VARIATION BENCHMARK DATASETS: UPDATE, CRITERIA, QUALITY AND APPLICATIONS

Short title: Variation benchmark datasets

Anasua Sarkar[1], Yang Yang[2,3] and Mauno Vihinen[1]

[1]Department of Experimental Medical Science, BMC B13, Lund University, SE-22 184 Lund, Sweden

[2]School of Computer Science and Technology, Soochow University, China

[3]Provincial Key Laboratory for Computer Information Processing Technology, Soochow University, China

Correspondence to Mauno Vihinen, Department of Experimental Medical Science, BMC B13, Lund University, SE-22 184 Lund, Sweden

ABSTRACT

Development of new computational methods and testing their performance has to be done on experimental data. Only in comparison to existing knowledge can method performance be assessed. For that purpose, benchmark datasets with known and verified outcome are needed. High-quality benchmark datasets are valuable and may be difficult, laborious and time consuming to generate. VariBench and VariSNP are the two existing databases for sharing variation benchmark datasets. They have been used for training and benchmarking predictors for various types of variations and their effects. There are 419 new datasets from 109 papers containing altogether 329003373 variants; however there is plenty of redundancy between the datasets. VariBench is freely available at http://structure.bmc.lu.se/VariBench/. The contents of the datasets vary depending on information in the original source. The available datasets have been categorized into 20 groups and subgroups. There are datasets for insertions and deletions, substitutions in coding and non-coding region, structure mapped, synonymous and benign variants. Effect-specific datasets include DNA regulatory elements, RNA splicing, and protein property predictions for aggregation, binding free energy, disorder and stability. Then there are several datasets for molecule-specific and disease-specific applications, as well as one dataset for variation phenotype effects. Variants are often described at three molecular levels (DNA, RNA and protein) and sometimes also at the protein structural level including relevant cross references and variant descriptions. The updated VariBench facilitates development and testing of new methods and comparison of obtained performance to previously published methods. We compared the performance of the pathogenicity/tolerance predictor PON-P2 to several benchmark studies, and showed that such comparisons are feasible and useful, however, there may be limitations due to lack of provided details and shared data.

AUTHOR SUMMARY

A prediction method performance can only be assessed in comparison to existing knowledge. For that purpose benchmark datasets with known and verified outcome are needed. High-quality benchmark datasets are valuable and may be difficult, laborious and time consuming to generate. We collected variation datasets from literature, website and databases. There are 419 separate new datasets, which however contain plenty of redundancy. VariBench is freely available at http://structure.bmc.lu.se/VariBench/. There are datasets for insertions and deletions, substitutions in coding and non-coding region, structure mapped, synonymous and benign variants. Effect-specific datasets include DNA regulatory elements, RNA splicing, and protein property predictions for aggregation, binding free energy, disorder and stability. Then there are several datasets for molecule-specific and disease-specific applications, as well as one dataset for variation phenotype effects. The updated VariBench facilitates development and testing of new methods and comparison of obtained performance to previously published methods. We compared the performance of the pathogenicity/tolerance predictor PON-P2 to several benchmark studies and showed that such comparisons are possible and useful when the details of studies and the datasets are shared.

INTRODUCTION

Development and testing of computational methods are dependent on experimental data. Only in comparison to existing knowledge can method performance be assessed. For that purpose, benchmark datasets with known and verified outcome are needed. During the last few years, such datasets have been collected for a number of applications in the field of variation interpretation. VariBench [1] and VariSNP [2] are the two existing databases for variation benchmark datasets. VariBench contains all kinds of datasets while VariSNP is a dedicated resource for variation sets from dbSNP database for short variations [3].

Benchmark datasets are used both for method training and testing. We can divide testing approaches into three categories (Figure 1). The most reliable are systematic benchmark studies. Quite often the initial method performance assessment is done on somewhat limited test data or not reporting all necessary measures. The third group includes studies for initial method and hypothesis testing typically with a limited amount of data. An example for this kind of testing is Critical Assessment of Genome Interpretation (CAGI, https://genomeinterpretation.org/), which has organized several challenges for method developers. These contests with blind data, when the participants do not know the true answer, have been important e.g. for testing new ideas and methods, as well for tackling novel application areas.

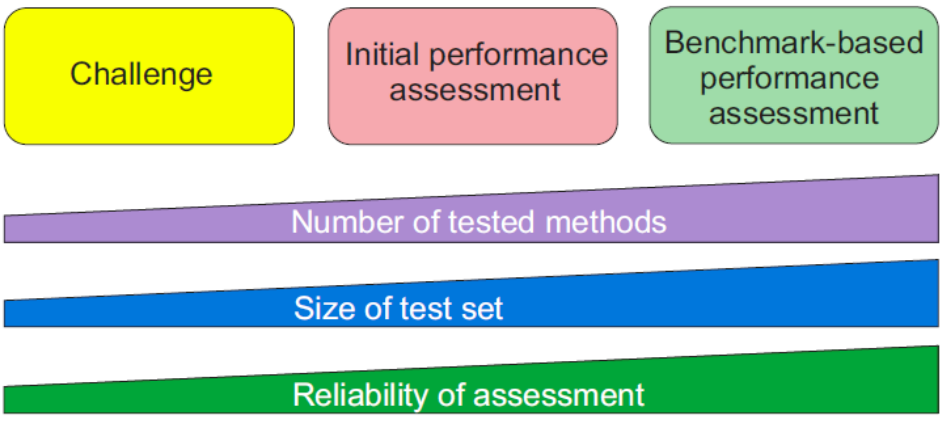

Figure 1. Types of method performance tests. The figure is adapted from [34].

High-quality benchmark datasets are valuable and may be difficult, laborious and time consuming to generate. Already from the point of view of reasonable use of resources it is important to share such datasets. Secondly, comparison of method performance is reliable only when using the same test dataset. According to the FAIR principles [4], research data should be made findable, accessible, interoperable and re-usable. VariBench and VariSNP provide variation data according to these principles.

It is still quite common that authors collect and use extensive datasets for their published papers, but do not share and make the datasets available. This prevents others from comparing additional tools to those used in the paper. Even when the data is made available, it may be in a format that makes re-use practically impossible. An example is the datasets used for testing the MutationTaster2 tolerance predictor [5]. They were published as figures and at barely legible resolution. Now, these datasets are available in VariBench.

CRITERIA FOR BENCHMARKS

We defined criteria for a benchmark when the VariBench database was first published [1]. These criteria were more extensive than previously used and have been found very useful and still form the basis for inclusion of data and for their representation in VariBench. The criteria are as follows.

Relevance. The dataset has to capture the characteristics of the investigated property. Not all available data may be relevant for the phenomenon or may be only indirectly related to it. The collected cases have to be for the specific effect or mechanism under study.

Representativeness. The datasets should cover the event space as well as possible, thus preferably containing examples from all the regions relevant to the effect. The actual number of cases for achieving this coverage may vary widely depending on the effect. The dataset should be of sufficient size to allow statistical studies but may not need to include all known instances.

Non-redundancy. This means excluding overlapping cases.

Experimentally verified cases. Method performance comparisons have to be based on experimental data, not on predictions, otherwise the comparison will be about the congruence of methods, not about their true performance.

Positive and negative cases. Comprehensive assessment has to be based both on positive (showing the investigated feature) and negative (not having effect) cases.

Scalability. It should be possible to test systems of different sizes.

Reusability. As datasets are expensive to generate they should be shared in such a way that they can be used for other investigations. This may mean similar applications or usage in new areas.

Most of the criteria are rather easy to fulfil, but some others are more difficult to take into account. We recently investigated the representativeness of 24 tolerance datasets from VariBench in the human protein universe by analysing the distribution and coverage of cases in chromosomes, protein structures, CATH domains and classes, Pfam families, Enzyme Commission (EC) categories and Gene Ontology annotations [6]. The outcome was that none of the datasets were well representative. When correlating the training data representativeness to the performance of predictors based on them, no clear correlation was found. However, it is apparent that representative training data would allow training of methods that have good performance for cases distributed throughout the event space.

Benchmark studies in relation to variation predictions have been made for variants affecting protein stability [7, 8], protein substitution tolerance/pathogenicity [9-14], protein localization [15], protein disorder [16], protein solubility [17], benign variants [18], transmembrane proteins [19], alternative splicing [20, 21] and phenotypes of amino acid substitutions [22]. Many of the datasets used in these studies are available for verification and reuse, but unfortunately e.g. the last one, which is unique, is not accessible.

To test relevance of the tolerance datasets, we investigated how many disease-causing variations could be found from neutral training data. A small number of such variants were found, 1.13 to 1.77 % [6]. These numbers are so small that they do not have a major impact on method performances. VariBench datasets are reusable and scalable, contain experimental cases, and are typically non-redundant. However, how redundancy should be defined may depend on the application. For example, when using domain features in variant predictors, variants even in related domain members would be redundant.

DATASET QUALITY

The quality of benchmark datasets is of utmost significance. This is naturally dependent on the quality of the data sources. There are not many quality schemes in this field. For locus specific variation databases (LSDBs) there is a quality scheme that contains close to 50 criteria in four main areas including database quality, technical quality, accessibility and timeliness [23]. However, these guidelines are not yet widely followed and similar criteria are missing for other types of variation data resources.

Systematics within datasets and databases can significantly improve their quality and usability. For variation data there are a number of systematics solutions available. These include systematic gene names available for human from the HUGO Gene Nomenclature Committee (HGNC) [24], Human Genome Variation Society (HGVS) variation nomenclature [25], Locus Reference Genomic (LRG) and [26] RefSeq reference sequences [27], and Variation Ontology (VariO) variation type, effect and mechanism annotations [28].

Quality relates to numerous aspects in the datasets, the correctness of variation and gene/protein and disease information, relevance of references, etc. We recently selected cases from ProTherm [29] to build an unbiased dataset for the protein variant stability predictor PON-tstab [30]. We were aware that the database had some problems, however, were surprised with the extent of problematic cases. While making the selection, we noticed numerous issues, such as cases of two-stage denaturation pathways where values for all the

steps and then the total value were provided; there were errors in sequences, variants, recorded measuring temperatures, ΔΔG values and their signs and units, and in indicated PDB structures; and so on. The uncorrected and wrong data have been used for development of tens of prediction methods. This is probably an extreme exception (ProTherm was taken away from the internet after our paper was published); however, this indicates that one has to be careful even when using popular data. VariBench has several quality controls, but lists also datasets that may contain problems e.g. numerous ProTherm sub-selections that have been published and sometimes used in several papers. They are included for comparative purposes.

HOW TO TEST PREDICTOR PERFORMANCE

The use of a benchmark dataset is just one of the requirements for systematic method performance assessment. Proper measures are needed to find out the qualities of performance. Most of the currently available prediction methods are binary, distributing cases into two categories. There are guidelines for how to test and report method performance [31-33]. There is also a checklist what to report when using such methods in publications.

Results for binary methods are presented in a contingency (also called for confusion) table out of which different measures can be calculated. The most important ones are the following six, which according to the guidelines [32] have to be provided for comprehensive assessments. Specificity, sensitivity, positive and negative predictive values (PPV and NPV) use half of the data in the matrix, while accuracy and Matthews correlation coefficient (MCC) use data from all the four data cells. Additional useful measures include area under curve (AUC) when presenting Receiver Operating Characteristic (ROC) curves, and Overall Performance Measure (OPM). Good methods display a balanced performance, their values for measures differ only slightly.

In case there is an imbalance in the number of cases in the classes, it has to be mitigated [31]. Several approaches are available for that. Cases used for testing method performance should not have been used for training them, otherwise there is circularity that overinflates

performance measures [14]. A scheme has been presented on how datasets should be split for training and testing as well as blind testing [34]. When there are more than two predicted classes additional measures are available [31, 32]. In addition to these measures, method assessment can contain other factors such as time required for predictions, as well as user friendliness and clarity of the service and results.

Datasets used for assessment have to be of sufficient size. There are a number of reasons for this requirement. Widely used machine learning methods are statistical by nature and require a relatively large number of cases for reliable testing. If we think the event space, in the case of proteins, there are 380 different amino acid substitution types, 150 of which are more likely due to happening because of a single nucleotide substitution within the coding region for a codon. These substitutions can appear in numerous different contexts, thus too small test datasets should be avoided. There are several performance assessments, especially for variants in a single protein or a small number of genes/proteins that do not have any statistical power. The smallest dataset we have seen contained just nine substitutions based on which a detailed analysis was performed to recommend the best performing tools!

Variation interpretation is often done in relation to human diseases. It is important to note that diseases are not binary states (benign/disease) instead there is a continuum and certain disease state can appear due to numerous different combinations of disease components, see the pathogenicity model [35]. This aspect has not been taken into account in benchmark datasets apart from training data for PON-PS [36] and clinical data for cystic fibrosis [37].

VARIATION DATASETS

We have collected from literature, websites and databases datasets, which have been used for training and benchmarking various types of variations and their effects (Table 1). The new datasets come from 109 papers. There are 419 new separate datasets containing altogether 329003373 variants. One paper can contain more than one dataset. The number of unique variants is smaller as many of the datasets are different subsets of commonly used datasets

such as ClinVar or ProTherm, or VariBench itself. The total number is dominated by VariSNP cases.

Table 1. New benchmark datasets added to VariBench

Origin of data	First used for	Number of variants	Reference
Variation type datasets			
Insertions and deletions (0/0)			
HGMD, 1000 GP	DDIG-In	659, 2008, 2479, 3861, 579, 2008, 2413, 3861	[38]
ClinVar, 1000 GP, ESP6500 SIFT-Indel	ENTPRISE-X	6513,5023,82, 366, 3171, 1604, 181, 1025	[39]
SwissProt, 100 GP, SM2PH	KD4i	2734	[40]
Sequence alignments	SIFT Indel	474, 9710	[41]
Substitutions, coding region (6/10)			
Training datasets			
Literature, patents	PredictSNP	10581, 5871, 43882, 32776, 3497, 11994	[11]
HGMD, SwissProt	FATHMM, FATHMM-XF	69141, 94995, 69141	[42, 43]
ClinVar, HGMD	MutationTaster	2600, 2199, 1100, 1100	[5]
HumDiv, UniProt, ClinVar	VIPUR	1542, 382, 949, 4992, 6555	[44]
Humsavar	BadMut	33483	[45]
HumVar, ExoVar, VariBenchSelected, SwissVarSelected	RAPSODY	21946	[46]
ClinVar, ESP	DANN	16627775, 49407057	[47]
SwissProt	NetSAP	5375, 1152	[48]

VariBench	PON-P2	10717, 13063, 1108, 1605, 6144, 8661, 656, 1053	[10]
Humsavar, VariBench	SuSPect	18633, 64163	[49]
CMG, DDD, ClinVar, ExoVar, 1000 GP, Hg19, Gencode, ESP6500	MAPPIN	64, 158, 3595, 15702, 512370, 51599, 11763, 1048544	[50]
Uniprot, 1000 GP, literature, VariBench, ARIC study	Ensemble predictor	36192, 238, 19520, 7953, 33511, 26962	[51]
ClinVar	PhD-SNP^{g}	48534, 1408	[52]
Multiple gene panel	MVP	1161	[53]
ADME genes LoF only	ADME optimized	337, 180	[54]
CinVar, NHGRI GWAS catalog, COSMIC, VariSNP	PredictSNP2	25480, 12050, 142722, 16716, 71674	[55]
Test datasets			
HumVar, ExoVar, VariBench, predictSNP, SwissVar	Circularity	40389, 8850, 10266, 16098, 12729	[14]
ClinVar, literature, PredictSNP	ACMG/AMP rules	14819, 1442, 4667, 6931, 5379, 12496, 14819, 4192, 16064, 10308, 7766	[56]
ClinVar, TP53, PPARG	Performance assessment	11995	[57]
UniProt	Guideline discordant/PRDIS	28474, 336730	[58]
ESP6500, HGMD	Compensated	1964	[59]

	pathogenic deviations		
VariBench	Representativeness	446013, 23671, 19335, 19459, 14610, 17623, 17525, 14647, 13096, 13069, 12584, 1605, 1301, 8664, 7152, 1053, 751, 16098, 10266, 8850, 40389, 21151, 22196, 75042	[6]
Structure mapped variants			
PDB, UniProt	PON-SC	349, 7795	[60]
3D	3D structure analysis	374	[61]
LSDBs, literature, ClinVar	Membrane proteins	2058	[19]
Synonymous			
ClinVar, GRASP, GWAS Catalog, GWASdb, PolymiRTS, PubMed, Web of Knowledge	dbDSM	2021	[62]
dbDSM, ClinVar, literature	IDSV	600, 5331	[63]
Benign			
dbSNP	VariSNP	446013, 956958, 470473, 3802, 9285, 3402, 5277, 11339, 588, 318967, 1804501, 610396, 25930776	[2]

ExAX	Assessment of benign variants	63197	[18]
Effect-specific datasets			
DNA regulatory elements			
Ensembl Compara, 1000 GP	Pathogenic regulatory variants	42, 142, 153, 43, 65, 3, 5	[64]
OMIM, ClinVar, VarDi, GWAS Catalog, HGMD, COSMIC, FANTOM5, ENCODE	Regulatory variants	27558, 20963, 43364	[65]
dbSNP, HGMP, HapMap, GWAS Catalog	Regulatory elements	225, 241910	[66]
ENCODE, NIH Roadmap Epigenomics	CAPE	7948, 4044, 2693, 51, 156, 56497, 2029	[67]
Whole genome sequences, GiaB, HGMD, ClinVar	CDTS	15741, 427, 10979, 67144812, 34687974, 30634572, 31893124, 61372584	[68]
Literature, OMIM, Epi4K	TraP	402, 97, 103	[69]
HGMD, 1000GP, ClinVar	ShapeGTP	4462, 1116	Malkowska et al. submitted
ClinVar, literature	NCBoost	655, 6550, 770	[70]
RNA splicing (1/1)			
Literature, LSDBs, HGP	DBASS3 and	307, 577	[71, 72]

	DBASS5		
HGMD, SpliceDisease database, DBASS, 1000 GP	dbscSNV	2959, 45, 2025	[21]
Experimental	BRCA1 and BRCA2	13, 15, 33, 38, 35, 73	[73]
Ensembl, UCSC Genome Browser	HumanSplicing Finder	424, 81, 15, 89	[74]
HGMD	MutPred Splice	2354, 638	[75]
hg19, GenBank, dbSNP	ASSEDA	41, 8, 12	[76]
Experimental	*RB1*	3, 17, 13, 6	[77]
Experimental	*LDLR*	18, 18	[78]
Experimental	*BRCA1* and *BRCA2*	6, 29, 6, 19	[79]
Experimental, LSDBs	*BRCA1* and *BRCA2*	53, 4, 4, 6, 5	[80]
Experimental	*BRCA1* and *BRCA2*	24, 22, 13, 10, 10, 5, 11	[81]
Experimental	Exon 1st nucleotide	25, 5, 9, 5, 5, 9, 30, 9	[82]
ClinVar, 1000GP	Splice site consensus region	222, 50	[83]
Protein aggregation (0/0)			
WALTZ-DB, AmylHex, AmylFrag, AGGRESCAN, TANGO	AmyLoad	1400	[84]

Experimental	WALTZ-DB	1089	[85]
Binding free energy			
Literature, ASEdb, PIN, ABbind, PROXiMATE, dbMPIKT	SKEMPI 2.0	7085	[86]
SKEMPI	Flex ddG	1249	[87]
Protein disorder (0/0)			
Literature	PON-Diso	103	[16]
Protein solubility (0/0)			
Literature	PON-Sol	443	[17]
Protein stability (4/6)			
Single variants			
ProTherm	PON-Tstab	1564	[30]
ProTherm	I-Mutant2.0	2087, 1948	[88]
ProTherm	Average assignment	1791, 1396, 2204	[89]
ProTherm	iPTREE-STAB	1859	[90]
ProTherm	SVM-WIN31 and SVM-3D12	1681, 1634, 499	[91]
ProTherm	PoPMuSiC-2.0	2648	[92]
ProTherm	sMMGB	1109	[93]
ProTherm	M8 and M47	2760, 1810	[94]
ProTherm	EASE-MM	238, 1676, 543	[95]
ProTherm	HoTMuSiC	1626	[96]
	SAAFEC	1262, 983	[97]
ProTherm	STRUM	3421, 306	[98]

ProTherm	Metapredictor	605	[99]
ProTherm	Automute	1962, 1925, 1749	[100]
TP53	TP53	42	[101]
ProTherm	S^{sym}	684	[102]
ProTherm, experimental data, ASEdb	Alanine scanning for binding energy	2971, 1005, 2154, 1210, 768, 380	[103]
ProTherm	Rosetta	1210	[104]
Double variants			
ProTherm	WET-STAB	180	[105]
Molecule-specific datasets (1/2)			
InSiGHT	PON-MMR2	178, 45	[106]
Literature	PON-mt-tRNA	145	[107]
BTKbase	PON-BTK	152	[108]
Kin-Driver, ClinVar, Ensembl	Kinact	384, 258	[109]
Literature	KinMutBase	1414	[110]
COSMIC	Kin-Driver	783, 648	[111]
OMIM, KinMutBase, HGMD	Protein kinases	1463, 999, 302	[112, 113]
UniProt, KinMutBase, SAAPdb, COSMIC	wKin-Mut	865, 2627	[114]
dbSNP, HGMD, COSMIC, literature	PTENpred	676	[115]
UniProt, Humsavar	Protein specific predictors	1872222 in 82 files	[12]

Literature	SAVER	187	[116]
Literature, experimental, dbSNP, ExAC, ESP	DPYD-Varifier	69, 295	[117]
Experimental	*BRCA1/2*	201, 68	[118]
Experimental	CFTR	20,11	[37]
CHAMP, literature	HApredictor	1138	[119]
Humsavar	MutaCYP	29, 285, 328	[120]
UniProt, HGMD, MutDB, dbSNP, literature	KvSNP	1259, 176	[121]
Disease-specific datasets (0/0)			
Literature, TP53 database, ClinVar, DoCM	Pan-cancer analysis	659, 65, 387	[122]
Literature, IARC TP53 Database, UMD BRCA1 and BRCA2	Cancer	3706	[123]
ICGC, TCGA, Pediatric Cancer Genome Project, dbSNP	Cancer	4690	[124]
Literature, LOVD, Inherited Arrhythmia Database	Long QT syndrome	90, 82, 8, 81, 113, 99, 14, 58, 55, 52, 28, 24, 109, 101, 8, 312	[125]
Experimental	PolyPhen-HCM	74, 78983	[126]
Functional assays	FASMIC	1049, 95, 40, 785, 21, 14, 35, 65, 22	[127]
Literature	dbCPM	941	[128]
cBioPortal, COSMIC, MSK-	OncoKB	4472	[129]

IMPACT cohort			
TCGA	DoCM	1364	[130]
Phenotype dataset (0/0)			
Literature, LSDBs	PON-PS	2527, 401	[36]

VariBench datasets are freely available at http://structure.bmc.lu.se/VariBench/ and can be downloaded separately. The website contains basic information about the datasets, their origin and for what purpose they were initially used for. Datasets are categorized similar to Table 1 for easy access. The contents of the datasets vary depending on information in the original source. We have enriched many of them e.g. by mapping to reference sequence or PDB structures, and some contain VariO annotations.

The available datasets have been categorized into 20 groups and subgroups as indicated in Figure 2. The figure shows also the relationships of the datasets in different categories. Variants are often described at three molecular levels (DNA, RNA and protein) and sometimes also at protein structural level, including relevant cross references and variant descriptions. VariBench utilizes and follows a number of standards and systematics including HGVS variation nomenclature, HGNC gene names (not in all databases due to mapping problems), and VariO annotations in some datasets.

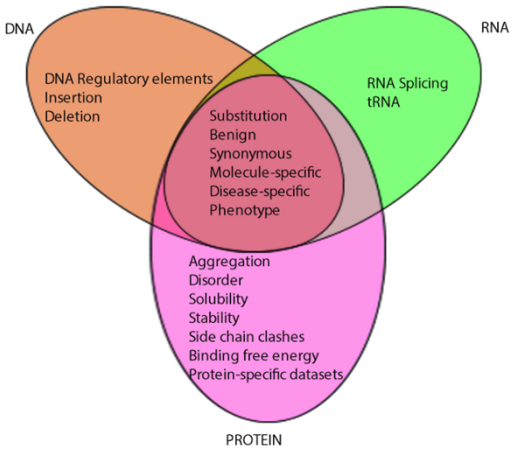

Figure 2. Types of benchmark datasets and their relations in VariBench.

Links are available to data in some external databases, including AmyLoad [84] and WALTZ-DB [85] for protein aggregation, DBASS3 and DBASS5 [71, 72] for splicing variants, SKEMPI [86], cancer datasets in KinMutBase [110], Kin-Driver [111], dbCPM [128], DoCM [130], and OncoKB [129], and tolerance predictor training set in DANN [47]. The latter has a link due to its huge size, the others since they are databases and as such easy to use directly and updated by third parties. We excluded datasets used in CAGI experiments, since they are available for registered participants only. LSDBs were excluded because data from these sources usually have to be manually selected before using as benchmark. Most of the time, there is no clear information for variant relevance to disease(s). Datasets for structural genomic variants were excluded, because they usually lack information about exact variation positions.

Unfortunately, many papers, even those reporting on benchmarking, do not contain and share the data, which does not allow others to extend the analyses and reuse the datasets.

Variation type datasets

Variation types include insertions and deletions, coding and non-coding region substitutions, which are divided into training and test datasets, structure mapped variants, as well as synonymous, and benign variants. There are now data from four amino acid insertion effect predictors, mainly for short alterations. Only datasets added after the release of the first version of VariBench are discussed here. In Table 1 it is shown how many datasets and publications in each category appeared in the first edition.

Training datasets have mainly been used for development of machine learning predictors, there are 17 new datasets. They typically also contain test sets. Six test datasets have been specifically designed for method assessments. These include a set for addressing circularity [14] and pathogenicity/tolerance method performance assessment [57]. The American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) has published guidelines for variant interpretation [131]. These include instructions

for use of prediction methods. A dataset was obtained for addressing concordance of prediction methods [56]. Another study addressed discordant cases [58]. Protein sequences of even closely related organisms contain differences and some of these are compensated variants where a disease-related variant in human is normal in another organism due to additional alteration(s) at other site(s). A dataset has been collected for such variants [59]. Unfortunately only the benign variants were made available. Analysis of the dataset representativeness, how well the datasets represent the variation space, was investigated for 24 datasets in VariBench and VariSNP [6]. These cases were mapped to reference sequence and are now available in the database.

Variations are mapped into protein three-dimensional structures in several datasets. Dedicated datasets contain those used for developing a method for predicting side chain clashes due to residue substitutions [60], analysis of effects on structures and functions of substitutions [61], and investigation of variations in membrane proteins [19].

There are two datasets for synonymous variants as well as two for benign ones.

Effect-specific datasets

These datasets are for various types of effects. On DNA level there are 8 sets for DNA regulatory elements, and on RNA level 13 datasets for splicing. Most of the splicing datasets are very small, but there are a few with substantially larger numbers. In the first version of VariBench, there were only protein stability datasets in this category, totally 6 datasets from 4 studies. Thus the growth has been substantial.

Many more sets are available for effects on protein level. Protein aggregation (2 datasets), binding free energy (2), disorder (1), solubility (1), and stability are the currently available categories. Among protein stability datasets, there are 18 new datasets for single variants, almost all originating from ProTherm, and one dataset for double variants.

Molecule-specific datasets

Thera are in VariBench 17 specific datasets for certain molecules. There is a set of variants used to train PON-mt-tRNA for substitutions affecting mitochondrial tRNA molecules [107]. This is of special interest as there are 22 unique mitochondrial tRNAs, which are implicated in a number of diseases.

The other datasets are protein specific. Kinact [109], Kin-Driver [111], KinMutBase [110], Kin-Mut [114] and the protein kinase dataset [112] contain variation information for protein kinases. The PON-BTK dataset was used to train a predictor for kinase domain variants in Bruton tyrosine kinase (BTK) [108]. There is a set for mismatch repair (MMR) proteins MLH1, MSH2, MSH6 and PMS2 and used to train PON-MMR2 [106].

Single amino acid substitutions were collected in 82 proteins to test whether there is a difference in performance for protein specific and generic predictors [12]. All the datasets contain at least ~100 variants. The results indicated vast differences in performances, the best generic predictors outperforming the specific predictors in most but not all cases.

The remaining datasets in this category are for variants in individual genes/proteins.

Disease-specific datasets

This category contains totally 9 datasets, six of which are for cancer, one for long QT syndrome [125] and another for hypertrophic cardiomyopathy [126].

Although there are numerous studies of cancer variations, the functional verification of the relevance of those variants for the disease is usually missing. VariBench contains three datasets for variants in cancer, which have been experimentally tested [122-124], and links to three other sources, namely dbCPM [128], DoCM [130], and OncoKB [129]. In addition, there is the FASMIC dataset for variants which are largely cancer related [127].

Phenotype dataset

One dataset contains information for disease phenotype, whether there is mild/moderate or severe disease due to substitutions. This dataset was used to train disease severity predictor called PON-PS [36].

BENCHMARK USE CASE

VariBench datasets have mainly been used for prediction method development and testing. As the benchmark studies typically have not contained all the best performing tools, we compared the performance of the variant tolerance/pathogenicity predictor PON-P2, since this tool has been the best or among the best performing methods in a number of previous investigations [10, 12, 18, 19, 58]. The setup was similar in all these studies: to test the outcome of a spectrum of methods. We extended the published benchmark studies by repeating the original analyses with PON-P2. To avoid circularity, we first excluded from the datasets all cases that had been used for training PON-P2. The results are shown in Table 2 and are reported according to the published guidelines [32] and including some additional measures.

The exercise indicated that reproducibility and reusability could not be achieved in a number of cases due to problems in reporting. We had to exclude some published benchmark studies. The dataset for pharmacogenetics variants [54] was too small for reliable estimation. The paper for compensated variants [59] did not share the disease-related variants, and thus could not be evaluated. Of the dataset used by [53] only 36 cases were not included to the PON-P2 training set, and therefore had to be excluded.

We were able to perform the analysis for six studies and we analysed altogether 17 datasets. Full comparison was not possible in all cases as some details were not available. Therefore we discuss and compare the performances based on the information in the original papers, but list all the details from our study in Table 2.

TABLE 2. Performance of PON-P2 on test datasets

Dataset	**TP**	**FP**	**TN**	**FN**	**Coverage**	**PPV**	**NPV**	**Sens**	**Spec**	**Acc**[a]	**MCC**	**OPM**
MutationTaster2, ClinVar [5]	544	9	959	32	0.685	0.99	0.947	0.944	0.991	0.968	0.936	0.910
MutationTaster2 [5]	407	10	803	63	0.635	0.986	0.881	0.866	0.988	0.927	0.860	0.810
Circularity, PredictSNPSelected [14]	5116	341	3173	590	0.623	0.940	0.770	0.900	0.860	0.880	0.730	0.606
Circularity, SwissVarSelected [14]	1551	818	3194	773	0.557	0.650	0.810	0.670	0.800	0.750	0.460	0.325
ACMG/AMP, MetaSVM [56]	2588	364	2457	192	0.503	0.878	0.927	0.931	0.871	0.901	0.803	0.733
ACMG/AMP, ClinVar_balanced [56]	841	136	608	69	0.455	0.835	0.915	0.924	0.817	0.871	0.746	0.666
ACMG/AMP, VaribenchSelected_Tolerance [56]	1727	171	2996	57	0.513	0.947	0.967	0.968	0.946	0.957	0.914	0.875
ACMG/AMP, predictSNPdsel [56]	3752	317	3071	427	0.539	0.906	0.899	0.898	0.906	0.902	0.804	0.734
ACMG/AMP, ClinVar_Sep2016 [56]	1050	215	1726	102	0.514	0.892	0.909	0.911	0.889	0.900	0.801	0.729
ACMG/AMP, Dominant_Recessive_Genes [56]	1284	98	619	52	0.506	0.875	0.957	0.961	0.863	0.912	0.828	0.769
ACMG/AMP, Oncogenes_TSG [56]	535	59	74	3	0.497	0.692	0.99	0.994	0.556	0.908 0.775(AN)	0.613	0.559
Variants in 3D structures [46]	5077	300	1060	266	0.337	0.812	0.94	0.95	0.779	0.865	0.741	0.676
ClinVar dataset [57]	1040	157	1200	169	0.541	0.881	0.864	0.86	0.884	0.872	0.745	0.664
TP53 dataset [57]	430	130	13	3	0.509	0.522	0.929	0.993	0.091	0.769 0.542(AN)	0.195	0.269
PPARG dataset [57]	131	1376	7	0	0.598	0.501	1.000	1.000	0.005	0.503	0.000	0.111
Cancer, functionally tested [123]	561	18	16	3	0.605	0.653	0.989	0.995	0.471	0.965 0.733(AN)	0.546	0.523
Cancer, non-COSMIC functionally tested [123]	108	10	14	3	0.455	0.700	0.956	0.973	0.583	0.904 0.778(AN)	0.604	0.549

[a]AN, after normalization.

For MutationTaster2 the published test data has not been previously available due to an inappropriate distribution format. MutationTaster 2 was originally compared to five tools and versions (MutationTaster1, PolyPhen humdiv and humvar, PROVEAN and SIFT) [5]. The accuracy and specificity are better for PON-P2 than the scores for the six tested tools and sensitivity is the second best. Only the measures given in the original article are discussed in here.

The study of circularity problems in variant testing was conducted on predictSNPSelected and SwissVarSelected datasets [14]. The performance of PON-P2 is superior compared to the eight tested predictors (MutationTaster2, PolyPhen, MutationAssessor, CADD, SIFT, LRT, FatHMM-U, FatHMM-W, Gerp++, and phyloP). In the test for predictSNPSelected dataset, NPV, PPV, sensitivity, accuracy and MCC are the best for PON-P2. Only for specificity it is the second best predictor, with a margin of 1%. In the data for SwissVarSelected, PON-P2 has the best score for PPV, accuracy and MCC. It is the second best for NPV and specificity, by 1-2% margin to the best, and for sensitivity. On both datasets, PON-P2 showed the most balanced performance.

25 tools were tested according to ACMG/AMP guidelines using several datasets [56]. The compared methods were REVEL, VEST3, MetaSVM, MetaLR, hEAt, Condel, MutPred, Mcap, Eigen, CADD, PolyPhen2, PROVEAN, SIFT, EA, MutationAssessor, MutationTaster, phyloP100way, FATHMM, DANN, LRT, SiPhy, phastConst100way, GenoCanyon, GERP, and Integrated_fitCons. Unfortunately, the results were not comprehensively reported. The paper contains data for AUC scores but they are presented as figures. The exact values were difficult to estimate, especially when results for 18 datasets were combined into single figures. In the end, we performed the test for 8 of these datasets. In the ClinVar balanced data the AUC of PON-P2 is either shared first or second, and in VariBenchselected data it has the best performance. Comparison for the six other datasets is not as reliable, but we can summarize

that the PON-P2 performance is among the best if not best for all of these. It is really a pity that exact numbers were not provided by the authors.

The performances of 23 methods (FATHMM, fitCons, LRT, MutationAssessor, MutationTaster, PlyPhen humdiv and humvar versions, PROVEAN, SIFT, VEST3, GERP++, phastCons, phyloP, SiPhy, CADD, DANN, Eigen, FATHMM-MKL, GenoCanyon, M-CAP, MetaLR, MetaSVM, REVEL) were tested on three datasets: ClinVar and two protein specific sets for TP53 and PPARG [57]. They had also a fourth set for autism spectrum diseases, but since there is no experimental evidence for the relation of these variations to the disease that set was excluded. Although the study was well performed and described, it seems that the authors have not corrected for class imbalance. For the methods to be comparable the measures should be calculated based on the same data and have equal numbers of positive and negative cases. If that is not the case, the imbalance has to be mitigated with one of the available solutions. Some of the other benchmark studies may suffer from the same problem, but we are not sure due to incomplete descriptions of the studies. None of the tools can predict all possible variations and thus they have predictions for different numbers. Therefore we present the results both for non-normalized and normalized data. We believe that the former was used by the authors. In the case of ClinVar data, PON-P2 has better PPV, accuracy and MCC than the other methods tested in the paper.

In the case of TP53 data, the PON-P2 accuracy is second best when the data are not normalized, on other measures PON-P2 is ranked the fourth or worse. All cancer variants, such as those in TP53, were excluded from the PON-P2 training data. This was done because the effects of variations in cancers usually have not been experimentally verified. A variant in TP53 is not “pathogenic” alone, several variants in different proteins are needed for cancer.

All the predictors are known to have variable performance depending on the tested protein, see the study of protein-specific predictors [12]. That study showed that PON-P2 had better performance for 85% of proteins, being the best of the five tested tools (PolyPhen-2, SIFT,

PON-P2, MutationTaster2, CADD). PPARG seems to be another example for which PON-P2 has poor performance [57]. An additional reason for poor performance may be that the PPARG data is not for pathogenicity, instead it is a "function score" that is based on the distribution of FACS sorted cells [132]. The same applies to the TP53 test data which is based on the protein function, not pathogenicity. Depending on a protein, the threshold for phenotype can be anything between 1 and 85% of the wild type activity (Vihinen, in preparation). We have previously tested PON-P2 in protein function prediction but with poor [133] or mixed (Kasak et al., submitted) outcome. This is because the method has not been trained and intended for this task. These results indicate the importance of applying computational tools to their intended purpose or at least testing the performance carefully before applied to new tasks.

Another study tested the performance of 14 tools (SEQ+DYN, SEQ, DYN, MutationTaster2, PolyPhen2, MutationAssessor, CADD, SIFT, LRT, FATHMM-U, Gerp++, phyloP, Condel, Logit) in relation to structural dynamics, which was used as a proxy for functional significance of amino acid substitutions [46]. PON-P2 has the best sensitivity, specificity, NPV and MMC, it is the second best for accuracy but only 13th for PPV. The explanation for the latter observation is that many of the tested tools are severely biased, having very high PPV but very low NPV, whereas the performance of PON-P2 was again balanced over all the measures.

The exercise indicated that it is possible to compare predictors to published results based on exactly the same datasets. The new performance results for PON-P2 are in line with several previously published studies that have indicated the methods to be a top performer on different benchmarks [10, 12, 18, 19, 58]. When choosing a method(s), one should look at consistent performance over several benchmarks.

Full comparisons were not always possible because of incomplete performance assessments. Therefore, authors should meticulously describe all details and procedures in the data analysis

as well as share the datasets used. Even if the data is taken from public sources, it is not possible for others to obtain exactly the same dataset as used in the papers even when applying the same selection criteria, as some important aspects seem always to be missing. In summary, it was possible to compare performances for methods not included into original studies. This is important in many ways, and contributes towards increased reproducibility and comparability. Good datasets are difficult to obtain, therefore VariBench will serve as a hub for sharing these important data.

REFERENCES

1. Nair PS, Vihinen M. VariBench: a benchmark database for variations. Hum Mutat. 2013;34(1):42-49. doi: 10.1002/humu.22204.

2. Schaafsma GC, Vihinen M. VariSNP, a benchmark database for vVariations from dbSNP. Hum Mutat. 2015;36(2):161-166. doi: 10.1002/humu.22727.

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

4. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018. doi: 10.1038/sdata.2016.18.

5. Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods. 2014;11(4):361-362. doi: 10.1038/nmeth.2890.

6. Schaafsma GC, Vihinen M. Representativeness of variation benchmark datasets. BMC Bioinformatics. 2018;19(1):461.

7. Potapov V, Cohen M, Schreiber G. Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. Protein Eng Des Sel. 2009;22(9):553-560. doi: 10.1093/protein/gzp030.

8. Khan S, Vihinen M. Performance of protein stability predictors. Hum Mutat. 2010;31(6):675-684. doi: 10.1002/humu.21242.

9. Thusberg J, Olatubosun A, Vihinen M. Performance of mutation pathogenicity prediction methods on missense variants. Hum Mutat. 2011;32(4):358-368. doi: 10.1002/humu.21445.

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

11. Bendl J, Stourac J, Salanda O, Pavelka A, Wieben ED, Zendulka J, et al. PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. PLoS Comput Biol. 2014;10(1):e1003440. doi: 10.1371/journal.pcbi.1003440.

12. Riera C, Padilla N, de la Cruz X. The complementarity between protein-specific and general pathogenicity predictors for amino acid substitutions. Hum Mutat. 2016;37(10):1012-1024. doi: 10.1002/humu.23048.

13. Masica DL, Karchin R. Towards increasing the clinical relevance of In silico methods to predict pathogenic missense variants. PLoS Comput Biol. 2016;12(5):e1004725. doi: 10.1371/journal.pcbi.1004725.

14. Grimm DG, Azencott CA, Aicheler F, Gieraths U, MacArthur DG, Samocha KE, et al. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. Hum Mutat. 2015;37(10):1013-1024. doi: 10.1002/humu.22768.

15. Laurila K, Vihinen M. Prediction of disease-related mutations affecting protein localization. BMC genomics. 2009;10:122-122. doi: 10.1186/1471-2164-10-122.

16. Ali H, Urolagin S, Gurarslan O, Vihinen M. Performance of protein disorder prediction programs on amino acid substitutions. Hum Mutat. 2014;35(7):794-804. doi: 10.1002/humu.22564.

17. Yang Y, Niroula A, Shen B, Vihinen M. PON-Sol: prediction of effects of amino acid substitutions on protein solubility. Bioinformatics. 2016;32(13):2032-2034. doi: 10.1093/bioinformatics/btw066.

18. Niroula A, Vihinen M. How good are pathogenicity predictors in detecting benign variants? PLoS Comput Biol. 2019;15(2):e1006481. doi: 10.1371/journal.pcbi.1006481.

19. Orioli T, Vihinen M. Benchmarking membrane proteins: Subcellular localization and variant tolerance predictors. BMC Genomics. 2019;(in press).

20. Desmet F, Hamroun G, Collod-Beroud G, Claustres M, Beroud C. Bioinformatics identification of splice site signals and prediction of mutation effects. In: Mohan RM, editor. Research Advances in Nucleic Acids Research. Kerala: Global Reseach Network; 2010. p. 1-16.

21. Jian X, Boerwinkle E, Liu X. In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Res. 2014;42(22):13534-13544. doi: 10.1093/nar/gku1206.

22. Anderson D, Lassmann T. A phenotype centric benchmark of variant prioritisation tools. NPJ Genom Med. 2018;3:5. doi: 10.1038/s41525-018-0044-9.

23. Vihinen M, Hancock JM, Maglott DR, Landrum MJ, Schaafsma GC, Taschner P. Human Variome Project Quality Assessment Criteria for Variation Databases. Hum Mutat. 2016;37(6):549-558. doi: 10.1002/humu.22976.

24. Gray KA, Yates B, Seal RL, Wright MW, Bruford EA. Genenames.org: the HGNC resources in 2015. Nucleic Acids Res. 2015;43(Database issue):D1079-1085. doi: 10.1093/nar/gku1071.

25. den Dunnen JT, Antonarakis SE. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. Hum Mutat. 2000;15(1):7-12. doi: 10.1002/(sici)1098-1004(200001)15:1<7::aid-humu4>3.0.co;2-n.

26. Dalgleish R, Flicek P, Cunningham F, Astashyn A, Tully RE, Proctor G, et al. Locus Reference Genomic sequences: an improved basis for describing human DNA variants. Genome Med. 2010;2(4):24. doi: 10.1186/gm145.

27. Rajput B, Pruitt KD, Murphy TD. RefSeq curation and annotation of stop codon recoding in vertebrates. Nucleic Acids Res. 2019;47(2):594-606. doi: 10.1093/nar/gky1234.

28. Vihinen M. Variation Ontology for annotation of variation effects and mechanisms. Genome Res. 2014;24(2):356-364. doi: 10.1101/gr.157495.113.

29. Kumar MD, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, Uedaira H, et al. ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. Nucleic Acids Res. 2006;34(Database issue):D204-206. doi: 10.1093/nar/gkj103.

30. Yang Y, Urolagin S, Niroula A, Ding X, Shen B, Vihinen M. PON-tstab: Protein variant stability predictor. Importance of training data quality. Int J Mol Sci. 2018;19(4). doi: 10.3390/ijms19041009.

31. Vihinen M. How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. BMC Genomics. 2012;13 Suppl 4:S2. doi: 10.1186/1471-2164-13-s4-s2.

32. Vihinen M. Guidelines for reporting and using prediction tools for genetic variation analysis. Hum Mutat. 2013;34(2):275-282. doi: 10.1002/humu.22253.

33. Walsh I, Pollastri G, Tosatto SC. Correct machine learning on protein sequences: a peer-reviewing perspective. Brief Bioinform. 2016;17(5):831-840. doi: 10.1093/bib/bbv082.

34. Niroula A, Vihinen M. Variation interpretation predictors: Principles, types, performance, and choice. Hum Mutat. 2016;37(6):579-597. doi: 10.1002/humu.22987.

35. Vihinen M. How to define pathogenicity, health, and disease? Hum Mutat. 2017;38(2):129-136. doi: 10.1002/humu.23144.

36. Niroula A, Vihinen M. Predicting severity of disease-causing variants. Hum Mutat. 2017;38(4):357-364. doi: 10.1002/humu.23173.

37. Masica DL, Sosnay PR, Raraigh KS, Cutting GR, Karchin R. Missense variants in CFTR nucleotide-binding domains predict quantitative phenotypes associated with cystic fibrosis disease severity. Hum Mol Genet. 2015;24(7):1908-1917. doi: 10.1093/hmg/ddu607.

38. Folkman L, Yang Y, Li Z, Stantic B, Sattar A, Mort M, et al. DDIG-in: detecting disease-causing genetic variations due to frameshifting indels and nonsense mutations

employing sequence and structural properties at nucleotide and protein levels. Bioinformatics. 2015;31(10):1599-1606. doi: 10.1093/bioinformatics/btu862.

39. Zhou H, Gao M, Skolnick J. ENTPRISE-X: Predicting disease-associated frameshift and nonsense mutations. PLoS One. 2018;13(5):e0196849. doi: 10.1371/journal.pone.0196849.

40. Bermejo-Das-Neves C, Nguyen HN, Poch O, Thompson JD. 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. 2014;15:111. doi: 10.1186/1471-2105-15-111.

41. Hu J, Ng PC. SIFT Indel: predictions for the functional effects of amino acid insertions/deletions in proteins. PLoS One. 2013;8(10):e77940. doi: 10.1371/journal.pone.0077940.

42. Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day IN, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics. 2015;31(10):1536-1543. doi: 10.1093/bioinformatics/btv009.

43. Shihab HA, Gough J, Mort M, Cooper DN, Day IN, Gaunt TR. Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. Hum Genomics. 2014;8:11. doi: 10.1186/1479-7364-8-11.

44. Baugh EH, Simmons-Edler R, Muller CL, Alford RF, Volfovsky N, Lash AE, et al. Robust classification of protein variation using structural modelling and large-scale data integration. Nucleic Acids Res. 2016;44(6):2501-2513. doi: 10.1093/nar/gkw120.

45. Korvigo I, Afanasyev A, Romashchenko N, Skoblov M. Generalising better: Applying deep learning to integrate deleteriousness prediction scores for whole-exome SNV studies. PLoS One. 2018;13(3):e0192829. doi: 10.1371/journal.pone.0192829.

46. Ponzoni L, Bahar I. Structural dynamics is a determinant of the functional significance of missense variants. Proc Natl Acad Sci U S A. 2018;115(16):4164-4169. doi: 10.1073/pnas.1715896115.

47. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. 2015;31(5):761-763. doi: 10.1093/bioinformatics/btu703.

48. Li Y, Wen Z, Xiao J, Yin H, Yu L, Yang L, et al. Predicting disease-associated substitution of a single amino acid by analyzing residue interactions. BMC Bioinformatics. 2011;12:14. doi: 10.1186/1471-2105-12-14.

49. Yates CM, Filippis I, Kelley LA, Sternberg MJ. SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features. J Mol Biol. 2014;426(14):2692-2701. doi: 10.1016/j.jmb.2014.04.026.

50. Gosalia N, Economides AN, Dewey FE, Balasubramanian S. MAPPIN: a method for annotating, predicting pathogenicity and mode of inheritance for nonsynonymous variants. Nucleic Acids Res. 2017;45(18):10393-10402. doi: 10.1093/nar/gkx730.

51. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet. 2015;24(8):2125-2137. doi: 10.1093/hmg/ddu733.

52. Capriotti E, Fariselli P. PhD-SNPg: a webserver and lightweight tool for scoring single nucleotide variants. Nucleic Acids Res. 2017;45(W1):W247-w252. doi: 10.1093/nar/gkx369.

53. Qian D, Li S, Tian Y, Clifford JW, Sarver BAJ, Pesaran T, et al. A Bayesian framework for efficient and accurate variant prediction. PLoS One. 2018;13(9):e0203553. doi: 10.1371/journal.pone.0203553.

54. Zhou Y, Mkrtchian S, Kumondai M, Hiratsuka M, Lauschke VM. An optimized prediction framework to assess the functional impact of pharmacogenetic variants. Pharmacogenomics J. 2018. doi: 10.1038/s41397-018-0044-2.

55. Bendl J, Musil M, Stourac J, Zendulka J, Damborsky J, Brezovsky J. PredictSNP2: A Unified Platform for Accurately Evaluating SNP Effects by Exploiting the Different

Characteristics of Variants in Distinct Genomic Regions. PLoS Comput Biol. 2016;12(5):e1004962. doi: 10.1371/journal.pcbi.1004962.

56. Ghosh R, Oak N, Plon SE. Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. Genome Biol. 2017;18(1):225. doi: 10.1186/s13059-017-1353-5.

57. Li J, Zhao T, Zhang Y, Zhang K, Shi L, Chen Y, et al. Performance evaluation of pathogenicity-computation methods for missense variants. Nucleic Acids Res. 2018;46(15):7793-7804. doi: 10.1093/nar/gky678.

58. de la Campa EA, Padilla N, de la Cruz X. Development of pathogenicity predictors specific for variants that do not comply with clinical guidelines for the use of computational evidence. BMC Genomics. 2017;18(Suppl 5):569. doi: 10.1186/s12864-017-3914-0.

59. Azevedo L, Mort M, Costa AC, Silva RM, Quelhas D, Amorim A, et al. Improving the in silico assessment of pathogenicity for compensated variants. Eur J Hum Genet. 2016;25(1):2-7. doi: 10.1038/ejhg.2016.129.

60. Calyseva J, Vihinen M. PON-SC - program for identifying steric clashes caused by amino acid substitutions. BMC Bioinformatics. 2017;18(1):531. doi: 10.1186/s12859-017-1947-7.

61. Bhattacharya R, Rose PW, Burley SK, Prlic A. Impact of genetic variation on three dimensional structure and function of proteins. PLoS One. 2017;12(3):e0171355. doi: 10.1371/journal.pone.0171355.

62. Wen P, Xiao P, Xia J. dbDSM: a manually curated database for deleterious synonymous mutations. Bioinformatics. 2016;32(12):1914-1916. doi: 10.1093/bioinformatics/btw086.

63. Shi F, Yao Y, Bin Y, Zheng CH, Xia J. Computational identification of deleterious synonymous variants in human genomes using a feature-based approach. BMC Med Genomics. 2019;12(Suppl 1):12. doi: 10.1186/s12920-018-0455-6.

64. Smedley D, Schubach M, Jacobsen JOB, Kohler S, Zemojtel T, Spielmann M, et al. A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. Am J Hum Genet. 2016;99(3):595-606. doi: 10.1016/j.ajhg.2016.07.005.

65. Ma M, Ru Y, Chuang LS, Hsu NY, Shi LS, Hakenberg J, et al. Disease-associated variants in different categories of disease located in distinct regulatory elements. BMC Genomics. 2015;16 Suppl 8:S3. doi: 10.1186/1471-2164-16-s8-s3.

66. Zhao Y, Clark WT, Mort M, Cooper DN, Radivojac P, Mooney SD. Prediction of functional regulatory SNPs in monogenic and complex disease. Hum Mutat. 2011;32(10):1183-1190. doi: 10.1002/humu.21559.

67. Li S, Alvarez RV, Sharan R, Landsman D, Ovcharenko I. Quantifying deleterious effects of regulatory variants. Nucleic Acids Res. 2017;45(5):2307-2317. doi: 10.1093/nar/gkw1263.

68. di Iulio J, Bartha I, Wong EHM, Yu HC, Lavrenko V, Yang D, et al. The human noncoding genome defined by genetic diversity. Nat Genet. 2018;50(3):333-337. doi: 10.1038/s41588-018-0062-7.

69. Gelfman S, Wang Q, McSweeney KM, Ren Z, La Carpia F, Halvorsen M, et al. Annotating pathogenic non-coding variants in genic regions. Nat Commun. 2017;8(1):236. doi: 10.1038/s41467-017-00141-2.

70. Caron B, Luo Y, Rausell A. NCBoost classifies pathogenic non-coding variants in Mendelian diseases through supervised learning on purifying selection signals in humans. Genome Biol. 2019;20(1):32. doi: 10.1186/s13059-019-1634-2.

71. Buratti E, Chivers M, Královicová J, Romano M, Baralle M, Krainer AR, et al. Aberrant 5' splice sites in human disease genes: mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization. Nucleic Acids Res. 2007;35(13):4250-4263. doi: 10.1093/nar/gkm402.

72. Vořechovský I. Aberrant 3' splice sites in human disease genes: mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization. Nucleic Acids Res. 2006;34(16):4630-4641. doi: 10.1093/nar/gkl535.

73. Houdayer C, Caux-Moncoutier V, Krieger S, Barrois M, Bonnet F, Bourdon V, et al. 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. 2012;33(8):1228-1238. doi: 10.1002/humu.22101.

74. Desmet FO, Hamroun D, Lalande M, Collod-Beroud G, Claustres M, Beroud C. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res. 2009;37(9):e67. doi: 10.1093/nar/gkp215.

75. Mort M, Sterne-Weiler T, Li B, Ball EV, Cooper DN, Radivojac P, et al. MutPred Splice: machine learning-based prediction of exonic variants that disrupt splicing. Genome Biol. 2014;15(1):R19. doi: 10.1186/gb-2014-15-1-r19.

76. Mucaki EJ, Shirley BC, Rogan PK. Prediction of mutant mRNA splice isoforms by information theory-based exon definition. Hum Mutat. 2013;34(4):557-565. doi: 10.1002/humu.22277.

77. Houdayer C, Dehainault C, Mattler C, Michaux D, Caux-Moncoutier V, Pages-Berhouet S, et al. Evaluation of in silico splice tools for decision-making in molecular diagnosis. Hum Mutat. 2008;29(7):975-982. doi: 10.1002/humu.20765.

78. Holla OL, Nakken S, Mattingsdal M, Ranheim T, Berge KE, Defesche JC, et al. Effects of intronic mutations in the LDLR gene on pre-mRNA splicing: Comparison of wet-lab and bioinformatics analyses. Mol Genet Metab. 2009;96(4):245-252. doi: 10.1016/j.ymgme.2008.12.014.

79. Vreeswijk MP, Kraan JN, van der Klift HM, Vink GR, Cornelisse CJ, Wijnen JT, et al. Intronic variants in BRCA1 and BRCA2 that affect RNA splicing can be reliably selected by splice-site prediction programs. Hum Mutat. 2009;30(1):107-114. doi: 10.1002/humu.20811.

80. Thery JC, Krieger S, Gaildrat P, Revillion F, Buisine MP, Killian A, et al. Contribution of bioinformatics predictions and functional splicing assays to the interpretation of unclassified variants of the BRCA genes. Eur J Hum Genet. 2011;19(10):1052-1058. doi: 10.1038/ejhg.2011.100.

81. Colombo M, De Vecchi G, Caleca L, Foglia C, Ripamonti CB, Ficarazzi F, et al. Comparative in vitro and in silico analyses of variants in splicing regions of BRCA1 and BRCA2 genes and characterization of novel pathogenic mutations. PLoS One. 2013;8(2):e57173. doi: 10.1371/journal.pone.0057173.

82. Grodecka L, Lockerova P, Ravcukova B, Buratti E, Baralle FE, Dusek L, et al. Exon first nucleotide mutations in splicing: evaluation of in silico prediction tools. PLoS One. 2014;9(2):e89570. doi: 10.1371/journal.pone.0089570.

83. Tang R, Prosser DO, Love DR. Evaluation of Bioinformatic Programmes for the Analysis of Variants within Splice Site Consensus Regions. Adv Bioinformatics. 2016;2016:5614058. doi: 10.1155/2016/5614058.

84. Wozniak PP, Kotulska M. AmyLoad: website dedicated to amyloidogenic protein fragments. Bioinformatics. 2015;31(20):3395-3397. doi: 10.1093/bioinformatics/btv375.

85. Beerten J, Van Durme J, Gallardo R, Capriotti E, Serpell L, Rousseau F, et al. WALTZ-DB: a benchmark database of amyloidogenic hexapeptides. Bioinformatics. 2015;31(10):1698-1700. doi: 10.1093/bioinformatics/btv027.

86. Jankauskaite J, Jimenez-Garcia B, Dapkunas J, Fernandez-Recio J, Moal IH. SKEMPI 2.0: An updated benchmark of changes in protein-protein binding energy, kinetics and thermodynamics upon mutation. Bioinformatics. 2018. doi: 10.1093/bioinformatics/bty635.

87. Barlow KA, S OC, Thompson S, Suresh P, Lucas JE, Heinonen M, et al. Flex ddG: Rosetta Ensemble-Based Estimation of Changes in Protein-Protein Binding Affinity upon Mutation. J Phys Chem B. 2018;122(21):5389-5399. doi: 10.1021/acs.jpcb.7b11367.

88. Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res. 2005;33(Web Server issue):W306-310. doi: 10.1093/nar/gki375.

89. Saraboji K, Gromiha MM, Ponnuswamy MN. Average assignment method for predicting the stability of protein mutants. Biopolymers. 2006;82(1):80-92. doi: 10.1002/bip.20462.

90. Huang LT, Gromiha MM, Ho SY. iPTREE-STAB: interpretable decision tree based method for predicting protein stability changes upon mutations. Bioinformatics. 2007;23(10):1292-1293. doi: 10.1093/bioinformatics/btm100.

91. Capriotti E, Fariselli P, Rossi I, Casadio R. A three-state prediction of single point mutations on protein stability changes. BMC Bioinformatics. 2008;9 Suppl 2:S6. doi: 10.1186/1471-2105-9-s2-s6.

92. Dehouck Y, Kwasigroch JM, Gilis D, Rooman M. PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. BMC Bioinformatics. 2011;12:151. doi: 10.1186/1471-2105-12-151.

93. Zhang Z, Wang L, Gao Y, Zhang J, Zhenirovskyy M, Alexov E. Predicting folding free energy changes upon single point mutations. Bioinformatics. 2012;28(5):664-671. doi: 10.1093/bioinformatics/bts005.

94. Yang Y, Chen B, Tan G, Vihinen M, Shen B. Structure-based prediction of the effects of a missense variant on protein stability. Amino Acids. 2013;44(3):847-855. doi: 10.1007/s00726-012-1407-7.

95. Folkman L, Stantic B, Sattar A, Zhou Y. EASE-MM: Sequence-based prediction of mutation-induced stability changes with feature-based multiple models. J Mol Biol. 2016;428(6):1394-1405. doi: 10.1016/j.jmb.2016.01.012.

96. Pucci F, Bourgeas R, Rooman M. Predicting protein thermal stability changes upon point mutations using statistical potentials: Introducing HoTMuSiC. Sci Rep. 2016;6:23257. doi: 10.1038/srep23257.

97. Getov I, Petukh M, Alexov E. SAAFEC: Predicting the Effect of Single Point Mutations on Protein Folding Free Energy Using a Knowledge-Modified MM/PBSA Approach. Int J Mol Sci. 2016;17(4):512. doi: 10.3390/ijms17040512.

98. Quan L, Lv Q, Zhang Y. STRUM: structure-based prediction of protein stability changes upon single-point mutation. Bioinformatics. 2016;32(19):2936-2946. doi: 10.1093/bioinformatics/btw361.

99. Broom A, Jacobi Z, Trainor K, Meiering EM. Computational tools help improve protein stability but with a solubility tradeoff. J Biol Chem. 2017;292(35):14349-14361. doi: 10.1074/jbc.M117.784165.

100. Masso M, Vaisman, II. Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis. Bioinformatics. 2008;24(18):2002-2009. doi: 10.1093/bioinformatics/btn353.

101. Pires DE, Ascher DB, Blundell TL. mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics. 2014;30(3):335-342. doi: 10.1093/bioinformatics/btt691.

102. Pucci F, Bernaerts K, Kwasigroch JM, Rooman M. Quantification of biases in predictions of protein stability changes upon mutations. Bioinformatics. 2018. doi: 10.1093/bioinformatics/bty348.

103. Kortemme T, Baker D. A simple physical model for binding energy hot spots in protein-protein complexes. Proc Natl Acad Sci U S A. 2002;99(22):14116-14121. doi: 10.1073/pnas.202485799.

104. Kellogg EH, Leaver-Fay A, Baker D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins. 2011;79(3):830-838. doi: 10.1002/prot.22921.

105. Huang LT, Gromiha MM. Reliable prediction of protein thermostability change upon double mutation from amino acid sequence. Bioinformatics. 2009;25(17):2181-2187. doi: 10.1093/bioinformatics/btp370.

106. Niroula A, Vihinen M. Classification of Amino Acid Substitutions in Mismatch Repair Proteins Using PON-MMR2. Hum Mutat. 2015;36(12):1128-1134. doi: 10.1002/humu.22900.

107. Niroula A, Vihinen M. PON-mt-tRNA: a multifactorial probability-based method for classification of mitochondrial tRNA variations. Nucleic Acids Res. 2016;44(5):2020-2027. doi: 10.1093/nar/gkw046.

108. Väliaho J, Faisal I, Ortutay C, Smith CIE, Vihinen M. Characterization of all possible single nucleotide change –caused amino acid substitutions in the kinase domain of Bruton tyrosine kinase. Hum Mutat. 2015;(in press).

109. Rodrigues CH, Ascher DB, Pires DE. Kinact: a computational approach for predicting activating missense mutations in protein kinases. Nucleic Acids Res. 2018;46(W1):W127-w132. doi: 10.1093/nar/gky375.

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

111. Simonetti FL, Tornador C, Nabau-Moreto N, Molina-Vila MA, Marino-Buslje C. Kin-Driver: a database of driver mutations in protein kinases. Database (Oxford). 2014;2014:bau104. doi: 10.1093/database/bau104.

112. Torkamani A, Schork NJ. Distribution analysis of nonsynonymous polymorphisms within the human kinase gene family. Genomics. 2007;90(1):49-58. doi: 10.1016/j.ygeno.2007.03.006.

113. Torkamani A, Schork NJ. Accurate prediction of deleterious protein kinase polymorphisms. Bioinformatics. 2007;23(21):2918-2925. doi: 10.1093/bioinformatics/btm437.

114. Izarzugaza JM, del Pozo A, Vazquez M, Valencia A. Prioritization of pathogenic mutations in the protein kinase superfamily. BMC Genomics. 2012;13 Suppl 4:S3. doi: 10.1186/1471-2164-13-s4-s3.

115. Johnston SB, Raines RT. PTENpred: A Designer Protein Impact Predictor for PTEN-related Disorders. J Comput Biol. 2016;23(12):969-975. doi: 10.1089/cmb.2016.0058.

116. Adebali O, Reznik AO, Ory DS, Zhulin IB. Establishing the precise evolutionary history of a gene improves prediction of disease-causing missense mutations. Genet Med. 2016;18(10):1029-1036. doi: 10.1038/gim.2015.208.

117. Shrestha S, Zhang C, Jerde CR, Nie Q, Li H, Offer SM, et al. Gene-Specific Variant Classifier (DPYD-Varifier) to Identify Deleterious Alleles of Dihydropyrimidine Dehydrogenase. Clin Pharmacol Ther. 2018;104(4):709-718. doi: 10.1002/cpt.1020.

118. Sadowski CE, Kohlstedt D, Meisel C, Keller K, Becker K, Mackenroth L, et al. BRCA1/2 missense mutations and the value of in-silico analyses. Eur J Med Genet. 2017;60(11):572-577. doi: 10.1016/j.ejmg.2017.08.005.

119. Hamasaki-Katagiri N, Salari R, Wu A, Qi Y, Schiller T, Filiberto AC, et al. A gene-specific method for predicting hemophilia-causing point mutations. J Mol Biol. 2013;425(21):4023-4033. doi: 10.1016/j.jmb.2013.07.037.

120. Fechter K, Porollo A. MutaCYP: Classification of missense mutations in human cytochromes P450. BMC Med Genomics. 2014;7:47. doi: 10.1186/1755-8794-7-47.

121. Stead LF, Wood IC, Westhead DR. KvSNP: accurately predicting the effect of genetic variants in voltage-gated potassium channels. Bioinformatics. 2011;27(16):2181-2186. doi: 10.1093/bioinformatics/btr365.

122. Niroula A, Vihinen M. Harmful somatic amino acid substitutions affect key pathways in cancers. BMC Med Genomics. 2015;8:53. doi: 10.1186/s12920-015-0125-x.

123. Martelotto LG, Ng CK, De Filippo MR, Zhang Y, Piscuoglio S, Lim RS, et al. Benchmarking mutation effect prediction algorithms using functionally validated cancer-related missense mutations. Genome Biol. 2014;15(10):484. doi: 10.1186/s13059-014-0484-1.

124. Goncearenco A, Rager SL, Li M, Sang QX, Rogozin IB, Panchenko AR. Exploring background mutational processes to decipher cancer genetic heterogeneity. Nucleic Acids Res. 2017;45(W1):W514-w522. doi: 10.1093/nar/gkx367.

125. Leong IU, Stuckey A, Lai D, Skinner JR, Love DR. Assessment of the predictive accuracy of five in silico prediction tools, alone or in combination, and two metaservers to classify long QT syndrome gene mutations. BMC Med Genet. 2015;16:34. doi: 10.1186/s12881-015-0176-z.

126. Jordan DM, Kiezun A, Baxter SM, Agarwala V, Green RC, Murray MF, et al. Development and validation of a computational method for assessment of missense variants in hypertrophic cardiomyopathy. Am J Hum Genet. 2011;88(2):183-192. doi: 10.1016/j.ajhg.2011.01.011.

127. Ng PK, Li J, Jeong KJ, Shao S, Chen H, Tsang YH, et al. Systematic Functional Annotation of Somatic Mutations in Cancer. Cancer Cell. 2018;33(3):450-462.e410. doi: 10.1016/j.ccell.2018.01.021.

128. Yue Z, Zhao L, Xia J. dbCPM: a manually curated database for exploring the cancer passenger mutations. Brief Bioinform. 2018. doi: 10.1093/bib/bby105.

129. Chakravarty D, Gao J, Phillips SM, Kundra R, Zhang H, Wang J, et al. OncoKB: A Precision Oncology Knowledge Base. JCO Precis Oncol. 2017;2017. doi: 10.1200/po.17.00011.

130. Ainscough BJ, Griffith M, Coffman AC, Wagner AH, Kunisaki J, Choudhary MN, et al. DoCM: a database of curated mutations in cancer. Nat Methods. 2016;13(10):806-807. doi: 10.1038/nmeth.4000.

131. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. 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. 2015;17:405-423. doi: 10.1038/gim.2015.30.

132. Majithia AR, Tsuda B, Agostini M, Gnanapradeepan K, Rice R, Peloso G, et al. Prospective functional classification of all possible missense variants in PPARG. Nat Genet. 2016;48(12):1570-1575. doi: 10.1038/ng.3700.

133. Niroula A, Vihinen M. PON-P and PON-P2 predictor performance in CAGI challenges: Lessons learned. Hum Mutat. 2017;38(9):1085-1091. doi: 10.1002/humu.23199.