Comparison of pathogenicity prediction tools on somatic variants

Voreak Suybeng, Florence Koeppel, Alexandre Harlé, Etienne Rouleau

Version of Record: https://www.sciencedirect.com/science/article/pii/S1525157820304839
Manuscript_a4e2f2f50b448d3a900db57129d10ef6

Comparison of pathogenicity prediction tools on somatic variants

Voreak Suybeng[1], Florence Koeppel[2], Alexandre Harlé[3], Etienne Rouleau[1]

[1] Gustave Roussy, Département de biologie et pathologie médicales, F-94805, Villejuif, France

[2] Gustave Roussy, Direction de la recherche, F-94805, Villejuif, France

[3] Université de Lorraine CNRS UMR 7039 CRAN, Service de Biopathologie, Institut de Cancérologie de Lorraine, F-54519, Vandœuvre-lès-Nancy, France

Corresponding author: Florence Koeppel

114, rue Édouard-Vaillant

94805 Villejuif Cedex

Email: florence.koeppel@gustaveroussy.fr

This research benefited from a funding by the French National Cancer Institute (INCa), grant number 2017-189.

Disclosures: None declared.

Abstract

Genomic sequencing has been increasingly used over the last decade as part of the management of patients with cancer. Interpretation of somatic variants and their pathogenicity is often complex. Pathogenicity prediction tools are commonly used as part of the expert interpretation of somatic variants, but most of these tools were initially developed for germline variants. The aim of this study was to benchmark their performance on somatic variants. To achieve this, we assembled a « gold standard » list of 4,319 somatic SNVs, classified as oncogenic (N=2,996) or neutral (N=1,323), based either on their presence in curated databases or on their allele frequency (AF) in the general population. We annotated these variants with the most commonly used prediction tools using dbNSFP and UMD-Predictor and we computed performance calculations. The stratification of the prediction tools based on Matthews correlation coefficient and area under the ROC curve allowed to identify the most performing ones, namely CADD, Eigen/Eigen-PC, Polyphen-2, PROVEAN, UMD-Predictor and REVEL. Interestingly, SIFT, which is a commonly used prediction tool for somatic variants, was ranked in the second performance category. Combining tools two by two only marginally improved performances, mainly because of the occurrence of discordant predictions.

Introduction

With the progress of sequencing technologies, the use of sequencing data for cancer patient management has been increasing. Next generation sequencing produces a huge amount of data. Sequencing data are first analyzed and annotated using bioinformatics pipelines. An expert interpretation is then mandatory in order to produce the clinical report, which will be used by the oncologist for discussion during a molecular tumor board and make the therapeutic decision. The interpretation is sometimes challenging. The difficulty is due to the complexity of the somatic context, but also to the larger number of variants typically observed, compared to germline variants. The stake of

a correct variant interpretation is double; first, preventing orientation of a patient towards a therapeutic option, when the tumor does not carry the corresponding pathogenic alteration; second, allowing the orientation of a patient towards a therapeutic option, when the tumor does carry the corresponding pathogenic alteration. Therefore, prediction tools can be very valuable to assist therapeutic decision.

Several computational tools are available to predict the pathogenicity of variants. These tools provide a pathogenicity prediction based on various criteria such as localization within the protein, conservation amongst species, biochemical properties of the mutant and wild-type residues, and the potential impact of the variation on mRNA. Some tools consist in a combination of prediction tools. Their performances have been evaluated for germline variants, variants from the IARC TP53 and from the ICGC databases[1] but to our knowledge, such evaluation is still missing for a large dataset of cancer somatic variants with curated pathogenicity classification. Results from prediction tools are often taken into account at least to some extent in the interpretation process of somatic variants. In some cases, prediction results are even used as inclusion criteria for clinical trials. It is important to understand how reliable these tools are and establish what role they should be given in the process of expert variant interpretation in the context of cancer management and clinical research. Increasing the quality of data interpretation is particularly crucial for variants of unknown significance. The key point is to retain variants potentially of interest in order to enrich associated therapeutic trials and therefore have opportunities to increase our knowledge of these variants. The aim is also to exclude variants which have enough evidence of inactivity.

In the current work, we are addressing the following questions. What is the reliability of prediction tools and databases most commonly used for interpretation of somatic variants? Should prediction tools be combined in order to increase reliability of the results? Is there an optimal

combination of the currently available tools? Do the prediction performances for somatic variants depend on whether the mutation is present on an oncogene or on a tumor suppressor gene?

Materials and Methods

Benchmark dataset of missense variants

Three independent data sources of missense variants were used to construct our benchmark dataset and compare the performance of 22 prediction tools (Table 1).

The first set of missense variants was sourced from the OncoKB database[18] (https://www.oncokb.org, last accessed on 2020, September 10). We downloaded the bulk data release (on 2019, February 23rd) and retained only missense variants classified as oncogenic (N=773), likely oncogenic (N=1,546) or likely neutral (N=501). Overall, the OncoKB database provided N=2,820 missense variants (Supplemental Table 1, worksheet 1). The second set of missense variants was sourced from CIViC database[19] (https://civicdb.org, last accessed on 2020, September 10). We downloaded the bulk data release (on 2019, March 1st) and retained only missense variants with somatic origin and evidence supporting a clinical significance (predictive of response to therapy, prognostic and/or diagnostic) with an A (validated association, N=4) or B (clinical evidence, N=47) evidence level. Overall, the CIViC database provided N=51 missense variants (Supplemental Table 1, worksheet 2). The third set of missense variants was sourced from DoCM database[19] (http://www.docm.info/, last accessed on 2020, September 10). We used the DoCM web interface to download variants (on 2019, March 9th) whose mutation types were missense and whose tags were pathogenic and/or likely pathogenic. Overall, the DoCM database provided N=1,134 missense variants (Supplemental Table 1, worksheet 3).

Before merging the three sets of missense variants into a single dataset, variant annotation was harmonized. Variants within a given gene could be annotated on multiple transcripts even if they were sourced from the same database. For example, *TP53* had variants annotated on five different Ensembl transcripts (https://www.ensembl.org/Homo_sapiens/Info/Index, last accessed on 2020, September 10) when it was sourced from DoCM (ENST00000269305, ENST00000359597, ENST00000413465, ENST00000455263, ENST00000604348). As a consequence, for each gene, we considered a unique Ensembl transcript which was assessed in the OncoKB or DoCM database (Supplemental Table 1, worksheet 4) and we excluded from the three sets all variants which were annotated on a different transcript. We also excluded all variants with conflicting pathogenicity (N=19, Supplemental Table 1, worksheet 5). Finally, we merged the three sets of missense variants into a single dataset of N=3,176 missense variants (Supplemental Table 1, worksheet 6) covering 230 genes from OncoKB, CIViC and DoCM (Figure 1). We divided the missense variants into two groups based on variant pathogenicity: (1) an oncogenic group of variants (N=2,686) which included retained OncoKB variants described as oncogenic or likely oncogenic, and all retained CIViC and DoCM variants; (2) a neutral group of variants (N=490) which included OncoKB variants described as likely neutral.

In this compiled dataset of 3,176 missense variants, only 490 variants covering 83 genes were labeled as neutral and all were retrieved from OncoKB. In order to avoid an unbalanced dataset, we expanded the number of neutral missense variants by adding polymorphisms as follows. Within the assessed transcripts of the 230 genes with oncogenic missense variants in our dataset, we used dbNSFP3.5[20] to identify missense variants whose respective SNVs had an allele frequency (AF) ≥ 0.01 in any subpopulation from gnomAD exomes[21]. Those missense variants were considered as polymorphisms and were included in the neutral group of variants if they were not already present. By adding variants with an AF ≥ 0.01 into our neutral group, additional conflicting variants arose again in our dataset and

consequently, they were excluded as well (N=25, Supplemental Table 1, worksheet 7) from both oncogenic and neutral group.

Thus, after mapping all our missense variants to genomic coordinates using dbNSFP3.5, our final dataset included 4,319 SNVs (Supplemental Table 2) divided into two levels of clinical significance, namely oncogenic variants (N=2,996) and neutral variants (N=1,323). Variants covered 230 genes (Supplemental Table 1, worksheet 4), which were classified as oncogene (N=98), tumor suppressor gene (TSG) (N=97), both (N=6) or unknown (N=29) based on their main mutation effect (gain of function or loss of function) or based on the Cancer Gene Census resources[22]. Figure 2 shows that our dataset contains oncogenic variants which are mostly extremely rare (AF<0.1% in general population) and neutral variants which are extremely rare, rare (between 0.1% and 1% in general population) or common (AF>1% in general population).

Prediction tools and prediction scores

We used dbNSFP v3.5a to annotate the 4,319 SNVs with 20 prediction tools including SIFT, Polyphen-2 HDIV, Polyphen-2 HVAR, LRT, MutationAssessor, FATHMM, PROVEAN, VEST3, MetaSVM, MetaLR, M-CAP, REVEL, MutPred, CADD, DANN, FATHMM-MKL, Eigen, Eigen-PC, GenoCanyon and fitCons (Table 1). We also used authors' publicly available website to annotate the 4,319 SNVs with UMD-Predictor[21]. Cut-off values recommended by dbNSFP or by the tools' authors were used to generate binary predictions for these SNVs (Supplemental Table 3). Note that although MutationTaster2 was also available through dbNSFP, it was not used in our study because it automatically predicts variants as disease causing when they are marked as pathogenic in ClinVar.

We added another prediction algorithm based on the position of the SNV within a functional domain or not. We used dbNSFP to assign the SNV position to a domain annotated in the Interpro

database. If the SNV was located in any known domain, then it was predicted as being oncogenic by the algorithm. On contrary, if the SNV was not located in a known domain, then it was predicted as being neutral. We named this qualitative prediction algorithm a functional domain-based tool.

Overall, we processed predictions from 22 different tools on 4,319 SNVs. Some prediction tools provided multiple scores or predictions for the same SNV due to multiple transcripts for the same gene. Most published articles used the most damaging scores[23] in their analyses or the average scores across all transcripts[11], but we chose to use only the scores specific to the Ensembl transcript assessed in the OncoKB or DoCM database. If a score for a given SNV could not be provided in this assessed transcript, then it was considered as a missing value. Thus, all tools could not provide scores for all SNVs in the assessed transcripts and in these cases, the scores and the predictions were considered as missing values.

When evaluating prediction tools on a dataset which comprises “likely” oncogenic SNVs or polymorphisms, a bias could be introduced in the results. Indeed, “likely” oncogenic SNVs may have been assessed by the very same tools that are being evaluated in this work. And in clinical routine, polymorphisms are generally not assessed by prediction tools, but they are rather filtered out by bioinformatic pipelines based on AF. Therefore, we also processed predictions on a subset of 1,440 SNVs retaining only “clearly” oncogenic SNVs (N=877) and likely neutral SNVs (N=563) from OncoKB database (Supplemental Table 4, worksheet 2).

Finally, we generated all possible pairwise combinations of prediction tools. For a given combination of two tools, a variant prediction was considered as neutral if both tools called it neutral and it was considered as oncogenic if both tools called it oncogenic. However, it was considered as a missing value if both tools within the combination did not provide the same prediction. Combinations were evaluated on the subset of 1,440 SNVs.

Performance measures

The performance of each tool was evaluated as a single tool and as a combination of tools two by two. Accuracy, Matthews correlation coefficient (MCC), sensitivity (Se), specificity (Sp), positive predictive value (PPV), negative predicted value (NPV), false positive rate (FPR) and false negative rate (FNR) were computed based on the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). We also plotted the Receiver Operating Characteristic (ROC) curves and computed the Area Under the Curve (AUC) for different tools by using the R "ROCR" package.

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

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

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

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

$$\text{NPV} = \frac{TN}{TN + FN}$$

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

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

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

Criteria of judgment

MCC is recommended over accuracy as it is less sensitive to the proportion of oncogenic and neutral variants in the benchmark dataset. The ROC curve exhibits the sensitivity and the specificity for

different cut-off values. The AUC value was not computed for LRT because its prediction is not solely determined by its score and for the functional domain-based tool because it is a qualitative prediction tool. When evaluating prediction tools individually, the percentage of missing values was considered acceptable when it was below 5%.

Results

Evaluating on a benchmark database of 4,319 SNVs

Among the SNVs included in this study, 3,559 were sourced from curated somatic cancer variant databases (OncoKB, DoCM, CIViC) to support their classification as either oncogenic (N=2,996) or neutral (N= 563). In addition, 760 supplemental SNVs were considered as neutral because they had an AF≥0.01 in any subpopulation from gnomAD exomes. Hence, the performance of the 22 prediction tools was assessed using these 4,319 SNVs as a set of « gold standard » list of variants based either on their presence in curated databases or on their AF. To our knowledge, none of these tools trained their model using variants sourced from the OncoKB, DoCM or CIViC databases. Missing values, accuracy, MCC, AUC, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), false positive rate (FPR) and false negative rate (FNR) were calculated to evaluate the performance of each prediction tool (Table 2).

For the 4,319 tested SNVs, the tools MutPred, M-CAP, VEST3 and LRT had 1,197 (27.71%), 772 (17.87%), 627 (14.52%), 395 (9.15%) missing prediction scores respectively. For VEST3, this can be explained by the fact that this tool could not provide prediction scores for some SNVs in the assessed Ensembl transcripts. For example, VEST3 could not provide scores for any SNVs in the Ensembl transcript of *ALK* which was assessed in the OncoKB database (ENST00000389048) but could provide scores in other Ensembl isoforms (ENST00000453137 and ENST00000453137) or RefSeq

(https://www.ncbi.nlm.nih.gov/refseq/ , transcript NM_004304). Consequently, performance of these four prediction tools (namely MutPred, M-CAP, VEST3 and LRT) were not further evaluated as their high percentage of missing values (above 5%) could introduce a bias in the interpretation of their results.

The analysis of the accuracy (Table 2, Supplemental Table 5) revealed that the proportion of SNVs classified correctly by the different prediction tools varied substantially (median 77.14%; range from 60.18% to 83.19%). UMD-Predictor was the most accurate (83.19%), while FATHMM was the least accurate (60.18%).

No prediction tools could achieve over 80% sensitivity and specificity simultaneously. FATHMM-MKL achieved highest sensitivity (97.33%) but at the cost of lowest specificity (30.91%). In general, for a given tool, sensitivity was much higher than specificity, suggesting the tendency for all tools to predict somatic SNVs to be oncogenic, as it was already reported for germline variants[1,23].

When evaluating overall performance, the best prediction tools would exhibit simultaneously highest MCC and AUC (Figure 3, Supplemental Figure 1). Hence, UMD-Predictor, CADD, Eigen/Eigen-PC, PROVEAN, REVEL and Polyphen HVAR/HDIV could be considered as the top performing tools (MCC ≥ 0.44 and AUC ≥ 0.78 simultaneously, Table 3). UMD-Predictor outperformed all predictions tools in the 2 metrics simultaneously (MCC = 0.58 and AUC = 0.84).

A second group of predictions tools with moderate overall performance (MCC ≥ 0.36 and AUC ≥ 0.73 simultaneously) could be described and includes SIFT, FATHMM-MKL, MetaSVM, DANN and MetaLR.

A last group of prediction tools with poorest overall performance (MCC ≤ 0.31 and AUC ≤ 0.72 simultaneously) could be described and includes the functional domain-based tool, Mutation Assessor, GenoCanyon, FATHMM and fitCons.

We selected from the benchmark dataset of 4,319 SNVs, either SNVs in oncogenes (N=2,023) or SNVs in TSGs (N=2,013) to assess whether prediction tools would have different performances. When SNVs in oncogenes only were assessed (Supplemental Table 6), UMD-Predictor achieved again best MCC (0.50) and best AUC (0.78). When SNVs in TSGs only were assessed (Supplemental Table 7), UMD-Predictor achieved also best MCC (0.64) whereas REVEL achieved best AUC (0.89). Overall, the three groups of performance level remained unchanged beside REVEL which performed much better for SNVs in TSGs relatively to SNVs in oncogenes in terms of MCC (0.61 vs. 0.36) and AUC (0.89 vs. 0.76). Interestingly, considering MCC and AUC, almost all predictors performed better in TSGs variants compared to oncogenes variants.

Evaluating on a benchmark database of 1,440 SNVs

In the above analysis, we merged SNVs with different levels of pathogenicity assertion (oncogenic and likely oncogenic SNVs from OncoKB database; A and B evidence level SNVs from CIViC database; pathogenic and likely pathogenic SNVs from DoCM database) into a single group of oncogenic variants and we included additional SNVs with AF ≥ 0.01 in the group of neutral variants. To identify if this approach could confound the analysis, performance of the prediction tools was also evaluated on a subset of 1,440 SNVs retaining only “clearly” oncogenic SNVs (N=877) and likely neutral SNVs (N=563) from OncoKB database (Supplemental Table 4, worksheet 2).

On this subset of 1,440 SNVs, the best overall performance tools based on AUC and MCC were confirmed (UMD-Predictor, Eigen, Eigen-PC, CADD, PROVEAN, REVEL, Polyphen-2 HVAR, Polyphen-2 HDIV) (Table 4). But interestingly, the ranking order of the prediction tools performance was not identical. Notably, PROVEAN ranked first with best MCC/AUC and Polyphen-2 HVAR ranked second. In addition, on this subset of 1,440 SNVs, almost all tools exhibited a performance drop in terms of MCC or AUC.

We also assessed whether the combination of prediction tools could improve the performance on this subset of 1,440 SNVs (Supplemental Table 4). Results for combinations of the best overall performance tools with missing values below 300/1,440 are shown in Table 5. In terms of accuracy and MCC, the three best combinations of prediction tools included PROVEAN: PROVEAN and Polyphen-2 HVAR/Eigen/UMD-Predictor. They could outperform the best individual prediction tool PROVEAN alone (accuracy 75.40%, MCC 0.47) due to a gain of sensibility but a loss in specificity. However, the number of missing values, mainly due to prediction discrepancies between two tools, increased drastically (about 20% vs. 0.63%).

Evaluating true concordance

When considering only the best overall performance tools, we found that 372/1,323 neutral SNVs were simultaneously correctly predicted neutral (104/466 SNVs from pure oncogenes and 214/702 SNVs from pure TSG) and 1,802/2,996 oncogenic SNVs were simultaneously correctly predicted oncogenic (902/1,557 SNVs from pure oncogenes and 831/1,311 SNVs from pure TSG).

Evaluating false concordance

There were 180/1,323 neutral SNVs distributed among 70 genes (30 oncogenes, 28 TSG, 2 both, 7 unknown) for which the best overall performance tools simultaneously mispredicted them as oncogenic (Supplemental Table 8, worksheet 1). Among them, 159 SNVs were considered neutral because their respective missense variants were sourced from OncoKB as Likely neutral and 21 SNVs were considered neutral because their AF was greater than 0.01 in any subpopulation from gnomAD exomes. Three of them were also confirmed as Likely benign or Benign by a ClinVar recognized expert panel (three-star review status).

There were 46/2,996 oncogenic SNVs distributed among 218 genes (96 oncogenes, 90 TSG, 6 both, 26 unknown) for which the best overall performance tools simultaneously mispredicted them as neutral (Supplemental Table 8, worksheet 2). All of them were considered oncogenic because their respective missense variants were sourced from OncoKB, CiVIC or DoCM. However, none of them were confirmed Likely pathogenic or Pathogenic by a Clinvar recognized expert panel. It is interesting to note that CDH1 p.T340A was considered Oncogenic by OncoKB but has been recently reviewed and interpreted as benign by an expert panel based on a BA1 criteria of the ACMG/AMP classification. To date, this SNV is still sourced as Oncogenic from OncoKB (data version: v2.7).

Discussion

Performance results

We have shown on a benchmark dataset of 4,319 SNVs that UMD-Predictor consistently achieved the best overall performance as compared to others prediction tools. It achieved best accuracy, best MCC and best AUC. Following UMD-Predictor, the performance was acceptable for Eigen/Eigen PC, CADD, PROVEAN, REVEL and Polyphen-2. A widely used prediction tool as SIFT ranked in the medium performance category.

Better performances were obtained by almost all predictors in TSGs compared to oncogenes. This might be explained by the broad range of inactivating mutations leading to loss-of-function often encountered in TSGs, while oncogenes usually exhibit well-localized hotspots of mutations leading to gain-of-function[22].

Another criterion which is widely used for the interpretation of somatic alterations is the localization of the variant with regards to the domains of the protein. Interestingly, the use of a

prediction tool based on functional domains yielded extremely poor results for oncogenes (MCC=0.09) and better results for TSGs (MCC=0.48). This result suggests that predictions in oncogenes should rather be based on critical hotspots and functional domains of the protein (*e.g.* Tyrosine kinase, protein kinase and SH2 domains) than on any InterPro domain.

We have confirmed on a subset of 1,440 SNVs, without "likely" oncogenic SNVs and additional polymorphisms, that the same prediction tools (CADD, Eigen/Eigen-PC, Polyphen-2, PROVEAN, UMD-Predictor and REVEL) still achieved best overall performance. However, almost all prediction tools exhibited a performance drop due to a loss in specificity without substantial gain in sensitivity. This could suggest a higher ability for prediction tools to assess polymorphisms. In particular, UMD-Predictor prediction relies also on a penalty score for SNVs which are described with a frequency above 0.001 in the general population. This could explain why this prediction tool does not achieve the best overall performance anymore when switching on a dataset without additional polymorphisms. In clinical routine, polymorphisms are generally automatically filtered out by bioinformatic pipelines and are not reported. Consequently, to avoid bias, evaluation of any prediction tools should be made on rare SNVs only. But such a benchmark dataset with a significant number of neutral SNVs is difficult to construct.

We have also highlighted on the subset of 1,440 SNVs, three combinations of prediction tools which all included PROVEAN: PROVEAN + Polyphen-2 HVAR/Eigen/UMD-Predictor. These combinations performed better than any individual prediction tools based on accuracy and MCC. However, this is achieved at the price of missing values (around 20%) and with hardly enough specificity for clinical routine.

Due to inconsistent choice of transcripts between knowledge databases to describe a somatic variant, we selected the Ensembl transcript which was assessed in OncoKB or DoCM database. However, all prediction tools could not always provide scores for this Ensembl transcript. For example, VEST3

could provide scores for *ALK* variants in RefSeq transcript (NM_004304) but could not provide scores for Ensembl transcript (ENST00000389048) and thus, it might have performed with fewer missing values if Refseq transcripts had been considered.

Finally, the risk of circularity needs to be considered. Our benchmark dataset has a limited number of SNVs (N=4,319) and amongst neutral variants (N=1,323), a large proportion (N=760) was included based on an AF ≥ 0.01 in any subpopulation from gnomAD exomes. This subgroup of frequent germline variants could have been included in the training of those prediction tools. To our knowledge, none of the 22 tools trained their model using variants sourced from the OncoKB, DoCM or CIViC databases although we did not examine the possibility of type 1 circularity[24]. In addition, due to the low number of neutral variants sourced in somatic databases, our benchmark dataset is particularly unbalanced towards oncogenic variants (N=2,996) vs. polymorphisms (N=1,323). For this reason, we cannot exclude type 2 circularity[24] either.

Quality of the sourced databases of missense variants

We restricted our benchmark dataset of missense variants to the high-confidence variants sourced from three well-curated somatic databases and to polymorphisms, resulting in a "gold standard" set of 4,319 SNVs, with 2,996 SNVs labeled as oncogenic and 1,323 SNVs labeled as neutral. In the collecting process of missense variants, we raised the question of the quality of the somatic databases due to the relatively high number of missense variants with conflicting pathogenicity and the low number of variants reviewed by an expert panel.

We found 19 missense variants with conflicting pathogenicity (Supplemental Table 1, worksheet 5) which were sourced as Likely neutral from OncoKB but Pathogenic and/or Likely pathogenic from DoCM. We also found 25 missense variants (Supplemental Table 1, worksheet 7) in our combined

dataset which were described as oncogenic but whose population allele frequency was greater than 0.01 in any subpopulation from gnomAD exomes. Among them, 24 were sourced from one single database (either OncoKB or DoCM) but 1 was sourced from both OncoKB and DoCM (ENST00000367921:G505S). Two missense variants, one in *ERCC4* and one in *PMS2*, were even sourced as Oncogenic from OncoKB although their AF was respectively greater than 0.02 in Finnish gnomAD exomes and 0.09 in East Asian gnomAD exomes.

Moreover, among the 4,319 SNVs in our combined dataset, only 139 variants were reviewed by an expert panel on ClinVar database (Supplemental Table 1, worksheet 8). These variants included 96 neutral variants sourced as Benign or Likely benign and 41 oncogenic variants sourced as Pathogenic, Likely pathogenic or Drug response. However, 1 neutral variant in *BRCA1* (ENST00000357654:C1787S) was sourced as Likely neutral from OncoKB but was interpreted as Pathogenic by ENIGMA expert panel, and 1 neutral variant in *TP53* (ENST00000269305:P72R), with an AF>0.73 in NFE gnomAD exomes, was interpreted as Drug response by PharmGKB expert panel.

Although all missense variants with conflicting pathogenicity were carefully filtered out to collect a "gold standard" set of SNVs, the quality of the benchmark dataset is also a limitation of all studies on pathogenicity prediction tools.

Overall, these findings demonstrate that despite huge efforts from well-curated databases, a lot of work remains to be done to catalog and curate somatic variants. To achieve this goal, a consensus guideline on somatic variant interpretation is necessary.

Conclusion

This work helps to better assess the performance of prediction tools for the somatic variant classification. Six prediction tools, namely CADD, Eigen/Eigen-PC, Polyphen-2, PROVEAN, UMD-Predictor

and REVEL, have shown good performance in comparison to others. Performances on tumor suppressor gene variants were better than on oncogene variants. The combination of tools weakly improved performances at the expense of increased missing values. Those tools have to be combined with other classification evidence in order to decide the classification of a variant. The standardization of the classification of somatic variants is really important to achieve, as some discrepancies were identified in the different databases used in this study.

References

1. Li J, Zhao T, Zhang Y, Zhang K, Shi L, Chen Y, Wang X, Sun Z. Performance evaluation of pathogenicity-computation methods for missense variants. Nucleic Acids Res, 2018, 46:7793–804

2. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc, 2009, 4:1073–81

3. Adzhubei I, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods, 2010, 7:248–9

4. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res, 2009, 19:1553–61

5. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: Application to cancer genomics. Nucleic Acids Res, 2011, 39

6. Shihab HA, Gough J, Cooper DN, Day INM, Gaunt TR. Predicting the functional consequences of cancer-associated amino acid substitutions. Bioinformatics, 2013, 29:1504–10

7. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the Functional Effect of Amino Acid Substitutions and Indels. PLoS One, 2012, 7:e466

8. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the Variant Effect Scoring Tool. BMC Genomics, 2013, 14:S3

9. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, Liu X. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet, 2015, 24:2125–37

10. Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN, Bernstein JA, Bejerano G. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nat Genet, 2016, 48:1581–6

11. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A, Li Q, Holzinger E, Karyadi D, Cannon-Albright LA, Teerlink CC, Stanford JL, Isaacs WB, Xu J, Cooney KA, Lange EM, Schleutker J, Carpten JD, Powell IJ, Cussenot O, Cancel-Tassin G, Giles GG, MacInnis RJ, Maier C, Hsieh CL, Wiklund F, Catalona WJ, Foulkes WD, Mandal D, Eeles RA, Kote-Jarai Z, Bustamante CD, Schaid DJ, Hastie T, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet, 2016, 99:877–85

12. Jain P, O'Roak BJ, Cooper GM, Witten DM, Shendure J, Kircher M. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet, 2014, 46:310–5

13. Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day INM, Gaunt TR, Campbell C. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics, 2015, 31:1536–43

14. Ionita-Laza I, Mccallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet, 2016, 48:214–20

15. Lu Q, Hu Y, Sun J, Cheng Y, Cheung KH, Zhao H. A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. Sci Rep, 2015, 5:10576

16. Gulko B, Hubisz MJ, Gronau I, Siepel A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. Nat Genet, 2015, 47:276–83

17. Frédéric MY, Lalande M, Boileau C, Hamroun D, Claustres M, Béroud C, Collod-Béroud G. UMD-

predictor, a new prediction tool for nucleotide substitution pathogenicity - Application to four genes: FBN1, FBN2, TGFBR1, and TGFBR2. Hum Mutat, 2009, 30:952–9

18. Chakravarty D, Gao J, Phillips S, Kundra R, Zhang H, Wang J, Rudolph JE, Yaeger R, Soumerai T, Nissan MH, Chang MT, Chandarlapaty S, Traina TA, Paik PK, Ho AL, Hantash FM, Grupe A, Baxi SS, Callahan MK, Snyder A, Chi P, Danila DC, Gounder M, Harding JJ, Hellmann MD, Iyer G, Janjigian YY, Kaley T, Levine DA, Lowery M, Omuro A, Postow MA, Rathkopf D, Shoushtari AN, Shukla N, et al. OncoKB: A Precision Oncology Knowledge Base. JCO Precis Oncol, 2017, 1:1–16

19. Griffith M, Spies NC, Krysiak K, McMichael JF, Coffman AC, Danos AM, Ainscough BJ, Ramirez CA, Rieke DT, Kujan L, Barnell EK, Wagner AH, Skidmore ZL, Wollam A, Liu CJ, Jones MR, Bilski RL, Lesurf R, Feng YY, Shah NM, Bonakdar M, Trani L, Matlock M, Ramu A, Campbell KM, Spies GC, Graubert AP, Gangavarapu K, Eldred JM, Larson DE, Walker JR, Good BM, Wu C, Su AI, Dienstmann R, et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat Genet, 2017, 49:170–4

20. Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. Hum Mutat, 2016. https://doi.org/10.1002/humu.22932

21. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD, Brand H, Solomonson M, Watts NA, Rhodes D, Singer-Berk M, England EM, Seaby EG, Kosmicki JA, Walters RK, Tashman K, Farjoun Y, Banks E, Poterba T, Wang A, Seed C, Whiffin N, Chong JX, Samocha KE, Pierce-Hoffman E, Zappala Z, O'Donnell-Luria AH, Minikel EV, Weisburd B, Lek M, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature, 2020. https://doi.org/10.1038/s41586-020-2308-7

22. Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat Rev Cancer, 2018, 18:696–705

23. Ghosh R, Oak N, Plon SE. Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. Genome Biol, 2017, 9:111

24. Grimm DG, Azencott CA, Aicheler F, Gieraths U, Macarthur DG, Samocha KE, Cooper DN, Stenson PD, Daly MJ, Smoller JW, Duncan LE, Borgwardt KM. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. Hum Mutat, 2015, 36:513–23

Figure Legends

Figure 1 : Venn diagram representing compiled dataset of missense variants from OncoKB, CIViC and DoCM (N=3,176).

Figure 2 : Distribution of Allele Frequency in general population for variants in our dataset (2,996 oncogenic variants and 1323 neutral variants).

Figure 3. ROC curves (including AUC) for six tools (two tools in each performance group).

Table 1. List of prediction tools.

Tools	Ref	Training data	Type	Criteria
SIFT	[2]	None	Function prediction	Conservation data
PolyphenHDIV	[3]	HumDiv	Function prediction	Conservation data, protein functional domain data and protein structural features
PolyphenHsVAR	[3]	HumVar	Function prediction	Conservation data, protein functional domain data and protein structural features
LRT	[4]		Function prediction	
MutationAssessor	[5]		Function prediction	
FATHMM	[6]	HGMD, SwissProt	Function prediction	Evolutionary conservation, for coding and non-coding variants
PROVEAN	[7]		Function prediction	
VEST3	[8]		Combination	
MetaSVM	[9]		Combination	SIFT, PolyPhen, MutationAssessor…
MetaLR	[9]		Combination	Very similar to MetaSVM, similar performance, more interpretable model
M-CAP	[10]		Combination	
REVEL	[11]	HGMD, rare EPS	Combination	
MutPred			Function prediction	
CADD	[12]	Simulated, Swissvar, HumVar	Combination	Integrates SIFT, GERP++, PolyPhen, CPG distance, GC content
DANN	[11]		Combination	
fathmm-MKL	[13]		Combination	
Eigen	[14]		Combination	
Eigen-PC	[14]		Combination	
GenoCanyon	[15]		Combination	
fitCons	[16]	none	Function prediction	Functional genomics data mainly sourced

				from chromatin analysis and evolutionary conservation data
UMD-Predictor	[17]		Function prediction	Localization within the protein, conservation, biochemical properties of the mutant and wild-type residues, and the potential impact of the variation on mRNA
Functional domain-based tool			Function prediction	Position of the SNV within a functional domain annotated in the Interpro database

Table 2. Performance measures of the prediction tools on a dataset of 4,319 SNVs.

	Missing value	Accuracy	MCC	AUC	Se	Sp	PPV	NPV	FPR	FNR
UMD-Predictor	23 (0.53%)	83.19%	0.58	0.84	93.40%	59.95%	84.15%	79.96%	40.05%	6.60%
Eigen	65 (1.50%)	80.61%	0.52	0.80	91.92%	55.09%	82.19%	75.16%	44.91%	8.08%
Eigen PC	65 (1.50%)	80.16%	0.50	0.80	93.11%	50.96%	81.06%	76.64%	49.04%	6.89%
CADD	0 (0.00%)	79.56%	0.48	0.80	94.29%	46.18%	79.87%	78.13%	53.82%	5.71%
PROVEAN	33 (0.76%)	77.60%	0.50	0.78	80.08%	71.93%	86.71%	61.23%	28.07%	19.92%
REVEL	1 (0.02%)	77.30%	0.49	0.83	80.01%	71.18%	86.29%	61.10%	28.82%	19.99%
Polyphen-2 HVAR	52 (1.20%)	77.41%	0.47	0.80	83.20%	64.30%	84.06%	62.86%	35.70%	16.80%
Polyphen-2 HDIV	52 (1.20%)	77.43%	0.44	0.78	87.83%	53.90%	81.17%	66.20%	46.10%	12.17%
SIFT	93 (2.15%)	76.08%	0.42	0.77	85.06%	55.85%	81.26%	62.42%	44.15%	14.94%
FATHMM-MKL	0 (0.00%)	76.99%	0.41	0.76	97.33%	30.91%	76.14%	83.64%	69.09%	2.67%
MetaSVM	1 (0.02%)	69.25%	0.38	0.78	66.92%	74.51%	85.61%	49.85%	25.49%	33.08%
DANN	0 (0.00%)	74.18%	0.38	0.73	82.41%	55.56%	80.77%	58.24%	44.44%	17.59%
MetaLR	1 (0.02%)	67.99%	0.36	0.79	65.85%	72.84%	84.61%	48.49%	27.16%	34.15%
Mutation Assessor	82 (1.89%)	67.64%	0.31	0.72	69.04%	64.49%	81.39%	48.08%	35.51%	30.96%
GenoCanyon	0 (0.00%)	68.84%	0.23	0.67	81.94%	39.15%	75.31%	48.91%	60.85%	18.06%
Functional domain	0 (0.00%)	72.59%	0.29		89.39%	34.54%	75.56%	58.97%	65.46%	10.61%
FATHMM	17 (0.39%)	60.18%	0.20	0.48	59.34%	62.10%	78.07%	40.18%	37.90%	40.66%
fitCons	65 (1.50%)	61.99%	0.14	0.59	69.12%	45.91%	74.23%	39.74%	54.09%	30.88%
VEST3	627 (14.52%)	79.52%	0.49	0.83	87.38%	60.65%	84.21%	66.67%	39.35%	12.62%
LRT	395 (9.14%)	80.45%	0.51		91.42%	55.43%	82.39%	73.91%	44.57%	8.58%
MutPred	1197 (27.71%)	76.11%	0.16	0.67	81.19%	39.31%	90.64%	22.41%	60.69%	18.81%
M.CAP	772 (17.87%)	80.97%	0.10	0.62	94.56%	12.29%	84.49%	30.90%	87.71%	5.44%

Table 3. Ranking of the evaluated prediction tools on the data set of 4,319 SNVs.

Best overall performance tools	UMD-Predictor Eigen/Eigen PC CADD PROVEAN REVEL Polyphen-2 HVAR/HDIV
Medium overall performance tools	SIFT FATHMM-MKL MetaSVM DANN MetaLR
Low overall performance tools	Mutation Assessor GenoCanyon Functional domain-based tool FATHMM fitCons
Excluded tools (missing values> 5%)	VEST3 LRT MutPred M.CAP

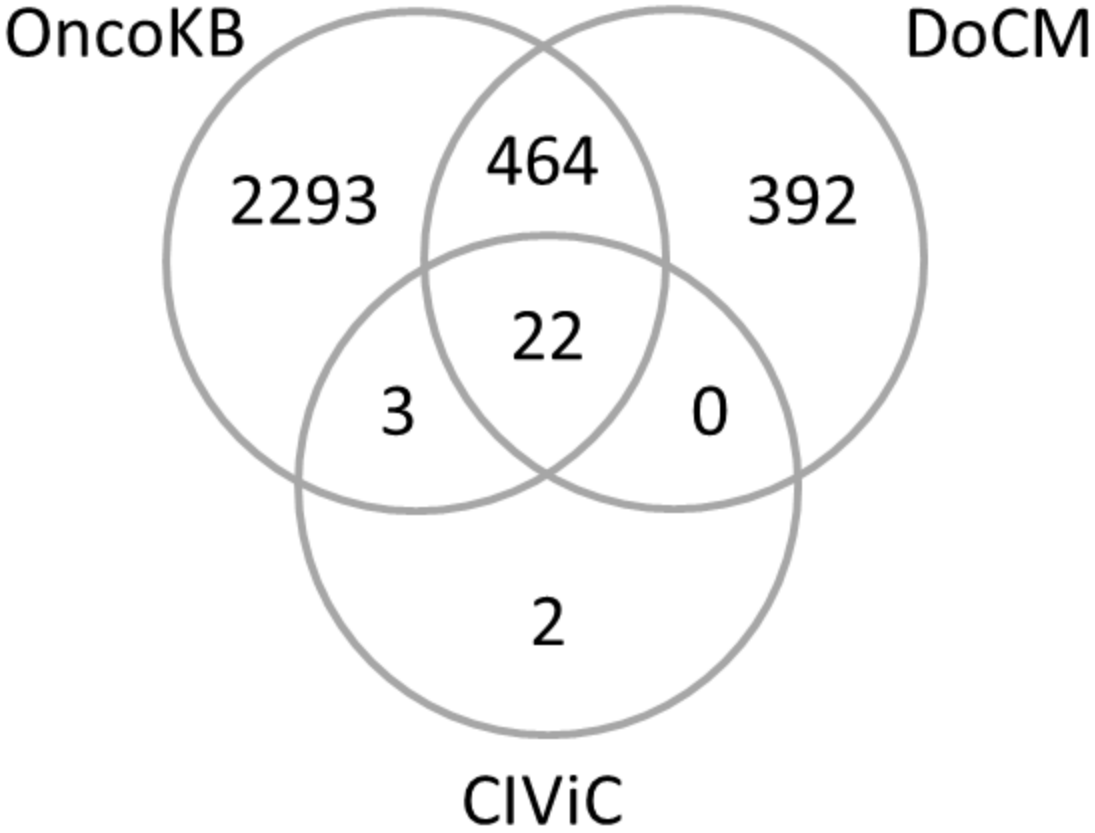
Table 4. Performance measures of the prediction tools on a dataset of 1,440 SNVs.

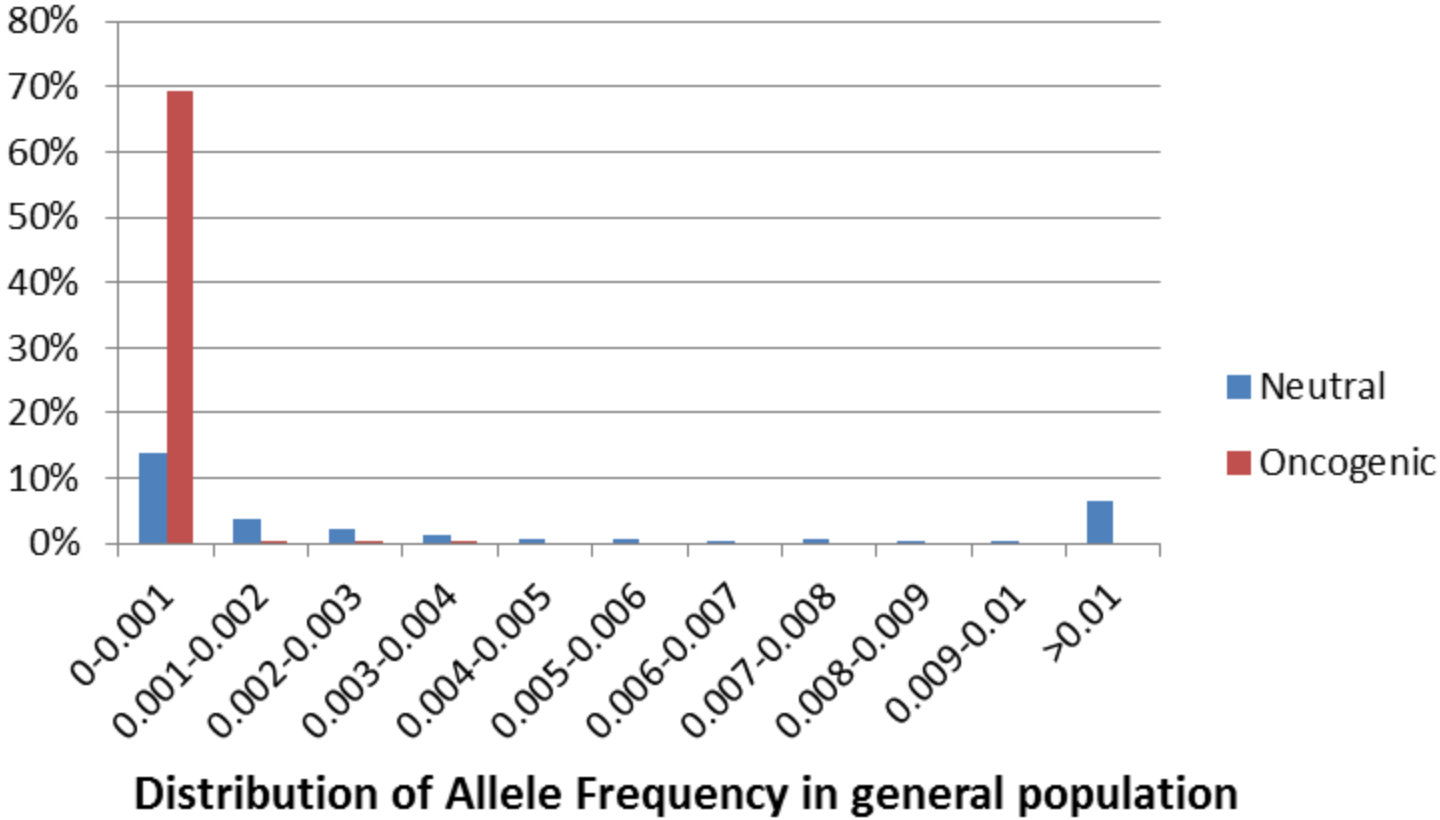
	Missing value	Accuracy	MCC	AUC	Se	Sp	PPV	NPV	FPR	FNR
UMD-Predictor	1 (0.07%)	70.88%	0.38	0.71	95.43%	32.68%	68.81%	82.14%	67.32%	4.57%
Eigen	27 (1.88%)	71.69%	0.40	0.70	94.17%	37.05%	69.75%	80.47%	62.95%	5.83%
Eigen PC	27 (1.88%)	71.83%	0.40	0.70	95.10%	35.97%	69.60%	82.64%	64.03%	4.90%
CADD	0 (0.00%)	69.03%	0.34	0.68	96.35%	26.47%	67.12%	82.32%	73.53%	3.65%
PROVEAN	9 (0.63%)	75.40%	0.47	0.74	86.40%	58.09%	76.44%	73.08%	41.91%	13.60%
REVEL	0 (0.00%)	66.94%	0.28	0.71	82.21%	43.16%	69.26%	60.90%	56.84%	17.79%
Polyphen-2 HVAR	22 (1.53%)	71.79%	0.39	0.72	86.64%	48.83%	72.36%	70.28%	51.17%	13.36%
Polyphen-2 HDIV	22 (1.53%)	69.75%	0.34	0.70	90.71%	37.34%	69.12%	72.22%	62.66%	9.29%
SIFT	18 (1.25%)	66.95%	0.27	0.65	86.21%	36.59%	68.18%	62.73%	63.41%	13.79%
FATHMM-MKL	0 (0.00%)	68.06%	0.32	0.68	98.18%	21.14%	65.98%	88.15%	78.86%	1.82%
MetaSVM	0 (0.00%)	61.32%	0.19	0.64	67.62%	51.51%	68.48%	50.52%	48.49%	32.38%
DANN	0 (0.00%)	66.94%	0.27	0.64	83.92%	40.50%	68.72%	61.79%	59.50%	16.08%
MetaLR	0 (0.00%)	58.82%	0.13	0.62	66.70%	46.54%	66.03%	47.29%	53.46%	33.30%
Mutation Assessor	47 (3.26%)	57.93%	0.13	0.60	62.90%	50.36%	65.88%	47.12%	49.64%	37.10%
GenoCanyon	0 (0.00%)	69.31%	0.33	0.62	89.51%	37.83%	69.16%	69.84%	62.17%	10.49%
Functional domain	0 (0.00%)	64.86%	0.21		92.70%	21.49%	64.78%	65.41%	78.51%	7.30%
FATHMM	1 (0.07%)	54.69%	0.06	0.55	60.73%	45.29%	63.33%	42.57%	54.71%	39.27%
fitCons	27 (1.88%)	52.02%	-0.01	0.51	60.68%	38.67%	60.39%	38.95%	61.33%	39.32%
VEST3	246 (17.08%)	64.24%	0.23	0.71	88.09%	30.78%	64.09%	64.83%	69.22%	11.91%
LRT	193 (13.40%)	73.22%	0.43		94.06%	41.02%	71.13%	81.71%	58.98%	5.94%
MutPred	254 (17.64%)	69.39%	0.23	0.66	83.52%	37.12%	75.22%	49.63%	62.88%	16.48%
M.CAP	35 (2.43%)	64.48%	0.17	0.62	97.14%	10.90%	64.15%	69.88%	89.10%	2.86%

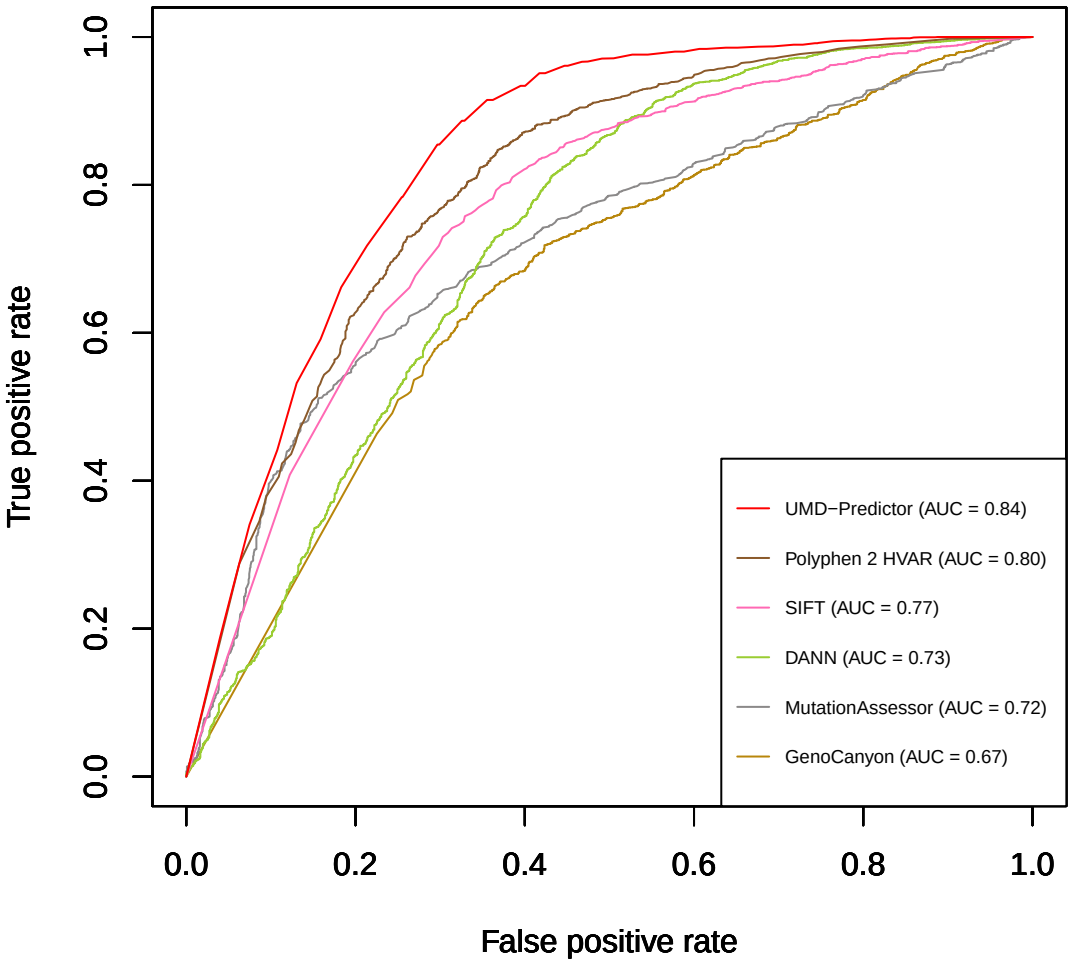
Table 5. Performance measures of the combinations of the best overall performance tools with missing values < 300/1,440.

Tools combination	Missing value	Accuracy	MCC	Se	Sp	PPV	NPV	FPR	FNR
SIFT + Polyphen-2 HDIV	288 (20.0%)	72.31%	0.38	95.05%	33.25%	70.97%	79.66%	66.75%	4.95%
SIFT + REVEL	277 (19.2%)	70.85%	0.33	90.38%	36.94%	71.34%	68.86%	63.06%	9.62%
SIFT + CADD	249 (17.3%)	71.70%	0.36	97.11%	26.57%	70.14%	83.82%	73.43%	2.89%
SIFT + Eigen	286 (19.9%)	73.48%	0.40	96.34%	33.09%	71.79%	83.64%	66.91%	3.66%
Polyphen-2 HDIV + Polyphen-2 HVAR	131 (9.1%)	72.50%	0.39	90.81%	42.19%	72.22%	73.50%	57.81%	9.19%
Polyphen-2 HDIV + CADD	203 (14.1%)	72.03%	0.36	96.99%	26.82%	70.59%	83.10%	73.18%	3.01%
Polyphen-2 HDIV + FATHMM-MKL	263 (18.3%)	72.64%	0.37	99.35%	21.14%	70.84%	94.44%	78.86%	0.65%
Polyphen-2 HDIV + Eigen	219 (15.2%)	73.30%	0.41	95.99%	34.15%	71.55%	83.15%	65.85%	4.01%
Polyphen-2 HDIV + Eigen PC	251 (17.4%)	74.01%	0.43	97.62%	32.95%	71.69%	88.82%	67.05%	2.38%
Polyphen-2 HDIV + UMD	235 (16.3%)	73.78%	0.41	97.56%	30.28%	71.90%	87.16%	69.72%	2.44%
Polyphen-2 HVAR + PROVEAN	278 (19.3%)	78.66%	0.52	92.07%	54.78%	78.38%	79.51%	45.22%	7.93%
Polyphen-2 HVAR + CADD	266 (18.5%)	74.45%	0.41	96.73%	32.93%	72.88%	84.38%	67.07%	3.27%
Polyphen-2 HVAR + Eigen	260 (18.1%)	75.34%	0.46	95.43%	41.06%	73.42%	84.04%	58.94%	4.57%
REVEL + CADD	292 (20.3%)	72.56%	0.36	96.24%	28.78%	71.41%	80.56%	71.22%	3.76%
REVEL + Eigen	293 (20.3%)	73.58%	0.40	94.41%	36.56%	72.57%	78.65%	63.44%	5.59%
PROVEAN + Eigen	292 (20.3%)	78.75%	0.51	95.50%	46.56%	77.44%	84.33%	53.44%	4.50%
PROVEAN + UMD	291 (20.2%)	79.03%	0.51	96.86%	43.64%	77.32%	87.50%	56.36%	3.14%
CADD + FATHMM-MKL	130 (9.0%)	70.38%	0.34	99.52%	18.82%	68.45%	95.70%	81.18%	0.48%
CADD + Eigen	143 (9.9%)	72.01%	0.39	97.43%	28.75%	69.95%	86.79%	71.25%	2.57%
CADD + Eigen PC	157 (10.9%)	72.33%	0.40	98.39%	27.85%	69.95%	91.03%	72.15%	1.61%
CADD + UMD	172 (11.9%)	72.63%	0.37	98.67%	23.98%	70.81%	90.60%	76.02%	1.33%
FATHMM-MKL + Eigen	179 (12.4%)	72.09%	0.36	98.53%	23.60%	70.28%	89.74%	76.40%	1.47%
FATHMM-MKL + Eigen PC	155 (10.8%)	71.75%	0.36	98.43%	23.75%	69.91%	89.34%	76.25%	1.57%
FATHMM-MKL + UMD	147 (10.2%)	71.69%	0.36	99.16%	21.62%	69.76%	93.40%	78.38%	0.84%
Eigen + Eigen PC	75 (5.2%)	72.53%	0.42	96.14%	36.01%	69.91%	85.78%	63.99%	3.86%
Eigen + UMD	200 (13.9%)	74.03%	0.43	98.36%	30.87%	71.63%	91.39%	69.13%	1.64%

Eigen PC + UMD	184 (12.8%)	73.81%	0.43	98.38%	30.55%	71.38%	91.45%	69.45%	1.62%


OncoKB
DoCM
2293
464
392
22
3
0
2
CIViC


80%
70%
60%
50%
40%
30%
20%
10%
0%
0-0.001
0.001-0.002
0.002-0.003
0.003-0.004
0.004-0.005
0.005-0.006
0.006-0.007
0.007-0.008
0.008-0.009
0.009-0.01
>0.01
Neutral
Oncogenic
Distribution of Allele Frequency in general population


True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
UMD−Predictor (AUC = 0.84)
Polyphen 2 HVAR (AUC = 0.80)
SIFT (AUC = 0.77)
DANN (AUC = 0.73)
MutationAssessor (AUC = 0.72)
GenoCanyon (AUC = 0.67)