A comprehensive and bias-free evaluation of genomic variant clinical interpretation tools

Minh Trang, Nguyen§
Center for Biomedical Informatics
Vingroup Big Data Institute
Hanoi, Vietnam
0000-0002-0438-8538

Anh Vu, Mai Nguyen§
Center for Biomedical Informatics
Vingroup Big Data Institute
Hanoi, Vietnam
0000-0003-1270-2512

Hoang Anh, Tran§
Center for Biomedical Informatics
Vingroup Big Data Institute
Hanoi, Vietnam
0000-0003-4814-7370

Nguyet Minh, Do§
Center for Biomedical Informatics
Vingroup Big Data Institute
Hanoi, Vietnam
0000-0002-0761-074X

Thanh Nguyen, Nguyen
Center for Biomedical Informatics
Vingroup Big Data Institute
Hanoi, Vietnam
0000-0001-9578-3420

Abstract—The advancement of Next Generation Sequencing (NGS) generates a huge pool of raw sequencing data and genomic variants, while the diverse selection of variant annotation tools adds even more confusion to the mix. Choosing the right tools for clinical interpretation of genomic variants is still challenging due to the lack of comprehensive evaluation studies in this field. Here, we introduced a bias-free analysis approach to assess ten well-known variant annotation tools in terms of clinical interpretation. Our results revealed notable correlations of contemporary methods when applied to the ClinVar dataset. Moreover, allele frequency is still a strong predictor, emphasizing the importance of biological insight in the prediction of clinical interpretation. Our analysis and evaluation scripts are available for public use at https://github.com/nmtrang00/VarAnnotEval.

Index Terms—human genome, genomic variants, variant annotation, feature evaluation, predictive power score, data curation

I. INTRODUCTION

The leap in Next Generation Sequencing (NGS) technology has leveled the ground for data generation and application of genomic sequencing around the world. Instead of a global effort, billion dollars cost, and years of experiment, the whole genome sequencing (WGS) now can be obtained in less than one day and cost roughly hundreds of dollars [1]. A typical WGS run can generate around 90-100 Gigabases of raw data per genome which yield around 5 million variants. In comparison, a whole-exome sequencing (WES) run can yield around 60 thousand variants from 10-15 Gigabases of raw data.

This rapid advancement, however, has caused the mining of meaningful biological variation to be lagged behind a massive pool of raw sequencing data. Extraction of biological information from NGS is not an easy task due to many reasons, e.g., a single variant can be the cause for one Mendelian disease [2]. To bridge the gap, tremendous efforts are spent on developing annotation tools for genomic variants [3], [4]. While each tool

§Equal contribution. Author ordering determined by a list randomizer

has its assumptions and regions of application, they can be divided into several main groups based on their predicting features: conservation score, deleteriousness, regulatory, and splicing site prediction. The annotation methods also evolve over time, from rule-based [5], [6] to ensemble machine learning [7]–[9] and deep learning neural network [10], [11], emphasizing a paradigm-shifting in development of variant annotation. Nevertheless, to some, more tools mean more data to consider, while others may be biased toward conventional tools, and vice versa.

Databases of variants become substantial sources of reference, especially for disease-causing variants. Launched in 2013, ClinVar is a public database focusing on human variation - phenotype relationship with supporting evidence [12]. Through 8 years of community submission and board curation, the database is now holding over 1.6 million submissions over 1 million unique variations. As a community-based effort, ClinVar has an assertion criteria system, referred to as ClinVar gold star. The value of this metric has a range of 1 to 4 stars for each variant, the higher the number of stars, the higher the reliability of the variant-phenotype interpretation.

In this work, we not only reviewed a variety of popular annotation tools but also presented a unified strategy for evaluation and assessment for their ease of use, regions of application, and correlation on the ClinVar dataset. Our evaluation should be helpful for not only choosing proper variant clinical interpretation tools for the field of interest but also improving these tools in terms of their annotation reliability.

II. MATERIALS AND METHODS

A. Variant Annotation Tools

In this study, we considered ten popular annotation tools divided into three groups according to their type of prediction feature: conservation, deleteriousness, and splicing site prediction. We evaluated them by their user interface, supported genome build, and target regions. Firstly, for user interface,
they all offer command-line programs or pre-computed data, which are ready to download from their website or GitHub, except for MutationTaster. Another option is web-based programs, where users can submit their variants and get results through the tools’ server. However, the latter option can be a bottleneck in high-throughput analysis, as their servers only support a limited number of variants in a single submission (around 100,000 variants for CADD, 5 million SNPs for PolyPhen-2, etc.) Secondly, the human genome GRCh37 build is supported by all mentioned tools, while only six out of ten have updates for GRCh38. The following describes the above-selected tools divided into three aforementioned types.

1) **Conservation scores:** Although conservation scores do not directly imply the pathogenicity of variants, they offer an insight into whether a variant lies within the functional region or not; they show us the conservation or vulnerability to changes of the reference allele under evolutionary constraints. A common approach of different tools is to identify sequences shared among various species, conducted by a multiple sequence alignment analysis. Given the phylogenetic information of these species, suitable models of evolution are implemented to calculate the final score. GERP++, PhyloP [14], [15] calculated scores for each individual alignment column, without consideration of neighboring sites. Each of them uses only a model (a model of substitution and a model of natural evolution ("non-conserved") respectively). Meanwhile, PhastCons [16]–[20] proposes two models: one for "conserved" and another for "non-conserved", which infers more initial assumptions.

2) **Deleteriousness scores:** A vast majority of prediction tools aim to predict whether a variant is harmful or detrimental to the overall function of the gene. The convention tools often apply various levels of assumption to simplify the interlaced network of different biological processes. An emerging method is an ensemble learning approach: it selects various types of metrics from existing tools (SIFT4G, Polyphen2, CADD, MutationTaster, Mutation Assessor, FATHMM, LRT, MetaLR, MetaSVM, RVIS, PhyloP, Phastcons, PAM250, Blosum62, SIPHY, GERP, etc.), especially the new conservation score matrices from Multiz [25]. MutationTaster [26] utilizes 5 specific ensemble models for variants in different regions of genes, for non-coding variants, variants causing only an amino acid (aa) change, variants causing more than one aa change, and variants in 3’ and 5’ UTR. For each model, MutationTaster tunes the number of Random Forest trees for optimal accuracy and run-time performance. However, if there is a conflict with the annotation of ClinVar Significance, MutationTaster takes ClinVar Significance as the final result.

Unlike other tools that rely on validated datasets, CADD [7], [8] defines its own training sets: "proxy-neutral" variants from phylogenetically conserved variants and "proxy-deleterious" de novo variants with no selective pressure. As a result, the fitting of the assembly feature and the scoring of both coding and non-coding variants are optimized. More than 60 annotations are used to train the logistic regression model. Judging by its performance on the ClinVar and ExAC dataset, it boasts an area under the receiver operating characteristic of 98.10% (SNV, MAF > 0.05) and 90.08 (gene-wise, SNV, MAF > 0.05), surpassing existing tools such as DANN, Eigen, FATHMM-XF, MutationTaster. Another fresh air for missense variant annotation is PrimateAI, a semi-supervised benign vs unlabeled training regimen that is completely different from the above tools.

3) **Splice effect scores:** Mutations at splice sites are of high importance because of their RNA splicing alternation impact, leading to the lengthen or shorten of the transcript and an altered protein-coding sequence. With the implementation of deep learning, SpliceAI [11], [27] determines the probability of every base within the pre-mRNA sequences is splice donor, splice acceptor, or neither. Another splicing alternation score is the AdaScore from dbscSNV [28]. Four in-silico tools for splicing regions including Position Weight Matrix (PWM) model, MaxEntScan (MES), Splice Site Prediction by Neural Network (NNSplice), GeneSplicer Human Splicing Finder
(HSF) were used as features in the AdaBoost model. They all have missing rates lower than 30%. Further improvement to AdaScore is made by adding PhyloP and CADD scores to the model, which continue to promote the new model AUC from 0.963 to 0.977. This boosting model outperforms all prediction tools when using individually.

B. Evaluation method

In order to assess tools’ performance, we evaluated their prediction results and their ClinVar interpretation correlation by model-based approaches. We treated the annotation scores as features, while the ClinVar label was the target, hence we could estimate the importance of each feature to the prediction target. The feature importance could shed light on the underlying relationships of the outcomes and a set of features, especially in the biomedical field [29], [30]. The interpretation of machine learning plays a crucial role in gaining the trust of novices and experts alike in the outcome of systems [31]–[33]. In this work, we chose two methods to determine the strength of association of features in our curated dataset. The first one was the predictive power score (PPS) [34] and the second one was the association score.

Essentially, PPS uses a simple decision tree to calculate relation scores. Applying a decision tree in practice comes with several benefits. Firstly, it does not require rigid data preprocessing. In other words, PPS could be run with data at the very first stage to have a quick glimpse through the data. Besides, the decision tree is the one that can easily handle outliers and prevent itself from overfitting [35], [36]. By using this algorithm, both numeric and class scores can be addressed in the same manner. In addition, PPS is asymmetric, it can reveal the relationship in both directions separately. The advantages of PPS come at a cost of the demanding computation resources.

The association score consisted of three measure kinds which are applied for continuous-continuous, categorical-continuous, and categorical-categorical cases. In the first case, Pearson’s r was used to find the correlation. In the latter, the association score was calculated by correlation ratio $\eta$, also a test of linearity [37]. In the last one, there were two measures: Cramér’s V [38] and Theil’s U uncertainty coefficient [39]. However, unlike PPS, this technique suffers from a drawback: it cannot detect non-linear bi-variate relationships between two numeric variables without extensive preprocessing.

III. PERFORMANCE ON CURATED DATASET

A. Data preprocessing

Toward a bias-free evaluation of annotation tools, we set out to standardize different scores and preprocess our reference set. We used ClinVar public dataset version 20210315 [12] as target for assessment. Specifically, only variants clinically interpreted as benign, likely benign, pathogenic, or pathogenic were selected from the ClinVar public dataset. Only approximately 380,000 variants passed this filter, and the remaining variants were searched for their presence in the M-CAP, CADD, MutationTaster, and the HGMD variant training dataset. The reason that variants existing in HGMD were filtered out is that a lot of tools use HGMD for their negative training set; for example, M-CAP used all HGMD variants with $MAF < 0.05$, FATHMM use all missense HGMD variants, and MutationTaster also use HGMD variants. The remaining 270,000 variants were annotated with selected tools and separated into SNP and Indel for evaluation. All variants were also annotated with their allele frequency queried from an unrelated subset of 1000 Genomes Project, resequenced at 30X coverage by New York Genomic Center [40].

As mentioned above, the final prediction of MutationTaster was taken from ClinVar interpretation if there were any conflicts. For fair evaluation, we took the simplified Mutation-Taster tree vote as the feature. For SIFT4G, the scoring and labeling were different among transcripts. In our evaluation, we only considered the most severe cases with the score closest to 1 if "tolerated" and 0 if "deleterious". Following M-CAP proposal [24], all missense variants not in the pre-computed database of M-CAP were scored 1. The missing allele frequency was filled by 0. For feature evaluation, ClinVar variants with labels not containing "Benign", "Likely Benign", "Pathogenic" or "Likely pathogenic" were excluded from the set. If the label contains both "Benign" and "Likely Benign", the label was simplified to "Likely Benign." A similar procedure was done with "Pathogenic". The standardized label for ClinVar interpretation was called "Clinsig model" in our analysis. For more information regarding our notation for each score name, please refer to Table I.

The number of missing entries was considerably high for some tools. To allow for an unbiased assessment of these tools, and because the PPS score cannot be used with missing values, all NAs were removed. The missing data can be caused largely by the fact that the tools were designed for specific regions of the genome. For example, M-CAP is designed to score the pathogenicity of rare alleles. Thus, it does not make sense to blindly fill all the missing values, as they are simply not relevant. One exception is M-CAP, whose missing missense data can be filled with 1 - M-CAP assumes that all high-frequency missense alleles are putatively neutral. In the next section, we will compare the performance of M-CAP, PrimateAI, PolyPhen2, SIFT4G with respect to the missense variants.

B. Annotation evaluation

In this work, we applied PPS to determine a set of important features. Further investigation was conducted using the association score. Each analysis was performed on the separated SNP and Indel sets. Table I shows PPS, correlation ratio $\eta$, Cramer’s V and Theil index for each dataset.

Overall, from the PPS of the Indel set, MT_treevote, CADD, and CADD_phred were observed as a collection of attributes strongly correlating with the Clinsig model. The association scores strengthened the evidence for MT_treevote, CADD, and CADD_phred while adding gERP_rs, phast-Con46, phyloP46, and AF_HC to the list. Consistently, MT_treevote, and CADD_phred were still among the
TABLE I: Annotation features and ClinVar interpretation relationship. Annotation tool groups: a) Conservation score, b) Deleteriousness, c) Splice site

<table>
<thead>
<tr>
<th>Tool</th>
<th>Feature names</th>
<th>Information</th>
<th>SNP</th>
<th>PPS</th>
<th>Correlation μ</th>
<th>Cramer’s V</th>
<th>Theil</th>
<th>Indel</th>
<th>PPS</th>
<th>Correlation μ</th>
<th>Cramer’s V</th>
<th>Theil</th>
</tr>
</thead>
<tbody>
<tr>
<td>GERP++</td>
<td>gERP.rs</td>
<td>GERP++ RS score</td>
<td>0.09</td>
<td>0.35</td>
<td>-</td>
<td>-</td>
<td>0.16</td>
<td>0.35</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PhyloP</td>
<td>phyloP46</td>
<td>Conservation scoring</td>
<td>0.12</td>
<td>0.56</td>
<td>-</td>
<td>-</td>
<td>0.08</td>
<td>0.38</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PhastCon</td>
<td>phastCon46</td>
<td>Probability conserved nucleotide</td>
<td>0.01</td>
<td>0.31</td>
<td>-</td>
<td>-</td>
<td>0.13</td>
<td>0.50</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Treevote</td>
<td>fathmm-xf</td>
<td>Diversity of the sequences used for prediction</td>
<td>0.01</td>
<td>0.08</td>
<td>-</td>
<td>-</td>
<td>0.01</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SIFT4G</td>
<td>SIFT_score</td>
<td>SIFT score</td>
<td>0.09</td>
<td>0.5</td>
<td>-</td>
<td>-</td>
<td>0.01</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SIFT4G</td>
<td>SIFT_piechart</td>
<td>C-score</td>
<td>0.01</td>
<td>0.08</td>
<td>-</td>
<td>-</td>
<td>0.01</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Polyphen-2</td>
<td>Polyphen-2.HumDiv</td>
<td>Polyphen2 score based on HumDiv.</td>
<td>0.13</td>
<td>-</td>
<td>0.31</td>
<td>0.16</td>
<td>0.01</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Polyphen-2</td>
<td>Polyphen-2.HumVar</td>
<td>Polyphen2 score based on HumVar</td>
<td>0.14</td>
<td>-</td>
<td>0.34</td>
<td>0.19</td>
<td>0.01</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MutationTaster</td>
<td>MT_treevote</td>
<td>Ratio of agreed trees</td>
<td>0.19</td>
<td>0.91</td>
<td>-</td>
<td>-</td>
<td>0.45</td>
<td>0.96</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CADD</td>
<td>CADD</td>
<td>CADD deleteriousness score</td>
<td>0.15</td>
<td>0.82</td>
<td>-</td>
<td>-</td>
<td>0.27</td>
<td>0.75</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CADD</td>
<td>CADD Scores</td>
<td>CADD relative rank</td>
<td>0.22</td>
<td>0.76</td>
<td>-</td>
<td>-</td>
<td>0.32</td>
<td>0.81</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FATHMM-MKL</td>
<td>fathmm-mkl_C.score</td>
<td>Deleterious probability as in coding region</td>
<td>0.03</td>
<td>0.01</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>0.02</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FATHMM-MKL</td>
<td>fathmm-mkl_NC.score</td>
<td>Deleterious probability as in non-coding region</td>
<td>0.02</td>
<td>0.07</td>
<td>-</td>
<td>-</td>
<td>0.03</td>
<td>0.06</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FATHMM-XF</td>
<td>fathmm-xf_C.score</td>
<td>Deleterious probability as in coding region</td>
<td>0.09</td>
<td>0.56</td>
<td>-</td>
<td>-</td>
<td>0.01</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FATHMM-XF</td>
<td>fathmm-xf_NC.score</td>
<td>Deleterious probability as in non-coding region</td>
<td>0.25</td>
<td>0.61</td>
<td>-</td>
<td>-</td>
<td>0.01</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PrimateAI</td>
<td>PrimAI</td>
<td>Pathogenicity probability</td>
<td>0.14</td>
<td>0.6</td>
<td>-</td>
<td>-</td>
<td>0.01</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>M-CAP</td>
<td>mcap_sensitivityv1.4</td>
<td>M-CAP sensitivity score</td>
<td>0.32</td>
<td>0.67</td>
<td>-</td>
<td>-</td>
<td>0.01</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SpliceAI</td>
<td>SpliceAI_DS_AG</td>
<td>Probability of Acceptor gain variant</td>
<td>0.03</td>
<td>0.3</td>
<td>-</td>
<td>-</td>
<td>0.01</td>
<td>0.03</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SpliceAI</td>
<td>SpliceAI_DS_AL</td>
<td>Probability of Donor loss variant</td>
<td>0.04</td>
<td>0.34</td>
<td>-</td>
<td>-</td>
<td>0.01</td>
<td>0.07</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SpliceAI</td>
<td>SpliceAI_DS_DL</td>
<td>Probability of Donor loss variant</td>
<td>0.04</td>
<td>0.34</td>
<td>-</td>
<td>-</td>
<td>0.01</td>
<td>0.07</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>AdaScore</td>
<td>Ada_score</td>
<td>Probability of splicing site variant by adaboost model</td>
<td>0.08</td>
<td>0.49</td>
<td>-</td>
<td>-</td>
<td>0.01</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Allele Frequency</td>
<td>AF_HC</td>
<td>Allele frequency from NYGC resequencing</td>
<td>0.46</td>
<td>0.34</td>
<td>-</td>
<td>-</td>
<td>0.24</td>
<td>0.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

highly predictive features in the SNP set, along with M-CAP_sensitivityv1.4, AF_HC, and fathmm-xf_NC.score. In association score, more features were considered as good predictors: phyloP46, SIFT_score, MT_treevote, fathmm-xf_C.score, fathmm-xf_NC.score, CADD, CADD_scores, PrimateAI, M-CAP_sensitivityv1.4, and AdaScore. Overall, MutationTaster and CADD showed an exceptionally high correlation with ClinVar interpretation. Both ensemble methods have unrivaled scores in both PPS and association analysis, followed by allele frequency and other conservation scores. Interestingly, SpliceAI’s PPS was a bit inferior when compared to AdaScore’s, while theoretically it outperforms other splicing site predictions. The reason could lie in its representation: instead of a single score for predicting splicing site disruption, SpliceAI delta score annotates the probability of four splicing site variation types.

Noticeably, there was more correlation in the SNP set than the indel set. From our speculation, this phenomenon could be explained by the difference in variant consequences. By altering the length of a coding sequence, an indel could cause a frameshift, which substantially modifies the translation product, frequently resulting in targeted mRNA degradation [41]. However, the effect of a SNP is more complicated: even only in the coding sequence, it could be a synonymous or nonsynonymous change, and both cases can cause disease or not based on other biological mechanisms. Thus, the clinical interpretation can take advance of more types of features.

Among missense annotators, M-CAP outperforms others, followed by PrimateAI, Polyphen-2, and SIFT4G. SIFT4G develops a scoring matrix based on numerous alignments that closely match the query one [21]. To make large-scale forecasts tractable, approaches like SIFT4G use a number of approximations and simplifications. These methods introduce systematic mistakes and result in lower prediction accuracy than expected [21]. M-CAP, however, employed the gradient boosting tree method where it utilized the “wisdom of crowds” strategy that is deduced iteratively to rectify previously misclassified components [24]. It also uses additional evolutionary conservation metrics for amino acids on coding variants [24], which was proven to be more efficient than their conventional metrics alone.

IV. CONCLUSION

In this work, we have reviewed and evaluated ten recently developed variant annotation tools using ClinVar data. Our evaluation illustrated that ensemble learning models tend to have higher PPS than other methods. This improvement follows the same trend in other fields, where complex models could solve the previously unsolvable problems, such as AlphaFold [42]. However, as more tools exploit the advantage of the black-box model, the risk of overlapping training sets rises and more work for interpretation is demanded. In contrast, allele frequency still proves to be one of the best predictors, despite its simpleness. We also found that while the annotation scores for Indels are not as popular as those for SNPs, the state-of-the-art tools still make accurate predictions. Our work provided a fresh perspective over a well-studied area and paved the way for next studies on the clinical interpretation of genomic variants.