MTBAN: An Enhanced Variant Effect Predictor Based on a Deep Generative Model

Ha Young Kim
Korea Advanced Institute of Science and Technology
Woosung Jeon
Korea Advanced Institute of Science and Technology
Dongsup Kim (kds@kaist.ac.kr)
Korea Advanced Institute of Science and Technology

Research Article

Keywords: MTBAN, human genetic diseases, variant effect, deep generative model, prediction tool

Posted Date: June 28th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-649705/v1

MTBAN: An enhanced variant effect predictor based on a deep generative model

Ha Young Kim[1], Woosung Jeon[1] and Dongsup Kim[1,*]

[1]Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology, Daejeon 34141, Republic of Korea
*Corresponding author (kds@kaist.ac.kr)

ABSTRACT

The development of an accurate and reliable variant effect prediction tool is important for research in human genetic diseases. A large number of predictors have been developed towards this goal, yet many of these predictors suffer from the problem of data circularity. Here we present MTBAN (Mutation effect predictor using the Temporal convolutional network and the Born-Again Networks), a method for predicting the deleteriousness of variants. We apply a form of knowledge distillation technique known as the Born-Again Networks (BAN) to a previously developed deep autoregressive generative model, mutationTCN, to achieve an improved performance in variant effect prediction. As the model is fully unsupervised and trained only on the evolutionarily related sequences of a protein, it does not suffer from the problem of data circularity which is common across supervised predictors. When evaluated on a test dataset consisting of deleterious and benign human protein variants, MTBAN shows an outstanding predictive ability compared to other well-known variant effect predictors. We also offer a user-friendly web server to predict variant effects using MTBAN, freely accessible at http://mtban.kaist.ac.kr. To our knowledge, MTBAN is the first variant effect prediction tool based on a deep generative model that provides a user-friendly web server for the prediction of deleteriousness of variants.

Introduction

While recent sequencing technologies have resulted in a tremendous amount of sequence variant data, the identification of deleterious variants is still a difficult problem. Development of a reliable computational tool to predict the effects of sequence variants would aid in the treatment of many human genetic diseases. To achieve this goal, many predictors have been developed based on different approaches. Among these methods, supervised methods learn from labelled variant data consisting of known deleterious and benign variants, and many of them show good predictive ability. However, many supervised methods face the problem of data circularity, which can be divided into two types according to Grimm *et al.*[1] The *type I circularity* arises due to the overlap between training data and test data. The *type II circularity* occurs when all variants in a given gene are labelled either all deleterious or all benign, which results in the model predicting the same label for all variants in that gene. Previous studies[1-3] have suggested that this problem of data circularity can result in an inflation of the reported performances of many supervised predictors. On the other hand, unsupervised methods do not learn from labelled variant data and learn solely from the evolutionary information contained in multiple sequence alignments. A recent study which carried out an extensive comparison of variant effect predictors claimed that a class of unsupervised models, namely the deep generative model, is a promising area of research for variant effect prediction[3].

Here, we introduce MTBAN (Mutation effect predictor using the Temporal convolutional network and the Born-Again Networks), an enhanced method to predict the deleteriousness of single amino acid variants. We previously developed a method called mutationTCN[4] based on a deep autoregressive generative model, and showed that it demonstrates state-of-the-art performances on the prediction of functional effects of variants. In this work, we apply a knowledge distillation technique called the Born-Again Networks (BAN)[5] to the mutationTCN model and develop an improved model called MTBAN. In machine learning, knowledge distillation is a process involving the transfer of knowledge learned from one machine learning model to another. In this scheme, the former model is referred to as the

"teacher network" and the latter is referred to as the "student network." Using the Born-Again Networks allows the student network to achieve an improved predictive power compared to the teacher network. When evaluated on human variant datasets with deleterious and benign variants, MTBAN shows superior predictive performances compared to other variant effect predictors. Our model is fully unsupervised and is not dependent on labelled data for training. This gives the model advantage over supervised predictors, for which data circularity is an inherent problem. We also offer a freely accessible web server for using MTBAN for variant effect prediction.

Methods

MTBAN model

We previously developed a deep autoregressive generative model for variant effect prediction, called mutationTCN[4]. As it is a generative model, it is trained by maximizing the likelihood of the training data, which is the evolutionarily related sequences of a given protein. The model is thus optimized by minimizing the negative log likelihood between the input sequence and the predicted output. After training, the model can predict the probability of observing a given protein sequence under the parameters of the trained model. The deep autoregressive generative model is implemented using the temporal convolutional network architecture[6], and is composed of an embedding layer followed by a series of dilated causal convolution layers, an attention layer, and a fully connected layer (Figure 1). We showed that this model can effectively capture information from evolutionarily related sequences and use this information to predict the functional effects of variations in a sequence[4].

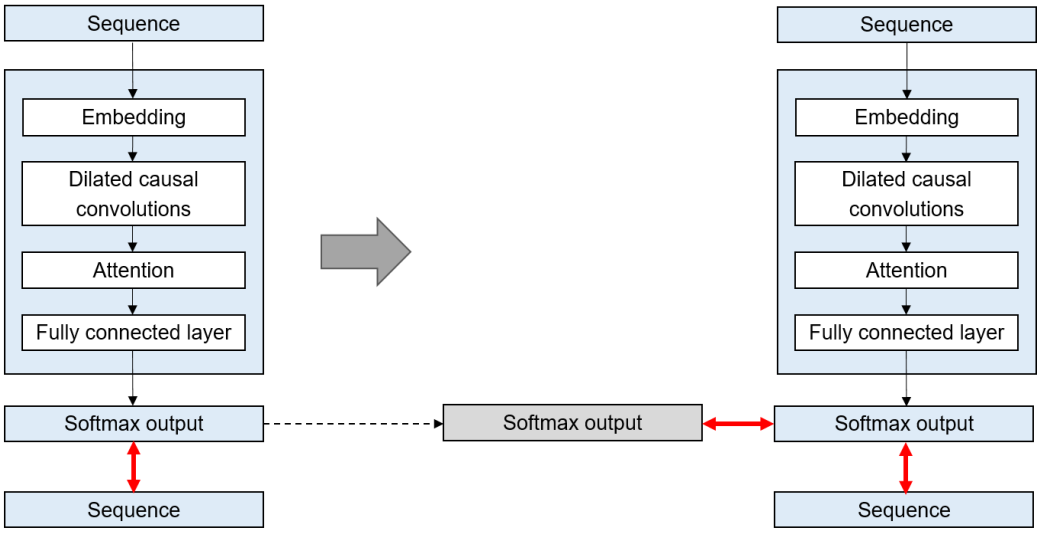

Figure 1. MTBAN model structure. We implemented BAN with mutationTCN as both the teacher and the student network. In the first step, only the teacher network is trained, with the loss function being the label loss, which refers to the cross entropy loss between the input sequence and the softmax output distribution of the teacher network. In the second step, only the student network is trained, with the loss being the sum of the label loss and the teacher loss. The teacher loss refers to the cross entropy loss between the softmax output distribution of the student network and the softmax output distribution of the teacher network.

MTBAN combines this model with a knowledge distillation technique in machine learning, known as the Born-Again Networks (BAN)[5]. Knowledge distillation is a process of model compression which involves transferring the knowledge from a teacher network to a student network with a smaller capacity[7]. This allows for the reduction of model size, while maintaining similar predictive power as the original model. In the setting of BAN, the student network is of the same capacity as the teacher network, which enables the student network to outperform the teacher network[5]. We found that the BAN framework in which both the teacher and the student network is implemented with mutationTCN outperforms the original mutationTCN model. The model structure of MTBAN is shown in Figure 1. In the first step, only the teacher network is trained, with the loss function being the *label loss*, which refers to

the cross entropy loss between the input sequence and the softmax output distribution of the teacher network. In the next step, only the student network is trained, with the loss being the sum of the *label loss* and the *teacher loss*. Here, the *teacher loss* refers to the cross entropy loss between the softmax output distribution of the student network and the softmax output distribution of the teacher network. This softmax output probability distribution p_i can be expressed as follows:

$$p_i = \frac{\exp(\frac{z_i}{T})}{\sum_j \exp(\frac{z_j}{T})}$$

where z_i is the logit computed for each class and T is the temperature parameter[7]. Using higher temperatures leads to more "softened" output distributions. In our implementation, we used a temperature of 4. By training the student network to learn the softened outputs of the teacher network, the student network can learn the knowledge that was previously learned by the teacher network. In our implementation, both teacher and student networks are trained for 500,000 iterations using the mini-batches with the size of 128. For both teacher and student networks, the learning rate is set to 0.001 when the number of training iterations is smaller than 3,000, and 0.0001 when it is greater than 3,000.

We computed the predictions of MTBAN for a total of 1,032 human protein alignments provided by Hopf *et al.*[8] These pre-computed predictions on the Hopf dataset were saved and used for obtaining the predictions of MTBAN on human protein variants.

Model Outputs

For a given variant, the model outputs the log probability score, the z-score, the probability of deleteriousness, and the predicted label. First, the log probability score is given by the following:

$$\log \frac{p(x^{mutant}|\theta)}{p(x^{wild-type}|\theta)}$$

where $p(x^{mutant}|\theta)$ and $p(x^{wild-type}|\theta)$ are the probability assigned to the mutant sequence and the wild-type sequence, respectively, by the generative model with parameters θ. The log probability score is easily computed as the negative of the loss, as the model loss function is the negative log likelihood[4]. The smaller the score, the more likely the variant has a deleterious effect. Second, the z-score is computed by normalizing the distribution of log probability scores for all possible missense variants against the target sequence of a protein. This normalization process is done due to the variations in the score distributions across different proteins. Third, the probability of deleteriousness for each variant, ranging from 0 to 1, is computed. This is determined from the set of variants in the Humsavar database (release 03/2020)[9] which overlap with our pre-computed model predictions for the Hopf dataset, which are 1221 deleterious and 1221 benign variants. We obtained the z-score distribution for this set of variants, divided the distribution into equal-length z-score intervals, and calculated the proportion of deleterious variants in each z-score interval. Finally, using the same z-score intervals, we determined a z-score threshold which maximizes the classification accuracy (Supplementary Fig. S1). This threshold is used to assign a predicted label, either deleterious or benign, to a given variant.

Evaluation Datasets

To evaluate the ability of the model to classify human protein variants as deleterious or benign, we created a test dataset by combining the variant data from datasets used by Grimm *et al*.[1] and Mahmood *et al*.[2] Details regarding the datasets can be found in Table 1. We used the HumVar dataset from Grimm *et al*., which contains human protein variants that are known to be disease-causing or neutral[1]. Also, we used the UniFun, BRCA1-DMS, and TP53-TA datasets from Mahmood *et al*., which contain deleterious and benign protein variants determined from direct in vitro functional assays, such as the deep mutational scanning experiment[2]. Mahmood *et al*. pointed out that commonly used disease-related variant datasets often overlap with the training data used by supervised predictors[2]. Because of this reason,

they created the functionally determined variant datasets in order to avoid the problem of data circularity and establish an independent test set for benchmarking[2]. Another study[3] also supports this claim and uses the data from deep mutational scanning experiments to benchmark a large number of variant effect predictors. Also, it is reported that the Critical Assessment of Genome Interpretation (CAGI), which aims to perform an unbiased assessment of variant effect predictors, uses data from deep mutational scanning experiments as part of their benchmark dataset[10]. Therefore, we use the functionally determined variant data from Mahmood *et al.* in addition to the disease-related variant data for comparing MTBAN with other predictors.

Reference	Dataset	Description	ND	NB
Grimm et al., 2015	HumVar	Disease-causing mutations from UniProtKB and common single nucleotide polymorphisms with major allele frequency > 1%[1]	772	772
	Total		772	772
Mahmood et al., 2017	UniFun	Deleterious and benign variants in UniProt which are derived from functional assays[2]	18	18
	BRCA1-DMS	Deleterious and benign variants derived from deep mutational scanning experiment measuring homology-directed DNA repair and tumor suppression activity[2]	41	41
	TP53-TA	Deleterious and benign variants derived from transactivation assay[2]	413	413
	Total		472	472
Total			1244	1244

Table 1. Test datasets used and the number of deleterious and benign variants for each dataset used for evaluation. ND stands for the number of deleterious variants, and NB stands for the number of benign variants.

We compared the performance of our model with other commonly used variant effect predictors, SIFT[11], PolyPhen-2[12], MutationAssessor[13], fathmm-MKL[14], MPC[15], GenoCanyon[16],

phastCons[17], DANN[18], GERP++[19], and phyloP[20]. The predictions of these tools on the test dataset were obtained from dbNSFP[21] via the Ensembl variant effect predictor[22].

We found variants among these datasets for which predictions exist in our pre-computed Hopf dataset, and used them for comparison with other methods. Since the number of deleterious variants was significantly larger than that of benign variants, we randomly selected variants from the deleterious variant data to match the data size of the deleterious variants and the benign variants. This resulted in a balanced test set consisting of 1244 deleterious and 1244 benign variants in total.

Evaluation Criteria

The following metrics were used for evaluating the classification ability of the variant effect predictors: ROC AUC (Receiver Operating Characteristic Area Under Curve), PR AUC (Precision-Recall Area Under Curve), accuracy, Matthews Correlation Coefficient (MCC), precision, specificity, sensitivity, F-score, and Negative Predictive Value (NPV). ROC-AUC and PR-AUC were calculated using z-scores, and other evaluation metrics were calculated using the predicted label. The following equations were used for computing the evaluation metrics:

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

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

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

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

$$\text{Sensitivity (Recall)} = \frac{TP}{TP + FN}$$

$$\mathrm{F-score} = 2\frac{Precision \times Recall}{Precision + Recall}$$

$$\text{Negative Predictive Value (NPV)} = \frac{TN}{TN + FN}$$

where TP, TN, FP, and FN are the number of true positives, true negatives, false positives, and false negatives, respectively.

Results

Evaluation on human protein variant datasets

We assessed MTBAN and other variant effect predictors on the task of classifying human protein variants as deleterious or benign. As described in Methods section, our test dataset combines the disease-associated variants from Grimm *et al.*[1] and functionally determined variants from Mahmood *et al*.[2], resulting in a total of 1244 deleterious and 1244 benign variants. A total of 11 predictors were compared in terms of ROC AUC and PR AUC. Among these 11 predictors, five predictors were compared in terms of accuracy, MCC, precision, specificity, sensitivity, F-score, and NPV, since other predictors only generated scores and not the predicted label. Our model achieved an ROC AUC of 0.876 and a PR AUC of 0.87 (Figure 2, Table 2), which were both the highest among 11 different variant effect predictors. It even outperformed the predictor PolyPhen-2, whose training dataset is known to have overlapping variants with the dataset used by Grimm *et al*.[1] Also, MTBAN achieved the highest accuracy, MCC, precision, specificity, and F-score, among all compared variant effect predictors. In addition, our model demonstrates a good balance between specificity and sensitivity, unlike fathmm-MKL which demonstrates good performance in only one of the two measures. When compared to other unsupervised predictors, GenoCanyon, phastCons, and MutationAssessor, our model shows a higher performance across all evaluation metrics.

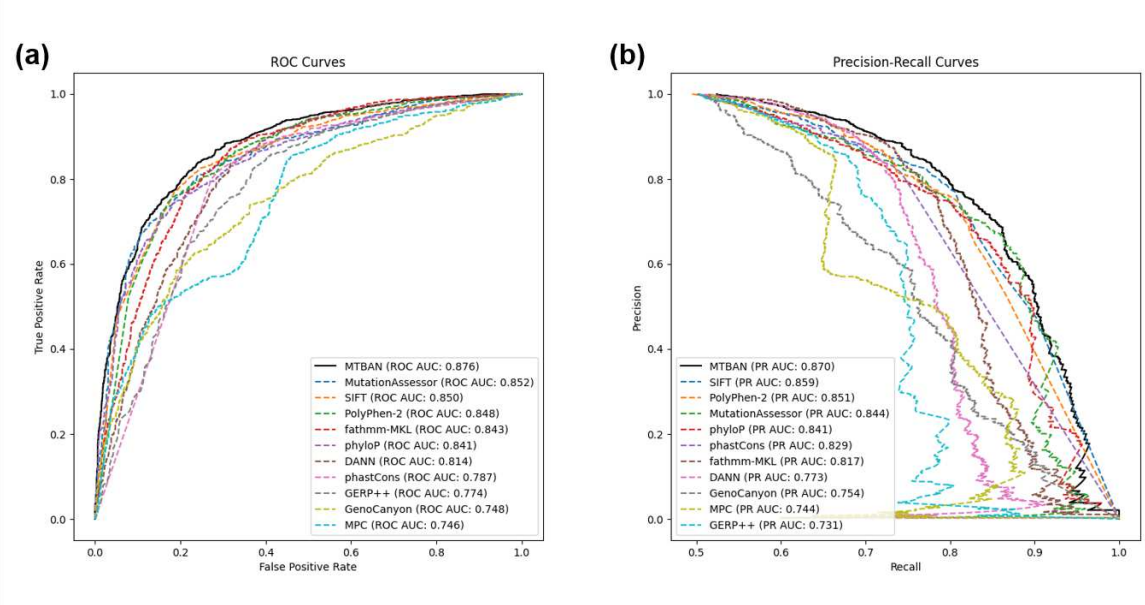

Figure 2. ROC Curves and Precision-Recall Curves for MTBAN and other predictors on the test dataset. (a) MTBAN achieved a ROC AUC (Receiver Operating Characteristic Area Under Curve) of 0.876, which is the highest among 11 variant effect predictors. (b) MTBAN achieved a PR AUC (Precision-Recall Area Under Curve) of 0.87, which is the highest among 11 variant effect predictors.

Predictor	ROC AUC	PR AUC	Accuracy	MCC	Precision	Specificity	Sensitivity	F-score	NPV
MTBAN	**0.876**	**0.87**	**0.785**	**0.583**	**0.737**	**0.684**	0.887	**0.805**	0.858
SIFT	0.85	0.859	0.764	0.542	0.714	0.645	0.881	0.789	0.844
PolyPhen-2	0.848	0.851	0.76	0.541	0.701	0.625	0.896	0.787	0.86
MutationAssessor	0.852	0.844	0.762	0.534	0.718	0.669	0.856	0.781	0.825
fathmm-MKL	0.843	0.817	0.734	0.504	0.671	0.544	**0.922**	0.777	**0.874**
MPC	0.746	0.744							
DANN	0.814	0.773							
phyloP	0.841	0.841							
phastCons	0.787	0.829							
GenoCanyon	0.748	0.754							
GERP++	0.774	0.731							

**ROC-AUC, Receiver Operating Characteristic Area Under Curve; PR-AUC, Precision-Recall Area Under Curve; MCC, Matthews Correlation Coefficient; NPV, Negative Predictive Value

Table 2. Performances of MTBAN and other predictors on the test dataset consisting of 1244 deleterious and 1244 benign variants. Note that MPC, DANN, phyloP, phastCons, GenoCanyon, and GERP++ only generated scores and not the predicted label.

In addition, we conducted further assessment using only the disease-associated variant data from Grimm *et al.*[1], and using only the functionally determined variant data from Mahmood *et al*.[2] When tested on the data from Grimm *et al.* consisting of 772 deleterious and 772 benign variants, our model achieved the highest ROC AUC, accuracy, MCC, and F-score (Supplementary Table S1). Also, when tested on the data from Mahmood *et al*. consisting of 472 deleterious and 472 benign variants, our model achieved the highest accuracy, MCC, precision, specificity, and F-score (Supplementary Table S2). Overall, MTBAN shows an outstanding classification ability in both disease-associated variant data and functional assay-derived variant data.

Web Server

We offer a user-friendly web server which predicts variant effects using MTBAN (Supplementary Fig. S2). The server takes in as input a protein UniProt accession and a list of amino acid variants. Upon receiving input, it determines the target protein sequence region, and checks if pre-computed predictions exist for the given variants. If they exist, the server immediately returns predictions to the user. Otherwise, it checks if a multiple sequence alignment of the target protein sequence region is present in the database. If an alignment is present, it uses that alignment for subsequent computations. If an alignment is not present, it generates one using a profile HMM homology search tool[23] and saves it in the database. During the computation, alignment columns that have more than 30% gaps are dropped. If some of the input variants belong to these un-aligned columns in the alignment, those variants are excluded from prediction and are indicated in the results. The next step is the computation of sequence weights, based on the similarity of sequences in the alignment. This step is included to reduce any sequence bias present in the multiple sequence alignment[4]. Afterwards, the prediction model is trained, and the server returns predictions to the user. After job processing, the predictions are saved so that the server can immediately return the results when the same set of mutations are later submitted as input. In the web server implementation,

due to time constraints, the MTBAN teacher network and student network are both trained for 200,000 iterations, with learning rate 0.001.

Discussion

Here, we have introduced MTBAN, an improved method for predicting the deleteriousness of single amino acid variants. As demonstrated in our previous work[4], the deep autoregressive generative model is a powerful tool for learning the distribution underlying the evolutionarily related sequences of a protein and predicting the effects of variations in a sequence. Combining the deep autoregressive generative model with a knowledge distillation method known as the Born-Again Networks further improves the predictive power of the model, by transferring the knowledge learned by the model to the second model of the same capacity. We conducted an assessment using the test set combining the disease-related variants from Grimm *et al.*[1] and the functionally determined variants from Mahmood *et al.*[2], and further assessment using each of the two variant sets. In all cases, MTBAN consistently shows outstanding predictive ability compared to other prediction tools. The results indicate that MTBAN is a reliable method for predicting the deleteriousness of human protein variants.

Previous works[1-3] have pointed out concerns regarding the problem of data circularity in many supervised predictors, which can lead to an inflation of the reported performances of these tools. Due to the fully unsupervised nature of MTBAN, it is not hindered by the problem of data circularity and can be considered to have higher generality compared to supervised models. Moreover, while we only considered human protein variants in this work, it is possible to predict the effects of protein variants in any other species if a multiple sequence alignment is available.

A potential limitation of MTBAN is that the training time is longer compared to mutationTCN alone for prediction. Although this model takes a longer time to train, it shows a

higher predictive performance compared to the previous model. Another potential limitation of this model is that it can only make predictions for variants which correspond to the conserved positions in the multiple sequence alignment of a protein. However, when we analyzed all of the 9,935 human protein multiple sequence alignments in the Hopf dataset, approximately 88% of the target sequences were conserved, which is a considerably large proportion.

The results of our work show that the deep generative model is a powerful tool for predicting the effects of sequence variations. We expect that deep generative models will continue to play an important role in discovering the effects of genetic variants. In addition, to our knowledge, MTBAN is the first variant effect prediction tool based on a deep generative model that provides a user-friendly web server for the prediction of deleteriousness of variants. This method is expected to be a useful tool for the prioritization and identification of variants involved in human genetic diseases.

Data availability statement

The datasets generated during and/or analysed during the current study are available at https://github.com/ha01994/MTBAN.

References

1 Grimm, D. G. *et al.* The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Human mutation* **36**, 513-523 (2015).

2 Mahmood, K. *et al.* Variant effect prediction tools assessed using independent, functional assay-based datasets: implications for discovery and diagnostics. *Human genomics* **11**, 1-8 (2017).

3 Livesey, B. J. & Marsh, J. A. Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. *Molecular systems biology* **16**, e9380 (2020).

4 Kim, H. Y. & Kim, D. Prediction of mutation effects using a deep temporal convolutional network. *Bioinformatics* **36**, 2047-2052 (2020).

5 Furlanello, T., Lipton, Z. C., Tschannen, M., Itti, L. & Anandkumar, A. Born again neural networks. *arXiv preprint arXiv:1805.04770* (2018).

6 Bai, S., Kolter, J. Z. & Koltun, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271* (2018).

7 Hinton, G., Vinyals, O. & Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

8 Hopf, T. A. *et al.* Mutation effects predicted from sequence co-variation. *Nature biotechnology* **35**, 128-135 (2017).

9 Consortium, U. UniProt: a worldwide hub of protein knowledge. *Nucleic acids research* **47**, D506-D515 (2019).

10 Hoskins, R. A. *et al.* Reports from CAGI: the critical assessment of genome interpretation. *Human mutation* **38**, 1039 (2017).

11 Sim, N.-L. *et al.* SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic acids research* **40**, W452-W457 (2012).

12 Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature methods* **7**, 248-249 (2010).

13 Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic acids research* **39**, e118-e118 (2011).

14 Shihab, H. A. *et al.* An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* **31**, 1536-1543 (2015).

15 Samocha, K. E. *et al.* Regional missense constraint improves variant deleteriousness prediction. *BioRxiv*, 148353 (2017).

16 Lu, Q. *et al.* A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. *Scientific reports* **5**, 1-13 (2015).

17 Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research* **15**, 1034-1050 (2005).

18 Quang, D., Chen, Y. & Xie, X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **31**, 761-763 (2015).

19 Davydov, E. V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* **6**, e1001025 (2010).

20 Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research* **20**, 110-121 (2010).

21 Liu, X., Li, C., Mou, C., Dong, Y. & Tu, Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome medicine* **12**, 1-8 (2020).

22 McLaren, W. *et al.* The ensembl variant effect predictor. *Genome biology* **17**, 122 (2016).

23 Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput Biol* **7**, e1002195 (2011).

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grants (2017M3A9C4065952, 2019R1A2C1007951) funded by the Korea Government (MSIT).

Author contributions

D.K. conceived the experiment(s), H.K. and W.J. developed the software, H.K. conducted the experiment(s), H.K. and D.K. analyzed the results, H.K. wrote the paper. All authors reviewed the manuscript.

Additional information

Competing interests

The author(s) declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SciRepHYKsupplfile.docx