AllelePred: A Simple Allele Frequencies Ensemble Predictor for Different Single Nucleotide Variants

Turki Sobahy [1] and Meshari Alazmi [2]

[1]King Faisal Specialist Hospital & Research Center-Jeddah
[2]Affiliation not available

October 30, 2023

Abstract

Genomic medicine stands to be revolutionized through the understanding of single nucleotide variants (SNVs) and their expression in single-gene disorders (mendelian diseases). Computational tools can play a vital role in the exploration of such variations and their pathogenicity. Consequently, we developed the ensemble prediction tool AllelePred to identify deleterious SNVs and disease causative genes. In comparison to other tools, our classifier achieves higher accuracy, precision, F1 score, and coverage for different types of coding variants. Furthermore, this research analyzes and structures 168,945 broad spectrum genetic variants from the genomes of the Saudi population to denote the accuracy of the model. When compared, AllelePred was able to structure the unlabeled Saudi genetic variants of the dataset to mimic the data characteristics of the known labeled data. On this basis, we accumulated a list of highly probable deleterious variants that we recommend for further experimental validation prior to medical diagnostic usage.

AllelePred: A Simple Allele Frequencies Ensemble Predictor for Different Single Nucleotide Variants

Turki M. Sobahy[1]*, Olaa Motwalli[2], and Meshari Alazmi[3]

Abstract— *Background & Objective*: Genomic medicine stands to be revolutionized by understanding single nucleotide variants (SNVs) and their expression in single-gene disorders (Mendelian diseases). Computational tools can play a vital role in the exploration of such variations and their pathogenicity. Consequently, we developed the ensemble prediction tool AllelePred to identify deleterious SNVs and disease causative genes.*Results:* The model utilizes different population genetics backgrounds, and restricted criteria for features selection to help generate high accuracy results. In comparison to other tools, such as Eigen, PROVEAN, and fathmm-MKL our classifier achieves higher accuracy (98%), precision (96%), F1 score (93%), and coverage (100%) for different types of coding variants. Furthermore, this research analyzes and structures 168,945 broad spectrum genetic variants from the genomes of the Saudi population to denote the accuracy of the model. When compared, AllelePred was able to structure the unlabeled Saudi genetic variants of the dataset to mimic the data characteristics of the known labeled data. On this basis, we accumulated a list of highly probable deleterious variants that we recommend for further experimental validation prior to medical diagnostic usage.*Conclusions*: The ensemble prediction tool AllelePred enables increased accuracy in recognizing deleterious SNVs and the genetic determinants in real clinical data.

Index Terms— Single nucleotide variants, Single-gene disorders, Predictive modeling

1 Introduction

As the enigma of the human genome begins to unravel, the study of genomics has found that many human diseases can be traced back to a genetic component [1],[2]. Both rare single-gene disorders (Mendelian diseases) and complex disorders (diabetes) can be linked to variations in the human genome [1], [3]. Therefore, it has become essential to identify disease-inducing genetic variations that can help diagnose human disease and understand its pathophysiological process [4], [5].

Genetic changes can be identified through human genome reference comparison and manifest in different types. These types can range from complex structural variations to simpler single nucleotide variants (SNVs). SNVs are the most common type of sequence change [6] and can be categorized into two main types; synonymous and non-synonymous. Synonymous SNVs involve the replacement of a codon. Contrarily, most non-synonymous SNVs (nsSNVs) result in an encoded protein with an amino acid change that affects and alters functionality [7], [8]. As a result, nsSNVs are the most common cause of Mendelian diseases and represent the majority of known disease-inducing genetic variations [7], [9]. Additionally, nsSNVs can be classified into three types of mutations; missense, nonsense, and frameshift. In a missense mutation, a DNA base pair is changed, resulting in the encodement of different amino acids. Similarly, a nonsense mutation involves a change in a DNA base pair; however, this produces a stop codon that terminates the protein synthesis process prematurely. Meanwhile, frameshift mutations involve the insertion or removal of a single nucleotide that changes the protein reading frame. These types of mutations can result in the compromise and dysfunctionality of encoded protein.

Moreover, SNVs can be classified based on the chemical properties of the changed nucleotide. In particular, the ring structure, which determines if a substitution is a transition or transversion [10]. Single-ring shaped molecules known as pyrimidine, include cytosine (C) and thymine (T). While, two-ring shaped molecules called purines include adenine (A) and guanine (G). An exchange of the same ring (purine to purine, pyrimidine to pyrimidine) is called a transition substitution. On the contrary, transversions include an exchange of different ring size (purine to pyrimidine, vice versa) [10]. Note that transversions are known to be more impactful because of the structural effects they pose to DNA [10], [11].

SNVs have been experimentally validated and studied through laboratory-based methods. Yet, a single human exome can harbor around 20,000 SNVs [12], making such laboratory verifications both time-consuming and expensive. As a result, computational tools have become valuable in exploring the footprint of genetic variations and their pathogenicity [8], [13], [14]. Thus, tools such as Mu-

1 King Faisal Specialist Hospital & Research Center-Jeddah (KFSHRC-J), Saudi Arabia (e-mail: tsobahy@kfshrc.edu.sa and ORCID ID: https://orcid.org/0000-0003-1797-3098)
2 Saudi Electronic University (SEU), College of Computing and Informatics, Madinah 41538-53307, Kingdom of Saudi Arabia (e-mail: o.motwalli@seu.edu.sa and ORCID ID: https://orcid.org/0000-0002-3392-5734)
3 Department of Information and Computer Science, College of Computer Science and Engineering, University of Ha'il, P.O. Box 2440, Ha'il, 81411, Saudi Arabia (e-mail: ms.alazmi@uoh.edu.sa and ORCID ID: https://orcid.org/0000-0001-9074-1029)
* The corresponding author

tationAssessor, LRT, SIFT, CADD, PolyPhen, ClinPred, fathmm-MKL, and Eigen were created [15]-[17].

The first class of computational methods, individual classifiers (MutationAssessor, LRT, SIFT, CADD, and PolyPhen). These methods rely on sequence-based features such as; chemical properties of the replacement, experimental data, sequence-based (e.g. CpG transition) and structure-based features, etc. [15], [16], [18]-[20] to generate a single classification model. On the other hand, the second class of methods, ensemble classifiers (VEST, ClinPred, fathmm-MKL, and Eigen). These methods incorporate more recent computational algorithms and individual predictor scores [17] to merge the predictions of multiple base models (refer to Figure 1). In general, ensemble classifiers are known for their high dimensionality, reduced model bias, and complex data structures [21].

Nonetheless, current ensemble predictors are designed to estimate the impact of only missense genetic variations, except for fathmm-MKL and Eigen, whose coverage extends to nonsense, frameshift, and synonymous variants. Even so, Eigen only predicts variants with full annotations, and fathmm-MKL does not consider one crucial feature [12], [18], [19]. The features of Allele Frequencies (AFs) are used to represent the population(s) background knowledge features [22]. The AFs are commonly applied during genetic variant interpretation workflows and can increase the precision of machine learning models [17], [23]. Thus, we used AFs to help create AllelePred; an accurate ensemble tool that helps in identifying deleterious SNVs, actionable variants to fine-tune workflows, and disease causative genes.

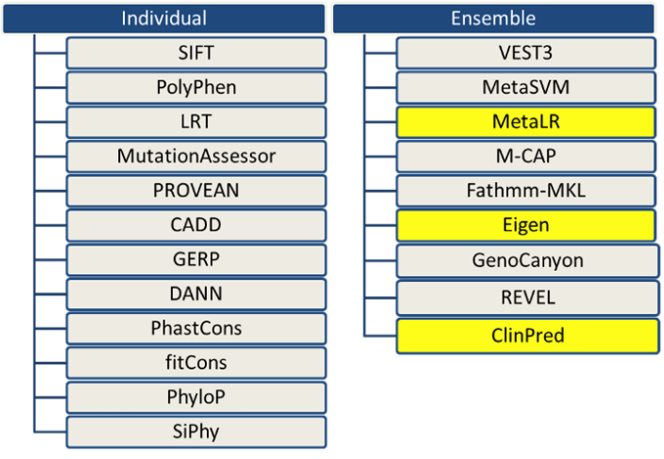


Fig.1. List of available predictive methods per class: individual and ensemble. Only three methods used AFs. Mainly, Eigen, and ClinPred employed AFs of different populations.

2 Materials And Methods

2.1 AllelePred Dataset

We collected 168,945 broad-spectrum genetic variants for the Saudi population from the SHGP portal (https://shgp.sa/index.en.html). However, there was a high error rate of the used sequencing platform (Ion Proton™). In predictive modeling, identifying reliable, balanced, and accurate sources of data is crucial. Thus, gnomAD was used to filter variants, leaving a total of 100,507 variants. Then, intronic, splicing, frameshift, and non-frameshift INDELS were removed. A final dataset of 56,172 genetic variants remained; missense (50%), nonsense (2%), and synonymous (48%). To classify the final dataset of genetic variants, the ClinVar database was downloaded in February 2019. ClinVar variants reviewed in January 2013 were used to label the variants into tolerated, likely tolerated, deleterious, or likely deleterious.

Consequently, only 9% of the final dataset was labeled by ClinVar and resulted in 5,123 SNVs. This final dataset resulted in a highly unbalanced labeled dataset and an uneven distribution of deleterious and tolerated variants within the different variant classes (e.g., synonymous variants had no true positive variants). To solve the unbalanced dataset issue, additional data was added by generating a ClinVar based testing dataset that accounts for all the investigated types of variations (missense, synonymous, and nonsense). All deleterious (86) and tolerated (31) in synonymous and nonsense variants were collected to even the distribution. A final 1,199 variants were then added to the dataset, 32 duplicate variants removed, and a final labeled dataset of 6,290 variants obtained (refer to Figure 2).

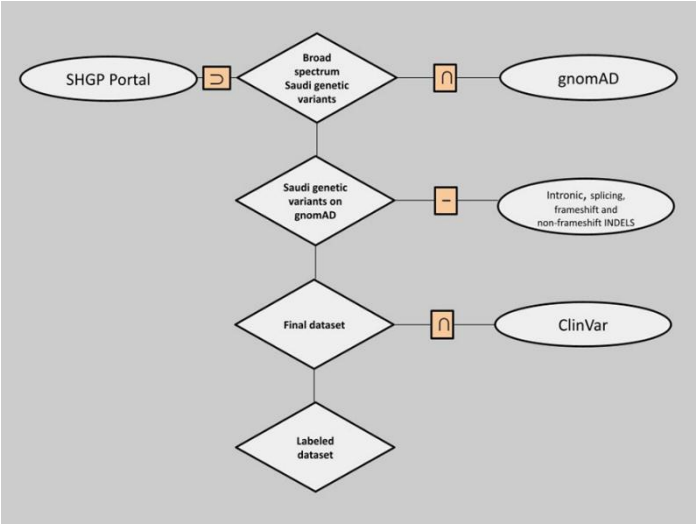


Fig.2. Dataset acquisition and pre-model training preparation.

2.2 Features

We used ANNOVAR to download nine prediction scores of different predictive models: CADD, ExAC_pLI, M-CAP, MetaSVM, MutationTaster, Polyphen2_HDIV, Polyphen2_HVAR, REVEL and SIFT, and the allele frequencies (AFs) of different populations (Supplementary Table 1). The Saudi AFs were collected from the SHGP web application. However, a challenge was encountered in finding functional features for synonymous SNVs (Table 1). Thus, features that returned no prediction scores were given a value of zero.

Variant Type	Number of Variants	Missing Features Count	Average per Variant
Missense	2,508	11,953	4.765948963
Synonymous	3,378	34,003	10.06601539
Nonsense	377	11,073	29.37135279
Other	27	117	4.333333333
Total	6,290	57,146	9.085214626

Table.1. The number of unavailable features in each type of variation per type.

Feature selection is a crucial component in training machine learning models and increasing model prediction accuracy and generalization [24]. To reduce our 41 features, we used the Pearson correlation coefficient to measure the strength of association between features. By defining a correlation of 0.98 as the threshold value for high degree collinearity, we removed 17 features. Furthermore, the chi-squared selection criteria were used to select highly dependent features on the target vector and provide a more stable model. Consequently, a final number of 20 features was selected. Through feature selection (supplementary table 1), we reduced the total number of features by 50% (20 of 41), resulting in a more reliable and generic model. The features were then normalized independently between 0 and 1 using min-max normalization to avoid behavioral changes.

2.3 AllelePred Model

The data was divided into 70% training and 30% testing datasets. Random Forest Classifier model was used on the training set, and the number of the decision trees was tuned via 5-fold cross-validation (Table 2). Meanwhile, the 30% unseen testing dataset was used to compare the performance of the final model (after tuning) with the state-of-the-art methods (Table 4).

3 Results

3.1 Saudi Genetic Variant Statistical Analysis

Statistical analysis was performed on the 100,507 genetic variants of the Saudi population and compared to other population variants.

First, the 95,416 unlabeled variants dataset were analyzed. Most of the variants were transition substitutions representing 73% of the data, while 27% were transversions. As shown in Figure 3(a), 36% of the transition substitutions were between C→T and T→C (pyrimidines), while 37% were between G→A and A→G (purines). Concurrently, in the transversion substitutions 13% were a purine ring to pyrimidine (A→C/T & G→C/T), and 14% a pyrimidine ring to a purine (C→A/G & T→A/G). This data is further analyzed after the AllelePred method was used to predict its label.

Labeled by ClinVar were 5,091 genetic variants; 80% were transition substitutions, and 20% transversions. The labeled data also presented a skewed distribution; 98.6% of the variants were tolerated, and 1.4% were deleterious. Of the 5,020 tolerated variants 80% were transition substitutions, and 20% were transversions (Figure 3(b)). In particular, 41% of the transitions were between C→T and T→C and 39% between G→A and A→G. While, of the transversions, 10% were substitutions of a purine ring to that of a pyrimidine, and 10% were of the opposite.

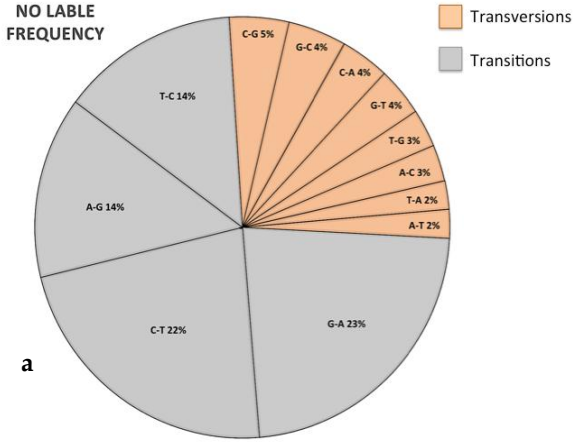
On the other hand, among the 71 deleterious variants, 91% were transitions, and 5% were transversions (Figure 3(c)). 57% of the transitions were between C→T and T→C, and 43% between G→A and A→G. Meanwhile, 2% of the transversions were of a purine ring to a pyrimidine, and 3% of the opposite. Note that only transversions of G→T, A→T, and C→G occurred.

It seems that the data shows favor to the occurrence of transition substitutions. Accordingly, most of the labeled deleterious variants, labeled tolerated variants, and unlabeled variants were transitions. The data is almost evenly split between same ring changes in purines and pyrimidines within the transition substitutions. Similarly, there is an almost even split between purine ring to pyrimidine and pyrimidine ring to purine in transversion substitutions.

AllelePred was then used to provide predictive labels to the 95,416 genetic variants in the unlabeled dataset. Similarly to the labeled dataset AllelePred predicted a high number of transitions and a lower number of transversions, 73% and 27%, respectively. A skewed distribution was also predicted; 0.48% of the variants were deleterious, while 99.52% tolerated.

Of the tolerated variants, 73% were transitions and 27% transversions. Moreover, 36% of the transition substitutions were between C→T and T→C, and 37% between G→A and A→G. While in the transversions, 13% were purine to pyrimidine, and 14% the opposite.

The deleterious variants, however, were 97% transition substitutions and 3% transversions. Of the transitions 49% were between C→T and T→C, and 48% between G→A and A→G. 1% of the transversions were purine to pyrimidine, and 2% of the opposite. Note that no transversions of A→C and G→C were found.



a

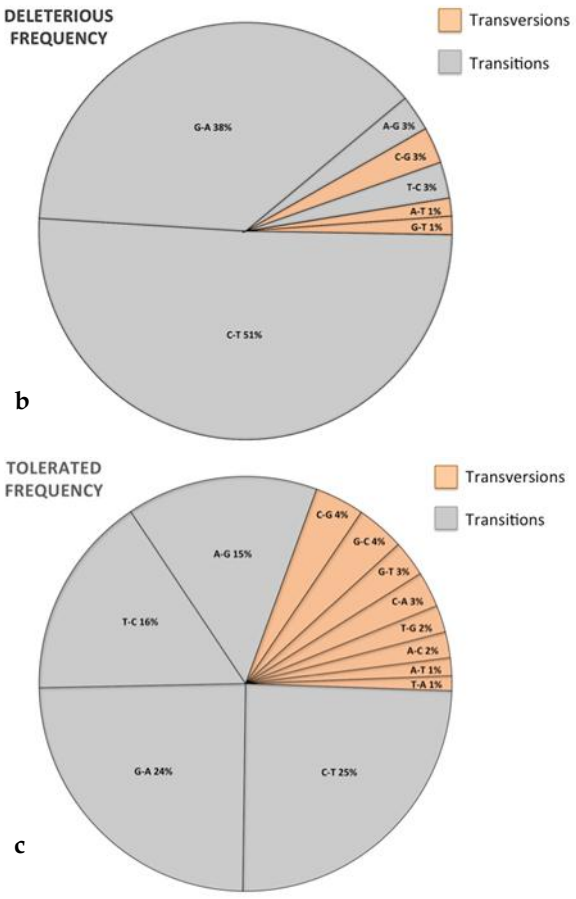


Fig.3. Frequency of the unlabeled, tolerated, and deleterious nucleic acid changes in the original dataset of 100,507 SNVs.

3.2 AllelePred Model Development Performance

These results were assessed through measuring the performance of the model using the following metrics: recall, precision, F1-score, and accuracy measures. Table 2 shows the average performance on the validation sets.

No.of trees	*Precision*	*Recall*	*F1-score*	*Accuracy*
1	0.89	0.90	0.89	0.97
10	0.91	0.91	0.91	0.98
100	0.93	0.91	0.92	0.98
1000	0.92	0.91	0.92	0.98

Table.2. Performance-based on the 5-fold cross-validation with decision tree number tuning.

3.3 Performance Comparison with Similar Tools

The results from the 30% unseen testing dataset were used to compare the performance of AllelePred with three other methods. The comparator tools were selected based on their ability to predict more than one type of mutation. Eigen and fathmm-MKL were the only ensemble classifiers that met this criterion. Thus, the individual classifier PROVEAN was also added. The deterministic cut-off for Eigen and fathmm-MKL was a default raw score of 0.5. Meanwhile, PROVEAN predicted definitive variant classes (neutral or damaging).

Additionally, AllelePred was tested against routine computational workflow. The routine workflow is a method that utilizes the overall AFs in gnomAD (less than 1%) and CADD (scaled C-score of at least 15) while including AF filtration and model prediction [25].

AllelePred demonstrated the best performance and coverage for all coding variants. Overall, Eigen did not return predictions for some variants (55%) and returned predictions for only 6% of the synonymous variants. Moreover, fathmm-MKL showed the lowest overall accuracy of 69.6%. In fact, for synonymous variants, fathmm-MKL only achieved 73.6% accuracy, while other methods (except Eigen) achieved a tight accuracy; AllelePred (98%), the routine workflow (99%), and PROVEAN (98%). In the nonsense category, Eigen returned predictions for 93% of the submitted variants with 71% accuracy; fathmm-MKL achieved an accuracy of 76%, and the routine workflow a 90% accuracy. AllelePred exceeded the other models with the highest accuracy of 99%. Note that PROVEAN was not designed to predict nonsense mutations. Finally, missense variants had full coverage by all models, except for Eigen which missed approximately 4% of the variants. AllelePred, again, achieved the highest accuracy of 99%. While, fathmm-MKL, the routine workflow, PROVEAN, and Eigen achieved 63%, 76%, 86.2%, and 91% respective accuracies (Figure 4, Table 3).

Method	Precision	Recall	F1-score	Accuracy	Coverage
AllelePred	0.96	0.90	0.93	0.98	1.00
Workflow	0.54	0.90	0.68	0.90	1.00
PROVEAN	0.50	0.69	0.58	0.93	0.94
Eigen	0.76	0.75	0.75	0.89	0.46
Fathmm-MKL	0.26	0.88	0.40	0.70	1.00

Table.3. The table shows performance of AllelePred with other methods.

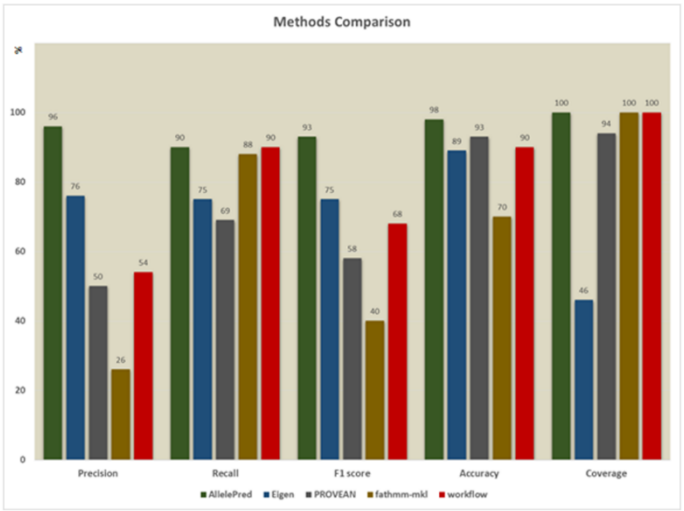


Fig.4. The results of the testing dataset (1,883 genetics variants) of AllelePred and comparative approaches.

3.3 Clinical Testing

AllelePred clinical applicability was evaluated in comparison with the "routine" workflow on two clinical WES datasets. First, two VCF files contain a single homozygous causative that was clinically verified and reported (NM_000466.2:c.3568C>T, NM_014780.3:c.2862+1G>A) were obtained from KFSHRC-R. The second WES (research set) we had at KFSHRC-J (no published data for the study yet) (ethical approval No. 2018-36). We used two variants (from Arab patients) that were published in peer-reviewed journals: NM_017988.4:c.106C>T, which a protein terminating variant with very low AF (< 1%) reported in a Saudi family; the parents are first cousins, diagnosed with arthrogryposis multiplex congenita (AMC); and NM_000933.3:c.1862G>A. This is a rare (not found on gnomAD) missense alteration that was reported in an Egyptian family with history of auriculocondylar syndrome (ARCND) with highly variable clinical phenotypes [26], [27]. The two variants were inserted into the two research WES files.

In total, the four exomes had no true positive variants on ClinVar, and shared 3,835 unique true negative variants. AllelePred and the "routine" workflow were evaluated based on detecting the causative variants, and the number of false-positive variants based on their classification in the reviewed ClinVar dataset. False negatives were not accounted for because no true positives were found on ClinVar. Both methods were able to predict the four causative variants. AllelePred only had one false positive variant, while the routine workflow had 58 unique false positive variants in all samples. AllelePred also predicted fewer variants as deleterious than the routine workflow (Tabel 4).

Sample	No. of variants	TP	Workflow			AllelePred		
			PP	TP	FP	PP	TP	FP
S1	84607	1	498	1	20	204	1	0
S2	86889	1	521	1	13	245	1	0
S3	56618	1	731	1	23	419	1	0
S4	54610	1	513	1	6	357	1	1

Table.4. Total number of variants per sample, number of positive predictive (PP), true positive (TP) & false positive variants (FP) by AllelePred, and the routine workflow.

4 Discussion

Elaborating on understanding genetic mutations is an integral aspect of medical genetics and is vital to providing opportunities to those with gene disorders [23]. To compensate for functional assays' costs and enable cost-effective experimental validation, a computational approach can be used to evaluate SNV molecular consequences [28]. Due to the high false-positive rates in the available predictive models [29], we designed AllelePred. This method uses different population's genetics backgrounds and restricted criteria for feature selection (AFs) to yield higher accuracy results. AFs enabled our method to display a high level of F1-score, precision, accuracy, and coverage compared to the other methods.

Infrequent variants are increasingly crucial in diagnosing single-gene and complex disorders and tend to be population specific. The rarity of a variant is often associated with an increased probability of variant causality [30]-[36]. This was denoted by the AF feature in AllelePred, and played a significant role in ensuring the method's high performance. In fact, The Population Architecture using Genomics and Epidemiology (PAGE) consortium project [37] showed that the usage of AFs from different populations resulted in the identification of 27 novel variants.

During the statistical analysis of the ClinVar labeled dataset, it was found that of the tolerated variants 80% were transition substitutions, and only 20% were transversion. The high number of transitions can be ascribed to the phenomenon called the transition bias. Since transition substitutions require less double-helix structure distortion, this makes them a more frequent occurrence than transversions [38]. Additionally, a high tolerated transition substitution rate and a low tolerated transversion rate were expected; transversions were significantly more detrimental than transitions [39]. This is attributed to the structure of the genetic code as transitions often have a lower probability of causing radical changes to the physicochemical properties of amino acids [40]. Surprisingly, however, in the deleterious variants 91% were transitions, and only 5% were transversions.

To further evaluate this particular pattern, we accumulat-

ed other population data (from previously downloaded gnomAD) and analyzed the percentage of its transition and transversion substitutions. Interestingly, the pattern of high deleterious transitions and low deleterious transversions was found among all other populations (with the exception of non-Finnish European population (NFE)) (Supplementary table 2). Furthermore, in the other populations, we found a high percentage of tolerated transitions and a low percentage of tolerated transversions. Thus, based on our data, we can conclude that this is a recurring pattern in both the Saudi population and most other populations. Moreover, we statistically analyzed the unlabeled Saudi genetic variants that were predicted by AllelePred. We found that the pattern of the labeled Saudi genetic variants, and other populations mentioned above, was mimicked in AllelePred's predictions.

In addition, during the analysis of the substitutions between the populations, it was found that the Saudi dataset showed the highest number of transversions and the lowest number of transitions. It can be hypothesized that this high transversion rate results from consanguineous marriages, which have been known to increase the prevalence of genetic disorders [41], [42]. Yet, surprisingly, the percentage of deleterious transversions is the lowest in the Saudi population.

From our findings, we were able to sum up the variants that have the highest probability of being deleterious (Table 5) and Saudi variants without reference labels (https://drive.google.com/file/d/1bT9t_dfbvyQ4wguhjTf7oIQRw1TC5oZ/view?usp=sharing). The listed variants were predicted to be deleterious in at least three prediction tools or more. Thus, we recommend further experiential validation to verify the effect of variants before their use for medical diagnostic purposes.

Chr.	Position	Variance		Tools			
1	17349144	G	A	AllelePred	Eigen	Fathmm-Mkl	Provean
1	17349219	G	A	AllelePred	Eigen	Fathmm-Mkl	Provean
13	48953760	C	T	AllelePred	Eigen	Fathmm-Mkl	
17	59821939	A	T	AllelePred	Eigen	Fathmm-Mkl	
22	30067836	C	T	AllelePred	Eigen	Fathmm-Mkl	
3	70008476	C	T	AllelePred	Eigen	Fathmm-Mkl	
13	32936740	G	A	AllelePred	Eigen	Fathmm-Mkl	
8	118849385	G	A	AllelePred		Fathmm-Mkl	Provean
17	7577085	C	T	AllelePred	Eigen	Fathmm-Mkl	Provean
17	29533378	C	T	AllelePred	Eigen	Fathmm-Mkl	
9	135796754	G	A	AllelePred	Eigen	Fathmm-Mkl	
3	41277290	C	T	AllelePred	Eigen	Fathmm-Mkl	
15	66729163	C	T	AllelePred	Eigen	Fathmm-Mkl	Provean
2	47702265	C	T	AllelePred	Eigen	Fathmm-Mkl	
11	32413566	G	A	AllelePred		Fathmm-Mkl	Provean
16	2114342	C	T	AllelePred	Eigen	Fathmm-Mkl	
10	89711899	C	T	AllelePred	Eigen	Fathmm-Mkl	Provean
5	112174631	C	T	AllelePred	Eigen	Fathmm-Mkl	
10	89717672	C	T	AllelePred	Eigen	Fathmm-Mkl	
11	108196143	C	T	AllelePred	Eigen	Fathmm-Mkl	Provean
5	112164616	C	T	AllelePred	Eigen	Fathmm-Mkl	
3	37042536	C	T	AllelePred	Eigen	Fathmm-Mkl	

Table.5. This table identifies the variants that were predicted to be deleterious in three or more tools.

5 Conclusion

The role of genetics in the diagnosis and treatment of diseases is steadily becoming a paramount one. To further such advancements, we developed a meta-predictive and straightforward model for the impact relevance of multiple types of SNVs. Random Forest classifier and comparative analysis were applied to measure feature relevance, select 20 features, and combine predicted functional annotations and AFs. The resulting model proved to have high coverage, accuracy, precisions, F1, and recall in comparison to other models. Additionally, analysis on broad-spectrum labeled Saudi genetic variants was performed. Wherein, AllelePred was able to mimic the high percentage of tolerated transitions and low percentage of tolerated transversions in predicting the unlabelled Saudi genetic variants. Based on this, our research suggests high probability deleterious Saudi genetic variants for further clinical trial and study. Future work could include larger Arab or Saudi annotated genetic datasets or burgeoning AFs applications for non-coding variants when reliable and curated sources become available.

5 Acknowledgments

For computational resources, this research used the resources of the Supercomputing Laboratory at King Abdullah University of Science & Technology (KAUST) in Thuwal, Saudi Arabia.

Allele frequencies for the Saudi population were made available by the Saudi Human Genome Program (SHGP) in King Abdulaziz City for Science & Technology (KACST).

KFSHRC-R (Thanks to the provider Dr. Fowzan S. Alkuraya) made the two clinical exome samples used in this research.

6 Data Availability

-Supplementary files are attached.
-gnomAD: https://gnomad.broadinstitute.org/
-ClinVar: https://www.ncbi.nlm.nih.gov/clinvar/
-ANNOVAR:
https://annovar.openbioinformatics.org/en/latest/
-SHGP: https://shgp.sa/index.en.html

References

1. Jackson M, Marks L, May GHW, Wil-son JB: The genetic basis of disease. Essays Biochem 2018, 62(5):643-723.

2. CHAPTER 1, GENETICS 101 In: Un-derstanding Genetics: A New York, Mid-Atlantic Guide for Patients and Health Professionals, Genetic Alliance; The New York-Mid-Atlantic Consortium for Genetic and Newborn Screening Services. Washington (DC); 2009.
3. Understanding Human Genetic Varia-tion. In: National Institutes of Health (US); Biological Sciences Curriculum Study, NIH Curriculum Supplement Series [Internet]. Bethesda (MD): National Institutes of Health (US); 2007.
4. Eichler EE: Genetic Variation, Compar-ative Genomics, and the Diagnosis of Dis-ease. N Engl J Med 2019, 381(1):64-74.
5. Talseth-Palmer BA, Scott RJ: Genetic variation and its role in malignancy. Int J Biomed Sci 2011, 7(3):158-171.
6. Kulkarni S PJ: Clinical Genomics. A Guide to Clinical Next Generation Sequenc-ing. London, UK: Elsevier Inc., Academic Press; 2015.
7. Li MX, Kwan JS, Bao SY, Yang W, Ho SL, Song YQ, Sham PC: Predicting mende-lian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. PLoS Genet 2013, 9(1):e1003143.
8. Sun H, Yu G: New insights into the pathogenicity of non-synonymous variants through multi-level analysis. Sci Rep 2019, 9(1):1667.
9. Katsonis P, Koire A, Wilson SJ, Hsu TK, Lua RC, Wilkins AD, Lichtarge O: Sin-gle nucleotide variations: biological impact and theoretical interpretation. Protein Sci 2014, 23(12):1650-1666.
10. Guo C, McDowell IC, Nodzenski M, Scholtens DM, Allen AS, Lowe WL, Reddy TE: Transversions have larger regulatory ef-fects than transitions. BMC Genomics 2017, 18(1):394.
11. Kristina Strandberg AK, Salter LA: A comparison of methods for estimating the transition:transversion ratio from DNA sequences. Mol Phylogenet Evol 2004, 32(2):495-503.
12. Shihab HA, Gough J, Mort M, Cooper DN, Day IN, Gaunt TR: Ranking non-synonymous single nucleotide polymor-phisms based on disease concepts. Hum Ge-nomics 2014, 8:11.
13. Abu-Elmagd M, Assidi M, Schulten HJ, Dallol A, Pushparaj P, Ahmed F, Scher-er SW, Al-Qahtani M: Individualized medi-cine enabled by genomics in Saudi Arabia. BMC Med Genomics 2015, 8 Suppl 1:S3.
14. Wu M, Wu J, Chen T, Jiang R: Prioriti-zation Of Nonsynonymous Single Nucleo-tide Variants For Exome Sequencing Studies Via Integrative Learning On Multiple Ge-nomic Data. Sci Rep 2015, 5:14955.
15. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J: A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 2014, 46(3):310-315.
16. Ng PC, Henikoff S: SIFT: Predicting amino acid changes that affect protein func-tion. Nucleic Acids Res 2003, 31(13):3812-3814.
17. Alirezaie N, Kernohan KD, Hartley T, Majewski J, Hocking TD: ClinPred: Predic-tion Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants. Am J Hum Genet 2018, 103(4):474-483.
18. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD: A spectral approach integrat-ing functional genomic annotations for cod-ing and non-coding variants. Nat Genet 2016, 48(2):214-220.
19. Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day IN, Gaunt TR, Campbell C: An integrative approach to pre-dicting the functional effects of non-coding and coding sequence variation. Bioinformat-ics 2015, 31(10):1536-1543.
20. Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN, Bernstein JA, Bejerano G: M-CAP elimi-nates a majority of variants of uncertain sig-nificance in clinical exomes at high sensitivi-ty. Nat Genet 2016, 48(12):1581-1586.
21. Pengyi Yang YHY, Bing B. Zhou, Al-bert Y. Zomaya: A Review of Ensemble Methods in Bioinformatics. Current Bioin-formatics 2010, 5(4).
22. Abouelhoda M, Sobahy T, El-Kalioby M, Patel N, Shamseldin H, Monies D, Al-Tassan N, Ramzan K, Imtiaz F, Shaheen R et al: Clinical genomics can facilitate coun-trywide estimation of autosomal recessive disease burden. Genet Med 2016, 18(12):1244-1249.
23. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E et al: Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genet-ics and Genomics and the Association for Molecular Pathology. Genet Med 2015, 17(5):405-424.
24. Mwangi B, Tian TS, Soares JC: A re-view of feature reduction techniques in neu-roimaging. Neuroinformatics 2014, 12(2):229-244.
25. Li MM, Datto M, Duncavage EJ, Kul-karni S, Lindeman NI, Roy S, Tsimberidou AM, Vnencak-Jones CL, Wolff DJ, Younes A et al: Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer: A Joint Consensus Rec-ommendation of the Association for Mo-lecular Pathology, American Society of Clin-ical Oncology, and College of American Pathologists. J Mol Diagn 2017, 19(1):4-23.
26. Nabil A, El Shafei S, El Shakankiri NM, Habib A, Morsy H, Maddirevula S, Alkuraya FS: A familial PLCB4 mutation causing auriculocondylar syndrome 2 with variable severity. Eur J Med Genet 2020, 63(6):103917.
27. Seidahmed MZ, Al-Kindi A, Alsaif HS, Miqdad A, Alabbad N, Alfifi A, Ab-delbasit OB, Alhussein K, Alsamadi A, Ib-rahim N et al: Recessive mutations in SCYL2 cause a novel syndromic form of arthrogryposis in humans. Hum Genet 2020, 139(4):513-519.
28. Bromberg Y, Kahn PC, Rost B: Neu-tral and weakly nonneutral sequence variants may define individuality. Proc Natl Acad Sci U S A 2013, 110(35):14255-14260.
29. Ghosh R, Oak N, Plon SE: Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. Genome Biol 2017, 18(1):225.
30. Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, Genomes P, Bustamante CD: Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci U S A 2011, 108(29):11983-11988.
31. Consortium STD, Estrada K, Aukrust I, Bjorkhaug L, Burtt NP, Mercader JM, Garcia-Ortiz H, Huerta-Chagoya A, More-no-Macias H, Walford G et al: Association of a low-frequency variant in HNF1A with type 2 diabetes in a Latino population. JA-MA 2014, 311(22):2305-2314.
32. Gudmundsson J, Sulem P, Gudbjarts-son DF, Masson G, Agnarsson BA, Ben-ediktsdottir KR, Sigurdsson A, Magnusson OT, Gudjonsson SA, Magnusdottir DN et al: A study based on whole-genome se-quencing yields a rare variant at 8q24 associ-ated with prostate cancer. Nat Genet 2012, 44(12):1326-1329.
33. Moltke I, Grarup N, Jorgensen ME, Bjerregaard P, Treebak JT, Fumagalli M, Korneliussen TS, Andersen MA, Nielsen TS, Krarup NT et al: A common Greenlandic TBC1D4 variant confers muscle insulin re-sistance and type 2 diabetes. Nature 2014, 512(7513):190-193.
34. Kenny EE, Timpson NJ, Sikora M, Yee MC, Moreno-Estrada A, Eng C, Huntsman S, Burchard EG, Stoneking M, Bustamante CD et al: Melanesian blond hair is caused by an amino acid change in TYRP1. Science 2012, 336(6081):554.
35. Manning A, Highland HM, Gasser J, Sim X, Tukiainen T, Fontanillas P, Grarup
N, Rivas MA, Mahajan A, Locke AE et al: A Low-Frequency Inactivating AKT2 Vari-ant Enriched in the Finnish Population Is Associated With Fasting Insulin Levels and Type 2 Diabetes Risk. Diabetes 2017, 66(7):2019-2032.
36. Han Y, Rand KA, Hazelett DJ, Ingles SA, Kittles RA, Strom SS, Rybicki BA, Nemesure B, Isaacs WB, Stanford JL et al: Prostate Cancer Susceptibility in Men of Af-rican Ancestry at 8q24. J Natl Cancer Inst 2016, 108(7).
37. Wojcik GL, Graff M, Nishimura KK, Tao R, Haessler J, Gignoux CR, Highland HM, Patel YM, Sorokin EP, Avery CL et al: Genetic analyses of diverse populations im-proves discovery for complex traits. Nature 2019, 570(7762):514-518.
38. Zou Z, Zhang J: Are Non-synonymous Transversions Generally More Deleterious than Non-synonymous Transitions? Mol Biol Evol 2021, 38(1):181-191.
39. Lyons DM, Lauring AS: Evidence for the Selective Basis of Transition-to-Transversion Substitution Bias in Two RNA Viruses. Mol Biol Evol 2017, 34(12):3205-3215.
40. Zhang J: Rates of conservative and radical non-synonymous nucleotide substitu-tions in mammalian nuclear genes. J Mol Evol 2000, 50(1):56-68.
41. Bittles A. H: Consanguinity and its rel-evance to clinical genetics. Clinical genetics, 2001, 60(2): 89-98.
42. El Mouzan, M. et al: Consanguinity and major genetic disorders in Saudi chil-dren: a community-based cross-sectional study. Annals of Saudi medicine, 2008, 28(3): 169-173.

Turki M. Sobahy received his Bachelors' degree in laboratory medicine from Umm Al-Qura University. He worked as medical technologist in molecular pathology. Then, he pursued and received his Masters' degree in bioinformatics from University of Sciences in Philadelphia. He currently works as research associate at King Faisal Specialist Hospital Research Center at Jeddah (KFSHRC-J). He is a junior bioinformatician and has contributed to the Saudi Human Genome Program for 3 years. His clinical laboratory experience and background drive his passion for developing new bioinformatics algorithms that would make a difference in clinical practice.

Olaa Motwalli's journey began by receiving her Bachelor's degree in Computer Science at King Abdul Aziz University in Jeddah. She later continued her Master's degree in Computer and Software Engineering at Widener University in the USA. Finally, She received her Ph.D. degree in computer science from King Abdullah University of Science and Technology (KAUST) in Saudi Arabia. Her hard work came to fruition when she became a consultant at Al Madinah Region Development Authority and is currently an assistant professor at the College of Computing and Informatics at Saudi Electronic University in Saudi Arabia. She is currently focusing on research in the area of bioinformatics and machine learning in the aims of continuously progressing and making a difference wherever she may be.

Meshari Alazmi received the BS, MS and Ph.D. degrees in computer science from University of Hail, University of Missouri, and King Abdullah University of Science and Technology (KAUST), Saudi Arabia, respectively. Currently, he is an assistant professor at the college of Computer Science and Engineering, University of Ha'il, Saudi Arabia. He currently serves as the vice dean for research and consulting studies institute at the University of Hail. His main areas of research include bioinformatics and machine learning and structural biology.