MetaRNN: Differentiating Rare Pathogenic and Rare Benign Missense SNVs and InDels Using Deep Learning

Chang Li[1], Degui Zhi[2], Kai Wang[3], Xiaoming Liu[1]*

[1]USF Genomics & College of Public Health, University of South Florida, Tampa, FL, USA
[2]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA
[3]Children's Hospital of Philadelphia & Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

*Correspondence to Xiaoming Liu, Ph.D., USF Genomics & College of Public Health, University of South Florida, 3720 Spectrum Boulevard, Suite 304, Tampa, FL 33612, USA; Phone: 1-813-974-9865; E-mail: xiaomingliu@usf.edu

Abstract

We present the pathogenicity prediction models MetaRNN and MetaRNN-indel to help identify and prioritize rare nonsynonymous single nucleotide variants (nsSNVs) and non-frameshift insertion/deletions (nfINDELs) using deep learning and context annotations. Employing independent test datasets, we demonstrate that these new models outperform state-of-the-art competitors and achieve a more interpretable score distribution. MetaRNN executables and precomputed scores are available at http://www.liulab.science/MetaRNN.

Main Text

Separating rare pathogenic and rare benign variants is an essential task in exome-sequencing-based Mendelian disease studies. This task is especially challenging for mutations that can cause changes in amino acid (AA) sequences, namely, nonsynonymous single nucleotide mutations (nsSNVs) and non-frameshift insertion/deletions (nfINDELs), while being exempt from definite and severe consequences, such as stop-gain mutations. Because experimentally validating the effects of these mutations is time-consuming and costly, computational predictive methods have

been developed for this purpose[1–10]. However, these methods generally suffer from two major limitations. First, most of these methods employ models trained with rare pathogenic variants and common benign variants, which cause them to be less optimized for separating rare pathogenic and rare benign variants. Second, most of the methods provide prediction scores for only nsSNVs or incomparable scores for nsSNVs and nfINDELs separately, making it infeasible to use these scores as weights in an integrated (nsSNV+nfINDELs) burden test for genotype-phenotype association analysis. To overcome these limitations, we developed the MetaRNN and MetaRNN-indel models, which enable users to easily annotate both nsSNVs and nfINDELs. MetaRNN and MetaRNN-indel integrated information from 16 high-level pathogenicity prediction scores, such as CADD[1] and M-CAP[5], 8 conservation scores and allele frequency information from the 1000 Genomes Project (1000GP)[11], ExAC[12], and gnomAD[13] (full list in **Supplementary Note**). The integrated information for all target mutations and their flanking base pairs, defined as a ±1 AA window, was extracted by a deep recurrent neural network (RNN) with multiple stacked bidirectional gated recurrent units[14] (GRUs) (**Supplementary Note**). We trained the RNN model with 26,517 rare nsSNVs and 2,057 rare nfINDELs from ClinVar[15] up to release 20190102 (**Supplementary Note**). The final prediction is the likelihood of a mutation being pathogenic. The model structure enabled us to employ flanking sequence annotations to improve target variant prediction. We found that the GRU-based models consistently outperformed models using only annotations of target mutations (**S. Figure 3**). The final output scores of both MetaRNN and MetaRNN-indel range from 0 to 1, and the score can be interpreted as the probability of the observed mutation being pathogenic. Since the prediction scores from MetaRNN and MetaRNN-indel share the same range and were constructed using the ClinVar dataset, predictions from these two models are comparable with their pathogenicity.

To evaluate the performance of the proposed models, we compared multiple state-of-the-art computational methods in the interpretation of sequence variants using independent test datasets. We observed that MetaRNN outperformed other competitors across all the test datasets, including two rare nsSNV test sets (RNTS) that was composed of rare pathogenic ClinVar nsSNVs after release 20190102 and location-matched rare nsSNVs from gnomAD, ExAC and the 1000 Genomes Project (n = 12,406 and n = 11,540) (selected comparisons with 8 tools in **Figure 1A**; all comparisons with 24 tools in **S. Figure 5** and **S. Figure 6**), two ClinVar only test sets composed of rare ClinVar pathogenic and rare benign nsSNVs after release 20190102 (n = 3,917 and n = 9,285) (**S. Figure 7** and **S. Figure 8**), an all-allele-frequency test dataset composed of both rare and common ClinVar nsSNVs after release 20190102 (n = 29,924) (**S. Figure 9**), and a functional test dataset for TP53 mutations (n = 824) (**S. Figure 10**). The results highlighted MetaRNN's increased ability relative to those of the other methods to separate not only rare pathogenic mutations from rare benign ones but also mutations with various degrees of

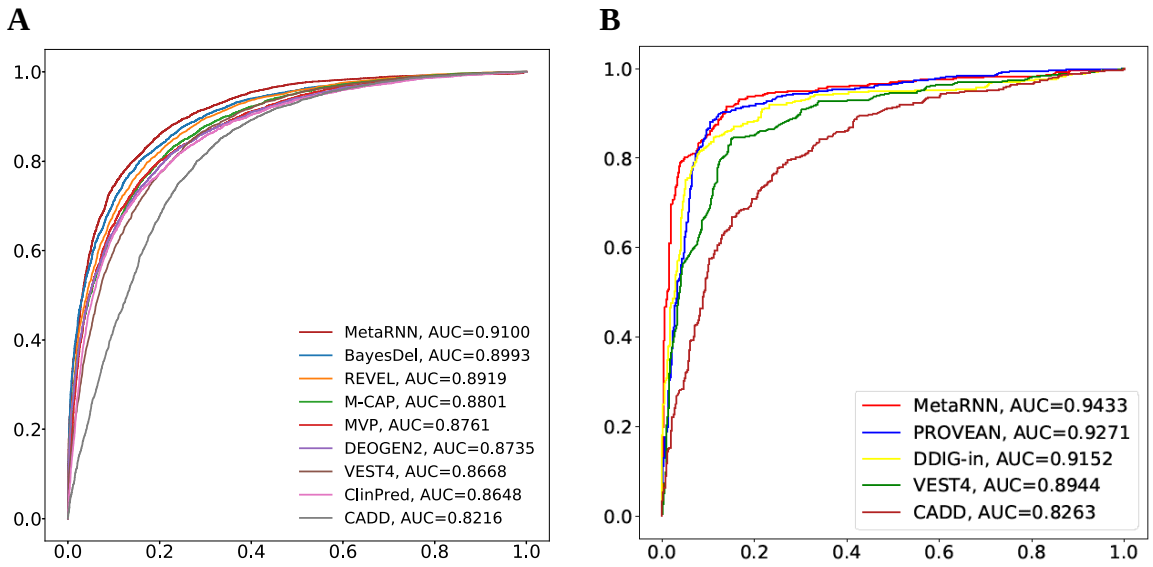

Figure 1. Comparisons of MetaRNN and MetaRNN-indel with other prediction tools. **A**: Performance comparison of MetaRNN and 8 other nsSNV prediction tools using the rare nsSNV test set (RNTS). **B**: Performance comparison of MetaRNN-indel and other nfINDEL prediction tools using ClinVar nfINDELs.

functional importance. With a test dataset that was composed of rare pathogenic ClinVar nfINDELs after release 20190102 (n = 989), MetaRNN-indel outperformed all competitors in ranking nfINDELs (**Figure 1B**), including two methods, VEST[10] and CADD, that showed good performance for nsSNVs. These results indicate that our training framework for MetaRNN and MetaRNN-indel consistently outperforms other methods concerning both nsSNVs and nfINDELs.

To explore the interpretability and usability of the proposed models, we first predicted scores for all nsSNVs in ClinVar that showed conflicting clinical interpretations (n = 20,337). These nsSNVs represent an important class, that is, variant of unknown significance (VUS) according to the ACMG-AMP guidelines[16], and the ability to distinguish and interpret VUSs is important to the clinical application of the proposed score to help improve the diagnostic rate. A score that shows sufficient dispersion enables further identification of relevant candidate variants. Additionally, these conflicting VUS variants are variants of interest with some evidence of being either pathogenic or benign. Among these variants, 15,788 (77.6%) showed conflicting interpretations between benign/likely benign and unknown significance (benign conflict group), whereas 4,110 (20.2%) showed conflicting interpretations between pathogenic/likely pathogenic and unknown significance (pathogenic conflict group). Based on the fact that the benign conflict group had approximately 4 times more variants than the pathogenic conflict group, we expect that variant prediction tools should be able to reflect this observation. MetaRNN's predictions showed a score distribution that fit these assumptions (**Figure 2A**), which peaked at the extremes of its score range and had approximately 4 times more extreme benign predictions than extreme pathogenic predictions. Compared with other scores, which either showed little change in the distribution across their predictions (e.g. CADD, VEST[17], REVEL[4]) or potentially

underestimated the proportion of VUSs at the extremes (BayesDel[9]), the proposed MetaRNN scores exhibited greater interpretability for sequencing data and VUSs. Additionally, we obtained nsSNVs observed in three large-scale population studies, namely, 1000GP, gnomAD, and ExAC, with at least two observed minor allele counts to remove the majority of potentially pathogenic variants. The distributions of the predicted scores are shown in **Figure 2B**. As expected, all methods showed a decreasing proportion of pathogenic variants. However, MetaRNN had a substantially lower proportion of variants predicted to be not benign (score > 0.5). This feature of MetaRNN can increase the power of genome-wide association studies by removing more genuinely benign variants. The proposed models are expected to be useful across various settings, ranging from filtering candidate SNVs for rare-variant association analysis to supporting disease gene identification with increased accuracy and interpretability. For example, MetaRNN and MetaRNN-indel scores are well calibrated to be combined for an integrated (nsSNV+nfINDELs) rare-variant burden test for genotype-phenotype association.

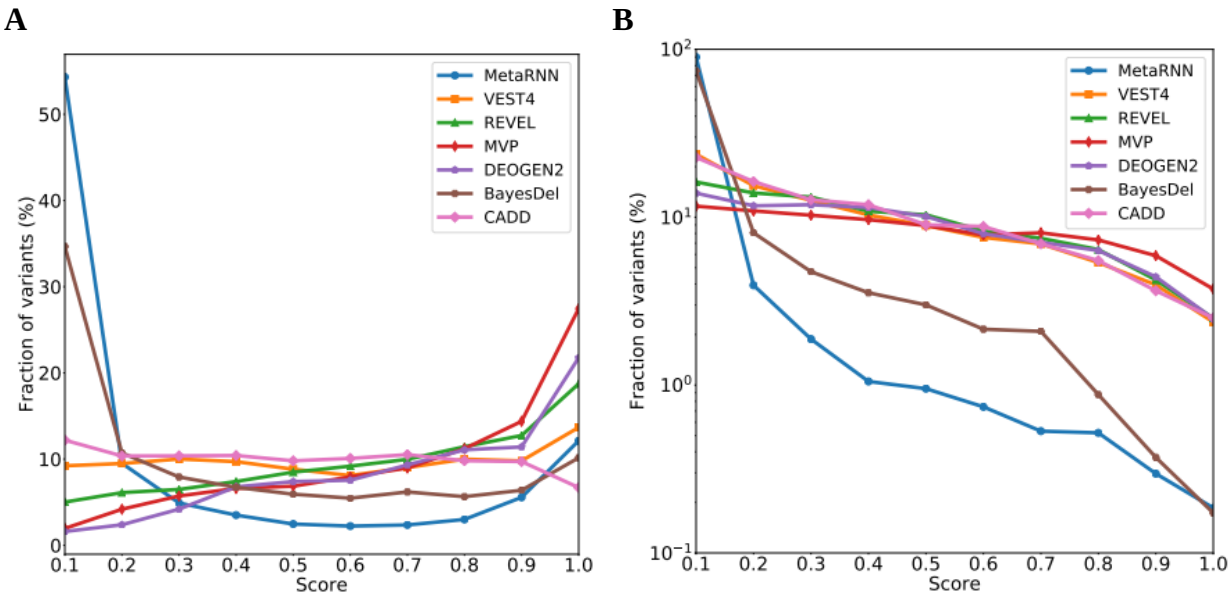

Figure 2. MetaRNN score distributions. **A**: Score distribution for ClinVar variants of unknown significance (VUSs). **B**: Score distribution for nsSNVs observed in the 1000 Genomes Project, gnomAD and ExAC with allele count > 1.

We provide predictions for all potential nsSNVs (~86 million) in the dbNSFP[18,19] database for rapid and user-friendly analysis. We provide a stand-alone Linux executable for the Linux environment for nfINDEL (and nsSNV) predictions. The executable takes a standard VCF file as input and provides variant pathogenicity scores in a transcript-specific manner as output (supported by ANNOVAR[20]). The average prediction time for a single insertion/deletion is approximately 0.2 seconds, which can support timely large-scale predictions.

We believe that with the improvements in prediction accuracy, score interpretability, and usability exhibited by our new method, it may provide more accessible and accurate annotation of rare VUSs in exome-sequencing-based Mendelian disease studies and integrated (nsSNV+nfINDELs) burden tests for common disease studies.

References

1. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310–315 (2014).
2. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE* **7**, e46688 (2012).
3. Hu, J. & Ng, P. C. Predicting the effects of frameshifting indels. *Genome Biology* **13**, R9 (2012).
4. Ioannidis, N. M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *The American Journal of Human Genetics* **99**, 877–885 (2016).
5. Jagadeesh, K. A. *et al.* M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat Genet* **48**, 1581–1586 (2016).
6. Qi, H. *et al.* MVP predicts the pathogenicity of missense variants by deep learning. *Nature Communications* **12**, 510 (2021).
7. Raimondi, D. *et al.* DEOGEN2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. *Nucleic Acids Res.* **45**, W201–W206 (2017).
8. Zhao, H. *et al.* DDIG-in: discriminating between disease-associated and neutral non-frameshifting micro-indels. *Genome Biology* **14**, R23 (2013).
9. Feng, B.-J. PERCH: A Unified Framework for Disease Gene Prioritization. *Human Mutation* **38**, 243–251 (2017).
10. Douville, C. *et al.* Assessing the Pathogenicity of Insertion and Deletion Variants with the Variant Effect Scoring Tool (VEST-Indel). *Hum Mutat* **37**, 28–35 (2016).
11. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

12. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).

13. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

14. Cho, K., van Merrienboer, B., Bahdanau, D. & Bengio, Y. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *arXiv:1409.1259 [cs, stat]* (2014).

15. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).

16. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).

17. Carter, H., Douville, C., Stenson, P. D., Cooper, D. N. & Karchin, R. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* **14 Suppl 3**, S3 (2013).

18. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. *Human Mutation* **32**, 894–899 (2011).

19. 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**, 103 (2020).

20. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).