Computers in Biology and Medicine 145 (2022) 105434

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Assessment of 13 *in silico* pathogenicity methods on cancer-related variants

Metin Yazar [a,b], Pemra Ozbek [a,*]

[a] Department of Bioengineering, Marmara University, Göztepe, İstanbul, Turkey
[b] Department of Genetics and Bioengineering, Istanbul Okan University, Tuzla, Istanbul, Turkey

ARTICLE INFO

Keywords:
Single nucleotide variants (SNVs)
Cancer-related variants
ClinVar
Protein function
Cancer genomics
In silico tools

ABSTRACT

Single nucleotide variants (SNVs) are single base substitutions that could influence many biological functions in the cell including gene expression, protein folding, and protein-protein interactions among many others. Thus, predictions of functional effects of cancer-related variants are crucial for drug responses and treatment options in clinical oncology. Experimental identification of these effects could be slow, inefficient, and inconvenient, hence *in silico* methods are gaining popularity in predicting the variants' effects. There are many studies on the cancer variants, however, up to date, none of these have been aimed to assess the performance metrics of *in silico* pathogenicity methods on functional relevance of cancer variants obtained from ClinVar. To this end, we examined the pathogenicity predictions of cancer-related variant datasets of 8 cancer types (bladder, breast, colon, colorectal, kidney, liver, lung, and pancreas cancer) retrieved from ClinVar using 13 different *in silico* methods including SIFT, CADD, FATHMM-weighted, FATHMM-unweighted, GERP++, MetaSVM, Mutation Assessor, MutationTaster, MutPred, PolyPhen-2, Provean, Revel and VEST4. A combination of statistical performance metric analysis, prediction distribution frequency data and ROC curve analysis results have suggested that; among all *in silico* prediction tools, top three tools with the highest discriminatory power were found to be MutPred (AUC = 0.677), MetaSVM (AUC = 0.645) and Revel (AUC = 0.637).

1. Introduction

The widespread use of next-generation sequencing (NGS) and single nucleotide variant array technologies in clinical diagnosis has resulted in the discovery of an increasing number of single nucleotide variants (SNVs) [1,2]. Both genome-wide association studies (GWAS) and candidate gene association studies provide valuable insights about SNVs [3], but other challenges, such as classification and prediction of functional effects of the variants have emerged [4]. As functional classification of variants with experimental methods can be time-consuming and labor-intensive, computational predictor tools and algorithms are gaining more popularity for identifying pathogenic variants based on biochemical and biological features [5]. Currently (February 2022), there are more than 1.9 million germline variants submitted in ClinVar [6], and more than 23 million somatic variants in COSMIC (the Catalogue Of Somatic Mutations In Cancer) [7]. Continuous increase in the submission of disease-related variants to the publicly available databases create the necessity of developing functional prediction computational tools [8,9]. Investigation of functional effects of variants on molecular mechanisms and disease pathogenesis is vital for genesis and prognosis of many diseases including cancer [10]. Relationship between cancer phenotype and molecular mechanism of variants remains unclear despite the presence of panel gene sequencing of cancer-related genes with NGS [11,12].

SNVs are the most common cause of differences in human genomes that can influence biological functions in the cell such as gene expression, disease susceptibility and protein-protein interactions among other things [13,14]. The variants can result in beneficial, neutral or negative effects on the phenotypic consequences and are defined as benign, neutral or pathogenic, respectively [15,16]. Benign and neutral variants generally have mild to tolerable or beneficial effects, whereas most pathogenic variants cause damaging effects on individuals and lead to decrease the frequency of these individuals in population [17]. Therefore, benign variants are transmitted through generations and are found at high frequencies among the population [18,19].

There are several country-specific consensus guidelines for the interpretation and classification of variants such as American College of Medical Genetics (ACMG) [20] and UK Association for Clinical Genomic Science (UK-ACGS) [21]. Other guidelines for the clinical interpretation of cancer related variants have been created by the Association for

* Corresponding author.
E-mail addresses: metin.yazar@okan.edu.tr (M. Yazar), pemra.ozbek@marmara.edu.tr (P. Ozbek).

https://doi.org/10.1016/j.compbiomed.2022.105434
Received 4 January 2022; Received in revised form 4 March 2022; Accepted 20 March 2022
Available online 26 March 2022
0010-4825/© 2022 Elsevier Ltd. All rights reserved.

Molecular Pathology (AMP), American Society of Clinical Oncology (ASCO) and College of American Pathologists (CAP) [22]. As a major classification, cancer variants are generally divided into somatic and germline variants based on cell type and location in the body [23]. In clinical oncology, somatic variants have received more attention than germline variants since they have a more influence on drug responses and treatment options [24]. Both somatic and germline variants have recently been thought to be important in drug sensitivity, toxicity and selection, and thus affect cancer pathogenesis [25,26]. Another valid classification for cancer variants is dividing them as 'driver' or 'passenger'. Driver variants are the ones that provide a biological benefit to the tumor cell, whereas passenger variants do not contribute to tumor progression and are only significant for increased mutation rates and loss of cell division control [24,25].

The ACMG, UK-ACGS, and AMP-ASCO-CAP guidelines all recommend using *in silico* tools and algorithms for the prediction and classification of variants in clinics. *In silico* predictor algorithms and tools use different methodologies that are classified as consensus-based tools, sequence-structure based tools, sequence homology-based tools and ensemble-supervised learning-based tools [16,27]. These methods generally include evolutionary, structure- and sequence-based parameters to gain insights on the impact of variants [28]. PolyPhen-2 [29], FATHMM [30], GERP++ [31] and MutationTaster [32] are examples of sequence-structure based tools. SIFT [33], Provean [34] and Mutation Assessor [35] are examples of sequence-homology based tools. CADD [36], MutPred [37], Revel [38], VEST4 [39], and meta-SVM [40] are examples of ensemble-supervised learning based tools. Meta-SNP [41] and PredictSNP [42] are example of consensus-based tools.

The need for a curated variant database arose upon generation of large amount of data using NGS [43]. Therefore, many variation databases have been created for the collection, curation and organization of the variants to help the bioinformaticians, clinicians and experimentalists [44,45]. The majority of disease-causing variants, including cancer variants, have been compiled in open variation databases such as ClinVar [6] and the genome aggregation database (gnomAD) [46,47]. ClinVar [6] is the most comprehensive variation database that includes clinical consequence data as well as genotype and disease information of variations. Furthermore, it gathers data from variant submissions with clinically or experimentally observations, and computes a cumulative interpretation to determine whether there is agreement or disagreement among submitters [6,48]. Although many cancer variants have been deposited into this database, a large group of these variants do not have true interpretation and classification [49]. Misinterpretation problem of cancer variants lead to several problems in clinics such as wrong clinical diagnosis, faulty pathogenicity effect mechanism of the variants and outdated submissions [50]. In addition, there are many variants in human variation databases whose clinical implications, interpretation, or classification have been unknown and are referred to as "variants of uncertain significance (VUS)" [51].

Performance assessment of *in silico* prediction tools [48,52]; [17,51,53,54] have been conducted with different benchmark variant datasets that use different set of tools. Several challenges were reported due to the high number of variants and prediction tools. A remarkable issue has arisen due to over-fitting of variant interpretation data, as the predictor tools' algorithms are trained with redundant data [48,55]. Another challenge is that the performance of a tool can vary significantly depending on the dataset type [17,56]. To eliminate both over-fitting and tool performance variability, datasets from online variation databases such as gnomAD [46], dbSNP [57], the 1000 Genomes Project [58] and ClinVar [6] have been proposed. Variant datasets were also classified according to their purpose such as effect-specific datasets, molecule-specific datasets and disease-specific datasets [59]. Cancer-specific datasets are disease-specific datasets that consist of cancer-related variations [60]. Since experimental verification of functional effects of cancer variations is generally missing, cancer-specific datasets do not contain many genes and variants [60–65]. Also, several molecular-specific datasets, including KinMutBase [66], KinMut [67], Kin-Driver [68], consist of several cancer variations.

There exist numerous studies on cancer variants [69–72] but, to date, none of these aimed to assess the functional relevance of cancer variants obtained from ClinVar [6] by using different *in silico* pathogenicity methods. To this end, we aimed to examine the functional effects of cancer-related variants with known clinical significance and VUSs from 8 different cancer types (bladder, breast, colon, colorectal, kidney, liver, lung and pancreas cancer) retrieved from ClinVar [6] using 13 different *in silico* methods. Also, we have evaluated the performance metrics of these tools on cancer-related variants from ClinVar.

2. Materials-methods

2.1. Retrieval of variants and formation of cancer datasets

Dataset in this study was created in January 2021 based on ClinVar [6] for 8 different cancer types; including bladder, breast, colon, colorectal, kidney, liver, lung, and pancreas cancers resulting in a total of 17549 variants. As a search keyword "cancer type" was used, "missense" molecular consequences were chosen, and variation type was set to "single nucleotide" when retrieving data from ClinVar. Furthermore, variants were filtered based on their ClinVar review status, with those with a "No assertion criteria" review status being removed. Duplicate variants were also excluded from the dataset. The dataset was split into two groups as variants with known clinical significance and VUSs. Details of the final dataset used in this study are given in Table 1.

2.2. Prediction scores of variants from in silico prediction tools

For each dataset, the functional effects of the variants were predicted using 13 *in silico* tools, including SIFT [33], CADD [36], FATHMM-weighted [30], FATHMM-unweighted [30], GERP++ [31], MetaSVM [40], Mutation Assessor [35], MutationTaster [32], MutPred [37], PolyPhen-2 [29], Provean [34], Revel [38] and VEST4 [39]. Pathogenicity scores of Provean [34], FATHMM-weighted [30] and FATHMM-unweighted [30] were obtained using these tools' own websites. Other tools were annotated via Variant Effect Predictor (VEP) [73]. While SIFT [33], CADD [36] and PolyPhen-2 [29] scores were generated through tools' own modules in VEP platform, remaining tools' scores were annotated using dbNSFP [74] modules in VEP. When the tools did not return any scores for a variant, it is indicated as "Not found". All the scores were converted into binary functional effect predictions as "Benign" and "Pathogenic" according to each tools' pathogenicity thresholds (Table 2). The pathogenicity prediction frequencies of variants in each cancer dataset were calculated based on this binary classification.

Table 1
The number of variants and genes of each cancer dataset retrieved from ClinVar [6].

Cancer type	Number of variants with known significance	Number of variants with uncertain significance	Number of total variants	Number of genes
Bladder	362	966	1328	50
Breast	1559	6101	7660	98
Colon	658	3931	4588	46
Colorectal	137	1888	2025	73
Kidney	246	440	686	42
Liver	85	115	201	29
Lung	114	788	902	12
Pancreas	86	74	160	24
Cancer	3246	14303	17550	257

Table 2
Pathogenicity score thresholds of 13 *in silico* prediction tools used in this study.

Tool Name	Pathogenicity Threshold	Ref.
SIFT	Pathogenic ≤0.05	[33]
CADD	Pathogenic ≥15	[36]
FATHMM-weighted	Pathogenic < −1.5	[30]
FATHMM-unweighted	Pathogenic < −3	[30]
GERP++	Pathogenic >0	[31]
MetaSVM	Pathogenic ≥0.08207	[40]
Mutation Assessor	Pathogenic >1.9	[35]
MutationTaster	Pathogenic ≥0.31733	[32]
MutPred	Pathogenic ≥0.5	[37]
PolyPhen-2	Pathogenic >0.446	[29]
Provean	Pathogenic <2.5	[34]
Revel	Pathogenic ≥0.5	[38]
VEST4	Pathogenic ≥0.75	[39]

2.3. Tool performance assessment and statistical analysis of predictions and scores

ClinVar's clinical consequence data is converted into a binary classification form as "Benign" and "Pathogenic" as well. For the performance assessment of *in silico* tools and statistical classification of variants in different cancer datasets, several metrics such as accuracy, precision, specificity, sensitivity, negative predictive value (NPV), Matthews correlation coefficient (MCC) and false positive rate were utilized with a confusion (contingency) matrix. These performance metrics were evaluated via comparing each tools' predictions with clinical consequence data obtained from ClinVar [6]. Performance metrics were only evaluated for the group of variants with known clinical significance, not for VUSs. The following equations were used, where true positives, true negatives, false positives, and false negatives were shortened as TP, TN, FP, and FN, respectively.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Specificty} = \frac{TN}{TN + FP}$$

$$\text{Sensitivty} = \frac{TP}{TP + FN}$$

$$\text{Negative predictive value (NPV)} = \frac{TN}{TN + FN}$$

$$\text{Matthews correlation coefficient (MCC)} = \frac{TP\ x\ TN - FN\ x\ FP}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}}$$

$$\text{False positive rate (FPR)} = 1 - \text{Specificity}$$

Receiver operating characteristic (ROC) curve analysis was performed using roc curve() function in Scikit-learn [75], a Python package for machine learning, to assess the discrimination power of each tool for the pathogenicity of variants. For this purpose, this analysis was carried out showing the performance metrics in different classification thresholds of each *in silico* method. ROC curve analysis was performed only for the group of variants with known clinical significance.

Correlations between the tools' prediction scores in both groups of variants with known clinical significance and VUSs were calculated using Spearman's rank correlation coefficient method, and heat-maps were created with the Seaborn statistical Python package [76].

3. Results and discussion

3.1. Information about variants in cancer datasets

The dataset (n = 17549) was created from ClinVar for 8 different cancer types including; bladder (n = 1328), breast (n = 7660), colon (n = 4588), colorectal (n = 2025), kidney (n = 686), liver (n = 200), lung (n = 902) and pancreas (n = 160) cancer. The dataset was split into two groups as variants with known clinical significance (n = 3246) and VUSs (n = 14303). For each cancer dataset, gene distribution data (Supplementary Figure-1 and Supplementary Tables 1–8) revealed that the number of tumor suppressor genes is greater than the number of oncogenes among the top 10 genes of each cancer datasets. These same genes were also found to exist in Cancer Gene Census (CGC) [77], which is a curated gene catalogue involving cancer driving mutations in human. This finding confirms the validity of using an open dataset as well.

Distribution of the clinical significance data of the variants as obtained from ClinVar is displayed in Figure-1 and Supplementary Figure-2. To detect the effect of VUS on the performance of the tools, variants with uncertain significance were filtered out in from the group of variants with known significance (Figure-1A). In the dataset, variants containing a 'Pathogenic' keyword in their clinical significance were considered as 'Pathogenic', while variants with 'Benign' keywords were considered as 'Benign'. The distribution of clinical significance in each dataset showed that the highest pathogenic variant percentage is observed in breast cancer (84%, n = 1304), while liver cancer dataset has the lowest (49%, n = 42).

3.2. Pathogenicity prediction frequencies and prediction scores of the variants obtained from in silico tools

For the evaluation of *in silico* prediction tools, we compared the features and sources used in these tools in detail (Table-3). Initially, features and sources were evaluated in five broad categories: DNA sequence, biological function, protein, epigenetics, and transcriptomics. Each category includes sub-topic(s) that are used as parameters in the prediction score calculation. All the other tools use a sequence conservation/identity score. On the other hand, only CADD [36] and GERP++ [31] use the epigenetic and transcriptomics sources and features.

From a general perspective, the prediction distribution frequencies of the variants obtained from *in silico* tools revealed that the benign variant frequencies of GERP++ (19%, n = 614) [31], Mutation Assessor (30%, n = 983) [35], CADD (31%, n = 998) [36] and MutPred (31%, n = 1000) [37] give the most similar results to ClinVar (22%, n = 713) in variants with known clinical significance dataset (Figure-2A). The only feature in scoring and classification that these tools have in common is the sequence conservation/identity score (Table-3), since most prediction tools rely on sequence conservation score in evolutionary aspect to predict the pathogenicity effects [14,78]. Functional annotation in variation databases and protein domain information are other common features among MutPred [37] and CADD [36] except GERP++ [31] and

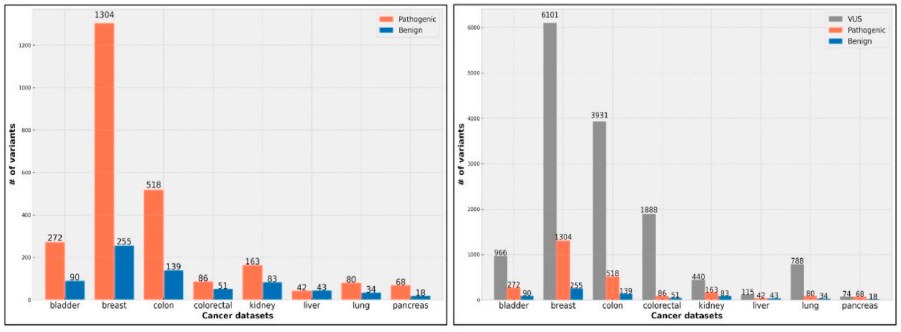

Fig. 1. The distribution of pathogenicity of each cancer datasets **A)** variants with known clinical significance group and **B)** all variants in each cancer datasets separately. Cancer datasets were formed via downloading a total of 17549 variants from ClinVar [6] for 8 different cancer types including bladder, breast, colon, colorectal, kidney, liver, lung and pancreas cancers. Variants containing a 'Pathogenic' keyword in their clinical significance were considered as 'Pathogenic' and variants with 'Benign' keywords were considered as 'Benign' in the dataset.

Table 3
Features and sources of *in silico* prediction tools' scoring and classification.

		SIFT	M.A*	M.P*	C*	Revel	V*	M.T*	M.S*	G*	Pro*	Poly*	F-W*	F-U*
DNA sequence	Sequence conservation/identity score	X	X	X	X	X	X	X	X	X	X	X	X	X
	Predicted mutational rate				X	X		X				X		
Biological Function	Functional annotation in variation database			X	X	X	X	X	X					
Protein	Protein domain			X	X	X		X	X			X	X	X
	Residue function information					X	X				X			
	Protein-specific functional properties			X	X	X	X					X		
	Physiochemical parameters				X	X	X					X		
Epigenetics	Epigenetic Factors				X									
Transcriptomics	Regulatory DNA/RNA sequence information				X					X				
	Gene expression information				X									

•M.A: Mutation Assessor, M.P: MutPred, C: CADD, V: VEST4, M.T: Mutation Taster, M.S: MetaSVM, G: GERP++, Pro: Provean, Poly:PolyPhen-2, F-W:FATHMM-weighted, F-U:FATHMM-unweighted.

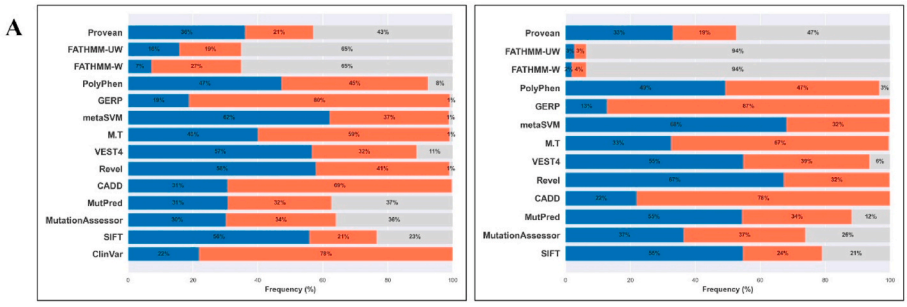

Fig. 2. The frequency distributions of pathogenicity predictions of **A)** variants with known clinical significance and **B)** VUSs in all cancer datasets. The pathogenicity scores of variants with known clinical significance (n = 3246) and VUSs (n = 14303) were retrieved from 13 different *in silico* tools. When no scores were returned from the tools for a variant, it is annotated as "Not found". After retrieval of pathogenicity scores, the scores were annotated as "Benign" (blue) and "Pathogenic" (red) according to each tools' pathogenicity thresholds.

Table 4
Performance assessment metrics of *in silico* tools and statistical classification of variants with known clinical significance group according to each tool's pathogenicity threshold.

	True Positive	True Negative	False Positive	False Negative	Accuracy	Precision	Sensitivity	Specificity	NPV	MCC	FPR(1-Sensitivity)
SIFT	540	493	129	1326	0.415	0.807	0.289	0.793	0.271	0.047	0.207
M.A	928	260	167	723	0.572	0.847	0.562	0.609	0.264	0.091	0.391
M.P	943	211	90	789	0.568	0.913	0.544	0.701	0.211	0.096	0.299
CADD	1800	267	440	731	0.638	0.804	0.711	0.378	0.268	0.067	0.622
Revel	1158	519	171	1362	0.522	0.871	0.46	0.752	0.276	0.107	0.248
VEST4	930	516	115	1326	0.501	0.89	0.412	0.818	0.28	0.116	0.182
M.T	1596	376	317	926	0.613	0.834	0.633	0.543	0.289	0.107	0.457
M.S	1060	561	127	1459	0.505	0.893	0.421	0.815	0.278	0.117	0.185
GERP	2059	151	543	463	0.687	0.791	0.816	0.218	0.246	0.038	0.782
Poly	1250	423	211	1113	0.558	0.856	0.529	0.667	0.275	0.103	0.333
F-W	637	88	255	151	0.641	0.714	0.808	0.257	0.368	0.088	0.743
F-U	478	206	137	310	0.605	0.777	0.607	0.601	0.399	0.156	0.399
Provean	573	276	104	897	0.459	0.846	0.39	0.726	0.235	0.055	0.274

M.A: Mutation Assessor, M.P: MutPred, M.T: Mutation Taster, M.S: MetaSVM, F-W:FATHMM-weighted, F-U:FATHMM-unweighted, NPV: Negative predictive value, MCC: Matthews correlation coefficient, FPR: False positivity rate.

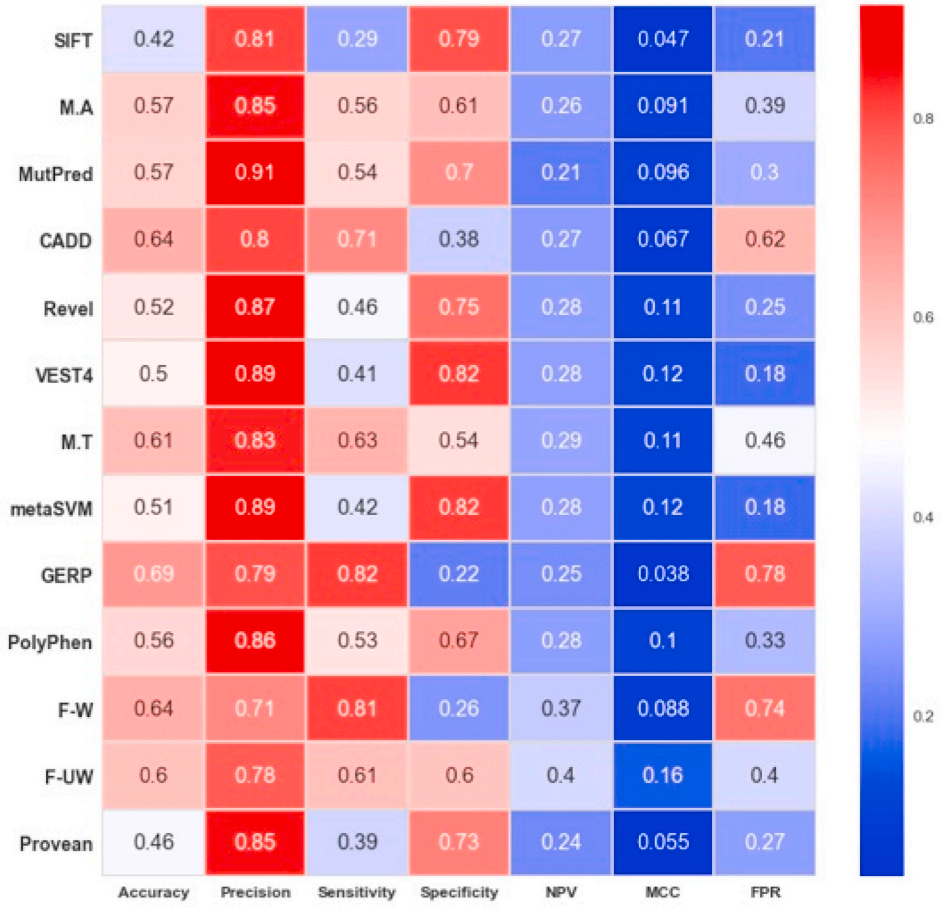

Fig. 3. Heatmap of performance assessment metrics of variants with known clinical significance. M.A: Mutation Assessor, M.P: MutPred, M.T: Mutation Taster, M.S: MetaSVM, F-W:FATHMM-weighted, F-UW:FATHMM-unweighted, NPV: Negative predictive value, MCC: Matthews correlation coefficient, FPR: False positivity rate.

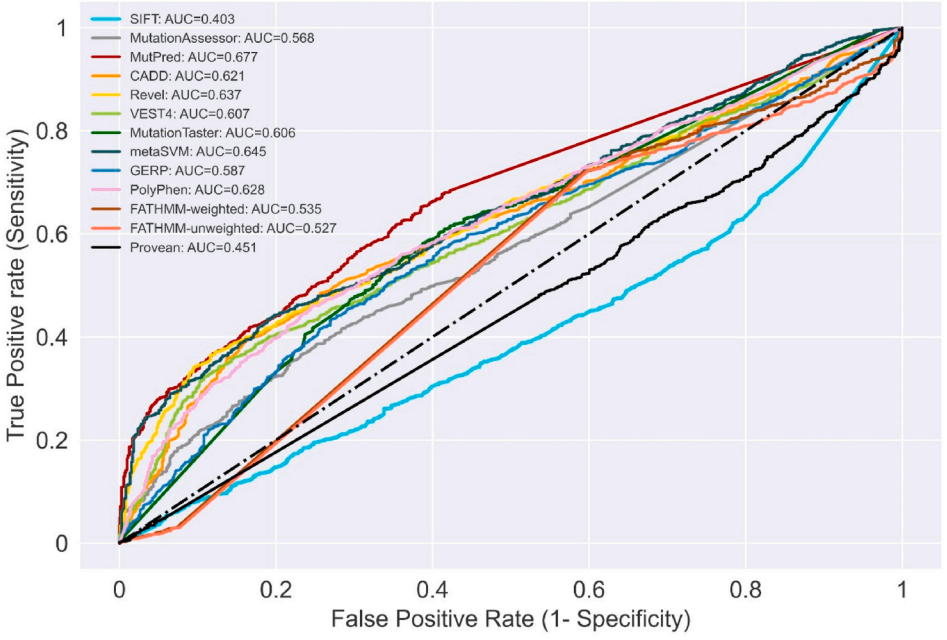

Fig. 4. Receiver operating characteristic (ROC) curve analysis of variants with known clinical significance group with 13 different *in silico* tools The ROC curve analysis was done via using Scikit-learn [75], a Python package for machine learning.

Mutation Assessor [35]. Functional annotation from biological function information was demonstrated to boost the variant effect analysis in both common and rare variants in genetic association studies, thus this feature was utilized in several tools [79]. Protein domain analysis can support variant pathogenicity prediction via investigation of variants in protein families with similar domain functionality [80]. In VUS group, the prediction frequencies of the variants are generally close to the known clinical significance group for each tool (Figure-2B).

3.3. The performance assessment and statistical classification of in silico tools in cancer associated variant datasets

Initially, performance assessment metrics of *in silico* tools and the statistical classification of variants were analyzed using pathogenicity thresholds defined for each tool as given in literature. The performance assessment metrics of variants with known clinical significance group were formed via confusion (contingency) matrix method in all cancer and different cancer types (Table-4,Supplementary Table-9). Heatmaps are used to display the metrics of both datasets (Figure-3, Supplementary Figure-4). Accordingly, the top three tools with the highest sensitivity values in variants with known clinical significance group were GERP++ (0.820), FATHMM-weighted (0.810) and CADD (0.710). However, the top three tools with highest specificity values were VEST4 (0.820), MetaSVM (0.820) and SIFT (0.790). Tools or tests with powerful performance should have higher specificity and sensitivity values at same time [81,82]. Besides this, these values in powerful tools should also be balanced indicating their ratio to each other is close to 1. Among 13 tools used in this study, MutPred, Mutation Assessor and MutationTaster displayed a balanced and high value of sensitivity-specificity, while also having high accuracy values. Thus, it can be deduced that the thresholds of the tools suggested by the authors would not be proper for this dataset due to the low specificity and sensitivity values, so more precise thresholds are needed for this kind of variant datasets [51,54]. To assess performance, Matthews Correlation Coefficient (MCC) is another metric that shows the degree of correlation between observed and predicted binary classification [83]. According to Figure-3, FATHMM-unweighted, MetaSVM and VEST4 have the highest MCC values thus these methods have stronger positive correlations between observed and predicted binary predictions of the methods.

The discriminatory power of the tools was compared via ROC curve analysis in variants with known clinical significance (Figure-4, Supplementary Figure-5). Among all *in silico* prediction tools, top three tools with the highest discriminatory power were found to be MutPred (AUC = 0.677), MetaSVM (AUC = 0.645) and Revel (AUC = 0.637). When ROC curve analysis, performance metrics and prediction distribution frequency results are combined, these tools are found to exhibit the best overall performance. We can conclude that ensemble and supervised learning-based methods have generally higher discriminatory powers than other prediction methods in the prediction of ClinVar cancer-related variant datasets.

3.4. Correlation of the normalized mean scores of the in silico methods

For both VUSs and variants with known clinical significance groups, we calculated Spearman's rank correlation of prediction scores to find out which tool's scores are correlated (Figure-5). First group consists of Mutation Assessor, MutPred, CADD, Revel, VEST4, MutationTaster, MetaSVM, GERP++, PolyPhen, and the other group comprises of Provean, SIFT, FATHMM-weighted and FATHMM-unweighted. In both datasets, 2 tool groups (in Figure-5, marked with dashed squared and straight line squared) were positively correlated within each other and negatively correlated with the other. The first group generally consists of ensemble and supervised learning-based tools, and second group contains sequence conservation or sequence-structure based tools or algorithms. In literature, ensemble and supervised learning methods like MetaSVM and Revel were suggested to be highly correlated with ClinVar deleterious variant dataset [17,51]. Also, another study on correlation analysis of cancer mutations indicate that the algorithms derived from the same study or features end up being correlated [11]. The superiority of the ensemble and supervised learning methods depend on using a combination of multiple predictive features but sequence-structure based methods can only utilize very few parameters such as DNA or protein sequence conservation, biochemical features etc. [38]. In the literature, ensemble and supervised learning methods were recommended to be used for disease-related variants to increase the predictive power [16,84,85].

4. Conclusion

Identifying the functional effects of cancer-related variants are the most common purpose in the era of precision cancer medicine. Many studies have aimed to assess the performance of prediction tools and cancer related variants [69–72] however, to date, none of these aimed to assess the functional relevance of cancer variants obtained from ClinVar [6]. In addition, detailed analysis of cancer related variants from ClinVar have not been performed computationally with *in silico* prediction tools for so many different cancer types up to date. For this purpose, we evaluated the functional relevance and performance metrics of 13

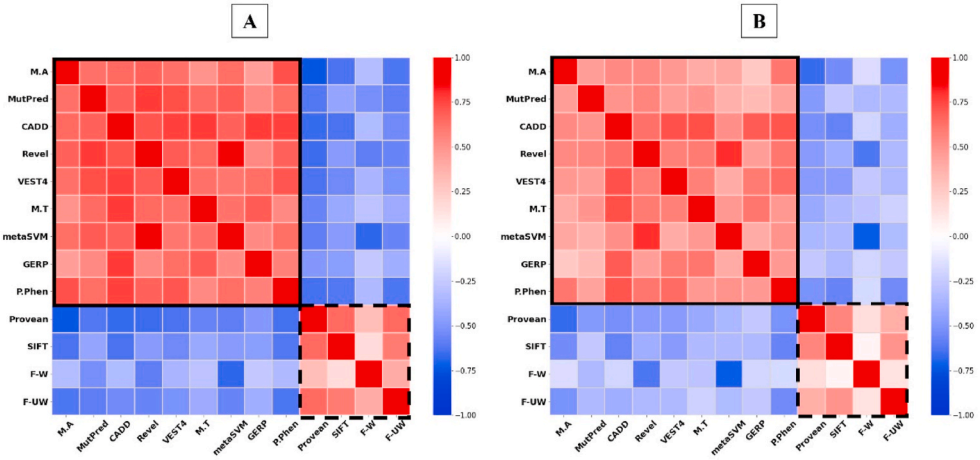

Fig. 5. Heatmap of Spearman's rank correlation of normalized prediction scores of **A)** variants with known significance group and **B)** VUS group in all cancer datasets combined obtained from *in silico* prediction tools. M.A: Mutation Assessor, M.P: MutPred, M.T: Mutation Taster, M.S: MetaSVM, F-W: FATHMM-weighted, F-UW: FATHMM-unweighted.

functional prediction methods on cancer-related variant datasets for 8 cancer types (bladder, breast, colon, colorectal, kidney, liver, lung and pancreas cancer) retrieved from ClinVar.

In comparison of both group of variants, it is shown that prediction distributions of the VUS group are generally close to the known clinical significance group of each tool. According to statistical performance metric analysis, prediction distribution frequency data and ROC curve results, MutPred has the highest discriminatory power for cancer-related variant datasets retrieved from ClinVar. Also, it can be referred that thresholds of each tool would be more flexible for different dataset in order to make an increase in specificity and sensitivity values. Ensemble and supervised learning-based methods have generally higher discriminatory power than other prediction methods for the prediction when ClinVar cancer-related variant datasets are used. Increased predictive performance of ensemble and supervised learning methods in variant pathogenicity prediction is the reason why these methods have higher discriminatory power. Boosted predictive performance of these methods create superiority over the sequence-structure based methods because they utilize a combination of multiple predictive parameters but sequence-structure based methods can only use very few features.

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declaration of competing interest

The authors declare that they have no conflicting interests.

Acknowledgement

PO acknowledges TUSEB Project number 3454.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2022.105434.

References

[1] A. Auton, R.A. Gibbs, E. Boerwinkle, H. Doddapaneni, Y. Han, V. Korchina, A. Rasheed, A global reference for human genetic variation, Nature 526 (7571) (2015) 68–74, https://doi.org/10.1038/nature15393.
[2] K.A. Rich, J. Roggenbuck, S.J. Kolb, Searching far and genome-wide: the relevance of association studies in amyotrophic lateral sclerosis, Front. Neurosci. 14 (January) (2021) 1–11, https://doi.org/10.3389/fnins.2020.603023.
[3] A. Gyulkhandanyan, A.R. Rezaie, L. Roumenina, N. Lagarde, V. Fremeaux-Bacchi, M.A. Miteva, B.O. Villoutreix, Analysis of protein missense alterations by combining sequence- and structure-based methods, Mol. Genet. Genom. Med. (2020) 1–28, https://doi.org/10.1002/mgg3.1166. November 2019.
[4] V. Tam, N. Patel, M. Turcotte, Y. Bossé, G. Paré, D. Meyre, Benefits and limitations of genome-wide association studies, Nat. Rev. Genet. 20 (8) (2019) 467–484, https://doi.org/10.1038/s41576-019-0127-1.
[5] T.G. Kucukkal, M. Petukh, L. Li, E. Alexov, Structural and physico-chemical effects of disease and non-disease nsSNPs on proteins, Curr. Opin. Struct. Biol. 32 (3) (2015) 18–24, https://doi.org/10.1016/j.sbi.2015.01.003.
[6] M.J. Landrum, J.M. Lee, M. Benson, G.R. Brown, C. Chao, S. Chitipiralla, D. R. Maglott, ClinVar: improving access to variant interpretations and supporting evidence, Nucleic Acids Res. 46 (D1) (2018) D1062–D1067, https://doi.org/10.1093/nar/gkx1153.
[7] J.G. Tate, S. Bamford, H.C. Jubb, Z. Sondka, D.M. Beare, N. Bindal, S.A. Forbes, COSMIC: the catalogue of somatic mutations in cancer, Nucleic Acids Res. 47 (D1) (2019) D941–D947, https://doi.org/10.1093/nar/gky1015.
[8] M.X. Li, J.S.H. Kwan, S.Y. Bao, W. Yang, S.L. Ho, Y.Q. Song, P.C. Sham, Predicting mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies, PLoS Genet. 9 (1) (2013) 1–11, https://doi.org/10.1371/journal.pgen.1003143.
[9] P. Sasidharan Nair, M. Vihinen, VariBench: a benchmark database for variations, Hum. Mutat. 34 (1) (2013) 42–49, https://doi.org/10.1002/humu.22204.
[10] L. Ponzoni, I. Bahar, Structural dynamics is a determinant of the functional significance of missense variants, Proc. Natl. Acad. Sci. U. S. A. 115 (16) (2018) 4164–4169, https://doi.org/10.1073/pnas.1715896115.
[11] H. Chen, J. Li, Y. Wang, P.K.S. Ng, Y.H. Tsang, K.R. Shaw, H. Liang, Comprehensive assessment of computational algorithms in predicting cancer driver mutations, Genome Biol. 21 (1) (2020) 1–17, https://doi.org/10.1186/s13059-020-01954-z.
[12] I. Martincorena, P.J. Campbell, Somatic mutation in cancer and normal cells, Science 349 (6255) (2015) 1483–1489, https://doi.org/10.1126/science.aab4082.
[13] F.S. Collins, M.S. Guyer, A. Chakravarti, Variations on a theme: cataloging human DNA sequence variation, Science 278 (5343) (1997) 1580–1581, https://doi.org/10.1126/science.278.5343.1580.
[14] A.J. Marian, Clinical interpretation and management of genetic variants, JACC (J. Am. Coll. Cardiol.): Basic Transl. Sci. 5 (10) (2020) 1029–1042, https://doi.org/10.1016/j.jacbts.2020.05.013.
[15] M. Petukh, T.G. Kucukkal, E. Alexov, On human disease-causing amino acid variants: statistical study of sequence and structural patterns, Hum. Mutat. 36 (5) (2015) 524–534, https://doi.org/10.1002/humu.22770.
[16] M. Yazar, P. Özbek, Silico tools and approaches for the prediction of functional and structural effects of single-nucleotide polymorphisms on proteins: an expert review, OMICS A J. Integr. Biol. 25 (1) (2021) 23–37, https://doi.org/10.1089/omi.2020.0141.
[17] A. Niroula, M. Vihinen, How good are pathogenicity predictors in detecting benign variants? BioRxiv 1–17 (2018) https://doi.org/10.1101/408153.
[18] P.C. Ng, S. Levy, J. Huang, T.B. Stockwell, B.P. Walenz, K. Li, J.C. Venter, Genetic variation in an individual human exome, PLoS Genet. 4 (8) (2008), https://doi.org/10.1371/journal.pgen.1000160.
[19] A. Telenti, L.C.T. Pierce, W.H. Biggs, J. Di Iulio, E.H.M. Wong, M.M. Fabani, J. C. Venter, Deep sequencing of 10,000 human genomes, Proc. Natl. Acad. Sci. U. S. A. 113 (42) (2016) 11901–11906, https://doi.org/10.1073/pnas.1613365113.
[20] S. Richards, N. Aziz, S. Bale, D. Bick, S. Das, J. Gastier-Foster, H.L. Rehm, Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of medical genetics and genomics and the association for molecular Pathology, Genet. Med. 17 (5) (2015) 405–424, https://doi.org/10.1038/gim.2015.30.
[21] A. Garrett, M. Durkie, A. Callaway, G.J. Burghel, R. Robinson, J. Drummond, C. Turnbull, Combining evidence for and against pathogenicity for variants in cancer susceptibility genes: CanVIG-UK consensus recommendations, J. Med. Genet. (2020) 1–8, https://doi.org/10.1136/jmedgenet-2020-107248.
[22] M.M. Li, M. Datto, E.J. Duncavage, S. Kulkarni, N.I. Lindeman, S. Roy, M. N. Nikiforova, Standards and guidelines for the interpretation and reporting of sequence variants in cancer: a joint consensus recommendation of the association for molecular Pathology, American society of clinical oncology, and College of American Pathologists, J. Mol. Diagn. 19 (1) (2017) 4–23, https://doi.org/10.1016/j.jmoldx.2016.10.002.
[23] A. Chatrath, R. Przanowska, S. Kiran, Z. Su, S. Saha, B. Wilson, A. Dutta, The pan-cancer landscape of prognostic germline variants in 10,582 patients, medRxiv 1–18 (2019), https://doi.org/10.1101/19010264.
[24] M.H. Bailey, C. Tokheim, E. Porta-Pardo, S. Sengupta, D. Bertrand, A. Weerasinghe, A. Mariamidze, Comprehensive characterization of cancer driver genes and mutations, Cell 173 (2) (2018) 371–385, https://doi.org/10.1016/j.cell.2018.02.060, e18.
[25] H. Carter, R. Marty, M. Hofree, A.M. Gross, J. Jensen, K.M. Fisch, T. Ideker, Interaction landscape of inherited polymorphisms with somatic events in cancer, Cancer Discov. 7 (4) (2017) 410–423, https://doi.org/10.1158/2159-8290.CD-16-1045.
[26] M.P. Menden, F.P. Casale, J. Stephan, G.R. Bignell, F. Iorio, U. McDermott, O. Stegle, The germline genetic component of drug sensitivity in cancer cell lines, Nat. Commun. 9 (1) (2018) 1–8, https://doi.org/10.1038/s41467-018-05811-3.
[27] T.G. Kucukkal, Y. Yang, S.C. Chapman, W. Cao, E. Alexov, Computational and experimental approaches to reveal the effects of single nucleotide polymorphisms with respect to disease diagnostics, Int. J. Mol. Sci. 15 (2014), https://doi.org/10.3390/ijms15069670.
[28] G. Thiltgen, R.A. Goldstein, Assessing predictors of changes in protein stability upon mutation using self-consistency, PLoS One 7 (10) (2012), https://doi.org/10.1371/journal.pone.0046084.
[29] I.A. Adzhubei, S. Schmidt, L. Peshkin, V.E. Ramensky, A. Gerasimova, P. Bork, S. R. Sunyaev, A method and server for predicting damaging missense mutations, Nat. Methods 7 (4) (2010) 248–249, https://doi.org/10.1038/nmeth0410-248.
[30] H.A. Shihab, J. Gough, D.N. Cooper, P.D. Stenson, G.L.A. Barker, K.J. Edwards, T. R. Gaunt, Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden markov models, Hum. Mutat. 34 (1) (2013) 57–65, https://doi.org/10.1002/humu.22225.
[31] E.V. Davydov, D.L. Goode, M. Sirota, G.M. Cooper, A. Sidow, S. Batzoglou, Identifying a high fraction of the human genome to be under selective constraint using GERP++, PLoS Comput. Biol. 6 (12) (2010) https://doi.org/10.1371/journal.pcbi.1001025.
[32] J.M. Schwarz, D.N. Cooper, M. Schuelke, D. Seelow, Mutationtaster2: mutation prediction for the deep-sequencing age, Nat. Methods 11 (2014) 361–362, https://doi.org/10.1038/nmeth.2890.
[33] N.L. Sim, P. Kumar, J. Hu, S. Henikoff, G. Schneider, P.C. Ng, SIFT web server: predicting effects of amino acid substitutions on proteins, Nucleic Acids Res. 40 (W1) (2012) 452–457, https://doi.org/10.1093/nar/gks539.
[34] Y. Choi, A.P. Chan, PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels, Bioinformatics 31 (16) (2015) 2745–2747, https://doi.org/10.1093/bioinformatics/btv195.
[35] B. Reva, Y. Antipin, C. Sander, Predicting the functional impact of protein mutations: application to cancer genomics, Nucleic Acids Res. 39 (17) (2011) 37–43, https://doi.org/10.1093/nar/gkr407.

[36] P. Rentzsch, D. Witten, G.M. Cooper, J. Shendure, M. Kircher, CADD: predicting the deleteriousness of variants throughout the human genome, Nucleic Acids Res. 47 (D1) (2019) D886–D894, https://doi.org/10.1093/nar/gky1016.

[37] V. Pejaver, J. Urresti, J. Lugo-Martinez, K.A. Pagel, G.N. Lin, H.-J. Nam, P. Radivojac, MutPred2: Inferring the Molecular and Phenotypic Impact of Amino Acid Variants, 2017, 134981, https://doi.org/10.1101/134981. *Doi.*Org.

[38] N.M. Ioannidis, J.H. Rothstein, V. Pejaver, S. Middha, S.K. McDonnell, S. Baheti, W. Sieh, REVEL: an ensemble method for predicting the pathogenicity of rare missense variants, Am. J. Hum. Genet. 99 (4) (2016) 877–885, https://doi.org/10.1016/j.ajhg.2016.08.016.

[39] H. Carter, C. Douville, P.D. Stenson, D.N. Cooper, R. Karchin, Identifying Mendelian disease genes with the variant effect scoring tool, BMC Genom. 14 (Suppl 3) (2013) S3, https://doi.org/10.1186/1471-2164-14-s3-s3. Suppl 3.

[40] J. Zaucha, M. Heinzinger, S. Tarnovskaya, B. Rost, D. Frishman, Family-specific analysis of variant pathogenicity prediction tools, NAR Genom. Bioinform. 2 (2) (2020) 1–8, https://doi.org/10.1093/nargab/lqaa014.

[41] E. Capriotti, R.B. Altman, Y. Bromberg, Collective judgment predicts disease-associated single nucleotide variants, BMC Genom. 14 (Suppl 3) (2013) S2, https://doi.org/10.1186/1471-2164-14-S3-S2.

[42] J. Bendl, J. Stourac, O. Salanda, A. Pavelka, E.D. Wieben, J. Zendulka, J. Damborsky, PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations, PLoS Comput. Biol. 10 (1) (2014) 1–11, https://doi.org/10.1371/journal.pcbi.1003440.

[43] D.K. Brown, Ö. Tastan Bishop, The role of structural bioinformatics in drug discovery via computational SNP analysis – a proposed protocol for analyzing variation at the protein level, Global Heart 12 (2) (2017) 151–161, https://doi.org/10.1016/j.gheart.2017.01.009.

[44] K. Ganesan, A. Kulandaisamy, S. Binny Priya, M. Michael Gromiha, HuVarbase: a human variant database with comprehensive information at gene and protein levels, PLoS One 14 (1) (2019) 1–7, https://doi.org/10.1371/journal.pone.0210475.

[45] K. Higasa, N. Miyake, J. Yoshimura, K. Okamura, T. Niihori, H. Saitsu, F. Matsuda, Human genetic variation database, a reference database of genetic variations in the Japanese population, J. Hum. Genet. 61 (6) (2016) 547–553, https://doi.org/10.1038/jhg.2016.12.

[46] K.J. Karczewski, L.C. Francioli, G. Tiao, B.B. Cummings, J. Alföldi, Q. Wang, D. G. MacArthur, The mutational constraint spectrum quantified from variation in 141,456 humans, Nature 581 (7809) (2020) 434–443, https://doi.org/10.1038/s41586-020-2308-7.

[47] M. Accetturo, N. Bartolomeo, A. Stella, In-silico analysis of NF1 missense variants in clinvar: translating variant predictions into variant interpretation and classification, Int. J. Mol. Sci. 21 (3) (2020) 1–19, https://doi.org/10.3390/ijms21030721.

[48] A.C. Gunning, V. Fryer, J. Fasham, A.H. Crosby, S. Ellard, E.L. Baple, C.F. Wright, Assessing performance of pathogenicity predictors using clinically relevant variant datasets, J. Med. Genet. (2020), https://doi.org/10.1136/jmedgenet-2020-107003 jmedgenet-2020-107003.

[49] M. Lek, K.J. Karczewski, E.V. Minikel, K.E. Samocha, E. Banks, T. Fennell, A. L. Williams, Analysis of protein-coding genetic variation in 60,706 humans, Nature 536 (7616) (2016) 285–291, https://doi.org/10.1038/nature19057.

[50] A. Stella, P. Lastella, D. Loconte, N. Bukvic, D. Varvara, M. Patruno, N. Resta, Accurate classification of NF1 gene variants in 84 Italian patients with neurofibromatosis type 1, Genes 9 (4) (2018) 216, https://doi.org/10.3390/genes9040216.

[51] J. Li, T. Zhao, Y. Zhang, K. Zhang, L. Shi, Y. Chen, Z. Sun, Performance evaluation of pathogenicity-computation methods for missense variants, Nucleic Acids Res. 46 (15) (2018) 7793–7804, https://doi.org/10.1093/nar/gky678.

[52] D.G. Grimm, C.A. Azencott, F. Aicheler, U. Gieraths, D.G. Macarthur, K. E. Samocha, K.M. Borgwardt, The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity, Hum. Mutat. 36 (5) (2015) 513–523, https://doi.org/10.1002/humu.22768.

[53] C. Riera, N. Padilla, X. de la Cruz, The complementarity between protein-specific and general pathogenicity predictors for amino acid substitutions, Hum. Mutat. 37 (10) (2016) 1013–1024, https://doi.org/10.1002/humu.23048.

[54] J. Thusberg, A. Olatubosun, M. Vihinen, Performance of mutation pathogenicity prediction methods on missense variants, Hum. Mutat. 32 (4) (2011) 358–368, https://doi.org/10.1002/humu.21445.

[55] J. Subramanian, R. Simon, Overfitting in prediction models – is it a problem only in high dimensions? Contemp. Clin. Trials 36 (2) (2013) 636–641, https://doi.org/10.1016/j.cct.2013.06.011.

[56] H. Tang, P.D. Thomas, Tools for predicting the functional impact of nonsynonymous genetic variation, Genetics 203 (2) (2016) 635–647, https://doi.org/10.1534/genetics.116.190033.

[57] S.T. Sherry, M.H. Ward, M. Kholodov, J. Baker, L. Phan, E.M. Smigielski, K. Sirotkin, DbSNP: the NCBI database of genetic variation, Nucleic Acids Res. 29 (1) (2001) 308–311, https://doi.org/10.1093/nar/29.1.308.

[58] Consortium, T. 1000 G. P, A global reference for human genetic variation, Nature 526 (7571) (2015) 68–74, https://doi.org/10.1038/nature15393.

[59] A. Sarkar, Y. Yang, M. Vihinen, Variation benchmark datasets: update, criteria, quality and applications, Database (2020) 1–16, https://doi.org/10.1093/database/baz117, 2020.

[60] A. Niroula, M. Vihinen, Harmful somatic amino acid substitutions affect key pathways in cancers, BMC Med. Genom. 8 (1) (2015) 1–12, https://doi.org/10.1186/s12920-015-0125-x.

[61] B.J. Ainscough, M. Griffith, A.C. Coffman, A.H. Wagner, J. Kunisaki, M. N. Choudhary, E.R. Mardis, DoCM: a database of curated mutations in cancer, Nat. Methods 13 (10) (2016) 806–807, https://doi.org/10.1038/nmeth.4000.

[62] D. Chakravarty, J. Gao, S. Phillips, R. Kundra, H. Zhang, J. Wang, N. Schultz, OncoKB: a precision oncology knowledge base, JCO Precis. Oncol. (1) (2017) 1–16, https://doi.org/10.1200/PO.17.00011.

[63] A. Goncearenco, S.L. Rager, M. Li, Q.X. Sang, I.B. Rogozin, A.R. Panchenko, Exploring background mutational processes to decipher cancer genetic heterogeneity, Nucleic Acids Res. 45 (W1) (2017) W514–W522, https://doi.org/10.1093/nar/gkx367.

[64] L.G. Martelotto, C.K.y. Ng, M.R. De Filippo, Y. Zhang, S. Piscuoglio, R.S. Lim, B. Weigelt, Benchmarking mutation effect prediction algorithms using functionally validated cancer-related missense mutations, Genome Biol. 15 (10) (2014) 484, https://doi.org/10.1186/s13059-014-0484-1.

[65] Z. Yue, L. Zhao, J. Xia, DbCPM: a manually curated database for exploring the cancer passenger mutations, Briefings Bioinf. 21 (1) (2018) 309–317, https://doi.org/10.1093/bib/bby105.

[66] C. Ortutay, J. Väliaho, K. Stenberg, M. Vihinen, KinMutBase: a registry of disease-causing mutations in protein kinase domains, Hum. Mutat. 25 (5) (2005) 435–442, https://doi.org/10.1002/humu.20166.

[67] J.M.G. Izarzugaza, A. del Pozo, M. Vazquez, A. Valencia, Prioritization of pathogenic mutations in the protein kinase superfamily, BMC Genom. 13 (Suppl 4) (2012) S3, https://doi.org/10.1186/1471-2164-13-S4-S3. Suppl 4.

[68] F.L. Simonetti, C. Tornador, N. Nabau-Moreto, M.A. Molina-Vila, C. Marino-Buslje, Kin-Driver: a database of driver mutations in protein kinases, Database (2014) 1–5, https://doi.org/10.1093/database/bau104, 2014.

[69] P. Ashford, C.S.M. Pang, A.A. Moya-García, T. Adeyelu, C.A. Orengo, A CATH domain functional family based approach to identify putative cancer driver genes and driver mutations, Sci. Rep. 9 (1) (2019) 1–15, https://doi.org/10.1038/s41598-018-36401-4.

[70] E. Kim, N. Ilic, Y. Shrestha, L. Zou, A. Kamburov, C. Zhu, W.C. Hahn, Systematic functional interrogation of rare cancer variants identifies oncogenic alleles, Cancer Discov. 6 (7) (2016) 714–726, https://doi.org/10.1158/2159-8290.CD-16-0160.

[71] D. Raimondi, A. Passemiers, P. Fariselli, Y. Moreau, Current cancer driver variant predictors learn to recognize driver genes instead of functional variants, BMC Biol. 19 (1) (2021) 1–13, https://doi.org/10.1186/s12915-020-00930-0.

[72] D. Sengupta, G. Bhattacharya, S. Ganguli, M. Sengupta, Structural insights and evaluation of the potential impact of missense variants on the interactions of SLIT2 with ROBO1/4 in cancer progression, Sci. Rep. 10 (1) (2020) 21909, https://doi.org/10.1038/s41598-020-78882-2.

[73] W. McLaren, L. Gil, S.E. Hunt, H.S. Riat, G.R.S. Ritchie, A. Thormann, F. Cunningham, The ensembl variant effect predictor, Genome Biol. 17 (1) (2016) 1–14, https://doi.org/10.1186/s13059-016-0974-4.

[74] X. Liu, C. Wu, C. Li, E. Boerwinkle, dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs, Hum. Mutat. 37 (3) (2016) 235–241, https://doi.org/10.1002/humu.22932.

[75] F. Pedregosa, G. Varoquaux, G. Alexandre, M. Vincent, B. Thirion, O. Grisel, E. Duchesnay, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. (12) (2011) 2825–2830, https://doi.org/10.1289/EHP4713.

[76] M. Waskom, Seaborn: statistical data visualization, J. Open Source Software 6 (60) (2021) 3021, https://doi.org/10.21105/joss.03021.

[77] Z. Sondka, S. Bamford, C.G. Cole, S.A. Ward, I. Dunham, S.A. Forbes, The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers, Nat. Rev. Cancer 18 (11) (2018) 696–705, https://doi.org/10.1038/s41568-018-0060-1.

[78] L. Azevedo, M. Mort, A.C. Costa, R.M. Silva, D. Quelhas, A. Amorim, D.N. Cooper, Improving the in silico assessment of pathogenicity for compensated variants, Eur. J. Hum. Genet. 25 (1) (2016) 2–7, https://doi.org/10.1038/ejhg.2016.129.

[79] X. Li, Z. Li, H. Zhou, S.M. Gaynor, Y. Liu, H. Chen, X. Lin, Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale, Nat. Genet. 52 (9) (2020) 969–983, https://doi.org/10.1038/s41588-020-0676-4.

[80] T.G. Richardson, H.A. Shihab, M.A. Rivas, M.I. McCarthy, C. Campbell, N. J. Timpson, T.R. Gaunt, A protein domain and family based approach to rare variant association analysis, PLoS One 11 (4) (2016) 1–12, https://doi.org/10.1371/journal.pone.0153803.

[81] L.A. McNamara, S.W. Martin, Principles of epidemiology and public health, in: Principles and Practice of Pediatric Infectious Diseases, Fifth Edit), 2018, https://doi.org/10.1016/B978-0-323-40181-4.00001-3.

[82] I.E. Sahin, A. Guclu-Gunduz, G. Yazici, C. Ozkul, M. Volkan-Yazici, B. Nazliel, M. A. Tekindal, The sensitivity and specificity of the balance evaluation systems test-BESTest in determining risk of fall in stroke patients, NeuroRehabilitation 44 (1) (2019) 67–77, https://doi.org/10.3233/NRE-182558.

[83] M. Vihinen, How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis, BMC Genom. 13 (Suppl 4) (2012), https://doi.org/10.1186/1471-2164-13-S4-S2. Suppl 4.

[84] C. Dong, P. Wei, X. Jian, R. Gibbs, E. Boerwinkle, K. Wang, X. Liu, Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies, Hum. Mol. Genet. 24 (8) (2015) 2125–2137, https://doi.org/10.1093/hmg/ddu733.

[85] N. Zhao, J.G. Han, C.R. Shyu, D. Korkin, Determining effects of non-synonymous SNPs on protein-protein interactions using supervised and semi-supervised learning, PLoS Comput. Biol. 10 (5) (2014), https://doi.org/10.1371/journal.pcbi.1003592.