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
0000-0002-7061-074X

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

2021 13th International Conference on Knowledge and Systems Engineering (KSE) | 978-1-6654-9975-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/KSE53942.2021.9648755

***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/Var_Annot_Eval.**

***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 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 evolves 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 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,

[§]Equal contribution. Author ordering determined by a list randomizer

978-1-6654-9975-0/21/$31.00 ©2021 IEEE

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 phylogenic information of these species, suitable models of evolution are implemented to calculate the final score. GERP++ [13] or 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 scores as features and implements an ensemble learning model for boosted accuracy.

SIFT, standing for "Sorting Intolerant From Tolerant", is one of the most conventional and trustworthy functional prediction tools for amino acid substitutions. SIFT4G [21] was born with improved run-time performance. SIFT4G starts by using a seeding detection and Smith-Waterman algorithm to find overlapping sequences in the protein databases, followed by choosing closely related sequences and calculating the probability of being tolerated of each variant. Similarly, Polyphen2 only predicts the deleteriousness of amino acid substitutions. Polyphen2's method, while having some features in common with other tools in the evaluation list, is unique as it also considers the unique physicochemical properties of the target proteins. Specifically, the mutant and the wild-type alleles are assessed for change of the physical features. Moreover, Polyphen2 makes use of the HumDiv, Humvar dataset (Uniprot, 8,946 non-damaging, and 13,032 disease-causing). We observed that the TPR of Polyphen2 is significantly improved from its first version, Polyphen (HumDiv's TPR 0.92, Humvar's TPR:0.73).

As its name suggests, FATHMM-MKL [22] implements multiple kernel learning (MKL) along with a Supported Vector Machine (SVM) to train its model. Its features vary from Conservation to Genome Segmentation/Footprints. In the non-coding model, FATHMM-MKL shows the accuracy in terms of auROC up to 1.7 times higher than each component prediction score. In the paper, the authors also emphasize the power and importance of conservation scores in pathogenicity prediction. Implementing a similar methodology to FATHMM-MKL, FATHMM-XF [23] boosts its performance of both coding and non-coding models by adding genome context features and leave one out cross-validation to prevent bias.

Using a gradient boosting tree classifier, M-CAP [24] only targets scoring the deleteriousness of missense variants with Minor Allele Frequency (MAF) $\leq 1\%$. Thus, any missense variant that is not scored can be considered as likely benign with a score of 1. For features, M-CAP uses a wide range 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 η, 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 MutationTaster 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 precomputed 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 η, 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, phastCon46, 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

Tool	Feature names	Information	SNP				Indel			
			PPS	Correlation η	Cramer's V	Theil	PPS	Correlation η	Cramer's V	Theil
GERP++[a]	gerp_rs	GERP++ RS score	0.09	0.35	-	-	0.16	0.35	-	-
PhastCons[a]	phastCon46	Probability conserved nucleotide	0.01	0.31	-	-	0.13	0.50	-	-
PhyloP[a]	phyloP46	Conservation scoring	0.12	0.56	-	-	0.08	0.38	-	-
SIFT4G[b]	SIFT_score	SIFT score	0.09	0.5	-	-	0.01	0	-	-
	SIFT_median	Diversity of the sequences used for prediction.	0.01	0.08	-	-	0.01	0	-	-
	SIFT_prediction	SIFT Label	0.08	-	0.44	0.13	0.01	-	0	0
Polyphen-2[b]	Polyphen-2.HumDiv	Polyphen2 score based on HumDiv.	0.13	-	0.31	0.16	0.01	-	0	0
	Polyphen-2.HumVar	Polyphen2 score based on HumVar.	0.14	-	0.34	0.19	0.01	-	0	0
MutationTaster[b]	MT_treevote	Ratio of agreed trees	0.19	0.91	-	-	0.45	0.96	-	-
CADD[b]	CADD	CADD deleteriousness score	0.15	0.82	-	-	0.27	0.75	-	-
	CADD_phred	CADD relative rank	0.22	0.76	-	-	0.32	0.81	-	-
FATHMM-MKL[b]	fathmm-mkl_C.score	Deleterious probability as in coding region	0.03	0.01	-	-	0	0.02	-	-
	fathmm-mkl_NC.score	Deleterious probability as in non-coding region	0.02	0.07	-	-	0.03	0.06	-	-
FATHMM- XF[b]	fathmm-xf_C.score	Deleterious probability as in coding region	0.09	0.56	-	-	0.01	0	-	-
	fathmm-xf_NC.score	Deleterious probability as in non-coding region	0.25	0.61	-	-	0.01	0	-	-
PrimateAI[b]	PrimateAI	Pathogenicity probability	0.14	0.6	-	-	0.01	0	-	-
M-CAP[b]	mcap_sensitivityv1.4	M-CAP sensitivity score	0.32	0.67	-	-	0.01	0	-	-
SpliceAI[c]	SpliceAI_DS_AG	Probability of Acceptor gain variant	0.03	0.23	-	-	0.01	0.05	-	-
	SpliceAI_DS_AL	Probability of Acceptor loss variant	0.04	0.34	-	-	0.01	0.07	-	-
	SpliceAI_DS_DG	Probability of Donor gain variant	0.02	0.18	-	-	0.01	0.08	-	-
	SpliceAI_DS_DL	Probability of Donor loss variant	0.05	0.37	-	-	0.02	0.11	-	-
AdaScore[c]	Ada_score	Probability of splicing site variant by adaBoost model	0.08	0.49	-	-	0.01	0	-	-
Allele Frequency	AF_HC	Allele frequency from NYGC resequencing	0.46	0.34	-	-	0.24	0.5	-	-

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_phred, 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 can 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.

REFERENCES

[1] S. Singh, "The hundred-dollar genome: a health care cart before the genomic horse," *CMAJ*, vol. 190, no. 16, pp. E514–E514, 2018.

[2] D. Botstein and N. Risch, "Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease," *Nature genetics*, vol. 33, no. 3, pp. 228–237, 2003.

[3] S. M. Harrison, L. G. Biesecker, and H. L. Rehm, "Overview of specifications to the acmg/amp variant interpretation guidelines," *Current protocols in human genetics*, vol. 103, no. 1, p. e93, 2019.

[4] S. Pabinger, A. Dander, M. Fischer, R. Snajder, M. Sperk, M. Efremova, B. Krabichler, M. R. Speicher, J. Zschocke, and Z. Trajanoski, "A survey of tools for variant analysis of next-generation genome sequencing data," *Briefings in bioinformatics*, vol. 15, no. 2, pp. 256–278, 2014.

[5] V. Ramensky, P. Bork, and S. Sunyaev, "Human non-synonymous snps: server and survey," *Nucleic acids research*, vol. 30, no. 17, pp. 3894–3900, 2002.

[6] P. C. Ng and S. Henikoff, "Sift: Predicting amino acid changes that affect protein function," *Nucleic acids research*, vol. 31, no. 13, pp. 3812–3814, 2003.

[7] P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, and M. Kircher, "CADD: predicting the deleteriousness of variants throughout the human genome," *Nucleic Acids Research*, vol. 47, no. D1, pp. D886—-D894, 2019, doi: 10.1093/nar/gky1016.

[8] J. S. . M. K. Philipp Rentzsch, Max Schubach, "CADD-splice—improving genome-wide variant effect prediction using deep learning-derived splice scores," *Genome Medicine*, vol. 13, no. 31, 2021, doi: 10.1186/s13073-021-00835-9.

[9] J. M. Schwarz, D. N. Cooper, M. Schuelke, and D. Seelow, "Mutationtaster2: mutation prediction for the deep-sequencing age," *Nature methods*, vol. 11, no. 4, pp. 361–362, 2014.

[10] L. Sundaram, H. Gao, Padigepati, and S. et al, "Predicting the clinical impact of human mutation with deep neural networks," *Nature Genetics*, vol. 50, pp. 1161–1170, 2018, doi: 10.1038/s41588-018-0167-z.

[11] J. K, K. P. S, M. JF, D. SF, K. D, L. YI, K. JA, A. J, C. W, S. GB, C. ED, K. E, G. H, K. A, B. S, S. SJ, and F. KK, "Predicting splicing from primary sequence with deep learning," *Cell*, vol. 176, no. 3, pp. 535–548, 2019, doi: 10.1016/j.cell.2018.12.015.

[12] M. J. Landrum, J. M. Lee, G. R. Riley, W. Jang, W. S. Rubinstein, D. M. Church, and D. R. Maglott, "Clinvar: public archive of relationships among sequence variation and human phenotype," *Nucleic acids research*, vol. 42, no. D1, pp. D980–D985, 2014.

[13] D. EV, G. DL, S. M, C. GM, S. A, and B. S, "Identifying a high fraction of the human genome to be under selective constraint using gerp++," *PLoS computational biology*, vol. 6, no. 12, 2010, doi: 10.1371/journal.pcbi.1001025.

[14] C. GM, S. EA, A. G, N. C. S. Program, G. ED, B. S, and S. A, "Distribution and intensity of constraint in mammalian genomic sequence," *Genome Res*, vol. 15, no. 7, pp. 901–913, 2005, doi: 10.1101/gr.3577405.

[15] A. Siepel, K. S. Pollard, and D. Haussler, "New methods for detecting lineage-specific selection," in *Research in Computational Molecular Biology*, A. Apostolico, C. Guerra, S. Istrail, P. A. Pevzner, and M. Waterman, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 190–205.

[16] F. J and C. GA, "A hidden markov model approach to variation among sites in rate of evolution," *Mol Biol Evol*, vol. 13, no. 1, pp. 93–104, 1996, doi: 10.1093/oxfordjournals.molbev.a025575.

[17] S. A, B. G, P. JS, H. AS, H. M, R. K, C. H, S. J, H. LW, R. S, W. GM, W. RK, G. RA, K. WJ, M. W, and H. D, "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes," *Genome Res.*, vol. 15, no. 8, 2005, doi: 10.1101/gr.3715005.

[18] A. Siepel and D. Haussler, "Computational identification of evolutionarily conserved exons," in *Proceedings of the Eighth Annual International Conference on Resaerch in Computational Molecular Biology*, ser. RECOMB '04. New York, NY, USA: Association for Computing Machinery, 2004, pp. 177—-186, doi: 10.1145/974614.974638.

[19] J. Thomas, J. Touchman, and R. e. a. Blakesley, "Comparative analyses of multi-species sequences from targeted genomic regions," *Nature*, vol. 424, no. 6950, pp. 788–793, 2003, doi: 10.1038/nature01858.

[20] Z. Yang, "Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites: Approximate methods," *Journal of Molecular Evolution*, vol. 39, pp. 306—-314, 1994, doi: 10.1007/BF00160154.

[21] R. Vaser, S. Adusumalli, S. N. Leng, M. Sikic, and P. C. Ng, "Sift missense predictions for genomes," *Nature Protocols*, vol. 11, no. 1, p. 1–9, 2015, doi: 10.1038/nprot.2015.123.

[22] H. A. Shihab, M. F. Rogers, J. Gough, M. Mort, D. N. Cooper, I. N. M. Day, T. R. Gaunt, and C. Campbell, "An integrative approach to predicting the functional effects of non-coding and coding sequence variation," *Bioinformatics*, vol. 31, no. 10, pp. 1536–1543, 2015, doi: 10.1093/bioinformatics/btv009.

[23] M. F. Rogers, H. A. Shihab, M. Mort, D. N. Cooper, T. R. Gaunt, and C. Campbell, "FATHMM-XF: accurate prediction of pathogenic point mutations via extended features," *Bioinformatics*, vol. 34, no. 3, pp. 511–513, 2017, doi: 10.1093/bioinformatics/btx536.

[24] K. A. Jagadeesh, A. M. Wenger, M. J. Berger, H. Guturu, P. D. Stenson, D. N. Cooper, and J. A. B. . G. Bejerano, "M-cap eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity," *Nature Genetics*, vol. 48, pp. 1581–1586, 2016, doi: 10.1038/ng.3703.

[25] R. M. Kuhn, D. Haussler, and W. J. Kent, "The ucsc genome browser and associated tools," *Brief Bioinform*, vol. 14, no. 2, pp. 144–161, 2012.

[26] R. Steinhaus, S. Proft, M. Schuelke, D. N. Cooper, J. M. Schwarz, and D. Seelow, "MutationTaster2021," *Nucleic Acids Research*, vol. 49, no. W1, pp. W446–W451, 04 2021, doi: 10.1093/nar/gkab266.

[27] J. K, K. P. S, M. JF, D. SF, K. D, L. YI, K. JA, A. J, C. W, S. GB, C. ED, K. E, G. H, K. A, B. S, S. SJ, and F. KK, "Spliceai: A deep learning-based tool to identify splice variants." [Online]. Available: https://github.com/Illumina/SpliceAI#spliceai-a-deep-learning-based-tool-to-identify-splice-variants

[28] X. Jian, E. Boerwinkle, and X. Liu, "In silico prediction of splice-altering single nucleotide variants in the human genome," *Nucleic Acids Research*, vol. 42, no. 22, pp. 13 534–13 544, 2014, doi: 10.1093/nar/gku1206.

[29] V. A. Huynh-Thu, Y. Saeys, L. Wehenkel, and P. Geurts, "Statistical interpretation of machine learning-based feature importance scores for biomarker discovery," *Bioinformatics*, vol. 28, no. 13, pp. 1766–1774, 2012.

[30] R. Rodríguez-Pérez and J. Bajorath, "Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions," *Journal of computer-aided molecular design*, vol. 34, no. 10, pp. 1013–1026, 2020.

[31] C. Molnar, *Interpretable machine learning*. Lulu. com, 2020.

[32] M. I. Razzak, S. Naz, and A. Zaib, "Deep learning for medical image processing: Overview, challenges and the future," *Classification in BioApps*, pp. 323–350, 2018.

[33] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. Moura, and P. Eckersley, "Explainable machine learning in deployment," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 648–657.

[34] F. Wetschoreck, T. Krabel, and S. Krishnamurthy, "8080labs/ppscore: zenodo release," Oct. 2020, doi: 10.5281/zenodo.4091345. [Online]. Available: https://doi.org/10.5281/zenodo.4091345

[35] S. Singh and P. Gupta, "Comparative study id3, cart and c4. 5 decision tree algorithm: a survey," *International Journal of Advanced Information Science and Technology (IJAIST)*, vol. 27, no. 27, pp. 97–103, 2014.

[36] M. Bramer, *Principles of data mining*. Springer, 2007, vol. 180.

[37] A. R. Crathorne, "Calculation of the correlation ratio," *Journal of the American Statistical Association*, vol. 18, no. 139, pp. 394–396, 1922, doi: 10.1080/01621459.1922.10502484.

[38] H. Cramér, "Mathematical methods of statistics, 1946," *Department of Mathematical SU*, 1946.

[39] A. Agresti, *Categorical data analysis*. John Wiley & Sons, 2003, vol. 482.

[40] M. Byrska-Bishop, U. S. Evani, X. Zhao, A. O. Basile, H. J. Abel, A. A. Regier, A. Corvelo, W. E. Clarke, R. Musunuri, K. Nagulapalli, S. Fairley, A. Runnels, L. Winterkorn, E. Lowy-Gallego, T. H. G. S. V. Consortium, P. Flicek, S. Germer, H. Brand, I. M. Hall, M. E. Talkowski, G. Narzisi, and M. C. Zody, "High coverage whole genome sequencing of the expanded 1000 genomes project cohort including 602 trios," *bioRxiv*, 2021, doi: 10.1101/2021.02.06.430068.

[41] B. Baumann, M. Potash, and G. Köhler, "Consequences of frameshift mutations at the immunoglobulin heavy chain locus of the mouse." *The EMBO Journal*, vol. 4, no. 2, pp. 351–359, 1985.

[42] M. AlQuraishi, "Alphafold at casp13," *Bioinformatics*, vol. 35, no. 22, pp. 4862–4865, 2019.