and aggregates are caused by irreversible structural alterations and are associated with neurodegenerative diseases such as Alzheimer, Huntington, and Parkinson disease. Normal proteins can become toxic upon fibrillation [Bucciantini et al., 2004] and other proteins not related to amyloid diseases can aggregate under destabilizing conditions [Chiti et al., 1999]. Numerous tools predict amyloidogenic regions in protein sequences [Tartaglia et al., 2008; Garbuzynskiy et al., 2010; Maurer-Stroh et al., 2010; Emily et al., 2013]. Amino acid substitutions can increase the tendency of amyloidogenic proteins to aggregate. Methods for predicting the effect of amino acid substitutions on amyloidogenic proteins include AGGRESCAN [Conchillo-Sole et al., 2007], PASTA2.0 [Walsh et al., 2014], TANGO [Fernandez-Escamilla et al., 2004], and Aggrescan3D [Zambrano et al., 2015] (Supp. Table S1). Aggrescan3D requires three-dimensional protein structure to predict the effect of amino acid substitutions to aggregation, whereas the others utilize peptide sequences.

Protein solubility predictors

Protein solubility has been of great interest because of its relevance to protein (over)expression and for protein structural studies. Both nuclear magnetic resonance and X-ray crystallography require the protein of interest to be soluble in a high concentration. A low protein solubility has also been associated with diseases. Methods for predicting the effects of variations on protein solubility include OptSolMut [Tian et al., 2010], CamSol [Sormanni et al., 2015], and PON-Sol [Yang et al., 2016] (Supp. Table S1). OptSolMut and CamSol require a known 3-D protein structure, whereas PON-Sol is based on sequence context, evolutionary conservation, properties of amino acids, and other sequence-based features. Solubility and aggregation are related physical phenomena. Aggregation is an irreversible process due to structural and other changes, whereas precipitated proteins may be dissolved by dilution [Arakawa and Timasheff, 1985].

Additional mechanisms that amino acid substitutions can affect and for which prediction methods are available include, electrostatics, residue-residue contacts, side chain rotamers, cavity formation, and sterical clashes [Thusberg and Vihinen, 2009; Vihinen, 2015c]. When a residue is replaced, its chemical and physical properties may be altered causing structural alterations, especially when the original residue is smaller than the substituting one.

Prediction of Cancer Variants

Cancer is a systemic disease where the accumulation of genetic variations leads to alterations of multiple phenomena and processes. Depending on the cancer, the numbers of somatic variations range from a few to more than one million. A majority of these are so-called passengers and have little or no effect on the disease. Driver variants essential for initiation, development, progression, and/or maintenance of tumors are typically small in number.

Large amounts of cancer genomic data are available from genomic projects including the Cancer Genome Project (CGP, https://www.sanger.ac.uk/research/projects/cancergenome/), The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/), and the International Cancer Genome Consortium (ICGC, https://icgc.org/). These projects collect and provide various types of genetic data for large numbers of cancer samples. These massive datasets provide unprecedented possibilities for data analysis. Various approaches have already been used to study mechanisms of tumorigenesis. Still, the vast majority of nonfunctional somatic variations remain a major challenge in cancer genomics.

Despite a large number of studies of genetic variation in numerous cancers, relatively little is known about the true cancer driver variants. The Cancer Gene Census (CGC, http://www.sanger.ac.uk/genetics/CGP/Census/) lists genes implicated in cancer [Futreal et al., 2004]. Variations have been identified in these genes, available in the COSMIC database [Forbes et al., 2011]. However, the variations have seldom been experimentally validated to be drivers. In the end, even in the genes listed in CGC, most variations are predicted to be harmless [Niroula and Vihinen, 2015b].

Several approaches have been taken to search for the driver variations, genes, networks, and pathways. Computational methods and tools for interpreting variations in cancers have been reviewed previously [Gonzalez-Perez et al., 2013; Ding et al., 2014; Raphael et al., 2014; Chen et al., 2015; Tian et al., 2015]. Here, we briefly discuss the commonly used methods and tools in cancer variation analysis. Tolerance prediction methods are often applied for analysis of somatic variations found in cancer genomes. Some of them have been used to develop cancer-specific predictors. Specific tools including CHASM [Wong et al., 2011], transFIC [Gonzalez-Perez et al., 2012], CanPredict [Kaminker et al., 2007], SPF-Cancer [Capriotti and Altman, 2011b], cancer-specific FATHMM [Shihab et al., 2013],

CanDrA [Mao et al., 2013], and others prioritize variants in cancer (Supp. Table S1).

Most of these methods are trained using frequent somatic variations in COSMIC and other cancer-related variations and putative neutral variations from various sources. Only a few databases contain validated cancer variants [Niroula and Vihinen, 2015b] (Table 1). The Database of Curated Mutations (DoCM) (http://docm.genome.wustl.edu/sources) contains disease-causing variations in cancers. The TP53 mutation database contains somatic variations in TP53 together with information about the effect on protein activity [Edlund et al., 2012]. Kin-Driver is a manually curated database of validated driver variations [Simonetti et al., 2014].

Frequencies and patterns of variations have been used to identify driver genes. These methods calculate the variation frequency compared with the expected background mutation rate. MutSigCV [Lawrence et al., 2013], MuSiC [Dees et al., 2012], and InVEx [Hodis et al., 2012] identify genes containing a significantly higher variation frequency compared with the background mutation rate. DrGaP [Hua et al., 2013] uses background mutation rates to identify significant genes and pathways, whereas OncodriveFM [Gonzalez-Perez and Lopez-Bigas, 2012] and MADGiC [Korthauer and Kendziorski, 2015] utilize the functional impact of variations. OncodriveCLUST [Tamborero et al., 2013] identifies genes that have significant clusters of variations in proteins. The 20/20 rule was derived to identify oncogenes and tumor-suppressor genes based on whether there are recurrent variations or the variations are distributed along the protein sequences [Vogelstein et al., 2013]. According to the rule, oncogenes bear at least 20% of the variations in the gene at recurrent sites and the variants lead to amino acid substitutions, and in tumor-suppressor genes, at least 20% of the total variants are inactivating. Based on similar concept, SomInaClust [Van den Eynden et al., 2015] and DOTS-Finder [Melloni et al., 2014] identify tumor-suppressor genes and oncogenes. ContrastRank [Tian et al., 2014] prioritizes putative driver genes based on mutation rates in the 1000 Genomes Project and normal samples in the TCGA, and ActiveDriver [Reimand and Bader, 2013] identifies genes containing significantly higher variation frequencies at active sites.

Certain methods recognize gene/protein interaction networks and pathways affected in cancer samples. NetBox identifies significant network modules and driver variations in cancer [Cerami et al., 2010]. VarWalker identifies sample-specific gene interactors and integrates them to prioritize driver genes [Jia and Zhao, 2014]. Mutual exclusivity of variations and/or genes across cancer samples has been used to identify pathways affected in large numbers of samples. This approach is based on the observation that variations affecting the same signaling pathways do not co-occur in the same samples [The Cancer Genome Atlas Research Network, 2008; Ciriello et al., 2012]. Dendrix [Vandin et al., 2012], MEMo [Ciriello et al., 2013], MultiDendrix [Leiserson et al., 2013], CoMEt [Leiserson et al., 2015], MEMCover [Kim et al., 2015], NBM [Wu et al., 2015], RME [Miller et al., 2011], CoMDP [Zhang et al., 2014], and others [Lu et al., 2015] use mutual exclusivity to identify affected pathways and genes.

As the large cancer genomic projects have collected heterogeneous data from large numbers of samples, data integration has become essential. Methods that integrate different types of data, for example, genome, transcriptome, proteome, and epigenome, include DawnRank [Hou and Ma, 2014], DriverNet [Bashashati et al., 2012], OncoIMPACT [Bertrand et al., 2015], and Pathway Relevance Ranking [Verbeke et al., 2015].

Cancer genomics is rapidly expanding and the numbers of new tools for analyzing cancer genomics are increasing rapidly. Predictors are important for prioritizing cancer-related variants, genes, networks, and pathways; however, they have limitations. Functional impact predictors have varying performance. Even minor differences in the performance lead to large numbers of differently predicted variations when applied to large datasets. Additional complexity arises because variations co-occur and their combined impact cannot be reliably predicted.

The frequency of a variation is not a direct indication of its relevance for disease. Some sites are just prone to variations, such as CpG sites [Ollila et al., 1996], but this does not necessarily mean that those variants are harmful. Variation spectra vary markedly between cancers. Those originating from actions of carcinogens such as melanoma and lung cancers accumulate large numbers of variants. In addition, the background mutation rate varies up to 100-fold in the genome [Hodgkinson and Eyre-Walker, 2011]. Large numbers of false positives could be identified due to high mutation rates in many cancers and many false negatives due to tumor heterogeneity [Lawrence et al., 2013]. These methods are less efficient in cancers that have a lower variation frequency. Even less frequent variations can equally affect the same pathways as frequent variations in cancer. Network- and pathway-based methods can point out the affected mechanisms. Since human protein interaction network and pathway databases are far from complete, these methods have limitations.

Method Performance Assessment

When choosing a prediction method, one should look at performance scores, preferably from independent benchmark studies. Three approaches can be used for testing method performance [Vihinen, 2012]. The first approach is challenge-based testing. Critical Assessment of Genome Interpretation (CAGI, http://genomeinterpretation.org/) is a challenge for method developers in the field of phenotypic impacts of genomic variation. Challenges do not aim for a systematic analysis of predictions, instead they assess what is currently doable, providing proof of concept, charting where to direct future efforts, and identifying new areas where predictive approaches would be needed.

The second test strategy is typically used by method developers. Most often the performance tests are not comprehensive, and the results are incomparable with those obtained from other methods due to using developer collected test sets. The third approach, systematic analysis, uses benchmark data and several evaluation measures to explain method performance.

Numerous measures should be used to describe predictor performance, since a single measure cannot capture all aspects of the performance. Typically, prediction methods are used as classifiers to define whether a case has the investigated feature or not. Results of this kind of binary predictions are presented in a 2×2 confusion or contingency table or matrix. Based on the four descriptors in the contingency table, several measures can be calculated (Fig. 3). Sensitivity, also called true positive rate (TPR) or recall, and specificity (true negative rate, TNR) show the ratio of the correctly predicted pathogenic and neutral cases. Positive predictive value (PPV/precision) and negative predictive value (NPV) are the conditional probabilities for pathogenic or neutral variants to be predicted as pathogenic or neutral, respectively. All these measures are calculated by using only half of the information in the contingency table. Accuracy and the Matthews correlation coefficient (MCC) utilize the whole matrix and are more balanced, representative, and comprehensive. To obtain a full picture of the predictor performance, it is important to evaluate all these six measures together [Vihinen, 2012].

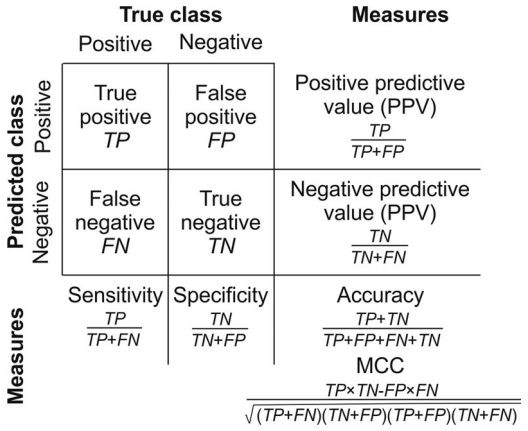

Figure 3. Recommended performance measures for binary predictors. Accuracy and MCC use all four values from the contingency table.

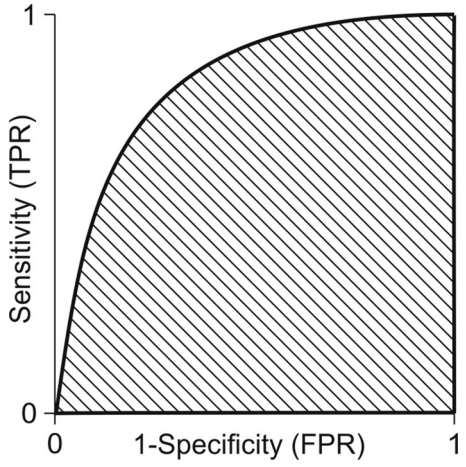

Figure 4. Example of a ROC curve. The shaded area is called area under the curve (AUC). It ranges from 0 to 1 where a random predictor will have an AUC 0.5 and a perfect predictor will have an AUC 1.

Unless sufficient and similar numbers of positive and negative cases are used, the values of NPV and PPV may be biased, even meaningless. Accuracy and the MCC are affected by class imbalance only in extreme cases. The usual requirement is that the numbers of positive and negative cases be equal. To overcome the class imbalance problem, different approaches can be taken [Vihinen, 2012]. It can be done by pruning the size of the bigger class to that of the smaller one. It is also possible to normalize values in the contingency table. Normalizing makes sense only if dataset is representative.

Receiver operating characteristic (ROC) analysis is a visualization of prediction performance. It indicates the tradeoffs between sensitivity and specificity. The ROC curve can be drawn when the predictor provides a score for the classification. The faster the curve rises and the higher it reaches in the beginning the better the method is (Fig. 4). The area under the ROC curve (AUC) has been used as a measure of goodness for predictions as it approximates the probability of ranking a randomly chosen positive instance higher than a randomly chosen negative one. A value of 0.5 indicates a random classification, whereas 1 would indicate a perfect binary classifier. The ROC curve shows the method's ranking potential, which is related to overall performance.

If there are more than two predicted classes, the measures discussed above cannot be applied [Vihinen, 2012]. The data can still be presented in an N x N contingency table. Many performance scores used for binary classification can be generalized to multiclass case. The data could be divided into several partitions of two categories. It is possible to calculate row- and column-wise ratios. The MCC is a special performance score for binary data of linear correlation coefficient, which can be used for several classes in its general format. Generalized squared correlation (GC^2) can be used for more than two classes [Baldi et al., 2000].

Method Performance Evaluations

Several studies have assessed the performance of prediction methods (Table 2). Here, we summarize results from them. There have been mainly two types of evaluations, larger benchmark-based studies and single gene, protein, or family-related investigations.

Protein tolerance predictors

A dataset of over 40,000 variants, 19,335 positive and 21,170 negative cases, was used to test the performance of nine widely used tolerance methods [Thusberg et al., 2011]. The programs were used to predict the entire dataset and the predictions were compared with the experimentally verified outcome. The results indicated the methods to have a wide variation in their performance. The best performing methods, SNPs&GO [Calabrese et al., 2009] and MutPred [Li et al., 2009], had MCC values close to 0.65. The performance differences have a big practical impact in exome and complete genome-wide analyses where the number of correctly predicted variants can vary by several hundreds or thousands per genome depending on the selected methods.

The predictors were further tested by addressing their performance on secondary structural elements and protein structural classes and by investigating effects of residue accessibility [Thusberg et al., 2011]. All programs predicted the effects of substitutions at different secondary structures with almost equal accuracy and precision as all variations in general. Solvent accessible surface areas of the positions did not markedly affect predictions; however, all methods were more sensitive to the effects of substitutions at buried positions. The performance varied significantly depending on the structural class of proteins. The results were generally better for proteins in the α-β class.

In another study, the authors generated benchmark datasets to evaluate the performance of eight methods and compared with their new tool, PredictSNP [Bendl et al., 2014]. Benchmark datasets were obtained by combining several previously used datasets and eliminating the variants used for training the evaluated methods. Three independent datasets were generated, the biggest containing 19,800 deleterious and 24,082 neutral variations. On the largest dataset, PhD-SNP and PredictSNP showed the best performance. The highest accuracy and MCC for both methods were 0.75 and 0.49, respectively.

The performance of seven prediction methods was evaluated using three different datasets and the methods were compared with PON-P2 [Niroula et al., 2015]. The two best performing methods were PON-P2 and MutationTaster2. The accuracy and MCC for PON-P2 were 0.86 and 0.71, respectively, in the PON-P2 test dataset. In the MutationTaster2 test dataset, the accuracy of PON-P2 were 0.95 and 0.90, respectively; for MutationTaster2, they were 0.89 and 0.79, respectively.

Table 2. Prediction method performance assessment studies

Number of variations[a]	Evaluated tools	Reference
Tolerance predictors		
40,505	MutPred, nsSNPAnalyzer, PANTHER, PhD-SNP, PolyPhen, PolyPhen-2, SIFT, SNAP, SNPs&GO	Thusberg et al. [2011]
43,882; 32,776; 3,497; 11,994	MAPP, nsSNPAnalyzer, PANTHER, PhD-SNP, PolyPhen, PolyPhen-2, PredictSNP, SIFT, SNAP	Bendl et al. [2014]
25,656; 2906; 1200; 15,818; 1,804	CADD, Condel, MutationTaster2, PolyPhen-2, PON-P, PON-P2, PROVEAN, SIFT, SNAP	Niroula et al. [2015]
40,389; 8,850; 10,226; 16,098; 12,729	CADD, FATHMM, GERP++, LRT, MutationAssessor, MutationTaster2, phyloP, PolyPhen-2, SIFT	Grimm et al. [2015]
30; 2,314	CADD, Condel, MutationAssessor, PANTHER, PolyPhen-2, SIFT	Miosge et al. [2015]
Protein stability predictors		
2,156	CC/PBSA, EGAD, FoldX, I-Mutant2.0, Rosetta, Hunter	Potapov et al. [2009]
2,716	CUPSAT, Dmutant, FoldX, I-Mutant2.0, I-Mutant3.0, Mupro, MultiMutate, Scide, SRide, Scpred	Khan and Vihinen [2010]
Splicing predictors		
272	ESE finder, Human Splicing Finder, MaxEntScan, NNSplice, RESCUE-ESE, Splice Site Finder	Houdayer et al. [2012]
623	GeneSplicer, GENSCAN, Human Splicing Finder, MaxEntScan, NNSplice, SplicePort, SplicePredictor, SpliceView, SROOGLE	Desmet et al. [2010]
2,959	GeneSplicer, GENSCAN, Human Splicing Finder, MaxEntScan, NetGene2, NNSplice, Position Weight Matrix, SplicePredictor	Jian et al. [2014]
Cancer variation predictors		
989	CanDrA (breast, lung, melanoma), CHASM (breast, lung, melanoma), Condel, FATHMM (cancer), FATHMM (missense), MutationAssessor, MutationTaster, PolyPhen-2, PROVEAN, SIFT, VEST	Martelotto et al. [2014]

[a]Separator when there is more than one dataset.

A study of 10 prediction methods that were evaluated with five datasets [Grimm et al., 2015] revealed that the performance of methods decreases when the variations present in the methods' training dataset were eliminated from the test data. It is well known that the performance of a method is overestimated when tested on cases present in its training data that is called "Type 1 circularity." Therefore, the guidelines for method testing demand the use of disjoint datasets for training and testing [Vihinen, 2013]. Some methods still had good performance after avoiding "Type 1 circularity"; however, it decreased significantly when variations in proteins present in the training dataset were excluded. When the evaluation data and the training data were disjoint at the protein level, the performance of all the methods was reduced and the highest MCC was 0.36. The differences in the performance scores were due to two types of circularity. The study showed that appropriate measures should be taken based on the features used to train the methods to obtain their actual performance. This study did not include PON-P2 that has been trained and tested on data where proteins in the same family were either present in the training or in the test dataset; thus, the results presented above for PON-P2 are without circularity. The datasets for all the studies mentioned above are in VariBench [Nair and Vihinen, 2013].

The prediction performance of six prediction methods was assessed using 30 variants in 23 primary immunodeficiency related proteins and 2,314 variants in the TP53 protein [Miosge et al., 2015]. The methods showed low false negative prediction rates but very high false positive prediction rates, that is, 42% for PolyPhen2 and 45% for CADD for the TP53 variants. A comparison of transcriptional activities showed that the false positive variants had a lower transcriptional activity, that is, 86% of the activity for true negative variants. Hence, the compared methods overestimated the deleteriousness of the tested variants. Full set of the measures described above was not used for the assessment.

Protein stability predictors

Six programs were compared [Potapov et al., 2009] by using 2,156 single variations from ProTherm (Table 2). The goal of the study was to compare the performance of the methods in $\Delta\Delta G$ prediction and thus the effect on protein function was not addressed. The only measure used was correlation between the experimental and predicted $\Delta\Delta G$ values. The highest correlation coefficient of 0.59 was for EGAD. Combination of the methods did not improve the correlation coefficients.

In another study, 11 methods were evaluated [Khan and Vihinen, 2010]. The dataset contained 1,784 variations in 80 proteins, with 1,154 positive cases, of which 931 were destabilizing ($\Delta\Delta G \geq 0.5$ kcal/mol), 222 were stabilizing ($\Delta\Delta G \leq -0.5$ kcal/mol), and 631 were neutral (0.5 kcal/mol $\geq \Delta\Delta G \geq -0.5$ kcal/mol). The overall performance was best for I-Mutant3.0 (structure version) [Capriotti et al., 2008], Dmutant [Zhou and Zhou, 2002], and FoldX [Schymkowitz et al., 2005] with accuracies from 0.54 to 0.64. However, the MCC values were poor for all the tested predictors, the best being I-Mutant3.0 (structure version) that had an MCC of 0.27.

All the programs succeeded better when predicting either stability-increasing or stability-decreasing variations separately [Khan and Vihinen, 2010]. The majority of the programs predicted variants in different secondary structural elements with almost equal accuracy. All the programs predicted the effects at buried sites with a higher accuracy and specificity than those at surface positions. When the original residue is replaced by a residue with a smaller volume, a cavity may form in the protein interior with a detrimental effect on the stability. Large volume changes were predicted better than small changes. For destabilizing variations, there were no significant performance deviations in the methods for different charge changes. The predictors performed similarly despite differences in the extent to which the volume or charge varied as functions of the original residue and the variant.

Splice-site predictors

Splicing effect predictors have been assessed in several studies [Houdayer et al., 2008; Holla et al., 2009; Vreeswijk et al., 2009; Thery et al., 2011; Houdayer et al., 2012; Grodecka et al., 2014]. The methods have high performances for variations occurring at the canonical AG/GT splice sites, and lower performances for variations further away from the splice site. Six splice-site predictors were tested using 17 variations at splice-site junctions and 22 variations outside the junction [Houdayer et al., 2008] (Table 2). The effects of variants at the splice-site junction were correctly predicted but there were greater inconsistencies in the scoring of variations outside the junctions. In another study, the performance of the same six tools was assessed using 272 variations in the *BRCA1* and *BRCA2* genes [Houdayer et al., 2012]. The variations were grouped into canonical AG/GT splice sites, splicing consensus region (11 base pairs at the 5' splice site and 14 base pairs at the 3' splice site), intronic region outside the consensus region, and exonic region outside the consensus region. Among them, 65 variants occurred within the consensus-splicing region. The differences between the scores for the reference sequence and the altered sequence (called as variant score) were used to define the splice-site defects. Due to differences in the scoring systems, the tools performed best at different cutoffs of the variant score. MaxEntScan (MES) had the best sensitivity (0.96) and specificity (0.74) at a 15% cutoff of the variant score, whereas splice site finder (SSF) [Shapiro and Senapathy, 1987] had the best sensitivity (0.91) and specificity (0.87) at a 5% cutoff of the variant score. Combination of MES and SSF had the highest sensitivity (0.96) and specificity (0.83) for variants in the consensus splicing region.

In another study, the performance of nine predictors was assessed using 538 variations affecting splicing and 85 variations that do not affect splice sites [Desmet et al., 2010]. Most of the tested tools showed high accuracies (as high as 100%) for variants at 3′ and 5′ invariant positions. The methods showed high accuracies for variants at +3 and +5 positions at the 5' end, but slightly lower accuracies for variants at other positions at the 5′ and 3′ ends. Among exonic variations, the methods predicted most accurately the variants at the last bases of the exons. MES showed the highest accuracy for intronic variants except those outside the immediate 5′ and 3′ ends but within 100 bp from the ends. MES also showed a higher false positive rate (25%) compared with the other methods. Overall, the methods showed very different prediction accuracies for variants at different locations.

Recently, a larger dataset containing 1,164 positive and 1,795 negative cases was used to evaluate eight splice-site predictors [Jian et al., 2014]. The dataset contained single-nucleotide substitutions in the splicing consensus region. Overall, the compared methods showed high accuracies ranging from 0.83 to 0.91. Position weight matrix and MES had the best performance with accuracies of 0.911 and 0.895, respectively. The performance assessments show that the predictors have high performances for variants in consensus splicing regions, as expected due to the high conservation of these sites. Since different performance measures have been used for the assessment and none of the studies reports all the recommended performance measures, the results remain descriptive.

Cancer variation impact predictors

The performance of both cancer-specific tools as well as generic tolerance predictors on cancer variants were assessed using putative variations with high minor allele frequencies and recurrent variations in the COSMIC database [Gnad et al., 2013]. There are problems with the design of the study, however. Recurrent variations from COSMIC have been used to train cancer-specific predictors, causing circularity. Relevance of high-frequency variations used as neutral dataset in this study is unknown in cancer. In another evaluation of 15 prediction methods, experimentally validated variants causing Li-Fraumeni syndrome, Li-Fraumeni-like syndrome, and early-onset breast and ovarian cancer were used [Martelotto et al., 2014]. In the assessment, most methods showed high PPVs but the NPVs were variable. CanDrA and CHASM can take into account the tissues of origin. Predictions of CanDrA for specific tissues showed variable performances, whereas CHASM showed similar performances for all three tested tissue types. Recently, a tolerance predictor, PON-P2, was assessed using 1,058 cancer variants where 69.3% (733) were predicted as harmful, 0.3% (4) as benign, and the remaining (321) as variants of unknown significance [Niroula and Vihinen, 2015b]. The method could not be tested with noncancer variants due to a lack of data. As experimental evidence for the involvement of specific variants in cancer is scarce, it has been difficult to develop and test performance of cancer predictors. This is further complicated by the fact that several alterations are simultaneously required for cancer development.

Problems and Remedies

Problems in the application of computational methods to variation analysis can appear at several stages and affect the interpretation of the results [Rogan and Zou, 2013; Vihinen, 2013, 2014a]. Before using any method, one should understand how it is used and for what purpose. As variation interpretation has become an ever more important topic, many methods have been developed. There has sometimes been a "me too attitude," with people applying methods that have been developed originally in some other domain. These may cause problems, especially when the application domain and the datasets are not well known to the developers. Some quite poor datasets have been used for training many predictors. Methods developed in one application area can be useful in another domain when properly trained.

Another problematic area has been overfitting of data, see, for example, Grimm et al. [2015]. If the training data and features are not representative of the full spectrum of true cases, the generalization ability of an ML algorithm is poor and the method is overfitted. Overfitting may also occur due to other reasons that were described in an earlier section of ML methods. The same data should never be used for both training and testing [Vihinen, 2012] since it introduces circularity [Grimm et al., 2015]. Even usage of highly similar cases, such as variants in the same protein or protein family, may bias the method and provide good test performance but poor performance when applied to unseen real-life cases [Capriotti and Altman, 2011a; Bendl et al., 2014; Niroula et al., 2015].

Many variation effects are rare or very rare; therefore, it may be difficult to obtain sufficient numbers of cases for method training. It has been shown that the best predictor performance is obtained when using a balanced training dataset, that is, data that have equal numbers of positive and negative cases [Wei and Dunbrack, 2013]. If this is not done, the predictor will be biased toward the bigger class. For example, if a predictor is trained on a dataset containing more positive cases than the negative cases, the predictor will have a high TPR but a low TNR. The training dataset can be balanced by downsampling the bigger class, oversampling the smaller class or by assigning class weights [Vihinen, 2012]. It is also recommended to use a balanced test dataset for performance evaluation. However, the dataset imbalance does not have as high impact in testing as in

training. Balanced performance scores such as balanced accuracy should be used in such cases.

It is not uncommon in the literature that authors present overblown statements about the quality of their methods [Vihinen, 2014a]. Therefore, it is important to look for independent quality assessments. Another problem is selective presentation of quality measures, showing only those that have a good performance for the tool. Many methods are not described in sufficient detail so that the readers could get a full picture about their quality. Many journals limit method details in articles such that the supplementary materials may be much longer than the actual article. In principle this is not problematic; however, quite often the supplements are not as carefully produced as the main text.

Human Mutation has published requirements both for prediction method developers and users [Vihinen, 2013] (see Boxes 1 and 2). For method users, these include the description of the choice and details of the method employed, its version, parameters and program options, and so on. In addition, program-generated statistical values should be provided. Special attention needs to be paid to the interpretation of the obtained results. Without all these details, readers cannot fully understand and evaluate the given results.

Box 1. Publication guidelines for prediction method developers

Adapted from Vihinen [2013] and *Human Mutation* submission requirements.

General requirements

- The method should be novel and have good performance; clearly improved performance; or describe novel useful parameters or applications and have reasonable performance

Method description

- Has to be in sufficient detail, including (depending on the approach): description of training, optimization, input parameters and their selection, details of applied computational tools, hardware requirements, and time complexity

Datasets

- Description of training and test datasets, preferably the largest high-quality benchmark datasets
- Description of how data were collected, quality controls, details of sources, sizes of datasets
- Distribution, preferably via a benchmark database

Method performance assessment

- Report of all appropriate performance measures. In the case of binary classifiers, six measures characterizing the contingency matrix (Fig. 3), if possible, including ROC analysis
- Statistically valid evaluation with sufficiently large and unbiased dataset(s) that contains both positive and negative cases and comparison to related methods
- Training and test sets have to be disjoint
- Dataset imbalance needs to be mitigated

Implementation

- Detailed description of the method including user guidelines and examples
- Availability of the program, either to download or as a Web service, and preferably open source
- Batch submission possibility is highly recommended

Sometimes methods are used outside their initial application area and without knowledge about their suitability to the new application [for more details, see Vihinen, 2014a]. One example is the use of generic protein disorder predictors to investigate effects of amino acid substitutions. Recently, the applicability of these tools was assessed and found to be very poor outside their primary use case [Ali et al., 2014]. Only PON-Diso, a tool dedicated for disorder affecting variants, had reasonable performance.

Box 2. Publication guidelines for prediction method users

Adapted from Vihinen [2013] and *Human Mutation* submission requirements.

- Choose a method designed for the task and with proven performance.
- Sometimes it may be beneficial to use several methods but beware of possible problems due to using similar or same data.
- Report in detail the method used including citation, URL, version, used parameters, and program options.
- Additional information such as user-generated multiple sequence alignment should be made available.
- Report output data plus P values, confidence intervals, and all other reliability measures provided by the tool(s).
- Be careful with data interpretation.
- Understand the principle of the method, its use, limitations, and applications.

Since many methods are available for predicting pathogenicity, it may be tempting to combine the results from several tools. This can be done, but there are certain caveats. The consensus approach can be problematic if the same or similar data dominate [see examples in Vihinen, 2014a]. Predictions from several methods can be combined and has been done in metapredictors like Condel [Gonzalez-Perez and Lopez-Bigas, 2011], PON-P [Olatubosun et al., 2012], MetaSNP [Capriotti et al., 2013], and PredictSNP [Bendl et al., 2014] that have trained ML methods to weigh the outputs of the individual predictors in different ways. Metapredictors should be trained on cases not used for training the constituent methods to avoid overfitting.

Choice of Tools

Whenever assessing variant pathogenicity, start by looking at known variants in databases including LSDBs, ClinVar and HGMD (especially if you have the commercial license), and others (Table 1). Many variants have heterogeneous effects; thus, these databases may not be fully conclusive. When using prediction methods, start with generic tolerance predictors or gene/protein/disease-specific predictors, if available (Supp. Table S1). In the next phase, mechanism/effect-specific tools can be applied to investigate the actual reason for the phenotype. Remember that no method is better than the input data; therefore, be aware of your variation data quality.

Choosing suitable methods can be a difficult task; however, it is important and will pay back. Reliable tools provide the best possible starting point for variation interpretation both for clinical and research purposes. Many tools are constantly being developed; therefore, it is important to follow the updates. Methods that were good a year or two ago may not have cutting-edge performance anymore. The community should not accept any black-box approaches; unfortunately, this is common with commercial solutions [Vihinen,

2015b]. Many companies do not provide sufficient details to allow understanding of how their methods work and what their true performance is. For a checklist of items to consider when choosing methods, see Box 3.

Box 3. Checklist for choosing methods

Adapted from Vihinen [2012].

Items to check when estimating method performance and comparing performance of different methods:

- Is there a comprehensive method description?
- Have established databases and benchmarks been used (if available)? If not, are the used datasets publicly available?
- Is the version of the method mentioned (if several versions exist)?
- Is the contingency table available with all measures and values included?
- Have the developers reported the six performance measures: sensitivity (or true positive rate, TPR), specificity (or true negative rate, TNR), positive predictive value, negative predictive value, accuracy, and Matthews correlation coefficient for binary predictors?
- Has-cross validation or some other partitioning method been used in method testing? Use of independent test dataset is highly recommended.
- Are the training and test datasets disjoint?
- Are the results in balance?
- Has the ROC curve been drawn based on the entire test data? Inspect the ROC curve and AUC.
- How does the method compare with others in all the measures?
- Does the method provide probabilities for predictions?

In addition to choosing the right method, the user has to be able to interpret the obtained results and understand the effect of adjustable program parameters. This means understanding the principles of the algorithm. This is probably the most difficult of the requirements for many biologists and medical scientists; however, they are instrumental and not impossible to master. The performance of the methods should be properly and comprehensively reported in articles. As a user, be sure that the method is applicable for the intended purpose. Use only methods with proven performance, preferably tested on established benchmark datasets or at least tested on the available datasets. Studies addressing method performance are in this respect very useful. Pathogenicity predictors provide one type of data for variants. Carefully consider the importance and relevance of the predictions and combine with other kinds of data, whether clinical signs or laboratory results.

The Way Ahead

Progress in variation interpretation has been fast and is not likely to slow down, especially since the need is growing due to the increased application of sequencing technologies in research and clinics. Gene- and protein-specific tools will likely become more common and they can be trained when more data become available. There will still be a need for accurate tolerance predictors. Benchmark datasets need to be improved and expanded, developed for new application areas, and used for predictor training and testing. Only experimentally validated cases should be used for training. To continue to sell their products, companies need to become more open and unveil details about how their methods have been developed and how they perform on benchmark data.

Predictors should provide information that can be easily used for diagnostic decision making. Integration of various data items from several sources will be one of the themes in the future. This could create additional requirements for method developers, as all the data items may not always be available. Approaches to handle incomplete data have to be devised. Standardization of variation data and annotation has to be improved. The use of HPO and other ontologies for pathogenicity will allow easy utilization of phenotype data and VariO annotations will facilitate generation of accurate datasets for training and test purposes. Other standards are needed and they have to be applied strictly. An example of failed use of a standard data format is the VCF format, of which there are probably dozens of different versions that cause problems when transferring data between programs.

Genome sequencing is increasingly provided as a direct-to-consumer service. There are several problems related to data interpretation and especially in the communication of the significance of the findings. This should be done by professional counsellors, a service that these companies do not provide.

Protein and likely also RNA structural information will be increasingly used for the interpretation of variant effects. There are already tools available, but the throughput speed is usually not suitable for large-scale studies. Structural insight will allow for new ways of understanding variation effects as well as for designing therapies to treat pathologies. For example, many variants that have stable structures affect molecular interactions [Sahni et al., 2015].

Bioinformatic tools are increasingly compiled into pipelines and workflows, such as Galaxy [Goecks et al., 2010], Taverna [Wolstencroft et al., 2013], and Anduril [Ovaska et al., 2010]. They have numerous benefits, especially in routine use. Users also have to be aware of situations when the pipelines are inappropriate to use. Cloud computing has been a buzz word in recent times. For most users of variation interpretation tools, it does not matter where the data are analyzed. But for clinical applications, security is a prime question. Due to potential security and privacy issues with distributed computing, interpretation via cloud services may need to be approached according to local and national laws. Indeed, improvements in online security are being made at constant pace.

Quality of interpretation in prediction algorithms will likely be the issue of greatest importance in the future. In fact, it applies to the whole process of genomic medicine. The smallest possible error rate must be achieved in sequencing, variant calling, annotation, and interpretation. Prediction methods have to be highly accurate, and at the same time very fast, to cope with exome and complete genome sequence data for numerous individuals. Interpretation has become and likely will remain the bottleneck for the use of variation data in clinical practice. Intelligent solutions are already available and will be further improved in the future. Computer-assisted variation interpretation will be part of the toolbox for everyone working on sequencing data generation and interpretation.

Acknowledgments

We thank Gerard Schaafsma for proofreading the manuscript.

Disclosure statement: There is no conflict of interest to declare.

References

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 467:1061–1073.

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 7:248–249.

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

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

Arakawa T, Timasheff SN. 1985. Theory of protein solubility. Methods Enzymol 114:49–77.

Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H. 2000. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16:412–424.

Bao L, Zhou M, Cui Y. 2005. nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. Nucleic Acids Res 33:W480–W482.

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: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:e1003440.

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.

Beroud C, Collod-Beroud G, Boileau C, Soussi T, Junien C. 2000. UMD (Universal Mutation Database): a generic software to build and analyze locus-specific databases. Hum Mutat 15:86–94.

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:e44.

Brookes AJ, Robinson PN. 2015. Human genotype-phenotype databases: aims, challenges and opportunities. Nat Rev Genet 16:702–715.

Bucciantini M, Calloni G, Chiti F, Formigli L, Nosi D, Dobson CM, Stefani M. 2004. Prefibrillar amyloid protein aggregates share common features of cytotoxicity. J Biol Chem 279:31374–31382.

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

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 30:1237–1244.

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 98:310–317.

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

Capriotti E, Fariselli P, Casadio R. 2004. A neural-network-based method for predicting protein stability changes upon single point mutations. Bioinformatics 20:i63–i68.

Capriotti E, Fariselli P, Casadio R. 2005. I-mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res 33(Web Server Issue):W306–W310.

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

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

Cartegni L, Chew SL, Krainer AR. 2002. Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet 3:285–298.

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

Celli J, Dalgleish R, Vihinen M, Taschner PE, den Dunnen JT. 2012. Curating gene variant databases (LSDBs): toward a universal standard. Hum Mutat 33:291–297.

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

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 29:852–860.

Chen CW, Lin J, Chu YW. 2013. iStable: off-the-shelf predictor integration for predicting protein stability changes. BMC Bioinformatics 14:S5.

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

Cheng J, Randall A, Baldi P. 2006. Prediction of protein stability changes for single-site mutations using support vector machines. Proteins 62:1125–1132.

Chiti F, Webster P, Taddei N, Clark A, Stefani M, Ramponi G, Dobson CM. 1999. Designing conditions for in vitro formation of amyloid protofilaments and fibrils. Proc Natl Acad Sci U S A 96:3590–3594.

Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. 2012. Predicting the functional effect of amino acid substitutions and indels. PLoS One 7: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) 6:80–92.

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

Ciriello G, Cerami E, Sander C, Schultz N. 2012. Mutual exclusivity analysis identifies oncogenic network modules. Genome Res 22:398–406.

Cline MS, Karchin R. 2011. Using bioinformatics to predict the functional impact of SNVs. Bioinformatics 27:441–448.

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 GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 15:901–913.

Cotton RG, Al Aqeel AI, Al-Mulla F, Carrera P, Claustres M, Ekong R, Hyland VJ, Macrae FA, Marafie MJ, Paalman MH, Patrinos GP, Qi M, et al. 2009. Capturing all disease-causing mutations for clinical and research use: toward an effortless system for the Human Variome Project. Genet Med 11:843–849.

Cotton RG, Auerbach AD, Brown AF, Carrera P, Christodoulou J, Claustres M, Compton J, Cox DW, De Baere E, den Dunnen JT, Greenblatt M, Fujiwara M, et al. 2007. A structured simple form for ordering genetic tests is needed to ensure coupling of clinical detail (phenotype) with DNA variants (genotype) to ensure utility in publication and databases. Hum Mutat 28:931–932.

Cotton RG, Sallee C, Knoppers BM. 2005. Locus-specific databases: from ethical principles to practice. Hum Mutat 26:489–493.

Dalgleish R, Flicek P, Cunningham F, Astashyn A, Tully RE, Proctor G, Chen Y, McLaren WM, Larsson P, Vaughan BW, Beroud C, Dobson G, et al. 2010. Locus reference genomic sequences: an improved basis for describing human DNA variants. Genome Med 2:24.

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:e1003382.

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 22:1589–1598.

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 25:2537–2543.

den Dunnen JT, Antonarakis SE. 2001. Nomenclature for the description of human sequence variations. Hum Genet 109:121–124.

den Dunnen JT, Sijmons RH, Andersen PS, Vihinen M, Beckmann JS, Rossetti S, Talbot CC Jr., Hardison RC, Povey S, Cotton RG. 2009. Sharing data between LSDBs and central repositories. Hum Mutat 30:493–495.

Desmet FO, Hamroun D, Collod-Beroud G, Claustres M, Beroud 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. pp. 1–14.

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

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

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 37:28–35.

Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z. 2002. Intrinsic disorder and protein function. Biochemistry 41:6573–6582.

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 109:9551–9556.

Emanuelsson O, Brunak S, von Heijne G, Nielsen H. 2007. Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2:953–971.

Emily M, Talvas A, Delamarche C. 2013. MetAmyl: a METa-predictor for AMYLoid proteins. PLoS One 8:e79722.

Fairbrother WG, Yeh RF, Sharp PA, Burge CB. 2002. Predictive identification of exonic splicing enhancers in human genes. Science 297:1007–1013.

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

Fechter K, Porollo A. 2014. MutaCYP: classification of missense mutations in human cytochromes P450. BMC Med Genomics 7:47.
Feenstra I, Fang J, Koolen DA, Siezen A, Evans C, Winter RM, Lees MM, Riegel M, de Vries BB, Van Ravenswaaij CM, Schinzel A. 2006. European Cytogeneticists Association Register of Unbalanced Chromosome Aberrations (ECARUCA); an online database for rare chromosome abnormalities. Eur J Med Genet 49:279–291.
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 22:1302–1306.
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 315:771–786.
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 32:557–563.
Folkman L, Stantic B, Sattar A. 2014. Feature-based multiple models improve classification of mutation-induced stability changes. BMC Genomics 15:S6.
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:D945–D950.
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 493:216–220.
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 4:177–183.
Garbuzynskiy SO, Lobanov MY, Galzitskaya OV. 2010. FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. Bioinformatics 26:326–332.
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 28:554–562.
Gilis D, Rooman M. 2000. Popmusic, an algorithm for predicting protein mutant stability changes: application to prion proteins. Protein Eng 13:849–856.
Giollo M, Martin AJ, Walsh I, Ferrari C, Tosatto SC. 2014. Neemo: a method using residue interaction networks to improve prediction of protein stability upon mutation. BMC Genomics 15 Suppl 4:S7.
Gnad F, Baucom A, Mukhyala K, Manning G, Zhang Z. 2013. Assessment of computational methods for predicting the effects of missense mutations in human cancers. BMC Genomics 14 Suppl 3:S7.
Goecks J, Nekrutenko A, Taylor J. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11:R86.
Goldgar DE, Easton DF, Byrnes GB, Spurdle AB, Iversen ES, Greenblatt MS. 2008. Genetic evidence and integration of various data sources for classifying uncertain variants into a single model. Hum Mutat 29:1265–1272.
Goldgar DE, Easton DF, Deffenbaugh AM, Monteiro AN, Tavtigian SV, Couch FJ. 2004. Integrated evaluation of DNA sequence variants of unknown clinical significance: application to BRCA1 and BRCA2. Am J Hum Genet 75:535–544.
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 4:89.
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 88:440–449.
Gonzalez-Perez A, Lopez-Bigas N. 2012. Functional impact bias reveals cancer drivers. Nucleic Acids Res 40: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 10:723–729.
Goren A, Ram O, Amit M, Keren H, Lev-Maor G, Vig I, Pupko T, Ast G. 2006. Comparative analysis identifies exonic splicing regulatory sequences—the complex definition of enhancers and silencers. Mol Cell 22:769–781.
Gray KA, Yates B, Seal RL, Wright MW, Bruford EA. 2015. Genenames.org: the HGNC resources in 2015. Nucleic Acids Res 43:D1079–D1085.
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 36:513–523.
Grodecka L, Lockerova P, Ravcukova B, Buratti E, Baralle FE, Dusek L, Freiberger T. 2014. Exon first nucleotide mutations in splicing: evaluation of in silico prediction tools. PLoS One 9:e89570.
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 320:369–387.
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 (Database issue):D514–D517.
Hebsgaard SM, Korning PG, Tolstrup N, Engelbrecht J, Rouze P, Brunak S. 1996. Splice site prediction in arabidopsis thaliana pre-mRNA by combining local and global sequence information. Nucleic Acids Res 24:3439–3452.
Hodgkinson A, Eyre-Walker A. 2011. Variation in the mutation rate across mammalian genomes. Nat Rev Genet 12:756–766.
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 150:251–263.
Holla OL, Nakken S, Mattingsdal M, Ranheim T, Berge KE, Defesche JC, Leren TP. 2009. Effects of intronic mutations in the LDLR gene on pre-mRNA splicing: comparison of wet-lab and bioinformatics analyses. Mol Genet Metab 96:245–252.
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 30:99–106.
Hornbeck PV, Kornhauser JM, Tkachev S, Zhang B, Skrzypek E, Murray B, Latham V, Sullivan M. 2012. PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic Acids Res 40:D261–D270.
Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K. 2007. WoLFPSORT: protein localization predictor. Nucleic Acids Res 35:W585–W587.
Hou JP, Ma J. 2014. DawnRank: discovering personalized driver genes in cancer. Genome Med 6:56.
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 33:1228–1238.
Houdayer C, Dehainault C, Mattler C, Michaux D, Caux-Moncoutier V, Pages-Berhouet S, d'Enghien CD, Lauge A, Castera L, Gauthier-Villars M, Stoppa-Lyonnet D. 2008. Evaluation of in silico splice tools for decision-making in molecular diagnosis. Hum Mutat 29:975–982.
Hu J, Ng PC. 2012. Predicting the effects of frameshifting indels. Genome Biol 13:R9.
Hu J, Ng PC. 2013. SIFT indel: predictions for the functional effects of amino acid insertions/deletions in proteins. PLoS One 8: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 93:439–451.
Huang LT, Gromiha MM. 2009. Reliable prediction of protein thermostability change upon double mutation from amino acid sequence. Bioinformatics 25:2181–2187.
Hunt RC, Simhadri VL, Iandoli M, Sauna ZE, Kimchi-Sarfaty C. 2014. Exposing synonymous mutations. Trends Genet 30:308–321.
Jia P, Zhao Z. 2014. VarWalker: personalized mutation network analysis of putative cancer genes from next-generation sequencing data. PLoS Comput Biol 10:e1003460.
Jian X, Boerwinkle E, Liu X. 2014. In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Res 42:13534–13544.
Johansen MB, Izarzugaza JM, Brunak S, Petersen TN, Gupta R. 2013. Prediction of disease causing non-synonymous snps by the artificial neural network predictor NetDiseaseSNP. PLoS One 8:e68370.
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 88:183–192.
Kaminker JS, Zhang Y, Watanabe C, Zhang Z. 2007. CanPredict: a computational tool for predicting cancer-associated missense mutations. Nucleic Acids (Web Server Issue):W595–W598.
Karchin R. 2009. Next generation tools for the annotation of human SNPs. Brief Bioinform 10:35–52.
Karchin R, Kelly L, Sali A. 2005. Improving functional annotation of non-synonomous SNPs with information theory. Pac Symp Biocomput 10:397–408.
Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. Genome Res 12:996–1006.
Khan S, Vihinen M. 2010. Performance of protein stability predictors. Hum Mutat 31:675–684.
Kim YA, Cho DY, Dao P, Przytycka TM. 2015. MEMCover: integrated analysis of mutual exclusivity and functional network reveals dysregulated pathways across multiple cancer types. Bioinformatics 31:i284–292.
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 46:310–315.
Kohonen-Corish MR, Al-Aama JY, Auerbach AD, Axton M, Barash CI, Bernstein I, Beroud C, Burn J, Cunningham F, Cutting GR, den Dunnen JT, Greenblatt MS, et al. 2010. How to catch all those mutations—the report of the third Human Variome Project Meeting, UNESCO Paris, May 2010. Hum Mutat 31:1374–1381.

Kondrashov FA. 2005. Prediction of pathogenic mutations in mitochondrially encoded human tRNAs. Hum Mol Genet 14:2415–2419.
Korthauer KD, Kendziorski C. 2015. MADGiC: a model-based approach for identifying driver genes in cancer. Bioinformatics 31:1526–1535.
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 28:150–158.
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 15:9670–9717.
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–D206.
Laimer J, Hofer H, Fritz M, Wegenkittl S, Lackner P. 2015. MAESTRO—multi agent stability prediction upon point mutations. BMC Bioinformatics 16:116.
Lancaster O, Beck T, Atlan D, Swertz M, Thangavelu D, Veal C, Dalgleish R, Brookes AJ. 2015. Cafe Variome: general-purpose software for making genotype-phenotype data discoverable in restricted or open access contexts. Hum Mutat 36:957–964.
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:D980–D985.
Lappalainen I, Lopez J, Skipper L, Hefferon T, Spalding JD, Garner J, Chen C, Maguire M, Corbett M, Zhou G, Paschall J, Ananiev V, et al. 2013. DbVar and DGVa: public archives for genomic structural variation. Nucleic Acids Res 41:D936–D941.
Laurila K, Vihinen M. 2009. Prediction of disease-related mutations affecting protein localization. BMC Genomics 10:122.
Laurila K, Vihinen M. 2011. PROlocalizer: Integrated web service for protein subcellular localization prediction. Amino Acids 40:975–980.
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 499:214–218.
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 47:955–961.
Leiserson MD, Blokh D, Sharan R, Raphael BJ. 2013. Simultaneous identification of multiple driver pathways in cancer. PLoS Comput Biol 9:e1003054.
Leiserson MD, Wu HT, Vandin F, Raphael BJ. 2015. CoMEt: a statistical approach to identify combinations of mutually exclusive alterations in cancer. Genome Biol 16:160.
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 25:2744–2750.
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 33:8–21.
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 37:235–241.
Lopes P, Dalgleish R, Oliveira JL. 2011. WAVe: web analysis of the variome. Hum Mutat 32:729–734.
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.
Lu S, Lu KN, Cheng SY, Hu B, Ma X, Nystrom N, Lu X. 2015. Identifying driver genomic alterations in cancers by searching minimum-weight, mutually exclusive sets. PLoS Comput Biol 11:e1004257.
MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. 2014. The database of genomic variants: a curated collection of structural variation in the human genome. Nucleic Acids Res 42(Database issue):D986–D992.
Macintyre G, Bailey J, Haviv I, Kowalczyk A. 2010. is-rSNP: a novel technique for in silico regulatory SNP detection. Bioinformatics 26:i524–i530.
Manke T, Heinig M, Vingron M. 2010. Quantifying the effect of sequence variation on regulatory interactions. Hum Mutat 31:477–483.
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:e77945.
Martelotto LG, Ng CK, De Filippo MR, Zhang Y, Piscuoglio S, Lim RS, Shen R, Norton L, Reis-Filho JS, Weigelt B. 2014. Benchmarking mutation effect prediction algorithms using functionally validated cancer-related missense mutations. Genome Biol 15:484.
Martinez-Fundichely A, Casillas S, Egea R, Ramia M, Barbadilla A, Pantano L, Puig M, Caceres M. 2014. InvFEST, a database integrating information of polymorphic inversions in the human genome. Nucleic Acids Res 42:D1027–D1032.
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 33:1267–1274.
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 23:683–687.
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 24:2–5.
Maurer-Stroh S, Debulpaep M, Kuemmerer N, Lopez de la Paz M, Martins IC, Reumers J, Morris KL, Copland A, Serpell L, Serrano L, Schymkowitz JW, Rousseau F. 2010. Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. Nat Methods 7:237–242.
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 26:2069–2070.
Melloni GE, Ogier AG, de Pretis S, Mazzarella L, Pelizzola M, Pelicci PG, Riva L. 2014. DOTS-Finder: a comprehensive tool for assessing driver genes in cancer genomes. Genome Med 6:44.
Miller CA, Settle SH, Sulman EP, Aldape KD, Milosavljevic A. 2011. Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. BMC Med Genomics 4:34.
Miller MP, Kumar S. 2001. Understanding human disease mutations through the use of interspecific genetic variation. Hum Mol Genet 10:2319–2328.
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:E5189–E5198.
Mitropoulou C, Webb AJ, Mitropoulos K, Brookes AJ, Patrinos GP. 2010. Locus-specific database domain and data content analysis: evolution and content maturation toward clinical use. Hum Mutat 31:1109–1116.
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:R19.
Nair PS, Vihinen M. 2013. VariBench: a benchmark database for variations. Hum Mutat 34:42–49.
Nalla VK, Rogan PK. 2005. Automated splicing mutation analysis by information theory. Hum Mutat 25:334–342.
Ng PC, Henikoff S. 2001. Predicting deleterious amino acid substitutions. Genome Res 11:863–874.
Niroula A, Urolagin S, Vihinen M. 2015. PON-P2: prediction method for fast and reliable identification of harmful variants. PLoS One 10:e0117380.
Niroula A, Vihinen M. 2015a. Classification of amino acid substitutions in mismatch repair proteins using PON-MMR2. Hum Mutat 36:1128–1134.
Niroula A, Vihinen M. 2015b. Harmful somatic amino acid substitutions affect key pathways in cancers. BMC Med Genomics 8:53.
Niroula A, Vihinen M. 2016. PON-mt-tRNA: a multifactorial probability-based method for classification of mitochondrial tRNA variations. Nucleic Acids Res 44:2020–2027.
O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, Bodily P, Tian L, Hakonarson H, Johnson WE, Wei Z, Wang K, et al. 2013. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med 5:28.
Olatubosun A, Väliaho J, Härkönen J, Thusberg J, Vihinen M. 2012. PON-P: integrated predictor for pathogenicity of missense variants. Hum Mutat 33:1166–1174.
Olivier M, Petitjean A, Teague J, Forbes S, Dunnick JK, den Dunnen JT, Langerod A, Wilkinson JM, Vihinen M, Cotton RG, Hainaut P. 2009. Somatic mutation databases as tools for molecular epidemiology and molecular pathology of cancer: proposed guidelines for improving data collection, distribution, and integration. Hum Mutat 30:275–282.
Ollila J, Lappalainen I, Vihinen M. 1996. Sequence specificity in CpG mutation hotspots. FEBS Lett 396:119–122.
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 25:435–442.
Ovaska K, Laakso M, Haapa-Paananen S, Louhimo R, Chen P, Aittomäki V, Valo E, Núñez-Fontarnau J, Rantanen V, Karinen S, Nousiainen K, Lahesmaa-Korpinen AM, et al. 2010. Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. Genome Med 2:65.
Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 40:1413–1415.

Papadopoulos P, Viennas E, Gkantouna V, Pavlidis C, Bartsakoulia M, Ioannou ZM, Ratbi I, Sefiani A, Tsaknakis J, Poulas K, Tzimas G, Patrinos GP. 2014. Developments in FINDbase worldwide database for clinically relevant genomic variation allele frequencies. Nucleic Acids Res 42:D1020–D1026.

Parthiban V, Gromiha MM, Schomburg D. 2006. CUPSAT: prediction of protein stability upon point mutations. Nucleic Acids Res 34:W239–W242.

Patrinos GP. 2006. National and ethnic mutation databases: recording populations' genography. Hum Mutat 27:879–887.

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

Perniola R, Musco G. 2014. The biophysical and biochemical properties of the autoimmune regulator (AIRE) protein. Biochim Biophys Acta 1842:326–337.

Pertea M, Lin X, Salzberg SL. 2001. GeneSplicer: a new computational method for splice site prediction. Nucleic Acids Res 29:1185–1190.

Peterson TA, Doughty E, Kann MG. 2013. Towards precision medicine: advances in computational approaches for the analysis of human variants. J Mol Biol 425:4047–4063.

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

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:W314–W319.

Plon SE, Eccles DM, Easton D, Foulkes WD, Genuardi M, Greenblatt MS, Hogervorst FB, Hoogerbrugge N, Spurdle AB, Tavtigian SV. 2008. Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. Hum Mutat 29:1282–1291.

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 22:553–560.

Poussu E, Vihinen M, Paulin L, Savilahti H. 2004. Probing the alpha-complementing domain of E. Coli beta-galactosidase with use of an insertional pentapeptide mutagenesis strategy based on Mu in vitro DNA transposition. Proteins 54:681–692.

Povey S, Al Aqeel AI, Cambon-Thomsen A, Dalgleish R, den Dunnen JT, Firth HV, Greenblatt MS, Barash CI, Parker M, Patrinos GP, Savige J, Sobrido MJ, et al. 2010. Practical guidelines addressing ethical issues pertaining to the curation of human locus-specific variation databases (LSDBs). Hum Mutat 31:1179–1184.

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

Reimand J, Bader GD. 2013. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. Mol Syst Biol 9:637.

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 17:405–424.

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

Robinson PN, Kohler S, Bauer S, Seelow D, Horn D, Mundlos S. 2008. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am J Hum Genet 83:610–615.

Rogan PK, Zou GY. 2013. Best practices for evaluating mutation prediction methods. Hum Mutat 34:1581–1582.

Safaei J, Manuch J, Gupta A, Stacho L, Pelech S. 2011. Prediction of 492 human protein kinase substrate specificities. Proteome Sci 9 Suppl 1:S6.

Sahni N, Yi S, Taipale M, Fuxman Bass JI, Coulombe-Huntington J, Yang F, Peng J, Weile J, Karras GI, Wang Y, Kovacs IA, Kamburov A, et al. 2015. Widespread macromolecular interaction perturbations in human genetic disorders. Cell 161:647–660.

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 12:683–691.

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

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

Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. 2005. The FoldX web server: an online force field. Nucleic Acids Res 33:W382–W388.

Shapiro MB, Senapathy P. 1987. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. Nucleic Acids Res 15:7155–7174.

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 17:267–276.

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 29:308–311.

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

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.

Sloan CA, Chan ET, Davidson JM, Malladi VS, Strattan JS, Hitz BC, Gabdank I, Narayanan AK, Ho M, Lee BT, Rowe LD, Dreszer TR, et al. 2016. ENCODE data at the ENCODE portal. Nucleic Acids Res 44:D726–D732.

Smith PJ, Zhang C, Wang J, Chew SL, Zhang MQ, Krainer AR. 2006. An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. Hum Mol Genet 15:2490–2508.

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

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 14:951–955.

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

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

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 133:1–9.

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

Storz JF, Zera AJ. 2011. Experimental approaches to evaluate the contributions of candidate protein-coding mutations to phenotypic evolution. Methods Mol Biol 772:377–396.

Sunyaev SR, Eisenhaber F, Rodchenkov IV, Eisenhaber B, Tumanyan VG, Kuznetsov EN. 1999. PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. Protein Eng 12:387–394.

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

Tartaglia GG, Pawar AP, Campioni S, Dobson CM, Chiti F, Vendruscolo M. 2008. Prediction of aggregation-prone regions in structured proteins. J Mol Biol 380:425–436.

Taschner PE, den Dunnen JT. 2011. Describing structural changes by extending HGVS sequence variation nomenclature. Hum Mutat 32:507–511.

Teng S, Srivastava AK, Wang L. 2010. Sequence feature-based prediction of protein stability changes upon amino acid substitutions. BMC Genomics 11:S5.

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

Thery JC, Krieger S, Gaildrat P, Revillion F, Buisine MP, Killian A, Duponchel C, Rousselin A, Vaur D, Peyrat JP, Berthet P, Frebourg T, et al. 2011. Contribution of bioinformatics predictions and functional splicing assays to the interpretation of unclassified variants of the BRCA genes. Eur J Hum Genet 19:1052–1058.

Thomas PD, Kejariwal A. 2004. Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects. Proc Natl Acad Sci U S A 101:15398–15403.

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 34:200–209.

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 34:255–265.

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 46:107–115.

Thorisson GA, Muilu J, Brookes AJ. 2009. Genotype-phenotype databases: challenges and solutions for the post-genomic era. Nat Rev Genet 10:9–18.

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

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

Tian R, Basu MK, Capriotti E. 2014. ContrastRank: a new method for ranking putative cancer driver genes and classification of tumor samples. Bioinformatics 30:i572–i578.
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.
Torkamani A, Schork NJ. 2007. Accurate prediction of deleterious protein kinase polymorphisms. Bioinformatics 23:2918–2925.
Uniprot Consortium. 2015. UniProt: a hub for protein information. Nucleic Acids Res 43:D204–D212.
Vacic V, Iakoucheva LM. 2012. Disease mutations in disordered regions—exception to the rule? Mol Biosyst 8:27–32.
Wagih O, Reimand J, Bader GD. 2015. MIMP: predicting the impact of mutations on kinase-substrate phosphorylation. Nat Methods 12:531–533.
Walsh I, Seno F, Tosatto SC, Trovato A. 2014. PASTA 2.0: an improved server for protein aggregation prediction. Nucleic Acids Res 42:W301–W307.
van Baal S, Zlotogora J, Lagoumintzis G, Gkantouna V, Tzimas I, Poulas K, Tsakalidis A, Romeo G, Patrinos GP. 2010. ETHNOS: a versatile electronic tool for the development and curation of national genetic databases. Hum Genomics 4:361–368.
Van den Eynden J, Fierro AC, Verbeke LP, Marchal K. 2015. SomInaClust: detection of cancer genes based on somatic mutation patterns of inactivation and clustering. BMC Bioinformatics 16:125.
Vandin F, Upfal E, Raphael BJ. 2012. De novo discovery of mutated driver pathways in cancer. Genome Res 22:375–385.
Wang Z, Moult J. 2001. SNPs, protein structure, and disease. Hum Mutat 17:263–270.
Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, Burge CB. 2004. Systematic identification and analysis of exonic splicing silencers. Cell 119:831–845.
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 37:36–42.
Wei Q, Dunbrack RL Jr. 2013. The role of balanced training and testing data sets for binary classifiers in bioinformatics. PLoS One 8:e67863.
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: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 31:28–33.
Vihinen M. 2012. How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. BMC Genomics 13:S2.
Vihinen M. 2013. Guidelines for reporting and using prediction tools for genetic variation analysis. Hum Mutat 34:275–282.
Vihinen M. 2014a. Majority vote and other problems when using computational tools. Hum Mutat 35:912–914.
Vihinen M. 2014b. Variation ontology for annotation of variation effects and mechanisms. Genome Res 24:356–364.
Vihinen M. 2015a. Muddled genetic terms miss and mess the message. Trends Genet 31:423–425.
Vihinen M. 2015b. No more hidden solutions in bioinformatics. Nature 521:261.
Vihinen M. 2015c. Types and effects of protein variations. Hum Genet 134:405–421.
Vihinen M, den Dunnen JT, Dalgleish R, Cotton RG. 2012. Guidelines for establishing locus specific databases. Hum Mutat 33:298–305.
Vihinen M, Hancock JM, Maglott DR, Landrum MJ, Schaafsma GC, Taschner P. 2016. Human Variome Project Quality Assessment Criteria for variation databases. Hum Mutat.[Epub ahead of print]
Wildeman M, van Ophuizen E, den Dunnen JT, Taschner PE. 2008. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. Hum Mutat 29:6–13.
Vitkup D, Sander C, Church GM. 2003. The amino-acid mutational spectrum of human genetic disease. Genome Biol 4:R72.
Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr., Kinzler KW. 2013. Cancer genome landscapes. Science 339:1546–1558.
Wolstencroft K, Haines R, Fellows D, Williams A, Withers D, Owen S, Soiland-Reyes S, Dunlop I, Nenadic A, Fisher P, Bhagat J, Belhajjame K, et al. 2013. The Taverna workflow suite: designing and executing workflows of web services on the desktop, web or in the cloud. Nucleic Acids Res 41:W557–W561.
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 27:2147–2148.
Woolfe A, Mullikin JC, Elnitski L. 2010. Genomic features defining exonic variants that modulate splicing. Genome Biol 11:R20.
Vreeswijk MP, Kraan JN, van der Klift HM, Vink GR, Cornelisse CJ, Wijnen JT, Bakker E, van Asperen CJ, Devilee P. 2009. Intronic variants in BRCA1 and BRCA2 that affect RNA splicing can be reliably selected by splice-site prediction programs. Hum Mutat 30:107–114.
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 31:2748–2750.
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 36:638–647.
Yang H, Wang K. 2015. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. Nat Protoc 10:1556–1566.
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 44:847–855.
Yang Y, Niroula A, Shen B, Vihinen M. 2016. PON-Sol: prediction of effects of amino acid substitutions on protein solubility. Bioinformatics. doi: 10.1093/bioinformatics/btw066.
Yates CM, Sternberg MJ. 2013. The effects of non-synonymous single nucleotide polymorphisms (nsSNPs) on protein-protein interactions. J Mol Biol 425:3949–3963.
Yeo G, Burge CB. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol 11:377–394.
Yip YL, Famiglietti M, Gos A, Duek PD, David FP, Gateau A, Bairoch A. 2008. Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase. Hum Mutat 29:361–366.
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:W306–W313.
Zhang J, Wu LY, Zhang XS, Zhang S. 2014. Discovery of co-occurring driver pathways in cancer. BMC Bioinformatics 15:271.
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:R23.
Zhou H, Zhou Y. 2002. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 11:2714–2726.
Zhou J, Troyanskaya OG. 2015. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods 12:931–934.
Zia A, Moses AM. 2011. Ranking insertion, deletion and nonsense mutations based on their effect on genetic information. BMC Bioinformatics 12:299.